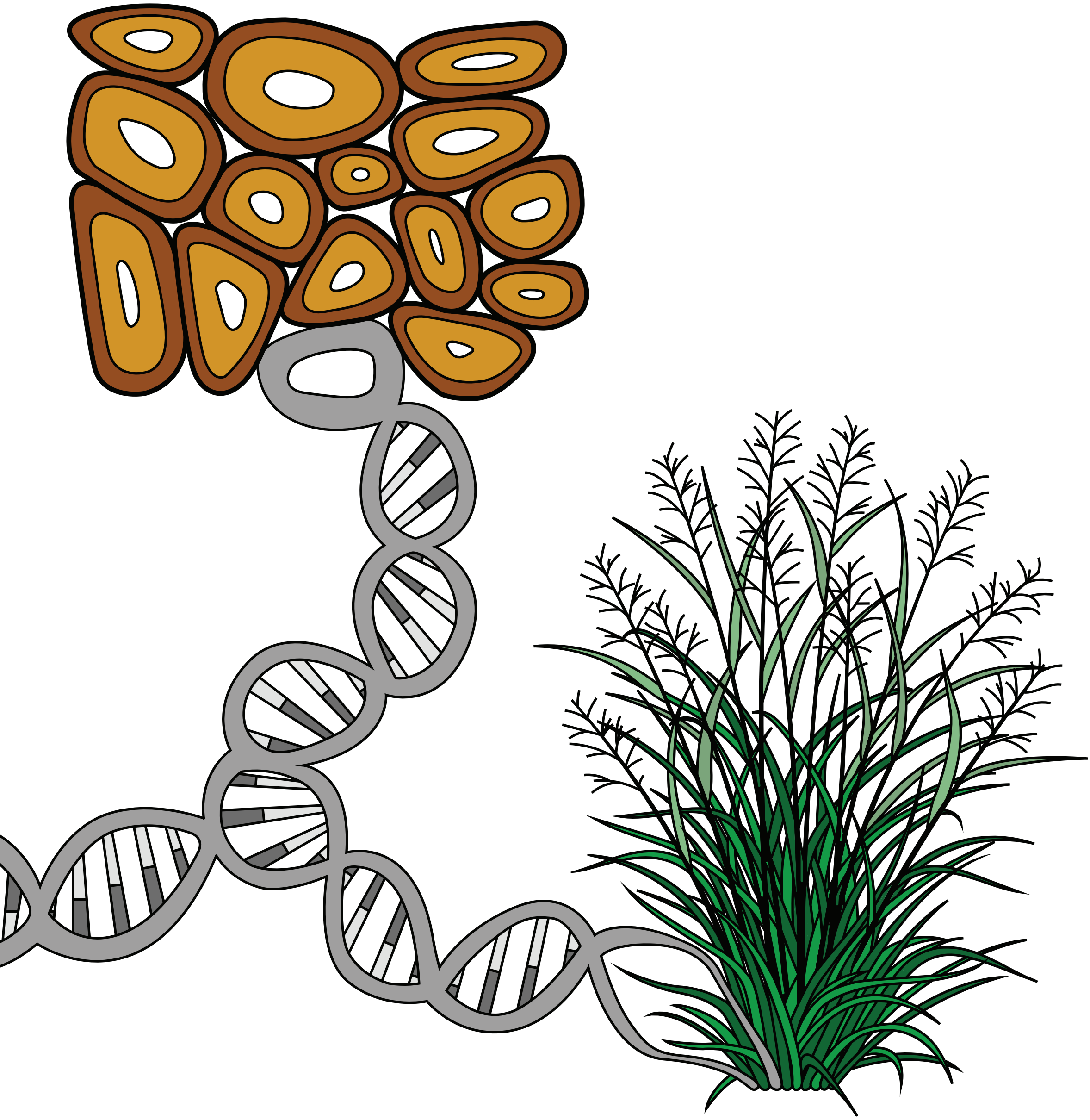
Plant cell walls: from evolutionary genomics to novel breeding tools



Francesco Pancaldi

Propositions

1. Secondary cell wall *Cellulose synthase* genes are derived from the primary cell wall ones. (this thesis)

2. Syntenic conservation of traits' genetic determinants across species allows to shorten crops breeding cycles. (this thesis)

3. The failure of schools to promote scientific thinking is at the basis of science distrust.

4. Speculation is the fuel for scientific discoveries with impact.

5. Electric vehicles keep the transportation sector unsustainable.

6. Social media are anti-social platforms.

Propositions belonging to the thesis, entitled

"Plant cell walls: from evolutionary genomics to novel breeding tools"

Francesco Pancaldi

Wageningen, 12th June 2023

Plant cell walls:
from evolutionary genomics to novel breeding tools

Francesco Pancaldi

Thesis committee

Promotors

Prof. Dr Luisa M. Trindade
Personal Chair, Plant Breeding
Wageningen University & Research

Prof. Dr M. Eric Schranz
Professor of Biosystematics
Wageningen University & Research

Co-promotor

Dr Eibertus N. van Loo
Senior Researcher, Plant Breeding
Wageningen University & Research

Other members

Dr E. Alexopoulou, Centre for Renewable Energy Sources and Saving, Pikermi Attiki, Greece
Prof. Dr R. Offringa, Leiden University
Prof. Dr P. Ulvskov, University of Copenhagen, Denmark
Prof. Dr B. J. Zwaan, Wageningen University & Research

This research was conducted under the auspices of the Graduate School Experimental Plant Sciences (EPS)

Plant cell walls: from evolutionary genomics to novel breeding tools

Francesco Pancaldi

Thesis
Submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Monday 12 June 2023
at 11 a.m. in the Omnia Auditorium.

Francesco Pancaldi
Plant cell walls: from evolutionary genomics to novel breeding tools,
248 pages

PhD thesis, Wageningen University, Wageningen, the Netherlands (2023)
With references, with summary in English

ISBN 978-94-6447-612-5
DOI https://doi.org/10.18174/588954

Table of contents

Chapter 1 General Introduction 7

Chapter 2 Genomic architecture and evolution of the *Cellulose synthase* gene superfamily as revealed by phylogenomic analysis 21

Chapter 3 Highly differentiated genomic properties underpin the different cell walls of Poaceae and eudicots 51

Chapter 4 Marginal lands to grow novel bio-based crops: a plant breeding perspective 85

Chapter 5 Detection and analysis of syntenic quantitative trait loci controlling cell wall quality in angiosperms 115

Chapter 6 Syntenic cell wall QTLs as versatile breeding tools: intra-specific allelic variability and predictability of biomass quality loci in target plant species 147

Chapter 7 General Discussion 183

References 202

Summary 232

Riassunto 236

Acknowledgements 240

About the author 244

List of publications 245

Education statement 246

Chapter 1

General Introduction

1 Reshaping agriculture toward a bio-based economy

At the peak of Middle Ages, the artist Ambrogio Lorenzetti (Siena, Italy, 1317-1348) painted his masterpiece "Allegory and Effects of the Good and Bad Government", showing that the downfall and prosperity of human societies depends on irresponsible and wise decisions, respectively, in the management of societies themselves (Argan, 1968). Seven hundred years later, humankind has to take some of the most crucial decisions ever to solve the grand challenges of our time: climate change, resource depletion, pollution. The "wise" decision that has been put foreword to solve these issues is to guide human systems toward environmental, economic, and social sustainability (Meadows et al., 1972, WCED, 1987, Giddings et al., 2002, Soergel et al., 2021). However, the deep changes that ambitious actions pose to stabilized socio-economic human dynamics are making the realization of a sustainable human society slow, difficult, and sometimes controversial (Soergel et al., 2021).

Agriculture is certainly one of the human activities mostly interconnected with the sustainability transition. On the one hand, the model of intensive agriculture that followed the Green Revolution and has nobly led to increased food security has shown its unintended environmental trade-offs: depletion of soil and water quality, pollution of ecosystems, and contribution to climate change (Robertson et al., 2000, Pingali, 2012, Tsiafouli et al., 2015). Therefore, a urgent change is requested to agriculture as well, to shift toward models based on agroecology, where food and feed production goes along with rational use of environmental resources, improvement of ecosystem services, and protection of human health and farmers' incomes (Tilman et al., 2011, Pe'er et al., 2020, Peeters et al., 2020). On the other hand, by changing production paradigms, agriculture can have an active role in operating the sustainability transition of other sectors, too, by providing biomass to produce energy, materials, and chemicals to fuel a global bio-based economy (McCormick and Kautto, 2013, Bennich and Belyazid, 2017).

A bio-based economy is an economic system where goods are produced by using renewable biological resources, including plant biomass (McCormick and Kautto, 2013, Priefer et al., 2017). Since plants use carbon dioxide to synthesize biomass during photosynthesis, bio-based value chains based on plant biomass are ideally climate-neutral (Priefer et al., 2017). Shifting economic systems toward bio-based ones has thus the potential to sustain human needs and simultaneously avoid climate alterations and other forms of negative environmental impact. Nevertheless, recent history and research showed that this holds true only if biomass production is cautiously planned to avoid issues as carbon positive land use change, or negative trade-offs on biodiversity and food prices (Tilman et al., 2009, Fritsche et al., 2010, Robertson et al., 2017). In fact, the advance of biofuels produced by using edible components of food crops as maize or sugarcane, or by cultivating biofuel crops on fertile land under intense tillage, contributed to the spike of prices of agricultural commodities that occurred in 2007-2008 (Collins, 2008, Mitchell, 2008). Moreover, it has raised concerns about the sustainability of bio-based agro-systems in relation to land use change (Searchinger et al., 2008). Following these controversies, research showed that the selection of land and crops used for biomass provision are the most critical factors to be considered when planning bio-based value-chains in agriculture (Fritsche et al., 2010).

The research described in this thesis is based on the context just described. Specifically, the European Union's Horizon 2020 project that funded it, MAGIC (Marginal Lands for Growing Industrial Crops), aimed at generating knowledge and tools to promote the use of agriculturally unproductive land with scarce natural value – the so-called marginal lands – for the production of biomass to fuel a bio-based economy (Von Cossel et al., 2019). While the establishment of non-food biomass crops on marginal lands can guarantee the sustainability of biomass production, most of the crops suitable for this system are largely under-domesticated perennial lignocellulosic crops (mainly grasses) (Zegada-Lizarazu et al., 2010, Robertson et al., 2017, Pancaldi and Trindade, 2020). By focusing on the breeding angle of this vision, a major goal of this thesis was to develop approaches to improve the efficiency of plant breeding in such under-domesticated biomass species. As it will be discussed in the next paragraphs, one of the most important traits in biomass crops is represented by the chemical and physical properties of plant cell walls (Trindade et al., 2010, van der Weijde et al., 2013, Wang et al., 2016). Therefore, this trait was selected as model for the development of breeding tools to advance novel biomass crops. Moreover, successful breeding strategies are very often based on detailed biological knowledge on target traits. In this sense, the multiple research gaps that are currently open with respect to the fundamental biology of plant cell walls allowed for the use of the data generated in this project to address some of them. Finally, as it is often the case in science, the results obtained from the fundamental sections of the research described in this thesis provided valuable complementary insights to the applied goals. In summary, plant cell walls represents the center of the work described in this PhD thesis, the common ground of the fundamental and breeding research described in this work. As such, the results of the fundamental and applied cell wall research performed are reported in this book.

2 Plant cell walls at the interface between plant biology and the bio-based economy

2.1 Current status of cell wall biology research

Plant cell walls are molecular networks that surround every plant cell and are composed of multiple polysaccharides – cellulose, hemicellulose, and pectins – along with lignin, structural proteins, and minor aromatic molecules (Carpita et al., 2001, Cosgrove, 2005) (**Figure 1**). Cellulose molecules are biochemically uniform, being constituted by chains of glucose moieties linked by β-1,4 glycosidic bonds (Zhong et al., 2019). However, they can vary in terms of degree of polymerization and crystallinity when aggregating into fibers (Torres et al., 2015b). Conversely, hemicellulose and pectins are rather heterogeneous classes of polymers. Specifically, the term "hemicellulose" groups multiple polysaccharides – mainly xylans, xyloglucans, and (galacto)(gluco)mannans – whose backbone is formed by different classes of 5- and 6-carbon sugars as xylose, arabinose, glucose, mannose, and galactose (Zhong et al., 2019). Pectins are instead galacturonic-acid-rich polysaccharides, mainly classified into Homogalacturonan, Rhamnogalacturonan I and Rhamnogalacturonan II based on the type of moieties forming pectin backbones – galacturonic acid and/or rhamnose – and on the presence and type of side chains (Atmodjo et al., 2013). Finally, lignin is a polyphenolic polymer formed by three different types of monolignols: p-hydroxyphenyl (H), guaiacyl (G) and syringyl (S) units (Zhong et al., 2019). Overall, cell walls play several important plant functions, including provision of tensile strength to support plant growth, determination of plant architecture, conduction of water and nutrients, and mediation of the interaction with biotic and abiotic agents (Sarkar et al., 2009, Franková and Fry, 2021). Moreover, they represent relevant biomass components for bio-based industries, as they contain the bulk of plant energy, polymers and chemicals to fuel modern biorefineries (Bhatia et al., 2017).

Over the decades, research produced a wealth of knowledge on the biochemistry, biophysics, evolution, and genetics of plant cell walls. The study of cell wall composition along plant development and across plant tissues revealed that plant cell walls can be classified into primary and secondary cell walls (**Figure 1**). Primary cell walls are deposited during cell division and extension, and are flexible layers that allow cell expansion thanks to the lack of lignin and to large amounts of pectins (Fry, 2004, Popper, 2008, Broxterman and Schols, 2018). Secondary cell walls are instead deposited in multiple plant organs – including tracheary elements, wood fibers, seed pods, and root endodermis – when cell growth ceases (Popper, 2008, Zhong et al., 2019). Moreover, they contain a substantial amount of lignin and are mainly involved

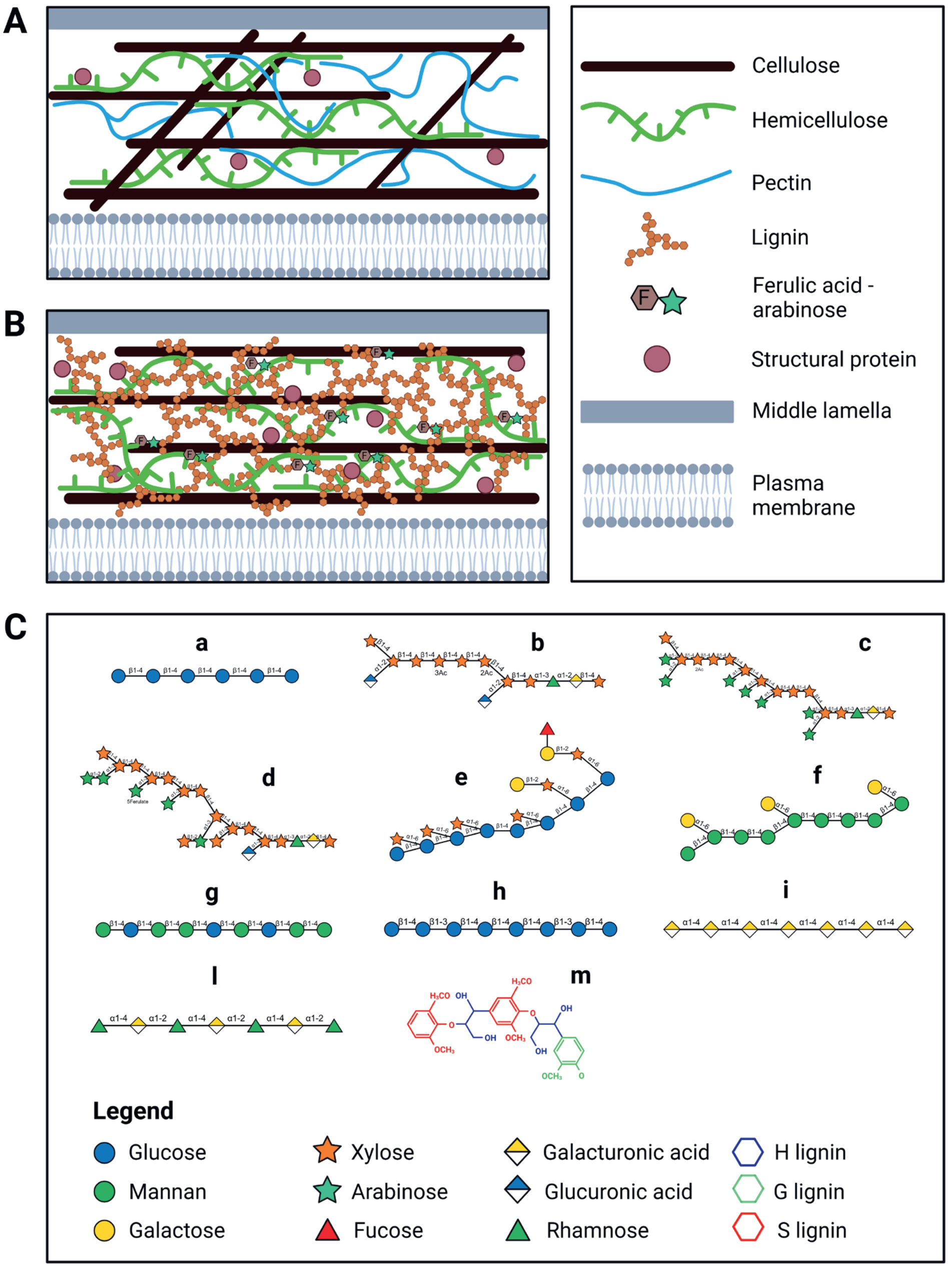


Figure 1. A-B) General representation of the primary (A) and secondary (B) cell walls of plants, the main components and their molecular arrangement within cell walls. C) Biochemistry and diversity of cell wall components: a) cellulose; b-h) hemicellulose (b: glucuronoxylan; c: arabinoxylan; d: glucuronoarabinoxylan; e: xyloglucan; f: galactomannan; g: glucomannan backbone; h: mixed-linkage glucan); i-l) pectin backbones of homogalacturonan (i) and rhamnogalacturonan I (l; note that rhamnogalacturonan can be decorated with a variety of side chains not shown here); m) molecular arrangement of multiple monolignols into a lignin polymer. Panels A-B are inspired from Nakano et al. (2015).

in providing mechanical strength and rigidity to support plant growth (Popper, 2008, Zhong et al., 2019).

The distinction between primary and secondary cell walls is a product of the complex evolution of cell walls themselves as plant structures, and of the co-evolution of cell walls and land plants (Sarkar et al., 2009, Banasiak, 2014). As an example, the major properties distinguishing primary and secondary cell walls – the presence and amount of lignin – have evolved simultaneously to the spread of vascular plants (Weng and Chapple, 2010, Renault et al., 2019). In fact, multiple mosses – an early-evolving clade of land plants that includes *Physcomitrella patens* – display non lignified, primary-like, cell walls (Sørensen et al., 2010, Norris et al., 2017), suggesting that this cell wall type appeared earlier during plant evolution. Notwithstanding, genetic analyses on the *CELLULOSE SYNTHASE* (*CesA*) gene family, which is responsible for cellulose synthesis in all plants (Richmond and Somerville, 2000) and in vascular species has specialized genes for cellulose synthesis in either primary or secondary cell walls (Kalluri and Joshi, 2004), suggested that secondary *CesA* genes evolved before the primary ones (Schwerdt et al., 2015, Little et al., 2018). This observation, along with the notorious challenge of inferring correlations between cell wall composition and plant phylogeny (Sørensen et al., 2010), makes the understanding of the early evolution of cell walls and of cell wall genetic determinants not fully resolved (Franková and Fry, 2021).

Irrespectively of the specific evolution of cell walls at the moment of land colonization by plants, basic cell wall features as the presence of primary and secondary cell walls, or of all the major cell wall components – cellulose, hemicellulose, pectins, and lignin – got stabilized and shared among the members of more recent plant clades, as the angiosperms (Sarkar et al., 2009, Meents et al., 2018). Still, evolution produced extensive cell wall diversification also within angiosperms, mostly at the level of content and chemical properties of the major cell wall molecules (Vogel, 2008, Sarkar et al., 2009, Burton et al., 2010). Specifically, a major differentiation is the one between type I and type II cell walls, with the former being specific to eudicots and non-commelinid monocots, and the latter to the Poaceae family (grasses) (Vogel, 2008, Penning et al., 2019). These two cell wall types differ in the content and type of several cell wall components, including the amount of lignin (typically higher in type II cell walls), the type of hemicellulosic molecules (dominated by xyloglucan in type I cell walls and by xylans and (1,3;1,4-β-glucans – also termed mixed-linkage glucans – in type II cell walls), the content of pectins (typically higher in type I primary cell walls) (Vogel, 2008, Penning et al., 2019), or the amount of ferulic acid to cross-link lignin and cell wall polymers (usually higher in type II secondary cell walls) (De Oliveira et al., 2015). Recent research showed that the enzymatic machinery necessary for synthesizing both type I and type II cell walls is overall conserved across Poaceae and eudicots, and so are the genes coding for these enzymes (Penning et al., 2019, Kozlova et al., 2020). Some exceptions occur, as the *CELLULOSE SYNTHASE-LIKE F* gene family, which is the major responsible for mixed-linkage glucan synthesis and is restricted to Poaceae within the angiosperm clade (Burton et al., 2006). Nevertheless, gene presence-absence variation alone does not convincingly explain the differences between type I and type II cell walls. While recent studies highlighted differential post-translational modifications or gene expression as important factors that might underly these differences (Penning et al., 2019, Kozlova et al., 2020), the genetic basis underpinning type I and type II cell wall differentiation remains not fully understood. Still, its study is highly relevant as it could unveil novel prospects for modifying cell wall composition in industrial biomass crops, or even reveal interesting dynamics at the basis of the evolution of highly complex traits as cell wall composition and morphology (Vogel, 2008).

By generalizing what just discussed for type I and type II cell walls, the high complexity of cell wall biosynthesis has made the study of its genetics a notoriously challenging task. Research aimed at this goal was initially focused on the identification of general gene families at the basis of cell wall biosynthesis. For example, the CAZy database was created in 1991 (Henrissat, 1991) to classify genes and enzymes responsible for carbohydrates metabolism in different functional groups based on sequence features (Drula et al., 2022). Similarly, screening of genomic libraries led to the identification and characterization of the first plant *CesA* genes and of the first arabidopsis *PHENYLALANINE AMMONIA-LYASE* (*PAL*) genes (which initiate the lignin pathway) more than 25 years ago (Ohl et al., 1990, Pear et al., 1996). The advent of tools for reverse genetic studies boosted the research on the genetic determinants of cell wall biosynthesis, allowing for the functional characterization of several genes. For example, expression profiling and T-DNA insertion mutants were used to characterize novel *IRREGULAR XYLEM* (*IRX*) and *COBRA* genes in the context of secondary cell wall formation in arabidopsis (Brown et al., 2005). Similarly, the creation and characterization of multiple arabidopsis and maize mutant lines extended the knowledge on cell wall genetics to families that were completely understudied until early 2000s (Yong et al., 2005, Penning et al., 2009). Overall, this type of research is still highly valuable and currently deployed for multiple purposes, including the functional characterization of homologous members of known cell wall gene families in understudied species (Liu et al., 2012, Peng et al., 2019, Song et al., 2019). Nevertheless, high-throughput methodologies for genetic and genomic research have also represented a valuable tool to move beyond single genes and gene families in our understanding of cell wall biosynthesis. As an example, the combination of forward and reverse genetic approaches with large-scale bioinformatic analysis of multiple -omics datasets allowed the modelling of the first comprehensive network of the regulation of secondary cell wall synthesis in arabidopsis (Taylor-Teeples et al., 2015). Similarly, bioinformatic analyses were pivotal to investigate the regulation of cell wall biosynthesis in different plant clades, including grasses (Rao and Dixon, 2018) and woody angiosperms (Zhang et al., 2018). To conclude, the further investigation of cell wall genomics through large-scale and comprehensive approaches is expected to further reveal relevant factors at the basis of cell wall biology (Yokoyama, 2020).

2.2 Plant cell walls as highly relevant biomass components for a bio-based economy

As anticipated earlier, plant cell walls are amongst the most important plant biomass components to obtain energy, polymers, chemicals, and several other raw materials to fuel a bio-based economy (Bhatia et al., 2017). The industrial relevance and versatility of plant cell walls depends on the molecular properties of these plant structures, as well as on how cell walls functionally evolved over plant evolution. Specifically, hemicellulose, cellulose, and lignin all display a high energy content, between 1.5 and 2.3 KJ/g (Novaes et al., 2010). This property makes cell walls a highly attractive feedstock for producing energy through biomass combustion or fermentation into cellulosic biofuels (van der Weijde et al., 2013, Torres et al., 2015b, Pang et al., 2018). Moreover, specialized cells in crops as cotton, hemp, or flax can produce secondary cell walls with an exceptionally high cellulose content, which are at the basis of fiber synthesis for production of textiles or composites (Van der Werf et al., 1994, Haigler et al., 2012, Zommere et al., 2013). Furthermore, the gelling and adhesive properties of pectins make this group of cell wall polymers highly attractive for the synthesis of films or hydrogels (Liu et al., 2006, Freitas et al., 2021). Additionally, the phenolic nature and the high carbon content of lignin makes it an excellent raw material for producing phenolics for medicines or bioplastics, lignin-based carbon fibers, or additives for construction materials (Agrawal et al., 2014). Finally, plant cell walls represent overall a barrier to the extraction of valuable raw materials as plant proteins that are encased within plant cells, making the study of cell walls pivotal for optimizing extraction procedures of these compounds (Sari et al., 2015, Kumar et al., 2021).

The amenability of plant biomass for all the aforementioned applications does not come as an inherent plant property, but is a highly complex trait that largely depends on the relative content and composition of the different cell wall components (Sari et al., 2015, Van der Weijde et al., 2017c, Petit et al., 2020a). This trait is referred to as "biomass quality" or "cell wall quality" throughout this thesis. An important aspect of it is that the ideal characteristics determining cell wall quality are defined based on the end use of plant biomass, with trade-offs existing between alternative end-uses.

For example, a high content of lignin in plant cell walls limits the efficiency of biomass conversion into cellulosic biofuels (Pauly and Keegstra, 2010, Van der Cruijsen et al., 2021). However, it represents a valuable property for whole-biomass combustion for combined heat and power generation (Lewandowski and Kicherer, 1997). Similarly, a high proportion of crystalline cellulose in plant cell walls is another unfavorable property for biomass fermentation into biofuels (Hall et al., 2010). However, it is a desired trait to obtain strong fibers from crops as hemp, flax, or cotton (Ward Jr, 1950, Marrot et al., 2013). Furthermore, the inherent molecular diversity of pectins is a major determinant of the different industrial end uses that these molecules can fulfil (Gawkowska et al., 2018). Finally, the relative cell walls content of (hemi)cellulosic components as arabinose, mannose, or glucose affects the extraction efficiency of high-value molecules from plants (Boulet et al., 2022).

The examples above highlight the importance of cell wall quality to create crop varieties tailored to specific biomass end uses, whose availability is pivotal to make several bio-based value chains economically profitable (Trindade et al., 2010, Pancaldi and Trindade, 2020). In this regard, it is relevant that cell wall quality is a largely genetically-determined trait, which makes the development of breeding tools (i.e. mining of target genes/alleles, molecular markers, and breeding strategies) to assist its improvement in biomass crops possible. Specifically, several studies demonstrated that multiple relevant cell wall compositional features for an industrial use of plant biomass can be highly heritable in different crops (Torres et al., 2015a, van der Weijde et al., 2017b, Petit et al., 2020b, Gulisano et al., 2022). Moreover, molecular markers and genomic regions associated with variability in biomass quality for different end uses were also mapped in different plant species (Torres et al., 2015a, van der Weijde et al., 2017b, Petit et al., 2020a, Gulisano et al., 2022). Finally, different genes with an impact on biomass quality have also been identified across plants (Penning et al., 2009, Brandon and Scheller, 2020). Overall, this knowledge shows that, despite the complexity of cell wall quality as a highly quantitative trait, it is possible to attain the development of crop varieties with superior performance in relation to specific bio-based end uses. However, this process is not straightforward, and poses significant challenges from a breeding point of view.

3. The challenge of sustainable biomass provision: breeding implications

Despite the theoretical feasibility of developing varieties of biomass crops tailored to specific end-uses, the actual development of such varieties in breeding programs is a complex process, typically time-consuming and cost-ineffective (Clifton-Brown et al., 2018). By recalling the vision of the MAGIC project discussed in section 1 (Von Cossel et al., 2019), the latter is especially true for under-domesticated biomass crops that could be cultivated on marginal lands thanks to their natural adaptation to suboptimal environments (Pancaldi and Trindade, 2020). In fact, breeding programs on crops as switchgrass, miscanthus, willow, or poplar can easily last up to 25 years before commercialization of elite genotypes is achieved (Clifton-Brown et al., 2018). Moreover, they can take even longer in other suitable bio-based crops for marginal lands that received less attention for investments so far, as *Lupinus mutabilis* (Gulisano et al., 2019) or guayule (Abdel-Haleem et al., 2019).

Among the steps to take in breeding programs of crops as the ones just mentioned, the study of the genetic architecture of target traits and the identification of molecular markers and favorable alleles to guide initial selection of genotypes, also referred as pre-breeding activities, are some of the most time-consuming ones (Clifton-Brown et al., 2018). Moreover, the need of extensive multi-trials to properly assess phenotypic variability and allow the mapping of the genetic background of such variability is also both time- and cost-ineffective. Therefore, the availability of tools to allow for a quicker screen of germplasm for genetic variants of interest underlying highly heritable traits would represent a valuable asset to increase the efficiency of plant breeding in the context just discussed.

4. Genomic prospects for cell wall biology and breeding in biomass crops

As previously mentioned, the study of the genomic background underlying cell wall biosynthesis could significantly expand our understanding of cell wall biology. As genomes contain the full genetic information of living organisms, including all their genes, their analysis is powerful to dissect the genetic properties governing biological traits. Specifically, comparing genomic properties across species through comparative genomic studies is extremely valid to understand the evolution and the genetic functioning of living organisms, and has allowed multiple key discoveries in this sense (Xia, 2013). In this regard, a highly relevant feature for comparative genomics is the concept of gene synteny. The word "synteny" has roots in the ancient Greek language, being the merged form of the terms *σύν, syn* = "along with", and *ταινία, tainiā* = "band", and translates to the expression "on the same ribbon" (Passarge et al., 1999). Accordingly, synteny was initially used to indicate groups of genes belonging to homologous chromosomes across species (Renwick, 1971, Passarge et al., 1999). Nevertheless, nowadays synteny is nearly entirely used to refer to conserved gene collinearity across species, indicating genomic regions showing multiple homologous genes in the same relative order in different species (**Figure 2**) (Liu et al., 2018a). This is also the conceptual meaning of synteny considered in this research, and genes located on syntenic regions are referred as "syntenic genes" or "syntelogs" throughout the thesis.

The reason why synteny is a highly important genomic property stands in how it implies not only inter-specific gene sequence similarity because of a likely common evolutionary origin – the concept of gene homology (Reeck et al., 1987) – but also the conservation of the local genomic area of genes – also referred as "genomic context"

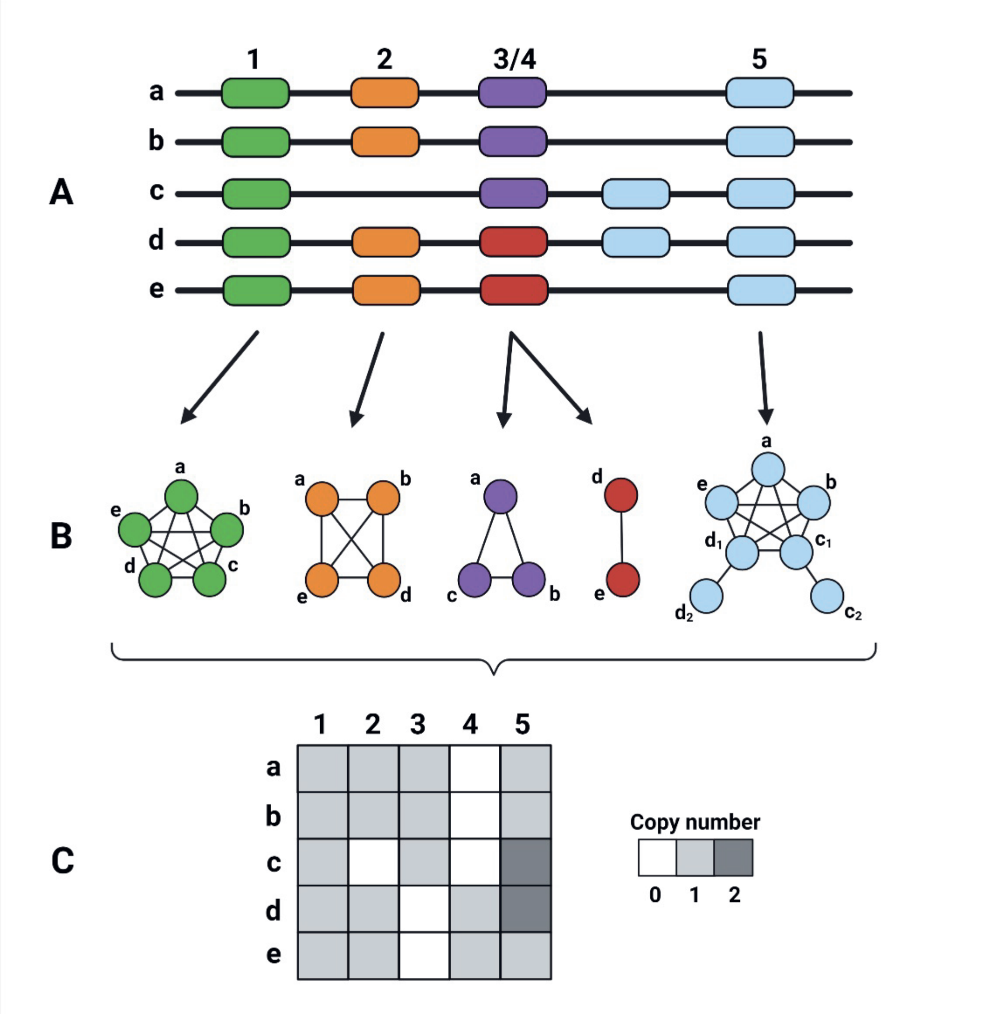


Figure 2. Representation of genome synteny as meant in the context of this PhD thesis. A) Representation of the local genomic collinearity of five fictious genes (1-5) across the genomes of five fictious species (a-e). The figure displays common situations that can be encountered during synteny analysis, as the deletion of gene 2 in species c, the coexistence of different genes – 3 and 4 – at the same genomic locus in different species, and the tandem duplication of gene 5 in species c and d. B) Network representation of the synteny dynamics shown in panel A, according to the theory and tools developed by Zhao and Schranz (2017). Each independent network represent an independent genomic gene context. C) Representation of the synteny information reported in panels A and B as a table reporting taxonomic profiling of gene copy number per genomic context across species. Both network synteny analysis and taxonomic profiling of synteny networks were frequently used in the research of this PhD thesis.

(Dewey, 2011). From a functional point of view, the conservation and diversification of genomic contexts turned intimately associated with the conservation and diversification of gene functions across species, respectively, in several studies (Zhao et al., 2017, Artur et al., 2019, Kerstens et al., 2020). Therefore, studying this property in large panels of plant species can be key to dissect complex trajectories of trait evolution that are ultimately genomic-dependent (Tang et al., 2008, Dewey, 2011, Zhao and Schranz, 2017). In this perspective, synteny data can be co-analyzed with other genomic and gene properties, as gene presence-absence and copy number variation across species, or diversification of gene sequences, in phylogenomic studies (Zhao and Schranz, 2017). While bioinformatic tools and genomic resources currently allow to perform these analyses at large-scale levels (Zhao and Schranz, 2017), they have been seldomly applied in the context of cell wall biology. Nevertheless, phylogenomic approaches to study cell wall genes as complex traits determinants are promising, as they led to highly interesting results in similar settings. For example, phylogenomic analyses on the MADS-box transcription factors from 51 plant genomes allowed to discriminate between gene diversification events that pre-dated diversification of seed plants from events specific to certain angiosperm clades (Zhao et al., 2017). These features were linked to key phenotypic adaptations of diverse plant clades with respect to complex traits as flowering regulation or root architecture (Zhao et al., 2017, Kerstens et al., 2020). At the moment of designing the research of this thesis, it was foreseen that this type of large-scale genomic research may have led to the identification of novel patterns at the basis of early cell wall evolution and of type I and type II cell wall differentiation.

Next to fundamental research, comparative genomics can be highly relevant also in the context of plant breeding. Specifically, this is true for the techniques that can be collectively referred as “translational genomics” (Salentijn et al., 2007, Kang et al., 2016). This concept was developed in the context of transferring genetic information from model species to under-domesticated or wild crops harboring interesting traits. Moreover, it has been extensively applied for identifying homologous genes across species based on similarity of sequence or of protein features as domain composition, or during genome annotation (Kang et al., 2016). Still, it has been envisioned that translational genomics could also be applied for transferring information harbored in quantitative trait loci (QTLs), by using syntenic conservation of QTLs across species (Kang et al., 2016). QTLs are highly important in breeding contexts as they are genomic regions mapped in association studies aimed at linking observed phenotypic variability in a species to its genetic determinants. As such, QTL regions are proved to be causative of trait variability in specific accession panels and/or environments. Therefore, their transfer to novel species/crops is foreseen to allow for faster and more efficient breeding efforts. However, it has been so far difficult to show that QTLs transfer between species can be successful in the way it can predict strong candidate loci of target traits in new species, at the same time pointing to regions displaying favorable allelic variability for breeding efforts. Nevertheless, the proof of this principle, along with the development of pipelines to operate such transfer, would certainly represent a great asset for advancing varieties of novel crops more effectively, also in the afore-mentioned context of biomass crops for marginal lands.

5. Layout of the thesis

Rooted in the context, concepts and knowledge described in all the previous paragraphs, the research efforts of this thesis encompassed a series of genomic analyses focused mainly on cell wall genes as major determinants of cell wall quality (as complex target trait for breeding novel biomass crops) and cell wall biosynthesis (from a fundamental perspective), with two major aims. First, to comprehensively study the genomic architecture of cell wall biosynthesis in land plants, by targeting preeminent research questions on cell wall biology. Second, to use knowledge on cell wall genomics and biology to design, develop, and validate novel tools to perform effective genomic translation of QTLs from model biomass species to new biomass crops to be cultivated on marginal lands. In agreement with these goals, the research performed during this PhD project has been structured as follows:

Chapter 2 reports on a detailed phylogenomic analysis of the *CELLULOSE SYNTHASE* gene superfamily in plants (*sensu lato*), including 242 genomes covering evolution from green algae to eudicots. The *CELLULOSE SYNTHASE* genes are both models to study cell wall evolution and highly relevant genetic determinants of cellulose and hemicellulose content and properties within cell walls. As such, the analysis of synteny, copy number, and phylogenetics of these genes yielded insights into the early evolution and later diversification of cell walls, and uncovered highly relevant genomic patterns that accompanied specific phenotypic innovations with potential practical importance for plant breeding.

In **Chapter 3**, a phylogenomic approach was applied to study variability in the genomic background of 150 different cell wall gene families across 169 angiosperm genomes. The results uncovered a substantial genomic differentiation of cell wall gene organization between Poaceae and dicots, with relevant implications for understanding the evolution of type I and type II cell walls, as well as for potentially design novel strategies to modulate cell wall composition in biomass crops.

Chapter 4 presents a perspective on how to use the available genetic knowledge on cell wall quality determination in order to design novel tools and strategies to attain an effective and faster improvement of promising under-domesticated biomass crops adapted to marginal lands. This perspective is accompanied by an extensive review on how marginal lands and under-domesticated perennial lignocellulosic crops can be highly valuable to support the development of bio-based value chains. Moreover, the major breeding targets to achieve a sustainable cultivation of biomass crops on marginal lands are also discussed.

In **Chapter 5**, a strategy for effectively translating genetic knowledge on cell wall quality in the form of QTLs between large sets of species is designed and implemented, based on genome synteny across plants. The application of this strategy led to the identification of several "syntenic QTLs" for cell wall quality (i.e. genomic regions syntenically conserved across species and harboring known cell wall quality QTLs in multiple model species) in a large panel of angiosperms, which appear inter-applicable across species. Moreover, the genomic characterization of syntenic QTLs in the context of biomass quality, as well as the discussion of how the strategy for syntenic QTLs detection can be potentially applied to other traits and breeding settings is also warranted.

The research on syntenic QTLs is continued in **Chapter 6**. First, by uncovering the existence of extensive intra-specific allelic variability of syntenic QTL regions through a pan-genomic approach involving six diverse species for which multiple genomes representing diverse accessions were collected. Second, by assessing the validity of syntenic QTL regions for predicting relevant genomic loci associated with cell wall quality variability in target accession panels and/or species that were not beforehand used in initial syntenic QTL detection. Overall, this research shows the usefulness of syntenic QTLs as novel tools in the breeder's toolbox.

Finally, **Chapter 7** provides a general discussion on the main findings of this thesis, from both a fundamental and applied perspective. Specifically, the novel considerations opened on cell wall evolution from the results of Chapters 2 and 3 are discussed in the context of the current knowledge on cell wall biology. Moreover, the relevance of genomic research for both the fundamental study of cell wall diversification and for novel breeding applications is discussed. Finally, the usefulness of syntenic QTLs as novel breeding tools and their potential applications are also debated, along with the perspectives that are opened on the future of plant breeding and on the role of biomass crops in a bio-based economy.

Chapter 2

Genomic architecture and evolution of the *Cellulose synthase* gene superfamily as revealed by phylogenomic analysis

Francesco Pancaldi[1], Eibertus N. van Loo[1],
M. Eric Schranz[2], Luisa M. Trindade[1]

[1]Plant Breeding, Wageningen University & Research, Wageningen,
The Netherlands

[2]Biosystematics, Wageningen University & Research, Wageningen,
The Netherlands

Published as: Pancaldi, F., van Loo, E.N., Schranz, M.E. and Trindade, L.M. (2022). Genomic Architecture and Evolution of the Cellulose synthase Gene Superfamily as Revealed by Phylogenomic Analysis. *Frontiers in plant science, 13*.
DOI: 10.3389/fpls.2022.870818

Abstract

The *Cellulose synthase* superfamily synthesizes cellulose and different hemicellulosic polysaccharides in plant cell walls. While much has been discovered about the evolution and function of these genes, their genomic architecture and its relationship with gene (sub-)functionalization and evolution remains unclear. By using 242 genomes covering plant evolution from green algae to eudicots, we performed a large-scale analysis of synteny, phylogenetic, and functional data of the *CesA* superfamily. Results revealed considerable gene copy number variation across species and gene families, as well as two patterns – singletons vs tandem arrays – in chromosomic gene arrangement. Synteny analysis revealed exceptional conservation of gene architecture across species, but also lineage-specific patterns across gene (sub-)families. Synteny patterns correlated with gene sub-functionalization into primary and secondary *CesAs* and distinct CslD functional isoforms. Furthermore, a genomic context shift of a group of cotton secondary *CesAs* was associated with peculiar properties of cotton fiber synthesis. Finally, phylogenetics suggested that primary *CesA* sequences appeared before the secondary *CesAs*, while phylogenomic analyses unveiled the genomic trace of the *CslD* duplication that initiated the *CslF* family. Our results describe in detail the genomic architecture of the *CesA* superfamily in plants, highlighting its crucial relevance for gene diversification and sub-functionalization, and for understanding their evolution.

1 Introduction

Plant cell walls (CWs) are versatile structures that surround plant cells and fulfil key plant functions, including providing tensile strength and mediating plant-environment interactions (Sarkar et al., 2009). About 60-90% of CWs dry weight is constituted by cellulose and hemicellulosic polymers (Pettolino et al., 2012), synthesized by the enzymes coded by *Cellulose synthase* (*CesA*) gene superfamily (Richmond and Somerville, 2000). The genes of this superfamily are present in all land plants and globally divided into 12 gene families (Richmond and Somerville, 2000, Yin et al., 2009, Little et al., 2018). Among these families, *CesA* genes were the first discovered members of the *CesA* superfamily and are involved in cellulose synthesis (Pear et al., 1996, Turner and Somerville, 1997, Arioli et al., 1998). These genes are a monophyletic group that the algal ancestors of land plants acquired through horizontal transfer from cyanobacteria and that expanded, diversified and sub-functionalized during plant evolution (Banasiak, 2014, Schwerdt et al., 2015). Specifically, angiosperms typically contain 10-20 *CesA* genes (Richmond and Somerville, 2000, Yin et al., 2009) and different CesA isoforms mediate cellulose synthesis in either primary or secondary cell walls (Kalluri and Joshi, 2004, Burton et al., 2004, Ranik and Myburg, 2006, Wang et al., 2010b, Carroll et al., 2012, Liu et al., 2012, Kaur et al., 2016). Moreover, different CesA members assemble into functional hetero-multimeric cellulose synthesis complexes that accomplish the actual cellulose synthesis (Haigler and Roberts, 2019, Carroll et al., 2012). Despite the recent advances in the knowledge of *CesA* biology, the evolutionary trajectories that led to the current *CesA* diversity are not fully understood (Little et al., 2018), nor the genomic architecture of this family and its relationship with CW biology.

The other *CesA* superfamily genes besides the *CesA* family are termed *Cellulose synthase-like* (*Csl*) genes. *Csl* genes participate to the synthesis of different hemicellulosic polysaccharides (Richmond and Somerville, 2000, Schwerdt et al., 2015, Little et al., 2018) and include 11 families, termed *CslA* to *CslM* (Little et al., 2018). The *CslB*/*D*/*E*/*F*/*G*/*H*/*J*/*M* families belong to the same monophyletic lineage of the *CesA* genes (Banasiak, 2014, Little et al., 2018), while the *CslA*/*C*/*K* families form a different clade that originated through an independent endosymbiosis in green algae (Yin et al., 2009, Banasiak, 2014, Little et al., 2018). Within the *CslB*/*D*/*E*/*F*/*G*/*H*/*J*/*M* families, the *CslD* and *CslF* clades are phylogenetically closer to *CesA* genes than the others (Yin et al., 2009, Schwerdt et al., 2015). Specifically, *CslDs* form a sister clade *CesAs*, while *CslFs* form a Poaceae-specific sister clade of *CslDs*. While phylogenetic analyses suggest that *CslF* genes originated from *CslD* duplication (Yin et al., 2009, Schwerdt et al., 2015, Little et al., 2018), it is unclear if such duplication took place after the divergence of the *CslD* clade, or in an ancestral clade of both the families

(Little et al., 2018). Similarly, the *CslB*/*E*/*G*/*H*/*J*/*M* families diverged independently of the *CslD* and *CslF* genes from the *CesA* family (Yin et al., 2009, Little et al., 2018), but the timing and modes of their evolution is largely undefined (Banasiak, 2014). Furthermore, the role of the genomic architecture of genes in the evolutionary trajectories of all these families is unclear.

Knowledge gaps are currently open also regarding the function of the *Csl* families. In this respect, the *CslF*, *CslH* and *CslJ* families are known to be involved in the synthesis of (1,3;1,4)-β-glucans (or mixed-linkage glucans), a group of polymers mainly found in Poaceae (grass) CWs and consisting of (1,3;1,4)-β linked glucosyl residues (Burton et al., 2006, Doblin et al., 2009, Little et al., 2018). While the involvement of the *CslF*, *CslH* and *CslJ* families in the synthesis of mixed-linkage glucans has been widely established over the last decades, two members of the barley *CslF* family – *HvCslF3* and *HvCslF10* – have recently been shown to synthesize (1,4)-β-linked glucoxylans (Little et al., 2019). This finding can challenge the concept that *CslF*, *CslH* and *CslJ* families synthesize a single type of polysaccharide. Nevertheless, the *CslF*, *CslH* and *CslJ* families remain the best-studied *Csl* families, and the other *Csl* genes are much less characterized. Some *CslA* genes synthesize mannans and glucomannans (Dhugga et al., 2004, Liepman et al., 2005), while certain CslC isoforms mediate xyloglucan biosynthesis (Cocuron et al., 2007, Kim et al., 2020b). However, the size and diversity of these families suggests that they could also synthesize other hemicellulosic polysaccharides, or different forms of mannans and xyloglucans (Liepman and Cavalier, 2012). *CslD* genes are believed to synthesize the non-crystalline single chains of cellulose in root hairs and pollen tubes (Doblin et al., 2001, Kim et al., 2007, Bernal et al., 2008), but their involvement in mannan synthesis has also been proposed (Verhertbruggen et al., 2011). In addition, *CslD* mutations affect pollen development (Bernal et al., 2008), root morphology (Wang et al., 2001, Kim et al., 2007, Bernal et al., 2008, Hu et al., 2018), and vegetative organ size (Li et al., 2009, Luan et al., 2011, Hunter et al., 2012, Li et al., 2018) in several plants, but the molecular basis of these effects is unclear. Finally, the function of the evolutionarily-related *CslB*/*E*/*G*/*M* families is currently unknown (Little et al., 2018). Moreover, as for the *CesA* genes, the role of the genetic architecture of the *Csl* families in determining gene function has not yet been investigated.

The study of the genomic architecture of the *CesA* superfamily in plants and of its relationship with the evolution and function of these genes is the focus of this research. To this aim, we performed a combined phylogenetic and synteny analysis (phylogenomic analysis) of the *CesA* superfamily genes from 242 species covering plant evolution from green algae to eudicots. Large-scale phylogenomic analyses are a rather novel approach, but have turned powerful to study the genetics and the evolution of complex gene families (Zhao et al., 2017, Kerstens et al., 2020). Moreover, if coupled with relevant phenotypic and functional data, these analyses can reveal relationships between the genomic architecture of target gene families and phenotypic adaptations of plants (Zhao et al., 2017, Kerstens et al., 2020). Large-scale phylogenomic analyses have become feasible thanks to the increasing availability of sequenced plant genomes and to the development of bioinformatic tools for their analysis, like network approaches for large-scale synteny computation (Zhao and Schranz, 2017). The application of these tools to the *CesA* superfamily highlighted interesting patterns in the genomic arrangement of these genes across plants, and relevant associations between phylogenomic patterns and key events in the evolution and sub-functionalization of different gene families, including *CesA*, *CslD*, and *CslF* genes. This also led to the formulation of novel hypotheses on the evolution of different *CesA* and *Csl* families.

2 Materials and methods

2.1 Genomic data sources

All the plant genomes (*sensu lato*) sequenced and published by 2018 and available with a scaffold-level assembly were searched for in online databases (**Supplementary Table 1**). For each genome, a GFF/BED file of gene positions and a protein FASTA file reporting the main protein coded by each gene were retrieved. Moreover, genomes were analysed for completeness and fragmentation by using the BUSCO Viridiplantae gene set (Seppey et al., 2019) and by assessing the number of scaffolds and the N50 statistics. In total, 242 genomes from 212 species were collected (**Supplementary Table 1**). For each species with an available genome, information on its ploidy level and the number of genome duplications were searched for online and on scientific literature (see especially Van de Peer et al. (2017) for information about genome duplications).

2.2 Identification of *CesA* and *Csl* genes

A group of 445 protein sequences of known *CesA*/*Csl* genes from 13 species spanning plant diversity were retrieved from literature and used as BLAST queries (Altschul et al., 1990) against the 242 proteomes of the study (**Supplementary Table 2**). In parallel, all the proteins of the 242 genomes were screened for the characteristic PFAM domains of the *CesA* superfamily – PF03553 and PF00535 (Little et al., 2018) – using HMMER (Eddy, 2009, El-Gebali et al., 2019). The outputs of the BLAST and HMMER searches were merged and filtered to exclude partial sequences not starting with Methionine and/or shorter than the residue length of the PFAM domains annotated onto them (total residues spanned by each PFAM; 170 AAs for PF00535 and 722 AAs for PF03552). The remaining sequences were assigned to *specific CesA*/*Csl* families through a second BLAST against the *CesA*/*Csl* genes from the initial 13 species for which a *specific CesA*/*Csl* function was known and by plotting them in a phylogenetic tree to identify wrongly annotated and spurious sequences (R, custom script).

2.3 Identification of primary and secondary *CesA* genes

For the phylogenomic analysis of primary and secondary *CesA* genes, a set of 49 experimentally validated primary and secondary *CesA* sequences was retrieved from literature (**Supplementary Table 3**) and used in a BLAST search (Altschul et al., 1990) against all the *CesA* genes found in the genomes of the study. BLAST results were used to preliminarily categorize all *CesA* genes as primary or secondary *CesAs*. This initial assignment was refined by checking the simultaneous presence of two motifs (CQIC and SVICEXWFA) previously shown to characterize only primary *CesA* proteins in a wide range of plants (Kaur et al., 2016). Moreover, the position of each *CesA* gene found in the BLAST search, relative to the clades where the initial 49 primary and secondary *CesAs* were placed, was manually inspected in each phylogenetic tree of this study to further help the categorization of primary and secondary *CesA* genes following the BLAST search.

2.4 Synteny network construction

The synteny network of the *CesA* superfamily was built by following the methodology of Zhao and Schranz (2017). In brief, we used Diamond (Buchfink et al., 2015) to perform BLAST-like alignments of all the proteins of each genome against all the other proteins of that genome and all the proteins of every other genome (Evalue = 1E−3). In this way, we identified homologous genes between different species. Subsequently, MCScanX (Wang et al., 2012b) was run with default parameters (except -s – the number of colinear genes to claim a syntenic block – set to 3) to detect gene synteny (i.e. conserved gene order across multiple genomes) The results of MCScanX were organized in a synteny network, in which each node is a gene and edges represent syntenic connections between genes. The set of edges in which at least one node was a *CesA*/*Csl* gene was extracted from the overall synteny network and represents the *CesA*/*Csl* synteny network (**Supplementary Table 4**).

2.5 Analysis of syntenic communities

The R package igraph and the Multi-level algorithm (Yang et al., 2016b) were used to detect syntenic communities within the *CesA*/*Csl* network (i.e. groups of *CesA*/*Csl* genes displaying a higher degree of synteny with each other than with the rest of the network) formed by at least four nodes. Communities were profiled to assess the type(s) of *CesA*/*Csl* families and of species included in each of them, and the gene copy number of each *CesA*/*Csl*-taxa combination found. Finally, communities were hierarchically clustered based on *CesA*/*Csl* copy number per species (R, hclust function, ward.2 algorithm).

2.6 Multiple sequence alignment and phylogenetic analysis

Phylogenetic trees were built for relevant groups of *CesA*/*Csl* genes. For each tree, full-length CesA/Csl protein sequences were aligned with MAFFT v7.453 (FFT-NS-2 algorithm) (Katoh and Standley, 2013), with default parameters except gap opening penalty, set at 1.0. MAFFT alignments were trimmed with TrimAl v1.2 (Capella-Gutiérrez et al., 2009), with manual optimization of the -gt and -cons flags to obtain alignments lengths similar to the median lengths of the initial proteins included in each tree. Phylogenetic trees were built from trimmed alignments using RAxML v8.2.9 (PROTCATBLOSUM62 substitution matrix; 100 bootstraps) (Stamatakis, 2014), and plotted and annotated using iTOL (Letunic and Bork, 2019).

2.7 Statistical analyses

All the statistical analyses (t-tests, ANOVAs, LSDs, correlations) described in this research were performed with SPSS (IBM Corp., Armonk, NY; Version 26.0).

3 Results

3.1 Genomic architecture of the *CesA* superfamily

Gene copy number

We used 242 plant genomes covering plant evolution from green algae to eudicots (**Supplementary Table 1**) to perform an extensive genomic and synteny analysis of the *CesA* superfamily in plants, of which the first step was the study of its genomic architecture. The 222 plant genomes with a BUSCO representation ≥75% contained 7997 *CesA* superfamily genes, with an average of 36 members per species (CV=48.6%; **Supplementary Table 5**; CV: Coefficient of Variation). The copy number of *CesA*/*Csl* genes correlates with both the ploidy level and the number of genome duplications of each species (ρ=0.57 and ρ=0.55, respectively; P<0.001 for both; **Supplementary Figure 1**). Moreover, the *CesA* superfamily size increased considerably during plant evolution, with an acceleration at the rise of angiosperms. In fact, while charophytes, bryophytes, lycophytes, ferns, and gymnosperms contain a relatively similar number of *CesA*/*Csl* genes (~18 per species, CV=36.9%), this figure doubles in angiosperms despite similar inter-species variability (~40 *CesA*/*Csl* genes per species, CV=35.9%, P<0.001; **Supplementary Figure 2**).

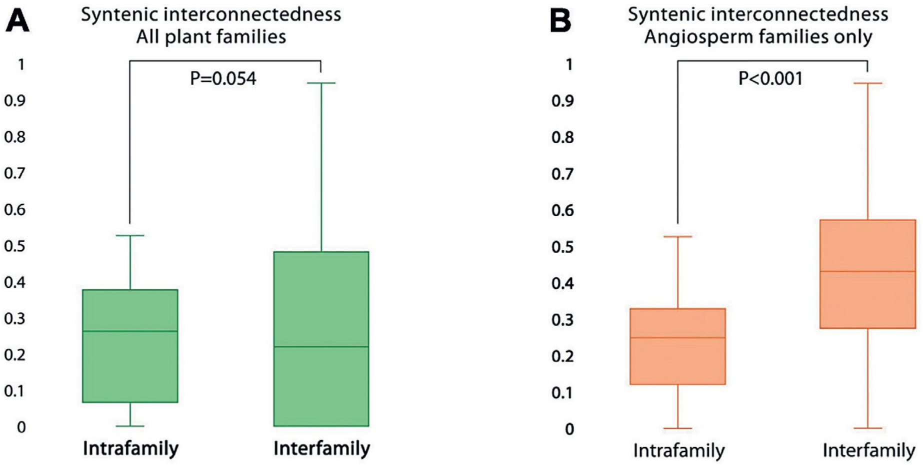
CesA is by far the most abundant *CesA*/*Csl* family, with 12 genes per species on average (CV=43.1%; **Supplementary Figure 3**). This is roughly two times the size of the *CslD*, *CslA* and *CslC* families (P<0.001), which all display ~6 genes per species. These three families are in turn significantly larger than the *CslB*, *CslE*, *CslG*, and *CslM* families (1-3 genes per species each; P<0.001). Finally, the monocot-specific *CslF* and *CslH* families typically display 7 and 1 genes per species on average, respectively. Overall, the copy number of all the *CesA*/*Csl* families varies extensively across plant families, with standard deviations often equal to the mean size of gene families across species.

Gene distribution along genomes

The relative positions of the *CesA/Csl* genes within each genome of the study were also assessed, highlighting two main patterns (**Supplementary Tables 6-8**). On the one hand, the *CesA*, *CslA*, *CslC* and *CslD* families display a scattered genomic distribution among multiple singleton loci. On the other hand, the *CslB*, *CslE*, *CslG*, and *CslM* clades are usually organized in tandem arrays of 2-4 genes. These arrays were detected in both monocots and dicots, and synteny analysis showed their extensive conservation across species (see Section 3.1.3). A large conserved gene array was also found for the *CslF* family in all the Poaceae species evaluated (as previously reported by Schwerdt et al. (2015) for rice, sorghum, brachypodium and barley only). However, we also found 20-30% of the *CslF* genes of each Poaceae species organized as conserved singletons or tandems at separated loci. Interestingly, this second group of *CslF* genes was syntenic with *CslD* members of several eudicot species (see Sections 3.1.3 and 3.3.2). Finally, the *CslH* and *CslJ* families displayed both configurations – tandem arrays and singleton loci. However, these families are present at too low frequencies and in too few species to allow generalizations.

Gene synteny

The syntenic conservation of the *CesA* superfamily was also assessed in detail by using network synteny analysis. Results indicated that *CesA* superfamily genes are highly syntenic across diverse plants, more than what observed for other plant gene families, including regulatory genes. In fact, of the 7005 *CesA* superfamily genes from the 193 genomes with a BUSCO representation ≥75% and at least five genes per scaffold on average, 6262 (89%) were included into the *CesA*/*Csl* synteny network (**Supplementary Table 9**). This is a remarkably high proportion, 9% and 24% higher than what found in comparable studies on the *MADS-box* and the *APETALA2* gene families, respectively, (Zhao et al., 2017, Kerstens et al., 2020). Furthermore, the synteny of the *CesA*/*Csl* genes is dense and extensive across diverse plant species and families. In fact, each *CesA*/*Csl* gene is syntenic to another 69 homologs from 46 different plant species on average. Moreover, the degree of gene synteny does not significantly differ between intra- and inter-family syntenic comparisons (P=0.054, **Figure 1A**). Remarkably, for angiosperms only, syntenic conservation is even higher between than within plant families (P<0.001, **Figure 1B**) and crosses the boundaries of monocots and dicots. This trend is clearly divergent from the patterns commonly observed for plant genes (Zhao and Schranz, 2019).



C

	CesA	*CslA*	*CslC*	*CslB/H*	*CslG/J/M*	*CslE*	*CslF*	*CslD*
CesA	107910	0	0	101	114	24	3	16
CslA	0	72967	71	0	0	0	0	0
CslC	0	71	53350	0	0	0	0	0
CslB/H	101	0	0	5626	5	6	0	0
CslG/J/M	114	0	0	5	4935	46	0	5
CslE	24	0	0	6	46	11520	1	1
CslF	3	0	0	0	0	1	863	580
CslD	16	0	0	0	5	4	580	58876

Figure 1. The degree and structure of *CesA*/*Csl* genes synteny. Synteny is, overall, high and extensive across plant families, but arranged differently across gene families. (A) Distribution of the coefficients representing the average number of syntenic connections per gene in the *CesA*/*Csl* synteny network with other genes from the same (intrafamily) or from different (interfamily) plant species. Coefficients were calculated using data from all the plant families included in the analysis. (B) Distribution of the coefficients representing the average number of syntenic connections per gene in the *CesA*/*Csl* synteny network with other genes from the same (intrafamily) or from different (interfamily) plant species. Coefficients were calculated using data from only angiosperm families. (C) Total number of syntenic connections detected between gene families within the *CesA* superfamily.

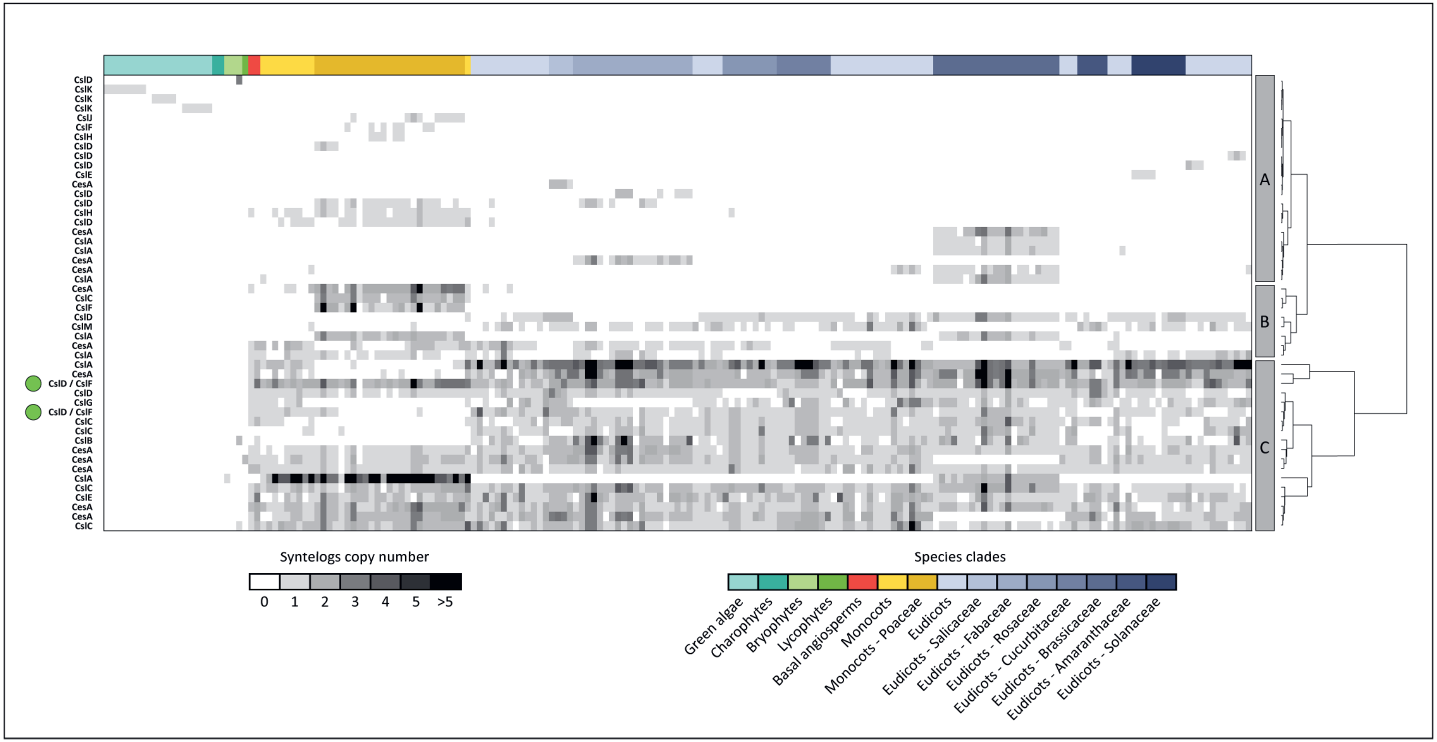


Figure 2. The 48 syntenic communities detected within the *CesA/Csl* synteny network. Syntenic communities are groups of genes displaying higher degrees of synteny with each other than with the rest of the network. Therefore, they constitute conserved architectural configurations of genes across genomes of different species. The figure highlights that *CesA/Csl* communities typically harbour genes belonging to the same *CesA/Csl* family, except for two communities grouping *CslF* and *CslD* genes (green dots on the left of the figure). *CesA/Csl* syntenic communities are also divided into three main clusters based on the spread of their genes across the plant kingdom (A–C panels and dendrogram on the right). In the figure, rows represent communities while columns represent species. Cells are coloured according to the number of syntenic genes harboured by each community and each species within communities. Row headers on the left side indicate the predominant gene family harboured by each community (>90% of the community members).

While synteny is dense and extensive across different plants, the same does not hold across different *CesA*/*Csl* families, which are organized in separate conserved genomic contexts. Accordingly, 99% of syntenic connections within the *CesA*/*Csl* synteny network takes place between genes from the same *CesA*/*Csl* family (**Figure 1C**). Moreover, each of the 48 syntenic communities identified by decomposing the *CesA*/*Csl* synteny network in groups of highly syntenic genes typically contains genes from one *CesA*/*Csl* family only (**Figure 2**). Furthermore, multiple communities were detected for most of the *CesA*/*Csl* families, revealing distinct conserved genomic contexts even at the level of gene subclades within *CesA*/*Csl* families (**Figure 2**). In this respect, it is noteworthy the marked differentiation in the genomic organization of the *CslA* genes between monocots (especially Poaceae) and dicots (group B and parts of Groups A and C, **Figure 2**). Moreover, syntenic communities for specific plant families were detected for some *CesA* and *CslA* genes within Salicaceae, Fabaceae, and Brassicaceae (Group A of **Figure 2**). Next to these examples of intra-family sub-organization of gene architecture, we also found communities that cross the boundaries of single *CesA*/*Csl* gene families (Group C of **Figure 2**). Specifically, two communities including 580 syntenic connections between *CslF* and *CslD* genes appeared particularly relevant, as they grouped ~40% of all the syntenic connections involving *CslF* genes (**Figure 2**).

3.2 Relationship between genomic architecture and gene properties

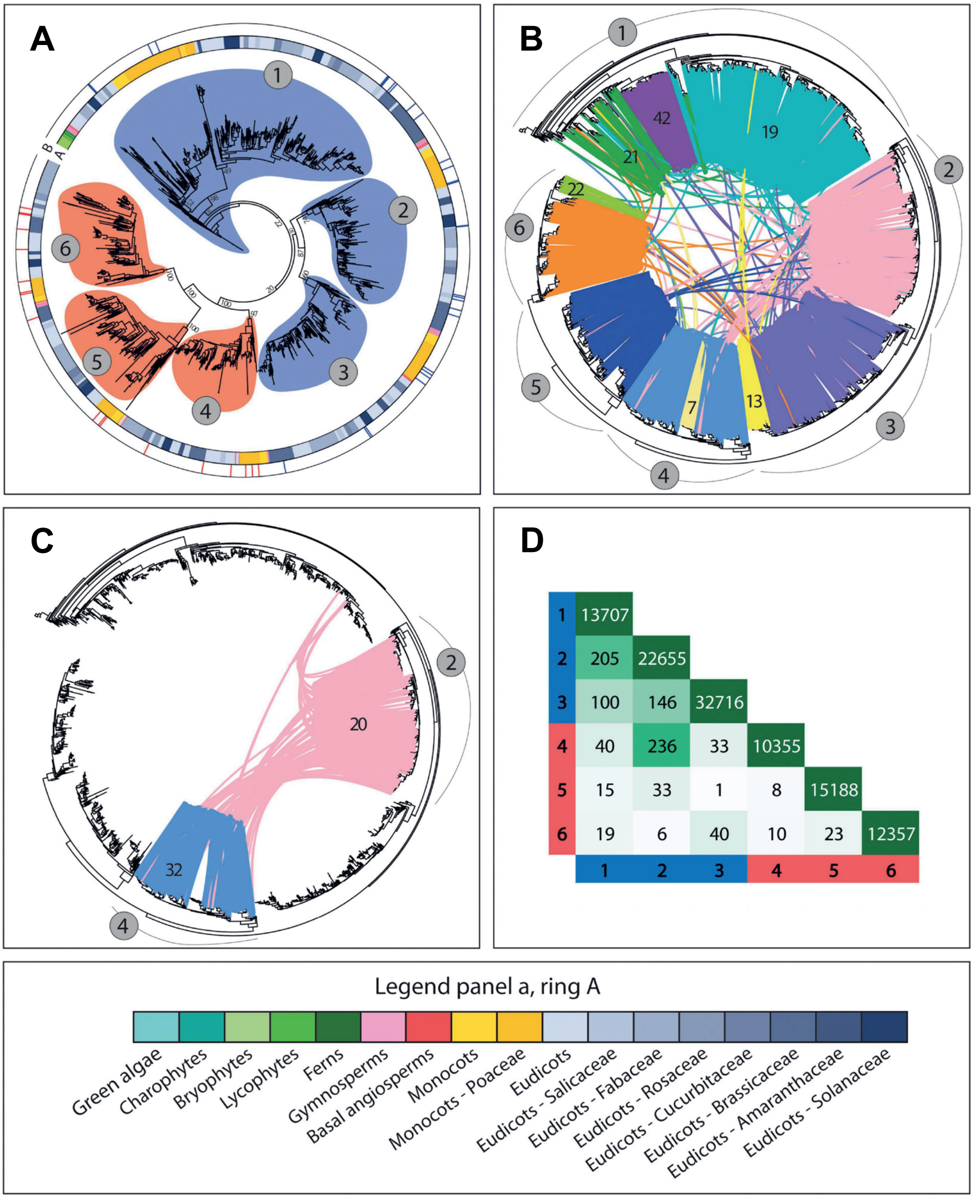
Distinct phylogenomic features for distinct CesA isoforms

To further characterize the genomic patterns described above and to study their relevance for the evolution and sub-functionalization of the *CesA*/*Csl* genes, we investigated how such patterns relate to phylogenetic and functional gene data. For the *CesA* family, our results revealed striking correspondence between the genomic organization of genes, the syntenic conservation of gene architecture, the phylogenetic relationships between genes, and the diversification of genes into distinct isoforms with different functional properties. Specifically, the *CesA* phylogeny was divided into six main clades supported by moderate-to-high bootstraps, corresponding to the six main groups of CesA isoforms known in *Arabidopsis thaliana* and *Oryza sativa* (**Figure 3A** and **5**). These six clades are grouped into two separated super-groups of three clades each corresponding to the subdivision into primary and secondary CW *CesA* genes (**Figure 3A**, **Figure 5** and **Supplementary Table 3**). Interestingly, each of the six phylogenetic clades is differentially organized and differentially conserved at the genomic level, by spanning only one of the six largest *CesA* syntenic communities found in the *CesA*/*Csl* synteny network (**Figures 3B** and **3D**). Overall, these six clades and communities span 84% of the syntenic *CesA* genes and cover 93% of the angiosperm families included in the synteny analysis. The correspondence between *CesA* phylogeny and synteny is therefore striking, with only one major exception found for clade 1 of **Figures 3A** and **3B**, which is subdivided into a main syntenic community largely dicot-specific (community 19 of **Figure 3b**) and other two smaller communities, of which one is specific to monocots (community 42 of **Figure 3B**) and one is a group of diverse *CesA* members (community 21 of **Figure 3B**). Besides this clade, only few other minor groups of genes specific to certain plant families deviate from the common phylogenetic and syntenic structure described above by being organized in distinct genomic contexts. Examples are a Fabaceae-specific community from clade 6 of **Figure 3B** (community 22 of **Figure 3B**), a Brassicaceae-specific community from clade 3 of **Figure 3B** (community 13 of **Figure 3B**), and a community from clade 4 of **Figure 3B** specific to Brassicaceae and Malvaceae (community 7 of **Figure 3B**).

A genomic context shift associated to specific CesA properties in cotton

Overall, the six different *CesA* phylogenomic clades are genomically independent from each other, with seldom inter-clade syntenic connections (**Figures 3B** and **3D**). An exception is represented by the 236 syntenic links between clades 2 (primary CW *CesAs*) and 4 (secondary CW *CesAs*) of **Figures 3B** and **3C**. A minor part of these connections is spread across different angiosperm families and can be regarded as background noise of the synteny analysis. However, 194 connections involve specifically the Malvaceae family and three cotton species therein: *G. hirsutum*, *G. arboreum* and *G. raimondii*. These connections mainly involve three primary and five secondary cotton *CesA* genes from branches 2 and 4, respectively, which are syntenic to a total of 104 *CesA* genes from 78 different angiosperm genomes placed in opposite phylogenetic groups (branch 4 for the cotton *CesAs* from branch 2, and branch 2 for the cotton *CesAs* on branch 4) (**Figure 3C**). This suggests that the cotton secondary CW *CesA* genes involved in these connections changed their genomic position relative to the other angiosperm genes from the same phylogenetic group, ending up in a genomic context typical of primary CW *CesAs* (clade 2 of **Figures 3B** and **3C**).

Remarkably, the *CesA* genes of *G. hirsutum* involved in the genomic inversion turned out to be associated with distinct functional properties at the basis of the massive fiber production of cotton. These are three different *G. hirsutum CesA* members: an already known *CesA7*, an already known *CesA8*, and a previously uncharacterized homolog of *GhCesA8* (99.6% sequence identity) (**Supplementary Tables 10-11**). Interestingly, these three isoforms have been proved to assemble together into functional rosettes for cellulose deposition in *G. hirsutum* fiber cells, with *CesA8* specifically acting as enhancer for massive fiber deposition (Li et al., 2016a, Li et al., 2013). In addition, the biological mechanism of fiber synthesis involving *G. hirsutum CesA7* and *CesA8* appears conserved across different cotton species, including the ones included in our analysis and showing the same genomic inversion (Li et al., 2013).



1	13707					
2	205	22655				
3	100	146	32716			
4	40	236	33	10355		
5	15	33	1	8	15188	
6	19	6	40	10	23	12357
	1	2	3	4	5	6

Figure 3. The phylogenomic structure of the *CesA* family. A) The phylogenetic tree of all the *CesA* genes from the 193 genomes with a BUSCO representation ≥75% and at least five genes per scaffold on average. The tree highlights the subdivision of the *CesA* sequences into six main phylogenetic groups corresponding to known primary and secondary cell wall *CesA* genes from different species. Rings A and B, respectively, represent the species taxonomy of the different genes (see legend) and the tree position of 49 experimentally validated primary (blue) and secondary (red) cell wall *CesA* genes. Circles around tree branches highlight the six major *CesA* clades detected in the tree, divided into three primary (blue) and three secondary (red) cell wall *CesA* clades. B) The phylogenetic tree of all the *CesA* genes as in panel A, with highlighted the syntenic connections between genes and the different syntenic communities detected. Lines connecting tree leaves indicate syntenic connections between genes, with different colours highlighting

different syntenic communities. Numbers in circles refer to gene clades as in panel A, while numbers on coloured syntenic connections indicate community numbers as referred in the article. The figure shows that the six major *CesA* clades of panel A are arranged into distinct conserved genomic contexts. C) Highlight of the syntenic connections between clades 2 (pink; primary cell wall *CesA* genes) and 4 (blue; secondary cell wall *CesA* genes). The pink links between the two phylogenetic clades indicate secondary cell wall *CesA* genes placed in genomic contexts typical of primary cell wall *CesAs*. These sequences belong mostly to Malvaceae species, and specifically to *G. hirsutum*, *G. arboreum*, and *G. raimondii*. D) Total number of syntenic connections detected between the six clades of *CesA* genes as represented in the tree of panel A. Row and column headers refer to the six clades of the tree in panel A.

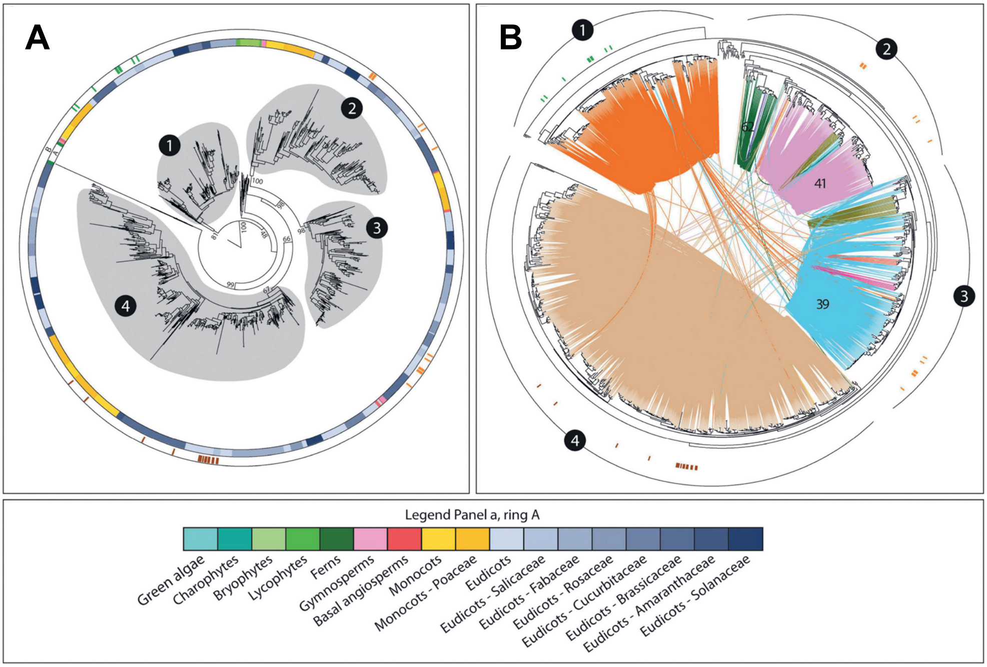


Figure 4. The phylogenomic structure of the *CslD* family. The trees highlight the correspondence between the phylogenetic and the conserved (syntenic) architectural structure of the family, and also the functional diversity of the *CslD* genes. A) The phylogenetic tree of all the *CslD* genes from the 193 genomes with a BUSCO representation ≥75% and at least five genes per scaffold on average. Ring A represents shows species taxonomy of the genes (see legend). Ring B represents the position of *CslD* genes known to be involved in the determination of vegetative organ size (green), pollen development (orange), or root hair formation (brown). Circles around tree branches highlight the four major *CslD* clades detected in the tree. Bootstrap values supporting clades subdivisions are reported. B) Syntenic structure of the four *CslD* phylogenetic clades. Lines connecting tree leaves indicate syntenic connections between genes, with different colours highlighting different syntenic communities. Numbers in black circles refer to the four clades of panel A, while numbers on syntenic connections indicate community numbers as referred in the article.

Accordingly, all the secondary CW CesA isoforms from *G. arboreum* and *G. raimondii* that are involved in the genomic context shift share 95.6% sequence identity with the *G. hirsutum CesA7* and *CesA8* for which functional data were available in scientific literature (**Supplementary Table 11**). Overall, these results suggest that the genomic positioning and organization of *CesA* genes might be critical for determining gene function, representing an important factor at the basis of cotton fiber deposition.

Differential genomic contexts for different gene functions in the CslD family

Striking correspondence between genomic organization, syntenic conservation, phylogenetic relationships, and functional diversification was also found within the *CslD* family. This family is structured into four main phylogenetic clades (**Figure 4A**) and previous studies demonstrated the involvement of different *CslD* members into three main plant processes – pollen development, determination of vegetative organ size, and root hair formation – across several species (Wang et al., 2001, Kim et al., 2007, Bernal et al., 2008, Li et al., 2009, Luan et al., 2011, Hunter et al., 2012, Hu et al., 2018, Li et al., 2018). Strikingly, the functional subdivision of the *CslD* members overlapped with the main phylogenetic *CslD* clades (**Figure 4**). In fact, known *CslD* genes involved in determination of vegetative organ size and root hair formation were only found within clades 1 and 4, respectively, while *CslD* members known to be involved in pollen development were only observed in clades 2 and 3 (**Figure 4**). In turn, the different phylogenetic and functional *CslD* groups corresponded to distinct and highly conserved *CslD* syntenic communities, highlighting their independent organization and conservation across plants (**Figure 4B**). However, while one syntenic community was found for each of the phylogenetic clades grouping *CslD* genes involved in organ size determination and root hair formation, respectively, multiple communities specific to different taxonomic groups were detected within the *CslD* branch associated to pollen development (**Figure 4B**). Of these, one is broad and spans both monocot and dicot species (community 39), one is large but restricted to dicots (community 41), and four are minor and restricted to specific taxonomic clades.

While the subdivision of *CslD* genes described above is shared by all angiosperms, the same does not hold for earlier land plants. In fact, *CslD* members from bryophytes and lycophytes are found in only a single phylogenetic group placed between the *CslD* branches associated to organ size determination and pollen development. Moreover, fern *CslDs* are divided into two groups that are closest to angiosperm *CslD* members involved in organ size determination and root hair formation, respectively. Finally, *CslD* genes from bryophytes, lycophytes, ferns and gymnosperms do not display any syntenic connection with the angiosperm ones.

Phylogenomic patterns in other Csl families

Commonalities between phylogenetic, syntenic, and functional patterns were also found in other *Csl* families. However, for the evolutionarily close *CslB/H/G/J/M/E* families such commonalities are observed for whole families rather than gene subclades within families (**Supplementary Figure 4**), with deviations observed only in a *CslM* syntenic community specific to Fabaceae (community 23 of **Supplementary Figure 4**) and a small *CslE* community with only *Capsicum* genes (community 17 of **Supplementary Figure 4**). Finally, *CslA* and *CslC* genes also display syntenic suborganization of their phylogenies (**Supplementary Figure 5**). However, the functional information available for these families are not abundant and do not reveal any clear correspondence with the phylogenetic and/or syntenic structure observed.

3.3 Evolutionary dynamics of specific gene families

Evolution of the CesA genes

Phylogenetic and synteny data were used to also study the evolution of the main *CesA*/*Csl* families. For the *CesA* genes, the taxonomic profiling of the six main *CesA* clades (see Section 2.2.1) revealed that *CesA* sequences of bryophytes and lycophytes are phylogenetically closest to the primary CW *CesA* branches, and specifically to the primary CW *CesA* genes homolog to the redundant arabidopsis *CesA2/5/6/9* and the *O. sativa CesA3/5/6* (clade 1 of **Figure 5**). This observation suggests that primary CW *CesAs* are the oldest *CesA* sequences of land plants, and that early land plants likely had only primary CW *CesA* (-like) genes. The positioning of fern *CesAs* suggests that the diversification of *CesA* genes towards the six main angiosperm clades described in Section 2.2.1 started somewhen during late lycophyte or early fern evolution. In fact, while a group of fern *CesAs* is also positioned close to the bryophyte and lycophyte sequences of clade 1 of **Figure 5**, another group of fern genes is preceding the split of the other two clades of angiosperm primary CW *CesAs* (clades 2 and 3 of **Figure 5**). Moreover, a third group of fern *CesAs* is closest to the three groups of angiosperms secondary CW *CesAs* (clades 4, 5, and 6 of **Figure 5**), just preceding their diversification. Therefore, *CesAs* diversification towards the different phylogenomic groups observed in higher plants progressed significantly during fern evolution. However, gymnosperms are the first group of plants whose *CesA* sequences are found in all the six *CesA* clades genomically conserved across all the angiosperms. To conclude, data on gene copy-number within each of the six *CesA* clades also support the evolutionary model just discussed. In fact, the early-diverging clade 1 contains by far the most *CesA* sequences in both **Figure 3A** and **Figure 5** (771 sequences in the full *CesA* tree of **Figure 3A**), followed by clades 2 and 3 of primary CW *CesAs* (394 genes on average, CV = 4.3%), and finally by the three secondary CW *CesA* clades (273 genes each on average, CV = 6.3%). These data, together with the presence of redundant arabidopsis *CesA* copies within two primary CW *CesA* clades (1 and 2 of **Figure 5**), agree with the view that oldest *CesA* families evolved first into multiple isoforms, of which some underwent full sub-functionalization within the cellulose

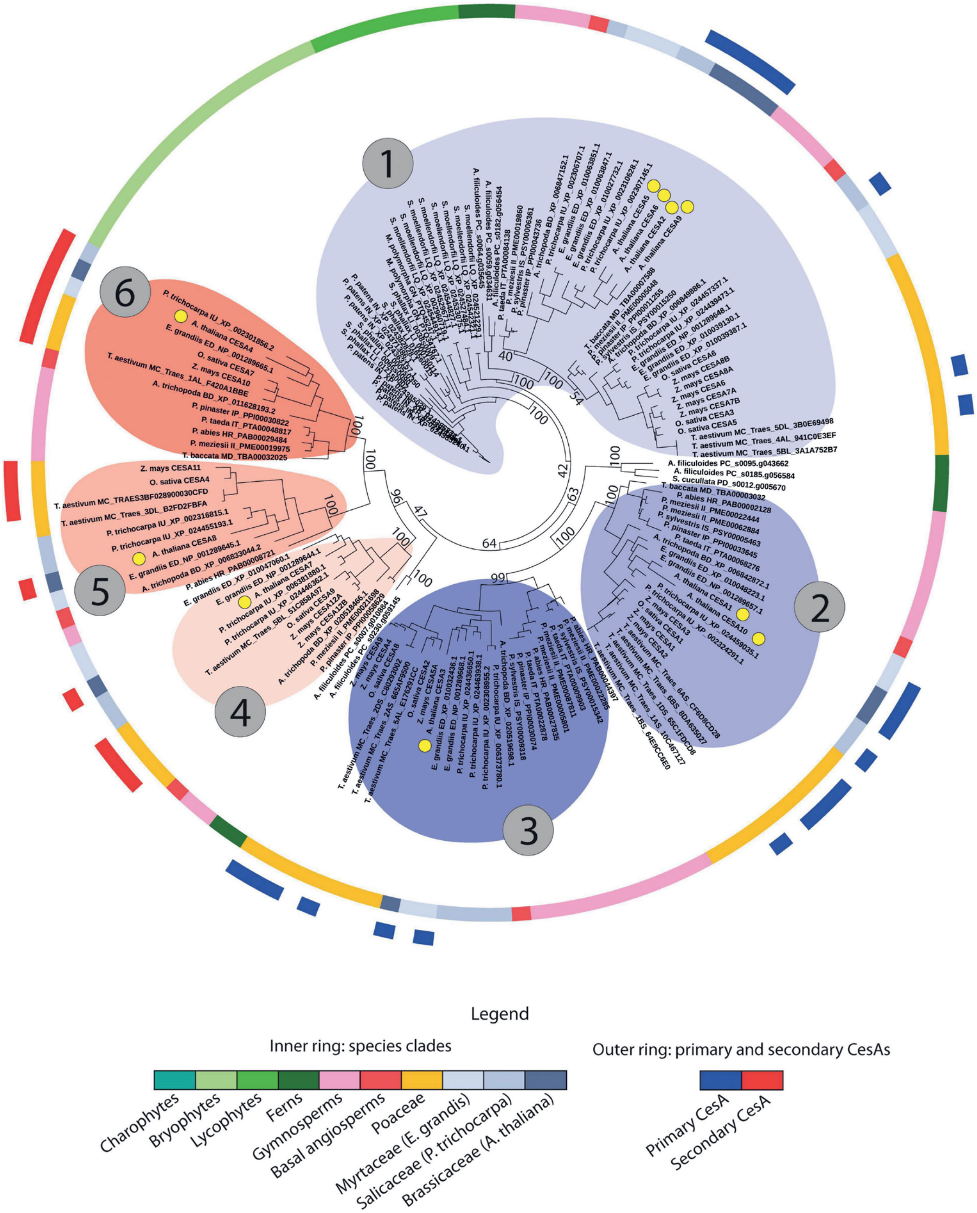
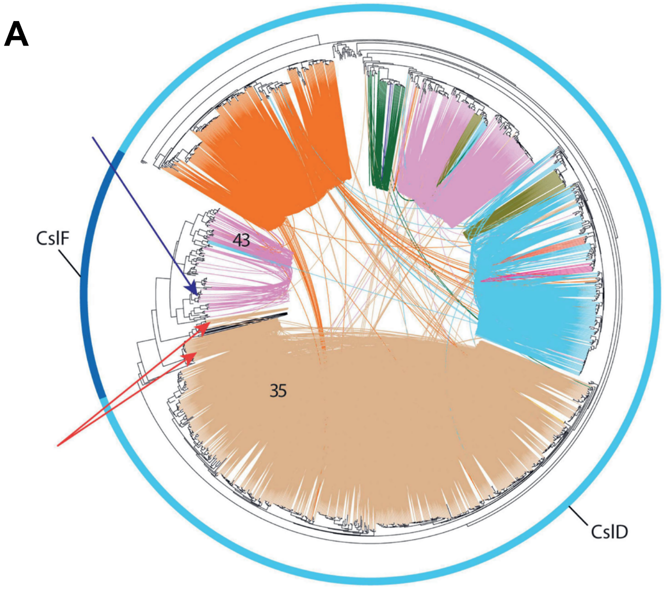
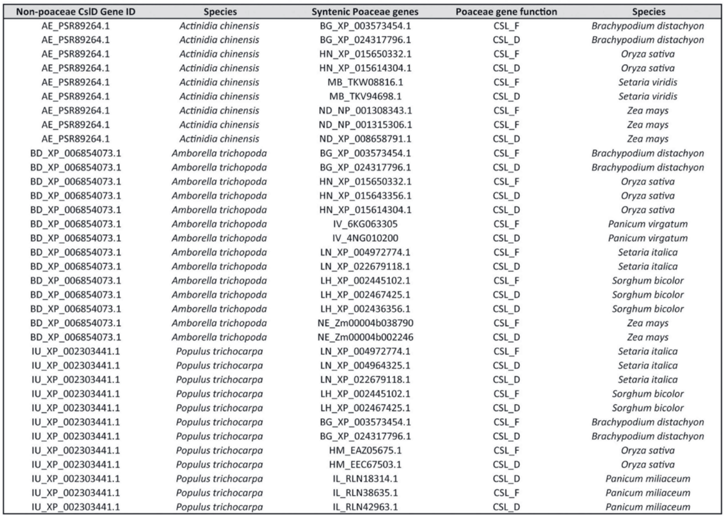


Figure 5. Phylogenetic tree of a subset of 151 *CesA* genes including sequences from model plant species, from species where primary and secondary cell wall *CesA* genes have been experimentally validated, along with all gymnosperm, fern, lycophyte, bryophyte, and charophyte *CesA* sequences. The tree shows that early-diverging plant sequences are only occurring closest to primary cell wall *CesA* branches (especially clade 1). Moreover, it highlights the subdivision of the *CesA* sequences into the six main phylogenetic groups corresponding to known primary and secondary cell wall CesA isoforms from different species. The inner ring indicates taxonomy, while the outer ring highlights the position of experimentally validated primary and secondary cell wall *CesA* genes (see legend at the bottom of the figure). Coloured clades around branches indicate the six main *CesA* clades, numbered in the same way as in Figure 3. Blue clades indicate primary cell wall *CesA* genes, while red clades indicate secondary cell wall *CesA* sequences. Bootstraps

supporting clade subdivisions are reported. Yellow dots indicate the position of the 10 *Arabidopsis thaliana CesA* genes.





Non-poaceae CslD Gene ID	Species	Syntenic Poaceae genes	Poaceae gene function	Species
AE_PSR89264.1	*Actinidia chinensis*	BG_XP_003573454.1	CSL_F	*Brachypodium distachyon*
AE_PSR89264.1	*Actinidia chinensis*	BG_XP_024317796.1	CSL_D	*Brachypodium distachyon*
AE_PSR89264.1	*Actinidia chinensis*	HN_XP_015650332.1	CSL_F	*Oryza sativa*
AE_PSR89264.1	*Actinidia chinensis*	HN_XP_015614304.1	CSL_D	*Oryza sativa*
AE_PSR89264.1	*Actinidia chinensis*	MB_TKW08816.1	CSL_F	*Setaria viridis*
AE_PSR89264.1	*Actinidia chinensis*	MB_TKV94698.1	CSL_D	*Setaria viridis*
AE_PSR89264.1	*Actinidia chinensis*	ND_NP_001308343.1	CSL_F	*Zea mays*
AE_PSR89264.1	*Actinidia chinensis*	ND_NP_001315306.1	CSL_F	*Zea mays*
AE_PSR89264.1	*Actinidia chinensis*	ND_XP_008658791.1	CSL_D	*Zea mays*
BD_XP_006854073.1	*Amborella trichopoda*	BG_XP_003573454.1	CSL_F	*Brachypodium distachyon*
BD_XP_006854073.1	*Amborella trichopoda*	BG_XP_024317796.1	CSL_D	*Brachypodium distachyon*
BD_XP_006854073.1	*Amborella trichopoda*	HN_XP_015650332.1	CSL_F	*Oryza sativa*
BD_XP_006854073.1	*Amborella trichopoda*	HN_XP_015643356.1	CSL_D	*Oryza sativa*
BD_XP_006854073.1	*Amborella trichopoda*	HN_XP_015614304.1	CSL_D	*Oryza sativa*
BD_XP_006854073.1	*Amborella trichopoda*	IV_6KG063305	CSL_F	*Panicum virgatum*
BD_XP_006854073.1	*Amborella trichopoda*	IV_4NG010200	CSL_D	*Panicum virgatum*
BD_XP_006854073.1	*Amborella trichopoda*	LN_XP_004972774.1	CSL_F	*Setaria italica*
BD_XP_006854073.1	*Amborella trichopoda*	LN_XP_022679118.1	CSL_D	*Setaria italica*
BD_XP_006854073.1	*Amborella trichopoda*	LH_XP_002445102.1	CSL_F	*Sorghum bicolor*
BD_XP_006854073.1	*Amborella trichopoda*	LH_XP_002467425.1	CSL_D	*Sorghum bicolor*
BD_XP_006854073.1	*Amborella trichopoda*	LH_XP_002436356.1	CSL_D	*Sorghum bicolor*
BD_XP_006854073.1	*Amborella trichopoda*	NE_Zm00004b038790	CSL_F	*Zea mays*
BD_XP_006854073.1	*Amborella trichopoda*	NE_Zm00004b002246	CSL_D	*Zea mays*
IU_XP_002303441.1	*Populus trichocarpa*	LN_XP_004972774.1	CSL_F	*Setaria italica*
IU_XP_002303441.1	*Populus trichocarpa*	LN_XP_004964325.1	CSL_D	*Setaria italica*
IU_XP_002303441.1	*Populus trichocarpa*	LN_XP_022679118.1	CSL_D	*Setaria italica*
IU_XP_002303441.1	*Populus trichocarpa*	LH_XP_002445102.1	CSL_F	*Sorghum bicolor*
IU_XP_002303441.1	*Populus trichocarpa*	LH_XP_002467425.1	CSL_D	*Sorghum bicolor*
IU_XP_002303441.1	*Populus trichocarpa*	BG_XP_003573454.1	CSL_F	*Brachypodium distachyon*
IU_XP_002303441.1	*Populus trichocarpa*	BG_XP_024317796.1	CSL_D	*Brachypodium distachyon*
IU_XP_002303441.1	*Populus trichocarpa*	HM_EAZ05675.1	CSL_F	*Oryza sativa*
IU_XP_002303441.1	*Populus trichocarpa*	HM_EEC67503.1	CSL_D	*Oryza sativa*
IU_XP_002303441.1	*Populus trichocarpa*	IL_RLN18314.1	CSL_D	*Panicum miliaceum*
IU_XP_002303441.1	*Populus trichocarpa*	IL_RLN38635.1	CSL_F	*Panicum miliaceum*
IU_XP_002303441.1	*Populus trichocarpa*	IL_RLN42963.1	CSL_D	*Panicum miliaceum*

Figure 6. The 1:2 syntenic connections between a group of angiosperm *CslD* genes and a class of Poaceae *CslD* and *CslF* genes. A) Phylogenetic tree of all the *CslD* and *CslF* genes from the 193 genomes with a BUSCO representation ≥75% and at least five genes per scaffold on average. The outer ring highlights *CslD* and *CslF* tree branches. Red arrows indicate the *CslF* genes syntenic to *CslD* isoforms involved in root morphology (same syntenic community as in Figure 4). Blue arrow indicates the positioning of *HvCslF3* and *HvCslF10*, the two *CslF* genes responsible for the synthesis of 1,4-β-linked glucoxylan in barley. B) Table reporting examples of syntenic connections between three *CslD* genes from three different angiosperm species and both *CslF* and *CslD* genes across different Poaceae species.

synthesis machinery (a model known as constructive neutral evolution, recently advanced by Haigler and Roberts (2019) for the *CesA* genes – see discussion).

Synteny reveals the genomic trace of CslF evolution

In the *CslF* family, our phylogenomic analysis detected 580 syntenic connections between the *CslF* genes organized as singletons or tandems in the Poaceae genomes (see Section 3.1.2) and a subset of *CslD* members, all included within the same syntenic community (community 35 in **Figure 6** and green dots in **Figure 2**). Interestingly, 327 of these connections involve *CslD* genes from 53 different eudicot species (27 different plant families) that display simultaneous synteny with a grass *CslD* gene and a grass *CslF* member in 17 of the 21 Poaceae genomes of our study (**Figure 6B**). The frequency of these connections is negatively correlated with the evolutionary distance of the species harbouring "seed" *CslD* sequences from Poaceae (r = –0.57). Moreover, the "seed" *CslD* genes all belong to the *CslD* phylogenomic clade grouping *CslD* genes involved in root hair formation (see Section 2.2.3). All together, these observations reveal the genomic trace of the *CslF* origin in grasses. On the one hand, they confirm that the *CslF* genes originated through the duplication of some *CslD* members (Yin et al., 2009, Schwerdt et al., 2015, Little et al., 2018). On the other hand, they show that the *CslF* family is nested within the *CslD* one (and specifically within the *CslD* clade involved in root hair formation). This in turns proves that the *CslF* genes took origin when the *CslD* family was already formed, thanks to a relaxed evolutionary pressure on duplicated *CslD* members in grasses. To conclude, it is noteworthy that no synteny was detected between the *CslF* genes syntenic to *CslDs* (red arrows in **Figure 6**) and the other *CslF* members included in community 43 (**Figure 6**), which form the conserved *CslF* genomic array (see Section 2.1.2). The fact that only some *CslF* genes – corresponding to one of the two architectural configurations of this family – display synteny with *CslD* members questions whether the *CslF* family originated through a single *CslD* duplication, or if the family underwent different rounds of expansion in Poaceae. In this respect, the phylogenomic positioning of the barley *CslF* members recently shown to synthesize (1,4)-β-linked glucoxylans was checked in the tree of **Figure 6A** (blue arrow). Interestingly, both *HvCslF3* and *HvCslF10* are positioned within the phylogenetic clade spanned by community 43, which does not display synteny with *CslD* members.

4 Discussion

4.1 The genomic architecture and the evolution of the *CesA* superfamily

In Section 2.1 we showed that the genomic architecture of the *CesA* superfamily varies considerably in terms of gene copy number, gene distribution along genomes and gene synteny across different plant and gene families. These results provide insights into the timing and modes of *CesA* superfamily evolution. Specifically, a key finding is that the *CesA*/*Csl* copy number variation correlates with the evolutionary timing of the *CesA* superfamily. Accordingly, we found that *CesA* is the largest *CesA*/*Csl* family, followed by the *CslD/F*, *CslA*, and *CslC* families, and finally by the *CslB*/*H*, *CslG*/*M*, and *CslE* genes. These groups represent respectively the oldest, the intermediate, and the most recent *CesA*/*Csl* families in evolutionary terms (Banasiak, 2014, Little et al., 2018). Moreover, we also found positive correlation between *CesA*/*Csl* copy number and whole genome duplications (WGDs) across species, suggesting that WGDs had a prominent role in driving *CesA*/*Csl* expansion and diversification. All together, these observations imply that novel duplicated *CesA*/*Csl* genes were typically retained by positive selection, contributing to expand, diversify and sub-functionalize the *CesA* superfamily. Interestingly, Schwerdt et al. (2015) analysed nucleotide substitution rates of *CesA*/*Csl* genes from four grasses and found evidence of past positive selection for several of them followed by genomic stabilization, supporting our conclusion.

Since positive selection takes typically place when recently-duplicated genes provide adaptive advantages (Demuth and Hahn, 2009, Jensen and Bachtrog, 2010), it is relevant to investigate the advantages brought by novel *CesA/Csl* genes to plants. In this respect, we demonstrated that the *CesA* and *CslD* diversification accompanied sub-functionalization into primary and secondary *CesA* isoforms and different *CslD* functional groups (i.e. pollen development, determination of vegetative organ size, and root hair formation), respectively. Moreover, conserved phylogenomic patterns were found across all the main *CesA/Csl* families. We therefore hypothesize that the expansion and diversification of the *CesA* superfamily sustained the increased cell wall diversity during plant evolution, which in turn drove the evolution of land plants (Sarkar et al., 2009, Sørensen et al., 2010). In fact, cellulose and hemicelluloses have been fundamental for both the rise of land plants (by serving as sinks of photosynthesized carbohydrates and by providing mechanical support) and the rise of vascular and flowering plants (by modulating cell wall composition for vascularization, evolution of flowers and fruits, and pathogen resistance) (Sarkar et al., 2009). Remarkably, while all the major *CesA*/*Csl* families have members in basal land plants, the largest expansion of all the families was observed at the rise of flowering plants, fitting the evolutionary model just discussed.

The importance of *CesA*/*Csl* diversification in higher plants evolution was highlighted also by synteny data. Specifically, a consistent part of *CesA*/*Csl* syntenic communities is conserved across most angiosperms (Group C in **Figure 2**), suggesting genomic fixation of the successful diversification of *CesA*/*Csl* families into a highly conserved genomic background. This is remarkable, since the synteny of most angiosperm genes is typically lineage-specific (Zhao and Schranz, 2019). On the other hand, the fact that lineage-specific syntenic communities were nonetheless found (Groups A and B in **Figure 2**) demonstrates that *CesA*/*Csl* evolution is still dynamic and novel evolutionary trajectories take place on top, rather than in replacement of, the common and stabilized genomic background. Lineage-specific gene arrangements may thus contribute to CW variability across plants. In this sense, the divergent genomic organization found in *CslA* genes between monocots and eudicots appears particularly marked, and we hypothesize that could be at the basis of the profound differences in mannan and glucomannan content between these groups of plants (Vogel, 2008). This is further supported by the absence of clear differences between monocots and eudicots in terms of mere *CslA* copy number or gene sequence variability.

4.2 Conserved genomic contexts in the *CesA* superfamily: biological implications

A major finding of this research is the association between phylogenetic patterns, conservation (synteny) of such patterns, and gene sub-functionalization found in the *CesA* and *CslD* families. Moreover, similar patterns were detected in other *Csl* families, suggesting that such associations might hold also for other genes, but the absence of sufficient functional data hampered their detection. These observations question whether synteny is merely a trace of past evolutionary dynamics of *CesA*/*Csl* genes or it can proactively contribute to determine gene function. In this respect, several elements support the second alternative. First, the exceptional extensiveness of *CesA*/*Csl* synteny compared to the typical lineage-specific synteny of angiosperm genes (Zhao and Schranz, 2019) suggests that selective constraints act against the positional genomic reshuffling of the *CesA* superfamily. In turn, this may indicate that synteny contributes to determine and conserve gene functions and adaptations, as hypothesized for other highly conserved gene families (Zhao et al., 2017, Kerstens et al., 2020). Furthermore, if this hypothesis is correct, positional constraints would likely be represented by tight mechanisms of gene regulation based on (epi)genetic properties variable along genomes and potentially disrupted by genome reshuffling (Gould et al., 2018, Crow et al., 2020, Lai et al., 2020). In this respect, several studies showed that cell wall biosynthesis is indeed hierarchically tightly regulated by the concerted action of complex plant gene networks (Taylor-Teeples et al., 2015, Rao and Dixon, 2018, Zhang et al., 2018), and some *CesA* genes were shown to be targets of defined loops within such networks (Taylor-Teeples et al., 2015). Therefore, positional reshuffling of the *CesA*/*Csl* genes may significantly affect the deposition of cellulose and hemicellulosic polymers, which is strikingly what we found for the cotton secondary *CesA* genes transposed to conserved genomic contexts typical of primary *CesAs* in other plant species. All together, these observations support a central role for the *CesA* superfamily genomic architecture in determining gene function. However, given the fragmented evidence across the different *CesA*/*Csl* families – which is mostly due to the scarce functional characterization of several *Csl* genes – such role should be better investigated in the future.

4.3 The evolution of the *CesA* family

Based on the phylogenomic analyses, we proposed a novel evolutionary model for the *CesA* family, which hypothesizes that primary CW *CesAs* evolved before the secondary CW *CesAs* (**Figure 5**). These findings are in contrast with previous studies that claimed secondary *CesA* genes as the oldest ones by observing the primary CW *CesAs* phylogenetically nested within them (Schwerdt et al., 2015, Little et al., 2018). In our case, all the *CesA* phylogenetic trees (**Figures 3** and **5**) positioned bryophyte and lycophyte *CesA* sequences within primary CW *CesA* branches (specifically clade 1 of **Figures 3** and **5**), providing strong evidence in support of our model. Furthermore, the currently most-accepted framework of plant and cell wall evolution also agrees with our model. First, the detection of three groups of primary and three groups of secondary CesA isoforms only in gymnosperms and angiosperms agrees with the pivotal role of secondary CW *CesAs* in the evolution of tall stems, wood, and vessels (Speck and Burgert, 2011, Cosgrove, 2012). Second, the fact that primary CWs alone sufficed the developmental needs of bryophytes and lycophytes (Sarkar et al., 2009) agrees with the finding of only primary *CesA* (-like) genes in these plants. Third, the intermediate evolutionary stage of ferns, with a group of sequences clustered within clade 1 of **Figure 5**, another group preceding the split of the other two primary *CesA* clades, and a final group basal to the division of the three secondary CW *CesA* clades, agrees with the finding of *ancestral* conductive vessels in these species, with marked differences from the conductive vessels of angiosperms (Carlquist and Schneider, 2001, Sarkar et al., 2009). To conclude, the overall positioning of bryophyte/lycophyte, fern, and gymnosperm/angiosperm genes in the trees of **Figures 3A** and **5** suggests that the *CesA* family evolved over a long time, with the likely recurrent stabilization of its members into different intermediate diversified forms. Interestingly, this agrees with the most recent hypotheses on the evolution of the Cellulose Synthase Complexes (CSCs), which state that the hetero-oligomeric CSCs of higher plants (i.e. formed by different CesA isoforms) arose through constructive neutral evolution of ancestral homo-oligomeric complexes (CSCs of chlorophytes are formed by linear combination of interchangeable CesA proteins) pushed to differentiation by the diversification of their CesA subunits (Haigler and Roberts, 2019). While the sequence diversity of our bryophyte and lycophyte CesA sequences does not allow us to conclude whether they represent a single isoform, their positioning basal to the primary CW *CesA* branch in **Figures 3A** and **5** does not exclude this possibility. Therefore, the real morphology of CSCs in early-diverging land plants should be better elucidated in the future to answer the latter question.

However, we think that the model of Haigler and Roberts (2019) overall agrees with the phylogenomic patterns observed in this research and further supports our hypotheses.

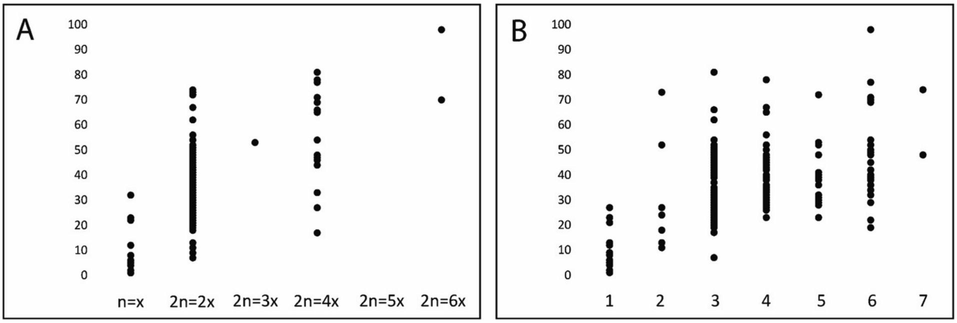
4.4 The evolutionary relationship between the *CslD* and *CslF* families

The syntenic relationships between some *CslD* and *CslF* genes represent a last important finding of our research. Specifically, while the origin of the *CslF* family through the duplication of *CslD* members had been already reported based on phylogenetics (Yin et al., 2009, Schwerdt et al., 2015, Little et al., 2018), our results unveiled the genomic trace of this evolutionary link. This trace lays in the 1:2 syntenic connections found between a set of eudicot *CslD* members and both *CslD* and *CslF* genes from Poaceae. Since synteny involves both eudicot and grass *CslD* members, our results demonstrate that *CslF* genes originated when the *CslD* family was already formed, thanks to the relaxed selection pressure on duplicated grass *CslD* copies. Moreover, since synteny between *CslD* and *CslF* genes is found across all the Poaceae species in our analyses, the duplication at the origin of the *CslF* family may be shared by all the Poaceae. Therefore, assuming that a WGD event was responsible for this, such event could be the so-called ρ WGD, which is at the origin of the whole Poaceae lineage and took place about 56-70 Mya (Clark and Donoghue, 2018, Lee et al., 2020). To conclude, it remains unclear whether the *CslD* duplication described in this research originated all or only a part of the *CslF* family. In fact, the *CslF* genes syntenic to *CslD* members are only the ones genomically organized as singletons or in small tandem arrays, while no synteny between *CslF* and *CslD* genes was detected for the *CslF* members arranged into the large conserved *CslF* array (see Section 2.1.2). Interestingly, such large conserved *CslF* array contains the two barley *CslF* genes that were recently found to synthesize a modified form of 1,4-linked polysaccharides termed (1,4)-β-linked glucoxylans (Little et al., 2019). Altogether, these observations indicate that (different parts of) the *CslF* family may have formed or expanded repeatedly during grass evolution, but the data produced in this research fail to highlight the specific modes of such evolution. Moreover, these dynamic evolutionary patterns involving differential genomic arrangement of homologous genes may be at the basis of the formation of novel cell wall polysaccharides, as for the (1,4)-β-linked glucoxylans of barley. However, the presence of this polysaccharide and of *CslF* genes responsible for its synthesis in species other than barley should be confirmed in the future to fully support the statement above. In conclusion, since the *CslF* genes syntenic with the *CslD* members are the phylogenetically oldest *CslFs* (**Figure 6**), we can confidently state that the *CslD* duplication was the first event in initiating the *CslF* family.

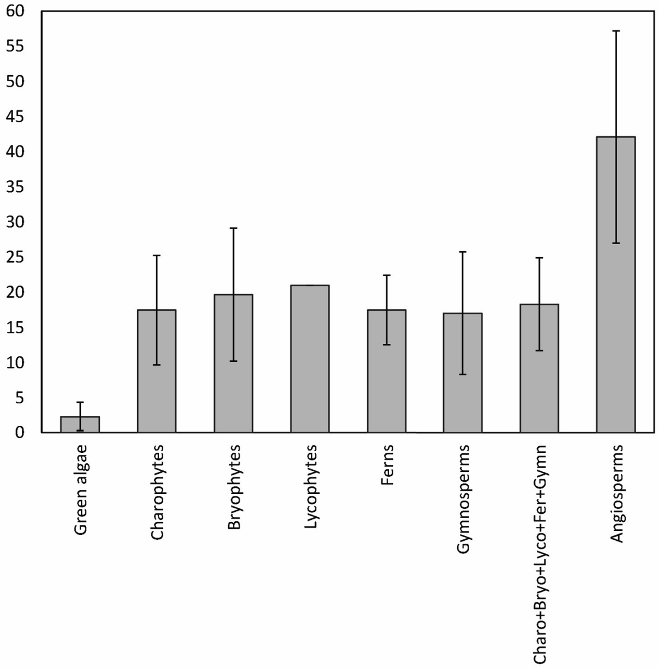
5 Conclusions

In this research, we took advantage of novel bioinformatic tools for large-scale synteny, genomic, and phylogenetic analysis to improve our knowledge on the genetics, evolution, and functional characterization of the *CesA* superfamily. The tools we used – especially the combined analysis of synteny, phylogenetic and functional data – provided the power to dissect the aspects just mentioned in the complex gene families at the basis of plant cell walls. However, the current level of functional characterization of many genes within the *CesA* superfamily is often hampering an effective use of the methodologies described here. Therefore, we foresee that progress in the functional characterization of cell wall genes through reverse genetics will be crucial to further advance our fundamental knowledge on these genes. In addition, future research can take advantage of our results to test the hypotheses we raise and close the open gaps in our understanding of the *CesA* superfamily.

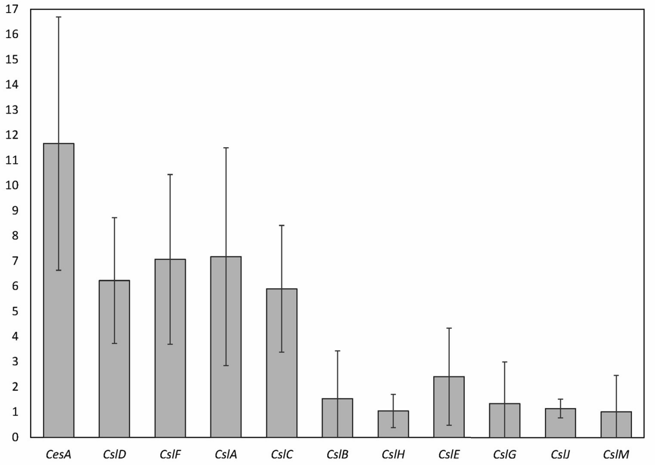
Supplementary Data



Supplementary Figure 1.
A) Scatterplot of the ploidy level (x-axis) against the total number of *CesA*/*Csl* genes for the 222 plant genomes of the study with a BUSCO representation >75% (Viridiplantae set; y-axis).
B) Scatterplot of the total number of WGDs to which a species has undergone along plant evolution (x-axis) against the total number of CesA/Csl genes for the 222 plant genomes of the study with a BUSCO representation >75% (Viridiplantae set; y-axis). WGDs data were retrieved from Clark and Donoghue (2018).



Supplementary Figure 2. Average number of *CesA*/*Csl* genes (y-axis) across major plant evolutionary clades (x-axis). Error bars represent the standard deviation of the mean. The data represented include the 222 plant genomes of the study with a representation >75% (Viridiplantae set).



Supplementary Figure 3. Average number of genes across the 222 plant genomes of the study with a BUSCO representation >75% (Viridiplantae set; y-axis) for the 11 major gene families within the *CesA* superfamily. Error bars represent the standard deviation of the mean. Data for the green algae-specific *CslK* family are not shown.

Supplementary Figure 4 (Page 45).

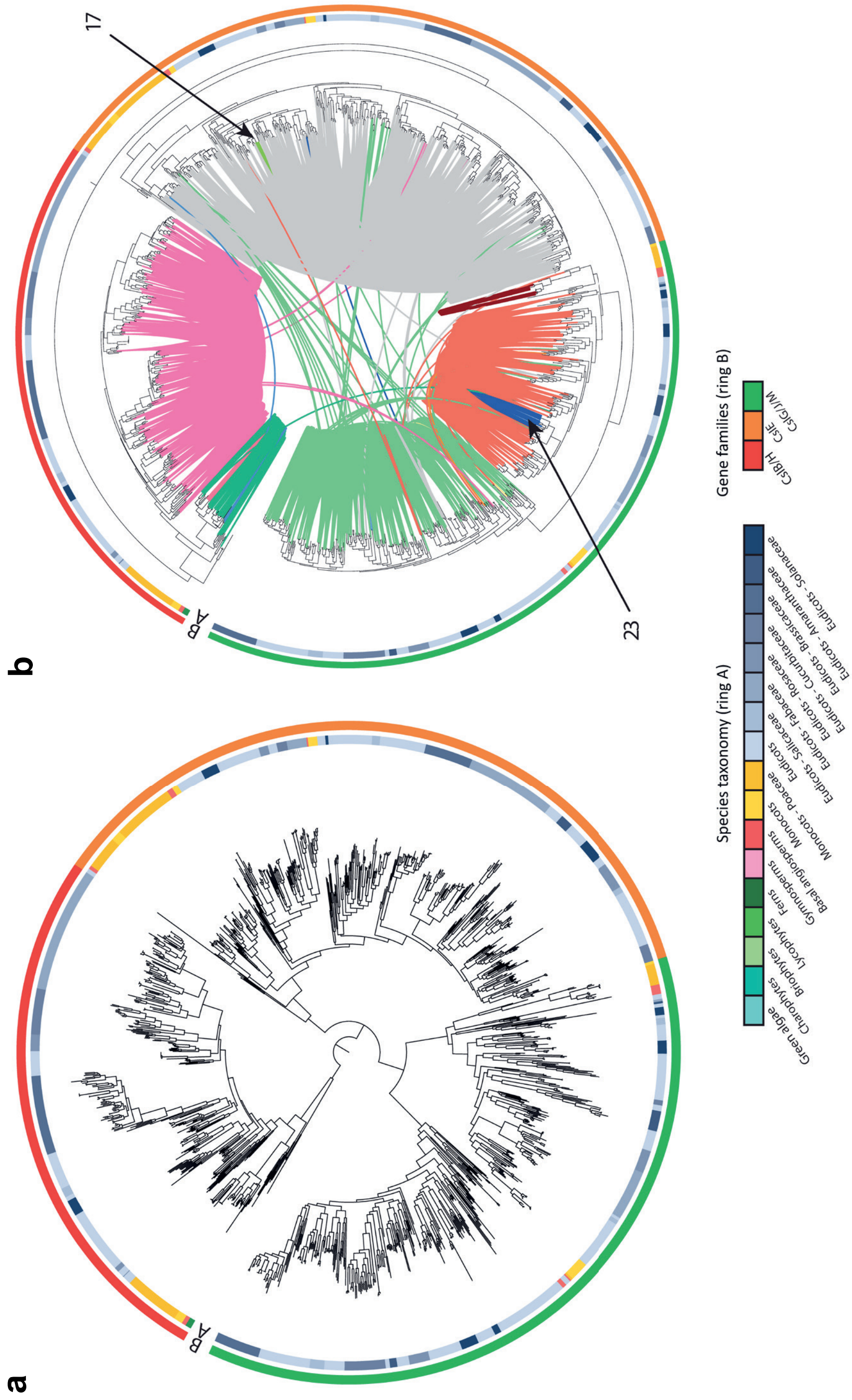
a) The phylogenetic tree of the *CslB*/*H*, *CslG*/*J*/*M* and *CslE* families. Ring A displays species taxonomy, while ring B displays gene families (see legend).

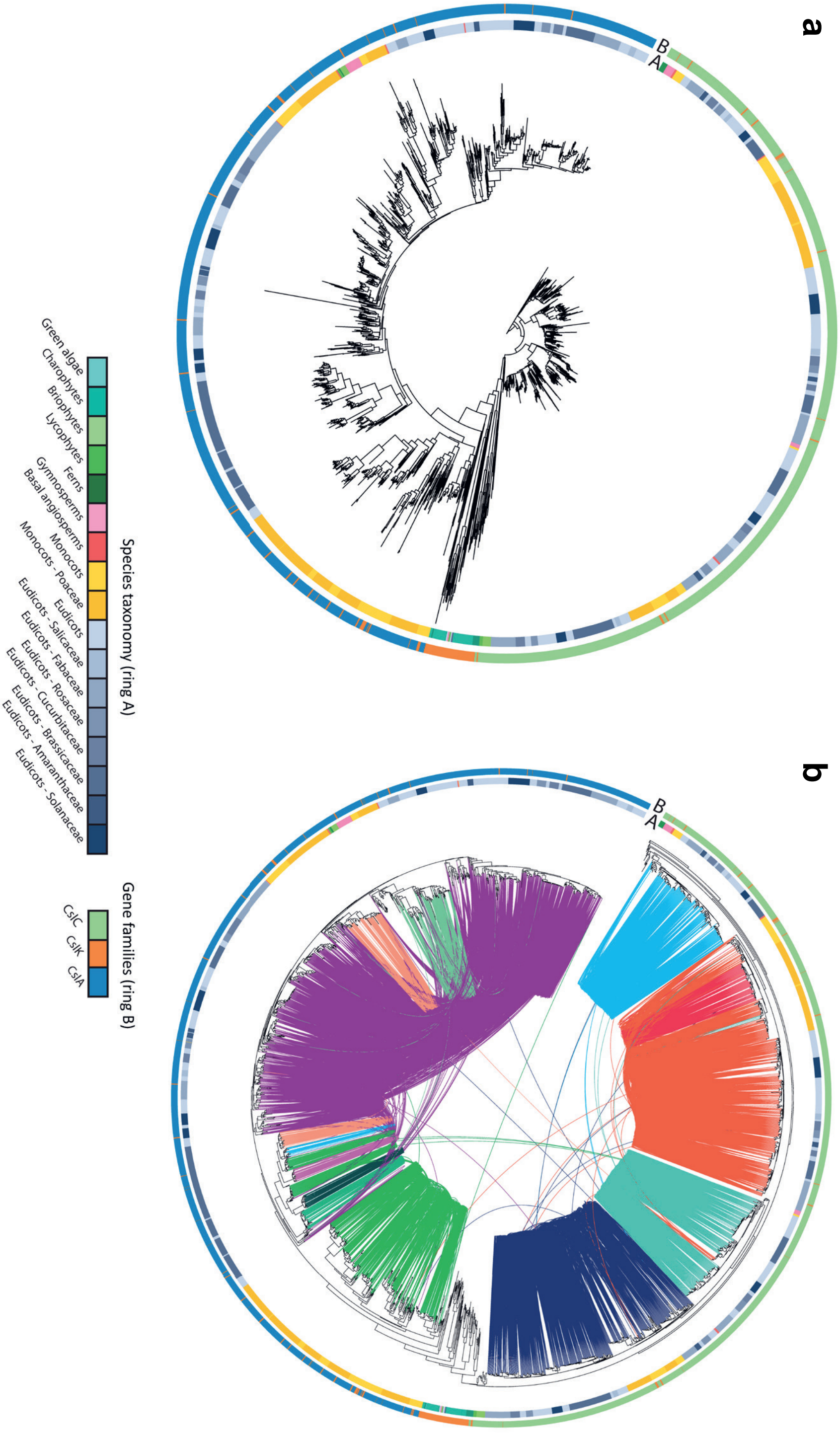
b) The phylogenetic tree of the *CslB*/*H*, *CslG*/*J*/*M* and *CslE* families with the annotation of syntenic relationships between genes grouped by syntenic communities. Like in panel a, ring A indicates gene taxonomy, while ring B indicates gene families (see legend). The different colours of inner connections indicate distinct syntenic gene communities. Black arrows indicate syntenic communities 17 and 23, which display deviations from the commonalities observed between phylogenetic, syntenic, and functional patterns of these genes.

Supplementary Figure 5 (Page 46).

a) The phylogenetic tree of the *CslA*/*C*/*K* families. Ring A displays species taxonomy, while ring B displays gene families (see legend). The orange bars within *CslA* and *CslC* areas indicate *Csl* genes that were not univoquely assigned to either *CslA* or *CslC* families by BLAST and/or HMMER search (and, therefore, regarded as *CslK* genes). However, phylogeny places them within the *CslA* and *CslC* groups. “True” *CslK* genes are represented by the group of orange *Csl* sequences on the right side of ring B.

b) The phylogenetic tree of the *CslA*/*C*/*K* families with the annotation of syntenic relationships between genes, grouped by syntenic gene communities. Ring A displays species taxonomy, while ring B displays gene families (see legend). The different colours of internal connections indicate distinct syntenic gene communities.


a
b
A
B
17
23
Species taxonomy (ring A)
Eudicots - Solanaceae
Eudicots - Amaranthaceae
Eudicots - Brassicaceae
Eudicots - Cucurbitaceae
Eudicots - Rosaceae
Eudicots - Fabaceae
Eudicots - Salicaceae
Eudicots
Monocots - Poaceae
Monocots
Basal angiosperms
Gymnosperms
Ferns
Lycophytes
Briophytes
Charophytes
Green algae
Gene families (ring B)
CslG/J/M
CslE
CslB/H


a
b
A
B
Species taxonomy (ring A)
Green algae
Charophytes
Briophytes
Lycophytes
Ferns
Gymnosperms
Basal angiosperms
Monocots
Monocots - Poaceae
Eudicots
Eudicots - Salicaceae
Eudicots - Fabaceae
Eudicots - Rosaceae
Eudicots - Cucurbitaceae
Eudicots - Brassicaceae
Eudicots - Amaranthaceae
Eudicots - Solanaceae
Gene families (ring B)
CslC
CslK
CslA

The following supplementary data can be accessed online at https://www.frontiersin.org/articles/10.3389/fpls.2022.870818/full#supplementary-material:

Supplementary Table 1: The 242 genomes used in this study.

Supplementary Table 2: List of the 445 *CesA* superfamily members retrieved from literature and used as BLAST queries to search for *CesA* superfamily genes in the 242 proteomes of the study.

Supplementary Table 3: The 49 experimentally validated primary and secondary *CesA* genes retrieved from literature.

Supplementary Table 4: The *CesA*/*Csl* synteny network.

Supplementary Table 5: List of the 7997 *CesA* superfamily genes found by BLAST and HMMER search across the 222 genomes with a BUSCO representation >75% (Viridiplantae set).

Supplementary Table 6: The annotation of gene tandem arrays for the *CesA*/*Csl* genes included in syntenic communities and from genomes with BUSCO representation >75% (Viridiplantae set) and at least 5 genes per scaffold on average.

Supplementary Table 7: The distribution of gene tandem arrays across species and gene families.

Supplementary Table 8: Summary of the distribution of gene tandem arrays across gene functions and main species clades.

Supplementary Table 9: List of the 6262 *CesA* genes included in the *CesA*/*Csl* synteny network

Supplementary Table 10: The 29 *CesA* genes of *G. hirsutum* detected in this study.

Supplementary Table 11: The 8 *CesA* genes from *G. hirsutum*, *G. arboreum*, and *G. raimondii* involved in the syntenic connections between clades 2 and 6 of Figure 3B and 3C, and the percentage of identity between them.

Chapter 3

Highly differentiated genomic properties underpin the different cell walls of Poaceae and eudicots

Francesco Pancaldi[1], M. Eric Schranz[2],
Eibertus N. van Loo[1], Luisa M. Trindade[1]

[1]Plant Breeding, Wageningen University & Research, Wageningen,
The Netherlands

[2]Biosystematics, Wageningen University & Research, Wageningen,
The Netherlands

Accepted for publication in *Plant Physiology*

Abstract

Plant cell walls of Poaceae and eudicots differ substantially, both in the content and composition of their components. However, the genomic and genetic basis underlying these differences is not fully resolved. In this research, we analyzed multiple genomic properties of 150 cell wall gene families across 169 angiosperm genomes. The properties analyzed include gene presence/absence, copy number, synteny, occurrence of tandem gene clusters and phylogenetic gene diversity. Results revealed a profound genomic differentiation of cell wall genes between Poaceae and eudicots, often correlated with cell wall diversity between these plant groups. For example, overall patterns of cell wall genes copy number variation and synteny are clearly divergent between Poaceae and eudicot species. Moreover, differential Poaceae-eudicot copy number and genomic contexts were observed for all the genes within the *BEL1-like HOMEODOMAIN 6* regulatory pathway, which respectively induces and represses secondary cell wall synthesis in Poaceae and eudicots. Similarly, divergent synteny, copy number, and phylogenetic gene diversification was observed for the major biosynthetic genes of xyloglucans, mannans, and xylans, potentially contributing to the hemicellulose differences of Poaceae and eudicot cell walls. Additionally, the Poaceae-specific tandem clusters and/or higher copy number of *PHENILALANINE AMMONYA-LYASE*, *CYNNAMOIL-CoA REDUCTASE* and *PEROXIDASE* genes may underly the higher cell wall lignin content of this plant clade. All these patterns are detailly discussed in this study, along with their evolutionary and biological relevance for cell wall (genomic) diversification between Poaceae and eudicots.

1 Introduction

All plant cells are surrounded by a cell wall, which mechanically supports plant growth and mediates plant-environment interactions (Somerville et al., 2004). The general plant cell wall architecture is conserved across angiosperms. Typically, primary cell walls are formed during plant cell expansion, and are composed of cellulose, different hemicellulosic polysaccharides, pectins, and structural proteins (Somerville et al., 2004, Sarkar et al., 2009). When cell growth ceases, secondary cell walls can be synthesized, in which pectins are absent or present at very low levels, while lignin is a major component (Zhong et al., 2019).

Despite the general cell wall patterns, extensive cell wall compositional and structural variation exists between plant taxa (Vogel, 2008, Sarkar et al., 2009). A major differentiation is between type I cell walls, which are specific to eudicots and non-commelinid monocots, and type II cell walls, which are found in grasses (Poaceae) (Vogel, 2008). Type I cell walls contain xyloglucan (XyG) as the major hemicellulose, along with relatively large amounts of (gluco)mannans. Moreover, they display large quantities of pectins and structural proteins (Vogel, 2008, Penning et al., 2019), while the lignin amount in type I secondary cell walls is typically lower than in the type II ones (Vogel, 2008). Conversely, type II cell walls contain xylans as the major hemicellulose, while XyGs and (gluco)mannans are only found in trace amounts. Furthermore, type II cell walls contain large quantities of (1,3;1,4)-β-glucans (also termed mixed linkage glucans – MLGs), a hemicellulose polysaccharide mainly restricted to grasses (Fincher and Stone, 1986, Vogel, 2008). Regarding non-hemicellulosic polysaccharides, type II cell walls contain much lower amounts of pectins and structural proteins than type I cell walls (Vogel, 2008, Penning et al., 2019). Finally, the lignin content of type II secondary cell walls is usually higher than that of type I cell walls (Vogel, 2008), while relatively large amounts of other phenylpropanoids – mainly ferulic acid – are present in type II cell walls to cross-link polysaccharides and lignin (De Oliveira et al., 2015, de Souza et al., 2018).

The differences between type I and type II cell walls are not absolute, and cell wall compositional variation exists within both Poaceae and eudicots (Vogel, 2008, Burton and Fincher, 2014a). Nevertheless, type I and type II cell walls represent established valuable models to describe the marked cell wall differentiations observed between these plant groups (Carpita and Gibeaut, 1993, Vogel, 2008, Carpita and McCann, 2020). This aspect, along with the agricultural relevance of Poaceae species and the importance of cell wall composition for the industrial utilization of plant biomass, makes the understanding of the genetics underlying type I and type II cell wall differences a valuable research target (Burton and Fincher, 2012, Pancaldi and Trindade, 2020).

The understanding of the genetics underpinning type I and type II cell walls is however far from being resolved. For MLGs, it was possible to associate their mostly Poaceae-specific occurrence with the grass-specific presence of *Cellulose synthase-like F* (*CslF*) genes (Burton et al., 2006). Notwithstanding, the complexity of cell wall biosynthesis, which relies on a multitude of genes with pleiotropic effects, hampers the elucidation of the genetic basis of type I and type II cell wall differentiation (Burton and Fincher, 2012, Yokoyama, 2020). In this context, Penning et al. (2019) analysed the occurrence of different carbohydrate-active enzymes in the genomes of arabidopsis, maize, and rice, showing their ubiquitous presence in all the genomes, irrespectively of the species' cell wall type. This result indicates that, with few exceptions such as *CslF*, gene presence-absence variation is not sufficient to explain type I and type II cell walls differentiation (Penning et al., 2019). Other researchers studied the regulatory differences between Poaceae and eudicot cell walls, showing that the cell wall regulatory machinery is overall conserved across these clades, even if few but relevant differences were observed (see Rao and Dixon (2018) for a review on this topic). For example, the master transcription factor *BEL1-like HOMEODOMAIN 6* (*BLH6*) has opposite function in eudicots and Poaceae, by repressing and inducing secondary cell wall deposition, respectively (Hirano et al., 2013, Liu et al., 2014a, Rao and Dixon, 2018). This observation highlights the importance of comparative genetic research to understand the cell wall differentiation between Poaceae and eudicots. However, these types of studies are seldom, and the question of what is the genetic-evolutionary basis of type I and type II cell walls is currently largely unresolved (Yokoyama, 2020).

In this research, the study of the genetics underlying type I and type II cell walls was tackled from the perspective of the genomic properties of cell wall genes across Poaceae and eudicots. This was performed by analysing patterns of gene copy number, synteny, tandem gene clusters, and phylogenetic relatedness of 150 different cell wall gene families across 169 angiosperm genomes representing plant cell wall diversity. This approach is on purpose large-scale and genomic-oriented, since recent research showed that such methodologies are very powerful to investigate complex genetic patterns at the basis of plant diversity (Zhao et al., 2017, Kerstens et al., 2020). To conclude, the data produced were on purpose analysed in comparisons between grasses and eudicots, as they are the taxonomic groups representing type I – type II cell wall differences.

2 Materials and methods

2.1 Collection of plant genomes

All the angiosperm genomes sequenced and published by the end of 2018 and available with at least a scaffold-level assembly were searched for in several online databases. For each genome, a BED file indicating gene positions and FASTA files reporting protein and nucleotide sequences from all the annotated protein-coding genes were retrieved. Genomes were checked for assembly completeness by using the BUSCO Viridiplantae gene set (Seppey et al., 2019) and for assembly fragmentation by assessing the number of scaffolds and the N50 statistics. Genomes with <75% BUSCO genes were excluded from further analyses. Through these criteria, a total of 169 genomes were collected (**Supplementary Table 2**). This set includes the genomes of two basal angiosperm species, 11 non-commelinid monocots, 24 Poaceae, and 132 eudicots from 39 different eudicot families.

2.2 Identification of cell wall genes in all the genomes used

The identification of cell wall genes within the 169 collected genomes was performed by following the approach published by Pancaldi et al. (2022b). In brief, the detailed gene annotation and the extensive cell wall research available for arabidopsis in scientific literature and online databases was used to create an initial list of 1312 genes proven to be involved in cell wall biosynthesis within this species (**Supplementary Table 6**). This list was further integrated with some Poaceae-specific cell wall genes for which functional information was available for maize and/or rice. The collected genes classified into 150 different cell wall-related functions, and were annotated for functional domain composition using HMMER3 (default parameters) (Mistry et al., 2013) and the Hidden Markov Models of all the protein domains available at the PFAM database (El-Gebali et al., 2019). Subsequently, the collected genes were aligned against the PEP files of all the 169 plant genomes of the study using BLAST (Evalue = 1E−3) (Altschul et al., 1990). This search led to the identification of all the potential homologs of the initial cell wall genes across all the collected genomes. The identified genes were also annotated for PFAM composition, and the BLAST outputs were then further filtered based on equal domain composition between BLAST queries and subjects. Finally, very large gene families for which it is known that not all the genes are involved in cell wall biosynthesis (e.g. *BAHD*) were further filtered by building phylogenetic trees with RAxML v8.2.9 (Stamatakis, 2014) and identifying clades containing genes from arabidopsis, maize, or rice for which cell wall functional validation is available in scientific literature. For gene filtering, RAxML was run with 100 bootstraps and by using the PROTCATBLOSUM62 substitution matrix. At the end, the search for cell wall gene homologs yielded a list of 320,005 genes across the 169 genomes and the 150 cell wall functions of the study (**Supplementary Table 1**).

2.3 Analysis of gene copy number variation

The number of gene copies present in each of the 169 genomes of the study was quantified for each of the 150 cell wall gene functions (custom R script). The average number of genes belonging to each cell wall function was determined for each genome, and t-tests were computed to assess significant differences in average copy number between Poaceae and eudicots, as well as between multiple other angiosperm families at a time. In addition, heatmaps were created to visualize patterns of copy number variation across both cell wall gene families and plant species. Principal Component Analysis (PCA) was also performed to assess the contribution of cell wall gene copy number variation to the differentiation of Poaceae, non-commelinid monocots, and eudicots (custom R script). Finally, quantitative data from all these analyses were crossed with literature information on gene function to identify classes of genes whose copy number patterns appear particularly relevant in the context of the differentiation between type I and type II cell walls.

2.4 Synteny analysis

The syntenic conservation of the 320,005 cell wall genes of the study across the 169 angiosperm genomes was analysed by following the methodology developed by Zhao and Schranz (2017) for large-scale network synteny analysis. Specifically, Diamond (Buchfink et al., 2015) was used to align all the proteins of each genome against all the other proteins of that genome and all the proteins of every other genome (default parameters; Evalue = 1E–3). The outputs of Diamond were processed with MCScanX (Wang et al., 2012b) to detect synteny (i.e. conserved gene order across multiple genomes) by evaluating the relative genomic position of pairs of homologous genes from each genome comparison. MCScanX was run with default parameters, except -s (number of colinear genes to claim a syntenic block) set to 3. The outputs of MCScanX were organized in a synteny network, in which each node is a gene and edges represent syntenic connections between genes. The synteny network was then filtered to retain only pairs of nodes linking cell wall genes (**Supplementary Table 4**), and decomposed into syntenic communities (i.e. groups of genes that display significantly higher synteny with each other than with the rest of genes in the network) of at least four nodes (k=4), by using the Infomap algorithm (Rosvall and Bergstrom, 2007, Rosvall et al., 2009). Syntenic communities of cell wall genes were taxonomically profiled, and the copy-number of syntenic genes across the species contained within each community was also assessed. Finally, the taxonomic and copy number data from different syntenic communities were used to evaluate the occurrence of a divergent genomic organization for each cell wall gene function between Poaceae, non-commelinid monocots, and eudicots.

2.5 Analysis of tandem gene clusters

The ordinal position of cell wall genes along genomic BED files was assessed for each genome of the study, to identify the occurrence of tandem gene clusters of homologous genes along chromosomes (custom R script). The results of this analysis were used to evaluate the proportion of clustered and singleton genes out of the total genes belonging to a certain cell wall function, and the mean size of the tandem clusters found in every species. Moreover, the occurrence of clustered genes within syntenic communities detected in the synteny analysis was also assessed.

2.6 Phylogenetic analyses

A group of gene families that displayed genomic patterns particularly relevant in the context of the differences between type I and type II cell walls were selected for a more detailed genetic study encompassing phylogenetic analysis (see Section 3). For this purpose, the protein sequences of the genes belonging to these gene families were aligned with MAFFT v7.453 (FFT-NS-2 algorithm) (Katoh and Standley, 2013), with default parameters except gap opening penalty, set to 1.0. MAFFT alignments were trimmed using TrimAl v1.2 (Capella-Gutiérrez et al., 2009), with default parameters. Finally, RAxML v8.2.9 (Stamatakis, 2014) was used to build phylogenetic trees out of trimmed alignments (PROTCATBLOSUM62 substitution matrix; 100 bootstraps). Phylogenetic trees were plotted and annotated using iTOL (Letunic and Bork, 2019). The TAIR database (Huala et al., 2001) was used to localize critical arabidopsis genes of the families analysed within each tree, while BLAST (Altschul et al., 1990) was used to find and localize relevant grass homologs of those trees within each tree (Evalue = 1E-3).

2.7 Analysis of selection pressure

Differences in the rates of selection pressure between the Poaceae and eudicot genes contained in both differentiated and shared genomic contexts of the genes included in the *BLH6* pathway were evaluated by using the EasyCodeML implementation (Gao et al., 2019) of the CodeML program from the PAML4.0 package (Yang, 2007). For each phylogenetic tree corresponding to a gene family within the *BLH6* pathway (see **Figure 4**), Poaceae clade(s) (corresponding to distinct syntenic communities from eudicots, or contained in shared syntenic communities with eudicots, depending on the gene family) have been set as foreground in a branch model (model = 2, NSsites = 0) to estimate dN:dS ratios specifically for those branches. A basic model was also run (model = 0, NSsites = 0) to estimate dN:dS ratios at the whole-tree level (background), and a likelihood ratio test (LRT, also implemented in EasyCodeML) was performed to test for significantly different selection pressure between foreground and background tree branches. For computational reasons, a random set of 80 leaves was selected from the set foreground and background branches in order to run CodeML models and tests

3 Results

3.1 Extensive copy number variation within the cell wall gene families of Poaceae and eudicots cell walls

Gene copy number was quantified across the 150 target cell wall gene families in each of the 169 angiosperm genomes (**Supplementary Tables 1 and 2** list the genes and genomes used). The data obtained were plotted onto a heatmap (**Figure 1A**) and analysed by principal component analysis (PCA; **Figures 1B and 1C**). Results revealed extensive copy number variation (CNV), across both gene families and plant species. Specifically, CNV between gene families ranged from singleton families in the majority of the species surveyed (e.g. the master transcription factor *E2Fc*, *KATANIN* genes involved in xylem development – *KTN*, or the homologs of arabidopsis *ALTERED XYLOGLUCAN 9* – *AXY9*) to families containing 100+ gene copies per genome on average (e.g. *PEROXIDASES* – *PRX*, *POLYGALACTURONASES* – *PG*, and *PECTIN ESTERASES* – *PE*) (**Figure 1A** and **Supplementary Table 3**). Moreover, CNV across plants highlighted species displaying deviations of gene copy number for several cell wall gene families. On the one hand, this was observed for specific plant clades known to share taxon-specific genome duplications, as the Salicaceae and Cucurbitaceae families, or the *Gossypium* and *Brassica* genera (dashed boxes of **Figure 1A**). On the other hand, copy number differences were also revealed between Poaceae and several eudicot species for multiple cell wall gene families (solid black boxes of **Figure 1A** and **Supplementary Table 4**). Remarkably, some of these gene families are particularly relevant for the differentiation of type I and type II cell walls. For example, they include the *IRREGULAR XYLEM* (*IRX*) *9* and *14* genes involved in xylan synthesis (CAZy GT43 family); the *PECTIN METHYL-ESTERASES* and associated inhibitors (*PME* and *PMEI*), which affect the pectin content of cell walls; or the *PHENYLALANINE AMMONIA-LYASE* genes (*PAL*), which catalyse the first step of the lignin pathway.

To better elucidate the relationship between CNV and angiosperm cell wall diversity, CNV data were also analysed by PCA (**Figure 1B**). The first two components capture 46% of the total CNV across all genes and species, and clearly separate Poaceae from the other angiosperms. This highlights the relevance of cell wall gene CNV as a major genomic property underlying Poaceae and eudicot cell wall diversity. Moreover, PCA

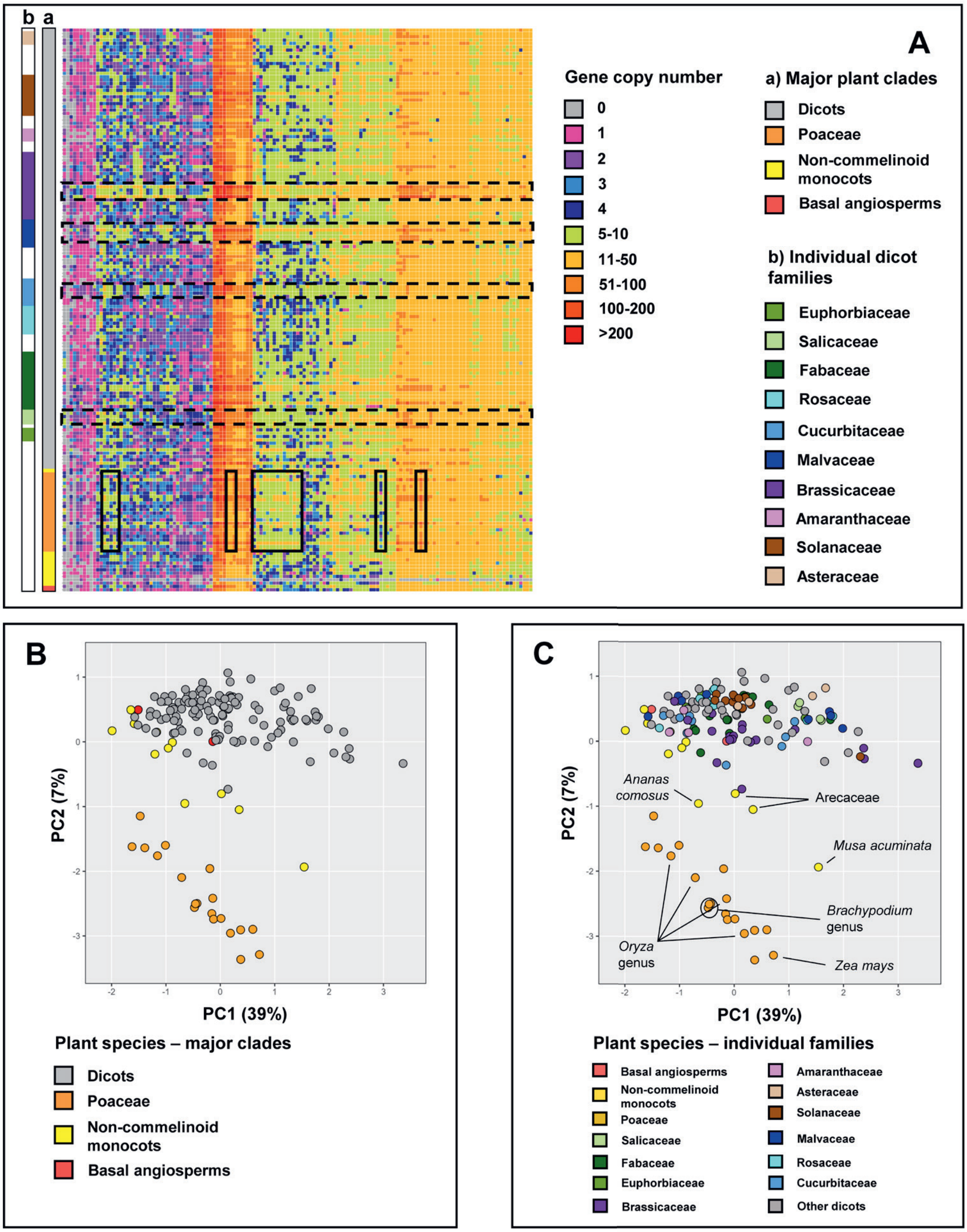


Figure 1. Copy number properties of cell wall genes. A) Heatmap showing the large CNV of cell wall gene families (columns) across the 169 genomes of the study (rows). Colors of heatmap cells represent the copy number of each species-gene combination (see legend). Left to the heatmap, genomes are categorized based on taxonomic clades. B) and C) PCA plots of the 169 genomes of the study based on CNV patterns. The two plots display PCA results at the level of general plant clades (B) and of individual plant families (C), respectively.

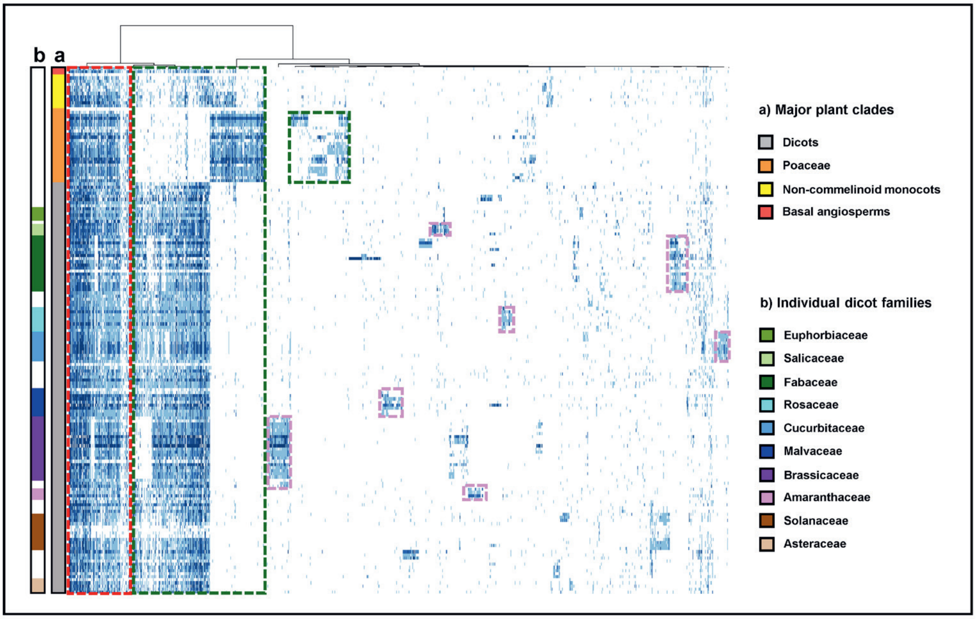


Figure 2. Heatmap displaying the taxonomic profiling of the 7634 syntenic communities detected across the 150 cell wall gene families and 169 plant genomes of this study. Heatmap cells are colored based on gene copy number of each community (columns) and species (rows) combination. Colors range from white (copy number = 0) to dark blue (max copy number). Communities are clustered based on patterns of taxonomic profiling. Dashed rectangles group communities representing genomic contexts that (i) are conserved across all (or most) angiosperms (red rectangle), (ii) are differentially conserved between Poaceae and (most of) eudicots (green rectangles), or (iii) display lineage-specific patterns of conservation (pink rectangles).

highlighted a generally large level of CNV also within Poaceae and eudicots themselves, as these species are extensively spread within the plots of **Figures 1B and 1C**. This is in line with both the relatively large Poaceae cell wall diversity (within the framework of the type II cell wall) (Vogel, 2008, Burton and Fincher, 2014a), and with the numerous (segmental) duplications and genomic translocations experienced by the different Poaceae species during grass evolution (Wang et al., 2015, Lee et al., 2020). Besides Poaceae, the majority of eudicots and of non-commelinid monocots form a single large cluster in the plots of **Figures 1B and 1C**. The different plant families within this cluster do not display further clear grouping patterns, suggesting that intra-family CNV in eudicots can be large and of similar magnitude as for eudicots as a whole group.

Given the clear correlation between cell wall gene CNV and type I – II cell walls differentiation, statistical tests were performed to assess which gene families display significant copy number differences between Poaceae and eudicots. The results showed that 70 of the 150 cell wall gene families analysed (47%) display significantly different copy number levels between these groups of plants (t-test, alpha=0.05; **Supplementary Table 4**). These 70 families were analysed for the magnitude of significant differences, as well as for their relevance for cell wall differentiation between Poaceae and eudicots based on scientific literature. This way, 20 particularly relevant families were identified (**Table 1**). These genes mediate critical steps in the biosynthesis of cell wall components that are variable between Poaceae and eudicot cell walls, and display CNV patterns in line with such differences, mainly in a perspective of gene dosage variability. The detailed explanations of the patterns found are reported in **Table 1**.

Table 1 – Gene families displaying significant copy number differences between Poaceae and eudicots that appear particularly relevant for the differentiation of type I and type II cell walls, in view of their known cell wall function and the direction of the copy number patterns observed.

Gene family	Eudicot mean copy number	Poaceae mean copy number	Functional relevance of copy number pattern for type I and type II cell walls	References
Mannanases (*MAN*)	4	1	Mannanases are associated with mannans synthesis and remodelling. Mannans are common in eudicots but seldom in Poaceae.	(Zhong et al., 2019, Vogel, 2008)
Irregular Xylem 9 (*IRX9/9L*; GT43)	6	13	Central gene for xylan synthesis. Complexes with *IRX10/10L* and *IRX14/14L*. Xylans are much more abundant in Poaceae than eudicots.	(Zhong et al., 2019, Vogel, 2008)
Irregular Xylem 10 (*IRX10/10L*; GT47)	33	26	Central gene for xylan synthesis. Complexes with *IRX10/10L* and *IRX14/14L*. Xylans are much more abundant in Poaceae than eudicots.	(Zhong et al., 2019, Vogel, 2008)
Irregular Xylem 14 (*IRX14/14L*; GT43)	5	7	Central gene for xylan synthesis. Complexes with *IRX10/10L* and *IRX14/14L*. Xylans are much more abundant in Poaceae than eudicots.	(Zhong et al., 2019, Vogel, 2008)
Cellulose synthase-like G (*CslG*)	3	1	At least one *CslG* transfers Glucuronic Acid (GlcA) during the synthesis of saponins, and *CslGs* could therefore act as GlcA transferases, also in the context of cell wall. GlcA substitutions of xylans highly differ between eudicots (highly substituted xylans) and grass (poorly substituted xyland).	(Jozwiak et al., 2020, Pena et al., 2016)
PARVUS	26	16	Involved in GlcA substitution of xylans. GlcA-enriched xylans are more abundant in eudicots.	(Pena et al., 2016, Vogel, 2008)
BEAT/AHCT/HCBT/DAT acyl-transferase (*BAHD*)	3	11	*BAHD* are involved in ferulic acid and p-coumaric acid substitution of several cell wall components. These molecules are much more common in grass cell walls.	(Vogel, 2008, Bartley et al., 2013, Molinari et al., 2013)

Glycosyl-transferase 61 (GT61)	5	28	Genes involved in the feruloylation of xylans in grasses	(Cenci et al., 2018)
Xylogalacturonan-deficient (*XGD*)	22	9	Central gene for the synthesis of backbone xylogalacturonan backbones. Pectins are much higher in eudicots primary cell walls compared to grass ones.	(Atmodjo et al., 2013, Vogel, 2008)
GAUT-like proteins (*GATL*)	11	5	Gene involved in pectin synthesis, even if with unclear role. Pectins are much higher in eudicots primary cell walls compared to grass ones.	(Atmodjo et al., 2013, Vogel, 2008)
Phenylalanine Ammonia-lyase (*PAL*)	5	10	First gene of the lignin pathway, affecting the total amount of substrates streamed to lignin synthesis. Lignin content is higher in Poaceae cell walls.	(Vogel, 2008, Zhong et al., 2019)
Peroxidase (*PRX*)	95	152	Mediate *in muro* lignin deposition. Lignin content is higher in Poaceae cell walls	(Vogel, 2008, Zhong et al., 2019)
Cinnamoyl CoA Reductase (*CCR*)	14	21	Central lignin gene shared by the branches of the lignin pathway leading to all monolignols. It can deeply influence final lignin amount of cell walls. Lignin content is higher in Poaceae cell walls.	(Tamasloukht et al., 2011, Zhong et al., 2019)
Caffeic acid O-methyltransferase (*COMT*)	10	5	*COMT* genes push the lignin pathway towards synthesis of S- and G- lignin subunits instead of H-lignin. Poaceae have a higher amount of H- subunits.	(Vogel, 2008, Zhong et al., 2019)
Pectin Lyase (*PLY*)	25	9	Central gene for pectin metabolism. Pectins are much higher in eudicots primary cell walls compared to grass ones.	(Atmodjo et al., 2013, Vogel, 2008)
Pectin Methylesterase (*PME*)	75	41	Central gene for pectin metabolism. Pectins are much higher in eudicots primary cell walls compared to grass ones.	(Atmodjo et al., 2013, Vogel, 2008)
Pectin Methylesterase-inhibitor (*PME*)	73	38	Central gene for pectin metabolism. Pectins are much higher in eudicots primary cell walls compared to grass ones.	(Atmodjo et al., 2013, Vogel, 2008)
Dynamin-related protein (*DRP/ADL*)	16	12	Genes likely involved in pectin trafficking to cell wall. Pectins are much higher in eudicots primary cell walls compared to grass ones.	(Atmodjo et al., 2013, Vogel, 2008)
BEL-like Homeodeomain 6 (*BLH6*)	2	4	Transcription factor displaying divergent function in grasses (inducer of secondary cell wall synthesis) and eudicots (repressor of secondary cell wall synthesis).	(Rao and Dixon, 2018)
Expansins (*EXP*)	39	70	Involved in cell wall expansion by remodelling hemicellulose polysaccharides and pectins, which are highly variable between grasses and eudicots.	(Atmodjo et al., 2013, Vogel, 2008)

3.2 Cell wall gene synteny reveals conserved and divergent genomic gene contexts between Poaceae and eudicots

Recent studies highlighted how differential gene synteny across plants might underlie traits variability and evolutionary adaptations, through changes of genomic gene contexts that can impact gene functions (Dewey, 2011, Zhao et al., 2017, Kerstens et al., 2020, Pancaldi et al., 2022a). To test if this is also the case for cell wall genes and type I and type II cell walls, the syntenic conservation of the 150 cell wall gene families across the 169 genomes of the study was examined. Gene synteny was analysed using the network approach developed by Zhao and Schranz (2017), which organizes large sets of syntenic genes from diverse genomes into networks where nodes represent genes and edges intergenic synteny. This way, synteny patterns can be dissected with statistical methods for networks analysis, including network decomposition into communities of nodes displaying significantly higher synteny within than between communities. Such syntenic communities represent independent gene positional configurations – or genomic contexts – occurring in specific groups of species.

The synteny analysis of the 320,005 cell wall genes (from the 150 gene families and 169 genomes) yielded a synteny network with 258,316 different nodes; 80.7% of the initial cell wall genes (**Supplementary Dataset 1**). The large number of cell wall genes retained in the network as nodes indicates a very high level of syntenic conservation of cell wall genes. Specifically, such percentage is higher than what detected in other studies of unrelated gene families (Zhao et al., 2017, Kerstens et al., 2020), but is in line with the level of synteny observed for the *Cellulose synthase* gene superfamily, a smaller distinct set of cell wall genes, in previous studies (Schwerdt et al., 2015, Pancaldi et al., 2022a). Moreover, each cell wall gene in the synteny network is on average syntenic with >50 other genes, irrespectively of it being a Poaceae or eudicot gene. This is remarkable, as while extensive gene synteny is commonly found in Poaceae (Gale and Devos, 1998), the same is rather rare in eudicots (Zhao and Schranz, 2019).

Synteny network decomposition yielded 7634 different syntenic communities of at least four nodes, each representing a specific genomic context of a specific cell wall gene type, conserved in a specific group of species (**Figure 2**). These syntenic communities were taxonomically and functionally profiled, revealing three main community groups. The first group contains 597 communities consisting of 87,905 total cell wall genes whose positional genomic organization is conserved across all or most of the angiosperms analysed, including Poaceae, eudicots, non-commelinid monocots, and basal angiosperms such as *Amborella trichopoda* (red box of **Figure 2**). Interestingly, functional profiling showed that these widely conserved syntenic communities contain a large proportion of transcription factors and lignin-related genes (66% and 52% of all the syntenic transcription factors and lignin genes, respectively). Conversely, the fraction of cellulose and hemicellulose genes within these communities is much smaller (14% and 33%, respectively). The second group includes 1,678 communities containing 98,538 total genes that display completely divergent syntenic conservation between eudicots, Poaceae, and non-commelinid monocots (green boxes of **Figure 2**). Of these communities, 718 are largely conserved across eudicots and non-commelinid monocots, but not in Poaceae; 492 communities are conserved within Poaceae but not in eudicots and non-commelinid monocots; 468 communities are conserved within Poaceae and most non-commelinid monocots, but not in eudicots. Concerning functional profiling, hemicellulose-related gene functions are the ones with the highest proportion of genes represented within the communities of group 2 of **Figure 2** (46% of all the hemicellulose-related syntenic genes), followed by cellulose-related genes (41%), lignin-related genes (35%), and transcription factors (29%). Finally, the third group of syntenic communities contains the remaining 5,359 communities, which include 34,138 total genes and display mostly lineage- or species-specific patterns of gene synteny (pink boxes of **Figure 2**). Among these are 144 communities specific to Brassicaceae, 87 communities specific to Fabaceae, and other minor groups of communities specific to Salicaceae, Rosaceae, and Cucurbitaceae.

Community data were also crossed-referenced with gene copy number results, showing that large CNV exists both between and within syntenic communities for several gene families. Specifically, CNV is often taxon- and/or community-specific. As an example, **Figure 3A** and **Figure 3B** show that two genes at the basis of sucrose supply for cellulose synthesis – *SUCROSE PHOSPHATE SYNTHASE* (*SPS*) and *SUCROSE SYNTHASE* (*SUS*) – display a Poaceae-specific increase in gene copy number (>3 gene copies per species on average) relative to eudicots (mostly singleton or double gene copies). Alternatively, **Figure 3C** displays that *PECTIN LYASE* (*PLY*) genes show both copy number and presence-absence variation between Poaceae and eudicots between and within several genomic contexts. Interestingly, the different genes reported in **Figure 3** are all known to affect the content of cellulose and pectins (Zhong et al., 2019), which both tend to vary considerably between type I and type II cell walls (Vogel, 2008).

To conclude, like CNV data, community data of each gene family were also analysed together with available information on gene function in the context of type I and type II cell walls from scientific literature. The next paragraphs report the results of this analysis for different gene families that are critical for secondary cell wall deposition and for hemicellulose biosynthesis. Moreover, **Table 2** reports a list of gene families that display either conservation or differentiation of genomic contexts across Poaceae and eudicots and appear relevant for type I – II cell walls differentiation.

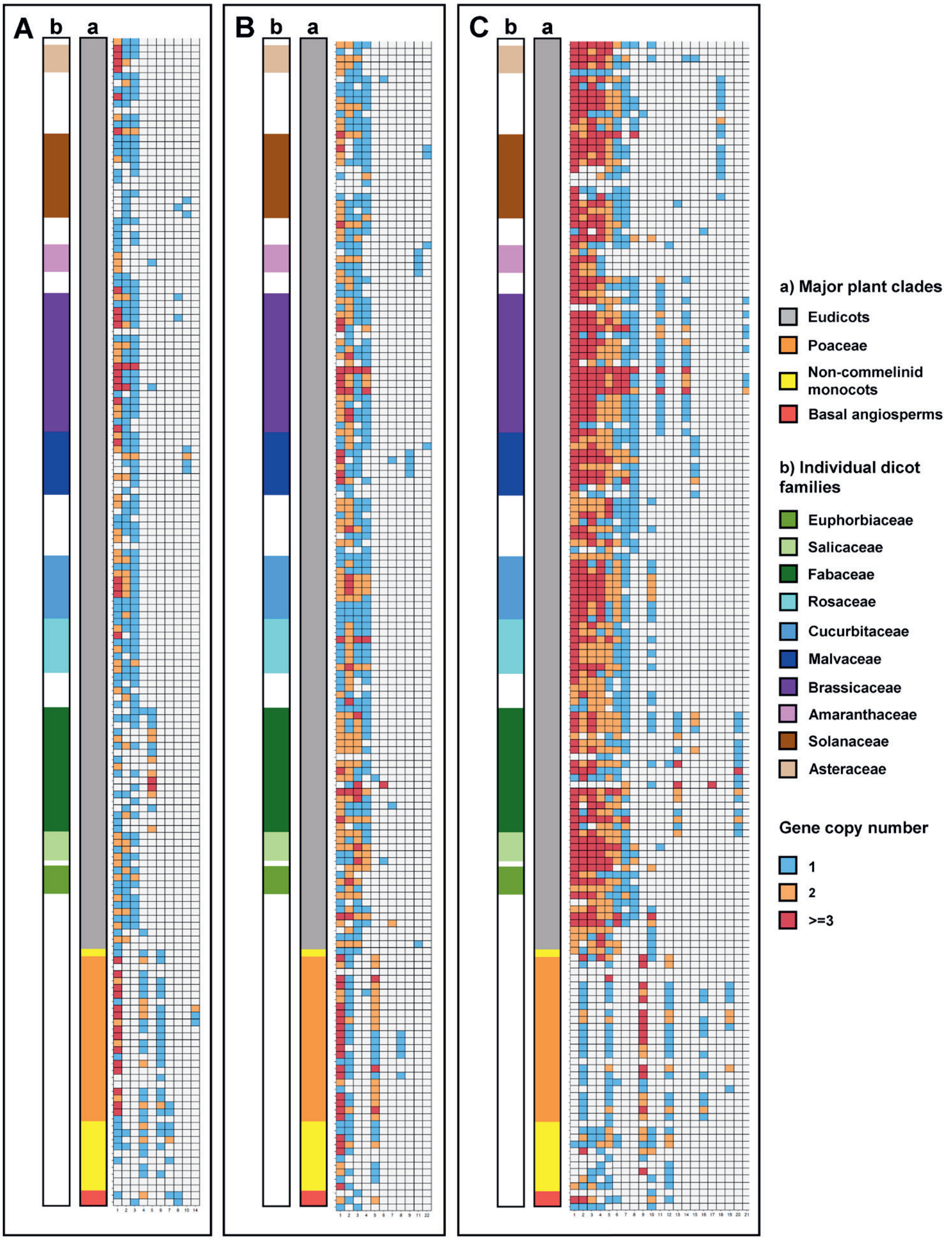


Figure 3. Examples of genomic context-specific patterns of gene CNV across three cell wall gene families. In all the heatmaps, columns represent syntenic communities detected for a specific gene family, while rows represent species, ordered taxonomically (see legend). Heatmap cells are coloured according to the gene copy number for every genomic context – species combination (see legend). A) *SUCROSE PHOSPHATE SYNTHASE* (*SPS*). B) *SUCROSE SYNTHASE* (*SUS*). C) *PECTIN LYASE* (*PLY*).

Table 2 – Gene families displaying either Poaceae/eudicot-specific or angiosperm-wide synteny that appear particularly relevant for the differentiation of type I and type II cell walls in view of their known cell wall function. Several structural genes involved in critical steps of the synthesis of polysaccharides whose content differs substantially between Poaceae and eudicots have their genes mostly contained in Poaceae- or eudicot-specific syntenic communities. Conversely, several transcription factors with conserved function across angiosperms are retained in syntenic communities displaying wide conservation across angiosperms.

Cell wall gene function	**Genes in Poaceae- and/or eudicot-specific communities (%)**	**Genes in angiosperm-wide communities (%)**	**Genes in other lineage-specific communities (%)**	**Functional relevance in the context of type I and type II cell walls**	**References**
F5H	93.9	0.0	6.1	Central gene in lignin synthesis, regulating total lignin production and monolignol ratio	(Zhong et al., 2019, Vogel, 2008)
CslA	75.1	0.3	24.7	Binds mannan residues during (gluco)mannan synthesis	(Zhong et al., 2019, Vogel, 2008)
CslC	61.5	29.8	8.7	Binds glucose residues of XyG molecules	(Zhong et al., 2019, Vogel, 2008)
AXY9	89.8	0.0	10.2	Mediates XyG O-acetylation	(Zhong et al., 2019, Vogel, 2008)
XXT	64.6	31.0	4.4	Binds xylose residues over XyG backbones	(Zhong et al., 2019, Vogel, 2008)
IRX9/9L (GT43)	73.0	15.0	12.1	Forms a biosynthetic complex with *IRX14/14L* and *IRX15/15L* during xylan synthesis	(Zhong et al., 2019, Vogel, 2008)
IRX14/14L (GT43)	64.0	23.0	13.0	Forms a biosynthetic complex with *IRX9/9L* and *IRX15/15L* during xylan synthesis	(Zhong et al., 2019, Vogel, 2008)
IRX15/15L	62.1	0.0	37.9	Forms a biosynthetic complex with *IRX9/9L* and *IRX14/14L* during xylan synthesis	(Zhong et al., 2019, Vogel, 2008)
XTH	60.7	0.0	39.3	Mediates the extension of XyG molecules	(Zhong et al., 2019, Vogel, 2008)
RGXT	61.4	0.2	38.4	Central gene for the synthesis of RGII	(Atmodjo et al., 2013, Vogel, 2008)

GALS	63.5	0.0	36.5	Involved in the synthesis of RGI side chains	(Atmodjo et al., 2013, Vogel, 2008)
XGD	61.8	0.5	37.8	Transfers xylose during Xylogalacturonan synthesis	(Atmodjo et al., 2013, Vogel, 2008)
BLH6	97.6	2.3	0.0	Inducer/repressor of secondary cell wall development	(Rao and Dixon, 2018)
E2FC	0.0	72.7	27.3	Key upstream regulator of cell wall biosynthesis.	(Taylor-Teeples et al., 2015, Rao and Dixon, 2018)
VND	19.2	69.9	10.9	First layer cell wall transcription factors regulating ectopic secondary cell wall deposition in vessels, with conserved function in angiosperms.	(Taylor-Teeples et al., 2015, Rao and Dixon, 2018)
SND	3.2	87.7	9.0	First layer cell wall transcription factors for secondary cell wall biosynthesis, with conserved function in angiosperms.	(Taylor-Teeples et al., 2015, Rao and Dixon, 2018)
NST1/2/3	0.0	98.4	1.6	First layer cell wall transcription factors for secondary cell wall biosynthesis, with conserved function in angiosperms.	(Taylor-Teeples et al., 2015, Rao and Dixon, 2018)
C2H2	0.0	85.1	14.9	Repressor of secondary cell wall development.	(Taylor-Teeples et al., 2015, Rao and Dixon, 2018)

Conserved and divergent genomic contexts for the pathways controlled by the functionally different BLH6/BLH9 transcription factors

BLH6 and *BLH9* are phylogenetically-close cell wall transcription factors (**Figure 4A**) that are present in all angiosperms but display diverse functionalization patterns, with *BLH9* being a repressor of lignification across all plants, and *BLH6* acting as repressor and inducer of secondary cell wall in eudicots and grasses, respectively (Rao and Dixon, 2018). This makes these genes a particularly interesting case to study the genetic factors underlying Poaceae-eudicot cell wall differences. Phylogenomic analysis revealed three distinct *BLH6* genomic contexts and two separate *BLH9* genomic contexts (**Figure 4A**). These genomic contexts correspond to distinct phylogenetic gene clades supported by high bootstrap (95-100) (**Figure 4A**). Interestingly, while both *BLH9* genomic contexts are conserved across all the angiosperms, of the three *BLH6* contexts one is specific to eudicots and the other two are restricted to grasses (**Figure 4A**). Moreover, the overall *BLH6* copy number is also different between grasses and eudicots, with Poaceae having 3.6 *BLH6* copies per species on average, compared to 2.2 of eudicots (t-test't P=0.000). This difference is not observed for *BLH9* (2.3 copies per grass species vs 1.9 in eudicots; t-test's P=0.119). Remarkably, the "surplus" *BLH6* grass copies are not equally spread across the two Poaceae-specific *BLH6* genomic contexts. In fact, the blue community of **Figure 4A** contains two *BLH6* copies per grass species on average, compared to only one copy for the green community of **Figure 4A**. Overall, these results show a clear correlation between the (phylo)genomic organization of *BLH6* and *BLH9* and their functional diversification in the context of type I and type II cell walls.

To further investigate the association between the genomic organization and functional specialization of *BLH6*/*BLH9* genes, phylogenomic analyses were extended to the other genes within the *BLH6* regulatory pathway. These include two downstream transcription factors – *OFP4* and *KNAT7* – and a major lignin structural gene – *F5H* (Rao and Dixon, 2018, Qin et al., 2020). Divergent phylogenomic patterns between Poaceae and eudicots were revealed for all these genes. Specifically, *OFP4* genes turned out to be organized into two main syntenic communities (**Figure 4B**). The largest one groups nearly all the eudicot *OFP4* copies, plus the majority of non-commelinid monocots *OFP4* and only two Poaceae copies. Conversely, the smaller community includes nearly all the Poaceae *OFP4*, and corresponds to an independent phylogenetic clade (bootstrap 99). Regarding *KNAT7*, its phylogenomic analysis also revealed the presence of one Poaceae-specific syntenic community and one eudicot-specific syntenic community (**Figure 4C**). Moreover, *KNAT7* genes were phylogenomically analysed together with the members of the *KNAT3* family (**Figure 4C**), which groups very close homologs involved in lignin deposition and regulated by the master cell wall transcription factors *NST1* and *NST2* (Qin et al., 2020). As for *BLH9*, the function of *NST* transcription factors and of *KNAT3* genes is largely conserved across grasses and eudicots (Rao and Dixon, 2018), and phylogenomic analyses showed that both *KNAT3* and *NST* genes display widespread syntenic conservation across both grasses and eudicots (**Figures 4C and 4D**). Finally, phylogenomic analysis of the last gene within the *BLH6* regulatory pathway – *F5H* – remarkably displayed differential syntenic and phylogenetic organization between Poaceae and eudicots (**Figure 4E**).

As a final step in the analysis of the genes belonging to the *BLH6* regulatory pathway, it was tested whether the gene clades corresponding to differential genomic contexts between Poaceae and eudicots displayed differential selection pressures between grass and eudicot clades. Selection pressure is defined by the ratio of nonsynonymous (dN) over synonymous (dS) substitutions between different gene nucleotide sequences (Schwerdt et al., 2015). Thus, differential dN:dS ratios between genes belonging to different taxonomic and/or phylogenomic clades highlight differential evolutionary rates associated with different genetic/genomic configurations. The analysis of selection pressure with the use of the branch model from the CodeML program (Yang, 2007) revealed that Poaceae genes organized in independent genomic contexts are systematically under significantly different (and positive – dN:dS>1) selection pressure as compared to their eudicot counterparts (LRT's P<0.01; **Supplementary Table 7**). This holds true for all the genes displaying differential syntenic conservation in Poaceae and eudicots, except for *KNAT7*. Vice versa, the Poaceae genes organized in syntenic communities shared with eudicot species (as for example for the two *BLH9* communities of the *KNAT3* genes) did not display significant differences in selection pressure as compared to eudicot genes within the same communities in all the tested performed (LRT's α = 0.01; **Supplementary Table 7**).

To conclude, all the data displayed in this paragraph highlight a striking correlation between the conservation/diversification of the positional organization and the conservation/diversification of gene function for all the genes within the *BLH6* pathway across Poaceae and eudicots, which is ultimately associated with similar or divergent effects of these genes on plant cell walls, respectively. Remarkably, this diversification is associated with a phylogenetic diversification of genes and extensive patterns of differential – and often positive – selection pressures, highlighting that nucleotide diversification of genes is likely favoured by and intimately associated to the observed large-scale genomic gene rearrangements.

Divergent genomic contexts between Poaceae and eudicots for multiple important hemicellulose-related genes

As the content of several hemicellulosic molecules differs substantially between type I and type II cell walls, the genes synthesizing the backbone of the most differing hemicellulosic polysaccharides between Poaceae and eudicots are also relevant targets for phylogenomic analyses. One of such polysaccharides is XyG, whose synthesis depends on *Cellulose synthase-like C* (*CslC*) genes – which bond the glucose residues of XyG backbones – and on *Alfa-1,6-Xylosyltransferases* (*XXT*) – which add the xylose moieties (Zabotina, 2012). For both these genes, phylogenomic analysis revealed distinct phylogenetic and syntenic patterns associated with their functional diversification between Poaceae and eudicots (**Figure 5**). Specifically, for the *XXT* family we found a total of six phylogenetic clades corresponding to six independent genomic contexts (**Figure 5A**). Three clades/communities grouped only eudicot genes and corresponded to the homologs of arabidopsis *XXT1*, *XXT2*, and *XXT3/5* genes, respectively. These are the most important *XXT* copies for XyG synthesis in arabidopsis, forming an active biosynthetic complex (Zabotina, 2012). Conversely, the other three clades/communities were specific to Poaceae and non-commelinid monocot genes, and included the maize homologs of *AtXXT1*, *AtXXT2*, and *AtXXT3/5* genes, respectively. To conclude, the eudicot *XXT1* and *XXT2* clades displayed a relatively large extent of interclade synteny and tree branches of similar sizes, highlighting relatively little phylogenetic differentiation. Conversely, grass *XXT1* and *XXT2* are independently syntenically organized and phylogenetically more distant (**Figure 5A**). Regarding *CslC* genes, phylogenomic analysis identified eight distinct syntenic communities, four of which were specific to either grasses or eudicots, while the other four included both eudicot and Poaceae genes (**Figure 5B**). The syntenic differentiation of *CslC* genes between Poaceae and eudicots is thus not absolute. However, BLAST analyses showed that the largest eudicot-specific *CslC* community included *AtCslC4*. This is the most active arabidopsis *CslC* gene and the only one highly expressed in all arabidopsis tissues (Zabotina, 2012). The two closest maize homologs of *AtCslC4* – XP_008662691.2 and XP_008657194.1 – were included in the two grass-specific *CslC* communities. To conclude, copy number community data indicated that eudicot- and grass-specific *CslC* communities comprise the majority of *CslC* members from all angiosperms (342 out of 596 total genes).

In addition to XyGs, (gluco)mannans content is very different between Poaceae and eudicot cell walls. (Gluco)mannans synthesis depends largely on *Cellulose synthase-like A* (*CslA*) genes, which bind the mannose residues (Zhong et al., 2019). Remarkably, phylogenomic analysis showed that *CslA* genes are genomically very differently organized between Poaceae and eudicots (**Figure 6**). Specifically, *CslA* members are divided into 10 different syntenic communities, of which five specific to eudicots (two restricted to Brassicaceae), and five specific to Poaceae and, partly, non-commelinid monocots. Of the five Poaceae-specific communities, two contain genes which are phylogenetically closer to eudicot *CslA*, while the members of the other three form a monocot-specific *CslA* phylogenetic clade (bootstrap = 97). Interestingly, Poaceae *CslA* copy number is highest in the largest community corresponding to such monocot-specific phylogenetic clade, with >3 *CslA* genes per species on average. Moreover, the three phylogenetically-distinct Poaceae-specific communities group the majority of Poaceae *CslA* genes (140 out of 207 copies).

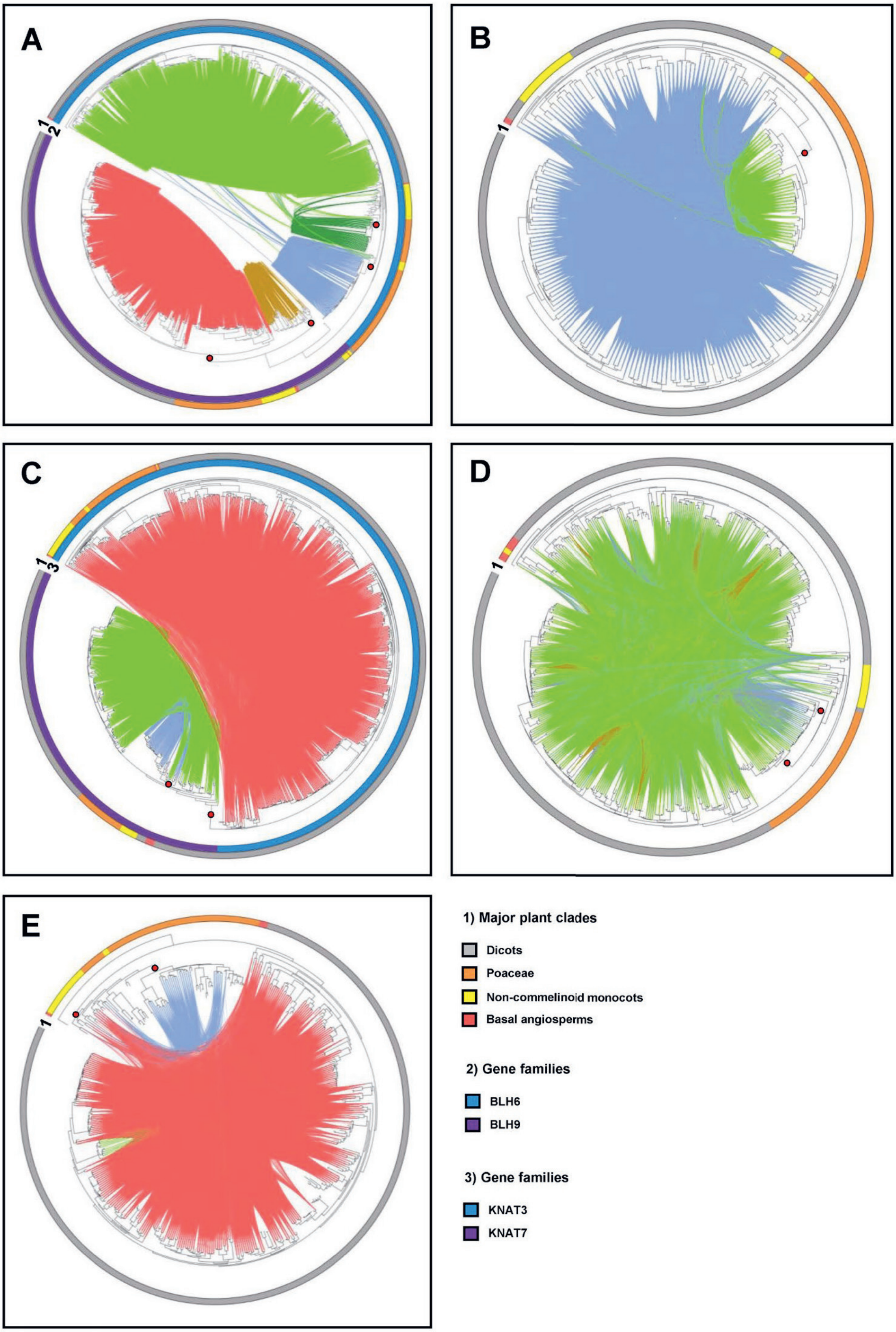


Figure 4. See caption at the top of next page.

Figure 4 (page 69). Phylogenetic trees displaying phylogenetic and syntenic relationships of the gene families analysed in relation to the *BLH6/BLH9* study case. Overall, the trees display that genes known to concur to the different *BLH6* function in Poaceae and eudicots are all differentially genomically (syntenically) organized between these species. Conversely, genes related to the *BLH6/BLH9* pathway but with conserved function across Poaceae and eudicots display conserved synteny across most or all angiosperms. In each plot, lines connecting tree leaves represent syntenic relationships between genes, classified according to the detected syntenic communities (different line colours indicate different syntenic communities). Rings around trees display the taxonomic profiling of the genes within each tree, or the gene family to which each gene in a tree belongs to (in the case the genes from multiple gene families were aligned together to build one tree) (see legends). Red dots on tree nodes indicate tree branches that are relevant for the phylogenomic diversification of gene sequences that are supported by bootstrap >=90. A) *BLH6/BLH9*. B) *OFP4*. C) *KNAT3/KNAT7*. D) *NST*. E) *F5H*.

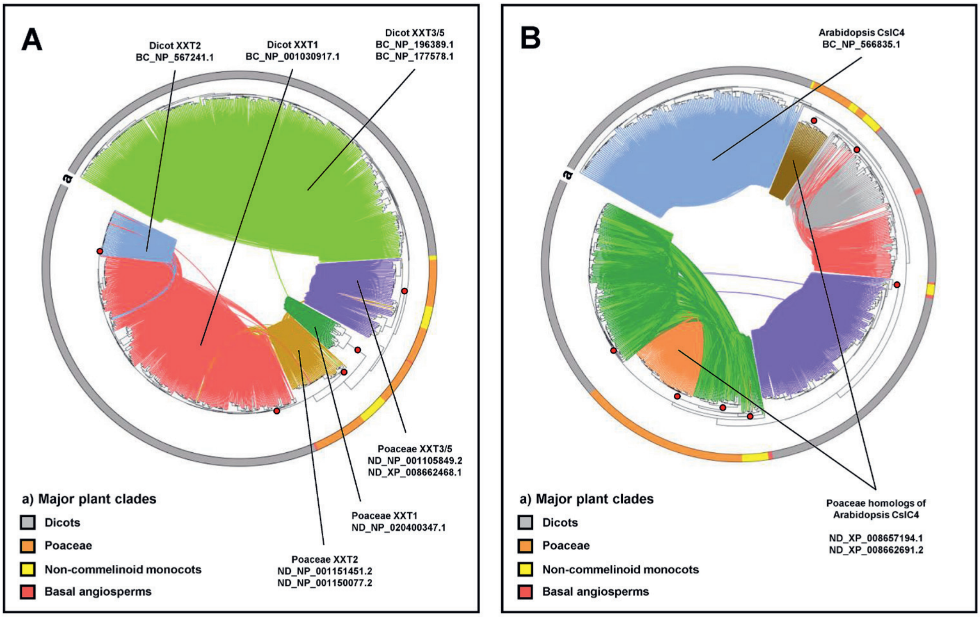


Figure 5. Phylogenetic trees displaying phylogenetic and syntenic relationships of the two main gene families involved in XyG biosynthesis: *XXT* (**A**) and *CslC* (**B**). The trees show that these genes tend to be differentially organized in Poaceae and eudicots from a genomic (synteny) point of view. Specifically, the functionally most important *XXT* and *CslC* members known in arabidopsis and their respective homologs in Poaceae (maize) are organized in completely different genomic contexts. In each plot, lines connecting tree leaves represent syntenic relationships between genes, classified according to the detected syntenic communities (different line colours indicate different syntenic communities). Rings around trees display the taxonomic profiling of the genes within each tree (see legends). Red dots on tree nodes indicate tree branches that are relevant for the phylogenomic diversification of gene sequences that are supported by bootstrap >=90. The syntenic communities containing arabidopsis and maize genes of interest are indicated (black arrows). Within gene IDs, "BC" indicates arabidopsis, while "ND" indicates maize.

Xylans represent a final relevant group of hemicellulosic molecules for type I and type II cell walls. Their most important biosynthetic genes are three different *IRX* families – *IRX9/9*-like (CAZy GT43 family), *IRX10/10*-like (CAZy GT47 family), and *IRX14/14*-like

(CAZy GT43 family) – that form a xylan biosynthetic complex (Zhong et al., 2019). As for the other hemicellulose-related genes above, phylogenomic analyses revealed substantial divergence in the positional organization, phylogenetic diversification, and copy number dynamics of *IRX* genes between Poaceae and eudicots (**Figure 7**). While two syntenic communities conserved across Poaceae, non-commelinid monocots, and eudicots were detected within each of the *IRX9/9L*, *IRX10/10L*, and *IRX14/14L* clades, 54% of all the *IRX* genes studied (1636 out of 3030) were included within 25 different syntenic communities either eudicot- or Poaceae-specific. Moreover, differential copy number representation of the *IRX* genes between Poaceae and eudicots was observed for all the *IRX* families within the syntenic communities conserved across both

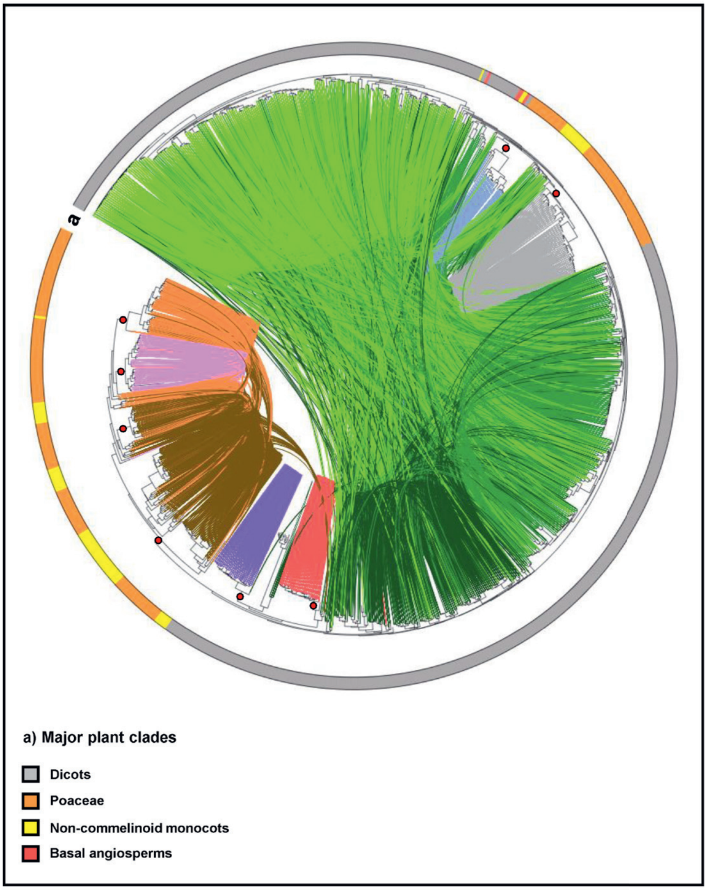


Figure 6. Phylogenetic tree displaying the phylogenetic and syntenic relationships of the main gene family involved in mannan biosynthesis: *CslA*. The tree shows that these genes are organized in (multiple) different genomic contexts in Poaceae and eudicots. Within the tree, lines connecting the leaves represent syntenic relationships between genes, classified according to the detected syntenic communities (different line colours indicate different syntenic communities). The ring around the tree displays the taxonomic profiling of the tree genes (see legends). Red dots on tree nodes indicate tree branches that are relevant for the phylogenomic diversification of gene sequences that are supported by bootstrap >=90.

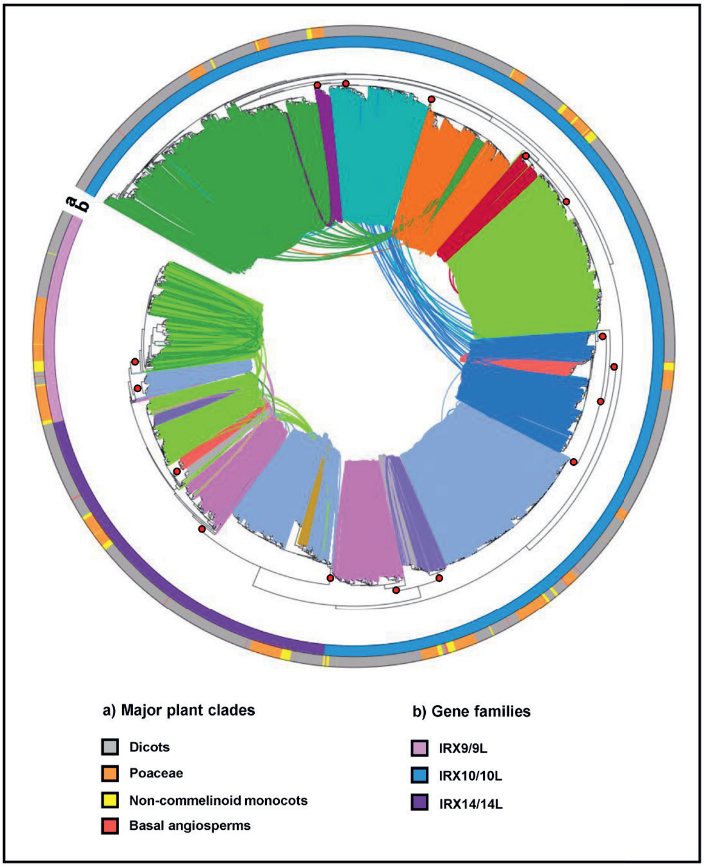


Figure 7. Phylogenetic tree displaying the phylogenetic and syntenic relationships of the three main gene family involved in xylan biosynthesis: *IRX9/9L*, *IRX10/10L*, and *IRX14/14L*. The tree shows that these genes are organized in (multiple) different genomic contexts in Poaceae and eudicots. Within the tree, lines connecting the leaves represent syntenic relationships between genes, classified according to the detected syntenic communities (different line colours indicate different syntenic communities). The rings around the tree display the taxonomic profiling of the tree genes and the genes belonging to each *IRX* gene family (see legends). Red dots on tree nodes indicate tree branches that are relevant for the phylogenomic diversification of gene sequences that are supported by bootstrap >=90.

Poaceae and eudicots (**Figure 8**), showing that the relative representation of shared *IRX* genomic contexts can differ substantially between Poaceae and eudicot genomes. To conclude, all the *IRX* communities identified are highly phylogenetically differentiated, corresponding to distinct phylogenetic clades supported by high bootstrap.

Cell wall genes are often organized in tandem gene clusters which are different between Poaceae and eudicots

To further characterize the cell wall genomic properties of Poaceae and eudicots, the (differential) occurrence and conservation of tandem cell wall gene clusters was also studied for the 150 gene families and the 169 genomes. The presence and evolution of tandem gene clusters affect plant traits in several ways, for example by influencing gene dosage or by facilitating gene sub- or neo-functionalization (Kono et al., 2018). Therefore, genomic variability of tandem gene arrays can be relevant also for type I – II cell walls differentiation.

Our analyses revealed that tandem gene clusters are relatively common for cell wall genes, as 44 of the 150 gene families analysed have (part of) their genes organized in tandem arrays of at least two members in 60% or more of the species surveyed. Conversely, 85 gene families (57% of the total) are mainly organized as (distinct) singleton loci (tandem arrays found in <40% of the species surveyed). The remaining 21 gene families displayed mixed singleton/tandem patterns, depending on the species (**Supplementary Table 5**). Overall, genes involved in cellulose, lignin, and pectin biosynthesis are the ones most often organized as tandem clusters. Conversely, cell wall transcription factors and genes involved in sugar supply for cell wall biosynthesis are the classes mostly arranged as singleton loci (**Table 3**).

Given our focus on the genetics underlying Poaceae and eudicot cell wall differences, the variability of cell wall gene tandem clusters between these two groups of species was studied more in detail. This analysis revealed that 20 of the 44 cell wall gene families with high occurrence of tandem gene clusters displays marked differences in the properties of such clusters between Poaceae and eudicots (**Table 4**). Specifically, such differences entail either the presence/absence variability of tandem clusters between Poaceae and eudicots (11 gene families), or the number of genes included in tandem arrays between Poaceae and eudicots (9 gene families). Remarkably, some of the gene families displaying variability of gene tandem arrays between Poaceae and eudicots are particularly relevant for the differences between type I and type II cell walls. For example, they include two critical families affecting lignin content: *PAL* – displaying one tandem cluster of five genes per species on average in Poaceae, but no clusters in eudicots – and *PRX* – which displays 17 tandem arrays of four genes per species on average in Poaceae and 9 clusters of three genes each on average per eudicot species. Moreover, four important pectin-related gene families – *XYLOGALACTURONAN XYLOSYLTRANSFERASE* (*XGD*), *PME/PMEI*, *PE*, and *PG* – all display a higher occurrence of tandem gene clusters (with also more genes per cluster) in eudicots compared to Poaceae (**Table 4**). Interestingly, for several gene families the variability of tandem gene clusters between Poaceae and eudicots goes in

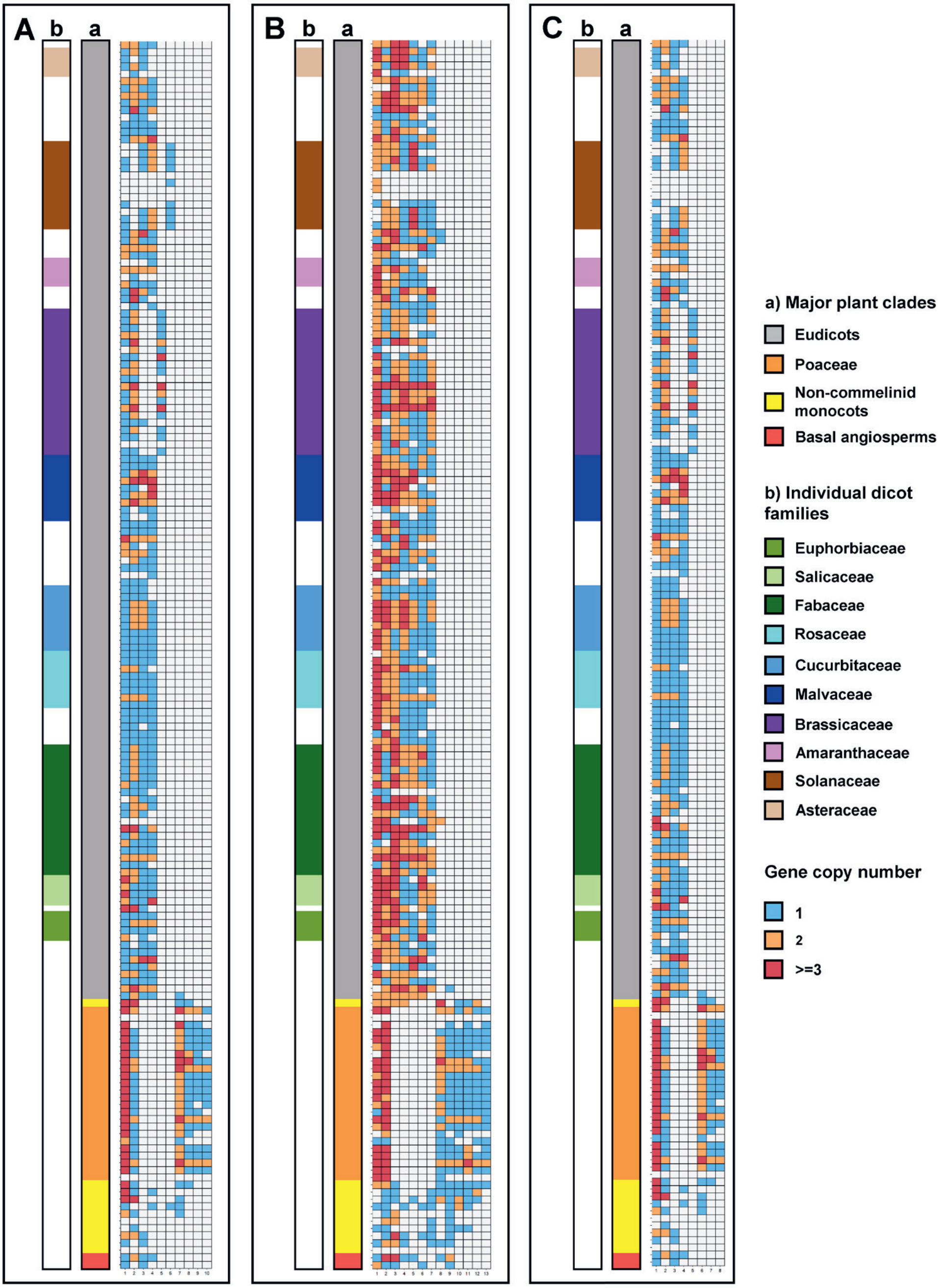


Figure 8. Heatmaps showing the genomic context- and taxonomi clade-specific patterns of gene CNV across the three *IRX* gene families involved in xylan synthesis. In all the heatmaps, columns represent syntenic communities detected for a specific gene family, while rows represent species, ordered taxonomically (see legend). Heatmap cells are coloured according to the gene copy number for every genomic context – species combination (see legend). A) *IRX9/9L.* B) *IRX10/10L.* C) *IRX14/14L.*

Table 3 – Statistics of gene tandem clusters presence across broad categories of cell wall gene functions.

Cell wall process	Gene families organized in tandem clusters (%)*	Gene families organized in tandem clusters (number) *	Gene families not organized in tandem clusters (%)**	Gene families not organized in tandem clusters (number)**	Gene families displaying unclear patterns (%)***	Gene families displaying unclear patterns (Number) ***
Callose synthesis	0.0	0	0.0	0	100.0	1
Cellulose synthesis	46.2	6	38.5	5	15.4	2
Glucose supply to cell wall	0.0	0	100.0	8	0.0	0
Hemicellulose metabolism	30.0	15	60.0	30	10.0	5
Lignin synthesis	42.1	8	31.6	6	26.3	5
Pectin metabolism	36.8	7	42.1	8	21.1	4
Transcription factors	0.0	0	94.1	16	5.9	1
Other cell wall genes	38.1	8	57.1	12	4.8	1

* Tandem clusters occur in >60% of the species analysed.
** Tandem clusters occurring in <40% of the species analysed.
*** Pattern observed does not fall in any of the two previous categories (* and **).

Table 4 – Gene families displaying large variability in cell wall gene tandem clusters between Poaceae and eudicots as either differential number of tandem clusters or presence/absence of tandem clusters.

Cell wall gene function	Average number of gene tandem clusters per species (Poaceae)	Average number of genes per cluster (Poaceae)	Average number of gene tandem clusters per species (Eudicots)	Average number of genes per cluster (Eudicots)
ARAD	3	3	1	2
MUR2/3/4	3	3	1	2
FUT	4	3	2	2
EXP/EXPL	8	4	2	3
PME/PMEI	2	2	9	3
PRX	17	4	9	3
PE	4	2	11	3
PG	7	3	11	3
BBE_like	1	5	3	7
GALT	2	3	0	2
ERF	1	5	0	2
XXT	1	3	0	2
PAL	1	5	0	3
HCT	0	-	1	3
DRP/ADL	0	-	1	3
FLA11/12	0	-	1	4
ESB1	0	-	1	3
PLY	0	-	1	3
THE1	0	-	1	3
XGD	0	-	2	3

parallel with the differences in gene copy-number variation (see commonalities between **Table 1 and 4**). This highlights once again the extensive interaction between the genomic factors considered in this study in shaping the genomic contexts of Poaceae and eudicot cell walls.

4 Discussion

The biological basis of the profound differences between the type I cell walls of eudicots and type II cell walls of Poaceae is not yet fully understood. Such differences entail nearly all the cell wall components, and represent the likely product of the combined evolution of the different angiosperm lineages and of cell walls themselves as plant structures (Vogel, 2008, Sarkar et al., 2009). Typically, such complex evolutionary trajectories leave traces in plant genes and genomes (Zhao et al., 2017, Kerstens et al., 2020). However, the occurrence, extent, and relevance of a genetic differentiation underlying type I and type II cell walls are still debated (Yokoyama and Nishitani, 2004, Penning et al., 2019, Yokoyama, 2020, Kozlova et al., 2020). This study revealed major differences between Poaceae and eudicots in the genomic organization of several cell wall genes. This differentiation involves multiple genomic properties, including presence/absence and copy-number variation of genes and gene tandem arrays, syntenic gene conservation, and gene sequences diversification. Moreover, the magnitude of such differentiation is exceptional, as the vast majority of the gene families analysed displays at least one pattern of genomic diversification between Poaceae and eudicots. Finally, for several gene families for which functional characterization is available, differential genomic patterns correlate strikingly with the cell wall differences between Poaceae and eudicots (**Figure 9**). Considering that genomic properties as differential gene synteny and copy number variation are known sources of functional gene diversification and phenotypic innovations (Flagel and Wendel, 2009, Zhao et al., 2017, Artur et al., 2019, Lye and Purugganan, 2019, Kerstens et al., 2020), the profoundly different cell wall genomic properties between Poaceae and eudicots are a potential major driver of the cell wall differentiation between these two plant clades (**Figure 9**).

In light of what just discussed, it is relevant to understand how the different cell wall “genomic landscapes” of Poaceae and eudicots got shaped and how they can lead to phenotypic cell wall diversification. In this regard, the differentiation of syntenic genomic contexts seemed to have played a major role. The fact that differentially-organized genomic gene contexts can facilitate the functional diversification of plant genes, leading to novel phenotypes, has been amply discussed, as for the *MADS-box* (Zhao et al., 2017), *APETALA2* (Kerstens et al., 2020), *LEA* (Artur et al., 2019), or *NRT* plant gene families (Zoghbi-Rodríguez et al., 2021). Typically, different (novel)

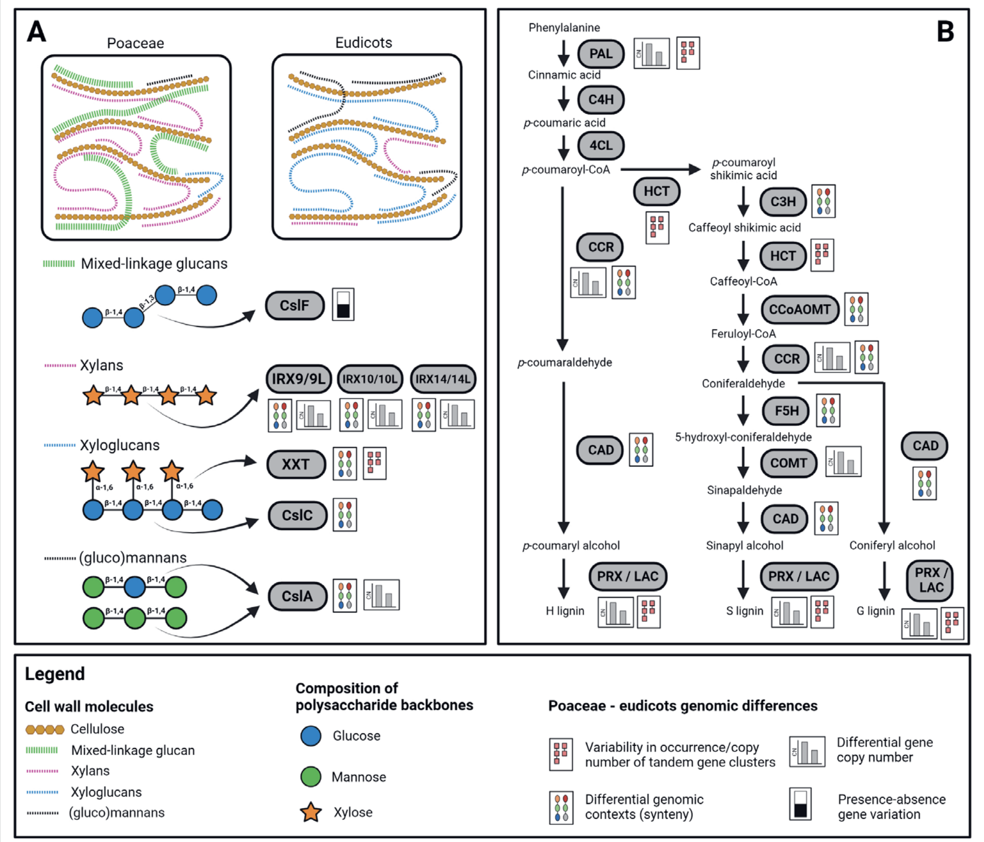


Figure 9. Summary of the differential genomic patterns between Poaceae and eudicots found for major hemicellulose and lignin genes. A) Differences in hemicellulose composition between Poaceae and dicot cell walls (upper part), with schematic representation of the backbones of the polysaccharides responsible of these differences, along with the indication of the major genes synthesizing polysaccharide backbones and the differential genomic patterns found for those genes between Poaceae and dicots. B) Representation of the lignin pathway, with the indication of genes and differential genomic patterns found for those genes between Poaceae and dicots.

genomic contexts can differentiate the modes of gene function, for example by determining different patterns of gene expression, or by favouring gene diversification and sub-functionalization (Dewey, 2011). In turn, novel functional modes can affect plant phenotypes and can be stabilized if selected over evolution (Dewey, 2011). In this study, the *BLH* case (Section 2.2.1) clearly demonstrated that this could be the case also for cell wall genes. In fact, the major genomic difference between the *BLH9* family – whose function is conserved in Poaceae and eudicots – and the *BLH6* family – which has opposite function in Poaceae and eudicots – was found to be the occurrence of differential *BLH6* genomic contexts between Poaceae and eudicots. Moreover, differential genomic contexts were also observed for all the genes directly controlled by *BLH6* and that concur to determine its differential function in Poaceae and eudicots, as *OFP4*, *KNAT7*, and *F5H*. Conversely, all the *BLH*-related genes whose function is believed to be conserved across all angiosperms – as *NST* transcription factors and *KNAT3* homologs (Rao and Dixon, 2018) – displayed conserved positional genomic organization. Finally, both CNV and phylogenetic differentiation supported by differential selection pressures were observed between the gene clades corresponding to distinct Poaceae and eudicot genomic contexts for the majority of the genes within the *BLH6* pathway. Overall, these findings suggest that the differential *BLH* genomic contexts may facilitate the different functionality of the *BLH6* pathway in Poaceae and eudicots, in ways similar to what reported in the studies referenced above. The fact that analogous patterns were also found for the major hemicellulose genes, as well as for many other cell wall structural genes – even if at a more general level – further corroborates this hypothesis, and extends it to several cell wall gene families.

Interestingly, the differentiation of genomic gene contexts – even if highly extensive across and within the 150 gene families analysed – was not absolute, especially for some of the hemicellulose-related genes that were analysed in Section 2.2.2. In light of the likely cell wall functional relevance of differentiated genomic landscapes, the presence of shared genomic gene configurations between Poaceae and eudicots, next to highly divergent genomic contexts, for some of the genes participating to the synthesis of polysaccharides that differ between grasses and the rest of angiosperms (as the *CslC* and *IRX* genes of Section 2.2.2) could indicate that not all the genes of these families perform the same function. In this regard, it is noteworthy that arabidopsis mutants at some of the *CslC* and *IRX* genes have been shown to display alterations of plant growth (Kim et al., 2020b) and of seed viability (Voiniciuc et al., 2015), suggesting the possibility for their direct or indirect involvement in multiple plant functions (Little et al., 2018). In this sense, the presence of conserved genomic gene contexts between Poaceae and dicots could be constrained by the involvement of some cell wall genes in particularly vital plant processes, next to their strict relationship with cell wall biosynthesis. In turn, this would once again further corroborate the hypothesis that diversification of genomic gene contexts facilitates the evolution of novel gene functions, as already advanced for other gene families (Dewey, 2011, Zhao et al., 2017, Kerstens et al., 2020). However, it has to be stressed that this remains currently only a hypothesis based on our results, and final functional proof to support it should be provided in future research.

In addition to differential syntenic gene organization, gene CNV emerged as another major force shaping the differential genomic landscapes of Poaceae and eudicots. As genomic context variability, also gene CNV represents a well-known source of functional variation that can lead to phenotypic innovations (Jiao et al., 2011, Kondrashov, 2012, Lye and Purugganan, 2019). While duplicated variants are usually deleterious and commonly undergo purifying selection (Lye and Purugganan, 2019), the extensive CNV between Poaceae and eudicots for a large part of cell wall gene families highlights the relevance of this process for the differentiation of the cell wall genomic landscape between these plant clades. Typically, CNV impacts gene function through gene dosage (Kondrashov, 2012). The combined analysis of CNV and gene functionality for the genes of this study suggests that this could be the case for several cell wall gene families. For example, lignin genes that are particularly important to determine the total cell wall lignin content were found in much higher copy number in Poaceae – which have higher lignin content – than in eudicots (**Table 1**). One of these is *PAL*, which initiates the lignin pathway and determines the efficiency of phenylalanine conversion into lignin precursors (Zhong et al., 2019). A second example is *CYNNAMOYL-CoA REDUCTASE* (*CCR*), which is the entry point of each of the separate branches leading to the synthesis of each monolignol within the lignin pathway (Tamasloukht et al., 2011). Finally, *PRX* genes, which mediate the *in muro* polymerization and deposition of monolignols (Zhong et al., 2019), were also found in much higher copy number in Poaceae compared to eudicots. All together, these results suggest that a higher copy number of genes mediating critical steps of lignin synthesis might reasonably "boost" the lignin pathway in Poaceae through a higher amount of gene products, leading to higher lignin production compared to eudicots. Beyond genes strictly involved in the lignin pathway, correlations between CNV and cell wall composition were found in several other cases (**Table 1**), including genes as *BAHD* and *GT61*, which are involved in the cross-linking of cell wall polymers and lignin via ferulic acid (de Souza et al., 2018, Cenci et al., 2018, Feijao et al., 2022). The feruloylation of cell wall molecules in grasses is highly important, as it increases biomass recalcitrance to industrial processing (de Souza et al., 2018). Remarkably, both *BAHD* and *GT61* genes were found in much higher copy number in grasses (**Table 1**), which could reasonably explain the higher prevalence of ferulic acid in Poaceae cell walls through higher gene dosage. In turn, this could open possibilities for the modification of cell wall feruloylation in grasses, through the silencing of certain *BAHD* or *GT61* members, or the analysis of natural variation for *BAHD* and *GT61* copy number among wild accessions of Poaceae species. To conclude, all these examples, along with all the other genes of **Table 1**, suggest that cell wall variation due to differential gene dosage dependent on CNV represents a likely important mechanism for the differentiation of Poaceae and eudicot cell walls.

Apart from the patterns reported in **Table 1**, CNV was also commonly observed between differentially conserved genomic contexts, different species clades within specific genomic contexts, or distinct phylogenetic clades. These patterns reflect an interaction between positional gene organization, CNV, and phylogenetic sequence diversification that took place during the evolution of cell wall genomic landscapes. Most likely, the combination of all these factors during the genomic differentiation of Poaceae and eudicots has contributed to deepen and stabilize the cell wall gene functional diversification between these plant clades. This appears reasonable as such "evolutionary boost" effect of the parallel differentiation of multiple genomic properties across diverse gene clades and families was already hypothesized for other genes, as the different *LEA* proteins (Artur et al., 2019). Moreover, such multiple differential genomic patterns were found in all the study cases of this research involving genes that are most likely behaving differently in Poaceae and eudicots, as the genes within the *BLH6* regulatory pathway, or the *XXT*, *CslC*, and *IRX* genes at the basis of hemicellulose synthesis.

A final genomic property analysed in this study is the occurrence and conservation of tandem cell wall gene arrays. To a certain extent, the variability of these configurations can have similar effects to CNV in terms of gene dosage, and can even overlap with CNV itself (Kono et al., 2018). However, the occurrence of gene tandem clusters can also facilitate gene diversification and trait variability (Picart-Picolo et al., 2020, Xu et al., 2020). Our results suggest that this could be the case also in the case of several cell wall genes between Poaceae and eudicots, especially considering that variability in tandem gene arrays goes in parallel with differentiation of genomic contexts and/or basic CNV. As an example, in the case of the previously mentioned *PAL* and *PRX* genes, the finding of a higher occurrence of gene tandem arrays in Poaceae may facilitate the combined expression of the clustered genes in these species, thus leading to higher dosage of gene products as previously hypothesized. Moreover, the combination of tandem arrays variability and the differentiation of the other genomic properties studied may further amplify and stabilize the differentiation of Poaceae and eudicot cell walls, as previously discussed for genomic contexts, CNV, and phylogenetic differentiation.

In conclusion, this study clearly showed that major genomic differences underly the divergent cell walls of Poaceae and eudicots. Such differences involve several major genomic properties, and hypotheses have been discussed regarding their evolutionary origin and the biological modes by which they can translate into different cell walls. At this moment, the scarce knowledge about the specific function of several of the genes considered in this research in other species than arabidopsis hampers a further interpretation of the patterns found beyond what discussed above. Therefore, we foresee that a further detailed characterization of cell wall genes in several species, together with the results reported in this study, will advance the investigation of the genetic basis of the different cell walls of Poaceae and eudicots, within the context of the genomic patterns found in this research. Moreover, the data of this study can offer opportunities for novel approaches of fundamental cell wall research in Poaceae and eudicots (e.g. genomic context engineering), as well as for the identification of gene targets to modify cell wall composition.

Supplementary Data

Supplementary Tables can be retrieved at: https://doi.org/10.7910/DVN/PHWYZV
These tables include:

Supplementary Table 1: The cell wall genes considered in this study for all the 169 genomes, classified into 150 gene families.

Supplementary Table 2: The 169 angiosperm genomes used in the study.

Supplementary Table 3: Complete gene copy number data for the 169 genomes and the 150 cell wall gene families studied.

Supplementary Table 4: Copy number data of the 150 cell wall gene families of the study, highlighting differences between Poaceae and dicots.

Supplementary Table 5: Tandem cell wall gene clusters statistics.

Supplementary Table 6: The 1312 arabidopsis cell wall genes used as seeds in the search for all the cell wall genes in all the 169 angiosperm genomes.

Supplementary Table 7: Results of PAML tests for selection pressure on the genes within the BLH6 pathway.

Supplementary Dataset 1 can be retrieved at https://doi.org/10.4121/21564756

Supplementary Dataset 2 can be retrieved at https://doi.org/10.4121/22068791

Chapter 4

Marginal lands to grow novel bio-based crops: a plant breeding perspective

Francesco Pancaldi[1], Luisa M. Trindade[1]

[1]Plant Breeding, Wageningen University & Research, Wageningen,
The Netherlands

Published as: Pancaldi, F. and Trindade, L.M. (2020). Marginal Lands to Grow Novel Bio-Based Crops: A Plant Breeding Perspective. *Frontiers in plant science*, *11:227*.
DOI: 10.3389/fpls.2020.00227

Abstract

The biomass demand to fuel a growing global bio-based economy is expected to tremendously increase over the next decades, and projections indicate that dedicated biomass crops will satisfy a large portion of it. The establishment of dedicated biomass crops raises huge concerns, as they can subtract land that is required for food production, undermining food security. In this context, perennial biomass crops suitable for cultivation on marginal lands (MALs) raise attraction, as they could supply biomass without competing for land with food supply. While these crops withstand marginal conditions well, their biomass yield and quality do not ensure acceptable economic returns to farmers and cost-effective biomass conversion into bio-based products, claiming genetic improvement. However, this is constrained by the lack of genetic resources for most of these crops. Here we first review the advantages of cultivating novel perennial biomass crops on MALs, highlighting management practices to enhance the environmental and economic sustainability of these agro-systems. Subsequently, we discuss the preeminent breeding targets to improve the yield and quality of the biomass obtainable from these crops, as well as the stability of biomass production under MALs conditions. These targets include crop architecture and phenology, efficiency in the use of resources, lignocellulose composition in relation to bio-based applications, and tolerance to abiotic stresses. For each target trait, we outline optimal ideotypes, discuss the available breeding resources in the context of (orphan) biomass crops, and provide meaningful examples of genetic improvement. Finally, we discuss the available tools to breed novel perennial biomass crops. These comprise conventional breeding methods (recurrent selection and hybridization), molecular techniques to dissect the genetics of complex traits, speed up selection, and perform transgenic modification (genetic mapping, QTL and GWAS analysis, marker-assisted selection, genomic selection, transformation protocols), and novel high-throughput phenotyping platforms. Furthermore, novel tools to transfer genetic knowledge from model to orphan crops (i.e. universal markers) are also conceptualized, with the belief that their development will enhance the efficiency of plant breeding in orphan biomass crops, enabling a sustainable use of MALs for biomass provision.

1 Introduction

A global bio-based economy where building-block materials, chemicals and energy are derived from biological biomass could significantly mitigate main environmental and social problems of our fossil-based society, including climate change, environmental pollution, and geopolitical tensions (McCormick and Kautto, 2013, Bennich and Belyazid, 2017). To address this challenge, more than forty governments worldwide have explicitly set up strategies to transit towards bio-based economic systems (Dietz et al., 2018), and projections indicate that the biomass demand to sustain bio-based production chains will amount to 6.7-13.4 Bln tons year^{-1} in 2050, with an increase of 198-396% compared to 2011 levels (3.4 t year^{-1}) (Piotrowski et al., 2015). Within the range of biomass sources to sustain such demand, dedicated herbaceous and woody crops have and will keep a prominent position (Piotrowski et al., 2015, OECD, 2017). This raises concerns, as the allocation of agricultural land to biomass production conflicts with the cultivation of food crops (Dauber et al., 2012, Gelfand et al., 2013), hampering food security, destabilizing food prices, and constraining the access to food, especially to poor rural communities (Mol, 2010, Ribeiro, 2013).

To avoid conflict with food production, dedicated bio-based crops could be grown on marginal lands (MALs) (Samson and Girouard, 1998, Tilman et al., 2006, Tilman et al., 2009, Fritsche et al., 2010, Gelfand et al., 2013, Post et al., 2013, Zhu et al., 2016, Mehmood et al., 2017, Carlsson et al., 2017), which are areas with marginal agronomic and economic potential for cultivation of food crops and currently not used by agriculture (Peterson and Galbraith, 1932, Dauber et al., 2012, Shortall, 2013, Post et al., 2013, Schmidt et al., 2015, Mehmood et al., 2017). Typical examples include abandoned lands (no longer used due to relocation of agriculture (Campbell et al., 2008)), degraded lands (no longer productive due to an intensive and unsustainable use (Daily, 1995, Campbell et al., 2008, Dauber et al., 2012)), or waste lands (with physical or environmental constraints to agriculture (Cai et al., 2010)). MALs comprise 247-729 Mha worldwide (Hoogwijk et al., 2003, Tilman et al., 2006, Smeets et al., 2007, Campbell et al., 2008, Field et al., 2008, Haberl et al., 2010, Haberl et al., 2011, Nijsen et al., 2012), thus retaining a great potential for growing dedicated biomass feedstocks. In fact, considering an average biomass yield of 7.9 t ha^{-1} from cultivation of herbaceous and woody biomass species on MALs (Nijsen et al., 2012), such acreages would be enough to supply 28-85% and 14-42% of the total biomass demand in 2050 under the 6.7 and the 13.4 Bln t scenarios, respectively.

To successfully grow bio-based crops on MALs it is critical to offset the factors that determine the marginality of these areas (Fritsche et al., 2010, Dauber et al., 2012). Most of these factors are physical and agronomic, such as adverse land morphology (e.g. steep slopes), unfavorable soil conditions (e.g. low fertility, salinity, acidity, erodibility, poor drainage, contamination by heavy metals), or hostile climate (e.g. recurrent droughts and floods, extreme temperatures) (Dauber et al., 2012, Gelfand et al., 2013, Jones et al., 2014, Smaliychuk et al., 2016, Meyfroidt et al., 2016). Others are socio-economic, as the lack of adequate transport infrastructures, undefined land ownership, or low mechanization levels and negative demographic conditions (e.g. low population density and educational level) in the regions where MALs are located (Fritsche et al., 2010, Dauber et al., 2012, Smaliychuk et al., 2016, Meyfroidt et al., 2016). Offsetting marginality through the planting of bioenergy crops requires therefore the design of cropping systems that match the site-specific conditions of each MAL at all the levels of the agricultural practice.

At the crop and agronomic level, this means adopting crops and practices that can sustain large yields of high-quality biomass (see section 3.2 for the definition of "biomass quality" in the context of biomass production on MALs) under unfavorable conditions, minimize input requirements and promote an ecological restoration of MALs (Zegada-Lizarazu et al., 2010, Dauber et al., 2012, Allwright and Taylor, 2016). Currently available crops hardly match these conditions, as they have been bred for centuries towards different targets than the production of bio-based commodities and the cultivation under MALs conditions (Trindade et al., 2010). As a result, their cultivation on MALs could even promote soil erosion and runoff of fertilizers and pesticides (National Research Council, 2008, Campbell et al., 2008), and would likely require high external inputs to sustain optimal biomass yields (de Vries et al., 2010, Parenti et al., 2018). An array of new crops tailored to marginal environments should therefore be developed, with a central role for plant breeding. This review aims at elucidating which crops are best to be sustainably grown on MALs, which traits should be prioritized in their improvement, and which tools are effective to advance those crops and traits.

2 Perennial lignocellulosic crops: a promising cropping system for biomass production on marginal lands

Several studies indicate mixtures of perennial lignocellulosic biomass crops grown under low-input agriculture as suitable cropping systems for biomass production on MALs (Tilman et al., 2006, Tilman et al., 2009, Dauber et al., 2012, Gelfand et al., 2013, Robertson et al., 2017, Mehmood et al., 2017, Carlsson et al., 2017). In this paragraph, we discuss why such systems fit well MALs conditions, and which factors are critical to ensure that their advantages are effectively delivered.

2.1 Advantages of cultivating mixed perennial lignocellulosic crops on marginal lands

Mixed perennial biomass crops (MPBCs) can couple the provision of lignocellulosic biomass with the improvement or the restoration of the ecological services of MALs at all the ecosystem levels.

Regarding soil properties, MPBCs enhance soil structure and reduce erosion (Blanco-Canqui et al., 2014, Cosentino et al., 2015, Blanco-Canqui, 2016, LeDuc et al., 2017, Fernando et al., 2018), owing to a dense and prolonged soil coverage, as well as deep and branched roots (Fernando et al., 2018), which hold large amounts of water and nutrients (Blanco-Canqui, 2016, Fernando et al., 2018). This, together with the high resource use efficiency of these crops (van der Weijde et al., 2013, Lewandowski, 2016, Carlsson et al., 2017) and their low or null fertilization requirements (Tilman et al., 2006, Robertson et al., 2017, Fernando et al., 2018), determines the very low levels of nutrient leaching (especially nitrogen – N) observed in MPBCs (Glover et al., 2010, Pérez-Suárez et al., 2014, Robertson et al., 2017, LeDuc et al., 2017, Fernando et al., 2018). For example, McIsaac et al. (2010) reported that, over a four-years comparison, unfertilized miscanthus and switchgrass leached on average 3.0 and 1.4 kg N ha^{-1} year^{-1}, respectively, far lower than a conventional maize-soybean rotation (40.4 kg N ha^{-1} year^{-1}). Similar conclusions have been reached in several other studies focused on different biomass perennials (e.g. poplar, willow, switchgrass, grass-legume), as thoroughly reviewed by Robertson et al. (2017). MPBCs improve also soil organic carbon (SOC) stocks (Anderson-Teixeira et al., 2009, Glover et al., 2010, Pérez-Suárez et al., 2014, Robertson et al., 2017, LeDuc et al., 2017, Fernando et al., 2018). In this regard, Chimento et al. (2016) showed that, over a six-years trial, perennial herbaceous (giant reed, miscanthus, switchgrass) and woody (poplar, black locust, willow) species accumulated on average 45% higher SOC than continuous tilled corn in the soil portion interested by roots. This large carbon (C) sequestration results from the continuous ground coverage, the low soil disturbance, and the large rooting systems of MPBCs (Carlsson et al., 2017), and is maximized when MPBCs are established over MALs that do not store large C stocks in their soils or natural vegetation (Tilman et al., 2009, Gelfand et al., 2013). Therefore, such "low-C MALs" should be prioritized for allocation to biomass production over areas that are naturally evolving to C-rich ecosystems (e.g. forests or wetlands) (Tilman et al., 2009, Robertson et al., 2017). An incautious use of C-rich MALs for biomass production could in fact even generate a large C debt (Schulze et al., 2012, Robertson et al., 2017), that could require decades or centuries for repaying (Fargione et al., 2008, Gibbs et al., 2008, Gelfand et al., 2011).

MPBCs can also contribute to preserve water resources. Their absolute water needs per ha generally equal – or even outweigh – annual crops as maize, wheat, or sorghum (Robertson et al., 2011c, van der Weijde et al., 2013, Hamilton et al., 2015, Robertson et al., 2017, Fernando et al., 2018), since their large biomass production and prolonged growing season implicate high evapotranspiration rates (Hamilton et al., 2015, Robertson et al., 2017). However, water use efficiencies (WUEs) of MPBCs are generally high (especially for C4 species) (Robertson et al., 2011c, Zeri et al., 2013, van der Weijde et al., 2013), which makes them best candidates for a water-effective production of biomass (Mehmood et al., 2017). This is especially true considering that several biomass perennials are drought tolerant (Mehmood et al., 2017), can be irrigated using wastewaters (Barbosa et al., 2015), and can suffice their water needs with seasonal rainfalls in temperate climates (Robertson et al., 2017). In addition, MPBCs improve water quality and water management, as their extensive roots, prolonged soil coverage, and positive effect on soil structure and porosity promote water penetration into soils (Fernando et al., 2018, Blanco-Canqui, 2016), minimizing water runoff and soil erosion (Blanco-Canqui, 2016, Acharya and Blanco-Canqui, 2018). Furthermore, the low leaching fluxes and little agrochemical needs of MPBCs minimize water pollution (Blanco-Canqui, 2016, Acharya and Blanco-Canqui, 2018). Finally, several perennial crops (e.g. willow, giant reed, miscanthus) can sequester contaminants (e.g. Cd, Pb, Zn, Cu) and pollutants from soils and water (Costa et al., 2016, Bandarra, 2013, Boléo et al., 2015), being therefore good options to remediate contaminated MALs and depurate polluted water streams (Acharya and Blanco-Canqui, 2018, Barbosa et al., 2015, Costa et al., 2016).

Biodiversity is also improved under the cultivation of MPBCs. Werling et al. (2014) showed that these cropping systems display a much wider diversity than monocultures of annual biomass crops as maize with respect to several taxonomic groups (microorganisms, arthropods, birds, and plants). These results are in line with several other studies that compared the diversity within specific biological clades in annual and perennial bioenergy crops, as Meehan et al. (2010) and Robertson et al. (2011b) (birds), Gardiner et al. (2010), Bennett et al. (2014) and Haughton et al. (2016) (insect communities), Levine et al. (2011) (methanotrophic bacteria), or Haughton et al. (2016) (plants). In addition, the specific change in land use from degraded MALs to MPBCs enhances biodiversity (Dauber et al., 2010, Meehan et al., 2010, Robertson et al., 2011a, Chauvat et al., 2014), and the enhancement is larger as the species diversity of the new cropping system is wider (Dale et al., 2010, Landis et al., 2018). The increase in biodiversity under biomass perennials boosts also fundamental ecosystem services for agriculture, as pollination (Bennett and Isaacs, 2014, Carlsson et al., 2017) or pest suppression (Werling et al., 2011, Carlsson et al., 2017), which allows for a reduced use of agrochemicals without compromising yields

(Carlsson et al., 2017). Moreover, these benefits are even extended to neighboring annual croplands, where pollination, pest suppression and yields can increase up to ~25%, thanks to the enhanced ecosystem functions of nearby MPBCs (Liere et al., 2015).

To conclude, MPBCs can also restore an economic value and an agrarian revenue to degraded lands, improving rural development. In developed countries, their cultivation contributes to diversify farm income from arable and grass lands, and offer new occupational perspectives to older or part-time farmers (Valentine et al., 2012). Alternatively, novel bioenergy production chains based on biomass from MALs can create employment opportunities in developing countries, as well as offer access to novel, clean, energy sources, which improves living and economic standards of local communities (Valentine et al., 2012).

2.2 Steps to establish mixed perennial biomass crops on MALs

To effectively deliver the benefits promised by MPBCs, the species and the management adopted should reflect the site-specific conditions of each target MAL (Zegada-Lizarazu et al., 2010, Blanco-Canqui, 2016, Robertson et al., 2017). Therefore, the preference should go for a wide array of dedicated and locally-adapted biomass perennials, each of which fitted to specific ecological niches, and globally suitable for diverse environmental scenarios (Robertson et al., 2017, Jones et al., 2015). **Table 1** reports a list of possible crops. These species typically withstand well the poor MALs conditions, showing constitutive resistance to several abiotic stresses, displaying high resource use efficiencies, and requiring low inputs (Zegada-Lizarazu et al., 2010, Dauber et al., 2012, Mehmood et al., 2017). However, most of them are novel or orphan crops, which did not undergo genetic improvement so far, especially with respect to biomass-related traits (Zegada-Lizarazu et al., 2010, Dauber et al., 2012, Jones et al., 2015, Zhu et al., 2016). As a consequence, their biomass yield and quality are highly variable (Zegada-Lizarazu et al., 2010), and often considerably lower than their genetic potential (Allwright and Taylor, 2016). This is critical, as the economic viability of cultivating MPBCs on MALs largely depends on their capacity of not only surviving structural and contingent suboptimal and low-input growing conditions, but of also producing, under those scenarios, large and high-quality biomass yields (Dauber et al., 2012, Blanco-Canqui, 2016). On the one hand, high-yielding and robust varieties significantly increase farmers' willingness of cultivating biomass perennials on MALs, by decreasing the opportunity cost of land and increasing profits (Soldatos, 2015). On the other hand, the provision of feedstocks with optimized lignocellulose composition in relation to the intended bio-based end uses (e.g. fermentation into biofuels, extraction of plant chemicals, or transformation into biomaterials as textile fibers) is critical to increase the profitability and competitiveness of the industrial use of plant biomass (Trindade et al., 2010). To enable the cultivation of MPBCs on MALs, it will be thus essential to breed varieties that couple robustness with optimal yields (Jones et al., 2015).

3 Target traits and genetic resources to tailor novel biomass crops to MALs

The improvement of biomass yield and quality in novel species for MALs depends on the enhancement of both biomass yield and quality *per se*, and biomass yield and quality *stability* under variable abiotic conditions. This paragraph illustrates which traits are critical to achieve these equally important breeding goals, which plant ideotypes are effective to guide the improvement of such traits, and which breeding resources are currently available.

3.1 Biomass yield

Biomass yield is a highly complex trait, overall determined by three factors: the efficiency of light interception by the crop canopy, the efficiency of light conversion into biomass, and the efficiency of biomass partitioning into target harvestable plant components (Monteith, 1972, Monteith, 1977). Thus, the enhancement of biomass yield encompasses breeding for all the morphological, physiological and phenological traits that are at the basis of these three factors (Jones et al., 2015). In addition, the characteristics of vegetative tissues should deserve particular attention, as in biomass crops they not only operate light interception and conversion into biomass, but are also the main harvest product.

Efficiency of light interception

The efficiency of light interception depends on plant architecture and duration and timing of crop growth (Jones et al., 2015). Plant architecture refers to crop growth habit (i.e. plant height and branching pattern) and leaves characteristics (i.e. number, size, shape, distribution and orientation) (Reinhardt and Kuhlemeier, 2002, Hollender and Dardick, 2015, Pan et al., 2017), and affects biomass yield by determining the plant density and biomass volume achievable per land unit, as well as the degree of soil coverage and photosynthetic area of the canopy. Tall plants, high tiller/stem number and density, thick stems, and upright, large and numerous leaves are all architectural characters positively correlated with biomass yield and light penetration and interception by the canopy in herbaceous biomass crops, as shown in **Figure 1** (Boe and Beck, 2008, Fernandez et al., 2009, Song et al., 2013). In woody species, vertical growth habit and production of sylleptic branches are also important to attain high plant densities and increase biomass yield per hectare (Rae et al., 2004, Marron et al., 2006, Dubouzet et al., 2013). For most of these traits, variability has been observed for crops suitable for MALs, as shown by e.g. Das et al. (2004) (switchgrass),

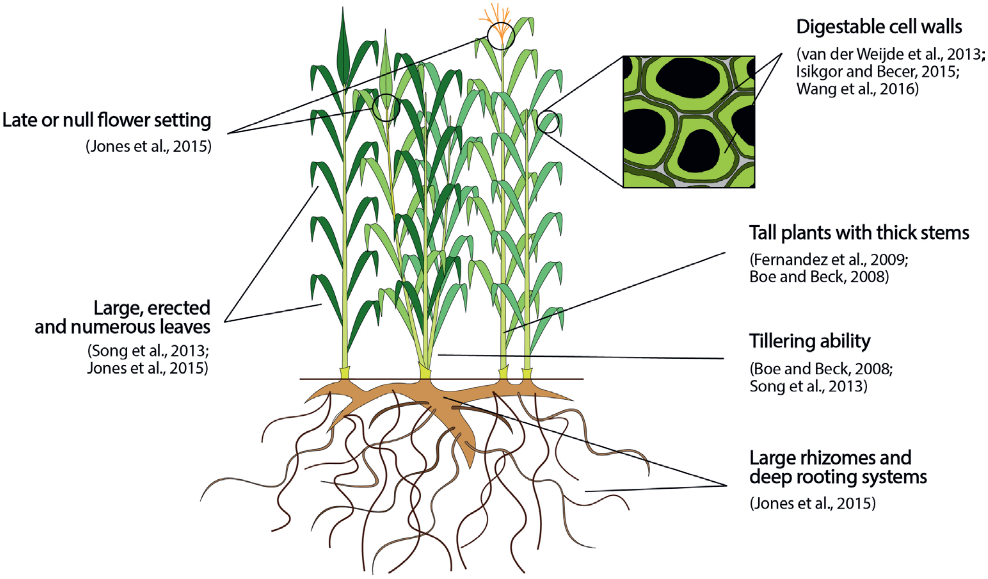


Figure 1. Preeminent architectural, phenological, and quality target traits to breed perennial lignocellulosic biomass crops for biomass production on MALs. The traits reported represent an ideotype to guide breeding activities, and the effective magnitudes of improvement attainable with respect to each trait can vary extensively depending on the species object of a breeding program.

Robson et al. (2013) (miscanthus), Cosentino et al. (2006) (giant reed), or Orlovic et al. (1998) and Rae et al. (2004) (poplar). In addition, QTLs underlying architectural traits have also been mapped in two of the most studied novel biomass species [miscanthus (Atienza et al., 2003, Gifford et al., 2015, Ge et al., 2018), and switchgrass (Serba et al., 2015)], as well as in model biomass crops as maize (e.g. Pan et al., 2017) or poplar (Bradshaw and Stettler, 1995, Zhang et al., 2006, Rae et al., 2008). These studies highlight the high genetic complexity of plant architecture, as particularly exemplified by Pan et al. (2017), who found nearly 800 QTLs associated with ten critical maize architectural traits, including plant height, number and length of branches, and leaf number, size, and orientation. Such a complexity hampers breeding for plant architecture in novel crops for which breeding tools are largely missing. Nevertheless, major-effect architectural loci have also been identified in model energy crops as maize, sorghum, or poplar, and in the model plant arabidopsis [see Wang and Li (2006), Fernandez et al. (2009), and Teichmann and Muhr (2015) for detailed reviews]. These loci could potentially be targeted in candidate gene approaches, by mining eventual homologous in novel biomass crops, screening allelic diversity, and selecting or introgressing favorable alleles. Preeminent examples are the *LG1* and *LG2* loci of maize, that, once mutated, induce upright leaves lacking of ligules and auricles, allowing for higher planting densities (Tian et al., 2011). Alternatively, four *DWARF* loci (*DW1*-*DW4*) largely control internode length – and therefore plant height – in sorghum in a pure Mendelian fashion (Hilley et al., 2016, Mullet, 2017).

The duration and timing of crop growth affects biomass yield by determining both the total amount of solar radiation that can be captured during crop growth, and the duration of vegetative growth prior to switching to the reproductive cycle, which typically terminates the synthesis of vegetative tissues (Fernandez et al., 2009, Jones et al., 2015). To maximize biomass yield, an ideal crop should thus display a long growing season, fully develop the canopy by the time solar radiation reaches its yearly maximum, and postpone or even avoid the setting of reproductive structures (**Figure 1**) (Jones et al., 2015). Achieving this ideotype implies breeding for an early leaf development, a delayed plant senescence, and a late flowering time. As a proof-of-concept, Dohleman and Long (2009) showed that the 4-weeks earlier canopy development and the 4-weeks longer canopy duration of miscanthus compared to maize largely explain the on average 60% higher biomass production of the former in US Midwest. The genetics underlying these developmental traits are complex, as a high number of genes showing pleiotropic and epistatic effects and interacting with environment are involved (see e.g. Wu et al. (2012), Gregersen et al. (2013), and Bluemel et al. (2015) for reviews). Nevertheless, some critical genes can be targeted in candidate gene or transgenic approaches. An example is the *IPT* gene from *Agrobacterium tumefaciens* that, once introgressed in several plants, boosts a rate-limiting step in the cytokinin biosynthesis and delays plant senescence by sustaining high cytokinin production along the whole growing season (Gregersen et al., 2013). Genes proven to evenly affect a critical trait across several species are undoubtedly important targets to advance novel crops without the need of *de novo* investigating the genetics underlying that trait in each species. However, they are generally rare for developmental traits whose genetic architecture relies on tens of small-effect loci (Buckler et al., 2009) and has been shaped by selective forces acting in a species- and environmental-specific manner during evolution, as discussed by Gifford et al. (2015) for flowering time in miscanthus. Therefore, conventional selection of superior genotypes in target MALs is nevertheless a time- and money-saving breeding approach for traits as earliness of leaf development, delayed plant senescence and late flowering time. In this regard, variation for all of them exists in several crops, as miscanthus (Farrell et al. (2006b), Jensen et al. (2011), Robson et al. (2012)), or switchgrass (Van Esbroeck et al., 1997, Bhandari et al., 2010). In addition, Van Esbroeck et al. (1997) and Bhandari et al. (2010) showed that the heritability of all these traits in switchgrass is moderate to high, promising success for selection. Finally, different ecotypic groups are often encountered in species that grow over wide geographical ranges (e.g. switchgrass Evans et al. (2017) or hemp Salentijn et al. (2015)), and represent an additional resource to enhance yield. For example, when short-day hemp cultivars are cultivated at northern latitudes native of long-day genotypes, flowering time is postponed and fiber yield increases (Salentijn et al., 2015).

Efficiency of light conversion into biomass

The efficiency of sunlight conversion into biomass is determined by the amount of biomass energy produced relative to the total sunlight energy intercepted by a crop over a specific period of time (Zhu et al., 2010), and depends directly on crop photosynthesis. Crops with C4 photosynthesis typically display higher conversion efficiencies than C3 species, owing to a CO_2-concentrating mechanism that prevents photorespiration by sustaining high CO_2 leaf concentrations for optimal Rubisco activity even at low atmospheric CO_2 levels (van der Weijde et al., 2013). Such higher photosynthetic efficiency translates into higher yield potentials of C4 relatively to C3 crops (van der Weijde et al., 2013), which make the former more suitable for cultivation on MALs. This is especially true considering that C4 plants generally outperform C3 species also in terms of nitrogen and water use efficiency, given the low levels of photosynthetic proteins in leaves and low stomatal conductance (Ghannoum et al., 2011).

Promising C4 perennials for biomass production on MALs include miscanthus, switchgrass, napiergrass or Indian grass. Moreover, research is investigating how to engineer C4 photosynthesis in C3 crops (Schuler et al., 2016, Zhu et al., 2010), which could hypothetically further enhance the biomass yield of already high-yielding C3 species suitable for MALs, as giant reed or tall wheatgrass. This appears a long-term goal though (Zhu et al., 2010), as the high complexity of C4 photosynthesis and the high number of genes involved in determining all its metabolic and physiological benefits did not allow so far to achieve satisfactory results (Schuler et al., 2016).

In addition, C4 plant species evolved as a result of convergent evolution in warm, tropical, climates (Schuler et al., 2016), and C3 plants can outperform C4 species for both biomass production and resource use efficiency at temperate latitudes (Carroll and Somerville, 2009, Karp and Shield, 2008). Where this condition applies, the intra- and inter-specific variation in photosynthetic capacity that has been nevertheless observed across several C3 species (Hikosaka, 2010, Lawson et al., 2012) represents a precious resource to improve the efficiency of light conversion in promising C3 crops for MALs. For example, Wullschleger (1993) assessed the photosynthetic capacity of 109 C3 species and found carboxylation rates in a range of 6-194 $\mu mol\ m^{-2}\ s^{-1}$. This variation is thought to arise mainly from different biochemical capacities of the photosynthetic machinery (e.g. antenna complex size or photosystem II photoprotection capacity), a variable degree of CO_2 diffusion in leaves, different rates of N supply to the photosynthetic systems, and changes in the activity of photosynthetic enzymes (Hikosaka, 2010, Zhu et al., 2010, Lawson et al., 2012). Moreover, it appears largely genetically controlled (Zhu et al., 2010, Flood et al., 2011), even if its genetic basis remains largely understudied (Flood et al., 2011, Lawson et al., 2012, Zhu et al., 2016). Recently, QTLs for photosynthetic efficiency have been identified in model species as arabidopsis (van Rooijen et al., 2015) or tomato (de Oliveira Silva et al., 2018). However, long time will be needed before critical genetic elements are identified underneath genomic regions associated with photosynthesis variation, and the validity of such genetic elements can be extended also to orphan crops for MALs. Nonetheless, this approach could lead to the discovery of key candidate genes to improve yield of biomass crops, and should therefore be explored in further detail.

Efficiency of carbon partitioning into vegetative tissues

The efficiency of C partitioning refers to the amount of fixed C invested in developing vegetative tissues over the total fixed C (Zhu et al., 2010). In biomass perennials, the seasonal production of vegetative tissues is initially fueled by the C stored into roots, until the new canopy develops sufficient photosynthetic capacity to take on this role. Therefore, two main crop characteristics are critical to maximize the biomass synthesis from a C partitioning perspective (Jones et al., 2015). Firstly, a preferential allocation of fixed C to the production of vegetative tissues over other C sinks. Secondly, a rooting system that develops its full size soon after crop establishment and capable of storing large C stocks, as displayed in **Figure 1**.

The preferential allocation of fixed C to the production of vegetative tissues can be improved through indirect selection for architectural and phenological traits correlated with biomass yield. In this sense, plant height and late flowering time are key, as they typically correlate with each other and with biomass yield, as shown in maize (Lübberstedt et al., 1997), sorghum (Murray et al., 2008, Ritter et al., 2008), switchgrass (Bhandari et al., 2010, Bhandari et al., 2011), or miscanthus (Gifford et al., 2015). Tall plants display a relaxed C demand by sinks different than vegetative tissues thanks to a postponed switch to reproductive growth, which translates into higher relative C investments in biomass production and larger yields. Importantly, these associations appear genetically-determined, as the co-localization of QTLs for plant height, flowering time, and biomass yield has also been observed, for example in sorghum (e.g. Lin et al. (1995), Murray et al. (2008), Ritter et al. (2008)). Alternatively, C partitioning can also be improved through transgenic approaches, by altering the expression of sucrose synthases or sucrose transporters (Yadav et al., 2015, Braun et al., 2013). For example, Poovaiah et al. (2015) considerably increased plant height (+37%), biomass yield (+13.6%), and tiller number (+79%) in switchgrass by overexpressing a constitutive Sucrose Synthase (*PvSUS1*) ubiquitously present in the plant. Despite the promising results, a major drawback of these approaches is that undesired side-effects on growth or physiological traits are often encountered, which calls for a deeper understanding of the genetics underlying C allocation into a plant system perspective (Yadav et al., 2015).

Improving roots and rhizomes as determinants of C allocation in biomass crops is challenging, as studying these traits is costly, time-consuming, and technically-demanding, especially in field conditions (Sartoni et al., 2015, Pierret et al., 2016). Thus, little is known on the genetics of root growth (Topp et al., 2013), especially in biomass crops. In this context, the studies on the genetic basis of rhizomatousness in sorghum using progeny from crosses of wild perennial and cultivated annual accessions (Paterson et al., 1995, Kong et al., 2015) appear particularly relevant. Following this strategy, Kong et al. (2015) mapped seven major QTLs for rhizomatousness, finding co-localizations with regions affecting tillering. The further investigation of causative genes underneath these QTLs could identify candidates to improve both rooting capacity and biomass yield components of sorghum and, possibly, other biomass perennials. In addition, novel phenotyping platforms [e.g. Nadezhdina et al. (2012), Topp et al. (2013), Sartoni et al. (2015)] are expected to reduce costs and destructiveness of root phenotyping, as well as grant assessment of root growth over multi-years trials, which is needed to understand root development in a C partitioning perspective. At the moment, the most feasible strategy to breed for fast-developing and large roots is by making use of known correlations between root properties and this trait ideotype. Specifically, above-ground biomass production in MPBCs positively correlates with rooting depth (Mueller et al., 2013), which in turn affects the total C sequestration capacity of roots and rhizomes (Jones et al., 2015). Therefore, root depth appears a promising target to breed for root systems capable of storing large C stocks and sustaining abundant yields.

3.2 Biomass quality

MPBCs produce large amounts of lignocellulosic biomass, which represents a highly attractive material for bio-based applications, as lignocellulose contains several classes of economically interesting compounds, including biopolymers and biochemicals (Trindade et al., 2010, Isikgor and Becer, 2015). The extraction of target molecules from lignocellulose is currently based on expensive and intensive post-harvest biomass treatments aimed at both loosening the structure and fractionating the components of plant cell walls, which are by far the major constituents of lignocellulose and contain the major part of attractive compounds (van der Weijde et al., 2013, Lauria et al., 2015, Isikgor and Becer, 2015, Wang et al., 2016). For this reason, the biomass quality of MPBCs entails primarily the recalcitrance of plant cell walls to deconstruction, as feedstocks with easily-destructible cell walls require milder and cheaper treatments to be processed into bio-based end-products (van der Weijde et al., 2013, Isikgor and Becer, 2015, Wang et al., 2016). Moreover, the relative content of molecules of interest within cell walls in relation to the end-use of the biomass is also a preeminent quality target in order to develop crop varieties tailored to specific bio-based production chains (Trindade et al., 2010). In this regard, **Table 2** reports a list of cell wall ideotypes that can fit the needs of different possible end-uses of the lignocellulosic biomass obtainable from MPBCs grown on MALs.

The general structure of cell walls is conserved across plants, consisting of cellulose fibers in a matrix of non-cellulosic polysaccharides (mainly hemicelluloses), lignin, structural proteins and mineral elements (Cosgrove, 2005, Vogel, 2008). However, the relative abundance, composition, and structure of cell wall components vary extensively across species, tissues and developmental stages (Vogel, 2008, Sarkar et al., 2009, Loque et al., 2015). Likewise, the occurrence and functionality of major cell wall synthetic genes is also conserved across plants (e.g. *CESAs* and *CSLs* as main players of cellulose and hemicellulose biosynthesis, respectively) (Vogel, 2008, Xu et al., 2009, Zhang et al., 2018), even if inter-specific differences exist also at the genetic level, which affect cell wall composition (e.g. the presence of *CSL-Fs* and *CSL-Hs* only in certain plant clades, which synthesize mixed-linkage glucans) (Vogel, 2008, Doblin et al., 2009). Taken together, these observations point out a large margin for the improvement of cell wall composition towards low-recalcitrant and purpose-made cell wall ideotypes. However, the extreme complexity of cell wall biosynthesis and regulation [~4000 genes are thought to be involved in arabidopsis (Wang et al., 2012a)] hampers breeding efforts, especially in novel crops lacking genetic tools. In this context, dissecting the trait “cell wall quality” into its main determinants (content, composition, and structure of lignin and cell wall polysaccharides) can help to better define the goals targetable by breeding, and the available biochemical and genetic knowledge to achieve them.

Lignin

Lignin is the cell wall component that mostly limits lignocellulose deconstruction (Li et al., 2016b). On the one hand, it cross-links with hemicelluloses forming a physical barrier that hides polysaccharides to degrading enzymes (Zhao et al., 2012, Li et al., 2016b). On the other hand, it irreversibly absorbs hydrolytic enzymes, inhibiting their activity (Zhao et al., 2012, Li et al., 2016b). To decrease biomass recalcitrance, an immediate strategy is to decrease the lignin content in cell walls (van der Weijde et al., 2013). However, as lignin provides mechanical support, stress response, and pathogen resistance to plants, decreased lignin contents can hamper plant stability and growth, and ultimately reduce biomass yield (Wang et al., 2016). Moreover, lignin itself represents an economically attractive compound, as it is a source of aromatic molecules that can find applications in the production of phenolics, carbon fibers, dispersants, and bio-plastics (Isikgor and Becer, 2015). Therefore, strategies aimed at modifying the lignin properties that affect biomass digestibility without decreasing the total lignin content – such as altering ratios of lignin subunits, relocating lignin deposition, and modifying lignin backbone and linkages with polysaccharides – also represent valid breeding approaches to improve biomass quality (Li et al., 2016b, Verma and Dwivedi, 2014, da Costa et al., 2019).

The lignin biosynthetic pathway is well characterized (Boerjan et al., 2003, Bonawitz and Chapple, 2010) and also highly conserved across vascular plants (Boerjan et al., 2003, Loque et al., 2015). These characteristics make candidate gene approaches particularly suitable to modify lignin properties, by identifying critical target genes within the lignin pathway, testing their effectiveness in model crops, and transferring successful approaches to other less-studied species, as orphan biomass perennials. For example, the relevance of down-regulating the Caffeoyl O-Methyltransferases (*COMTs*) to decrease lignin content and improve biomass digestibility for biofuel production has been firstly shown in model forage crops as maize, tall fescue, and alfalfa (Hisano et al., 2011). These results have pushed the successful reproduction of this approach in switchgrass (Fu et al., 2011), while gene cloning and *in silico* protein alignments led to the identification of *COMT* genes in miscanthus (Dwiyanti et al., 2014) and eucalyptus (Carocha et al., 2015), which can become future targets of genetic modification to improve biomass quality. According to this approach, the improvement of lignin properties in novel lignocellulosic perennials for MALs can benefit from the long list of successful modifications of lignin-related genes in model plants and staple biomass crops (see Liu et al. (2018b) for a recent overview). Overall, these studies also highlight four important principles for lignin modification. First, lignin content is decreased more effectively when genes acting early in lignin biosynthesis are targeted (Chen and Dixon, 2007). Second, down-regulation of genes is more effective than knock-outs to minimize side-effects on plant growth (Xie and Peng, 2011). Third, pathway cross-talks and gene redundancy need to be carefully considered to exclude mechanisms that can limit the gains from targeted approaches (Torres et al., 2015b). Fourth, to reproduce successful transgenic approaches in novel biomass crops, the availability of transformation and regeneration protocols is critical (Clifton-Brown et al., 2018).

Besides genetic modification, large natural variation in lignin content and composition exists within biomass perennials promising for MALs, as miscanthus (Zhao et al., 2014), switchgrass (Yan et al., 2010), or willow and poplar (Kenney et al., 1990). Moreover, studies investigating the sources of such variability showed that it is typically highly heritable (Torres et al., 2015a, van der Weijde et al., 2017b), and therefore constitutes an important breeding resource. Alternatively, intra-specific crosses between accessions showing contrasting lignin profiles have been also successfully used to map lignin-related QTLs across a range of species (van der Weijde et al., 2017b, Thumma et al., 2010, Torres et al., 2015a), which open prospects for marker-assisted selection (MAS) of genotypes showing superior lignin profiles and high biomass degradability.

Cell wall polysaccharides

Cellulose and hemicelluloses represent attractive polymers for bio-based applications, as they constitute the bulk of energy contained into lignocellulosic biomass and can be used as platforms to produce several classes of valuable biochemicals (van der Weijde et al., 2013, Isikgor and Becer, 2015). Therefore, increasing the cellulose and hemicellulose content in cell walls and modifying their molecular properties that promote recalcitrance are also effective targets to improve biomass quality (Torres et al., 2015b). Cellulose recalcitrance depends on the degree of cellulose crystallinity and polymerization (Wang et al., 2016), and reducing these two parameters is critical to improve lignocellulose degradability (van der Weijde et al., 2013, Torres et al., 2015b, Allwright and Taylor, 2016). Conversely, hemicelluloses affect recalcitrance through their total content in cell walls and their degree of branching (Torres et al., 2015b, Wang et al., 2016). Specifically, as hemicelluloses cross-link cellulose and lignin, high hemicellulose content reduces cellulose crystallinity (Wang et al., 2016). At the same time, low xylan branching ensures an easy separation of hemicelluloses, cellulose and lignin during saccharification (Torres et al., 2015b).

The molecular alteration of cell wall polysaccharides is challenged by our limited knowledge on cellulose and hemicellulose biosynthesis and regulation. Research in this field is still at the level of functional genetics in model species as arabidopsis, while attempts of candidate gene studies in biomass crops remain few, mainly restricted to poplar and – to a less extent – maize (Wang et al., 2016). Several studies assessed the effect of modifying cellulose synthases (*CESAs*) on cellulose properties. For example, Joshi et al. (2011) decreased cellulose crystallinity in poplar wood by overexpressing a constitutive *CESA* (*PtdCESA8*). Alternatively, Harris et al. (2009) mutated the *AtCESA3*, which led to a 34% reduction of cellulose crystallinity and a 151% increase of biomass degradability. Besides *CESAs*, numerous other genes affect the molecular properties of cellulose and hemicelluloses. Glass et al. (2015) reduced cellulose crystallinity and increased biomass yield in arabidopsis by down-regulating a class C endoglucanase (*AtGH9C2*). Furthermore, Mortimer et al. (2010) mutated the two glycosyltransferases-coding loci *GUX1* and *GUX2* in arabidopsis, achieving substitutions-free hemicelluloses and easily extractable xylans. Overall, these studies highlight promising targets to improve cell wall polysaccharides, for which homologs could be searched in biomass crops and considered for genetic modification or screening of allelic diversity for conventional breeding programs (Torres et al., 2015b). However, in order to better predict/avoid the (negative) pleiotropic side-effects that are often encountered in these studies, it is critical to develop a better, systemic, understanding of the genetic complexity of cellulose and hemicellulose biosynthesis and regulation, and their interplay with other plant metabolic pathways. For example, along with the decreased cellulose crystallinity, the poplar transgenic lines produced by Joshi et al. (2011) displayed also a decreased cellulose content, increased lignin, and stunt growth, which are all unwanted traits for elite lines of perennial biomass crops.

As for lignin, variability in the composition and structure of cell wall polysaccharides across species constitutes an important breeding resource. In this regard, Harris and DeBolt (2008) reported large variation in cellulose crystallinity across a set of 35 plant species, while Vandenbrink et al. (2012) and Porth et al. (2013) found considerable intra-specific variation in sorghum and poplar association panels, respectively. Alternatively, Torres et al. (2013) found variable hemicellulose composition and substitution patterns in a maize doubled-haploid population, which correlated with differences in biomass digestibility. To conclude, when molecular markers are available or can be easily developed, variability can also be used for association mapping, leading to the identification of genomic regions associated with high biomass quality, as performed by Torres et al. (2015a) in maize or van der Weijde et al. (2017b) in miscanthus.

3.3 Biomass yield and quality stability under abiotic stresses

Abiotic stresses as water surplus or deficit, extreme temperatures, and soil salinity are common constraints to agriculture on MALs, and are expected to intensify as a result of climate change (Quinn et al., 2015, Jones et al., 2015). These stresses discourage the cultivation of MALs, as they can hamper crop growth, reduce biomass yield and quality, and ultimately hinder a stable biomass supply, which would volatilize prices of raw biomass and bio-based commodities. Developing genotypes that show stability of biomass yield and quality under adverse abiotic conditions is thus pivotal to successfully allocate worldwide MALs to biomass production. Specifically, since abiotic stresses often occur in combination (e.g. heat and drought) or succession (e.g. waterlogging followed by drought) (Mickelbart et al., 2015), resistant varieties should possibly combine different sources of resistance to withstand multiple stresses at once.

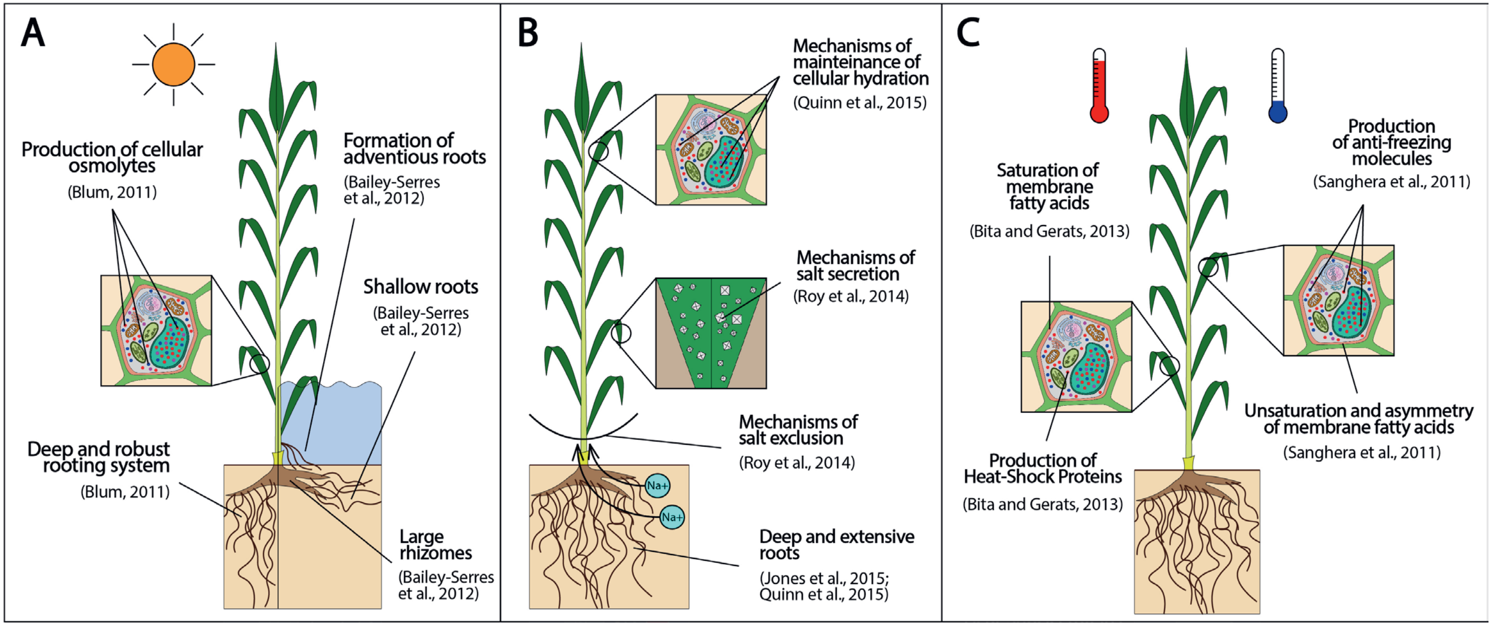

A
Production of cellular osmolytes
(Blum, 2011)
Deep and robust rooting system
(Blum, 2011)
Formation of adventious roots
(Bailey-Serres et al., 2012)
Shallow roots
(Bailey-Serres et al., 2012)
Large rhizomes
(Bailey-Serres et al., 2012)
B
Mechanisms of mainteinance of cellular hydration
(Quinn et al., 2015)
Mechanisms of salt secretion
(Roy et al., 2014)
Mechanisms of salt exclusion
(Roy et al., 2014)
Na+
Na+
Deep and extensive roots
(Jones et al., 2015; Quinn et al., 2015)
C
Saturation of membrane fatty acids
(Bita and Gerats, 2013)
Production of Heat-Shock Proteins
(Bita and Gerats, 2013)
Production of anti-freezing molecules
(Sanghera et al., 2011)
Unsaturation and asymmetry of membrane fatty acids
(Sanghera et al., 2011)

Figure 2. Preeminent target traits to equip perennial lignocellulosic biomass crops with effective resistance to common abiotic stresses of MALs: drought (panel A, left), flooding (panel A, right), salinity (panel B), and extremely warm (panel C, left) or cold (panel C, right) temperatures. The traits reported represent an ideotype to guide breeding activities, and the effective magnitudes of improvement attainable with respect to each trait can vary extensively depending on the species object of a breeding program.

Suboptimal water availability

Water shortage inhibits cellular expansion, hydration and photosynthesis, with negative impacts on plant germination, establishment, growth, nutrient assimilation and transport (Quinn et al., 2015), and ultimately biomass yield (Emerson et al., 2014). Flooding determines instead stomata closure and uptake of toxic compounds released by anaerobic microorganisms in anoxic soils, which inhibit nutrient transport and photosynthesis, damaging plant growth and yield (Quinn et al., 2015). Drought and waterlogging are major causes of yearly yield losses worldwide and are acquiring more-than-ever importance as a consequence of extreme atmospheric events in a changing climate (Jones et al., 2015). Moreover, common MALs characteristics as steep slopes, high erodibility, or poor drainage amplify the occurrence and effects of these stresses (Quinn et al., 2015), calling for crops able to maintain normal metabolism, growth, and yield under drought and flooding.

Several plant traits can be targeted to develop drought- and flood- tolerant crops, as visually summarized in **Figure 2A**. For drought, deep and robust roots, able to penetrate harsh soils, are important characters to reach deep water in dry areas (Blum, 2011). Moreover, the ability of accumulating cellular osmolites to avoid dehydration (osmotic adjustment) is also important, particularly at the seedling stage when roots are still underdeveloped (Blum, 2011), and can be improved by selecting against leaf rolling and in favor of green canopy under drought, which are two easily-scorable traits (Amelework et al., 2015). For flood tolerance, a shallow root system, with a thick root epidermis, well-developed aerenchymatous tissues, and adventitious roots are critical traits as they facilitate aeration (Bailey-Serres et al., 2012). Moreover, large rhizomes are also favorable to provide starch to sustain optimal growth under prolonged flooding (Bailey-Serres et al., 2012).

Natural variation in tolerance to suboptimal water conditions is an important breeding resource, especially considering the high polygenicity of this trait (Mickelbart et al., 2015), and the difficulty of phenotyping roots in field conditions (see 3.1.3). Different water use efficiencies (WUEs) between species are a first important variable trait, given the direct correlation between WUE and drought tolerance (Allwright and Taylor, 2016). In this regard, as previously discussed (see 3.1.2), C4 crops display overall higher WUEs than C3 species, and are thus best options on dry MALs. In addition, intra-specific variation in the capacity of undergoing drought or floods without substantial yield penalties has been observed across several biomass perennials. For example, Barney et al. (2009) showed that upland and lowland switchgrass ecotypes display significantly different levels of drought tolerance. This observation is relevant, as it could enable to simultaneously improve drought tolerance and other traits displaying analogous latitudinal variation, such as flowering time, biomass yield, and cold resistance (Clifton-Brown et al., 2018). In miscanthus, Van der Weijde et al. (2017a) have found large variation in the capacity of 50 accessions to produce biomass under drought, as their drought-constrained yields amounted to 30-110% of the yields under control conditions. The study assessed also biomass degradability under drought, demonstrating that it increased across all the 50 accessions, irrespectively of the level of drought tolerance, probably through an increase of the hemicellulose/cellulose ratio in cell walls (Van der Weijde et al., 2017a). This is a remarkable finding, as it potentially allows to develop genotypes that are not only drought-tolerant, but also capable of turning the drought burden into an advantage for biomass quality. Finally, in black locust, Zhang (2010) observed that tetraploid accessions withstand better drought than diploid genotypes, an information that can be useful at the moment of choosing superior lines to be used as parents in breeding programs (even if polyploidy can complicate breeding).

A better understanding of the genetics underlying drought and flood tolerance paves also the way to improve these traits through candidate gene approaches. Transcription factors (TFs) related to hormone metabolism are consistently accumulated under both drought and flooding across a range of plants (Mickelbart et al., 2015, Yin et al., 2014), and appear thus relevant to mine candidate genes. For example, Yang et al. (2015) and Zhao et al. (2016) enhanced drought, cold, and salt tolerance in arabidopsis by ectopically overexpressing two NAC TFs from *Miscanthus lutarioriparius* (*MlNAC5*, *MlNAC9*) involved in the ABA-dependent signalling pathway. These results highlight the relevance of these genes to equip miscanthus (and possibly other perennial grasses) with tolerance to multiple abiotic stresses, and their validity should be now investigated in target biomass crops. Alternatively, ERF-VIIs are TFs upregulated upon flooding across several plants (Mickelbart et al., 2015). They include the rice *SUB1A* gene, which provides flood tolerance to rice by quitting growth upon flooding, limiting resource dispersal (Xu et al., 2006). Manipulating the expression of ERF-VII TFs could thus be key to provide flood tolerance (Mickelbart et al., 2015), and deserves deeper investigation, especially in biomass crops. In fact, even if proper flood tolerance has not yet been improved through this approach (Peña-Castro et al., 2011), the expression of *OsSUB1A* in arabidopsis not only led to flowering inhibition (mimicking the quiescence survival strategy observed in rice) (Peña-Castro et al., 2011), but also improved the saccharification efficiency of the biomass (Nunez-Lopez et al., 2015). Therefore, possibilities to improve both flood tolerance and biomass quality by manipulating ERF-VIIs could be furtherly explored. More in general, the interplay between cell wall quality and tolerance to abiotic stresses could be key to breed biomass crops, as cell wall composition and expression of cell wall genes are deeply changed under abiotic stresses, including drought and flooding (Houston et al., 2016, Le Gall et al., 2015). Unravelling genes acting upstream of both cell wall changes and induction of mechanisms of abiotic stress tolerance could thus represent a promising strategy to breed for both these traits in parallel.

Extreme temperatures

Freezing (<–1°C), chilling (0-18°C), or high (>35-45°C, depending on species sensitivity) temperatures can physiologically and physically damage plants, penalizing biomass yield (Quinn et al., 2015). Cold temperatures reduce germination rates, growth, and tillering (Yadav, 2010), which delay or hamper biomass production (Jones et al., 2015). Moreover, cell wall composition is also affected by frost, probably with negative effects on biomass recalcitrance, as lignin content is typically increased (Le Gall et al., 2015). Heat stress shortens the growing season by accelerating feedstock maturation and inhibits photosynthesis, causing sugar catabolism (Quinn et al., 2015). Thus, both biomass yield (Bita and Gerats, 2013) and quality (Ananda et al., 2011) are penalized.

Tolerance to extreme temperatures starts by cultivating crops adapted to the temperature ranges of a target area. In fact, while temperate species – like many C3 crops (Jones et al., 2015) – are capable of cold acclimation to withstand chilling temperatures without impacts on crop performance, tropical and subtropical plants typically lack of such mechanisms, showing sensitivity even to chilling conditions (Sanghera et al., 2011). Conversely, crops from tropical and subtropical regions – such as the majority of C4 species (Schuler et al., 2016) – withstand better heat stress (van der Weijde et al., 2013), and could thus be preferentially planted in warm environments, where could eventually display combined tolerance to heat and drought. Once crops are correctly allocated, plant breeding can enhance freezing and heat tolerance to species already adapted to cool or warm environments, respectively. To this aim, high unsaturation and asymmetry in membrane lipids (which lower the membrane freezing point), and production of protective molecules (antioxidants, chaperones, deposition of apoplastic sugar) are important traits to prevent cellular damages from freezing (**Figure 2C**, right) (Sanghera et al., 2011). Conversely, heat tolerance is enhanced by production of heat-shock proteins (HSPs), secondary metabolites to protect against oxidative damage, and membrane saturated fatty acids to increase membranes melting point (**Figure 2C**, left) (Bita and Gerats, 2013).

Because of the complex molecular and genetic nature of the traits just mentioned, selection of tolerant accessions within pools showing variability for cold and heat resistance represents an effective breeding strategy (Jones et al., 2015, Bita and Gerats, 2013). In this respect, measuring electrolyte leakage following freezing is a rapid screening method for frost tolerance (Jones et al., 2015). Vice versa, assessing morphological traits that reduce heat load [e.g. pubescent, vertically-oriented, or light-colored leaves (Quinn et al., 2015)], as well as physiological characters as maintenance of photosynthesis, chlorophyll content, and stomatal conductance under heat stress (Bita and Gerats, 2013) allows quick screening of heat tolerance. Furthermore, variability for both cold and heat tolerance has been observed for several biomass perennials (Quinn et al., 2015). For example, the level of frost tolerance is variable across diverse miscanthus genotypes (Clifton-Brown and Lewandowski, 2000, Farrell et al., 2006a), and cold-tolerance could be coupled with breeding for earlier germination, which is also variable in miscanthus (Clifton-Brown and Jones, 1997) and is a target trait to increase biomass yield (see 3.1.1).

Besides selection within intra-specific variation, sources of both heat and cold tolerance can be also introgressed through crosses with wild relatives evolved under stressed environments, or by transgenic modification. The first approach has been successfully applied in wheat (Mohammed et al., 2014), and could be used in biomass crops in which different species show different levels of tolerance to extreme temperatures, as poplar [*Populus euphratica* is tolerant to both heat and drought, and candidate genes and mechanisms underneath tolerance have been identified (Ferreira et al., 2006, Tang et al., 2013)] or sugarcane [the wild species *Saccharum spontaneum* shows higher cold tolerance than the most cold tolerant commercial varieties of sugarcane (Hale et al., 2014)]. The second approach has instead been successfully applied in eucalyptus, where the transgenic introgression of *CBF* genes [a class of TFs that promote cold acclimation across a range of plants (Mickelbart et al., 2015)] has greatly improved cold tolerance (Hinchee et al., 2011). As several genes involved in cold and heat tolerance are known across plants [see Sanghera et al. (2011) for a review], this approach could be reproduced also with other species and other targets.

Salinity

Soil salinity hampers plant growth as it reduces osmotic potential – which challenges water uptake and solute movement – and causes ion cytotoxicity as salt ions compete with potassium to occupy enzymatic active sites (Quinn et al., 2015, Mickelbart et al., 2015). In biomass crops, salinity consistently decreases biomass yield (Quinn et al., 2015) and remodels cell wall composition, which is thought to negatively affect biomass quality, as lignin content is typically increased (Le Gall et al., 2015).

Not all biomass crops are salt sensitive, and the allocation of tolerant species to salty MALs is a preeminent strategy to avoid salt-induced penalties to biomass production (Jones et al., 2015). In this regard, some herbaceous biomass perennials as giant reed or *Pennisetum purpureum* display striking levels of salt tolerance, comparable with halophytes (Quinn et al., 2015). Furthermore, biomass trees as *Eucalyptus camaldulensis*, several poplar and willow hybrids, and tetraploid *Robinia pseudacacia* accessions also show high tolerance to salinity (Quinn et al., 2015). Globally, all these species represents good options for salty MALs.

Crops can also be equipped with mechanisms of salt tolerance through plant breeding. Since salty and dry soils are functionally similar, plant traits benefitting drought tolerance are also effective to improve salt tolerance (Quinn et al., 2015). Specifically, deep and large root systems and the capacity of maintaining cellular hydration appear particularly critical characters, as shown in **Figure 2B**. These traits allow plants to search for water and nutrients in deep soil layers less affected by salinity, and to keep water-to-salt ratios in cells at acceptable levels (Quinn et al., 2015, Jones et al., 2015). In addition, mechanisms of salt exclusion from shoots, and secretion of excess salt uptake are other effective characters to provide salt tolerance (**Figure 2B**) (Roy et al., 2014), and can be effectively phenotyped by screening ion levels in plant tissues (i.e. total Na^+ and Cl^- contents and K^+/Na^+ ratio) (Chen, 2018). For example, Chen (2018) has successfully applied this method to screen levels of salt tolerance across 70 miscanthus genotypes, revealing a relatively high degree of genetic diversity in the ability of withstanding salinity among the accessions. The study has specifically showed the presence of genotypes capable of excluding salt ions from shoots, preventing leaf senescence, and sustaining biomass production on salty soils (Chen, 2018). These genotypes could be used as parents in future miscanthus breeding programs (Chen, 2018). Besides miscanthus, genetic variability for biomass production under salinity has been observed also in other biomass perennials, as switchgrass (Liu et al., 2014b), black locust (Wang et al., 2013c), or poplar (Sixto et al., 2005), and could be used for breeding. However, the assessment of the capacity of maintaining biomass quality under salinity – in addition to biomass yield – is often skipped by studies investigating salt tolerance. This is a major shortcoming that can lead to negative selection biases (i.e. selection of salt tolerant genotypes that eventually turn out to be highly recalcitrant to biomass processing), as studies in miscanthus have shown that best-performing accessions for salt tolerance do not coincide with genotypes showing highest biomass quality under salt stress (Chen, 2018).

4 Tools and strategies to advance promising lignocellulosic perennial crops for MALs

4.1 Available breeding tools

Until now, only a limited set of lignocellulosic biomass crops promising for MALs have undergone breeding (Allwright and Taylor, 2016), while the majority lies in a state close to (selected) wild germplasm (see **Table 1**). This small set of species includes miscanthus, switchgrass, willow, poplar, and eucalyptus, whose improvement history dates back to the second half of the twentieth century (Grattapaglia and Kirst, 2008, Allwright and Taylor, 2016, Clifton-Brown et al., 2018). All these species are outcrossing, and until early 2000s their genetic improvement has entirely relied on phenotypic selection within breeding populations created through recurrent inter-mating of wild and advanced germplasm, or intra- and inter-specific hybridization (Clifton-Brown et al., 2018, Grattapaglia and Kirst, 2008). These methods allow to combine favorable traits from distant genetic pools into elite lines and to exploit heterosis (given that the genetic structure of target populations and heterotic compatibility are analyzed prior to design breeding schemes) (Acquaah, 2012). However, the release of commercial varieties takes long time (i.e. 11-26 years), which significantly delays the adoption into real agricultural contexts of the improvements achieved in breeding programs (Clifton-Brown et al., 2018). This is due to the long period (sometimes several years) that perennial crops often require to phenotypically express traits of interest, the long breeding cycles (especially in biomass trees), and the prolonged time to commercially upscale elite lines of highly heterozygous outcrossing species (Clifton-Brown et al., 2018).

Over the last decades, research investments have permitted to develop advanced tools to assist the improvement of these crops at all the breeding levels (Clifton-Brown et al., 2018). The first tools created were genetic maps which, together with the use of mapping populations segregating for a target character that is phenotypically divergent between the founder individuals, have enabled to map quantitative trait loci (QTLs) for growth, quality, and agronomic traits (see **Table 1** for relevant references). Initially, QTL detection has exclusively relied on bi-parental crosses, but more recently multi-parental approaches to mine polymorphisms linked to loci of interest – which allow to enhance the allelic diversity and variety of genetic backgrounds included in a study – have begun to be applied also in biomass crops (Mandrou et al., 2014). The availability of markers and QTLs has in turn enabled marker-assisted selection (MAS) (Clifton-Brown et al., 2018). MAS encompasses genotyping breeding material for the alleles harbored at marker loci associated with QTLs of interest, and selecting accessions carrying positive alleles at those marker loci. Therefore, selection can take place already at early developmental stages, even before a trait is phenotypically expressed, which can significantly accelerate breeding gains. For example, Thavamanikumar et al. (2018) applied MAS to improve wood density, pulp yield, and total plant growth in eucalyptus, showing that it can reduce by 50% the duration of breeding cycles (from 10-15 to 5-7 years), while breeding gains can be achieved at a 2-3 fold higher rate than by conventional selection.

The drop of genotyping costs brought by genotyping-by-sequencing (GBS) technologies, coupled with the release of the genome sequences of all the species referred in here [Tuskan et al. (2006) (poplar), Myburg et al. (2014) (eucalyptus), Dai et al. (2014) (willow), http://phytozome.jgi.doe.gov/ (miscanthus, switchgrass)], has permitted the development of dense arrays of single nucleotide polymorphisms

(SNPs) covering the whole genome, which have in turn paved the way to genome-wide association studies (GWAS) and genomic selection (GS) schemes (Allwright and Taylor, 2016, Clifton-Brown et al., 2018). GWAS is a powerful tool to detect marker-traits associations using genotypic collections inclusive of long recombination histories, which promises to achieve deep mapping resolutions, and save the time needed to set up experimental populations for QTL mapping (Bernardo, 2016). Because of their high genetic diversity, undomesticated status, and generally fast linkage disequilibrium (LD) decay, lignocellulosic crops – especially biomass trees – appear ideal for GWAS (Du et al., 2018), and several studies have thus used this approach to reveal loci underlying biomass-related traits. For example, GWAS has been used in poplar to detect several marker-trait associations for quality characters as lignin content and composition (Porth et al., 2013), as well as for phenology traits as canopy duration or flowering date (McKown et al., 2014). Despite GWAS promises, Fahrenkrog et al. (2017) have pointed out how rare allele variants, whose detection can be quite often missed by GWAS analyses (Bernardo, 2016), can be particularly relevant to explain genetic variation for bioenergy traits as cell wall composition. Therefore, good experimental designs (e.g. adequate sample size and geographical sampling of accessions to give a balanced representation of the variability for a trait of interest in the panel used (Brachi et al., 2011)) are pivotal to successfully perform a GWAS (Du et al., 2018). As for QTL mapping, GWAS results can be directly used for MAS (Allwright and Taylor, 2016). However, genome-wide marker allelic effects from GWAS analyses can also be used to calculate breeding values for every individual plant in a breeding population, which is the concept standing behind genomic selection (GS) (Heffner et al., 2009). GS is particularly suited for crops showing large phenotypic and genetic variability, as miscanthus (Allwright and Taylor, 2016), where the feasibility of applying this strategy has started to be explored (Davey et al., 2017). To conclude, marker arrays are also useful to screen the genetic diversity of novel germplasm collections, which is a common need of pre-breeding research in orphan crops (Clifton-Brown et al., 2018). Diversity screenings can be informative to establish the geographical origin and the relatedness with other plant material, which are important information to take breeding decisions (Narasimhamoorthy et al., 2008, Lu et al., 2013).

Research on miscanthus, switchgrass, poplar, willow and eucalyptus has also aimed at developing transformation protocols to insert genes underlying traits hardly found in extant accessions. Clifton-Brown et al. (2018) and Kendurkar and Rangaswamy (2018) have recently reviewed the progress achieved in this field across these five crops, highlighting that stable protocols are available for switchgrass, miscanthus, and poplar, while willow and eucalyptus can display recalcitrance to transformation. In addition, several studies demonstrated the efficacy of genetic modification to improve traits for which critical candidate genes are known, as discussed in section 3. Overall, the public acceptance of genetic modification for biomass crops grown for bio-based applications could be higher than that for food crops (van der Weijde et al., 2013). However, when transgenic lines of outcrossing species were effectively cultivated, measures of gene confinement should be designed, as (trans)gene flow to relative wild species could be an issue (Clifton-Brown et al., 2018).

To conclude, fast and cost-effective phenotyping is also an asset to improve understudied crops, for which screening large germplasm collections is fundamental to evaluate variability for breeding programs (Clifton-Brown et al., 2018). Recent advances in high-throughput phenotyping open promising prospects in this regard. Fernandez et al. (2017) developed a robotic workstation that can be used to phenotype yield-related traits in tall herbaceous biomass crops as sorghum. The system has been successfully used to phenotype stem diameter and plant height in a GWAS sorghum panel, and the data collected allowed the detection of known QTLs for these traits, demonstrating the efficacy of this platform (Fernandez et al., 2017). Near-infrared spectroscopy (NIRS) technologies offer instead a viable option for high-throughput phenotyping of cell wall compositional traits, and robust protocols for their application have been recently developed and successfully used to phenotype a mapping population of miscanthus (van der Weijde et al., 2017b). Finally, thermal aerial imaging constitutes a high-throughput option to screen abiotic stress tolerance, and Ludovisi et al. (2017) have reported its successful application to phenotype the drought response in a large black poplar population consisting of 4603 individuals (503 genotypes). These examples clearly highlight how novel phenotyping technologies can widen the scale and enhance the speed of breeding programs, and need to be considered when improving novel biomass crops.

4.2 Prospects for the improvement of orphan lignocellulosic biomass species

The tools above are effective for crop improvement, but they are available just for a handful of lignocellulosic species, and their *de novo* development for orphan crops would require time and adequate research investments [even if the drop of sequencing costs will soon allow association mapping and whole-genome sequencing also to novel crops (Allwright and Taylor, 2016)]. Conversely, we have seen that classical breeding alone is also time-consuming and not very effective. In this scenario, genetic tools to transfer genetic knowledge from model species to less-studied crops and to meaningfully coalesce genetic information on relevant traits across crops can be key to bridge the gap between advanced and orphan biomass species.

Translational genomics offer a possibility in this perspective, as it allows the identification of candidate genes underlying a trait of interest in a “target” organism based on its homologue(s) in a model species (Salentijn et al., 2007). Such candidate genes can then be targeted through genetic modification to obtain a desired phenotype (see section 3 for examples). This approach allowed the identification and modification of several of the candidate genes discussed in section 3, and is particularly powerful for plant clades that share high levels of genome synteny between members and include model bioenergy crops (van der Weijde et al., 2013), as grasses (Bennetzen and Freeling, 1997, Carpita and McCann, 2008). To facilitate translational genomics, several tools have been developed over the years, in the form of both genomic databases [e.g. PlantGDB (www.plantgdb.org) (Dong et al., 2004), Plaza 4.0 (https://bioinformatics.psb.ugent.be/plaza/) (Van Bel et al., 2017), or other grass-specific databases reviewed by van der Weijde et al. (2013)] and platforms specifically designed for orphan crops lacking of a sequenced genome, but for which transcriptomes can be developed [e.g. Orphan Crops Browser (http://www.bioinformatics.nl/denovobrowser/db/species/index) (Kamei et al., 2016)].

Most of the traits discussed in section 3 are highly quantitative, and the available knowledge on the genetics underlying them in model species is typically in the form of QTLs with no validated or known candidate genes (Barrière et al., 2007). In these cases, tools to meaningfully coalesce the information on relevant QTLs between species and make it inter-applicable in a way immediately usable in MAS or GS contexts would be very useful. Combining meta-QTL analysis approaches (Goffinet and Gerber, 2000) extended beyond species boundaries with the development of "universal markers" that are present across species but can assay intra-specific diversity for traits of interest (Ranade and Yadav, 2014) could offer promising possibilities in this direction. Such universal markers could effectively allow to project known QTLs to breeding material not included in the original panels used for QTL mapping, or even to possibly other (orphan) species, on which MAS based on universal markers could take place, without the need of *de novo* producing species-specific knowledge. The extensive occurrence of common genetic factors underlying complex biomass-related traits across evolutionary distant plant species (as exemplified for cell wall recalcitrance in section 3) promises success from the application of the approach just described. However, research is needed to define to which extent common genetic determinants of traits of interest show positional conservation of their genomic organization to allow inter-species projection of QTLs and universal markers. In this direction, novel high-throughput tools to assess overall syntenic relationships between genetic elements underlying critical traits across large sets of plant genomes not even always displaying high levels of collinearity can offer promising prospects (Zhao and Schranz, 2017, Zhao et al., 2017).

Universal markers as just defined would represent a very useful tool to overcome the condition of orphan crops in which several promising species for marginal lands lay. However, these tools – as well as all the other genomic, molecular, and biotechnological approaches discussed in this review – do not represent *per se* a “finish line” in breeding novel promising perennial crops for marginal lands. Their effectiveness will ultimately depend by the specific ways in which breeders will integrate these tools in well-planned and modern “knowledge-based” breeding programs. Specifically, such programs will continue to largely rely on pre-breeding activities (i.e. germplasm development, dissection of the genetics underlying the traits discussed in section 3, and development of markers), conventional crossing of promising accessions and selection within progenies, as well as ongoing schemes of population improvement through recurrent selection (especially in open-pollinated species) (Clifton-Brown et al., 2018). Nevertheless, the tools discussed in this review will allow to speed up major steps of such programs (from the genetic characterization of breeding material and the dissection of the genetic determinants of target traits, to phenotyping, marker development, or the targeting of critical genes by genetic modification), as well as precisely guide breeding activities in crops that so far have been poorly studied. This aspects will ultimately be key to ensure that novel lignocellulosic perennials will be advanced at a sufficient level for commercialization in a reasonable time, which is currently the major priority for using MALs to provide biomass for a bio-based economy.

5 Concluding remarks

MALs have great potential to sustainably supply a large proportion of the biomass needed to fuel a global bio-based economy. However, the lack of crop varieties that can couple sustainability of biomass production with optimal biomass yield and quality to ensure profitable cultivation of MALs and cost-effective biomass conversion into bio-based commodities currently impedes to realize this vision. We firmly believe that plant breeding will be key to break through this impasse, and in this article we have dissected the problem of biomass provision using MALs from a plant breeding perspective. What emerges is that great progress has been made over the years in understanding the genetics underlying biomass traits. Moreover, the development of tools to study these aspects on larger scales and through quicker approaches will expand this knowledge in the future. Progress is however uneven among crops. While a few model species can count on an array of breeding tools and genetic knowledge to support their improvement but are unsuitable for sustainable cultivation on MALs, a wide range of locally-adapted crops cannot be readily improved being paradoxically orphan in the genomic era. Therefore, our ability of creating tools to effectively transfer and coalesce the genetic knowledge on traits of interest across crops and to integrate such tools into modern, “knowledge-based” breeding programs will ultimately represent a key factor to enable the development of biomass crops tailored to the needs of MALs and to a bio-based economy.

Chapter 5

Detection and analysis of syntenic quantitative trait loci controlling cell wall quality in angiosperms

Francesco Pancaldi[1], Dennis Vlegels, Hugo Rijken, Eibertus N. van Loo, Luisa M. Trindade[1]

[1]Plant Breeding, Wageningen University & Research, Wageningen, The Netherlands

Published as: Pancaldi, F., Vlegels, D., Rijken, H., van Loo, E.N. and Trindade, L.M. (2020). Detection and analysis of syntenic quantitative trait loci controlling cell wall quality in angiosperms. *Frontiers in plant science*, *13*.
DOI: 10.3389/fpls.2022.855093

Abstract

Translational genomics can enable a quicker improvement of orphan crops toward novel agricultural applications, including the advancement of orphan biomass species for cultivation on marginal lands. In this sense, cell wall quality is a preeminent breeding target. However, tools to efficiently project genetic data on target traits across large sets of species are currently missing. This study aimed at closing this gap by developing a strategy to project a set of cell wall QTLs across a large group of plants by using genome synteny. This strategy is suited for large-scale analyses and detected 362 syntenic cell wall QTLs (SQTLs) across 74 angiosperms, including several (orphan) biomass species. SQTLs analyses revealed that they span large portions of the initial cell wall QTLs and are extensively conserved across diverse species. Moreover, numerous QTLs cell wall genes were conserved through SQTLs, including genes displaying allelic variation associated with cell wall composition. Functional analyses showed that highly conserved genes of SQTLs include important cell wall transcription factors and genes involved in the remodelling of cell wall polymers. For some of these gene families, SQTLs indicated the presence of differentially conserved genomic contexts for different gene members, highlighting their utility as a tool to pinpoint gene targets that maximize the likelihood of functional gene conservation. Overall, the results of this study can facilitate "universal" approaches for breeding (orphan) biomass crops, while the strategy for QTLs translation can be applied to other sets of traits and species, helping to unlock the potential of orphan species.

1 Introduction

A key challenge faced by modern plant breeding is the effective valorisation of the genetic resources of orphan crops (Kamei et al., 2016). These crops are currently underutilized in agriculture, despite being relevant for the subsistence of many local and regional communities, being promising for emerging agricultural markets, and carrying valuable traits (Kamei et al., 2016, Tadele, 2019). Orphan crops include plants that can potentially fit different agricultural applications, from diverse food species (cereals, legumes, vegetables, tubers) to several industrial crops (oilseeds and biomass species, as well as plants suitable for multipurpose biorefineries) (see Tadele (2019) and Pancaldi and Trindade (2020) for an extensive list of orphan species). Moreover, many orphan crops are native of areas that face adverse climatic conditions (Pancaldi and Trindade, 2020). Therefore, their advancement could contribute to diversify agricultural markets, improve food diets, as well as mine useful traits to adapt staple crops to climate change (Kamei et al., 2016, Tadele, 2019). One of the main reasons for the underutilization of orphan crops in agriculture is the scarcity of genetic tools available for their improvement (Pancaldi and Trindade, 2020). This aspect, together with the niche markets currently shared by most of these crops, hampers their effective utilization for the purposes just discussed.

Several studies identified a turning point for the effective valorization of orphan crops in the rapid advancement of genome sequencing technologies, together with the dropping of their costs (Salentijn et al., 2007, Kang et al., 2016). In fact, *de novo* genome assemblies do not represent a constraint to plant research anymore, and the increasing availability of these resources for new species calls for approaches where they can be effectively used for transferring knowledge on traits of interest from model species to new crops (Salentijn et al., 2007, Pancaldi and Trindade, 2020). These approaches can be broadly referred as "translational genomics", and generally encompass the analysis of genetic information on traits of interest from multiple species and their inter-species projection through genomic analyses (Kang et al., 2016). Typically, genetic data on the determinants of traits of interest are available as lists of candidate genes detected through reverse genetics, or as markers delimiting quantitative trait loci (QTLs) mapped in forward genetic studies (Pflieger et al., 2001, Rafalski, 2010). In principle, if the physical genomic positions of target genetic elements are available, genes, markers and QTLs could be projected between species by using bioinformatic tools such as the analysis of gene synteny (i.e. the conservation of the type and order of genes across the genomes of different species) (Zhao and Schranz, 2019, Pancaldi and Trindade, 2020). However, because of the current lack of pipelines suited for whole-trait analyses on large sets of species, these methodologies have been hardly applied at a large-scale level so far (Pancaldi and Trindade, 2020).

Among all the sectors where orphan crops could provide precious resources to plant breeding and agriculture, the field of the bio-based economy and of biomass crops is certainly one whose market opportunities are expected to grow more rapidly over the next decades (Piotrowski et al., 2015). In this sense, it has been shown that biomass crops for bio-based applications could be extensively grown on marginal lands (i.e. lands not used by agriculture and of poor natural value) to avoid competition with food crops (Carlsson et al., 2017, Mehmood et al., 2017, Pancaldi and Trindade, 2020). However, this requires crop varieties able to produce large and high-quality yield under the suboptimal (a)biotic conditions of marginal lands (Dauber et al., 2012, Blanco-Canqui, 2016). Interestingly, several perennial biomass species are native of marginal areas, and are naturally able to withstand the stresses that can be encountered there (Carlsson et al., 2017, Mehmood et al., 2017). However, the amount and quality of their biomass yield is not optimal, as most of these crops lay in an orphan state and never underwent breeding so far (Pancaldi and Trindade, 2020). Examples of these crops include miscanthus, switchgrass, poplar or willow, which share a sequenced genome; but also herbaceous and woody biomass species for which genetic resources are even more scarce, as giant reed, reed canary grass, black locust and siberian elm (Pancaldi and Trindade, 2020). These crops should therefore be improved, and a preeminent target trait is certainly cell wall quality, which is the major determinant of biomass quality for bio-based applications (van der Weijde et al., 2013, Isikgor and Becer, 2015, Van der Cruijsen et al., 2021). Specifically, the total content of the major cell wall components – cellulose, hemicelluloses, and lignin – as well as cellulose cristallinity, degree and type of hemicellulose substitutions, and the monolignol composition of the lignin polymers are all target characters at the basis of biomass quality (van der Weijde et al., 2013, Van der Cruijsen et al., 2021). Applying translational genomics to the improvement of these traits in orphan biomass crops is expected to significantly speed up the development of varieties that combine resilience to marginal land conditions with production of large and good-quality biomass yields (Pancaldi and Trindade, 2020).

Within the context above, this study aimed at setting up a strategy for the inter-species projection of a set of 610 QTLs controlling cell wall quality previously mapped in 8 diverse plant species (detailed information in **Supplementary Table 1**) across a wide group of angiosperms, by using genome synteny. In this strategy, cutting-edge bioinformatic tools for network synteny analysis in large sets of genomes were applied to infer the syntenic conservation of the QTLs across all the plants of the study, leading to the detection of numerous conserved syntenic QTL regions. Syntenic cell wall QTLs were then characterized in terms of extensiveness among plants, fragmentation of their syntenic conservation, and conserved candidate genes. This allowed to make general inferences on the functional relevance of our translational genomics approach, to improve our knowledge on the conservation of critical genes at the basis of plant cell walls, and to highlight important candidate genes for further studies in (orphan) plant species. The strategy developed in this study can be applied to other traits, as well as other sets of QTLs and species, representing a novel tool to assist breeding research in orphan crops.

2 Materials and methods

2.1 Collection of cell wall QTLs

Scientific literature was searched for all the QTLs related to cell wall quality traits mapped in diverse species and delimited by molecular markers whose physical genomic position was reported or could be retrieved by BLAST (Altschul et al., 1990) or regression of genetic to physical genomic maps. This search was made in August 2019 and retrieved 610 QTLs for different traits related to cell wall quality from 19 different publications and 8 diverse plant species (*Arabidopsis thaliana*, *Eucalyptus grandis*, *Glicine max*, *Miscanthus sinensis*, *Oryza sativa*, *Populus trichocarpa*, *Sorghum bicolor*, *Zea mays*) (see **Supplementary Table 1** for the full list of QTLs, traits, references, and QTL positions in the genome of each species).

2.2 Collection of plant genomes

All the angiosperm genomes sequenced and published by the end of 2018 and available with at least a scaffold-level assembly were searched for in several online databases. For each genome, a BED file indicating gene positions and a FASTA file reporting protein sequences coded by all the annotated genes were retrieved. Genomes were checked for assembly completeness by using the BUSCO Viridiplantae gene set (Seppey et al., 2019) and for assembly fragmentation by assessing the number of scaffolds and the N50 statistics. To select reliable genome assemblies for synteny analysis, genomes with <75% BUSCO genes and <10 genes per scaffold on average have been excluded from the collected set. Through these criteria, a total of 151 genomes from 134 species were collected and used in all further analyses (**Supplementary Table 2**).

2.3 Identification of candidate cell wall genes in all the genomes used

Scientific literature was searched for all the genes known to play a role in plant cell wall synthesis and functioning (**Supplementary Table 3**). As the vast majority of cell wall genes turned out to be discovered or studied in *Arabidopsis thaliana*, the proteome of this species was downloaded from UniProt (UP000006548) (UniProt Consortium, 2018) and filtered for the identified cell wall functions based on UniProt

(UniProt Consortium, 2018), TAIR (Berardini et al., 2015), and NCBI (NCBI Resource Coordinators, 2018) protein annotations. This filtering led to a list of 1311 arabidopsis cell wall proteins, which were then extracted from the PEP file of the arabidopsis genome and annotated for their domain architecture using HMMER (default parameters) (Wheeler and Eddy, 2013) and all the HMM alignments of the PFAM database (Finn et al., 2014). The 1311 arabidopsis cell wall proteins were then used in a search for homologs based on both BLAST (Evalue = 1E−3) and HMMER (for matching PFAM protein architecture; default parameters) across the 151 genomes of the study. The set of candidate cell wall proteins identified was further adjusted by an iterative search of cell wall gene homologs based on BLAST and HMMER with the use of the identified candidate cell wall proteins as queries. In addition, known cell wall genes specific to certain species were used separetly from the set of 1311 arabidopsis cell wall genes as queries for homologs search as performed with the arabidopsis genes. All together, these searches yielded 252471 candidate cell wall genes with functional annotation across the 151 genomes of the study (**Supplementary Table 4**)

2.4 Construction of the cell wall QTLs synteny network

Genome synteny was analysed across all the 151 genomes of the study by following the methodology developed by Zhao and Schranz (2017) for large-scale network synteny analysis. In brief, Diamond (Buchfink et al., 2015) was used to perform BLAST-like alignments of all the proteins of each genome against all the other proteins of that genome and all the proteins of every other genome (Evalue = 1E−3). In this way, homologous genes between different species were identified across all pairs of genomes. Subsequently, MCScanX (Wang et al., 2012b) was used to detect synteny (i.e. conserved gene order across multiple genomes) by evaluating the positions of the homologous genes from each genome comparison. MCScanX was run with default parameters except -s (number of colinear genes to claim a syntenic block) set to 3. The outputs of MCScanX were organized in a synteny network, in which each node is a gene and edges represent syntenic connections between genes. The synteny network was then filtered to retain only pairs of nodes in which at least one of the genes was included in the initial 610 cell wall QTLs (R, custom script). The output of this filtering was in turn further subset as described in section 2.5 and constitutes the network of syntenic relationships of all the genes included in the 610 cell wall QTLs across all the 151 genomes of the study (**Supplementary Table 5**).

2.5 Filtering the syntenic QTL network

To identify conserved syntenic QTL regions representing reliable conservation of the initial 610 cell wall QTLs, it was assessed with which plant families the QTLs of each initial species displayed the highest synteny levels (calculated as the average percentage of genes of each initial QTL syntenic to the genomes of each plant family). The distributions of synteny levels were plotted in separate boxplots for each species/family for which cell wall QTLs were retrieved (**Supplementary Figure 1**). These boxplots were then used to filter the QTL synteny network to retain only the syntenic connections between genes of the initial QTLs and other genes belonging to genomes included in the upper quartile of each boxplot distribution (R, custom script). **Table 1** illustrates the groups of families selected for each species for which initial cell wall QTLs were available. As a next step, the fragmentation of QTL synteny across the genomes included in each group of **Table 1** was also assessed, to discriminate between cases in which small QTL fragments were syntenic toward several different genomic regions of a target species and cases in which large QTL(s) segment(s) were syntenic toward a single region in a target species (implying higher likelihood of QTLs functional conservation). To this aim, the synteny level of each initial QTL against each of the chromosomes of the species included in the groups of **Table 1** was assessed, and the syntenic QTL network was filtered to retain only syntenic connections between QTL genes and other genes located on chromosomes on which at least 50% of QTL's genes showed synteny. The filtered syntenic QTL network is included in **Supplementary Table 5** and contains 494026 genes (nodes) from 87 genomes.

Table 1 – Lists of plant families used for SQTLs detection for each of the plant families for which initial QTLs were retrieved. The selection of plant families was based on the synteny level of the different initial QTLs against the 151 species of the project, as described in paragraph 2.5.

Initial QTLs families	Plant families selected for SQTL detection
Brassicaceae	Brassicaceae, Malvaceae, Cleomaceae, Anacardaceae, Actinidiaceae, Myrtaceae
Fabaceae	Fabaceae, Salicaceae, Moraceae, Rhamnaceae, Linaceae, Cannabaceae, Euphorbiaceae, Rosaceae, Vitaceae, Cucurbitaceae, Crassulaceae, Nelumbonaceae, Ranunculaceae, Myrtaceae, Papaveraceae, Lythraceae, Anacardaceae, Rutaceae, Malvaceae, Brassicaceae, Cleomaceae, Amaranthaceae, Actinidiaceae, Convolvulaceae, Solanaceae, Rubiaceae, Oleaceae, Pedaliaceae, Phrymaceae, Asteraceae, Apiaceae
Myrtaceae	Lythraceae, Anacardaceae, Rutaceae, Malvaceae, Cucurbitaceae, Rosaceae, Rhamnaceae, Fabaceae, Actinidiaceae, Salicaceae, Rubiaceae, Vitaceae, Oleaceae, Pedaliaceae, Phrymaceae, Apiaceae
Salicaceae	Salicaceae, Linaceae, Fabaceae, Euphorbiaceae, Moraceae, Rhamnaceae, Vitaceae, Cannabaceae, Rosaceae, Crassulaceae, Cucurbitaceae, Nelumbonaceae, Ranunculaceae, Papaveraceae, Myrtaceae, Lythraceae, Anacardaceae, Rutaceae, Malvaceae, Brassicaceae, Cleomaceae, Amaranthaceae, Lauraceae, Theaceae, Amborellaceae, Actinidiaceae, Convolvulaceae, Solanaceae, Rubiaceae, Oleaceae, Pedaliaceae, Phrymaceae, Asteraceae, Apiaceae
Poaceae	Arecaceae, Araceae, Bromeliaceae, Asparagaceae, Orchidaceae, Musaceae, Poaceae

2.6 Identification of syntenic cell wall QTLs

The filtered syntenic QTL network was used to detect syntenic cell wall QTLs within each group of **Table 1** by following a "double-clustering" approach. First, to identify sets of genes highly syntenic with each other and with initial cell wall QTL(s), the R igraph package was used to identify all the communities of at least 10 nodes within the QTL synteny network (Louvain algorithm; 16644 communities detected; modularity = 0.99). These communities turned out to typically include all the members of a single gene type that are syntenic across the genomes inspected. Therefore, to detect syntenic QTL *regions* from single-homologs syntenic communities, a second clustering was applied to the detected communities. In this step, the identifier(s) of the QTL(s) harboured by (some of) the genes included within each detected community were annotated to the communities themselves. The annotated communities were then used to calculate the all-vs-all similarity of the communities themselves based on the initial cell wall QTLs represented in each community (R, custom script; Jaccard similarity algorithm). Similarities between communities were saved into a network and their distribution was plotted in a boxplot. This network was then filtered to retain only the connections between communities supported by a similarity >0.6 (upper quartile of the distribution of similarities). In turn, the QTL synteny network was then also filtered to retain only syntenic relationships between genes included in the filtered communities. The genomic regions whose genes are included in the communities contained in the filtered QTL synteny network represent the syntenic cell wall QTLs (SQTLs).

2.7 Analysis of syntenic cell wall QTLs

The SQTLs detected through the methodology above were analyzed in terms of extensiveness, fragmentation, frequency and size across species, as well as for the co-localization of functionally-different initial QTLs. Moreover, the conservation of cell wall genes through syntenic cell wall QTLs was also analyzed. In this respect, all the analyses performed are described in the next sections of the manuscript, and were performed by using R, Excel, or SPSS.

3 Results

3.1 Preliminary analysis of cell wall QTLs and cell wall gene data

To develop an effective methodology for projecting cell wall QTLs across a wide set of plants with the use of gene synteny, 610 cell wall QTLs previously mapped in arabidopsis (Brassicaceae), soybean (Fabaceae), poplar (Salicaceae), eucalyptus (Myrtaceae), miscanthus, maize, sorghum, and rice (Poaceae) were collected from scientific literature (Section 2.1 and **Supplementary Table 1**). In addition, ~250000 candidate cell wall genes were identified across >150 angiosperm genomes through a combined BLAST- (Altschul et al., 1990) and HMMER-based (Wheeler and Eddy, 2013) search on a large set of characterized cell wall genes retrieved from scientific literature (Section 2.3 and **Supplementary Table 4**). To assess the feasibility of genomically translating the QTLs and the candidate cell wall genes therein across a wide set of species through gene synteny, QTLs and cell wall genes were initially assessed for QTL gene content, QTL length variability, and general synteny of both QTLs and candidate cell wall genes.

Since some of the 610 cell wall QTL intervals retrieved from scientific literature referred to genetic maps, the QTLs were first translated to physical genomic positions (Section 2.1 and **Supplementary Table 1**), and the gene content of the genomic QTL regions was analysed. Knowing the gene content of target regions to be projected between species through the use of genome synteny is highly relevant, since synteny is defined at the gene level (i.e. conservation of the type and order of genes between species) (Zhao and Schranz, 2017). The analysis of QTLs gene content revealed that 16 QTLs are located on genomic regions without genes, 37 QTLs do not span any candidate cell wall gene, and 50 QTLs include only one candidate cell wall gene (**Supplementary Table 1**). The latter group raises a particular interest, as the candidate cell wall gene harboured by each of those QTLs may represent the causative gene of each QTL. Therefore, these genes were collected and analysed, revealing that they vary considerably in terms of cell wall function(s) and process(es) in which they play a role (**Supplementary Table 6**).

By using the physical QTL ranges, the QTLs length variation was also assessed (meant as both nucleotide length of QTL ranges and total number of genes within QTLs). This is another important parameter to be considered in synteny analysis, since the length of target genomic regions for synteny analysis is known to potentially affect the fragmentation of the syntenic regions obtained (Liu et al., 2018a). In this sense, ANOVA results indicated that QTLs collected for the Poaceae species and, to a lesser extent, for eucalyptus, span significantly longer nucleotide regions than the QTLs of arabidopsis, poplar, and soybean, whose length ranges do not differ significantly from each other (Bonferroni's LSDs; $\alpha=0.05$). However, given the highly significant differences between the species for which QTLs were retrieved in terms of gene density of the QTL regions ($P<0.000$), arabidopsis QTLs displayed a significantly higher gene content than the QTLs from all the other species ($P<0.000$). In turn, QTLs from Poaceae, poplar, eucalyptus, and soybean did not differ from each other in terms of gene content, with the only exception of poplar and eucalyptus QTLs (Bonferroni's LSDs; $\alpha=0.05$). The patterns observed for the overall QTLs gene content are similar to what observed for the QTLs cell wall gene content. Accordingly, the correlation between total QTL gene content and QTL cell wall gene content turned out to be particularly high ($\rho=0.91$, $P<0.000$).

In addition to QTL gene content and QTL length variability, the general synteny of the candidate cell wall genes and of the 594 cell wall QTLs spanning at least one gene was also assessed, in order to estimate the overall feasibility of using gene synteny for inter-species QTLs projection. These analyses were performed by using the general synteny network of the 151 genomes of the study and the filtered QTL synteny network, respectively (see Section 2.4 and 2.5). Results revealed that candidate cell wall genes are significantly more syntenic than other genes not related with cell wall across all the genomes analysed. This pattern holds true both when assessed across whole genomes and when assessed over cell wall QTL regions only. Specifically, t-tests showed that each of the 252471 candidate cell wall genes identified across the 151 angiosperm genomes of the study displays synteny with other 101 homologs in other species on average, compared to 68 average syntenic connections for the non-cell wall genes (P<0.000) (**Figure 1A**). Within QTLs, these figures amount to 107 and 67 syntenic connections, respectively (t-test's P<0.000) (**Figure 1B**). To conclude, the synteny level of QTL genes (both cell wall and non-cell wall) does not significantly differ from the one of genes outside QTL regions (t-test; α=0.05) (**Figure 1C**). Nevertheless, with an average of 69 syntenic connections per gene, QTL genes can overall be considered highly syntenic (**Figure 1C**).

3.2 Detection and descriptive analysis of syntenic cell wall QTLs

As previously mentioned, the main goal of this study was to develop a methodology to efficiently project an initial set of 610 cell wall QTLs across 151 angiosperm genomes by using gene synteny, to identify conserved syntenic cell wall QTLs (SQTLs). The high synteny of the cell wall QTLs collected from scientific literature and of the candidate cell wall genes therein (Section 3.1) promised success for reaching this goal, and the approach described in **Figure 2** and Sections 2.4-2.6 was therefore designed. In this pipeline, the genes contained in the 594 cell wall QTLs spanning at least one gene were evaluated for syntenic conservation across all the genomes of the study by building a synteny network of QTL genes where each node represents a gene and edges connect syntenic genes (**Figure 2B**). The network was first used to evaluate the degree of synteny of each initial QTL across all the genomes of the study, leading to the identification of the groups of **Table 1**, which list species among which the synteny of the different initial QTLs is maximized, increasing the likelihood of QTLs' functional conservation across species. In turn, each group of **Table 1** was used to cluster the corresponding genes of the QTL synteny network first into single-locus communities (**Figure 2C**), and second into groups of communities representing syntenic regions sharing same initial QTLs (**Figure 2D**). These groups constitute the syntenic cell wall QTLs (SQTLs), which can be defined as genomic regions conserved across multiple species and spanning (a part of) one or more known cell wall QTLs in at least one species. **Figure 3** shows a meaningful example of SQTL.

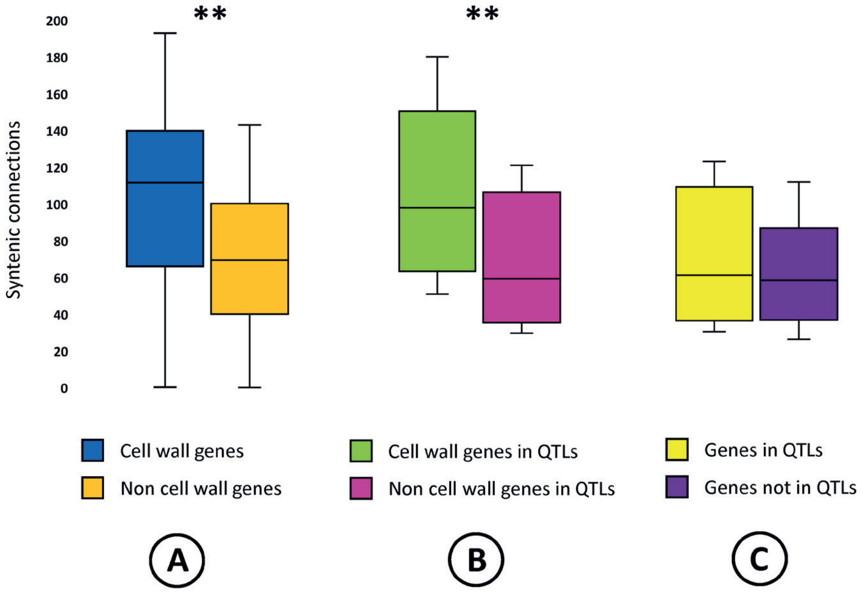


Figure 1. The level of synteny of different relevant classes of genes. A) Boxplots of the number of syntenic connections per gene of cell wall genes and of non-cell wall genes from the 151 genomes of the study. B) Boxplots of the number of syntenic connections per gene of cell wall genes and of non-cell wall genes from the 594 cell wall QTLs of the study spanning at least one gene. C) Boxplots of the number of syntenic connections per gene of cell wall QTL genes and of genes not included in cell wall QTLs from the 151 genomes of the study. **Significant at alpha = 0.01.

The pipeline above led to the detection of 362 SQTLs across the five groups of angiosperm genomes of **Table 1** (see **Supplementary Table 7** for a list of all the SQTLs and all their genes). These SQTLs span a total of 398231 genes (81% of the genes within the QTL synteny network) and of 74 different species (55% of the species used in the study). On average, each SQTL groups 1100 genes (CV=152%) from 18 different species (CV=37%) with a mean of 62 genes per species (CV=169%). Of all the genes within the 362 SQTLs, 24987 are candidate cell wall genes (83% of the cell wall genes of the QTL synteny network), with an average of 69 cell wall genes per SQTL (CV=160%) and of 4 cell wall genes per species within each SQTL (CV=168%). Out of the 362 SQTLs, 92 do not contain any candidate cell wall gene. These 92 SQTLs contain significantly less genes and span significantly less species than the other SQTLs (t-test; P<0.05 for both). To conclude, within the 270 SQTLs containing candidate cell wall genes, these represent on average 9% of the total SQTLs genes (CV=85%, range=1-63%).

To evaluate the validity of the approach followed for SQTLs detection, as well as to gain insights into the patterns of conservation of the cell wall QTLs across the 74 angiosperm species represented within SQTLs, the 362 SQTLs were detailly characterized for several attributes. These include the overall representation and fragmentation of initial cell wall QTLs across SQTLs, the frequency and extensiveness of SQTLs across relevant (groups of) plant species, the SQTL size (both overall and across distinct plants), and the general patterns of candidate cell wall gene conservation through SQTLs.

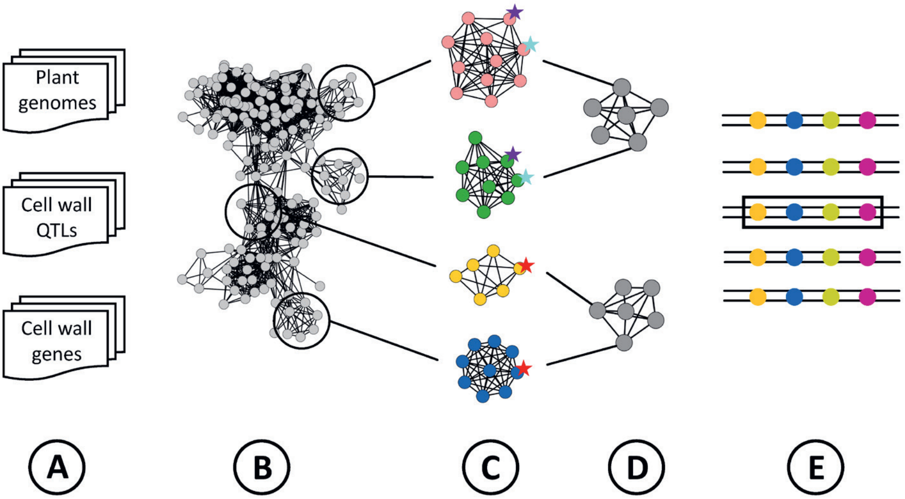


Figure 2. The pipeline followed for detecting syntenic cell wall QTLs (SQTLs). A) The starting data sets of 151 angiosperm genomes, of 610 cell wall QTLs from 8 diverse plant species, and of ~250000 candidate cell wall genes identified across all the genomes of the study. B) A synteny network of QTL genes across the species of the study was created, where each node (grey circles) represents a gene and black edges represent syntenic connections between genes. C) Syntenic communities grouping the members of specific gene families that are syntenic across (groups of) plants were detected and annotated with the initial cell wall QTLs harboured by (some of) their members. Coloured circles represent genes, and circles with the same colour are genes belonging to the same syntenic community. Coloured stars represent initial cell wall QTLs, with different colours indicating different initial cell wall QTLs. D) Syntenic communities were clustered into SQTLs based on similarity of the cell wall QTLs harboured by each community. A SQTLs network was therefore obtained, where each node represent a syntenic community (grey circles), while black edges between communities indicate a similarity >0.6 between two syntenic communities in terms of initial cell wall QTLs harboured. E) Representation of the genomic meaning of SQTLs. Each double strip with circles represents the genome of a species, and coloured circles represent genes of different types (different colours). A SQTL represents a genomic region syntenic between multiple species (same genes in the same order) and that in at least one species spans at least (a part of) one initial cell wall QTL (black rectangle).

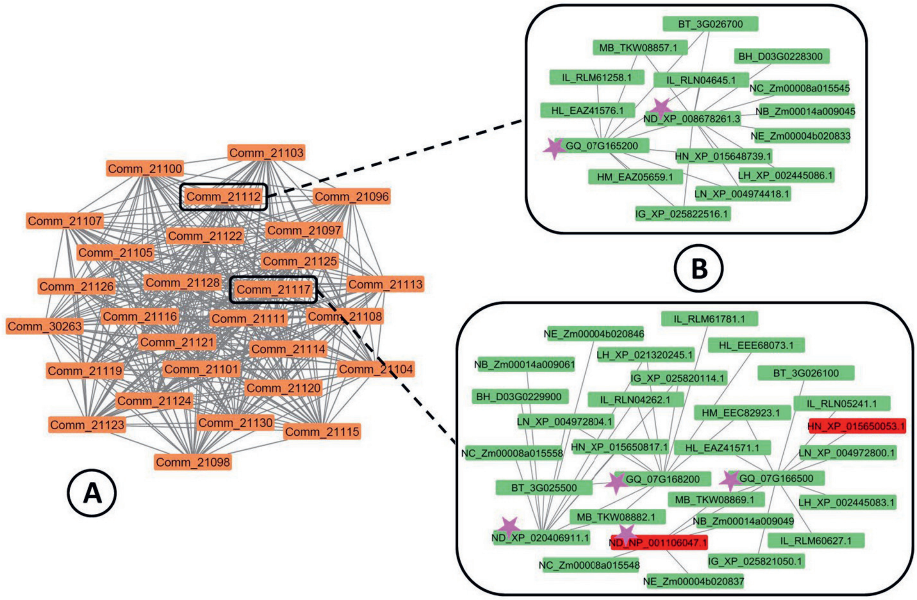


Figure 3. A) Example of a SQTL (SQTL_196). The orange nodes represent all the syntenic communities of SQTL_196, each of them grouping the genes of a syntenic genomic region, of which at least some are retained within initial cell wall QTLs. B) Two of the syntenic communities that constitute SQTL_196. Green rectangles represent community nodes. Pink stars indicate maize (ND) and miscanthus (GQ) nodes included within initial cell wall QTLs. Red nodes indicate a gene from maize (ND_NP_001106047.1) and a gene from rice (HN_XP_015650053.1) corresponding to the maize *Brown Midrib 3* locus and the rice *CAldOMT1* locus, respectively. Both these loci encode a *COMT* gene involved in lignin synthesis and known to display allelic variation associated to phenotypic variation in cell wall quality in maize and rice populations.

Regarding the representation and fragmentation of initial cell wall QTLs within SQTLs, our analyses revealed that 512 of the 594 initial cell wall QTLs spanning at least one gene are involved in SQTLs. Of these 512 QTLs, 90 are involved into one SQTL only, which includes 39% of the initial QTLs' genes on average (range = 0.4%-84%). The other 422 QTLs are instead involved into two or more SQTLs (average of 5; range = 2-31), with each SQTL spanning 13% of the initial QTLs' genes on average (CV=68%; range=0.3%-42%). QTL fragmentation over multiple SQTLs is not even across the five plant groups of **Table 1**. Specifically, while the QTLs from the dicot species are divided over 2.7 SQTLs on average, for the Poaceae QTLs this average is significantly higher (5.2 SQTLs; t-test's P<0.000), revealing a higher level of fragmentation (**Figure 4A**). In addition, the length of QTL fragments conserved through SQTLs is significantly shorter in Poaceae, Myrtaceae and Salicaceae (34, 11, 43 genes on average respectively) than in the Brassicaceae and Fabaceae groups (163 and 113 genes on average, respectively; ANOVA; P<0.000). To conclude, the representation of initial cell wall QTLs within SQTLs was analysed also from the point of view of SQTLs. This revealed that each of the 362 SQTLs spans 6 different initial cell wall QTLs (CV=105%;

range=1-55) that come from 2 different initial QTL species (CV=49%; range=1-5) on average.

Table 2 – The extensiveness of SQTLs across the groups of species used for their detection

SQTL group	Total species included in SQTL group	Total number of SQTLs	Average species spanned by SQTLs	Percentage over total species
Brassicaceae	29	9	6	21%
Fabaceae	111	16	31	28%
Myrtaceae	59	6	21	35%
Salicaceae	112	50	25	22%
Poaceae	27	281	17	62%

As previously mentioned, other analyses were performed to inspect the frequency and extensiveness of SQTLs across species. In this regard, SQTLs turned out to be numerous across all the 5 initial groups of **Table 1**, even if with substantial differences reflecting the asimmetry in the initial QTLs availability between plant groups and the intrinsic genomic relatedeness of the species within each group. In fact, while 281 SQTLs were detected in the Poaceae (~75% of the total; the group with the highest number of initial QTLs and a well-known example of highly-syntenic plant family), the SQTLs detected across the other groups of **Table 1** range from 6 for Myrtaceae to 50 for Salicaceae (**Table 2**). Despite the diversity in the absolute SQTL abundance across plant groups, SQTLs from all the groups are extensive across the species of each group, as each SQTL spans on average 54% of the species included in the group of **Table 1** used for its detection. The extensiveness of SQTLs is again highest in the Poaceae group, where each SQTL spans on average 62% of the "Poaceae" species from **Table 1** (**Table 2**). Nevertheless, the percentages found in the dicot groups are also relatively high (**Table 2**), considering the higher evolutionary diversity of these species compared to the Poaceae group.

Another parameter that was assessed is SQTLs size (meant as total number of genes spanned by SQTLs in each of the species included in SQTLs). Results revealed that SQTLs size is highest in SQTLs mapped within the Brassicaceae group (267 genes/species/SQTL), being significantly higher than in all the other groups of **Table 1** (51 genes/species/SQTL as average across all other groups; ANOVA's $P<0.000$) (**Figure 4B**). Moreover, all the other groups were found to not significantly differ between each other in terms of SQTLs size (ANOVA's $\alpha=0.05$). Interestingly, Brassicaceae is the group whose initial QTLs display the highest gene content (see Section 3.1), and QTL gene content was positively correlated with SQTLs size across all initial QTLs species ($\rho=0.91$, $P<0.000$).

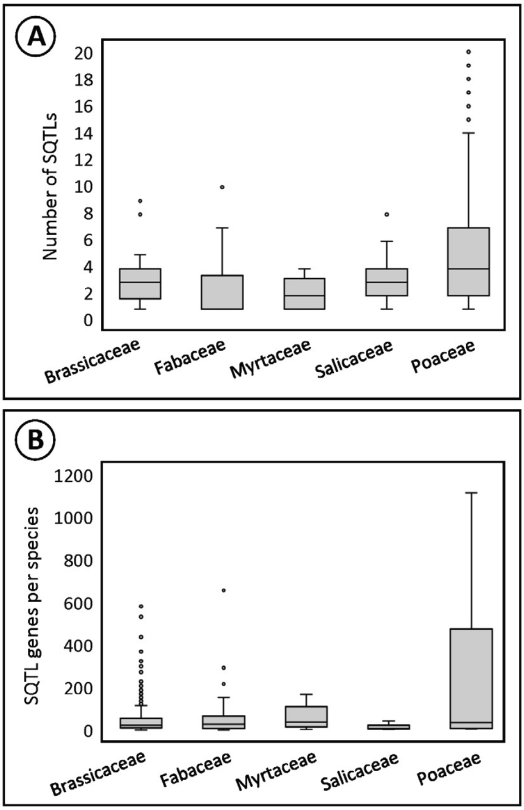


Figure 4. A) The level of fragmentation of the initial cell wall QTLs from the different groups of **Table 1** over the SQTLs in which they are involved. Fragmentation is expressed as number of different SQTLs over which different initial QTLs resulted involved in (y-axis), per species group (x-axis). B) The variability of SQTL length (expressed as number of genes per SQTL and per species involved in a SQTL) of the SQTLs detected for the plant groups of **Table 1**.

To conclude, the representation of candidate cell wall genes from the initial QTLs within the 362 SQTLs was also analysed. This revealed that 28 of the 512 initial QTLs represented in SQTLs have their candidate cell wall genes not conserved at all through SQTLs. Interestingly, these 28 QTLs span significantly less genes and cell wall genes than the other QTLs (t-test, $P<0.05$ for both). However, their nucleotide length does not significantly differ from the one of the other QTLs (t-test, $\alpha=0.05$). The 484 QTLs with candidate cell wall genes represented within SQTLs have instead on average 72% of their candidate cell wall genes conserved in SQTLs (CV=37%; range=2-100%). In conclusion, in terms of traits and functions of the candidate cell wall genes, chi-square tests revealed that the candidate cell wall genes conserved through SQTLs are significantly enriched in transcription factors (+1-2%) and significantly de-enriched in lignin genes (–4-6%; $P<0.000$). Specifically, this holds true when the frequencies of candidate cell wall gene functions within SQTLs are compared with the corresponding frequencies in the full list of candidate cell wall genes, in the list of candidate cell wall genes within the QTL synteny network, and in the list of candidate cell wall genes within the 594 initial QTLs spanning at least one gene.

Table 3 – The 22 SQTLs selected for candidate gene inspection based on a high co-localization of initial QTLs and the inclusion of diverse plant species

SQTL	**SQTL group**	**Number of QTLs co-localizing in SQTL**	**Total species in SQTL**	**Species included in SQTL**
SQTL_217	Fabaceae	6	2	*Eucalyptus grandis*; *Glycine max*
SQTL_23	Myrtaceae	9	2	*Eucalyptus grandis*; *Populus trichocarpa*
SQTL_121	Myrtaceae	7	3	*Arabidopsis thaliana*; *Eucalyptus grandis*; *Populus trichocarpa*
SQTL_169	Myrtaceae	7	4	*Eucalyptus grandis*; *Glycine max*; *Populus trichocarpa*; *Populus trichocarpa*
SQTL_187	Salicaceae	14	3	*Arabidopsis thaliana*; *Eucalyptus grandis*; *Populus trichocarpa*
SQTL_246	Salicaceae	11	3	*Arabidopsis thaliana*; *Eucalyptus grandis*; *Populus trichocarpa*
SQTL_50	Salicaceae	8	3	*Arabidopsis thaliana*; *Eucalyptus grandis*; *Populus trichocarpa*
SQTL_174	Salicaceae	8	3	*Eucalyptus grandis*; *Populus trichocarpa*; *Populus trichocarpa*
SQTL_14	Salicaceae	8	2	*Arabidopsis thaliana*; *Populus trichocarpa*
SQTL_39	Salicaceae	7	3	*Arabidopsis thaliana*; *Eucalyptus grandis*; *Populus trichocarpa*
SQTL_53	Salicaceae	6	4	*Arabidopsis thaliana*; *Eucalyptus grandis*; *Populus trichocarpa*; *Populus trichocarpa*
SQTL_2	Poaceae	45	4	*Miscanthus sinensis*; *Oryza sativa*; *Sorghum bicolor*; *Zea mays*
SQTL_160	Poaceae	25	4	*Miscanthus sinensis*; *Oryza sativa*; *Sorghum bicolor*; *Zea mays*
SQTL_188	Poaceae	24	4	*Miscanthus sinensis*; *Oryza sativa*; *Sorghum bicolor*; *Zea mays*
SQTL_245	Poaceae	23	4	*Miscanthus sinensis*; *Oryza sativa*; *Sorghum bicolor*; *Zea mays*
SQTL_20	Poaceae	20	2	*Sorghum bicolor*; *Zea mays*
SQTL_91	Poaceae	20	4	*Miscanthus sinensis*; *Oryza sativa*; *Sorghum bicolor*; *Zea mays*
SQTL_47	Poaceae	20	4	*Miscanthus sinensis*; *Oryza sativa*; *Sorghum bicolor*; *Zea mays*
SQTL_56	Poaceae	19	3	*Oryza sativa*; *Sorghum bicolor*; *Zea mays*
SQTL_69	Poaceae	18	2	*Miscanthus sinensis*; *Zea mays*
SQTL_73	Poaceae	16	4	*Miscanthus sinensis*; *Oryza sativa*; *Sorghum bicolor*; *Zea mays*
SQTL_60	Poaceae	16	3	*Miscanthus sinensis*; *Sorghum bicolor*; *Zea mays*

3.3 Analysis of the candidate cell wall genes within syntenic cell wall QTLs

The 362 SQTLs described in paragraph 3.2 represent genomic regions that are syntenic across crops and that are known to display patterns of allelic variation associated with variation in cell wall composition in the species where they span known cell wall QTLs. The study of the candidate cell wall genes included in SQTLs could therefore contribute to identify relevant targets for crop improvement, to improve our knowledge on the degree of conservation of critical cell wall genes, and to design novel breeding strategies. In line with these goals, SQTLs were detailly functionally analysed as described in the next sections.

Cell wall genes within the SQTLs with highest co-localization of initial QTLs

As previously mentioned, the 362 SQTLs were checked for extensiveness across both species and initial QTLs (section 3.2). By assessing these parameters, we identified the 22 SQTLs that fall in the upper quartile of SQTLs distribution for both number of initial cell wall QTLs and number of diverse plant species represented within SQTLs (**Table 3**). Because of these properties, these 22 SQTLs could likely represent relevant regions for conserved mechanisms of cell wall quality control across several species, and the types of candidate cell wall genes that they contain were analysed.

In total, 1493 candidate cell wall genes were extracted from the 22 selected SQTLs (**Supplementary Table 8**). Notably, the proportions of candidate cell wall genes belonging to different cell wall processes observed within this gene set differ substantially from the ones observed across all the candidate cell wall genes of the study and all the candidate cell wall genes of the initial cell wall QTLs. Specifically, transcription factors (TFs) and genes involved in lignin and hemicellulose metabolism are the categories showing the largest variation. On the one hand, TFs constitute 27% of all the candidate cell wall genes from the 22 selected SQTLs (404 of the 1493 genes) (**Figure 5A**), a proportion that is 6 and 5 fold higher than what observed among the candidate cell wall genes from the 151 angiosperm genomes (**Figure 5C**) and from the 610 cell wall QTLs (**Figure 5B**), respectively. On the other hand, lignin and hemicellulose genes represent 7% (111 genes) and 13% (199 genes) of the cell wall genes from the 22 selected SQTLs, respectively (**Figure 5A**). These percentages are considerably lower than what observed in the candidate cell wall genes from the 151 genomes (20% for lignin and 25% for hemicellulose) (**Figure 5C**) and the initial cell wall QTLs (22% for lignin and 26% for hemicellulose) (**Figure 5B**).

In addition to the proportions above, the types of specific cell wall gene functions that were most represented among the candidate cell wall genes from the 22 selected SQTLs were also evaluated. Regarding TFs, the most represented categories turned out to be *Vascular-related NAC Domain* (*VND*) genes, certain cell wall related *MYB* TFs

(*MYB4*, *MYB6*, *MYB7*, *MYB21*, and *MYB32*), *WRKY12*, *NAC Secondary cell wall Thickening* (*NST*), *Ovate Family Protein 4* (*OFP4*), and cell wall related *Ethylene-responsive factors* (*ERF*). All these TFs are either known master regulators of cell wall synthesis in several species, or known regulators of the lignin pathway. Furthermore, some of them are known to display allelic variation associated with significant variation in cell wall properties across different species (see paragraph 3.3.2 and **Supplementary Table 9**). Concerning lignin genes, our analyses showed that the 22 selected SQTLs contain a relatively high amount of *peroxidases* (*PRX*), *mediator complex subunits* (*MED*), *caffeoyl CoA O-methyltransferases* (*CCoAOMT*) and *caffeoyl shikimate esterases* (*CSE*). The hemicellulose genes from the 22 selected SQTLs displayed instead a relatively large proportion of *mannan synthesis-related* (*MSR*) genes and of different gene families involved in the substitution and remodelling of hemicellulose molecules. These include *BAHD acyltransferases* (*BAHD*), *beta-xylosidases* (*BXL*), *eskimo* genes (*ESK*), *reduced wall acetylation* genes (*RWA*), and *arabinogalactan methylesterases* (*AGM*). Finally, the 22 selected SQTLs harbour also a relatively large amount of genes involved in cell wall remodelling, including *expansin/expansin-like* genes (*EXP/EXPL*), *extensins* (*EXT*), and *polygalacturonases/pectin lyases* (*PG/PL*).

The synteny networks of the gene families just discussed were extracted from the 22 selected SQTLs, to analyse the specific patterns of syntenic conservation through SQTLs of all these genes (**Figure 6**). This analysis showed that the genes above display extensive positional conservation across diverse plant species through SQTLs, as large syntenic communities exist across both monocots and dicots for all these genes (**Figure 6** and **Supplementary Table 8**). Moreover, these communities span diverse species, including important biomass crops, and include several gene members from the initial cell wall QTLs (**Figure 6**). Interestingly, for several gene families not all the members included in the genomes of the species used for SQTLs detection resulted included in the communities of **Figure 6.** For example, out of the seven *VND* TFs of arabidopsis, only one member (*AtVND7*) is included in the dicot *VND* syntenic community of **Figure 6A**. Similarily, of the five *RWA* genes of maize, only one is conserved in the monocot syntenic community of **Figure 6G**. Overall, these patterns reveal that for several gene families from the 22 selected SQTLs, only a fraction of their members from diverse species are exact positional orthologs of the genes that were originally included in cell wall QTLs. In this sense, the detected SQTLs represent a useful tool to readily identify such positional orthologs, increasing the likelyhood of complete functional gene conservation.

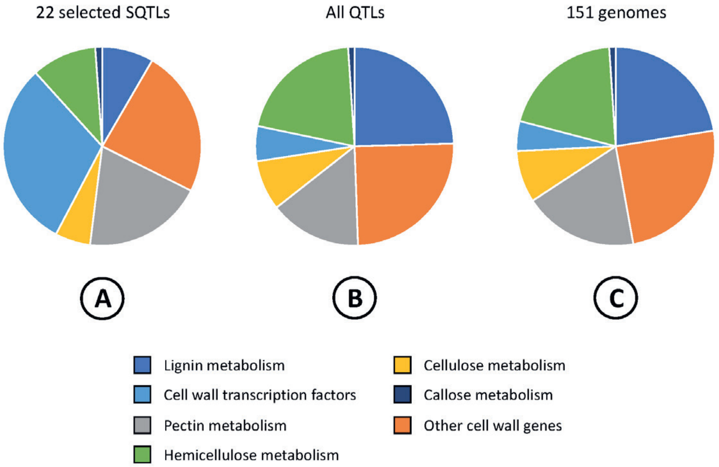


Figure 5. The proportions of cell wall genes participating to different cell wall processes observed among a set of 22 SQTLs showing an exceptionally high co-localization of initial cell wall QTLs from diverse plant species and highly conserved across different angiosperms (A), among the genes included in all the initial 594 cell wall QTLs spanning at least one gene (B), and among all the genes from the 151 angiosperm genomes of the study (C).

The conservation of critical cell wall related loci through SQTLs

Another functional analysis that was performed on the SQTLs encompassed a comparison between all the genes harbored by the 362 SQTLs and a set of 139 cell wall genes from maize, sorghum, rice, arabidopsis and poplar (**Supplementary Table 9**) that are known from scientific literature to display patterns of allelic/mutational variation with a significant impact on plant cell walls. Since these genes constitute an important set for breeding biomass crops, this analysis aimed at assessing the relevance of SQTLs for crop improvement based on the extent of the conservation of critical cell wall genes through SQTLs. Interestingly, 85% of the genes collected for the grass species (64 out of 75 genes) turned out to be included in SQTLs. Among others, these genes include different *brown-midrib* loci of maize and sorghum (*ZmBM1*, *ZmBM2*, *ZmBM3*, *ZmBM4*, *SbBM2*, *SbBM3*), 15 of the 17 *brittle culm* and *brittle culm-like* loci of rice, as well as several critical TFs involved in grass cell wall regulation (**Supplementary Table 9**). In total, the 64 grass genes are involved in 34 different SQTLs. Moreover, the syntenic communities at the basis of these 34 SQTLs revealed that the genomic organization of these genes is extensively conserved across monocot species (**Figure 7**). In fact, those communities typically span staple crops as maize, rice, barley and sorghum; biomass crops as miscanthus and switchgrass; less utilized relatives of grass cereals as wild rice species, *Panicum miliaceum*, and *Setaria italica*; as well as important species for plant research as *Brachypodium dystachyon* (**Figure 7**). Finally, syntenic communities also revealed interesting differences in the copy

5

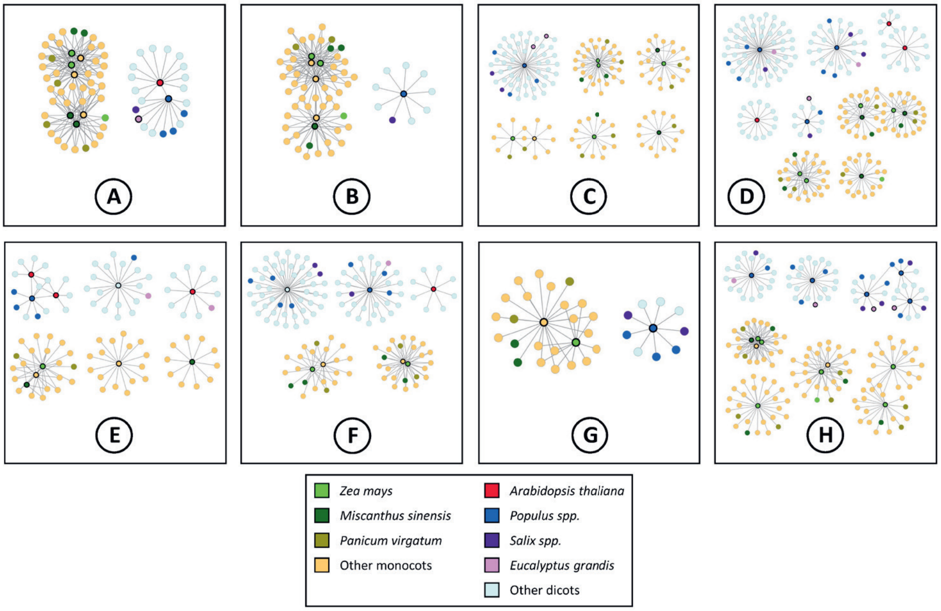


Figure 6. The structure of the syntenic communities of some of the highly conserved cell wall gene families included within 22 SQTLs showing an exceptionally high co-localization of initial cell wall QTLs from diverse plant species and a high extensiveness across angiosperms. Each independent network represents a gene community conserved across a set of species. Network nodes indicate the gene members of syntenic communities from different species (see legend). Network nodes with black bold edges indicate genes contained within initial cell wall QTLs. A) *Vascular-related NAC domain* (*VND*); B) *WRKY12*; C) *NAC Secondary cell wall Thickening factor* (*NST*); D) ***C**affeoyl **S**hikimate **E**sterases* (*CSE*); E) ***C**affeoyl CoA O-**M**ethyltransferases* (*CCoAOMT*); F) **M**ediator complex subunits (*MED*); (G) *Reduced Wall Acetylation* **(*RWA*);** H) *BAHD Acyltransferase*.

number of conserved positional orthologs between species, as well as the occurrence of some of those positional orthologs within initial cell wall QTLs (**Figure 7**).

For the dicot genes, the situation is opposite, as only 29% of the genes collected for arabidopsis and poplar (19 out of 64 genes) are included in SQTLs (across 11 different SQTLs). Still, the genes contained in SQTLs are relevant, as they include the arabidopsis *IRX14* gene (an essential *xylan synthase* for the synthesis of hemicellulose backbones), an arabidopsis *COBL4* homolog of the sorghum *brittle culm 1* locus, an arabidopsis *CCR*, two poplar *laccases* (*LAC*) and different TFs. Moreover, detailed analyses of gene networks showed that positional conservation through SQTLs is extensive also for these dicot genes (**Figure 7F and 7G**). For example, *AtIRX14* turned out to be syntenic through SQTLs to one homolog from Brassicaceae, another one from *Eucalyptus grandis* (included in an initial cell wall QTL), two from *Fragaria vesca*, two from *Rosa chinensis*, and one from *Theobroma cacao* (**Figure 7G**). In other cases, synteny was restricted to specific plant families, but still extensive within them, as in

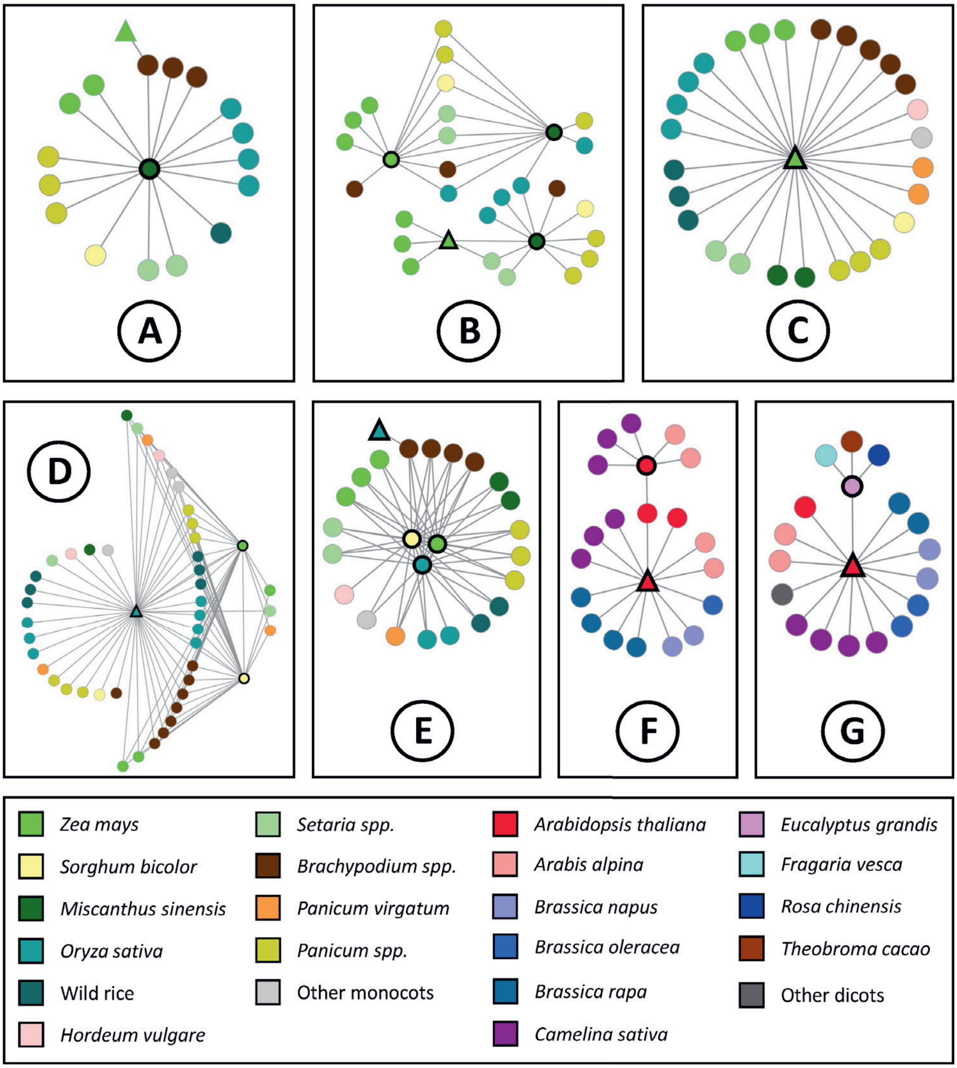


Figure 7. The structure of the syntenic communities of some of the monocot and dicot cell wall genes known to display allelic/mutational variation associated with variation in cell wall quality phenotypes and included in SQTLs. Each independent network represents a gene community conserved across a set of species. Network nodes indicate the gene members of syntenic communities from different species (see legend). Network nodes with black bold edges indicate genes contained within initial cell wall QTLs. Triangular nodes indicate that the gene is known from literature to display allelic/mutational variation associated with variation in cell wall quality phenotypes. Circular nodes are positional orthologs of those genes. A) *Zea mays Brown Midrib 1* (*ZmBM1*); B) *Zea mays Brown Midrib 3* (*ZmBM3*); C) *Zea mays Brown Midrib 4* (*ZmBM4*); D) *Oryza sativa Brittle culm-like 8* (*OsBCL8*); E) *Oryza sativa Brittle culm-like 9* (*OsBCL9*); F) *Arabidopsis thaliana Cobra-like 4* (*AtCOBL4*); G) *Arabidopsis thaliana Irregular xylem 14* (*AtIRX14*).

the case of *AtCOBL4*, which showed extensive conservation through SQTLs across six different species of the Brassicaceae family (**Figure 7F**).

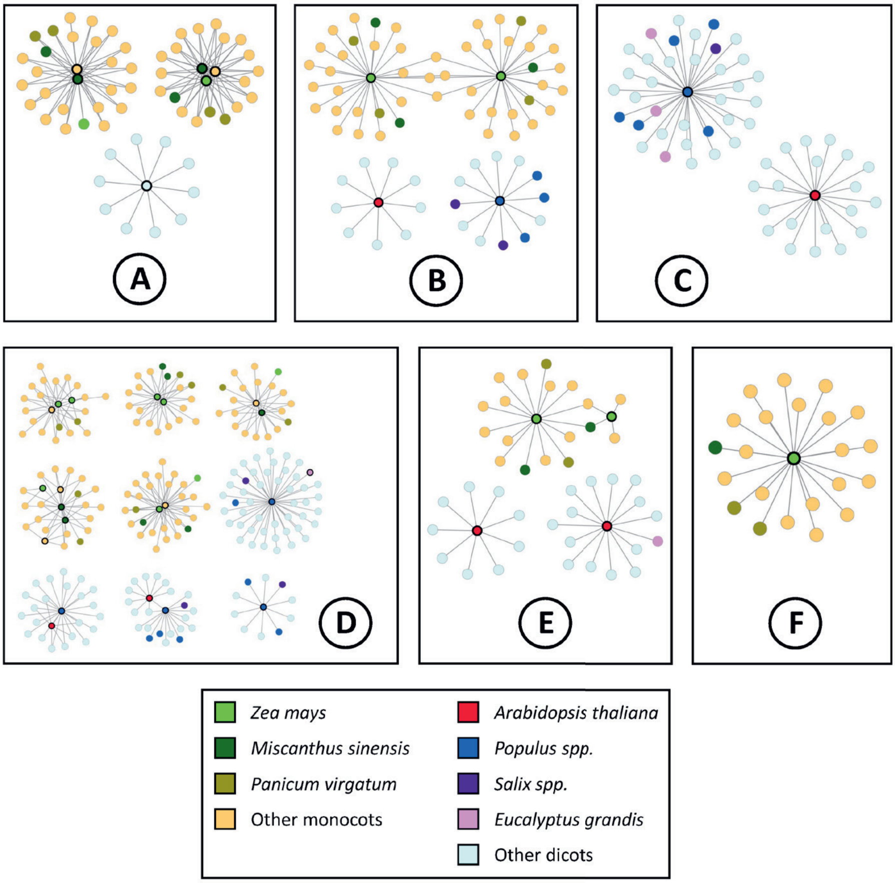


Figure 7. The structure of the syntenic communities of some of the gene families included in initial cell wall QTLs and overall highly conserved through SQTLs. Each independent network represents a gene community conserved across a set of species. Network nodes indicate the gene members of syntenic communities from different species (see legend). Network nodes with black bold edges indicate genes contained within initial cell wall QTLs. A) *Xylem NAC domain* (*XND*); B) *BEL1-like Homeodomain 9* (*BLH9*); C) *Knotted1-like homeobox gene 3* (*KNAT3*); D) *PARVUS*; E) *Ferulate 5-hydroxylase* (*F5H*); F) *STELLO* (*STL*).

Most and least conserved genes through SQTLs: overall patterns

A final functional analysis entailed the evaluation of the degree of conservation through SQTLs of all the candidate cell wall genes included within the 594 initial QTLs spanning at least one gene and used for SQTLs detection. The aim was to identify cell wall gene classes displaying relevant changes in their relative abundance in SQTLs compared to initial QTLs. In this way, we identified 21 functions that can be considered “poorly conserved” (they are much more abundant in QTLs compared to SQTLs) and 27 functions that can be defined “highly conserved” (their abundance displays little or no change in SQTLs compared to initial QTLs, meaning that the majority of their members from initial QTLs are conserved through SQTLs). **Table 4** displays these results. Interestingly, out of the 27 highly conserved functions, 15 are different types of TFs that play key roles in the regulation of cell wall quality in plants. These include *Xylem NAC Domain* (*XND*) TFs, different *NST* members, *WRKY12*, *BLH9*, *KNAT3*, *MYB46*/*52*/*54*/*83*/*58*/*63*, *ERF* genes, and *WND* TFs. In addition, *F5H* lignin genes from initial cell wall QTLs are also highly conserved in SQTLs. Finally, other highly conserved genes include different hemicellulose and pectin genes (*PARVUS* and *IRX8*), and the *STELLO* proteins. Concerning the group of "poorly conserved" genes, they include several lignin-related functions (*BGLU*, *COMT*, *DIR*, *G4H*, *CAD*, *FMT*). Moreover, three main cell wall TFs resulted also included in this group: *E2FC*, *SND2*, and *SND3*.

The genes that were shown to be highly conserved through SQTLs were also analysed in terms of genomic patterns of their syntenic conservation through SQTLs. In this sense, their syntenic communities from SQTLs revealed that these genes are extensively positionally conserved across diverse species, both in monocot and dicot plants (**Figure 8**). Moreover, several members of those syntenic communities are genes that were initially included in cell wall QTLs from different species (**Figure 8**). Finally, multiple syntenic communities corresponding to different "genomic contexts" in which different members of the gene families above are located were often identified for specific gene groups. For example, the four miscanthus *XND* genes of **Figure 8A** are grouped into two different syntenic communities. Alternatively, the eight maize copies of the *PARVUS* genes of **Figure 8D** are divided into five different syntenic communities. This highlights the existence of divergent (conserved) genomic contexts for different gene members of target gene families, which may be revelatory of different evolutionary trajectories or of functional diversification and subfunctionalization within those families. As already mentioned in Section 3.3.1, the distinct genomic contexts for different members of specific gene families highlights once again that SQTLs can be used to identify exact positional orthologs of specific gene copies across a wide range of species.

Table 4 – The cell wall related gene classes from the initial cell wall QTLs whose conservation through SQTLs resulted highest and lowest. *CS: Cellulose synthesis; TF: Transcription factor; HM: Hemicellulose metabolism; PM: Pectin metabolism; LM: Lignin metabolism; CaS: Callose synthesis; CSS: Sugar supply for cellulose synthesis. **CN: Copy Number.

Gene function	Cell wall process*	SQTL conservation	CN** in cell wall QTLs	CN** in SQTLs	CN** decrement SQTLs/QTLs (%)
STL	CS	Highly conserved	1	1	0
BLH9	TF	Highly conserved	4	4	0
KNAT3	TF	Highly conserved	2	2	0
MYB46	TF	Highly conserved	4	4	0

Table 4 *(continued)*

Gene function	Cell wall process*	SQTL conservation	CN** in cell wall QTLs	CN** in SQTLs	CN** decrement SQTLs/QTLs (%)
MYB52	TF	Highly conserved	5	5	0
MYB54	TF	Highly conserved	5	5	0
MYB83	TF	Highly conserved	4	4	0
CESA (II)	CS	Highly conserved	10	9	10
UAfT	HM	Highly conserved	10	9	10
UXT	HM	Highly conserved	10	9	10
UUAT	HM	Highly conserved	19	17	11
NST1	TF	Highly conserved	8	7	13
NST2	TF	Highly conserved	8	7	13
NST3	TF	Highly conserved	8	7	13
WRKY12	TF	Highly conserved	8	7	13
GATL	PM	Highly conserved	14	12	14
RHM	PM	Highly conserved	14	12	14
XND1	TF	Highly conserved	7	6	14
UGE	HM	Highly conserved	12	10	17
MYB58	TF	Highly conserved	6	5	17
MYB63	TF	Highly conserved	6	5	17
UGP	CS	Highly conserved	5	4	20
F5H	LM	Highly conserved	5	4	20
WND	TF	Highly conserved	39	31	21
ERF	TF	Highly conserved	34	27	21
IRX8	HM	Highly conserved	54	42	22
QUA1	PM	Highly conserved	54	42	22
PARVUS	HM	Highly conserved	44	34	23
PGI	PM	Poorly conserved	72	21	71
FMT	LM	Poorly conserved	14	4	71
UGD	HM	Poorly conserved	4	1	75
SND2	TF	Poorly conserved	4	1	75
SND3	TF	Poorly conserved	4	1	75
CAD	LM	Poorly conserved	52	11	79
C4H	LM	Poorly conserved	5	1	80

***Table 4** (continued)*

Gene function	Cell wall process*	SQTL conservation	CN** in cell wall QTLs	CN** in SQTLs	CN** decrement SQTLs/QTLs (%)
DIR	LM	Poorly conserved	44	8	82
UGT75B1	CaS	Poorly conserved	12	2	83
SPS	CSS	Poorly conserved	6	1	83
AGAL	Other	Poorly conserved	6	1	83
XYN	Other	Poorly conserved	12	2	83
COMT	LM	Poorly conserved	19	3	84
BGLU45	LM	Poorly conserved	20	3	85
KTN1	Other	Poorly conserved	2	0	100
MANT	Other	Poorly conserved	1	0	100
PNT	Other	Poorly conserved	1	0	100
AXY9	HM	Poorly conserved	1	0	100
MAN	HM	Poorly conserved	4	0	100
E2FC	TF	Poorly conserved	2	0	100
MYB75	TF	Poorly conserved	3	0	100

4 Discussion

The major goal of this study was to develop novel tools to translate genomic regions known to control biomass (cell wall) quality from model species to a large set of (orphan) crops. This goal was reached by setting up a strategy for inter-species QTLs projection based on gene synteny, which led to the detection of a large number of SQTLs within different groups of plant species. The detected SQTLs turned out to be extensive across species and to span large portions of the initial cell wall QTLs. Moreover, they highlighted key genes for cell wall quality variation that are positionally conserved across model and orphan species. On the one hand, these results indicate the validity of the approach followed in this research for genomically translating genetic information on traits of interest at a large-scale level and in an effective manner. On the other hand, they also open important considerations on the SQTLs found, with implications for the methodology developed in this research, cell wall genomics, and breeding of (novel) biomass crops. In this section, all these aspects are detailly discussed.

4.1 A novel strategy for translating (cell wall) QTLs across diverse plant species

A major delivery of this research is the development of a strategy to project QTL regions between plant species by using gene synteny. Specifically, our pipeline detected 362 SQTL regions that span a total of 74 different angiosperm species, starting from a set of 610 cell wall QTLs previously mapped in eight plant species. The detailed analysis of SQTLs properties showed that SQTLs are relatively large in size, are conserved across numerous and diverse plant species (including both monocots and dicots), and overlap with relatively large portions of the initial QTLs used for their detection (see Section 3.2). All together, these observations indicate that synteny can be successfully used to project QTLs between species in a “translational genomics” framework, and that, because of their high conservation, SQTLs may resemble the functionality of the initial QTLs. In addition, our strategy for SQTL detection – based on network analysis of QTL synteny – is easily scalable to even larger sets of genomes and QTLs, as well as to other traits than the ones used in this study. Therefore, we believe that the approach developed in this research retains high potential for translational genomics also in other contexts than cell wall research.

Nevertheless, since our main focus was on cell wall quality and biomass crops, it is noteworthy that the 74 species included in the detected SQTLs are of high interest for this field of research. In fact, they include important (orphan) biomass crops, like *Miscanthus sinensis*, *Panicum virgatum*, *Eucalyptus grandis*, *Populus trichocarpa*, and *Salix purpurea*; relevant fiber (orphan) species, like different cotton species and *Cannabis sativa*; and some important general bio-based crops for which cell wall quality is relevant for the extraction of bioresources, as *Beta vulgaris*, *Lupinus angustifolius*, *Chenopodium quinoa*, and *Camelina sativa*. Moreover, some vegetable and fruit crops for which cell wall composition is a major determinant of the quality of their food parts are also included in SQTLs, such as *Actinidia chinensis*, *Malus domestica*, *Prunus persica*, and *Solanum lycopersicum*. In all these crops, the regions spanned by SQTLs and the candidate genes therein represent relevant targets for further reverse genetics studies and/or breeding programs. Specifically, some of the species above are currently lacking genetic resources for the improvement of their biomass composition, and the availability of candidate genomic regions and genes from SQTLs could significantly speed up pre-breeding research efforts.

As reported in Section 3.2, the majority of the SQTLs was detected in the “Poaceae” group of **Table 1**, which is a notoriously highly syntenic group of plants (Bennetzen and Freeling, 1997). Therefore, the overall synteny of target species on which the initial QTLs are projected appears to be a critical parameter to successfully apply the strategy developed in this study. Moreover, our results highlighted that the overall synteny of the QTLs and of the candidate genes therein is also pivotal to successfully detect SQTLs, and needs to be evaluated beforehand. In the case of cell wall research, the fact that cell wall genes and QTLs turned out to be overall highly syntenic across all angiosperms (Section 3.1) and that Poaceae species include the majority of biomass crops certainly had a positive influence on SQTLs identification. Nevertheless, our results demonstrated that genomic QTL translation can be successfully achieved also in several groups of eudicot species, such as the ones reported in **Table 1**. In this case, the level of inter-species synteny appeared more stringent in determining the level of taxonomic distance from species for which initial QTLs were available until which SQTLs identification is possible. Nevertheless, SQTLs were successfully detected across all the eudicot groups of **Table 1**, highlighting the feasibility of this approach also in these plants. Moreover, the fragmentation of initial QTLs over multiple SQTLs was minimized in smaller groups of eudicots used for SQTLs detection (e.g. Brassicaceae), while SQTLs size was maximized. Furthermore, the fact that cell wall QTLs from poplar and eucalyptus were relatively easily translated across several species (thanks to the high level of synteny found for these species across several eudicot families) is very promising for biomass and cell wall research in eudicots and, specifically, biomass trees. In fact, the improvement of perennial crops such as trees is notoriously time consuming given the long breeding cycles (Clifton-Brown et al., 2018), and translational genomics through genome synteny may therefore significantly speed up pre-breeding research efforts in these species.

To conclude, the analysis of the cell wall gene classes mostly represented within SQTLs also highlighted important considerations on the strategy developed in this study. Specifically, SQTLs turned out to be enriched in cell wall TFs, and several of these TFs belong to the highest layers of cell wall regulation, highlighting the overall relevance of the regions spanned by SQTLs for the regulation of cell wall synthesis, and therefore for controlling variability in cell wall composition. In parallel, the abundant occurrence within SQTLs of genes known to display allelic and mutational variation associated with cell wall composition also indicates the relevance of SQTL regions for controlling cell wall quality. Overall, the general patterns observed in candidate cell wall genes highlight the validity of using gene synteny to project QTLs between species in a meaningful way from the point of view of the genetic architecture of traits.

4.2 Conserved determinants of cell wall variability as revealed by SQTLs

The functional analyses conducted on SQTLs (Section 3.3) highlighted both the presence of critical cell wall candidate genes conserved through SQTLs across large sets of species, as well as classes of genes represented within the initial cell wall QTLs but poorly conserved through SQTLs. On the one hand, the candidate cell wall genes that are highly conserved through SQTLs may represent interesting targets for setting up "universal" approaches to improve biomass crops. Moreover, they also give insights on what can be considered "universal" across sets of plants in terms of the genomic architecture of the trait "cell wall quality". On the other hand, the combined analysis of conserved and non-conserved candidate cell wall genes through SQTLs open interesting considerations on the most effective mechanisms to manipulate cell wall composition.

A first result of the functional analyses of SQTLs is the consistent occurrence of cell wall related TFs within SQTLs (see Section 3.3.1 and 3.3.3). Since TFs are important players in the regulation of plant traits, including plant cell wall (Zhang et al., 2018, Rao and Dixon, 2018), and are often causative genes at the basis of QTLs (Barrière et al., 2012, Courtial et al., 2014), their consistent conservation through SQTLs highlights the value of these tools to pinpoint relevant conserved candidate genes across diverse species. The analysis of the cell wall TFs included in SQTLs revealed that they include master regulators of cell wall biosynthesis, such as the *VND* and *NAC* TFs. These TFs regulate the global deposition of secondary cell walls in plant vessels and fibers, respectively, and are able to bind to critical structural genes at the basis of cellulose, xylan and lignin biosynthesis (Zhong and Ye, 2015, Taylor-Teeples et al., 2015). In addition, their functionality is hypothesized to be largely conserved across diverse plant species (Nakano et al., 2015, Zhong and Ye, 2015), and genetic modifications of these genes resulted in plant phenotypes with altered cell wall composition and quality, including improvement of biomass saccharification (Iwase et al., 2009, Yoshida et al., 2013). All together, this evidence highlights the relevance of *VND* and *NAC* TFs for the improvement of cell wall composition in plants. In this context, SQTLs can be used to readily detect sets of *VND* and *NST* orthologs located on conserved genomic contexts across species. In addition, SQTLs could also be used to discriminate between the different copies of these genes when mining the exact positional orthologs of specific gene members with a higher functional relevance, to maximize the likelihood of a complete gene functional conservation (Dewey, 2011). For example, it has been demonstrated that out of the seven *VND* copies of arabidopsis, one – *AtVND7* – is the major player of the *AtVND* family, by acting as the transcriptional terminus of these genes and by impacting cell wall deposition the most (Yamaguchi et al., 2011, Endo et al., 2015). Interestingly, *AtVND7* is the *AtVND* member that resulted conserved through SQTLs (red dot circled with bold border in **Figure 6A**). The other dicot genes of **Figure 6A** syntenically connected to *AtVND7* represent therefore the exact positional orthologs of *AtVND7* – and not of the other *AtVND* copies – in other species. Therefore, SQTLs can discriminate between the different members of critical gene families when defining gene targets for plant research.

In addition to *VND* and *NAC* TFs, *WRKY12* is another master regulator of cell wall biosynthesis that acts as repressor of lignin deposition (Wang et al., 2010a) and that was largely conserved through SQTLs across both monocots and dicots. Mutations at *WRKY12* heavily affect the relative content of lignin, cellulose, and hemicellulose, as well as the production of total stem biomass in arabidopsis (Wang et al., 2010a). Moreover, a *WRKY12* gene from *Miscanthus lutarioriparius* was shown to promote flowering when inserted in arabidopsis (Yu et al., 2013). This information highlights once again that SQTLs can pinpoint relevant genes for improving biomass crops, that thanks to syntenic conservation can be easily mapped across the different species of the study. Moreover, because of its properties, WRKY12 could specifically represent an attractive target for the parallel modification of cell wall quality, biomass production, and flowering time.

Other TFs that were highly conserved through SQTLs include *OFP4*, *ERFs*, *BLH9*, *KNAT3* and several *MYB* genes. These genes are all important regulators of cell wall deposition across several species, even if their role is less central than the one of the *VND*, *NST*, and *WRKY12* TFs (Zhong and Ye, 2015). Nevertheless, their functional redundancy with other TFs (Taylor-Teeples et al., 2015) may underly a higher chance of finding useful allelic variation at the loci coding for these genes in target crops, as the selection pressure exerted on these genes might have been relatively relaxed. For all these genes, SQTLs may again be used to readily identify homologs laying in conserved genomic contexts across the species of this study, including monocot and dicot biomass species (miscanthus, switchgrass, poplar, willow) across which the genomic organization of these genes appeared extensively conserved. In species of interest, allelic variation at the loci mapped through SQTLs may then be studied with novel methodologies for targeted sequencing (Scaglione et al., 2019), eventually leading to the detection of favourable germplasm sources to be used in breeding programs.

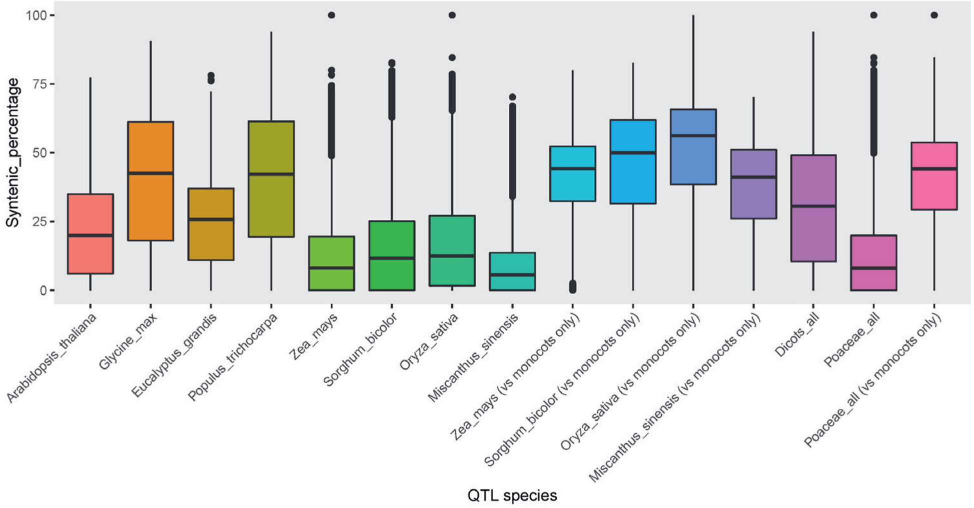
In addition to TFs, SQTLs contained a large amount of genes involved in the substitution and/or remodelling of cell wall polymers. This class of genes is also of high importance for the improvement of biomass crops, as both the degree of substitution of cell wall polymers with a variety of chemical moieties and the re-building of cell wall polymers during cell wall metabolism are pivotal processes to determine the amenity of plant cell walls to deconstruction (van der Weijde et al., 2013, Torres et al., 2015b). In this perspective, SQTLs may be used to identify critical candidate genes in crops with scarce genetic resources thanks to positional orthology with key genes included in cell wall QTLs in model species. Specifically, some of the genes involved in the remodelling of cell wall polymers and conserved through SQTLs are typically found in single or low copy-number in plant genomes, including *PARVUS* and *ESK*. Both these genes contribute to xylans synthesis, with *PARVUS* being likely involved in the synthesis of the xylans reducing ends (which in turn are presumably primers for total xylan synthesis) (York and O'Neill, 2008), and *ESK* being involved in xylans mono-acetylation (Yuan et al., 2013). Even if the precise functioning of these genes is far from being understood (Smith et al., 2017), the traits on which they are presumably involved – xylans amount and xylans substitutions – are preeminent targets for improvement of biomass crops (van der Weijde et al., 2013, Torres et al., 2015b). Therefore, the members of the *PARVUS* and *ESK* gene families retained in SQTLs could certainly represent interesting targets for further reverse genetic studies, also because of the presence of some of their members within the initial cell wall QTLs used for SQTLs detection. In contrast to *PARVUS* and *ESK*, other genes involved in the remodelling of cell wall molecules and highly conserved through SQTLs belong to large families, like the *BAHD*, *BXL*, *RWA*, *EXT*, and *PG*/*PL* genes. These genes perform different functions within plant cell walls (see Zhong et al. (2019) for a review), but have all been indicated as candidate genes for modifying biomass quality by changing the content and biochemical properties of cell wall molecules (Bartley et al., 2013, Biswal et al., 2014, Pawar et al., 2017). Interestingly, for several of these gene families from model species it has been shown that different members can perform different functions, or can exert their functions in different plant organs or developmental stages (Nakhamchik et al., 2004, Tuominen et al., 2011, Cao, 2012). In this sense, as previously discussed for *VND* TFs, SQTLs may again both allow the quick identification of conserved positional orthologs across diverse species and help to discriminate between multiple family members to decide which copies to target in plant research or plant breeding.

To conclude, in addition to the gene classes mentioned so far, the analysis of SQTLs highlighted the presence of gene families that, despite being represented in the initial cell wall QTLs used for SQTLs detection, revealed to be poorly conserved through SQTLs (section 3.3.3). Interestingly, several lignin structural genes involved in the pathway leading to monolignol synthesis were included in this category. This observation, together with the fact that several lignin TFs were instead found highly conserved through SQTLs, suggests that targeting TFs may be a more successful and more interapplicable strategy for modifying lignin across different species compared to the targeting of lignin structural genes. In this context, the *ferulate 5-hydroxylase* gene (*F5H*) represents an exception, as it was found highly conserved through SQTLs (**Figure 8**). The genetic manipulation of this gene is known to alter monolignol ratios and can substantially improve biomass saccharification in several species (Stewart et al., 2009, Weng et al., 2010). Therefore, the copies included in SQTLs may represent interesting breeding targets. In addition, the extensive positional conservation of *F5H* across several monocot and dicot species suggests that targeting this gene may represent a “universal” approach to biomass improvement.

5 Conclusions

The present study is the first research, to our knowledge, to develop a successful strategy to project a set of (cell wall) QTLs across a large set of species in a translational genomics framework and through the use of gene synteny. The approach developed in this study represents a novel tool to assist breeding of (orphan) lignocellulosic biomass crops, and can potentially be applied also to other sets of species and traits than the ones used here. The functional analysis of SQTLs demonstrated that those regions retain conserved critical genes for cell wall quality – as *VND*, *NAC* and *WRKY12* transcription factors, *PARVUS*, *RWA*, or *ESK* genes involved in cell wall remodelling, and several *F5H* copies – which could represent targets for “universal” approaches for biomass improvement. In this sense, future research efforts may be directed to evaluate the allelic variation of SQTL regions across diverse species and to further validate the relevance of the candidate genes found through reverse genetics.

Supplementary Data



Supplementary Figure 1.
The distributions of synteny levels (as percentage of genes of certain species syntenic with all the other genes from all the other species) for relevant species/families inspected for determining plant groups to be used for SQTLs detection (**Table 1**).

The following supplementary data can be accessed online at https://www.frontiersin.org/articles/10.3389/fpls.2022.855093/full#supplementary-material:

Supplementary Table 1: The list of 610 cell wall related QTLs retreived through literature search.

Supplementary Table 2: The 151 angiosperm genomes used in the study.

Supplementary Table 3: The list of generic gene functions known to play a role in plant cell walls retrieved from scientific literature.

Supplementary Table 4: The candidate cell wall genes identified in the 151 genomes of the study.

Supplementary Table 5: The syntenic cell wall QTLs network.

Supplementary Table 6: The 50 QTLs spanning only one candidate cell wall gene.

Supplementary Table 7: The 362 syntenic cell wall QTLs and all the genes included within them.

Supplementary Table 8: The 1493 candidate cell wall genes included in the 22 SQTLs showing large co-localization of initial QTLs and spanning diverse species.

Supplementary Table 9: The 139 genes known to display allelic/mutational variation associated with variation in cell wall quality that were collected in maize, rice, sorghum, arabidopsis, and poplar.

Chapter 6

Syntenic cell wall QTLs as versatile breeding tools: intraspecific allelic variability and predictability of biomass quality loci in target plant species

Francesco Pancaldi[1], Eibertus N. van Loo[1], Sylwia Senio[1], Mohamad Al Hassan[1], Kasper van der Cruijsen[1], Maria-João Paulo[2], Oene Dolstra[1], M. Eric Schranz[3], Luisa M. Trindade[1]

[1]Plant Breeding, Wageningen University & Research, Wageningen, The Netherlands

[2]Biometris, Wageningen University & Research, Wageningen, The Netherlands

[3]Biosystematics, Wageningen University & Research, Wageningen, The Netherlands

Published as: Pancaldi, F., van Loo, E. N., Senio, S., Al Hassan, M., van der Cruijsen, K., Paulo, M. J., Dolstra, O., Schranz, M. E. and Trindade, L.M. (2023). Syntenic cell wall QTLs as versatile breeding tools: intraspecific allelic variability and predictability of biomass quality loci in target plant species. *Plants*, *12*(4).
DOI: 10.3390/plants12040779

Abstract

Syntenic cell wall QTLs (SQTLs) can identify genetic determinants of biomass traits in understudied species based on results from model crops. However, their effective use in plant breeding requires SQTLs to display intra-specific allelic variability and to predict causative loci in other populations/species than the ones used for SQTLs identification. In this study, genome assemblies from different accessions of arabidopsis, rapeseed, tomato, rice, brachypodium and maize were used to evaluate intra-specific variability of SQTLs. In parallel, a genome-wide association study (GWAS) on cell wall quality traits was performed in miscanthus to verify co-localization between GWAS loci and miscanthus SQTLs. Finally, an analogous approach was applied on a set of switchgrass cell wall QTLs retrieved from literature. These analyses revealed large SQTLs intra-specific genetic variability, ranging from presence-absence gene variation, to SNPs/INDELs and changes in coded proteins. Cell wall genes displaying gene dosage regulation, as PAL and CAD, displayed presence-absence variation in brachypodium and rapeseed, while protein INDELs were detected for the brachypodium homologs of the rice Brittle-culm-like 8 locus, which might likely impact cell wall quality. Furthermore, SQTLs significantly co-localized with the miscanthus and switchgrass QTLs, with relevant cell wall genes being retained in co-localizing regions. Overall, SQTLs are useful tools to screen germplasm for relevant genes and alleles to improve biomass quality, and can increase the efficiency of plant breeding in understudied biomass crops.

1 Introduction

The projection of known quantitative trait loci (QTLs) across species through genome synteny was recently shown to allow the quick identification of conserved genomic regions underlying traits of interest in large panels of plants, including novel, under-domesticated, crops (Pancaldi et al., 2022b). On the one hand, this is possible thanks to the established relationship between the occurrence of gene synteny (i.e. the conservation of gene presence and gene order across genomes) and the conservation of gene function across diverse living organisms (Dewey, 2011, Zhao and Schranz, 2017, Kerstens et al., 2020). On the other hand, the co-localization of syntenic regions with previously mapped QTLs ensures the relevance of the regions identified for the involvement in the trait of interest, since the occurrence of QTLs proves that genomic regions are causative of trait variability in particular species. The genomic regions identified across several species by combining synteny and QTLs information can be referred to as syntenic quantitative trait loci (SQTLs) (Pancaldi et al., 2022b).

The concept of SQTLs was demonstrated for plant cell wall compositional traits (Pancaldi et al., 2022b). These traits entail the relative amounts and the chemical-physical properties of the polysaccharides that constitute plant cell walls and underly feedstock quality of biomass crops – mainly cellulose, hemicelluloses, pectins, and lignin – (van der Weijde et al., 2013, Pancaldi and Trindade, 2020). The generally high syntenic conservation of cell wall genes within previously mapped cell wall QTLs allowed for the identification of numerous SQTLs, highlighting the potential of this strategy for projecting important conserved loci underlying biomass quality across multiple plants (Pancaldi et al., 2022b). Furthermore, cell wall SQTLs were shown to allow for a "fine-mapping" of the initial cell wall QTLs, by assessing the overlap of multiple initial QTLs on specific syntenic segments across multiple genomes (Pancaldi et al., 2022b). Finally, cell wall SQTLs contain relevant cell wall genes, known from previous studies to affect cell wall quality in different species, and therefore likely representing (some of) the conserved causative genes of the initial cell wall QTLs (Pancaldi et al., 2022b).

What just discussed demonstrates that SQTLs represent valuable tools during pre-breeding steps of crop improvement. Specifically, their availability, combined with the dropping of genome sequencing costs, allows to potentially circumvent (part of) the pre-breeding studies on trait genetics needed to start breeding programs in under-domesticated crops (Pancaldi et al., 2022b). Given the wealth of genetic resources represented by under-domesticated plant species (Tadele, 2019), the availability of tools to speed up their improvement is an extremely important asset for agriculture (Salentijn et al., 2007, Armstead et al., 2009, Kamei et al., 2016). As an example, the improvement of biomass crops – which are all the plant species that can produce biomass to sustain bio-based value-chains (Trindade et al., 2010, Mehmood et al., 2017) – could greatly benefit from this prospect. In fact, they include several under-domesticated species (see Mehmood et al. (2017) and Pancaldi and Trindade (2020) for a comprehensive list), whose breeding cycles are highly time-consuming (Clifton-Brown et al., 2018), while several complex plant traits should be improved in these species to allow their cultivation on marginal lands to avoid competition with food production (Pancaldi and Trindade, 2020). Nonetheless, the use of SQTLs in breeding contexts requires the availability of intra-specific allelic variability with a potential impact on traits of interest for the SQTL regions themselves, as genetic variability is the prerequisite for selection. Moreover, SQTLs must be able to predict the localization of causative loci of the traits for which they were mapped for also in other plant populations or species than the ones used for SQTLs identification.

The two latter aspects are the focus of this study. On the one hand, the intra-specific allelic genomic variability of cell wall SQTLs was assessed in six angiosperm species – *Arabidopsis thaliana*, *Solanum lycopersicum*, *Brassica napus*, *Zea mays*, *Oryza sativa*, and *Brachypodium distachyon* – for which multiple genomes representing diverse plant accessions were available. On the other hand, the SQTLs value for predicting relevant loci in novel populations/species was tested by assessing the co-localization between cell wall SQTLs from *Miscanthus sinensis* and *Panicum virgatum* and genomic regions associated with cell wall variability identified through association mapping in miscanthus and switchgrass, respectively. Miscanthus and switchgrass are C4 biomass crops with high potential to fulfil diverse industrial applications, but at the same time still largely under-domesticated (van der Weijde et al., 2013, Clifton-Brown et al., 2018). The study of the genetics underlying cell wall composition in these crops is thus pivotal to bring them out of their current state (Pancaldi and Trindade, 2020, Van der Cruijsen et al., 2021). The results obtained in this study demonstrate the validity of SQTLs for this purpose.

2 Materials and methods

2.1 Intra-specific SQTL alignment

Multiple chromosome-level genome assemblies representing different accessions of *Arabidopsis thaliana*, *Brassica napus*, *Solanum lycopersicon*, *Brachypodium distachyon*, *Oryza sativa*, and *Zea mays* have been collected from online databases after literature search (**Supplementary Table 8**). These six species were chosen because of the availability of multiple good-quality chromosome-level genome assemblies from either pan-genomic studies (Gordon et al., 2017, Hurgobin et al., 2018, Jiao and Schneeberger, 2020) or genomic databases, as well as their relevance for plant research. Moreover, they include both eudicots and grasses for which SQTLs were available from our previous study (Pancaldi et al., 2022b). The reference assembly of each of the six species above on which SQTLs were initially detected was used to extract the reference nucleotide sequences of cell wall SQTLs. These sequences were then aligned against the accessions collected by using the NUCmer package of the MUMmer software (Delcher et al., 2002, Kurtz et al., 2004). NUCmer was run with the following parameters: --minmatch 100 and --mincluster 200. The NUCmer show-snps command was also called (default parameters) to detect SNPs and INDELs between reference SQTL sequences and aligning regions on target accessions.

2.2 Analysis of intra-specific SQTL alignments

Custom R scripts were developed to process the NUCmer outputs and extract different information. First, the aligning regions and coordinates of SQTLs in every target accession. Second, SQTLs coverage for each alignment. Third, the SQTL (cell wall) genes included in each alignment. Finally, the SNPs and INDELs found along alignments. To extract these data, the custom R scripts made also use of a previously-developed list of cell wall genes from 169 angiosperm genomes (Pancaldi et al., 2022b), as well as of the IRanges and GenomicRanges R packages (Lawrence et al., 2013). Moreover, to quantify gene PAV in SQTLs alignments, the data on missing genes in alignment produced by NUCmer were validated by a BLAST search (Altschul et al., 1990) of the genes identified as missing in specific alignments against the assemblies involved in the alignments themselves (Evalue = 1E-3). Finally, all the statistical analyses reported in this manuscript have been performed in R or SPSS v27.0 (IBM Corp, Armonk, NY).

2.3 Analysis of cell wall protein sequence changes

The assessment of the effects of SQTLs genomic variability on protein sequences was performed by using a custom R script were the coordinates of the exons of every SQTL gene included in alignments and concurring to code the main gene transcripts were used to retrieve gene CDS and translated proteins in the target genome assemblies. On the retrieved proteins, sequence changes between reference and target assemblies were assessed with ClustalW (Chenna et al., 2003). Moreover, HMMsearch (Eddy, 2011) was used to annotate protein functional domains available on the PFAM database (El-Gebali et al., 2019). Furthermore, protein signal peptides, including N- and C-terminus and related functional signals were annotated by integrating the predictions provided by SignalP v6.0 (Teufel et al., 2022), DeepTMHMM (Hallgren et al., 2022), NetGPI (Gíslason et al., 2021) and DeepLoc v2.0 (Thumuluri et al., 2022). Finally, changes in protein structure and properties due to sequence changes were assessed by using NetSurfP v3.0 (Høie et al., 2022).

2.4 Genome-wide association study on *Miscanthus sinensis*

The GWAS on *Miscanthus sinensis* was performed by using a miscanthus collection established in 2013 in Wageningen (The Netherlands) and composed of 94 accessions originated from various international gene banks and breeding programs. Genotypes were planted in square-like plots with 16 clonal replicas, and the four central plants of each plot were harvested every spring for five years, starting in 2017.

Phenotyping was performed for eight cell wall quality traits, including NDF cell wall as percentage of dry matter, ADF cell wall as percentage of dry matter, ADL lignin as percentage of dry matter, Lignin as percentage of NDF, Cellulose as percentage of dry matter, Cellulose as percentage of NDF, Hemicellulose as percentage of dry matter, and Hemicellulose as percentage of NDF (**Supplementary Table 9**). Phenotyping was performed by first chopping the harvested miscanthus stems into pieces of 4 cm, and by drying (60°C, 48h) and weighting stem pieces to determine dry matter content. The dried stems were then milled, and a subset of samples was used for training a Near Infrared Spectrometry (NIRS) model for cell wall composition, by performing NDF, ADF, and ADL biochemical analyses following the ANKOM Technology protocols (ANKOM Technology Corporation). The NIRS model was in turn used to phenotype the cell wall traits on the milled feedstock of all the genotypes of the collection.

The genomic DNA of all the accessions was isolated from random young leaves from the four central plants of all the plots in the collection, following a CTAB-based protocol (Tai and Tanksley, 1990). Extracted DNA from every sample was digested by using the restriction enzyme EcoR1, ligated to unique adapters, pooled, purified, amplified, and finally sequenced using an Illumina HiSeq X10/4000 system. Sequencing was performed by BGI (Shenzen, Guangdong, China), and generated 371.52 Gb of cleaned data. Reads were aligned to the *Miscanthus sinensis* reference genome (Mitros et al., 2020), resulting in the identification of ~7.0 million SNPs. SNPs were filtered for only biallelic SNPs displaying 100% call rate across the 94 miscanthus accessions and a minor allele frequency >20%. Moreover, following chromosome-wide LD analysis with the LD.decay function from R package Sommer (Covarrubias-Pazaran, 2016) (**Supplementary Figure 5**), SNPs were further filtered to keep only one marker for genomic bins corresponding to one third of the average chromosomal LD distance. This way, a set of 57891 markers relatively evenly distributed over the 19 miscanthus chromosomes was obtained.

The filtered SNPs were used to estimate population structure by using Van Raden kinship (VanRaden, 2008) and Principal Coordinate Analysis (PCoA; ape R package – Paradis and Schliep (2019)) (**Supplementary Figure 6**). A dendrogram of the kinship matrix was also produced (ape R package – Paradis and Schliep (2019)). Moreover, patterns of population structure among accessions were compared with the ones inferred from a principal component analysis (PCA) on the phenotypic data (**Supplementary Figure 7**), to assess co-variation between population and phenotypic accession clusters (**Supplementary Figure 8**).

Genome-wide associations between the filtered SNPs and the eight cell wall traits above were performed by using a Linear Mixed Model (LMM) incorporating SNP data and the kinship matrix, as implemented in the statgenGWAS R package (van Rossum et al., 2020). FDR correction (1%) was used to account for multiple testing (Benjamini and Hochberg, 1995), while QQ-plots of observed vs expected p-values of associations were computed to assess the effectiveness of population structure correction (**Supplementary Figure 9**). GWAS analyses were performed separately for each trait. Chromosome-scale LD windows were used to define significant regions around the significant markers found, within which candidate genes were looked for, by using the set of angiosperm cell wall genes developed in our previous SQTL study (Pancaldi et al., 2022b) and the arabidopsis- and rice-based annotations of the miscanthus genes from Phytozome. Moreover, a separate GWAS was also performed without incorporating the kinship matrix in the model, to perform the study described in Paragraph 2.5.

2.5 Co-localization of SQTLs and cell wall loci mapped on *Miscanthus sinensis*

Co-localization between SQTLs and the 91 QTLs found by the miscanthus GWAS was performed by developing 100 sets of 91 random QTL regions from the miscanthus genome mirroring the size distribution of the QTLs from GWAS results (custom R script). The proportion of QTLs co-localizing for >50% of their bp length with SQTLs was then calculated for every set, and binomial tests were performed to assess if random QTLs co-localized with SQTLs significantly less than the QTLs from the GWAS (custom R script).

In addition to calculating the statistical significance of the co-localization between miscanthus SQTLs and QTLs, the cell wall genes in co-localizing regions were identified by using the set of angiosperm cell wall genes developed in our previous SQTL study (Pancaldi et al., 2022b). Moreover, angiosperm-wide syntenic conservation of those genes was analysed by retrieving their syntenic homologs from the synteny network developed by Pancaldi et al. (2022a) and Pancaldi et al. (2022b).

2.6 Retrieval of the *Panicum virgatum* cell wall QTLs and analysis of their co-localization with SQTLs

A total of 56 QTLs related to cell wall traits in switchgrass were retrieved from the results of Ali et al. (2020) (**Supplementary Table 5**). Co-localization between these QTLs and the switchgrass SQTLs was analysed with an analogous procedure to miscanthus. Specifically, 100 sets of 56 random QTL regions from the switchgrass genome mirroring the size distribution of the 56 QTLs from Ali et al. (2020) were computed. The proportion of QTLs co-localizing for >50% of their bp length with SQTLs was then calculated for every set, and binomial tests were performed to assess presence and significance of a decrement in such proportion (custom R script).

As performed in miscanthus, the cell wall genes in co-localizing regions were identified by using the set of angiosperm cell wall genes developed in our previous SQTL study (Pancaldi et al., 2022b). Moreover, angiosperm-wide syntenic conservation of those genes was analysed by retrieving their syntenic homologs from the synteny network developed by Pancaldi et al. (2022a) and Pancaldi et al. (2022b).

3 Results

3.1 General SQTLs alignment patterns across multiple accessions of six angiosperm species

To investigate intra-specific genetic variability of SQTLs, the nucleotide sequences of 1184 cell wall SQTLs previously identified in six angiosperm species (arabidopsis, rapeseed, tomato, maize, rice, and brachypodium) (Pancaldi et al., 2022b) were aligned against 111 genome assemblies representing different plant accessions of the six species themselves (**Figure 1**). On average, alignments covered 89.3% of SQTLs initial bp length across all the genome assemblies tested (CV=5.8%; **Figure 2**). **Figure 2** shows that inter-specific syntenic regions spanned by SQTLs are, as expected, overall very well conserved also at the intra-specific level. Nevertheless, it also highlights that minor SQTLs portions display presence/absence variation (PAV) patterns between reference and specific target assemblies. Specifically, from the boxplots of **Figure 2**, these patterns appear particularly relevant in the three Poaceae species and, to a minor extent, tomato, where several SQTLs have relatively large portions of their total length not represented in alignments. On the one hand, this might be due to technological reasons as the sequencing methodology of the target assemblies (e.g. de novo vs reference-based re-sequencing). On the other hand, it could also depend on local intra-specific genomic rearrangements involving SQTLs regions. This hypothesis was tested in maize, by using the annotation of the transposon regions from the reference maize genome assembly used for initial SQTLs

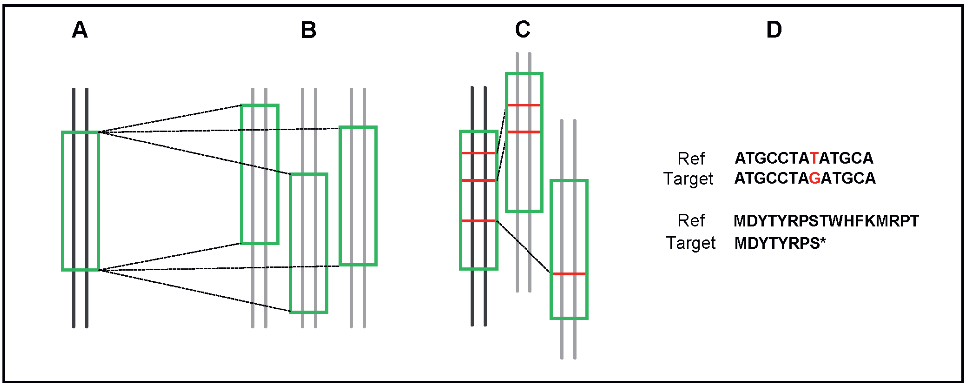


Figure 1. Schematic representation of the workflow followed to analyze the intra-specific allelic variability of SQTLs. SQTL nucleotide sequences were extracted from reference genome assemblies of six species (A) and aligned against multiple genome assemblies representing diverse accessions of each species (B). Nucleotide polymorphisms and gene presence-absence variation were quantified and analyzed across accessions and SQTLs (C). Finally, the effect of genomic poly-morphisms on gene coding sequences and protein sequences and strictures was also assessed (D) and compared with known mutations responsible of relevant biomass phenotypes.

detection (B73 version 4.0). This way, it was found out that maize SQTL regions span a large number of transposons (74 transposons per SQTL on average), which may be involved in intra-specific structural genomic variability, leading to the alignment patterns found.

Irrespectively of the source of alignment length variability, the fact that SQTLs sometimes do not entirely align to target genome assemblies might lead to PAV of SQTL genes across the assemblies tested. In this sense, a BLAST validation of the SQTL genes displaying PAV according to the outputs of the intra-specific SQTL alignments revealed that 4225 SQTL genes are missing in one or more target assemblies across all the species tested (3.3% of all the genes contained in SQTLs). The majority of these genes is absent in only one target assembly (2086 genes across all the species), while 475 genes across all the species tested are missing in >50% of the target assemblies assessed for a particular species (**Supplementary Table 1**). Finally, of all the genes displaying PAV patterns between reference and target assemblies, 178 are cell wall genes (4.2%; **Supplementary Table 1**). Among these, some appear relevant for cell wall quality variability. These include an endoglucanase-coding gene from maize homolog to arabidopsis KORRIGAN (ND_NP_001288520.1, which is absent in four maize target assemblies), five PAL genes from *Brachypodium distachyon* involved in lignin synthesis (BG_XP_003575404.1, BG_XP_003575365.1, BG_XP_003575403.1, BG_XP_003575240.2 and BG_XP_003575238.1, which are absent in 11, 4, 2, 1 and 1 assembl(ies), respectively), and two CAD genes operating the *in muro* monolignol polymerization from *Brassica napus* (BL_XP_013702441.1, BL_XP_013702446.1, both

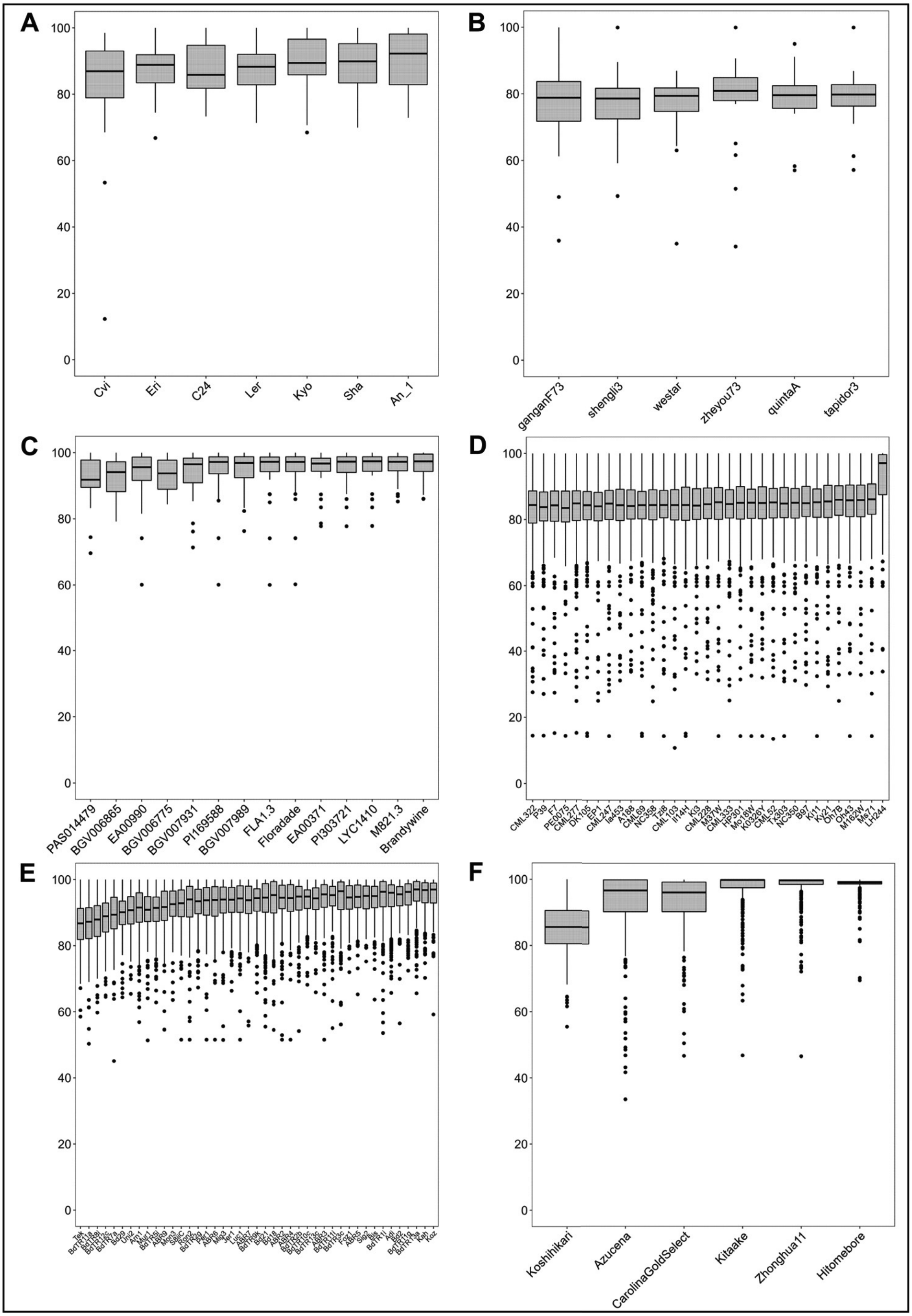


Figure 2. Boxplots representing the coverage of reference SQTL segments across the target accessions used for SQTL alignment in every species. Data points are percentages of SQTL segments contained in a target assembly (one data point per SQTL, per species). A) *Arabidopsis thaliana*; B) *Brassica napus*; C) *Solanum lycopersicum*; D) *Zea mays*; E) *Brachypodium distachyon*; F) *Oryza sativa*.

absent in one target assembly) (**Supplementary Table 1**). To conclude, in spite of the genes just mentioned, the overall relatively low level of intra-specific PAV agrees with both the close relatedness of intra-specific accessions and the high level of syntenic conservation of SQTL regions.

3.2 Large nucleotide variation within the intra-specific SQTL aligning regions

SQTL alignments were also used to quantify intra-specific nucleotide variation at SQTL regions consisting of single nucleotide polymorphisms (SNPs) and insertions/deletions (INDELs, <100 bp) between reference and target assemblies (**Figure 1**). The results revealed extensive intra-specific nucleotide variability of SQTL regions (**Figure 3**). Specifically, SQTLs displayed on average 2.2 SNPs/kbp and 0.3 INDELs/kbp across all the species and assemblies assessed. These figures correspond to an absolute mean of 2643 SNPs and 295 INDELs per SQTL, considering the average SQTL length across all species (1176 kbp).

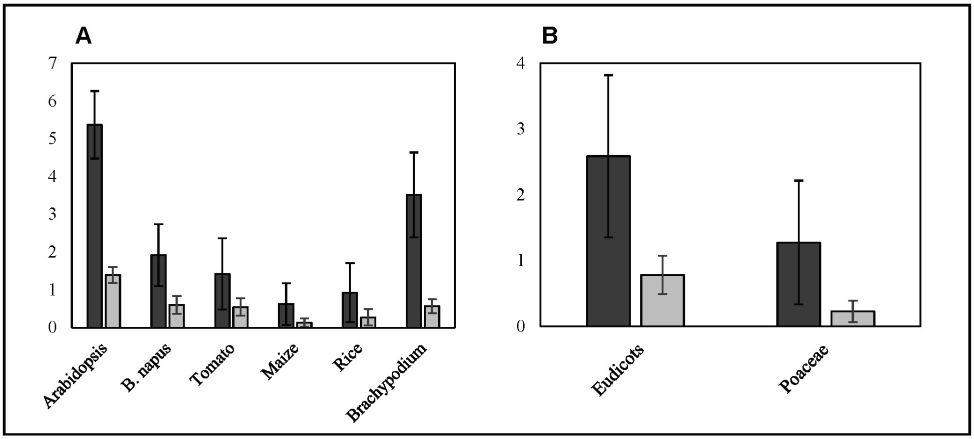


Figure 3. Average number of SNPs/kbp (dark grey) and INDELs/kbp (light grey) over SQTLs and per species (A) and plant clade (eudicots vs Poaceae) (B).

A more detailed analysis revealed that the number of SNPs/kbp and INDELs/kbp varies substantially between species (ANOVA's P=0.000; **Figure 3A**). Specifically, arabidopsis and brachypodium SQTLs displayed a particularly high number of SNPs/kbp between reference and target assemblies compared to the other species (LSD's P<0.001; **Figure 3A**). Conversely, regarding INDELs/kbp, only arabidopsis SQTLs displayed a substantially higher number of INDELs compared to the average across all SQTLs from all species (LSD's P<0.001; **Figure 3A**). Finally, both SNPs and INDELs occur with much higher frequency in dicot SQTLs (2.6 SNPs/kbp and 1.3 INDELs/kbp) compared to Poaceae SQTLs (0.8 SNPs/kbp and 0.2 INDELs/kbp; t-test's

P<0.001 for both; **Figure 3B**), in spite of the opposite trend observed for overall coverage of SQTL alignments (Paragraph 2.1).

The occurrence of SNPs and INDELs on SQTLs between reference and target assemblies was also assessed with respect to SQTL portions specifically spanned by genes. In this regard, it was found out that the majority of nucleotide polymorphisms occurs within intergenic genomic areas, as only 20.3% of SQTL SNPs and 11.5% of SQTL INDELs was located on SQTL genes (**Supplementary Table 2**). Moreover, the majority of all polymorphisms occurring on gene regions is located on non-coding gene segments, as only 44.4% of the SNPs on gene regions and 36.7% of the INDELs on gene regions were specifically located on exons. Still, in absolute terms, the latter two percentages correspond to an average of 159 SNPs and 23 INDELs per SQTL located on gene exons across all target species and assemblies, which are considerable numbers for the potential trait effects that these polymorphisms may cause. Finally, when assessed on a single-gene level, our results showed that 60207 SQTL genes displayed at least one polymorphism in one species and against one target assembly (47% of all QTL genes). Of these, 3156 are cell wall related genes (5.2% of total polymorphic genes). As average, each of these genes retains 16 SNPs and 3 INDELs across all species and all target assemblies, of which 7 SNPs and 1 INDEL are on exon regions.

3.3 Intra-specific SQTL variability leads to changes in cell wall protein sequences with a potential functional impact

The final step of the analysis of SQTL variability across the six species and related target genome assemblies consisted in assessing the effect of SNPs and INDELs on translated protein sequences. Overall, across all SQTLs, all species, and all target assemblies, 30654 SQTL genes (23% of all SQTL genes) displayed polymorphisms leading to one or more protein sequence changes between reference and one or more target assemblies (**Supplementary Dataset 1**). Of these, 1861 are cell wall genes. On average, each of the SQTL genes displaying protein-impacting polymorphisms across target assemblies retained 5.1 SNPs and 0.2 INDELs with an effect on the translated protein sequences. These polymorphisms cause a mean of 4.1 amino acid changes per coded protein sequences. Finally, 1126 SQTL genes of all the ones displaying protein-impacting polymorphisms (0.8% of all SQTL genes) have SNPs or INDELs between reference and target assemblies that lead to stop codons and (likely) truncated proteins. Of these, 421 are cell wall genes.

To assess the potential effect of the intra-specific protein sequence variability on cell wall quality traits, it was studied how such patterns impact functionally relevant and highly syntenically-conserved candidate SQTL genes identified in our previous SQTL

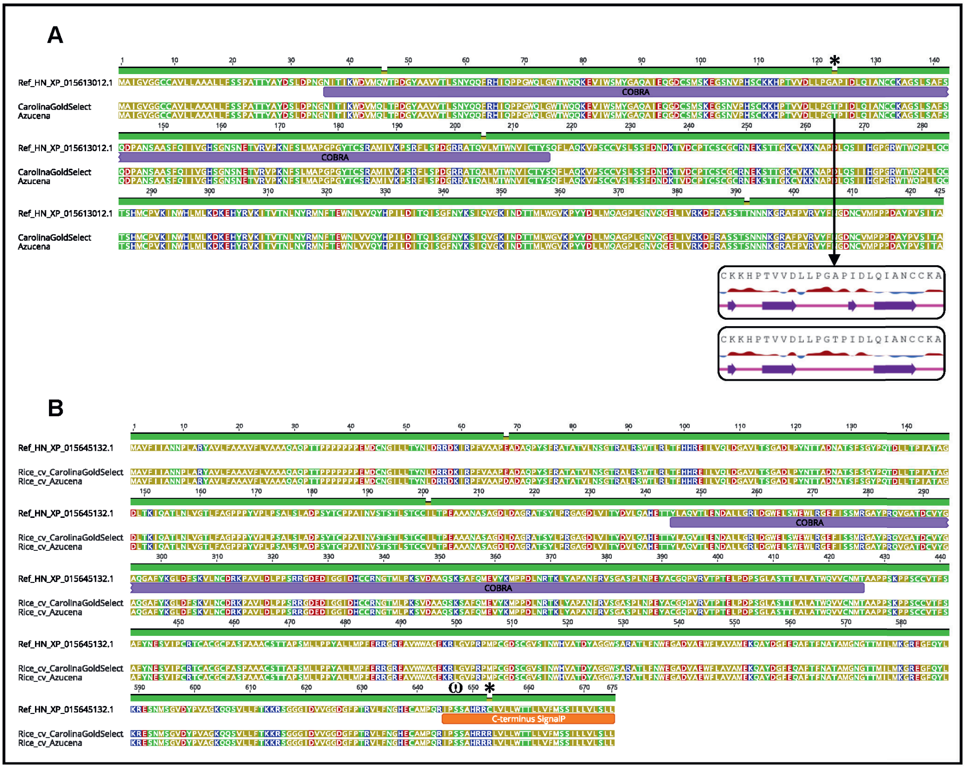


Figure 4. Amino acid changes and their effects on protein structure for three *Brittle culm-like* loci between the reference rice cv. "Nipponbare" and the two cultivars "Azucena" and "CarolinaGoldSelect". A) *OsBCL9*; B) *OsBCL8*. In each figure, protein sequences are colored according to amino acid polarity, and protein domains and signal peptides are annotated. Amino acid changes indicated with * indicate a change in polarity, while sites annotated with ω indicate the predicted GPI-anchoring-related omega sites.

study (Pancaldi et al., 2022b). These genes include members of the *COBRA* (*COB*) and *COBRA-like* (*COBL*) gene families responsible of the Brittle-culm cell wall rice mutants (Zhang and Zhou, 2011); important genes associated with lignin content variability in cell walls, as *FERULATE 5-HYDROXYLASE* (*F5H*), *CINNAMOYL-CoA O-METHYLESTERASE* (*CCOAOMT*), and *PHENYLALANINE AMMONIA-LYASE* (*PAL*) (Zhong et al., 2019); key hemicellulose- and cellulose-related genes for cell wall polysaccharides metabolism, as members of the *CELLULOSE SYNTHASE* (*CESA*), *CELLULOSE SYNTHASE-LIKE* (*CSL*), and *IRREGULAR XYLEM* (*IRX*) gene families (Brown et al., 2011, Little et al., 2018); important transcription factors for secondary cell wall development as *WRKY12*, *NAC SECONDARY WALL THICKENING FACTOR1* (*NST1*), and *C3H14* (Zhong et al., 2008, Rao and Dixon, 2018). For all of these genes, protein sequences were aligned and annotated for functional domains and motifs, allowing for the detection of protein sequence variability across assemblies, as well as for the assessment of the effect of variability on protein polarity, hydrophobicity, and

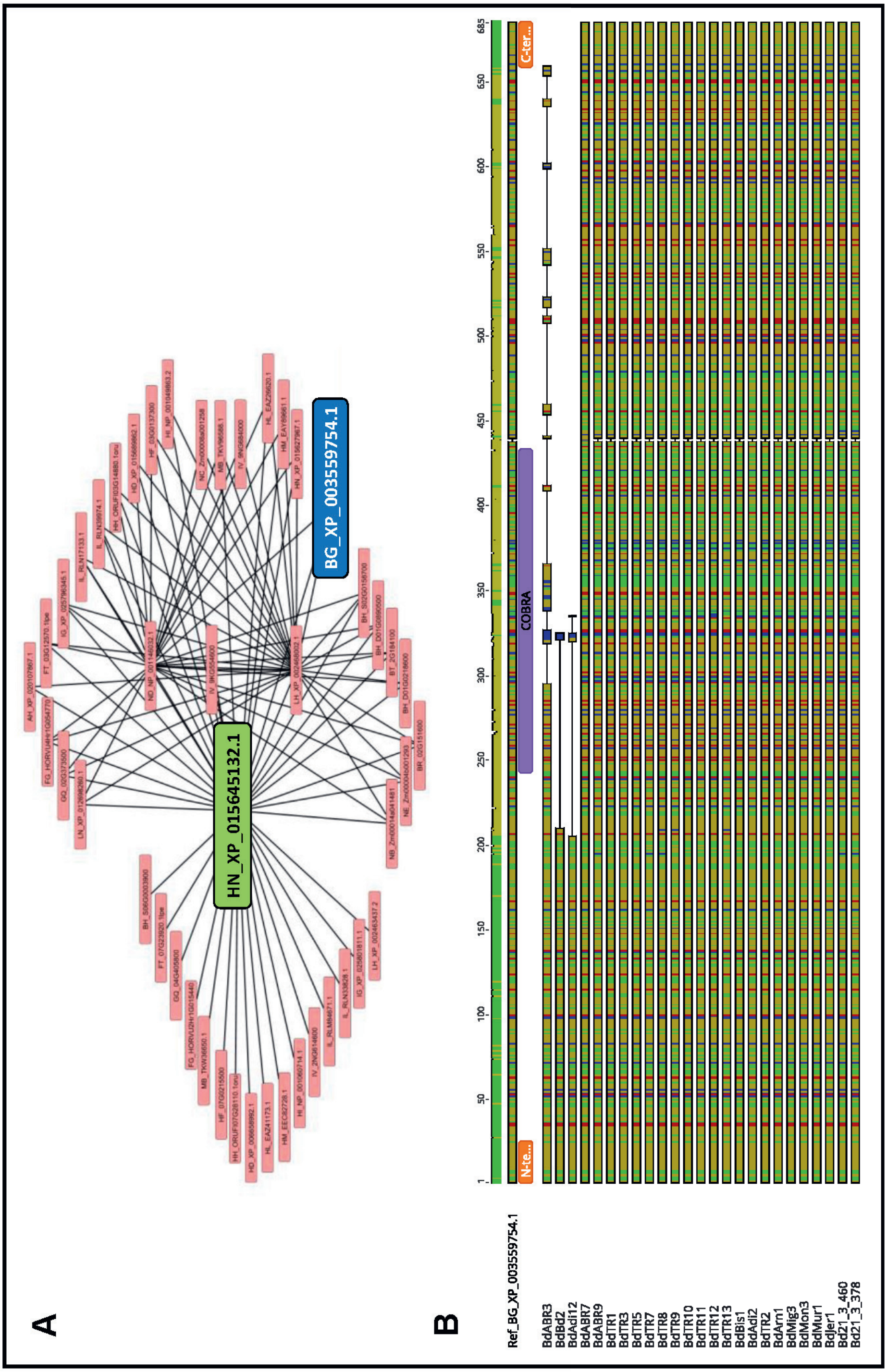


Figure 5. See the caption at the top of next page.

Figure 5 (page 160). The syntenic conservation of the rice locus *OsBCL8* (HN_XP_015645132.1) across Poaceae through SQTLs (A), and the intra-specific allelic variability with an impact on protein sequence of the *Brachypodium distachyon* syntenic homolog of *OsBCL8* conserved through SQTLs (BG_XP_003559754.1), across 25 brachypodium accessions compared to the reference genome (B). In Figure 5A, connections between genes indicate synteny through SQTLs, as described in Pancaldi et al. (2022b). In Figure 5B, protein sequences are colored according to amino acid polarity, and protein domains and signal peptides are annotated.

functional domain properties. **Figures 4 and 5** and **Supplementary Figures 1-4** display the results of these analyses.

Among all the alignments computed, the ones of the three members of the rice *BRITTLE CULM-LIKE* loci *OsBCL1*, *OsBCL8*, and *OsBCL9*, and of their related syntenic homologs conserved through SQTLs (Pancaldi et al., 2022b), yielded particularly interesting results. In fact, all the proteins coded by these three rice genes displayed one or more amino acid changes with effect on the polarity, charge and/or hydrophobicity of the proteins themselves. Specifically, these changes took place in the rice cultivars “Carolina Gold Select” and “Azucena” compared to the reference rice “Nipponbare”. In this regard, *OsBCL9* displayed three amino acid changes within the *COBRA* domain of the protein, of which one involving a change in polarity (Alanine to Threonine at position 123) and leading to the removal of a β-sheet domain in the 2D protein configuration (**Figure 4A**). A similar pattern was observed for *OsBCL1* (**Supplementary Figure 1**). Finally, *OsBCL8* displayed a Cysteine to Arginine change

leading to loss of amino acid charge within the protein C-terminus (position 653), right next to the GPI-anchoring-related ω-site (**Figure 4B**). In addition to the rice *BCL* loci, the other *BCL* genes from brachypodium, sorghum, and maize syntenic to *OsBCL* genes and retained within SQTLs were also assessed for protein variability. Intra-specific alignments revealed that the brachypodium homolog of the *OsBCL8* gene (XP_003559754.1) displayed massive protein rearrangements in three of the 45 brachypodium genome assemblies analysed (“ABR3”, “Bd2”, and “Adi12”). These changes involve large deletions of the *BdBCL8* protein sequence, including a consistent part of the *COBRA* domain in the line “ABR3”, and the complete *COBRA* domain in the lines “Bd2” and “Adi12” (**Figure 5**).

In addition to the particularly relevant examples found for the *BRITTLE CULM-LIKE* genes, intra-specific amino acid changes and INDELs impacting protein properties were found in several of the other proteins coded by the candidate SQTL genes inspected (**Supplementary Figures 2-4**). As for the rice Brittle culm loci, protein changes were observed both within and outside protein functional domains or motifs. For example, a maize *IRX9* protein involved in xylans biosynthesis (NP_001147664.1) displayed two regions with relatively large INDELs and different amino acid substitutions within the GT43 functional domain (**Supplementary Figure 2**). These changes take place in 12 of the 33 maize assemblies assessed and have effects on protein polarity and/or charge. Similarly, the enzyme coded by the brachypodium secondary cell wall *CESA7* gene (XP_003574029.1), also retained within SQTLs, displayed large deletions within the Cellulose synthase catalytic domain in the accession "TR9" compared to the reference brachypodium genome (**Supplementary Figure 3**). Moreover, amino acid substitutions within the protein functional domain with an effect on protein polarity were also observed. Overall, these patterns were linked to major changes in the secondary structure of the *CESA7* protein, including the deletion of seven α-helix and one β-sheet domains (**Supplementary Figure 3**). Finally, multiple INDELs and amino acid substitutions with impact on protein chemical properties were also detected across the arabidopsis accessions for the *AtC3H14* gene (a transcription factor regulating secondary cell wall thickening), across different brachypodium accessions for the *BdWRKY12* gene (a major transcription factor regulating cell wall biosynthesis in plant vessels), and between the rice reference genome and three rice cultivars (Koshihikari, Kitaake, and Azucena) for the *OsNAC43* gene (also a major transcription factor for secondary cell wall biosynthesis in different plant tissues) (**Supplementary Figure 4**).

3.4 SQTLs are valid tools to predict important cell wall genomic loci in *Miscanthus sinensis* and *Panicum virgatum*

In addition to evaluating the level and relevance of the intra-specific genetic variability of SQTLs in diverse species, another aim of this study was to assess the validity of SQTLs for predicting genomic loci associated with biomass quality traits in novel plant populations and species. This aspect was studied by assessing the degree of co-localization between the SQTLs previously detected in *Miscanthus sinensis* and *Panicum virgatum* (Pancaldi et al., 2022b) and cell wall related QTLs mapped by genome-wide association analysis (GWAS) in a *Miscanthus sinensis* collection and QTLs mapped by other researchers in an F1 progeny of a biparental cross of two switchgrass lines diverging for cell wall quality traits (Ali et al., 2020). These populations represent respectively a different intra-specific population than the one from which QTLs used for SQTLs mapping came from (miscanthus), and a separate species than the ones from which initial QTLs have been selected at the moment of SQTLs detection (thus, a hypothetical novel, under-domesticated, species; switchgrass).

The GWAS conducted on the miscanthus population identified a total of 91 QTLs associated to eight traits related to cell wall content and composition (**Figure 6** and **Supplementary Table 2**). These 91 QTLs cover 6.8% of the miscanthus genome and include a total of 148 cell wall genes. First, it was assessed the general degree of overlap between the GWAS QTL regions and the 254 SQTLs previously detected in miscanthus (which cover 32.7% of the miscanthus genome; **Supplementary Table 3**). This analysis revealed that 67 SQTLs (26.4% of all the miscanthus SQTLs) co-localized (for parts of their regions) with the 91 GWAS QTLs. Moreover, it was observed that 35 of the 91 GWAS QTLs (38%) co-localized for >50% of their bp length with genomic regions where miscanthus SQTLs are also present (**Figure 6**). To test if these figures highlight significant co-localization of the QTL loci identified by GWAS with the miscanthus SQTLs, a permutation analysis was performed by constructing 100 sets of 91 random genomic regions mirroring the bp size distribution of the GWAS QTLs (**Supplementary Dataset 2**). For each set, the proportion of random QTLs co-localizing with SQTLs was computed, and a binomial test was performed to assess if this proportion was significantly lower than the one observed for the real GWAS QTLs. The results showed that, as average across the 100 random QTLs sets analyzed, 17 QTLs of the 91 included in each set (19%) co-localized for >50% of their length with SQTLs. This figure is significantly lower compared to what observed in real QTLs (Binomial test significant in 91 of the 100 tests performed at α=0.01), highlighting that SQTLs co-localize significantly with the GWAS QTLs.

An analogous procedure to the one just described for miscanthus was performed in switchgrass to test the co-localization between the 56 cell wall related QTLs identified by Ali et al. (2020) (**Supplementary Table 5**), and the 254 SQTLs previously detected in switchgrass and conserved across Poaceae (**Supplementary Table 6**). In switchgrass, it was shown that 53 SQTLs (20.8% of all the switchgrass SQTLs) (partially) co-localize with the 56 cell wall QTLs, while 33 QTLs (59%) co-localize for >50% of their length with SQTLs (**Figure 7**). Permutation analysis demonstrated that the co-localization of QTLs and SQTLs is significant, as random QTL sets produced significantly lower proportions of co-localizing QTLs (Binomial test significant in 99% of iterations at α=0.01).

Since co-localization between SQTLs and QTLs was demonstrated for both miscanthus and switchgrass, as a final step we analyzed which cell wall genes are retained in co-localizing regions between SQTLs and QTLs, as well as the level of syntenic conservation of those genes in angiosperms. In this regard, miscanthus results showed that several important cell wall genes were retained within the co-localizing regions between SQTLs and GWAS QTLs. These include, among others, different central cell wall transcription factors, as one of the two *MsBLH6* (GQ_01G471400 and GQ_02G130600), the *MsWRKY12* (GQ_12G168300), the miscanthus homolog of arabidopsis *VND4* (GQ_12G158800), and the *MsMYB103* (GQ_07G174800) (**Figure 6** and **Table 1**). In addition, they comprise important lignin genes, as a miscanthus *PAL*

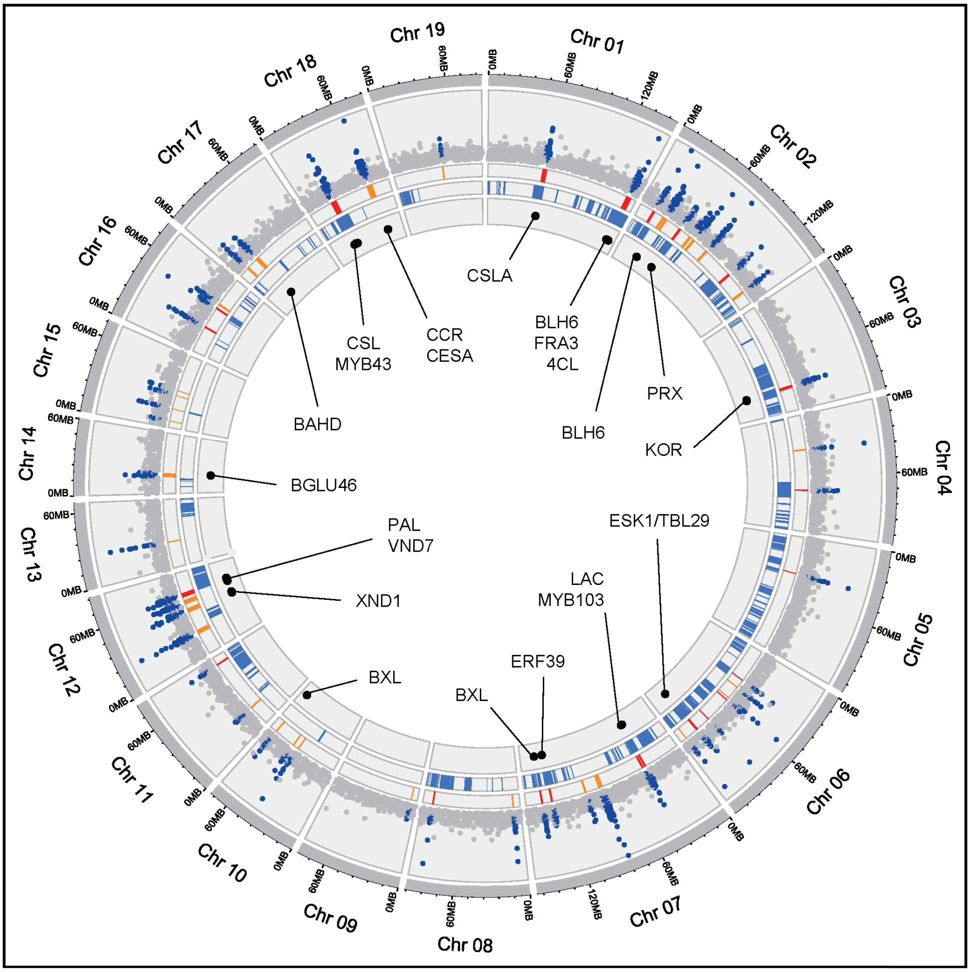


Figure 6. The results of the miscanthus GWAS and of the co-localization between miscanthus SQTLs and QTLs plotted onto the miscanthus genome. From outside to inside, the first strip shows the LOD scores of the markers from the GWAS, with the markers included in the 91 QTLs colored in blue; the second strip displays the genomic ranges of the 91 QTLs from the GWAS, with the QTLs co-localizing with SQTLs for >50% of their length colored in red; the third strip displays the positions of miscanthus SQTLs; the fourth strip highlights the most relevant candidate genes identified from the GWAS analysis.

and *CCOAOMT* copy (GQ_12G150500 and GQ_07G169800), or the miscanthus homolog of the maize *BRITTLE-STALK 2* locus (*COBL* gene; GQ_12G165800) (**Figure 6** and **Table 1**). Finally, synteny analysis showed that each of these genes is syntenic to other 84 genes on average (range 26-190). Most of the synteny is, as expected, toward Poaceae (**Figure 8A-B**). However, some syntenic connections involve also relatively distant eudicot species, as rapeseed or tomato, which have copies of their *WRKY12*, *VND4*, and *CCOAOMT* genes syntenic to miscanthus.

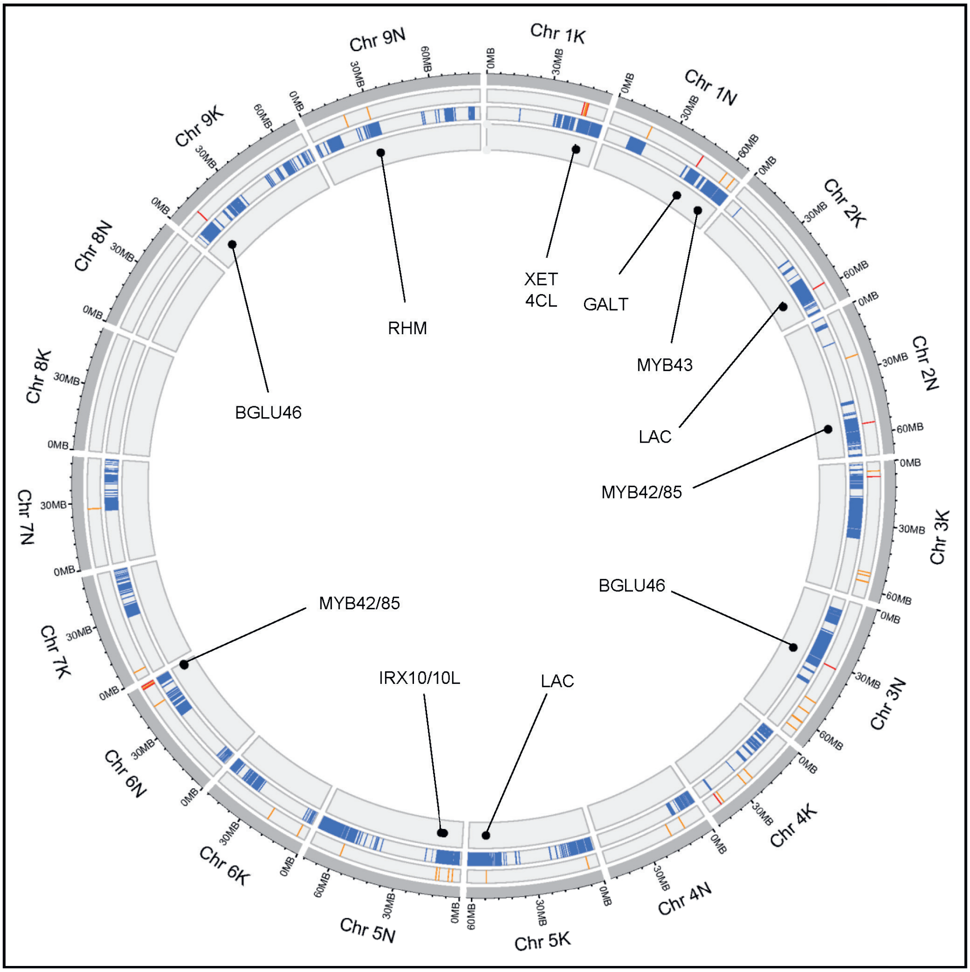


Figure 7. The genomic map of the cell wall QTLs collected for switchgrass from the work of Ali et al. (2020) and of the switchgrass SQTLs. From the outside, the first band represents the position of the 56 switchgrass QTLs collected from Ali et al. (2020), with the QTLs co-localizing with SQTLs colored in red. The second band represents the positions of the switchgrass SQTLs. The third band displays relevant cell wall candidate genes found in the co-localizing regions between SQTLs and QTLs.

For switchgrass, the most relevant cell wall genes found in co-localizing regions between SQTLs and QTLs from Ali et al. (2020) are summarized in **Figure 7** and **Table 2.** Among others, these include a homolog of the arabidopsis *FRAGILE FIBRE 1* locus (*FRA1*; IV_6NG323500) and a homolog of the arabidopsis *KOR* gene (IV_1NG408000). Moreover, an *IRX* gene involved in xylan synthesis (IV_5NG144100), and two

MYB42/MYB85 genes important for lignin biosynthesis (IV_2NG449200 and IV_6NG352900). All these genes are retained in QTLs from Ali et al. (2020) that are associated to traits for which the genes themselves appear highly functionally relevant

(**Table 2**). Moreover, synteny analysis of these genes showed that, as for miscanthus, they are highly syntenic, even if synteny is again mostly restricted to Poaceae (**Figure 8C-D**).

Table 1 – List of cell wall genes found within the miscanthus QTLs from the GWAS and conserved through SQTLs. The IDs of SQTL are reported as in Pancaldi et al. (2022b).

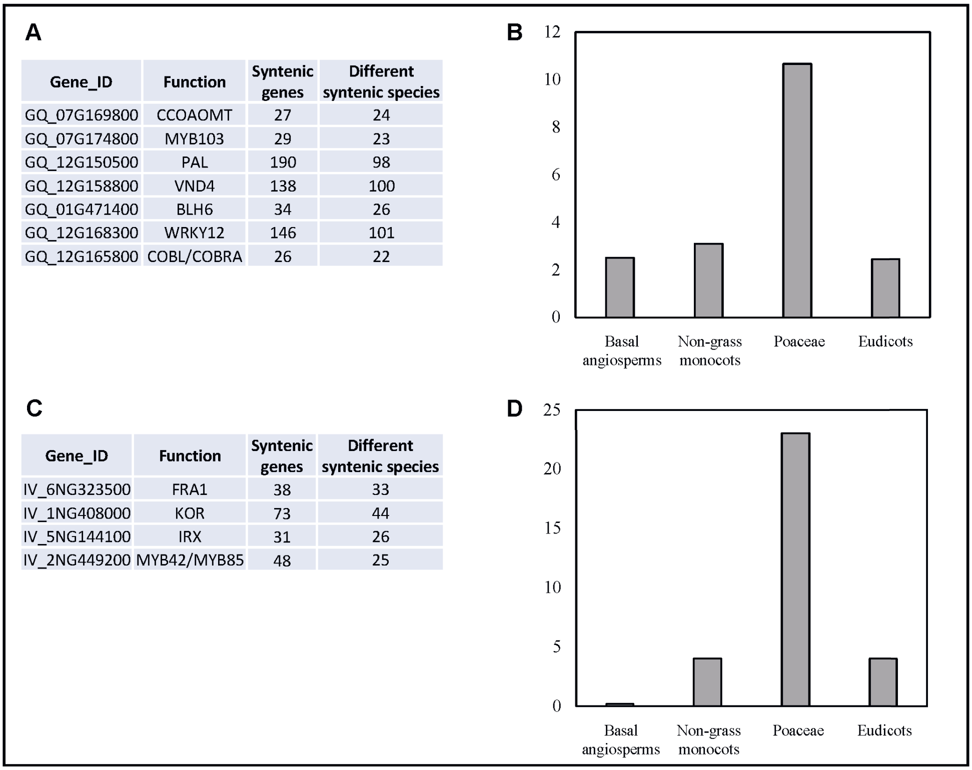
Gene ID	Chromosome	Conserved in SQTL	Cell wall function	Notes	References
GQ_01G229400	Chr01	MSQTL_122	CSL	Homolog of AtCSLA9, involved in the synthesis of mannans.	(Liepman et al., 2005)
GQ_01G468200	Chr01	MSQTL_122	CSL	Homolog of AtCSLA9, involved in the synthesis of mannans.	(Liepman et al., 2005)
GQ_01G471400	Chr01	MSQTL_5	BLH6	Transcription factor promoting secondary cell wall synthesis in grasses	(Hirano et al., 2013, Rao and Dixon, 2018)
GQ_01G471500	Chr01	MSQTL_5	BGAL	Beta-galactosidase involved in the degradation of several polysaccharides and cell wall remodelling	(Chandrasekar and van der Hoorn, 2016)
GQ_01G474600	Chr01	MSQTL_122	FRA3	Homolog of *AtFRA3*, which coordinates actin organization during cellulose deposition	(Zhong et al., 2004)
GQ_01G478800	Chr01	MSQTL_212	4CL	Important gene of the lignin synthesis pathway	(Zhong et al., 2019)
GQ_03G236400	Chr03	MSQTL_216	Endoglucanase/KOR	Important genes for cellulose and plant cell wall metabolism	(Zuo et al., 2000, Glass et al., 2015)
GQ_03G240500	Chr03	MSQTL_248	Endoglucanase/KOR	Important genes for cellulose and plant cell wall metabolism	(Zuo et al., 2000, Glass et al., 2015)
GQ_07G169800	Chr07	MSQTL_180	CCOAOMT	Central gene for lignin synthesis	(Zhong et al., 2019)
GQ_07G169900	Chr07	MSQTL_180	LAC	Gene involved in lignin deposition	(Zhong et al., 2019)

Table 1 *(continued)*

Gene_ID	Chromosome	Conserved in SQTL	Cell wall function	Notes	References
GQ_07G174800	Chr07	MSQTL_180	MYB103	Regulates F5H expression and S-lignin deposition in arabidopsis	(Öhman et al., 2013)
GQ_07G435600	Chr07	MSQTL_245	ERF39	Binds to promoters of CESA1/3/6, synthesizing primary-like cell walls	(Saelim et al., 2019)
GQ_07G435700	Chr07	MSQTL_245	ERF39	Binds to promoters of CESA1/3/6, synthesizing primary-like cell walls	(Saelim et al., 2019)
GQ_07G477100	Chr07	MSQTL_445	BXL	Inhibits xylan synthesis	(Goujon et al., 2003)
GQ_12G086100	Chr12	MSQTL_118	XND1/ WND1A	Major transcription factor for secondary cell wall	(Zhong and Ye, 2015)
GQ_12G091100	Chr12	MSQTL_118	Endoglucanase/KOR	Important genes for cellulose and plant cell wall metabolism	(Zuo et al., 2000, Glass et al., 2015)
GQ_12G150500	Chr12	MSQTL_38	PAL	First step of the lignin pathway	(Zhong et al., 2019)
GQ_12G153400	Chr12	MSQTL_38	UGT72E3	Genes influencing kinetics of lignin deposition	(Baldacci-Cresp et al., 2020)
GQ_12G158800	Chr12	MSQTL_2	VND4	Master regulator of secondary cell walls	(Zhong and Ye, 2015, Rao and Dixon, 2018)
GQ_12G165800	Chr12	MSQTL_91	COBL/ COBRA	Homolog of Brittle-stalk2 locus of maize	(Sindhu et al., 2007)
GQ_12G168300	Chr12	MSQTL_2	WRKY12	Involved in regulation of secondary cell wall and flowering in *Miscanthus lutarioriparius*	(Yu et al., 2013)
GQ_18G102600	Chr18	MSQTL_178	CSL	Gene involved in hemicellulose biosynthesis	(Little et al., 2018)
GQ_18G114300	Chr18	MSQTL_185	MYB20/ MYB43	MYB20, MYB43, and MYB85 regulate secondary cell wall formation	(Zhong and Ye, 2015, Rao and Dixon, 2018, Geng et al., 2020)

Table 2 – List of cell wall genes from the switchgrass QTLs mapped by Ali et al. (2020) and conserved through SQTLs. QTL traits refer to the traits associated to the QTLs from Ali et al. (2020) where the genes were found as retained. The SQTL IDs are reported as in Pancaldi et al. (2022b).

Gene_ID	Chromosome	QTL trait(s) (Ali et al., 2020)	Conserved in SQTL	Cell wall function
IV_1KG461400	Chr_01K	Glucose;Total cell wall sugar	MSQTL_195	4CL
IV_6NG354300	Chr_06N	Total cell wall sugar	MSQTL_80	BGAL
IV_3NG135100	Chr_03N	Glucose	MSQTL_25	BGLU46
IV_9KG034800	Chr_09K	Xylose	MSQTL_102	BGLU46
IV_1NG408000	Chr_01N	Total cell wall sugar	MSQTL_185	Endoglucanase/KOR
IV_6NG323500	Chr_06N	Glucose;Total cell wall sugar	MSQTL_45	FRA1
IV_1NG306100	Chr_01N	Total cell wall sugar	MSQTL_18	GALT/HPGT
IV_2NG448300	Chr_02N	Klason lignin	MSQTL_54	GAUT1
IV_2KG434900	Chr_02K	Glucose	MSQTL_230	GH17
IV_5NG130900	Chr_05N	Klason lignin	MSQTL_123	GLCAT
IV_5NG144100	Chr_05N	Glucose;Xylose	MSQTL_130	IRX10/10L
IV_2KG435600	Chr_02K	Glucose	MSQTL_230	LAC
IV_5KG613400	Chr_05K	Glucose	MSQTL_350	LAC
IV_1NG407200	Chr_01N	Total cell wall sugar	MSQTL_185	MED/REF
IV_1NG407100	Chr_01N	Total cell wall sugar	MSQTL_185	MYB20/MYB43
IV_6NG354000	Chr_06N	Total cell wall sugar	MSQTL_80	MYB4/MYB6/MYB7/ MYB21/MYB32
IV_2NG449200	Chr_02N	Klason lignin	MSQTL_248	MYB42/MYB85
IV_6NG352900	Chr_06N	Total cell wall sugar	MSQTL_80	MYB42/MYB85
IV_6NG354100	Chr_06N	Total cell wall sugar	MSQTL_80	PRX
IV_9KG292900	Chr_09N	Klason lignin	MSQTL_122	RHM
IV_1KG460000	Chr_01K	Glucose;Total cell wall sugar	MSQTL_164	XET/XTH



A

Gene_ID	Function	Syntenic genes	Different syntenic species
GQ_07G169800	CCOAOMT	27	24
GQ_07G174800	MYB103	29	23
GQ_12G150500	PAL	190	98
GQ_12G158800	VND4	138	100
GQ_01G471400	BLH6	34	26
GQ_12G168300	WRKY12	146	101
GQ_12G165800	COBL/COBRA	26	22

C

Gene_ID	Function	Syntenic genes	Different syntenic species
IV_6NG323500	FRA1	38	33
IV_1NG408000	KOR	73	44
IV_5NG144100	IRX	31	26
IV_2NG449200	MYB42/MYB85	48	25

Figure 8. Analysis of the syntenic relationships of the most relevant genes from miscanthus (A-B) and switchgrass (C-D) that are conserved through SQTLs and co-localizing with the QTLs described in Paragraph 2.4. A) Absolute number of syntenic genes and of different species to which each miscanthus gene is syntenic. B) Average number of syntenic relationships with species belonging to different angiosperm clades across the genes in panel A. C) Absolute number of syntenic genes and of different species to which each switchgrass gene is syntenic. D) Average number of syntenic relationships with species belonging to different angiosperm clades across the genes in panel C.

3.5 SQTLs as tools to circumvent limitations of genetic mapping approaches

In the final step of this research, it was evaluated whether SQTLs, given the incorporation of information from multiple genomic loci previously shown to determine variability in a trait of interest (in this case cell wall quality), can be used to overcome limitations of genetic mapping approaches. Specifically, a common issue encountered in association mapping – and especially GWAS – is that if patterns of population structure overlap with patterns of phenotypic variability across accessions for the trait(s) evaluated, population structure correction can hide (part of) the relevant loci that govern the trait(s) (Brachi et al., 2011). The GWAS conducted in *Miscanthus sinensis* was therefore used to study if SQTLs can, at least partly, overcome this limitation. First, the analysis of population structure computed to perform the "standard" GWAS described in Paragraph 2.4 was co-analysed with the phenotypic

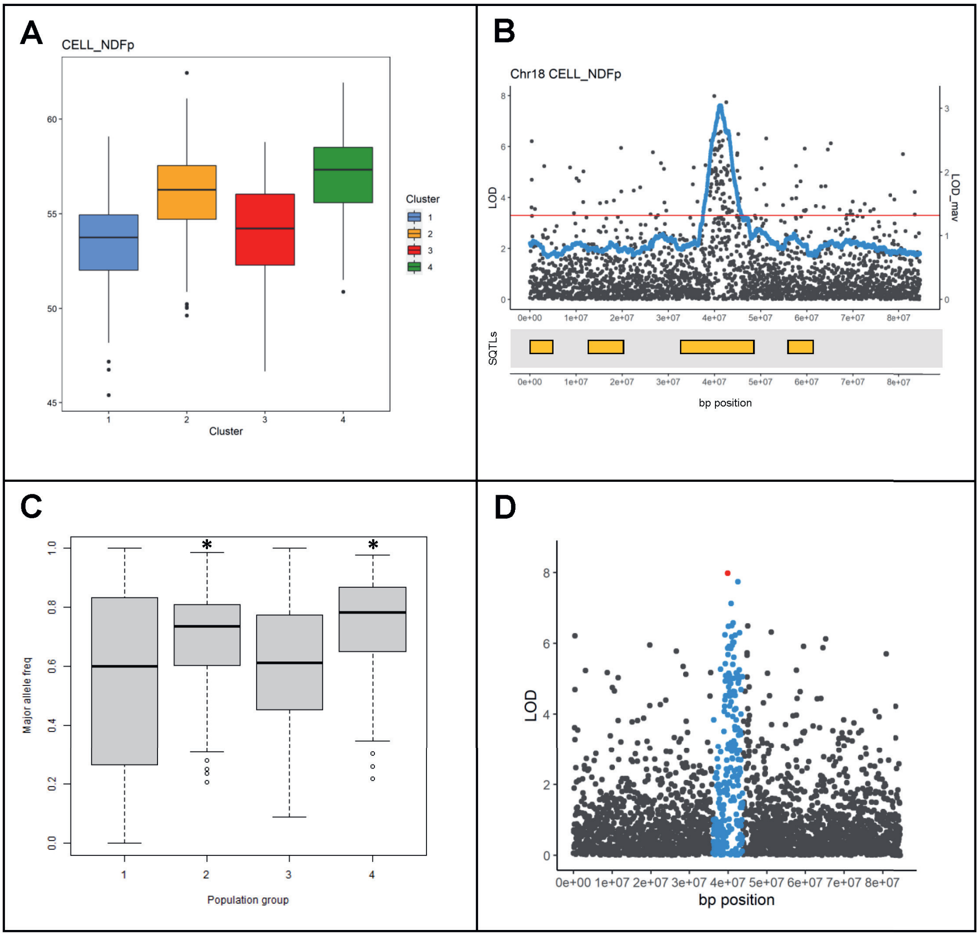


Figure 9. The steps followed for the detection of the "extra" GWAS QTLs. A) Significant variation for one trait across the four main population structure groups of the GWAS panel. B) A LOD peak from the GWAS without population structure correction for the same trait as in panel A. This peak was not included in the "standard" GWAS and co-localizes with a known SQTL region (yellow bars). C) Assessment of the variation in the major allele frequency across the four population structure groups as in panel A, for all the markers included within the LOD peak of panel B. MAF differs significantly across population structure groups (*: significant differences at α=0.05). D) Identification of the "extra QTL".

data on cell wall traits to assess co-variation between population structure and phenotypic traits across the miscanthus accessions of the GWAS panel (**Supplementary Figures 5-6**). Interestingly, this co-variation was found for multiple traits tested, including total cell wall (NDF; dry matter percentage), total ADF (dry matter percentage), cellulose (dry matter percentage), cellulose (NDF percentage), hemicellulose (dry matter percentage), and hemicellulose (NDF percentage) (**Supplementary Figures 5-6**). A GWAS without population structure correction was therefore run, and co-localization between SQTLs and the LOD peaks found by this

GWAS and not already included among the 91 QTLs from the “standard” GWAS with population structure correction (Paragraph 2.4) was assessed. These peaks would normally be discarded as false positive associations due to population structure. However, their co-localization with SQTLs indicates that in other species, QTLs for similar traits were found on these exact genomic regions, highlighting their potential relevance. As extra control, the SNP markers included within the LOD peaks co-localizing with SQTLs and not included among the 91 QTLs of the “standard” GWAS were extracted and tested with ANOVA for significant differences in the major allele frequency (MAF) between the different population structure groups depicted in **Supplementary Figure 5**. Upon detection of significant MAF differences between co-varying population structure and phenotypic groups, the peak regions were declared as “extra QTLs” found by co-localization with SQTLs (**Figure 9**).

In total, 17 “extra QTLs” were detected (**Supplementary Table 7**), which contain a total of 13 cell wall genes. Among these, the most relevant ones are a miscanthus homolog of the arabidopsis *KOR* gene (GQ_05G142200) that was found within a QTL associated to Cellulose dry matter content. Moreover, a *MsCSL* gene (GQ_19G142500), homolog of the arabidopsis and maize *CslD5*, was also found within four co-localizing “extra QTLs” associated to total cell wall dry matter content, ADF dry matter content, Cellulose dry matter content, and Cellulose NDF percentage. Finally, two peroxidases (*PRX*; GQ_06G157300, GQ_19G143200), and a *XYLOGLUCAN ENDO-TRANS-GLYCOSYLASE/HYDROLASE* (*XTH*; GQ_19G142300).

4 Discussion

4.1 SQTLs as reservoirs of allelic variation with a potential use for biomass improvement

In a previous publication, it was shown that SQTLs are useful breeding tools to project known genetic information on the architecture of traits of interest – specifically cell wall quality traits – from model species to understudied crops (Pancaldi et al., 2022b). This research aimed at assessing whether SQTLs can potentially guide breeding activities by spanning genomic regions displaying allelic variability for target traits and by pinpointing relevant loci associated to biomass quality variability in novel plant populations and crops. Regarding the first point, this study clearly showed that

SQTL regions are reservoirs of intra-specific allelic variability. A part – minor – of this variability entails PAV of SQTL genes. Gene PAV is commonly found in intra-specific comparative genomic studies, as pangenome analyses, and is a known source of trait variability (Gordon et al., 2017, Song et al., 2020). In this sense, the SQTL cell wall genes displaying intra-specific PAV may potentially impact biomass traits in target accessions in several ways. For example, in the case a SQTL gene is part of a multi-gene family, its PAV could affect overall gene copy number and lead to differential gene dosage across accessions (Kondrashov, 2012, Gabur et al., 2019). Alternatively, in more extreme cases, PAV may lead to a loss-of-function, with a likely large impact on plant traits (Gabur et al., 2019). In this study, the SQTL genes showing PAV between reference and target assemblies included *PAL* and *CAD* genes, which displayed PAV among some of the *Brachypodium distachyon* and *Brassica napus* accessions assessed. *PAL* and *CAD* are multi-gene families, and gene dosage is thought to be important for their functionality (Wagner et al., 2012, Preisner et al., 2014, Lu et al., 2019). Moreover, in the close relative of *Brassica napus*, *Brassica rapa*, intra-specific copy number and PAV of *PRX* genes, which together with *CAD* determine the efficiency of lignin production at the final steps of the lignin pathway (Liu et al., 2018b), was suggested to affect lignin metabolism and related morphological traits (Lin et al., 2014). Therefore, intra-specific PAV of SQTL genes might be relevant for determining variability in cell wall quality traits, and genomic analysis of SQTL regions in (novel) crop panels might quickly highlight promising target accessions for inclusion in breeding programs based on the assessment of gene PAV.

In addition to PAV, nucleotide polymorphisms were also found in high numbers in the intra-specific analyses of SQTL regions. Specifically, multiple genes that were previously defined as relevant SQTLs candidates (Pancaldi et al., 2022b) display intra-specific nucleotide polymorphisms that impact, sometimes considerably, the sequences of their proteins. In this sense, the examples reported for the rice *BCL8* locus and its brachypodium syntenic homolog identified through SQTLs are particularly relevant. In fact, the protein modifications reported for these genes sometimes break the integrity of the *COBRA* domain of the protein, or affect the membrane-anchoring domains at the C-terminus. These mutations are highly similar to the ones found in the well characterized *OsBC1* and *OsBCL1* mutants, which are close homologs of the *OsBCL8* gene (Li et al., 2003, Dai et al., 2009). Both these mutants display alterations of the plant mechanical strength and of cell wall compositional properties as effect of the *COBL* gene mutations (Li et al., 2003, Dai et al., 2009, Zhang and Zhou, 2011). Thus, the mutations found in this study might potentially lead to similar effects. Interestingly, the rice cv. “Azucena” – which is one of the rice cultivars displaying a change in protein polarity in the C-terminus of the *BCL8* protein – displays differences in the relative content of cell wall components, including cellulose, compared to the reference of this study, cv. “Nipponbare” (Jahn et al., 2011). Nevertheless, future research is needed to determine a relationship between these mutations and cell wall composition. The same goes for all the other mutations discussed in Paragraph 2.3.

To conclude, irrespectively of the functional relevance of the specific patterns of genomic variability observed for the SQTL regions in the intra-specific comparisons performed, it is still highly relevant that SQTL genes were proven to show such patterns. This demonstrates that the high conservation of gene presence and order across diverse genomes does not preclude the existence of intra-specific allelic variability for those highly-conserved regions. Therefore, synteny and SQTLs can be effectively used to mine interesting alleles, by genomically analysing SQTL regions in target species and searching for specific mutations previously identified as particularly relevant in model crops. If this step would be performed during the screening of useful material for breeding programs, it would lead to faster and more effective breeding activities in novel (under-domesticated) biomass species.

4.2 SQTLs as valid tools toward the improvement of breeding activities in novel crops

In the second part of this study, it was tested if SQTLs represent effective tools to predict relevant loci associated to biomass quality in novel mapping panels or crop species. The results showed that two sets of QTLs from *Miscanthus sinensis* and *Panicum virgatum*, respectively, co-localized significantly with SQTL regions of these two species as compared to random genomic loci. This result demonstrates that, in a breeding context involving novel species, SQTLs can be valuable tools to pinpoint relevant target loci that are likely responsible of variability in traits of interest. Moreover, by using the annotations of SQTL genes from the multiple species that are represented within SQTLs – as performed here for cell wall genes – it is possible to filter candidate genes based on functional and literature information. Finally, syntenic conservation of target loci through SQTLs can quickly display the functional conservation of interesting target genes through multiple (model) species for which studies might be available, to better evaluate the functional relevance of cell wall genes within QTLs and SQTLs.

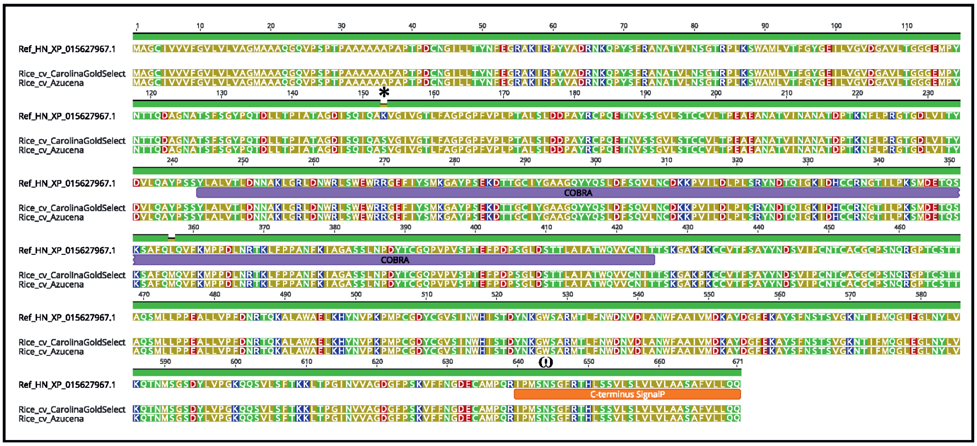
By applying the approach just described, in addition to the co-localization between SQTLs and miscanthus and switchgrass QTLs, several genes involved in these co-localizations were identified that represent valuable candidates for determining variability in biomass composition in these species. Specifically, some of the genes found from the miscanthus GWAS and conserved through SQTLs appear of particular interest. For example, the transcription factor *BLH6* is known to deeply affect the properties of plant cell walls, not only within different species (Hirano et al., 2013, Liu et al., 2014a), but also between different species clades, as Poaceae vs eudicots (Rao and Dixon, 2018). Therefore, the finding of this gene within GWAS QTLs co-localizing with SQTLs makes it a very good candidate for modulating the relative ratio of different cell wall polysaccharides and lignin in miscanthus. In addition to *BLH6*, genes as *WRKY12*, *VND4*, and *MYB103* are all central cell wall genes, demonstrated to modulate cell wall regulation and composition across a range of species (Zhong et al., 2008, Wang et al., 2010a, Öhman et al., 2013, Zhou et al., 2014, Yang et al., 2016a, Rao and Dixon, 2018, Wu et al., 2021). All these genes are thus important candidates for the cell wall quality variability observed in the miscanthus panel used for the GWAS described in this manuscript. Ideally, this information might be therefore used to map relevant mutations at these genes and screen plant material for those mutations. Remarkably, the intra-specific genomic comparisons performed in this research revealed that nucleotide variability leading to amino acid substitutions and protein INDELs were observed among the brachypodium accessions. These two approaches – SQTL-guided association mapping and genomic screening/prediction of mutations at target candidate genes – could be combined in novel species to enable quicker (pre)breeding for biomass compositional traits.

In the last part of this study it was also tested if SQTLs can be used to make the population structure correction more efficient in GWA studies. While it needs to be clearly stated that the accounting of population structure in GWAS is pivotal to enable accurate mapping, the results obtained here indicate that the inspection of co-localization between SQTLs and significant SNPs normally discarded as false-positives after population structure correction could represent a useful complement of a "standard" GWAS pipeline. Specifically, as they are defined, SQTLs represent genomic regions that are syntenically conserved between species and in multiple model species were shown to be associated to QTLs for a trait of interest (Pancaldi et al., 2022b). Therefore, they could help in retrieving some of the "missing heritability" that can get blurred below the threshold of false discovery (Brachi et al., 2011), in the case the SQTL areas are significantly co-varying with a target trait in other (related) crops. However, the "eye" of the researcher is crucial in order to be able to evaluate the relevance of potential candidate genes found within the "extra GWAS QTLs" pinpointed by using SQTLs. In this sense, in our miscanthus panel, some of the candidate genes retained within the 17 detected "extra QTLs" of paragraph 2.5 appear highly relevant in the context of cell wall variability. Therefore, these genes might be considered, depending on the needs, in breeding settings to improve miscanthus biomass quality, together with all the other ones discussed in the previous paragraph(s).

5 Conclusions

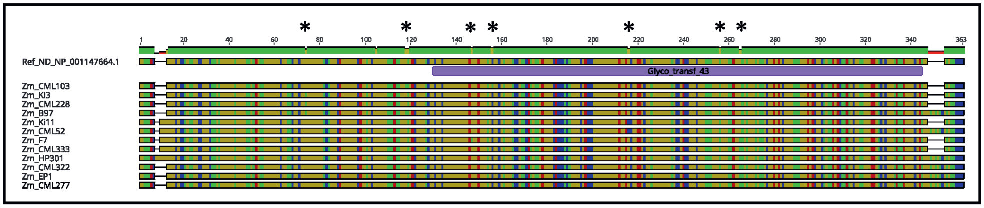
This study aimed at assessing the validity of SQTLs to assist breeding of (novel) biomass crops, by (i) analysing the intra-specific level of SQTLs allelic genomic variability in panels of available good-quality plant genomes, (ii) assessing the SQTLs predictive value of relevant biomass-quality-related loci in miscanthus and switchgrass, and (iii) suggesting possible side-uses of SQTLs to complement standard approaches of genetic mapping. The results showed that SQTLs are valuable tools to improve breeding activities in novel crops, where they can be applied in different ways to either screen plant material for genes or alleles of interest, or to pinpoint relevant loci based on the information from model species in novel accession panels or crops. Moreover, the research performed to achieve the goals above allowed the finding of relevant alleles at candidate SQTL loci, as well as of important candidate cell wall genes for miscanthus biomass improvement. Future research could investigate the phenotypic relevance hypothesized for the intra-specific patterns of SQTLs variability revealed in this study, by phenotyping cell wall composition in the plant lines showing promising variability. Moreover, the miscanthus loci highlighted in this study could be targeted by genetic modification to study their specific relevance for miscanthus biomass quality. Finally, SQTLs might start to be included in breeding programs of under-domesticated biomass crops as proposed in this research, to test their usefulness in novel, important, breeding settings.

Supplementary Data


1
10
20
30
40
50
60
70
80
90
100
110
Ref_HN_XP_015627967.1 MAGCIVVVFGVLVLVAGMAAAQGQVPSPTPAAAAAAPAPTPDCNGILLTYNFEGRAKIRPYVADRNKQPYSFRANATVLNSGTRPLKSWAMLVTFGYGEILVGVDGAVLTGGGEMPY
Rice_cv_CarolinaGoldSelect MAGCIVVVFGVLVLVAGMAAAQGQVPSPTPAAAAAAPAPTPDCNGILLTYNFEGRAKIRPYVADRNKQPYSFRANATVLNSGTRPLKSWAMLVTFGYGEILVGVDGAVLTGGGEMPY
Rice_cv_Azucena MAGCIVVVFGVLVLVAGMAAAQGQVPSPTPAAAAAAPAPTPDCNGILLTYNFEGRAKIRPYVADRNKQPYSFRANATVLNSGTRPLKSWAMLVTFGYGEILVGVDGAVLTGGGEMPY
120
130
140
150
*
160
170
180
190
200
210
220
230
Ref_HN_XP_015627967.1 NTTQDAGNATSFSGYPQTDLLTPIATAGDISQIQAKVGIVGTLFAGPGPFVPLPTALSLDDPAYRCPQETNVSSGVLSTCCVLTPEAEANATVINANATDPTKNFLPRGTGDLVITY
Rice_cv_CarolinaGoldSelect NTTQDAGNATSFSGYPQTDLLTPIATAGDISQIQASVGIVGTLFAGPGPFVPLPTALSLDDPAYRCPQETNVSSGVLSTCCVLTPEAEANATVINANATDPTKNFLPRGTGDLVITY
Rice_cv_Azucena NTTQDAGNATSFSGYPQTDLLTPIATAGDISQIQASVGIVGTLFAGPGPFVPLPTALSLDDPAYRCPQETNVSSGVLSTCCVLTPEAEANATVINANATDPTKNFLPRGTGDLVITY
240
250
260
270
280
290
300
310
320
330
340
350
Ref_HN_XP_015627967.1 DVLQAYPSSYLALVTLDNNAKLGRLDNWRLSWEWRRGEFIYSMKGAYPSEKDTTGCIYGAAGQYYQSLDFSQVLNCDKKPVILDLPLSRYNDTQIGKIDHCCRNGTILPKSMDETQS
COBRA
Rice_cv_CarolinaGoldSelect DVLQAYPSSYLALVTLDNNAKLGRLDNWRLSWEWRRGEFIYSMKGAYPSEKDTTGCIYGAAGQYYQSLDFSQVLNCDKKPVILDLPLSRYNDTQIGKIDHCCRNGTILPKSMDETQS
Rice_cv_Azucena DVLQAYPSSYLALVTLDNNAKLGRLDNWRLSWEWRRGEFIYSMKGAYPSEKDTTGCIYGAAGQYYQSLDFSQVLNCDKKPVILDLPLSRYNDTQIGKIDHCCRNGTILPKSMDETQS
360
370
380
390
400
410
420
430
440
450
460
Ref_HN_XP_015627967.1 KSAFQLQVFKMPPDLNRTKLFPPANFKIAGASSLNPDYTCGQPVPVSPTEFPDPSGLDSTTLAIATWQVVCNITTSKGAKPKCCVTFSAYYNDSVIPCNTCACGCPSNQRGPTCSTT
COBRA
Rice_cv_CarolinaGoldSelect KSAFQMQVFKMPPDLNRTKLFPPANFKIAGASSLNPDYTCGQPVPVSPTEFPDPSGLDSTTLAIATWQVVCNITTSKGAKPKCCVTFSAYYNDSVIPCNTCACGCPSNQRGPTCSTT
Rice_cv_Azucena KSAFQMQVFKMPPDLNRTKLFPPANFKIAGASSLNPDYTCGQPVPVSPTEFPDPSGLDSTTLAIATWQVVCNITTSKGAKPKCCVTFSAYYNDSVIPCNTCACGCPSNQRGPTCSTT
470
480
490
500
510
520
530
540
550
560
570
580
Ref_HN_XP_015627967.1 AQSMLLPPEALLVPFDNRTQKALAWAELKHYNVPKPMPCGDYCGVSINWHISTDYNKGWSARMTLFNWDNVDLANWFAAIVMDKAYDGFEKAYSFNSTSVGKNTIFMQGLEGLNYLV
Rice_cv_CarolinaGoldSelect AQSMLLPPEALLVPFDNRTQKALAWAELKHYNVPKPMPCGDYCGVSINWHISTDYNKGWSARMTLFNWDNVDLANWFAAIVMDKAYDGFEKAYSFNSTSVGKNTIFMQGLEGLNYLV
Rice_cv_Azucena AQSMLLPPEALLVPFDNRTQKALAWAELKHYNVPKPMPCGDYCGVSINWHISTDYNKGWSARMTLFNWDNVDLANWFAAIVMDKAYDGFEKAYSFNSTSVGKNTIFMQGLEGLNYLV
590
600
610
620
630
640
ω
650
660
671
Ref_HN_XP_015627967.1 KQTNMSGSDYLVPGKQQSVLSFTKKLTPGINVVAGDGFPSKVFFNGDECAMPQRIPMSNSGFRTHLSSVLSLVLVLAASAFVLLQQ
C-terminus SignalP
Rice_cv_CarolinaGoldSelect KQTNMSGSDYLVPGKQQSVLSFTKKLTPGINVVAGDGFPSKVFFNGDECAMPQRIPMSNSGFRTHLSSVLSLVLVLAASAFVLLQQ
Rice_cv_Azucena KQTNMSGSDYLVPGKQQSVLSFTKKLTPGINVVAGDGFPSKVFFNGDECAMPQRIPMSNSGFRTHLSSVLSLVLVLAASAFVLLQQ

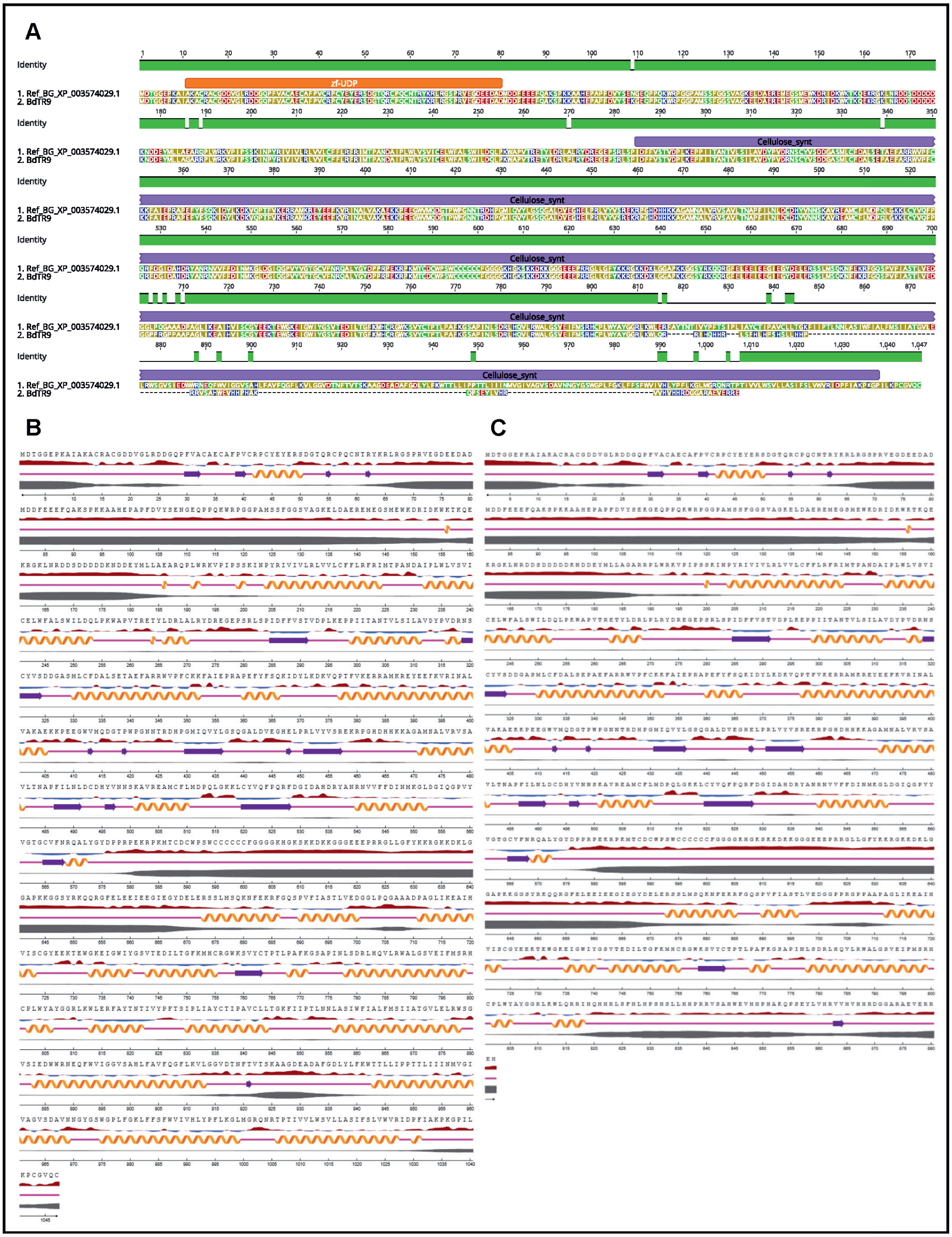
Supplementary Figure 1.

Amino acid changes and their effects on the protein structure of *OsBCL1* between the reference rice cv. “Nipponbare” and the two cultivars “Azucena” and “CarolinaGoldSelect”. Colouring of amino acids reflects amino acid polarity, and protein domains and signal peptides are annotated. Amino acid changes indicated with * indicate a change in polarity, while ω sites indicate the predicted GPI-anchoring-related omega sites.



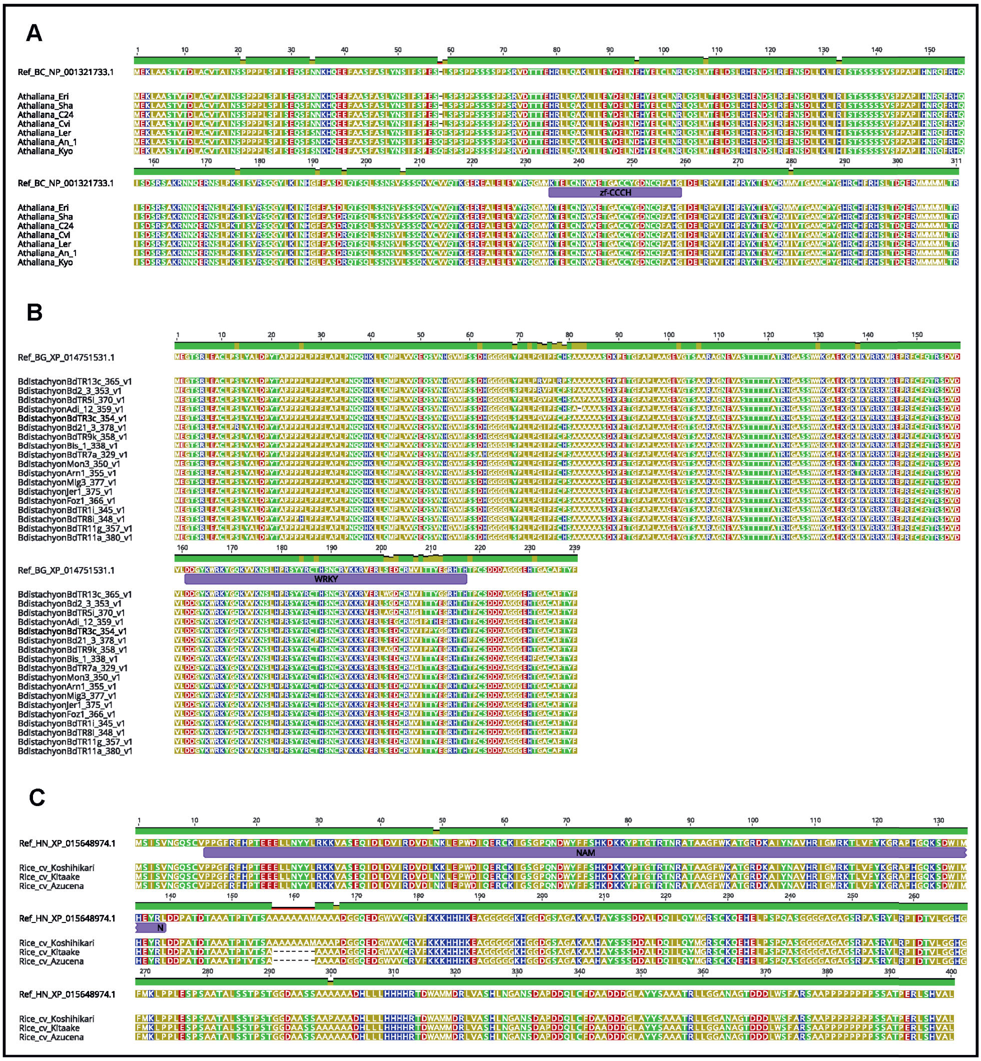
Supplementary Figure 2.

Amino acid substitutions and INDELs of *ZmIRX9* between the reference maize B73 genome and 12 maize accessions. Colouring of amino acids reflects amino acid polarity, and the GT43 protein domain is annotated. Amino acid changes indicated with * indicate a change in polarity/charge in some of the accessions compared to the reference genome.



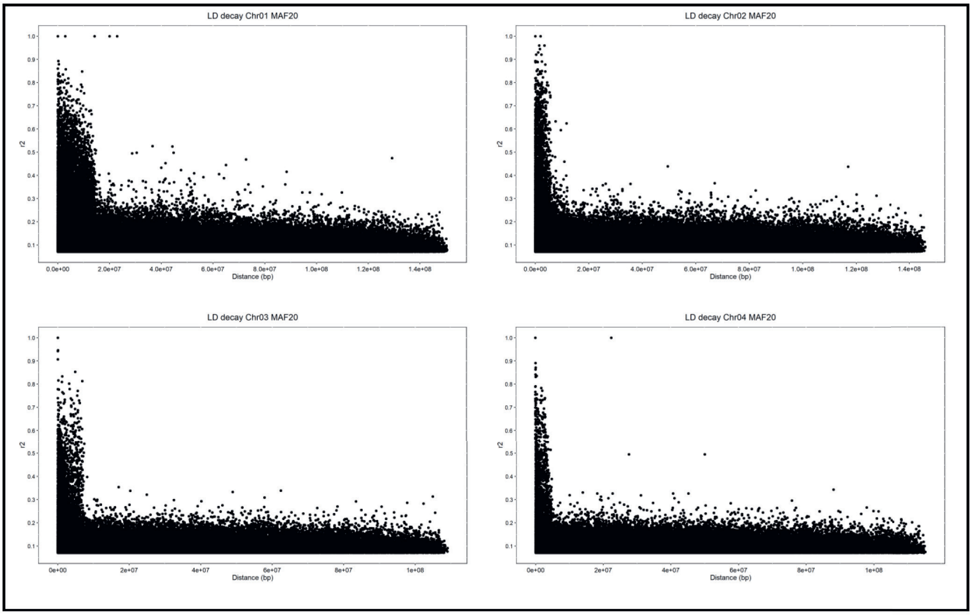
Supplementary Figure 3.

Amino acid substitutions and INDELs of *BdCESA7* between the reference brachypodium genome and the BdTR9 assembly. A) Multiple protein sequence alignment displaying substitutions and INDELs. Colouring of amino acids reflects amino acid polarity, and the Cellulose synthase protein domain is annotated. B) Predicted 2D protein structure of the reference *BdCESA7* protein. C) Predicted 2D protein structure of the *BdCESA7* protein from line BdTR9.



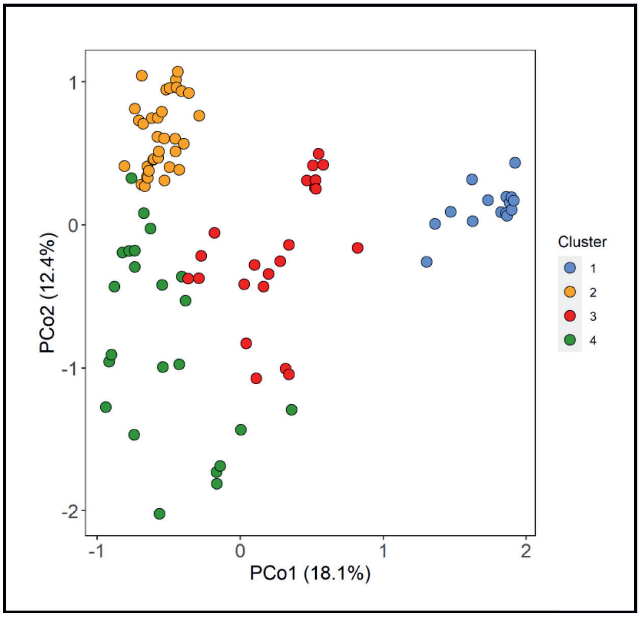
Supplementary Figure 4.

Amino acid substitutions and INDELs of *AtC3H14*, *BdWRKY12*, and *OsNAC43* between the reference genome assemblies and multiple target genome assemblies displaying polymorphisms for the proteins coded by these genes. Colouring of amino acids reflects amino acid polarity, and the functional protein domains are annotated. **A)** *AtC3H14*; **B)** BdWRKY12; **C)** *OsNAC43*.



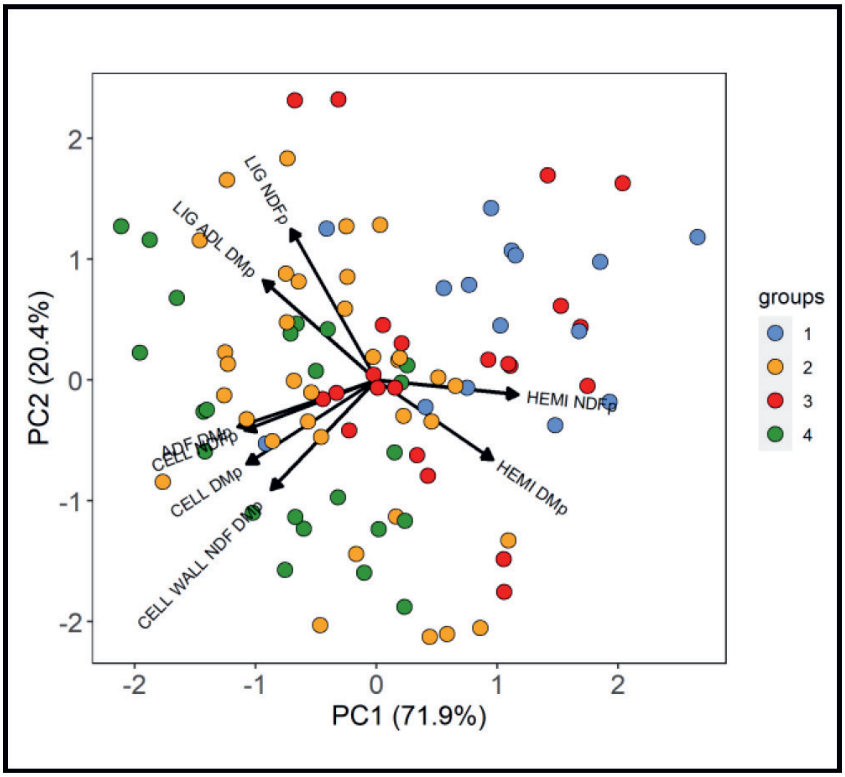
Supplementary Figure 5

Linkage disequilibrium decay plots of the first four *Miscanthus sinensis* chromosomes.

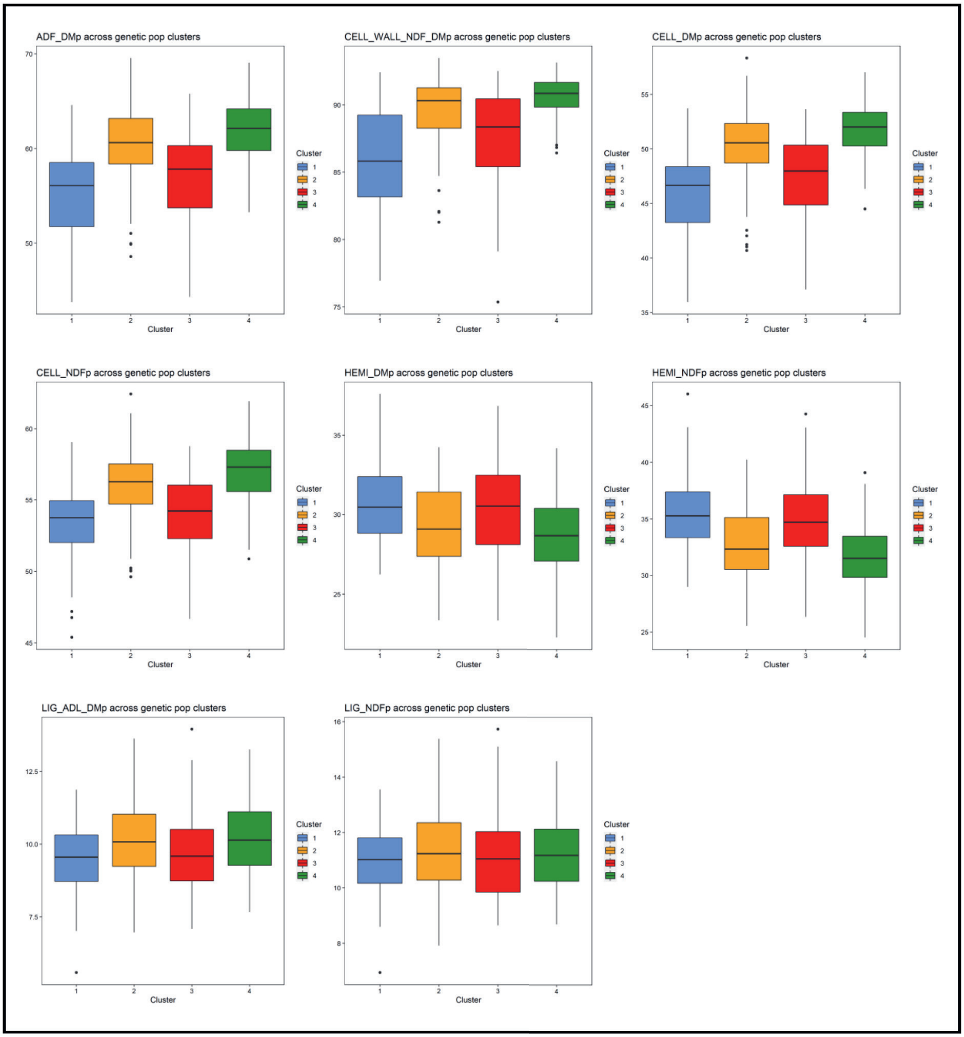


Supplementary Figure 6.

PCoAof the kinship matrix of all the miscanthus GWAS accessions, representing the population structure within the miscanthus collection. Four main population structure groups emerge (different colours). The definition of the groups was based also on hierarchical clustering of the accessions based on SNP data.

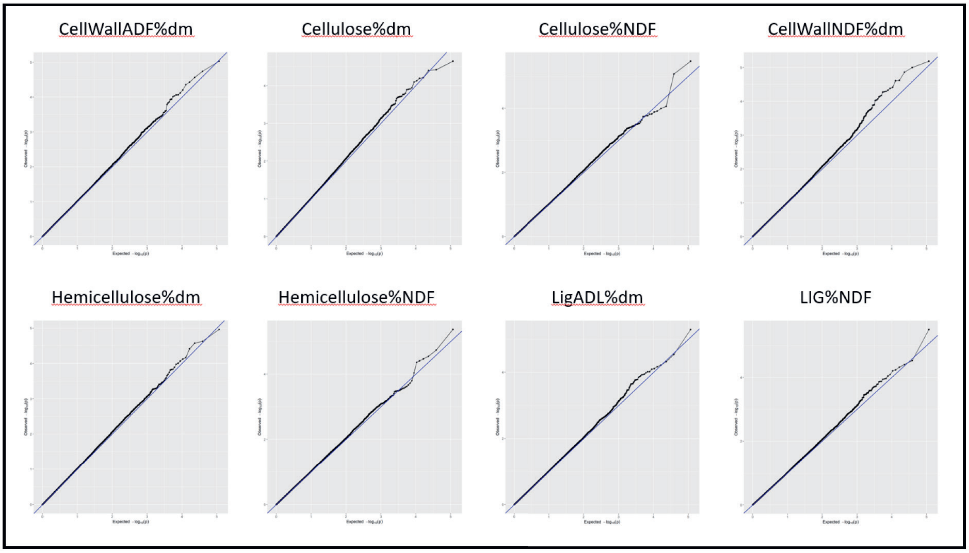


Supplementary Figure 7.
PCA of miscanthus GWAS accessions based on phenotypic data. Accession points are coloured based on the four identified population structure groups (see **Supplementary Figure 6**).



Supplementary Figure 8.

Phenotypic variability per trait across the four population structure groups. Colors reflect the population structure groups defined as in **Supplementary Figure 6**.



Supplementary Figure 9.
QQ-plots of expected (x-axes) vs observed (y-axes) p-values of SNP associations after correction for population structure.

The following supplementary data can be accessed online at https://doi.org/10.3390/plants12040779

Supplementary Table 1: Missing SQTL genes across all the target assemblies analysed.

Supplementary Table 2: All the data on SNPs and INDELs found on SQTL genes.

Supplementary Table 3: The QTL intervals identified through the GWAS in miscanthus.

Supplementary Table 4: The miscanthus SQTLs intervals.

Supplementary Table 5: The Switchgrass cell wall QTLs identified by Ali et al. (2020).

Supplementary Table 6: The Switchgrass SQTLs intervals.

Supplementary Table 7: The 17 "extra QTLs" mapped in the GWAS run without population structure correction.

Supplementary Table 8: The genome assemblies used in the comparative genomics study.

Supplementary Table 9: Full phenotypic data used as input for the miscanthus GWAS.

Supplementary Dataset 1 can be retrieved at: https://doi.org /10.4121/21896757

Supplementary Dataset 2 can be retrieved at: https://doi.org / 10.4121/22068593

Chapter 7

General Discussion

1 Foreword

The research performed in this PhD thesis focused mostly on plant cell walls, meant both as important biological entities of plants and as major targets to improve biomass crops toward the needs of a bio-based economy. The different chapters of this book were therefore designed to both close relevant gaps in our knowledge on cell wall biology, and to use such knowledge – along with other types of genetic data – to develop tools to support breeding of under-domesticated biomass crops adapted to marginal lands. These goals were achieved by performing diverse types of genomic and bioinformatic analyses on >150 cell wall gene families and >200 plant genomes. The results obtained unveiled the important role of genomic dynamics impacting genomic gene arrangement to sustain cell wall diversification during plant evolution (**Chapters 2-3**). Moreover, by exploiting the syntenic conservation of the genomic architecture of cell wall quality, it was possible to develop novel tools to make breeding efforts in under-domesticated biomass crops more effective (**Chapters 4-6**). By building on the results achieved, this final chapter discusses the impact of the outcomes of this thesis on our understanding of cell wall biology and on the breeding of under-domesticated crops. Overall, the work of this PhD thesis adds precious insights on our understanding of cell wall evolution. Moreover, it also opens novel prospects for designing strategies to breed biomass crops for a bio-based economy, and to breed under-domesticated crops in general. All these aspects are discussed in the next paragraphs.

2 Novel perspectives on the evolution and genetics of plant cell walls

2.1 The early evolution of cellulose biosynthesis and of plant cell walls

A key aspect of cell wall biology research that is currently not fully resolved entails the specific modes through which cell walls evolved in the first, early-diverging, land plants (Franková and Fry, 2021). In fact, plant species as the bryophytes lack vascular tracheary elements with lignified cells (Sørensen et al., 2010, Norris et al., 2017), while the presence of lignin is a discriminant trait between the primary and secondary cell walls of seed plants (see **Chapter 1**). Therefore, the lack of lignified cells in the mosses questions whether these plants display a true differentiation between primary and secondary cell walls. Moreover, it opens questions on how the cell wall characteristics of angiosperms arose from the cell wall phenotypes of the early-diverging land plants.

Despite the complete absence of lignin in moss cell walls, different cell types of these organisms display large variability in cell wall thickness between different cells. In fact, the cell walls of the support cells of mosses – called stereids – are much thicker than the ones of the (ancestral) water-conducting cells of these species – called hydroids (Hébant, 1977, Kenrick and Crane, 1997, Norris et al., 2017). This diversification might represent an early structural/functional differentiation in the direction of the distinction between primary and secondary cell walls observed in seed plants. Moreover, it largely depends on variability in the cellulose content and properties between these two moss cell types, which is in turn determined by differential specificity and functioning of distinct CesA isoforms (Norris et al., 2017, Li et al., 2022).

This PhD research contains the largest phylogenomic study conducted so far on the *CesA* superfamily, including the *CesA* genes from 242 plant genomes covering evolution from green algae to angiosperms (**Chapter 2**). This analysis revealed that all the *CesA* genes from early-diverging land plants as bryophytes and lycophytes are basal to the primary cell wall *CesA* genes of seed plants (**Figure 5, Chapter 2**). Moreover, they are specifically positioned closest to the cluster of primary *CesA* genes of angiosperms containing the redundant arabidopsis CesA2, CesA5, CesA6 and CesA9 isoforms. Overall, this result suggests that only primary cell walls were formed in the cells of bryophytes and lycophytes, and that homo-oligomeric *CesA* complexes (i.e. complexes of CesA enzymes formed by only equal isoforms) were responsible for the synthesis of these cell walls. These considerations have important consequences for our understanding of cell wall biology. On the one hand, they imply that the variability observed in the thickness of the cell walls of species as *Physcomitrella patens* is confined within the boundaries of seed plants primary cell walls. On the other hand, they confirm a recent theory proposed to explain the diversification of plant *CesA* genes from bryophytes to angiosperms based on a constructive neutral evolution (CNE) model (Haigler and Roberts, 2019).

The CNE theory of Haigler and Roberts (2019) postulates that the distinct isoforms of the hetero-oligomeric cellulose synthase complexes (CSCs) responsible for cellulose synthesis in the primary and secondary cell walls of angiosperms arose from the expansion, diversification, and functional stabilization of an initial pool of highly similar *CesA* genes that formed homo-oligomeric CSCs. In our view, the bryophyte and lycophyte *CesA* genes that are all basal to the angiosperm primary cell wall CesA isoforms of cluster 1 of **Figure 5 in Chapter 2** could likely represent this early stage of the CNE of plant *CesA* genes. In turn, the cycles of recurrent expansion and functional stabilization into hetero-oligomeric CSCs might have produced intermediate complexes in the ferns, and later in gymnosperms and angiosperms. This is because the *CesA* genes from these plant clades are gradually more phylogenomically diversified across the fern- and angiosperm-specific clusters of **Figure 5 of Chapter 2** as compared to the bryophyte and lycophyte *CesA* genes.

With regard to what just discussed, it is noteworthy that Li et al. (2022) functionally proved the existence of homo-oligomeric CSCs in *P. patens*, which are specifically involved in cellulose synthesis in the hydroid cells (Li et al., 2022). *P. patens* is up-to-date the only land plant where the occurrence of homo-oligomeric CSCs is documented, highlighting that (some of) the primary-like *P. patens CesA* genes are indeed involved in homo-oligomeric CSCs and likely represent a starting point in the evolution of all other plant CesA isoforms. However, Li et al. (2022) also demonstrated the occurrence of *P. patens* hetero-oligomeric complexes, next to the homo-oligomeric ones, which are specifically involved in synthesizing cellulose in the thicker cell walls of stereids. This observation shows that genetic and functional *CesA* diversification already occurred in *P. patens*, at least for some CesA isoforms (Li et al., 2022). In this regard, the extremely wide sampling of plant species for the *CesA* phylogenomic analysis performed in **Chapter 2** allows for contextualizing this diversification at proper evolutionary scales. Specifically, **Figure 5 of Chapter 2** reveals that, despite being all placed at the basis of angiosperms primary *CesA* genes, *P. patens* CesA isoforms are subdivided into two distinct minor clades that are likely responsible for the different *P. patens* CSCs as described by Li et al. (2022). However, this diversification is of much lower magnitude than the one observed between the different *CesA* components of the hetero-oligomeric CSCs of ferns and higher plants. The evidence for hetero-oligomeric CSCs in *P. patens* (Li et al., 2022) highlights that both these different degrees of differentiation can lead to the formation of hetero-oligomeric complexes, even if the evolutionary trajectories responsible for their occurrence in *P. patens* and angiosperms are probably distinct (Li et al., 2019). However, the differential degree of genetic differentiation of *CesA* genes from basal plants to angiosperms appears in line with the higher degree of cell wall complexity in the latter group of plants, where different CesA isoforms are specific to either *true* primary or *true* secondary cell walls (Turner and Somerville, 1997, Arioli et al., 1998, Taylor et al., 2004).

In conclusion, based on the phylogenomic data shown in **Chapter 2** and the current scientific knowledge, it is possible to state that CNE as theorized by Haigler and Roberts (2019) has been the driver of *CesA* diversification into multiple conformations of CSCs and associated cell wall phenotypes. However, such CNE has been a highly complex process, and gave rise to multiple levels of heterogeneity of the CSC members at different plant evolutionary stages. Specifically, even minor genetic variability in the CSC constituents as compared to the one observed among angiosperms might be sufficient to sustain the formation of hetero-oligomeric CSCs and specific phenotypic cell wall changes – as the thicker and thinner cell walls of stereids and hydroids (Li et al., 2022). Nevertheless, it is likely that the level of genetic diversification observed in *P. patens*, together with the absence of lignin, was not yet sufficient to sustain the formation of truly secondary cell walls in this species, or to evolve toward a higher level of genetic complexity in a mutual evolution with more diversified cell walls. Irrespectively of the specific paths followed by *CesA* evolution, both our results and scientific literature highlight a correlation between the genetic and functional diversification of *CesA* genes, and the phenotypic diversification of cell walls at different scales of complexity from bryophytes to angiosperms (Turner and Somerville, 1997, Arioli et al., 1998, Taylor et al., 2004, Li et al., 2022). As such, we foresee that detailed functional characterization of CSCs in other bryophytes than *P. patens*, as well as in lycophytes and ferns, will be key to finely associate *CesA* diversification with different levels of cell wall variability that took place during plant evolution. In turn, this information might become crucial to genetically modulate cell wall – and specifically cellulose – properties in industrial biomass crops by genetic modification of CSCs that mirror specific plant evolutionary trajectories.

2.2 Genomic reshuffling as a driver of plant cell walls diversification

Most of the fundamental research on cell wall biology that was performed in this PhD thesis entailed the study of the genomic dynamics at the basis of plant cell walls diversification, by taking an on-purpose large-scale approach in terms of the number of genomes and gene families analysed (**Chapters 2 and 3**). A main aim of this research was to identify major dynamics that shaped the genomic architecture of cell walls during plant evolution. In this sense, the results obtained indicate that genomic reshuffling (i.e. all mechanisms of genomic/chromosomal rearrangement) leading to changes of genomic contexts of cell wall genes between plant clades, as well as the parallel phylogenetic diversification and/or differential duplication of genes in those genomic contexts, represent major modes by which the genomic background of plant cell walls evolved. This conclusion is supported by several findings presented in this thesis. For example, the analysis of *CesA* genes in **Chapter 2** revealed that a reciprocal translocation of three primary and five secondary *CesA* genes in cotton relative to other angiosperms could underpin the extreme cellulose deposition observed in cotton fibers (**Figure 3C of Chapter 2**). Moreover, correlation between the diversification of genomic gene contexts and patterns of gene sub-functionalization were found for the *CslD* genes (**Figure 4 of Chapter 2**). Furthermore, the duplication of an entire genomic segment containing a *CslD* gene and the subsequent diversification of genes across the duplicated genomic contexts initiated the *CslF* family in Poaceae (**Figure 6 of Chapter 2**). Finally, correlations between the differentiation of multiple genomic properties between Poaceae and eudicots and specific phenotypic characteristics of either type I or type II cell walls were extensively found across 150 cell wall gene families analysed in **Chapter 3**.

The existence of synteny across plant species – which refers to the colinearity of genes across different genomes (Liu et al., 2018a) and allows to infer conservation of genomic gene contexts across species (Dewey, 2011) – was shown before this PhD thesis for multiple cell wall gene families and specific plant clades. For example, Wang et al. (2013a) showed that several *PECTIN METHYL-ESTERASE* (*PME*) genes display synteny between multiple pairs of angiosperm species. Similarily, Schwerdt et al. (2015) revealed that members of both the *CesA* and *CslF* gene families are highly syntenically conserved across the genomes of four grasses. Nevertheless, the finding that the *diversification* of the syntenic organization of cell wall genes – and, thus, of their genomic contexts (Dewey, 2011) – across plant species correlates with key events of plant cell walls diversification that occurred during plant evolution is a novel and relevant outcome of this thesis. Similarily, the scale at which conserved and divergent genomic patterns of cell wall genes were described in **Chapters 2 and 3** – with more than 200 total genomes and 150 gene families surveyed – is unprecedented, and allowed the inference of detailed insights on cell wall genomic dynamics.

In the context just described, the results presented in **Chapter 3** in relation to the differentiation of the type I and type II cell walls of eudicots and Poaceae are some of the most relevant ones. Overall, they clearly show that a profound genomic diversification underlies these two cell wall types. While it is evident that type I and type II cell walls are highly biochemically divergent (Carpita and Gibeaut, 1993, Vogel, 2008, Penning et al., 2019), and specific biochemical dynamics as the recycling or the post-Golgi modification of cell wall polysaccharides sustain such diversification (Penning et al., 2019, Kozlova et al., 2020), the existence and extent of genomic variability underlying type I-II cell wall diversification is still largely debated (Penning et al., 2019, Yokoyama, 2020). In this scenario, our results confirm and bring to a new level the hypotheses advanced by the works of Yokoyama and Nishitani (2004), Xu et al. (2009), Little et al. (2018), or Bulone et al. (2019), which all assume an important role for the (evolutionary) genomic dynamics of cell wall genes – beyond simple presence-absence variation, as in the case of the *CslF* family (Burton et al., 2006) – in shaping cell wall diversity (especially between Poaceae and eudicots). Specifically, our results provide evidence to hypothesize a complete evolutionary model that can explain the trajectories followed by cell wall evolution in Poaceae and eudicots, beyond the level of single gene families, and based on multiple genomic and genetic features. Specifically, we believe that macro-evolutionary events as the complex patterns of genomic rearrangements that occurred within the Poaceae family since ~70 Mya (Lee et al., 2020) could have provided the foundation for the large-scale genomic differentiation of cell wall genes that was reported for this plant clade as compared to eudicots in **Chapter 3**. Following large-scale genomic rearrangements, novel genomic gene contexts could have been stabilized for multiple cell wall gene families in the early genomes of the Poaceae clade, where they could have played an important role for the adaptation to the novel ecological niches that grasses colonized (Strömberg, 2011). In parallel, novel genomic configurations for multiple grass cell wall gene families could have determined and subsequently tightened specific pleoiotropic relationships among cell wall genes themselves, for example through specific expression modes of the genes interested by such configurations (Dewey, 2011, Kondrashov, 2012). Finally, this would have further stabilized the grass cell wall type into the characteristic phenotypes of type II cell walls, through mechanisms as the expansion or reduction of the gene families sizes in specific genomic gene contexts. These types of patterns were extensively found among the gene families analysed in **Chapter 3**, and are well known mechanisms of gene and trait evolution, also beyond plant cell walls (Demuth and Hahn, 2009, Kondrashov, 2012, Zhao et al., 2017, Kerstens et al., 2020).

The evolutionary model just described clearly highlights genomic reshuffling as a major driver for the diversification of plant cell walls. However, it is important to note that functional evidence is needed in the future to provide a definitive proof that the genomic patterns of **Chapters 2 and 3** are undoubtely at the basis of the phenotypic cell wall differentiations described in the same chapters. In this perspective, the cloning of the genes involved in the genomic dynamics described in this PhD thesis, their repeated transformation at random genomic positions in target plants, and the evaluation of the effects of those transformations could represent an effective strategy for providing such a proof. In fact, such an approach could reveal that the sole change of genomic gene contexts can cause phenotypic differences related with the function of specific cell wall genes. Further, the reciprocal translocation between primary and secondary *CesA* genes that was found in cotton (**Chapter 2**) could be used as a model system to functionally evaluate the effect of changes of genomic gene contexts on cell wall characteristics. Specifically, the transformation of the cotton *CesA7* and *CesA8* genes that are placed in a genomic context typical of primary cell wall *CesA* (**Figure 3 of Chapter 2**) back to the secondary cell wall *CesA* genomic context commonly observed in other angiosperms could readily reveal if such a shift of genomic gene context can be indeed responsible for the particularly high activity of these genes during cotton fiber deposition (Li et al., 2013, Li et al., 2016a). Such a result would not only provide a strong proof of how genomic contexts diversification can functionally affect cell wall phenotypes, but would also deliver significant insights from an application point of view. This is because changes in genomic contexts of *CesA* genes similar to the one observed for the cotton *CesA7* and *CesA8* members could be potentially replicated in other fiber crops by cis- or trans-genesis, to modulate fiber production toward higher amounts or better properties. To conclude, a last aspect that also remains to be functionally demonstrated entails whether different genomic contexts of cell wall genes can indeed produce different gene expression profiles in Poaceae and dicots. To this aim, specific gene configurations typical of either Poaceae or eudicots could be engineered in eudicot or Poaceae model species, respectively. Subsequently, comparative gene expression profiling could be conducted on the transformed plant lines of eudicot and Poaceae species simultaneously, under equal experimental systems, and over the whole plant cycle. In this regard, the genes identified in **Chapter 3** provide relevant candidates to be considered for this type of studies, including the *BLH6*/*BLH9* genes and related pathway, or the members of the *XXT*, *CslC*, and *IRX* gene families. In fact, these genes were not only found to be differentially genomically organized in Poaceae and eudicot plants, but are also known to be involved in metabolic pathways that lead to somewhat highly divergent phenotypic effects on the cell walls of Poaceae and eudicots.

3 From genome synteny to novel tools for a more efficient breeding of under-domesticated (biomass) crops

3.1 Syntenic QTLs are novel versatile breeding tools

Next to the fundamental study of the genomic properties underlying plant cell walls, a major goal of this PhD thesis was to develop tools that allow for a faster and more efficient improvement of under-domesticated biomass crops for marginal lands. As discussed in **Chapter 1** and **Chapter 4**, the availability of these tools would be a precious asset to boost the development of bio-based value chains that do not compete with food production (Clifton-Brown et al., 2018, Von Cossel et al., 2019). Moreover, if such “universal” breeding tools could be applied to multiple traits and crops, they would greatly benefit the whole plant breeding sector, by allowing for a more effective exploitation of the bulk of wild material and under-domesticated accessions/species that is so important to introgress (novel) traits to existing plant varieties and/or develop new crops (Salentijn et al., 2007, Kamei et al., 2016, Tadele, 2019). To address this research goal, syntenic QTLs (SQTLs) for cell wall quality traits were developed in **Chapters 5 and 6**.

The concept behind SQTLs is simple: exploiting the syntenic conservation of genomes across species to map relevant loci responsible for variability in target trait(s) in novel crops, by using the genetic information available in model species. The idea of SQTLs came up thanks to previously available scientific observations that the loci responsible for biomass quality appeared genomically conserved in groups of related species. For example, Wang et al. (2013b) discovered that genomic loci associated to cell wall quality largely co-localize with conserved genomic regions between maize and sorghum. Similarily, van der Weijde et al. (2017b) observed synteny between

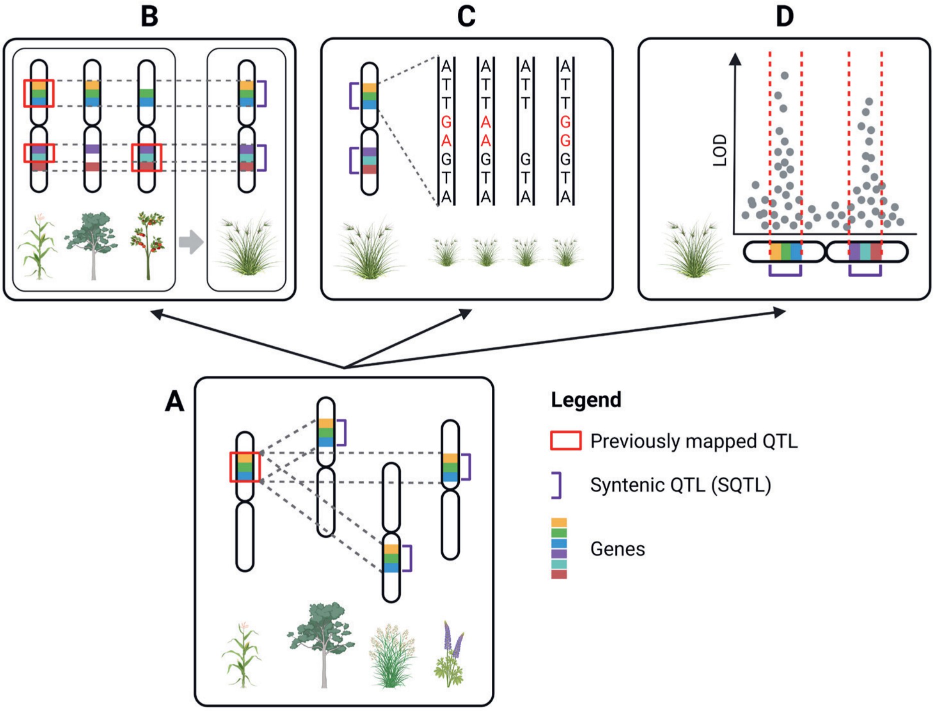


Figure 1. Syntenic QTLs (SQTLs) and their applications for breeding research. A) SQTLs are genomic regions that are syntenically conserved across different species – possibly including both closely-related and evolutionary-distant plant species – and co-localize with previously mapped QTLs in at least one (model) species. B-D) Possible SQTLs applications in breeding research. B) SQTLs can be used to reconstruct the genetic architecture of a trait of interest in a (novel) plant species. This way, cost- and time-consuming pre-breeding steps can be potentially skipped. C) SQTLs can guide the screening of (favourable) allelic variation at target loci controlling a trait of interest in plant populations, e.g. combined with targeted sequencing methods. D) SQTLs can aid the detection of relevant loci at the basis of a trait of interest in genetic mapping studies, or the refinement of genetic mapping results, by analysing the co-localization and genomic co-linearity between mapped loci and previously identified QTLs from other species.

miscanthus QTLs for biomass quality traits and a set of independently mapped sorghum QTLs for similar traits (Gifford et al., 2015). While these observations promised success for a strategy aimed at using genome colinearity to map relevant genetic loci across species, the possibility to implement such a strategy in a structured and scalable pipeline was not an option at the start of this thesis. Moreover, the feasibility of detecting co-linearity of genetic loci across evolutionary-distant species was also questionable. **Chapter 5** closes this methodological gap, by proposing and implementing an effective strategy for SQTL detection. The strategy proposed can make use of a large number of initial QTL regions and of plant genomes to project relevant loci associated to cell wall quality between large sets of crops, including closely-related and evolutionary-distant species.

The possibility of mapping SQTLs opens novel prospects for a faster breeding of under-domesticated (biomass) crops (**Figure 1**). Specifically, the research of **Chapters 5 and 6** demonstrates that SQTLs can be effectively used to carefully reconstruct the (conserved) genomic architecture of a trait of interest across plant species, as they allowed for the mapping of several syntenically-conserved cell wall genes involved in multiple initial QTLs across different species. When aiming at developing a novel crop toward specific breeding ideotypes, the genomic regions identified through SQTLs could be easily extended to that new crop, by simply sequencing the genome of the target novel species and performing synteny analysis between SQTLs and the new genome. For this purpose, standard programs for synteny computation as MCScanX (Wang et al., 2012b) or DAGChainer (Haas et al., 2004) might be used. Alternatively, in the case good nucleotidic conservation of SQTLs in the new crop is expected, programs for the nucleotidic alignment of whole genomic regions could also be used, as e.g. MUMMER (Marçais et al., 2018), which do not require prior gene annotation and/or inference of protein sequences over the sequenced target genome. The mapping of SQTLs in a new crop would in turn allow for the targeted sequencing of those regions across large germplasm collections, by using technologies as the Allegro-SPET system (Scaglione et al., 2019). This step would allow the identification of favourable SNPs or even complete allelic gene variants at known relevant genomic loci conserved through SQTLs. This is similar to what was achieved in **Chapter 6**, where the intra-specific analysis of SQTL regions through the use of available genomic resources revealed the existence of allelic variation at important cell wall genes, as the rice *Brittle culm-like 8* locus. Since **Chapter 5** also demonstrated that SQTL regions can (partly) predict the localization of relevant QTLs in miscanthus and switchgrass, the knowledge about favourable alleles at SQTL loci in target accession panels could be used to select plant lines to be advanced in a breeding program by minimizing phenotyping efforts. Overall, the combination of targeted sequencing and of minimized phenotyping would significantly reduce the costs and time of a breeding program in a novel crop, achieving the goal of a faster and more efficient breeding of under-domesticated (biomass) crops, which drove the SQTL research of this thesis.

In addition to the validity of SQTLs for making screening and selection of material in breeding programs more effective, these tools can also assist and improve “standard” pipelines of trait analysis. In this regard, **Chapter 5** showed that SQTLs can help in overcoming limitations of mapping approaches. Specifically, SQTL regions allowed the identification of potentially significant loci from a GWAS for cell wall quality traits in miscanthus that would have normally been removed from GWAS results by applying a – needed – correction for population structure (**Chapter 5**). This represents an important result, since correlation between population structure patterns and phenotypic variability among accessions can lead to overcorrection and removal of relevant genomic loci from mapping results (Brachi et al., 2011, Lawson et al., 2020). In this respect, the availability of SQTLs in the hands of a plant breeder could allow for the genomic analysis of extra loci from a mapping study because of their co-localization with perfectly conserved QTLs for similar traits previously mapped in other species – as is the case for the 17 "extra" QTLs identified in **Chapter 5** in miscanthus. This information, when complemented by relevant information on the function of genes included in such "extra" loci, could lead to the selection of extra material for a breeding program that could have otherwise been discarded. This can represent a great advantage for improving the accuracy of the breeding process in different crops and make sure that selection models are based on the most of information that can be retrieved from the results of mapping studies.

3.2 The strategy for syntenic QTLs detection: methodological considerations

The previous paragraph discussed how SQTLs represent useful tools for more efficient breeding activities, mainly in under-domesticated crops. Nevertheless, it is important to note that the validity of SQTLs as breeding tools comes with important considerations about the methodology developed for their identification in **Chapters 5 and 6**.

A first important element to discuss is the fact that SQTLs can be developed only for genomic segments involved in previously mapped QTLs that are syntenically conserved between species. However, non-syntenic portions of QTLs used for SQTLs detection could also contain – or constitute themselves – highly important genetic elements at the basis of a trait. Specifically, this appears particularly relevant as previous sections of this chapter discussed how genomic reshuffling – including diversification of synteny patterns across species – appears a major factor underpinning phenotypic cell wall variability in plants. Moreover, several studies highlighted the importance of structural genome variability for trait adaptation, as differential gene duplication or loss (Kondrashov, 2012, Panchy et al., 2016, Flagel and Wendel, 2009), differential genome loading with transposons or mobile elements (Lisch, 2013, Casacuberta and González, 2013), or larger chromosomal and genomic rearrangements between different species clades (Qiao et al., 2019, Zhang et al., 2020). The fact that SQTLs could fail to capture important genetic sources of trait variability due to their reliance on gene synteny might thus appear a major drawback of this approach. In this regard, however, it is important to point out that structural genome variability with an impact on gene order and trait expression tends to increase in parallel to the taxonomic distance of plant species with each other (Stelkens and Seehausen, 2009, Zhao and Schranz, 2019). As such, it is typically found at higher frequency between evolutionary-distant plant clades, as it was the case in **Chapter 3** for several cell wall gene families between Poaceae and eudicots. At smaller evolutionary distances, until the level of intra-specific accessions, more common sources of genetic variability are constituted by different gene alleles determined by single-nucleotide polymorphisms (SNPs) or INDELs, whose detection is at the basis of association mapping (Zhu et al., 2008, Brachi et al., 2011). Because of all these reasons, the failure to detect important loci in novel species when reconstructing the genomic architecture of a target trait with SQTLs is actually expected to occur infrequently. This is because SQTLs were mapped separately for specific groups of plant families, whose definition was based on taxonomic distance (either Poaceae or eudicot species) and on a preliminary assessment of the average degree of synteny among the members of each species group (see **Chapter 5**). Moreover, the synteny of cell wall genes was specifically assessed in different chapters of this thesis, leading to the conclusion that those genes – which include the most likely candidates of the initial QTLs used for SQTLs detection – are highly syntenic, typically at higher degrees than other gene families (see **Chapters 2, 3 and 5**). These two types of analyses – the synteny analysis of predicted candidate genes (when possible) and the assessment of overall QTL synteny across plant families considered for SQTL detection – should always be performed prior to SQTLs mapping for any target trait, to minimize the possibility of not covering relevant target QTL genes in SQTLs because of genomic reshuffling. To conclude, it is also important to point out that the utilization of genome synteny for the projection of genetic elements across species provides also objective advantages to ensure functional conservation of initial QTL regions between species. This is because a prerequisite of genome synteny is the homology of genes in syntenic regions, on top of which synteny itself ensures the exact similarity of longer genomic stretches on which genes are located (Dewey, 2011, Liu et al., 2018a, Kerstens et al., 2020). The latter is typically a sign of conserved genomic and gene regulation (Dewey, 2011, Kerstens et al., 2020). Therefore, similar gene alleles detected on syntenic regions between species reasonably maximize the chance of translating knowledge between species with a functional meaning for a target trait.

Another important methodological consideration regarding SQTLs identification entails the quality of the genome assemblies to be used during SQTLs detection, and also of novel target species if one aims to use SQTLs for trait mapping in novel crops. In this sense, it is important to note that both the coverage and the fragmentation (i.e. N50 statistics) of genome assemblies are critical parameters that can significantly affect the precision of inter-projecting genetic information between species by using SQTLs. This is because incomplete or highly fragmented genome assemblies could lead to the failure of detecting syntenically conserved genetic elements in initial QTLs, because those elements are simply missing in a genome assembly, or are harboured by short scaffolds for which synteny computation may fail (Liu et al., 2018a). At the time the genomes to be used for our analyses were retrieved (late 2018 and early 2019), several releases of gymnosperm genomes were showing both the issues just mentioned, especially genomic fragmentation, with an average number of genes per scaffold typically <2 and a representation of BUSCO Viridiplantae genes sometimes <40%. To circumvent these issues, we therefore decided to apply stricter thresholds for including genomes in the pipeline for SQTLs detection, including a genomic representation of the BUSCO Viridiplantae genes >75% and and an average number of genes per scaffold >10 over a full genome assembly (see **Chapter 5**). We believe that these parameters should be carefully considered when performing SQTL detection, as their relaxation could significantly lower the quality of the results obtained. As an example, in **Chapter 2** it was not possible to infer synteny dynamics between angiosperms and gymnosperms exactly because of the high fragmentation of the gymnosperm genome assemblies. Thus, the inclusion of similar assemblies in SQTLs detection would certainly lead to the problems raised above. In general, it is important to note that genome assemblies developed by using the newest long-read sequencing technologies allow for much improved assemblies quality (Amarasinghe et al., 2020). For example, re-sequencing of the gymnosperm *Ginkgo biloba* genome using Single-Molecule Real Time (SMRT) technology led to a massive improvement of the genome assembly, increasing the N50 metric of this highly-repetitive and extremely large genome from 48 kb to 775 Mbp (Liu et al., 2021). Such an improvement would certainly greatly positively impact in the estimation of synteny between this species and other plants. As such, the preference for using genome assemblies produced with the last sequencing technologies should be carefully considered when approaching SQTLs detection.

3.3 SQTLs and cell wall quality: targets and strategies to breed biomass crops for marginal lands

In addition to conceptualizing SQTLs, this PhD thesis also identified and characterized SQTLs for cell wall quality across a large panel of plant species – including multiple biomass crops – with the aim of providing effective tools to breed biomass crops for marginal lands. This research revealed that cell wall genes are overall highly syntenic between the species targeted in this study, and several important ones are contained in previously mapped cell wall QTLs and conserved through the SQTLs of **Chapters 5 and 6**. In this regard, an important outcome of the thesis it that SQTLs appeared enriched of highly-conserved cell wall transcription factors, including master regulators as different *VND* and *NAC/NST* factors, the *WRKY12* gene, and several cell wall *MYB* transcription factors (**Chapter 5**) (Zhong et al., 2010, Taylor-Teeples et al., 2015). All these genes were previously shown to be responsible for the occurrence of cell wall QTLs in different species, as maize (Barrière et al., 2012, Barrière et al., 2015, Barrière et al., 2017), sorghum (Wang et al., 2013b), poplar (Ranjan et al., 2010), or potato (Yogendra et al., 2017), making them important targets for designing standardized strategies of crop improvement in (under-domesticated) biomass crops, as proposed in **Chapter 4**. Specifically, SQTLs could be effectively used to identify these genes in new crops, even at the isoform level, and to identify relevant alleles for breeding by following the approaches proposed in the previous sections. In this perspective, in **Chapter 5** we showed that SQTLs successfully identified the exact positional orthologs of the arabidopsis *VND7* gene – which is the most important *VND* factor in terms of expression and cell wall effects (Yamaguchi et al., 2011, Endo et al., 2015) – in several eudicot plants, including biomass species as poplar, eucalyptus, or salix (**Figure 6 of Chapter 5**). Moreover, a *VND7* homolog from miscanthus was contained in a QTL identified by GWAS in **Chapter 6** which spanned a genomic area that was already predicted as relevant thanks to co-localization with SQTLs (**Figure 6 of Chapter 6**). Finally, the genomic research performed in **Chapter 6** showed that a *WRKY12* gene from brachypodium spanned by SQTLs display allelic variability with effects on the polarity of the WRKY protein domain (**Supplementary Figure 5 in Chapter 6**). Interestingly, *WRKY12* can simultaneously affect secondary cell wall properties and flowering time in *Miscanthus lutarioriparius* (Yu et al., 2013), thus representing a good candidate for improving multiple traits related to biomass quality simultaneously. This appears particularly relevant in the context of marginal lands, as this gene appears also involved in conferring tolerance to drought or heavy metals (Han et al., 2019, Shi et al., 2018).

In addition to central cell wall transcription factors as the ones just discussed, the analysis of SQTL cell wall genes in angiosperms revealed other targets for standardized strategies of biomass improvement. For example, several *COBRA* and *COBRA-like* (*COBL*) genes known to be at the basis of maize brown-midrib and rice brittle culm mutants were shown to be highly conserved through SQTLs across Poaceae species (**Chapter 5**). These genes have been recognized as highly relevant to design solid strategies for cell wall and biomass quality improvement in grass species, even beyond maize and rice, which are the species where they have been first discovered and characterized (Barrière et al., 2004, Sattler et al., 2010, Zhang and Zhou, 2011). In this context, SQTLs allowed to again quickly identify the exact positional orthologs of different members of these groups of genes across Poaceae. Moreover, their genomic analysis readily uncovered interesting alleles for the brachypodium homolog of the rice *Brittle culm-like 8* gene, with large rearrangements of the COBRA protein domains. In the future, functional studies could be conducted in brachypodium to evaluate the effective relevance of the alleles discovered for modulating cell wall quality and/or biomass properties. Finally, positive results of these types of analyses could pave the way to the quick selection of functionally-proven alleles at these critical loci in under-domesticated biomass crops, through the application of SQTLs.

The strategy just proposed could be applied to the several other relevant genes identified as conserved through SQTLs in **Chapters 5 and 6**, including central lignin genes as *F5H*, *CSE*, or *CCoAOMT*; homologs of the arabidopsis *IRX* genes; or positional orthologs of the *BLH6/9* transcription factors – which were both predicted by SQTLs in the genomic areas spanned by the QTLs detected by the GWAS in miscanthus (**Chapter 6**) and showed to be critical in the context of the cell wall differentiation between Poaceae and eudicots (**Chapter 3**). Overall, all these genes represent relevant targets for biomass improvement, and the fact that they were found conserved through SQTLs opens prospects for their targeting in "universal" breeding programs to advance under-domesticated biomass crops, following gene identification and allele mining as previously described. In this sense, a concluding important remark entails the expectations that can be harboured concerning the extent of breeding advancement attainable by targeting these genes in under-domesticated crops as proposed. Overall, cell wall quality traits at the basis of cell wall quality are usually highly heritable, as shown in different crops as maize (Torres et al., 2015a), miscanthus (van der Weijde et al., 2017b), poplar (Klasnja et al., 2003), or hemp (Petit et al., 2020b). As large heritability is typically associated with steady trait responses across environments (Acquaah, 2009), it is reasonable to expect that similar effects for the selection of same alleles through SQTLs across crops. This would be highly positive for applying SQTLs to the improvement of biomass crops in standardized breeding strategies. However, differential pleiotropic gene effects across species could also occur (Burton and Fincher, 2014b, Garzon et al., 2022), which could potentially lead to variation in the effect of same genes across crops also in cases of maximized heritabilities.

4 From syntenic QTLs to the future of plant breeding

4.1 Integrating multiple large data sources for better trait improvement

A long-lasting research goal in plant breeding is represented by the development of methods to increase the speed, efficiency, and accuracy of breeding programs. This goal (in)directly guided most of the advancements achieved in plant breeding since when humans domesticated the first plants ~10000 years ago, at the dawn of agriculture (Hancock, 2012). In this context, the Mendelian genetic laws represented a major milestone in the history of plant breeding, as they provided the theoretical framework to enable the first genetically-informed selection strategies to maximize varietal gains (Lee et al., 2015). In turn, the study and use of genetics in the context of plant breeding paved the way to several breakthroughs that changed the modes and the efficiency of breeding practices. These include the development of hybrids, the exploitation of heterosis, the design of increasingly more efficient schemes of conventional selection as backcrossing and recurrent selection, the development of molecular markers and of marker-assisted selection, the implementation of association mapping strategies, and the development of methodologies of genetic modification and engineering of plant traits (Reeves and Cassaday, 2002, Acquaah, 2009, Lee et al., 2015). Altogether, these milestones brought the plant breeding sector to its current status, a point in which the effective integration of multiple and heterogeneous (big) data sources coming from high-throughput analyses of genotypes, phenotypes, and environments is envisioned as key to further advance breeding programs (Harfouche et al., 2019, Kim et al., 2020a, Niazian and Niedbała, 2020).

At the core of the use of big data in plant breeding stands the fact that high-throughput technologies for genotyping – genome sequencing and markers mining – and, more recently, phenotyping – automated and continuous analysis of multiple plant traits in lab and field experiments – are becoming increasingly versatile, precise, scalable and cheaper (Shakoor et al., 2019, Yang et al., 2020, Purugganan and Jackson, 2021). Thus, they open novel prospects for the early selection of plant material, to finely predict plant performance over multiple environments, genotypes, and/or species, or to screen wild plant material for trait introgression (Harfouche et al., 2019, Kim et al., 2020a, Xu et al., 2022). It is evident that the effective implementation of these strategies in breeding programs would enourmously increase the speed and reduce costs of breeding practice. Nevertheless, a key challenge of this prospect entails the development of methods and platforms to effectively allow the combined analysis of big data sources from different trials, materials and experiments in standardized pipelines to support breeding (Harfouche et al., 2019). In this regard, the work of this PhD thesis represents a possible way to go. Specifically, we think that the demonstration of the possibility to integrate large genomic and genetic data (i.e. numerous genome assemblies and QTLs) from different studies in a single analysis that led to the successful extraction of coalesced information on a complex trait as biomass quality represents a highly valuable achievement. In fact, this shows both the attainability and the relevance of integrating large data sets for better trait improvement.

Nevertheless, we foresee that other steps need to be taken in the future to equip breeders with a complete toolbox for precise data-driven breeding. These steps include the inclusion of phenotypic data in pipelines as the one developed in this PhD thesis, so that effects of collected QTLs and of alleles mined can be better weighted over each other in the context of target traits, achieving higher accuracy in the reconstruction of the genomic architecture of a trait, in an informative way for selection. Moreover, pipelines as the one developed in this PhD thesis could be transformed into structured platforms that require only raw data on genomes, QTLs and phenotypes as inputs. Finally, the development of protocols to standardize genomic and phenotypic data uploaded onto public repositories in a way that allows their instant merged processing will also represent a key element to fully exploit the potential of big data integration for plant breeding. To conclude, we believe that the realization of these steps will deliver tools in the hands of plant breeders that will effectively enable the data revolution of breeding practices that is largely advocated.

4.2 More efficient breeding strategies to solve key challenges of our times

The efficient and large-scale exploitation of genotypic, phenotypic, and environmental high-throughput (big) data from multiple crops, trials, and traits in integrative and innovative breeding strategies would certainly benefit the whole plant breeding sector. Nevertheless, these approaches could represent major breakthroughs especially for some of the main challenges currently faced by agriculture. These include the need for quick and flexible adaptation of crops to climate change (Harfouche et al., 2019, Anderson et al., 2020), the fight against biodiversity loss (Ceccarelli, 2009), or the capacity to sustain bio-based circular economic systems where agriculture supplies energy and materials next to food and feed (Trindade et al., 2010). This is because plant breeding is key to solve these challenges, by tailoring crop varieties to specific environmental conditions, by developing multi-purpose crops that can simultaneously satisfy the production of food and industrial compounds, and by discovering and introgressing unexploited traits from wild crop relatives in crop varieties (Trindade et al., 2010, Anderson et al., 2020). However, as discussed in **Chapter 4**, the complexity of traits as the resistance to drought, heat, or floodings, the long time and large costs of (pre-)breeding activities in under-domesticated crops, and the lack of tools for quick allele screening are slowing down these processes.

In this context, concepts and methods as the ones developed in this PhD thesis are envisioned to offer promising prospects for a faster, cheaper, and accurate improvement of crops toward the objectives and traits above. Faster and cheaper, as entire steps of conventional breeding programs – as the need for large, multi-environment trials – could be downsized or skipped. Moreover, the genetic background of target traits could be hypothetically directly translated to specific crops, starting from model species. This is highly relevant, as research on relevant traits for climate change adaptation takes usually place in model crops, as arabidopsis (Chew and Halliday, 2011), different grasses (Ngara and Ndimba, 2014, Dawson et al., 2015), or species with natural tolerance to specific abiotic stresses as quinoa (Roman et al., 2020), or takes anyway advantage of alleles that are mined in wild crop relatives as they got removed from the gene pools of major crops by genetic bottlenecks (Huang et al., 2016, Mammadov et al., 2018). As such, platforms to support breeding based on multi-data integration could greatly benefit the process of trait or alleles translation between crops. In turn, accuracy of breeding programs could also be improved, as the integration of data from multiple species in tools as SQTLs can allow for a precise identification of relevant genetic elements, alleles, or traits, both in crops and crop relatives. The selected material or the target genetic elements could be then readily included in novel breeding programs or existing breeding schemes based, for example, on recurrent selection (Clifton-Brown et al., 2018). These considerations are extremely important, as a boost in activities for climate change mitigation and adaptation, or for establishing bio-based value chains, is required to keep the environmental degradation of our Planet within boundaries that allow human life (Qin et al., 2021). In this respect, we strongly believe that carrying on research on the integration of genetic and crop data into novel pipelines for approaching data-driven breeding will ultimately lead to such a boost and to a contribution to solving these challenges.

5 Concluding remarks

Starting from the goal of developing novel tools to breed more effectively novel biomass crops for marginal lands, the genomic and genetic data produced in this PhD thesis allowed to gain new insights on the biology of plant cell walls, to set up landmarks for the improvement of (novel) biomass crops, and to envision strategies for more efficient breeding programs in the future. Throughout all the previous sections we have commented on the relevance of the various outcomes obtained, highlighting the biological hypotheses, the methodological considerations, and the applied prospects that they open. While being the precious end-point of this research, all the outcomes obtained represent also a starting point for future studies. Such studies could further test the hypotheses advanced about the evolution of plant cell walls and lead to the implementation in real breeding programs of the tools and concepts developed. Novel appealing prospects are ahead for plant cell walls research and (big-)data-driven breeding of (biomass) crops!

References

ABDEL-HALEEM, H., LUO, Z. & RAY, D. 2019. Genetic improvement of guayule (Parthenium argentatum A. Gray): an alternative rubber crop. *Advances in Plant Breeding Strategies: Industrial and Food Crops.* Springer.

ACHARYA, B. S. & BLANCO-CANQUI, H. 2018. Lignocellulosic-based bioenergy and water quality parameters: a review. *GCB Bioenergy,* 10**,** 504-533.

ACQUAAH, G. 2009. *Principles of plant genetics and breeding*, John Wiley & Sons.

ACQUAAH, G. 2012. Breeding cross-pollinated species. *Principles of Plant Genetics and Breeding.* John Wiley & Sons, Ltd.

AGRAWAL, A., KAUSHIK, N. & BISWAS, S. 2014. Derivatives and applications of lignin–an insight. *SciTech J,* 1**,** 30-36.

ALI, S., SERBA, D. D., WALKER, D., JENKINS, J., SCHMUTZ, J., BHAMIDIMARRI, S. & SAHA, M. C. 2020. Genome-wide quantitative trait loci detection for biofuel traits in switchgrass (Panicum virgatum L.). *GCB Bioenergy,* 12**,** 923-940.

ALLWRIGHT, M. R. & TAYLOR, G. 2016. Molecular breeding for improved second generation bioenergy crops. *Trends in Plant Science,* 21**,** 43-54.

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. 1990. Basic local alignment search tool. *Journal of molecular biology,* 215**,** 403-410.

AMARASINGHE, S. L., SU, S., DONG, X., ZAPPIA, L., RITCHIE, M. E. & GOUIL, Q. 2020. Opportunities and challenges in long-read sequencing data analysis. *Genome biology,* 21**,** 1-16.

AMELEWORK, B., SHIMELIS, H., TONGOONA, P. & LAING, M. 2015. Physiological mechanisms of drought tolerance in sorghum, genetic basis and breeding methods: a review. *African Journal of Agricultural Research,* 10**,** 3029-3040.

ANANDA, N., VADLANI, P. V. & PRASAD, P. V. 2011. Evaluation of drought and heat stressed grain sorghum (Sorghum bicolor) for ethanol production. *Industrial Crops and Products,* 33**,** 779-782.

ANDERSON-TEIXEIRA, K. J., DAVIS, S. C., MASTERS, M. D. & DELUCIA, E. H. 2009. Changes in soil organic carbon under biofuel crops. *Gcb Bioenergy,* 1**,** 75-96.

ANDERSON, R., BAYER, P. E. & EDWARDS, D. 2020. Climate change and the need for agricultural adaptation. *Current opinion in plant biology,* 56**,** 197-202.

ARGAN, G. C. 1968. *Storia dell'arte italiana,* Milano.

ARIOLI, T., PENG, L., BETZNER, A. S., BURN, J., WITTKE, W., HERTH, W., CAMILLERI, C., HÖFTE, H., PLAZINSKI, J. & BIRCH, R. 1998. Molecular analysis of cellulose biosynthesis in Arabidopsis. *Science,* 279**,** 717-720.

ARMSTEAD, I., HUANG, L., RAVAGNANI, A., ROBSON, P. & OUGHAM, H. 2009. Bioinformatics in the orphan crops. *Briefings in Bioinformatics,* 10**,** 645-653.

ARTUR, M. A. S., ZHAO, T., LIGTERINK, W., SCHRANZ, E. & HILHORST, H. W. 2019. Dissecting the genomic diversification of late embryogenesis abundant (LEA) protein gene families in plants. *Genome biology and evolution,* 11**,** 459-471.

ATIENZA, S. G., SATOVIC, Z., PETERSEN, K. K., DOLSTRA, O. & MARTIN, A. 2003. Identification of QTLs influencing agronomic traits in Miscanthus sinensis Anderss. I. Total height, flag-leaf height and stem diameter. *Theoretical and Applied genetics,* 107**,** 123-129.

ATMODJO, M. A., HAO, Z. & MOHNEN, D. 2013. Evolving views of pectin biosynthesis. *Annu Rev Plant Biol,* 64**,** 747-779.

BAILEY-SERRES, J., LEE, S. C. & BRINTON, E. 2012. Waterproofing crops: effective flooding survival strategies. *Plant Physiology,* 160**,** 1698-1709.

BALDACCI-CRESP, F., LE ROY, J., HUSS, B., LION, C., CRÉACH, A., SPRIET, C., DUPONCHEL, L., BIOT, C., BAUCHER, M. & HAWKINS, S. 2020. UDP-GLYCOSYLTRANSFERASE 72E3 plays a role in lignification of secondary cell walls in Arabidopsis. *International journal of molecular sciences,* 21**,** 6094.

BANASIAK, A. 2014. Evolution of the cell wall components during terrestrialization. *Acta Societatis Botanicorum Poloniae,* 83.

BANDARRA, V. D. L. 2013. *Efeito da aplicação de águas residuais na produtividade e na qualidade de três genótipos de Miscanthus.* Faculdade de Ciências e Tecnologia.

BARBOSA, B., COSTA, J., FERNANDO, A. L. & PAPAZOGLOU, E. G. 2015. Wastewater reuse for fiber crops cultivation as a strategy to mitigate desertification. *Industrial Crops and Products,* 68**,** 17-23.

BARNEY, J. N., MANN, J. J., KYSER, G. B., BLUMWALD, E., VAN DEYNZE, A. & DITOMASO, J. M. 2009. Tolerance of switchgrass to extreme soil moisture stress: ecological implications. *Plant Science,* 177**,** 724-732.

BARRIÈRE, Y., COURTIAL, A., SOLER, M. & GRIMA-PETTENATI, J. 2015. Toward the identification of genes underlying maize QTLs for lignin content, focusing on colocalizations with lignin biosynthetic genes and their regulatory MYB and NAC transcription factors. *Molecular breeding,* 35**,** 1-23.

BARRIÈRE, Y., GUILLAUMIE, S., DENOUE, D., PICHON, M., GOFFNER, D. & MARTINANT, J. P. 2017. Investigating the unusually high cell wall digestibility of the old INRA early flint F4 maize inbred line. *Maydica,* 62**,** 1-21.

BARRIÈRE, Y., MÉCHIN, V., LEFEVRE, B. & MALTESE, S. 2012. QTLs for agronomic and cell wall traits in a maize RIL progeny derived from a cross between an old Minnesota13 line and a modern Iodent line. *Theoretical and Applied Genetics,* 125**,** 531-549.

BARRIÈRE, Y., RALPH, J., MÉCHIN, V., GUILLAUMIE, S., GRABBER, J. H., ARGILLIER, O., CHABBERT, B. & LAPIERRE, C. 2004. Genetic and molecular basis of grass cell wall biosynthesis and degradability. II. Lessons from brown-midrib mutants. *Comptes Rendus Biologies,* 327**,** 847-860.

BARRIÈRE, Y., RIBOULET, C., MÉCHIN, V., MALTESE, S., PICHON, M., CARDINAL, A., LAPIERRE, C., LUBBERSTEDT, T. & MARTINANT, J.-P. 2007. Genetics and genomics of lignification in grass cell walls based on maize as model species. *Genes Genomes Genomics,* 1**,** 133-156.

BARTLEY, L. E., PECK, M. L., KIM, S.-R., EBERT, B., MANISSERI, C., CHINIQUY, D. M., SYKES, R., GAO, L., RAUTENGARTEN, C. & VEGA-SÁNCHEZ, M. E. 2013. Overexpression of a BAHD acyltransferase, OsAt10, alters rice cell wall hydroxycinnamic acid content and saccharification. *Plant Physiology,* 161**,** 1615-1633.

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

BENNETT, A. B. & ISAACS, R. 2014. Landscape composition influences pollinators and pollination services in perennial biofuel plantings. *Agriculture, ecosystems & environment,* 193**,** 1-8.

BENNETT, A. B., MEEHAN, T. D., GRATTON, C. & ISAACS, R. 2014. Modeling pollinator community response to contrasting bioenergy scenarios. *PloS one,* 9**,** e110676.

BENNETZEN, J. L. & FREELING, M. 1997. The unified grass genome: synergy in synteny. *Genome Research,* 7**,** 301-306.

BENNICH, T. & BELYAZID, S. 2017. The route to sustainability—Prospects and challenges of the bio-based economy. *Sustainability,* 9**,** 887.

BERARDINI, T. Z., REISER, L., LI, D., MEZHERITSKY, Y., MULLER, R., STRAIT, E. & HUALA, E. 2015. The Arabidopsis information resource: making and mining the "gold standard" annotated reference plant genome. *genesis,* 53**,** 474-485.

BERNAL, A. J., YOO, C.-M., MUTWIL, M., JENSEN, J. K., HOU, G., BLAUKOPF, C., SØRENSEN, I., BLANCAFLOR, E. B., SCHELLER, H. V. & WILLATS, W. G. 2008. Functional analysis of the cellulose synthase-like genes CSLD1, CSLD2, and CSLD4 in tip-growing Arabidopsis cells. *Plant Physiology,* 148**,** 1238-1253.

BERNARDO, R. 2016. Bandwagons I, too, have known. *Theoretical and Applied Genetics,* 129**,** 2323-2332.

BHANDARI, H., SAHA, M., FASOULA, V. & BOUTON, J. 2011. Estimation of genetic parameters for biomass yield in lowland switchgrass (Panicum virgatum L.). *Crop science,* 51**,** 1525-1533.

BHANDARI, H., SAHA, M., MASCIA, P., FASOULA, V. & BOUTON, J. 2010. Variation among half-sib families and heritability for biomass yield and other traits in lowland switchgrass (Panicum virgatum L.). *Crop science,* 50**,** 2355-2363.

BHATIA, R., GALLAGHER, J. A., GOMEZ, L. D. & BOSCH, M. 2017. Genetic engineering of grass cell wall polysaccharides for biorefining. *Plant biotechnology journal,* 15**,** 1071-1092.

BISWAL, A. K., SOENO, K., GANDLA, M. L., IMMERZEEL, P., PATTATHIL, S., LUCENIUS, J., SERIMAA, R., HAHN, M. G., MORITZ, T. & JÖNSSON, L. J. 2014. Aspen pectate lyase Ptxt PL1-27 mobilizes matrix polysaccharides from woody tissues and improves saccharification yield. *Biotechnology for biofuels,* 7**,** 1-13.

BITA, C. & GERATS, T. 2013. Plant tolerance to high temperature in a changing environment: scientific fundamentals and production of heat stress-tolerant crops. *Frontiers in plant science,* 4**,** 273.

BLANCO-CANQUI, H. 2016. Growing dedicated energy crops on marginal lands and ecosystem services. *Soil Science Society of America Journal,* 80**,** 845-858.

BLANCO-CANQUI, H., GILLEY, J. E., EISENHAUER, D. E., JASA, P. J. & BOLDT, A. 2014. Soil carbon accumulation under switchgrass barriers. *Agronomy Journal,* 106**,** 2185-2192.

BLUEMEL, M., DALLY, N. & JUNG, C. 2015. Flowering time regulation in crops—what did we learn from Arabidopsis? *Current opinion in biotechnology,* 32**,** 121-129.

BLUM, A. 2011. Drought resistance and its improvement. *Plant breeding for water-limited environments.* Springer.

BOE, A. & BECK, D. L. 2008. Yield components of biomass in switchgrass. *Crop Science,* 48**,** 1306-1311.

BOERJAN, W., RALPH, J. & BAUCHER, M. 2003. Lignin biosynthesis. *Annual review of plant biology,* 54**,** 519-546.

BOLÉO, S., FERNANDO, A., BARBOSA, B., COSTA, J., DUARTE, M. & MENDES, B. 2015. Remediation of soils contaminated with zinc by Miscanthus. *WASTES***,** 37-42.

BONAWITZ, N. D. & CHAPPLE, C. 2010. The genetics of lignin biosynthesis: connecting genotype to phenotype. *Annual review of genetics,* 44**,** 337-363.

BOULET, J.-C., ABI-HABIB, E., CARRILLO, S., ROI, S., VERAN, F., VERBAERE, A., MEUDEC, E., RATTIER, A., DUCASSE, M.-A. & JØRGENSEN, B. 2022. Focus on the relationships between the cell wall composition in the extraction of anthocyanins and tannins from grape berries. *Food Chemistry***,** 135023.

BRACHI, B., MORRIS, G. P. & BOREVITZ, J. O. 2011. Genome-wide association studies in plants: the missing heritability is in the field. *Genome biology,* 12**,** 232.

BRADSHAW, H. & STETTLER, R. F. 1995. Molecular genetics of growth and development in populus. IV. Mapping QTLs with large effects on growth, form, and phenology traits in a forest tree. *Genetics,* 139**,** 963-973.

BRANDON, A. G. & SCHELLER, H. V. 2020. Engineering of bioenergy crops: dominant genetic approaches to improve polysaccharide properties and composition in biomass. *Frontiers in plant science,* 11**,** 282.

BRAUN, D. M., WANG, L. & RUAN, Y.-L. 2013. Understanding and manipulating sucrose phloem loading, unloading, metabolism, and signalling to enhance crop yield and food security. *Journal of Experimental Botany,* 65**,** 1713-1735.

BROWN, D., WIGHTMAN, R., ZHANG, Z., GOMEZ, L. D., ATANASSOV, I., BUKOWSKI, J. P., TRYFONA, T., MCQUEEN-MASON, S. J., DUPREE, P. & TURNER, S. 2011. Arabidopsis genes IRREGULAR XYLEM (IRX15) and IRX15L encode DUF579-containing proteins that are essential for normal xylan deposition in the secondary cell wall. *The Plant Journal,* 66**,** 401-413.

BROWN, D. M., ZEEF, L. A., ELLIS, J., GOODACRE, R. & TURNER, S. R. 2005. Identification of novel genes in Arabidopsis involved in secondary cell wall formation using expression profiling and reverse genetics. *The Plant Cell,* 17**,** 2281-2295.

BROXTERMAN, S. E. & SCHOLS, H. A. 2018. Interactions between pectin and cellulose in primary plant cell walls. *Carbohydrate polymers,* 192**,** 263-272.

BUCHFINK, B., XIE, C. & HUSON, D. H. 2015. Fast and sensitive protein alignment using DIAMOND. *Nature methods,* 12**,** 59-60.

BUCKLER, E. S., HOLLAND, J. B., BRADBURY, P. J., ACHARYA, C. B., BROWN, P. J., BROWNE, C., ERSOZ, E., FLINT-GARCIA, S., GARCIA, A. & GLAUBITZ, J. C. 2009. The genetic architecture of maize flowering time. *Science,* 325**,** 714-718.

BULONE, V., SCHWERDT, J. G. & FINCHER, G. B. 2019. Co-evolution of enzymes involved in plant cell wall metabolism in the grasses. *Frontiers in Plant Science,* 10**,** 1009.

BURTON, R. A. & FINCHER, G. B. 2012. Current challenges in cell wall biology in the cereals and grasses. *Frontiers in plant science,* 3**,** 130.

BURTON, R. A. & FINCHER, G. B. 2014a. Evolution and development of cell walls in cereal grains. *Frontiers in plant science,* 5**,** 456.

BURTON, R. A. & FINCHER, G. B. 2014b. Plant cell wall engineering: applications in biofuel production and improved human health. *Current Opinion in Biotechnology,* 26**,** 79-84.

BURTON, R. A., GIDLEY, M. J. & FINCHER, G. B. 2010. Heterogeneity in the chemistry, structure and function of plant cell walls. *Nature chemical biology,* 6**,** 724-732.

BURTON, R. A., SHIRLEY, N. J., KING, B. J., HARVEY, A. J. & FINCHER, G. B. 2004. The CesA gene family of barley. Quantitative analysis of transcripts reveals two groups of co-expressed genes. *Plant physiology,* 134**,** 224-236.

BURTON, R. A., WILSON, S. M., HRMOVA, M., HARVEY, A. J., SHIRLEY, N. J., MEDHURST, A., STONE, B. A., NEWBIGIN, E. J., BACIC, A. & FINCHER, G. B. 2006. Cellulose synthase-like CslF genes mediate the synthesis of cell wall (1, 3; 1, 4)-ß-D-glucans. *Science,* 311**,** 1940-1942.

CAI, X., ZHANG, X. & WANG, D. 2010. Land availability for biofuel production. *Environmental science & technology,* 45**,** 334-339.

CAMPBELL, J. E., LOBELL, D. B., GENOVA, R. C. & FIELD, C. B. 2008. The global potential of bioenergy on abandoned agriculture lands. *Environmental science & technology,* 42**,** 5791-5794.

CAO, J. 2012. The pectin lyases in Arabidopsis thaliana: evolution, selection and expression profiles.

CAPELLA-GUTIÉRREZ, S., SILLA-MARTÍNEZ, J. M. & GABALDÓN, T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics,* 25**,** 1972-1973.

CARLQUIST, S. & SCHNEIDER, E. L. 2001. Vessels in ferns: structural, ecological, and evolutionary significance. *American Journal of Botany,* 88**,** 1-13.

CARLSSON, G., MÅRTENSSON, L. M., PRADE, T., SVENSSON, S. E. & JENSEN, E. S. 2017. Perennial species mixtures for multifunctional production of biomass on marginal land. *Gcb Bioenergy,* 9**,** 191-201.

CAROCHA, V., SOLER, M., HEFER, C., CASSAN-WANG, H., FEVEREIRO, P., MYBURG, A. A., PAIVA, J. A. & GRIMA-PETTENATI, J. 2015. Genome-wide analysis of the lignin toolbox of Eucalyptus grandis. *New Phytologist,* 206**,** 1297-1313.

CARPITA, N., TIERNEY, M. & CAMPBELL, M. 2001. Molecular biology of the plant cell wall: searching for the genes that define structure, architecture and dynamics. *Plant Cell Walls.* Springer.

CARPITA, N. C. & GIBEAUT, D. M. 1993. Structural models of primary cell walls in flowering plants: consistency of molecular structure with the physical properties of the walls during growth. *The Plant Journal,* 3**,** 1-30.

CARPITA, N. C. & MCCANN, M. C. 2008. Maize and sorghum: genetic resources for bioenergy grasses. *Trends in Plant Science,* 13**,** 415-420.

CARPITA, N. C. & MCCANN, M. C. 2020. Redesigning plant cell walls for the biomass-based bioeconomy. *Journal of Biological Chemistry,* 295**,** 15144-15157.

CARROLL, A., MANSOORI, N., LI, S., LEI, L., VERNHETTES, S., VISSER, R. G., SOMERVILLE, C., GU, Y. & TRINDADE, L. M. 2012. Complexes with mixed primary and secondary cellulose synthases are functional in Arabidopsis plants. *Plant physiology,* 160**,** 726-737.

CARROLL, A. & SOMERVILLE, C. 2009. Cellulosic biofuels. *Annual review of plant biology,* 60**,** 165-182.

CASACUBERTA, E. & GONZÁLEZ, J. 2013. The impact of transposable elements in environmental adaptation. *Molecular ecology,* 22**,** 1503-1517.

CECCARELLI, S. 2009. Evolution, plant breeding and biodiversity. *Journal of Agriculture and Environment for International Development (JAEID),* 103**,** 131-145.

CENCI, A., CHANTRET, N. & ROUARD, M. 2018. Glycosyltransferase family 61 in liliopsida (Monocot): the story of a gene family expansion. *Frontiers in Plant Science,* 9**,** 1843.

CHANDRASEKAR, B. & VAN DER HOORN, R. A. 2016. Beta galactosidases in Arabidopsis and tomato–a mini review. *Biochemical Society Transactions,* 44**,** 150-158.

CHAUVAT, M., PEREZ, G., HEDDE, M. & LAMY, I. 2014. Establishment of bioenergy crops on metal contaminated soils stimulates belowground fauna. *biomass and bioenergy,* 62**,** 207-211.

CHEN, C.-L. 2018. *Genetic diversity and mechanisms of salt tolerance of Miscanthus.* Doctoral degree PhD Thesis, Wageningen University.

CHEN, F. & DIXON, R. A. 2007. Lignin modification improves fermentable sugar yields for biofuel production. *Nature biotechnology,* 25**,** 759.

CHENNA, R., SUGAWARA, H., KOIKE, T., LOPEZ, R., GIBSON, T. J., HIGGINS, D. G. & THOMPSON, J. D. 2003. Multiple sequence alignment with the Clustal series of programs. *Nucleic acids research,* 31**,** 3497-3500.

CHEW, Y. H. & HALLIDAY, K. J. 2011. A stress-free walk from Arabidopsis to crops. *Current Opinion in Biotechnology,* 22**,** 281-286.

CHIMENTO, C., ALMAGRO, M. & AMADUCCI, S. 2016. Carbon sequestration potential in perennial bioenergy crops: the importance of organic matter inputs and its physical protection. *Gcb Bioenergy,* 8**,** 111-121.

CLARK, J. W. & DONOGHUE, P. C. 2018. Whole-genome duplication and plant macroevolution. *Trends in plant science,* 23**,** 933-945.

CLIFTON-BROWN, J. & JONES, M. 1997. The thermal response of leaf extension rate in genotypes of the C4–grass Miscanthus: an important factor in determining the potential productivity of different genotypes. *Journal of Experimental Botany,* 48**,** 1573-1581.

CLIFTON-BROWN, J. & LEWANDOWSKI, I. 2000. Overwintering problems of newly established Miscanthus plantations can be overcome by identifying genotypes with improved rhizome cold tolerance. *The New Phytologist,* 148**,** 287-294.

CLIFTON-BROWN, J., HARFOUCHE, A., CASLER, M. D., DYLAN JONES, H., MACALPINE, W. J., MURPHY-BOKERN, D., SMART, L. B., ADLER, A., ASHMAN, C. & AWTY-CARROLL, D. 2018. Breeding progress and preparedness for mass-scale deployment of perennial lignocellulosic biomass crops switchgrass, miscanthus, willow and poplar. *Gcb Bioenergy*.

COCURON, J.-C., LEROUXEL, O., DRAKAKAKI, G., ALONSO, A. P., LIEPMAN, A. H., KEEGSTRA, K., RAIKHEL, N. & WILKERSON, C. G. 2007. A gene from the cellulose synthase-like C family encodes a β-1, 4 glucan synthase. *Proceedings of the National Academy of Sciences,* 104**,** 8550-8555.

COLLINS, K. J. 2008. *The role of biofuels and other factors in increasing farm and food prices: a review of recent developments with a focus on feed grain markets and market prospects*, K. Collins.

COSENTINO, S. L., COPANI, V., D'AGOSTA, G. M., SANZONE, E. & MANTINEO, M. 2006. First results on evaluation of Arundo donax L. clones collected in Southern Italy. *Industrial Crops and Products,* 23**,** 212-222.

COSENTINO, S. L., COPANI, V., SCALICI, G., SCORDIA, D. & TESTA, G. 2015. Soil erosion mitigation by perennial species under Mediterranean environment. *BioEnergy Research,* 8**,** 1538-1547.

COSGROVE, D. C. 2012. Comparative structure and biomechanics of plant primary and secondary cell walls. *Frontiers in plant science,* 3**,** 204.

COSGROVE, D. J. 2005. Growth of the plant cell wall. *Nature reviews molecular cell biology,* 6**,** 850.

COSTA, J., BARBOSA, B. & FERNANDO, A. L. Wastewaters reuse for energy crops cultivation. Doctoral Conference on Computing, Electrical and Industrial Systems, 2016. Springer, 507-514.

COURTIAL, A., MÉCHIN, V., REYMOND, M., GRIMA-PETTENATI, J. & BARRIÈRE, Y. 2014. Colocalizations between several QTLs for cell wall degradability and composition in the F288× F271 early maize RIL progeny raise the question of the nature of the possible underlying determinants and breeding targets for biofuel capacity. *BioEnergy Research,* 7**,** 142-156.

COVARRUBIAS-PAZARAN, G. 2016. Genome-assisted prediction of quantitative traits using the R package sommer. *PloS one,* 11**,** e0156744.

CROW, T., TA, J., NOJOOMI, S., AGUILAR-RANGEL, M. R., RODRÍGUEZ, J. V. T., GATES, D., RELLAN-ALVAREZ, R., SAWERS, R. & RUNCIE, D. 2020. Gene regulatory effects of a large chromosomal inversion in highland maize. *PLoS genetics,* 16**,** e1009213.

DA COSTA, R. M., PATTATHIL, S., AVCI, U., WINTERS, A., HAHN, M. G. & BOSCH, M. 2019. Desirable plant cell wall traits for higher-quality miscanthus lignocellulosic biomass. *Biotechnology for biofuels,* 12**,** 85.

DAI, X., HU, Q., CAI, Q., FENG, K., YE, N., TUSKAN, G. A., MILNE, R., CHEN, Y., WAN, Z. & WANG, Z. 2014. The willow genome and divergent evolution from poplar after the common genome duplication. *Cell research,* 24**,** 1274.

DAI, X., YOU, C., WANG, L., CHEN, G., ZHANG, Q. & WU, C. 2009. Molecular characterization, expression pattern, and function analysis of the OsBC1L family in rice. *Plant molecular biology,* 71**,** 469-481.

DAILY, G. C. 1995. Restoring value to the world's degraded lands. *Science,* 269**,** 350-354.

DALE, V. H., KLINE, K. L., WIENS, J. & FARGIONE, J. 2010. *Biofuels: implications for land use and biodiversity*, Ecological Society of America Washington, DC.

DAS, M. K., FUENTES, R. G. & TALIAFERRO, C. M. 2004. Genetic variability and trait relationships in switchgrass. *Crop science,* 44**,** 443-448.

DAUBER, J., BROWN, C., FERNANDO, A. L., FINNAN, J., KRASUSKA, E., PONITKA, J., STYLES, D., THRÄN, D., VAN GROENIGEN, K. J. & WEIH, M. 2012. Bioenergy from "surplus" land: environmental and socio-economic implications. *BioRisk: Biodiversity & Ecosystem Risk Assessment,* 7.

DAUBER, J., JONES, M. B. & STOUT, J. C. 2010. The impact of biomass crop cultivation on temperate biodiversity. *Gcb Bioenergy,* 2**,** 289-309.

DAVEY, C. L., ROBSON, P., HAWKINS, S., FARRAR, K., CLIFTON-BROWN, J. C., DONNISON, I. S. & SLAVOV, G. T. 2017. Genetic relationships between spring emergence, canopy phenology, and biomass yield increase the accuracy of genomic prediction in Miscanthus. *Journal of experimental botany,* 68**,** 5093-5102.

DAWSON, I. K., RUSSELL, J., POWELL, W., STEFFENSON, B., THOMAS, W. T. & WAUGH, R. 2015. Barley: a translational model for adaptation to climate change. *New Phytologist,* 206**,** 913-931.

DE OLIVEIRA, D. M., FINGER-TEIXEIRA, A., RODRIGUES MOTA, T., SALVADOR, V. H., MOREIRA-VILAR, F. C., CORREA MOLINARI, H. B., CRAIG MITCHELL, R. A., MARCHIOSI, R., FERRARESE-FILHO, O. & DANTAS DOS SANTOS, W. 2015. Ferulic acid: a key component in grass lignocellulose recalcitrance to hydrolysis. *Plant biotechnology journal,* 13**,** 1224-1232.

DE OLIVEIRA SILVA, F. M., LICHTENSTEIN, G., ALSEEKH, S., ROSADO-SOUZA, L., CONTE, M., SUGUIYAMA, V. F., LIRA, B. S., FANOURAKIS, D., USADEL, B. & BHERING, L. L. 2018. The genetic architecture of photosynthesis and plant growth-related traits in tomato. *Plant, cell & environment,* 41**,** 327-341.

DE SOUZA, W. R., MARTINS, P. K., FREEMAN, J., PELLNY, T. K., MICHAELSON, L. V., SAMPAIO, B. L., VINECKY, F., RIBEIRO, A. P., DA CUNHA, B. A. & KOBAYASHI, A. K. 2018. Suppression of a single BAHD gene in Setaria viridis causes large, stable decreases in cell wall feruloylation and increases biomass digestibility. *New Phytologist,* 218**,** 81-93.

DE VRIES, S. C., VAN DE VEN, G. W., VAN ITTERSUM, M. K. & GILLER, K. E. 2010. Resource use efficiency and environmental performance of nine major biofuel crops, processed by first-generation conversion techniques. *Biomass and Bioenergy,* 34**,** 588-601.

DELCHER, A. L., PHILLIPPY, A., CARLTON, J. & SALZBERG, S. L. 2002. Fast algorithms for large-scale genome alignment and comparison. *Nucleic acids research,* 30**,** 2478-2483.

DEMUTH, J. P. & HAHN, M. W. 2009. The life and death of gene families. *Bioessays,* 31**,** 29-39.

DEWEY, C. N. 2011. Positional orthology: putting genomic evolutionary relationships into context. *Briefings in bioinformatics,* 12**,** 401-412.

DHUGGA, K. S., BARREIRO, R., WHITTEN, B., STECCA, K., HAZEBROEK, J., RANDHAWA, G. S., DOLAN, M., KINNEY, A. J., TOMES, D. & NICHOLS, S. 2004. Guar seed ß-mannan synthase is a member of the cellulose synthase super gene family. *Science,* 303**,** 363-366.

DIETZ, T., BÖRNER, J., FÖRSTER, J. & VON BRAUN, J. 2018. Governance of the bioeconomy: A global comparative study of national bioeconomy strategies. *Sustainability,* 10**,** 3190.

DOBLIN, M. S., DE MELIS, L., NEWBIGIN, E., BACIC, A. & READ, S. M. 2001. Pollen tubes of Nicotiana alata express two genes from different β-glucan synthase families. *Plant Physiology,* 125**,** 2040-2052.

DOBLIN, M. S., PETTOLINO, F. A., WILSON, S. M., CAMPBELL, R., BURTON, R. A., FINCHER, G. B., NEWBIGIN, E. & BACIC, A. 2009. A barley cellulose synthase-like CSLH gene mediates (1, 3; 1, 4)-β-D-glucan synthesis in transgenic Arabidopsis. *Proceedings of the National Academy of Sciences,* 106**,** 5996-6001.

DOHLEMAN, F. G. & LONG, S. P. 2009. More productive than maize in the Midwest: how does Miscanthus do it? *Plant physiology,* 150**,** 2104-2115.

DONG, Q., SCHLUETER, S. D. & BRENDEL, V. 2004. PlantGDB, plant genome database and analysis tools. *Nucleic acids research,* 32**,** D354-D359.

DRULA, E., GARRON, M.-L., DOGAN, S., LOMBARD, V., HENRISSAT, B. & TERRAPON, N. 2022. The carbohydrate-active enzyme database: functions and literature. *Nucleic acids research,* 50**,** D571-D577.

DU, Q., LU, W., QUAN, M., XIAO, L., SONG, F., LI, P., ZHOU, D., XIE, J., WANG, L. & ZHANG, D. 2018. Genome-wide association studies to improve wood properties: challenges and prospects. *Frontiers in plant science,* 9.

DUBOUZET, J. G., STRABALA, T. J. & WAGNER, A. 2013. Potential transgenic routes to increase tree biomass. *Plant Science,* 212**,** 72-101.

DWIYANTI, M. S., STEWART, J. R. & YAMADA, T. 2014. Forages for feedstocks of biorefineries in temperate environments: review of lignin research in bioenergy crops and some insight into Miscanthus studies. *Crop and Pasture Science,* 65**,** 1199-1206.

EDDY, S. R. 2009. A new generation of homology search tools based on probabilistic inference. *Genome Informatics 2009: Genome Informatics Series Vol. 23.* World Scientific.

EDDY, S. R. 2011. Accelerated profile HMM searches. *PLoS computational biology,* 7**,** e1002195.

EL-GEBALI, S., MISTRY, J., BATEMAN, A., EDDY, S. R., LUCIANI, A., POTTER, S. C., QURESHI, M., RICHARDSON, L. J., SALAZAR, G. A. & SMART, A. 2019. The Pfam protein families database in 2019. *Nucleic acids research,* 47**,** D427-D432.

EMERSON, R., HOOVER, A., RAY, A., LACEY, J., CORTEZ, M., PAYNE, C., KARLEN, D., BIRRELL, S., LAIRD, D. & KALLENBACH, R. 2014. Drought effects on composition and yield for corn stover, mixed grasses, and Miscanthus as bioenergy feedstocks. *Biofuels,* 5**,** 275-291.

ENDO, H., YAMAGUCHI, M., TAMURA, T., NAKANO, Y., NISHIKUBO, N., YONEDA, A., KATO, K., KUBO, M., KAJITA, S. & KATAYAMA, Y. 2015. Multiple classes of transcription factors regulate the expression of VASCULAR-RELATED NAC-DOMAIN7, a master switch of xylem vessel differentiation. *Plant and Cell Physiology,* 56**,** 242-254.

EVANS, J., SANCIANGCO, M. D., LAU, K. H., CRISOVAN, E., BARRY, K., DAUM, C., HUNDLEY, H., JENKINS, J., KENNEDY, M. & KUNDE-RAMAMOORTHY, G. 2017. Extensive genetic diversity is present within North American switchgrass germplasm. *The plant genome*.

FAHRENKROG, A. M., NEVES, L. G., RESENDE, M. F., VAZQUEZ, A. I., CAMPOS, G., DERVINIS, C., SYKES, R., DAVIS, M., DAVENPORT, R. & BARBAZUK, W. B. 2017. Genome-wide association study reveals putative regulators of bioenergy traits in Populus deltoides. *New Phytologist,* 213**,** 799-811.

FARGIONE, J., HILL, J., TILMAN, D., POLASKY, S. & HAWTHORNE, P. 2008. Land clearing and the biofuel carbon debt. *Science,* 319**,** 1235-1238.

FARRELL, A., CLIFTON-BROWN, J., LEWANDOWSKI, I. & JONES, M. 2006a. Genotypic variation in cold tolerance influences the yield of Miscanthus. *Annals of Applied Biology,* 149**,** 337-345.

FARRELL, A. E., PLEVIN, R. J., TURNER, B. T., JONES, A. D., O'HARE, M. & KAMMEN, D. M. 2006b. Ethanol can contribute to energy and environmental goals. *Science,* 311**,** 506-508.

FEIJAO, C., MORREEL, K., ANDERS, N., TRYFONA, T., BUSSE-WICHER, M., KOTAKE, T., BOERJAN, W. & DUPREE, P. 2022. Hydroxycinnamic acid-modified xylan side chains and their cross-linking products in rice cell walls are reduced in the Xylosyl arabinosyl substitution of xylan 1 mutant. *The Plant Journal,* 109**,** 1152-1167.

FERNANDEZ, M. G. S., BAO, Y., TANG, L. & SCHNABLE, P. S. 2017. A high-throughput, field-based phenotyping technology for tall biomass crops. *Plant physiology,* 174**,** 2008-2022.

FERNANDEZ, M. G. S., BECRAFT, P. W., YIN, Y. & LÜBBERSTEDT, T. 2009. From dwarves to giants? Plant height manipulation for biomass yield. *Trends in plant science,* 14**,** 454-461.

FERNANDO, A. L., COSTA, J., BARBOSA, B., MONTI, A. & RETTENMAIER, N. 2018. Environmental impact assessment of perennial crops cultivation on marginal soils in the Mediterranean Region. *Biomass & Bioenergy,* 111**,** 174-186.

FERREIRA, S., HJERNØ, K., LARSEN, M., WINGSLE, G., LARSEN, P., FEY, S., ROEPSTORFF, P. & SALOMÉ PAIS, M. 2006. Proteome profiling of Populus euphratica Oliv. upon heat stress. *Annals of botany,* 98**,** 361-377.

FIELD, C. B., CAMPBELL, J. E. & LOBELL, D. B. 2008. Biomass energy: the scale of the potential resource. *Trends in ecology & evolution,* 23**,** 65-72.

FINCHER, G. & STONE, B. 1986. Cell walls and their components in cereal grain technology. *Advances in cereal science and technology,* 8**,** 207-295.

FINN, R. D., BATEMAN, A., CLEMENTS, J., COGGILL, P., EBERHARDT, R. Y., EDDY, S. R., HEGER, A., HETHERINGTON, K., HOLM, L. & MISTRY, J. 2014. Pfam: the protein families database. *Nucleic acids research,* 42**,** D222-D230.

FLAGEL, L. E. & WENDEL, J. F. 2009. Gene duplication and evolutionary novelty in plants. *New Phytologist,* 183**,** 557-564.

FLOOD, P. J., HARBINSON, J. & AARTS, M. G. 2011. Natural genetic variation in plant photosynthesis. *Trends in plant science,* 16**,** 327-335.

FRANKOVÁ, L. & FRY, S. C. 2021. Hemicellulose-remodelling transglycanase activities from charophytes: towards the evolution of the land-plant cell wall. *The Plant Journal,* 108**,** 7-28.

FREITAS, C. M. P., COIMBRA, J. S. R., SOUZA, V. G. L. & SOUSA, R. C. S. 2021. Structure and applications of pectin in food, biomedical, and pharmaceutical industry: A review. *Coatings,* 11**,** 922.

FRITSCHE, U. R., SIMS, R. E. & MONTI, A. 2010. Direct and indirect land-use competition issues for energy crops and their sustainable production–an overview. *Biofuels, Bioproducts and Biorefining,* 4**,** 692-704.

FRY, S. C. 2004. Primary cell wall metabolism: tracking the careers of wall polymers in living plant cells. *New phytologist,* 161**,** 641-675.

FU, C., MIELENZ, J. R., XIAO, X., GE, Y., HAMILTON, C. Y., RODRIGUEZ, M., CHEN, F., FOSTON, M., RAGAUSKAS, A. & BOUTON, J. 2011. Genetic manipulation of lignin reduces recalcitrance and improves ethanol production from switchgrass. *Proceedings of the National Academy of Sciences,* 108**,** 3803-3808.

GABUR, I., CHAWLA, H. S., SNOWDON, R. J. & PARKIN, I. A. 2019. Connecting genome structural variation with complex traits in crop plants. *Theoretical and applied genetics,* 132**,** 733-750.

GALE, M. D. & DEVOS, K. M. 1998. Comparative genetics in the grasses. *Proceedings of the National Academy of Sciences,* 95**,** 1971-1974.

GAO, F., CHEN, C., ARAB, D. A., DU, Z., HE, Y. & HO, S. Y. W. 2019. EasyCodeML: A visual tool for analysis of selection using CodeML. *Ecology and Evolution***,** 3891-3898.

GARDINER, M. A., TUELL, J. K., ISAACS, R., GIBBS, J., ASCHER, J. S. & LANDIS, D. A. 2010. Implications of three biofuel crops for beneficial arthropods in agricultural landscapes. *BioEnergy Research,* 3**,** 6-19.

GARZON, C. D., LEQUART, M., CHARRAS, Q., FOURNET, F., BELLENGER, L., SELLIER-RICHARD, H., GIAUFFRET, C., VERMERRIS, W., DOMON, J.-M. & RAYON, C. 2022. The maize low-lignin brown midrib3 mutant shows pleiotropic effects on photosynthetic and cell wall metabolisms in response to chilling. *Plant Physiology and Biochemistry,* 184**,** 75-86.

GAWKOWSKA, D., CYBULSKA, J. & ZDUNEK, A. 2018. Structure-related gelling of pectins and linking with other natural compounds: A review. *Polymers,* 10**,** 762.

GE, C., AI, X., JIA, S., YANG, Y., CHE, L., YI, Z. & CHEN, C. 2018. Interspecific genetic maps in Miscanthus floridulus and M. sacchariflorus accelerate detection of QTLs associated with plant height and inflorescence. *Molecular Genetics and Genomics***,** 1-11.

GELFAND, I., SAHAJPAL, R., ZHANG, X., IZAURRALDE, R. C., GROSS, K. L. & ROBERTSON, G. P. 2013. Sustainable bioenergy production from marginal lands in the US Midwest. *Nature,* 493**,** 514.

GELFAND, I., ZENONE, T., JASROTIA, P., CHEN, J., HAMILTON, S. K. & ROBERTSON, G. P. 2011. Carbon debt of Conservation Reserve Program (CRP) grasslands converted to bioenergy production. *Proceedings of the National Academy of Sciences,* 108**,** 13864-13869.

GENG, P., ZHANG, S., LIU, J., ZHAO, C., WU, J., CAO, Y., FU, C., HAN, X., HE, H. & ZHAO, Q. 2020. MYB20, MYB42, MYB43, and MYB85 regulate phenylalanine and lignin biosynthesis during secondary cell wall formation. *Plant Physiology,* 182**,** 1272-1283.

GHANNOUM, O., EVANS, J. & VON CAEMMERER, S. 2011. *Nitrogen and Water Use Efficiency of C4 Plants*.

GIBBS, H. K., JOHNSTON, M., FOLEY, J. A., HOLLOWAY, T., MONFREDA, C., RAMANKUTTY, N. & ZAKS, D. 2008. Carbon payback times for crop-based biofuel expansion in the tropics: the effects of changing yield and technology. *Environmental research letters,* 3**,** 034001.

GIDDINGS, B., HOPWOOD, B. & O'BRIEN, G. 2002. Environment, economy and society: fitting them together into sustainable development. *Sustainable development,* 10**,** 187-196.

GIFFORD, J. M., CHAE, W. B., SWAMINATHAN, K., MOOSE, S. P. & JUVIK, J. A. 2015. Mapping the genome of Miscanthus sinensis for QTL associated with biomass productivity. *Gcb Bioenergy,* 7**,** 797-810.

GÍSLASON, M. H., NIELSEN, H., ARMENTEROS, J. J. A. & JOHANSEN, A. R. 2021. Prediction of GPI-anchored proteins with pointer neural networks. *Current Research in Biotechnology,* 3**,** 6-13.

GLASS, M., BARKWILL, S., UNDA, F. & MANSFIELD, S. D. 2015. Endo-β-1, 4-glucanases impact plant cell wall development by influencing cellulose crystallization. *Journal of integrative plant biology,* 57**,** 396-410.

GLOVER, J. D., CULMAN, S. W., DUPONT, S. T., BROUSSARD, W., YOUNG, L., MANGAN, M. E., MAI, J. G., CREWS, T. E., DEHAAN, L. R. & BUCKLEY, D. H. 2010. Harvested perennial grasslands provide ecological benchmarks for agricultural sustainability. *Agriculture, Ecosystems & Environment,* 137**,** 3-12.

GOFFINET, B. & GERBER, S. 2000. Quantitative trait loci: a meta-analysis. *Genetics,* 155**,** 463-473.

GORDON, S. P., CONTRERAS-MOREIRA, B., WOODS, D. P., DES MARAIS, D. L., BURGESS, D., SHU, S., STRITT, C., ROULIN, A. C., SCHACKWITZ, W. & TYLER, L. 2017. Extensive gene content variation in the Brachypodium distachyon pan-genome correlates with population structure. *Nature communications,* 8**,** 1-13.

GOUJON, T., MINIC, Z., EL AMRANI, A., LEROUXEL, O., ALETTI, E., LAPIERRE, C., JOSELEAU, J. P. & JOUANIN, L. 2003. AtBXL1, a novel higher plant (Arabidopsis thaliana) putative beta-xylosidase gene, is involved in secondary cell wall metabolism and plant development. *The Plant Journal,* 33**,** 677-690.

GOULD, B. A., CHEN, Y. & LOWRY, D. B. 2018. Gene regulatory divergence between locally adapted ecotypes in their native habitats. *Molecular ecology,* 27**,** 4174-4188.

GRATTAPAGLIA, D. & KIRST, M. 2008. Eucalyptus applied genomics: from gene sequences to breeding tools. *New phytologist,* 179**,** 911-929.

GREGERSEN, P. L., CULETIC, A., BOSCHIAN, L. & KRUPINSKA, K. 2013. Plant senescence and crop productivity. *Plant molecular biology,* 82**,** 603-622.

GULISANO, A., ALVES, S., MARTINS, J. N. & TRINDADE, L. M. 2019. Genetics and breeding of Lupinus mutabilis: An emerging protein crop. *Frontiers in Plant Science,* 10**,** 1385.

GULISANO, A., DECHESNE, A., PAULO, M. J. & TRINDADE, L. M. 2022. Investigating the potential of Andean lupin as a lignocellulosic feedstock for Europe: First genome-wide association study on Lupinus mutabilis biomass quality. *GCB Bioenergy*.

HAAS, B. J., DELCHER, A. L., WORTMAN, J. R. & SALZBERG, S. L. 2004. DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics,* 20**,** 3643-3646.

HABERL, H., BERINGER, T., BHATTACHARYA, S. C., ERB, K.-H. & HOOGWIJK, M. 2010. The global technical potential of bio-energy in 2050 considering sustainability constraints. *Current Opinion in Environmental Sustainability,* 2**,** 394-403.

HABERL, H., ERB, K.-H., KRAUSMANN, F., BONDEAU, A., LAUK, C., MÜLLER, C., PLUTZAR, C. & STEINBERGER, J. K. 2011. Global bioenergy potentials from agricultural land in 2050: Sensitivity to climate change, diets and yields. *Biomass and bioenergy,* 35**,** 4753-4769.

HAIGLER, C. H., BETANCUR, L., STIFF, M. R. & TUTTLE, J. R. 2012. Cotton fiber: a powerful single-cell model for cell wall and cellulose research. *Frontiers in plant science,* 3**,** 104.

HAIGLER, C. H. & ROBERTS, A. W. 2019. Structure/function relationships in the rosette cellulose synthesis complex illuminated by an evolutionary perspective. *Cellulose,* 26**,** 227-247.

HALE, A. L., VIATOR, R. P. & VEREMIS, J. C. 2014. Identification of freeze tolerant Saccharum spontaneum accessions through a pot-based study for use in sugarcane germplasm enhancement for adaptation to temperate climates. *Biomass and bioenergy,* 61**,** 53-57.

HALL, M., BANSAL, P., LEE, J. H., REALFF, M. J. & BOMMARIUS, A. S. 2010. Cellulose crystallinity–a key predictor of the enzymatic hydrolysis rate. *The FEBS journal,* 277**,** 1571-1582.

HALLGREN, J., TSIRIGOS, K. D., PEDERSEN, M. D., ARMENTEROS, J. J. A., MARCATILI, P., NIELSEN, H., KROGH, A. & WINTHER, O. 2022. DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks. *bioRxiv*.

HAMILTON, S., HUSSAIN, M., BHARDWAJ, A., BASSO, B. & ROBERTSON, G. 2015. Comparative water use by maize, perennial crops, restored prairie, and poplar trees in the US Midwest. *Environmental Research Letters,* 10**,** 064015.

HAN, Y., FAN, T., ZHU, X., WU, X., OUYANG, J., JIANG, L. & CAO, S. 2019. WRKY12 represses GSH1 expression to negatively regulate cadmium tolerance in Arabidopsis. *Plant Molecular Biology,* 99**,** 149-159.

HANCOCK, J. F. 2012. *Plant evolution and the origin of crop species*, CABI.

HARFOUCHE, A. L., JACOBSON, D. A., KAINER, D., ROMERO, J. C., HARFOUCHE, A. H., MUGNOZZA, G. S., MOSHELION, M., TUSKAN, G. A., KEURENTJES, J. J. & ALTMAN, A. 2019.

Accelerating climate resilient plant breeding by applying next-generation artificial intelligence. *Trends in biotechnology,* 37**,** 1217-1235.

HARRIS, D. & DEBOLT, S. 2008. Relative crystallinity of plant biomass: studies on assembly, adaptation and acclimation. *PLoS One,* 3**,** e2897.

HARRIS, D., STORK, J. & DEBOLT, S. 2009. Genetic modification in cellulose-synthase reduces crystallinity and improves biochemical conversion to fermentable sugar. *Gcb Bioenergy,* 1**,** 51-61.

HAUGHTON, A. J., BOHAN, D. A., CLARK, S. J., MALLOTT, M. D., MALLOTT, V., SAGE, R. & KARP, A. 2016. Dedicated biomass crops can enhance biodiversity in the arable landscape. *GCB Bioenergy,* 8**,** 1071-1081.

HÉBANT, C. 1977. conducting tissues of bryophytes.

HEFFNER, E. L., SORRELLS, M. E. & JANNINK, J.-L. 2009. Genomic selection for crop improvement. *Crop Science,* 49**,** 1-12.

HENRISSAT, B. 1991. A classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochemical journal,* 280**,** 309-316.

HIKOSAKA, K. 2010. Mechanisms underlying interspecific variation in photosynthetic capacity across wild plant species. *Plant Biotechnology,* 27**,** 223-229.

HILLEY, J., TRUONG, S., OLSON, S., MORISHIGE, D. & MULLET, J. 2016. Identification of Dw1, a regulator of sorghum stem internode length. *PLoS One,* 11**,** e0151271.

HINCHEE, M., ZHANG, C., CHANG, S., CUNNINGHAM, M., HAMMOND, W. & NEHRA, N. 2011. Biotech Eucalyptus can sustainably address society's need for wood: the example of freeze tolerant Eucalyptusin the southeastern U.S. *BMC Proceedings,* 5**,** I24.

HIRANO, K., KONDO, M., AYA, K., MIYAO, A., SATO, Y., ANTONIO, B. A., NAMIKI, N., NAGAMURA, Y. & MATSUOKA, M. 2013. Identification of transcription factors involved in rice secondary cell wall formation. *Plant and Cell Physiology,* 54**,** 1791-1802.

HISANO, H., NANDAKUMAR, R. & WANG, Z.-Y. 2011. Genetic modification of lignin biosynthesis for improved biofuel production. *Biofuels.* Springer.

HØIE, M. H., KIEHL, E. N., PETERSEN, B., NIELSEN, M., WINTHER, O., NIELSEN, H., HALLGREN, J. & MARCATILI, P. 2022. NetSurfP-3.0: accurate and fast prediction of protein structural features by protein language models and deep learning. *Nucleic Acids Research*.

HOLLENDER, C. A. & DARDICK, C. 2015. Molecular basis of angiosperm tree architecture. *New Phytologist,* 206**,** 541-556.

HOOGWIJK, M., FAAIJ, A., VAN DEN BROEK, R., BERNDES, G., GIELEN, D. & TURKENBURG, W. 2003. Exploration of the ranges of the global potential of biomass for energy. *Biomass and bioenergy,* 25**,** 119-133.

HOUSTON, K., TUCKER, M. R., CHOWDHURY, J., SHIRLEY, N. & LITTLE, A. 2016. The plant cell wall: a complex and dynamic structure as revealed by the responses of genes under stress conditions. *Frontiers in plant science,* 7**,** 984.

HU, H., ZHANG, R., DONG, S., LI, Y., FAN, C., WANG, Y., XIA, T., CHEN, P., WANG, L. & FENG, S. 2018. AtCSLD3 and GhCSLD3 mediate root growth and cell elongation downstream of the ethylene response pathway in Arabidopsis. *Journal of experimental botany,* 69**,** 1065-1080.

HUALA, E., DICKERMAN, A. W., GARCIA-HERNANDEZ, M., WEEMS, D., REISER, L., LAFOND, F., HANLEY, D., KIPHART, D., ZHUANG, M. & HUANG, W. 2001. The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic acids research,* 29**,** 102-105.

HUANG, L., RAATS, D., SELA, H., KLYMIUK, V., LIDZBARSKY, G., FENG, L., KRUGMAN, T. & FAHIMA, T. 2016. Evolution and adaptation of wild emmer wheat populations to biotic and abiotic stresses. *Annual review of phytopathology,* 54**,** 279-301.

HUNTER, C. T., KIRIENKO, D. H., SYLVESTER, A. W., PETER, G. F., MCCARTY, D. R. & KOCH, K. E. 2012. Cellulose Synthase-Like D1 is integral to normal cell division, expansion, and leaf development in maize. *Plant physiology,* 158**,** 708-724.

HURGOBIN, B., GOLICZ, A. A., BAYER, P. E., CHAN, C. K. K., TIRNAZ, S., DOLATABADIAN, A., SCHIESSL, S. V., SAMANS, B., MONTENEGRO, J. D. & PARKIN, I. A. 2018. Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid Brassica napus. *Plant biotechnology journal,* 16**,** 1265-1274.

ISIKGOR, F. H. & BECER, C. R. 2015. Lignocellulosic biomass: a sustainable platform for the production of bio-based chemicals and polymers. *Polymer Chemistry,* 6**,** 4497-4559.

IWASE, A., HIDENO, A., WATANABE, K., MITSUDA, N. & OHME-TAKAGI, M. 2009. A chimeric NST repressor has the potential to improve glucose productivity from plant cell walls. *Journal of biotechnology,* 142**,** 279-284.

JAHN, C. E., MCKAY, J. K., MAULEON, R., STEPHENS, J., MCNALLY, K. L., BUSH, D. R., LEUNG, H. & LEACH, J. E. 2011. Genetic variation in biomass traits among 20 diverse rice varieties. *Plant Physiology,* 155**,** 157-168.

JENSEN, E., FARRAR, K., THOMAS-JONES, S., HASTINGS, A., DONNISON, I. & CLIFTON-BROWN, J. 2011. Characterization of flowering time diversity in Miscanthus species. *Gcb Bioenergy,* 3**,** 387-400.

JENSEN, J. D. & BACHTROG, D. 2010. Characterizing recurrent positive selection at fast-evolving genes in Drosophila miranda and Drosophila pseudoobscura. *Genome biology and evolution,* 2**,** 371-378.

JIAO, W.-B. & SCHNEEBERGER, K. 2020. Chromosome-level assemblies of multiple Arabidopsis genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *Nature Communications,* 11**,** 1-10.

JIAO, Y., WICKETT, N. J., AYYAMPALAYAM, S., CHANDERBALI, A. S., LANDHERR, L., RALPH, P. E., TOMSHO, L. P., HU, Y., LIANG, H. & SOLTIS, P. S. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature,* 473**,** 97-100.

JONES, M. B., FINNAN, J. & HODKINSON, T. R. 2015. Morphological and physiological traits for higher biomass production in perennial rhizomatous grasses grown on marginal land. *Gcb Bioenergy,* 7**,** 375-385.

JONES, R., LE-BAS, C., NACHTERGAELE, F., ROSSITER, D., VAN, J., ORSHOVEN, R. S. & VAN VELTHUIZEN, H. 2014. Updated common bio-physical criteria to define natural constraints for agriculture in Europe. Definition and scientific justification for the common criteria. *JRC Science and Policy Reports*.

JOSHI, C. P., THAMMANNAGOWDA, S., FUJINO, T., GOU, J.-Q., AVCI, U., HAIGLER, C. H., MCDONNELL, L. M., MANSFIELD, S. D., MENGESHA, B. & CARPITA, N. C. 2011. Perturbation of wood cellulose synthesis causes pleiotropic effects in transgenic aspen. *Molecular plant,* 4**,** 331-345.

JOZWIAK, A., SONAWANE, P. D., PANDA, S., GARAGOUNIS, C., PAPADOPOULOU, K. K., ABEBIE, B., MASSALHA, H., ALMEKIAS-SIEGL, E., SCHERF, T. & AHARONI, A. 2020. Plant terpenoid metabolism co-opts a component of the cell wall biosynthesis machinery. *Nature Chemical Biology,* 16**,** 740-748.

KALLURI, U. C. & JOSHI, C. P. 2004. Differential expression patterns of two cellulose synthase genes are associated with primary and secondary cell wall development in aspen trees. *Planta,* 220**,** 47-55.

KAMEI, C. L. A., SEVERING, E. I., DECHESNE, A., FURRER, H., DOLSTRA, O. & TRINDADE, L. M. 2016. Orphan Crops Browser: a bridge between model and orphan crops. *Molecular Breeding,* 36.

KANG, Y. J., LEE, T., LEE, J., SHIM, S., JEONG, H., SATYAWAN, D., KIM, M. Y. & LEE, S. H. 2016. Translational genomics for plant breeding with the genome sequence explosion. *Plant biotechnology journal,* 14**,** 1057-1069.

KARP, A. & SHIELD, I. 2008. Bioenergy from plants and the sustainable yield challenge. *New Phytologist,* 179**,** 15-32.

KATOH, K. & STANDLEY, D. M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution,* 30**,** 772-780.

KAUR, S., DHUGGA, K. S., GILL, K. & SINGH, J. 2016. Novel structural and functional motifs in cellulose synthase (CesA) genes of bread wheat (Triticum aestivum, L.). *PLoS One,* 11**,** e0147046.

KENDURKAR, S. V. & RANGASWAMY, M. 2018. Genetic Transformation in Eucalyptus. *Biotechnologies of Crop Improvement, Volume 2.* Springer.

KENNEY, W., SENNERBY-FORSSE, L. & LAYTON, P. 1990. A review of biomass quality research relevant to the use of poplar and willow for energy conversion. *Biomass,* 21**,** 163-188.

KENRICK, P. & CRANE, P. R. 1997. The origin and early evolution of plants on land. *Nature,* 389**,** 33-39.

KERSTENS, M. H., SCHRANZ, M. E. & BOUWMEESTER, K. 2020. Phylogenomic analysis of the APETALA2 transcription factor subfamily across angiosperms reveals both deep conservation and lineage-specific patterns. *The Plant Journal,* 103**,** 1516-1524.

KIM, C. M., PARK, S. H., JE, B. I., PARK, S. H., PARK, S. J., PIAO, H. L., EUN, M. Y., DOLAN, L. & HAN, C.-D. 2007. OsCSLD1, a cellulose synthase-like D1 gene, is required for root hair morphogenesis in rice. *Plant Physiology,* 143**,** 1220-1230.

KIM, K. D., KANG, Y. & KIM, C. 2020a. Application of genomic big data in plant breeding: Past, present, and future. *Plants,* 9**,** 1454.

KIM, S.-J., CHANDRASEKAR, B., REA, A. C., DANHOF, L., ZEMELIS-DURFEE, S., THROWER, N., SHEPARD, Z. S., PAULY, M., BRANDIZZI, F. & KEEGSTRA, K. 2020b. The synthesis of xyloglucan, an abundant plant cell wall polysaccharide, requires CSLC function. *Proceedings of the National Academy of Sciences,* 117**,** 20316-20324.

KLASNJA, B., ORLOVIC, S., GALIC, Z. & IVANISEVIC, P. 2003. Mass volume of poplar wood as a factor of increasing quantity of manufactured fibers.

KONDRASHOV, F. A. 2012. Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proceedings of the Royal Society B: Biological Sciences,* 279**,** 5048-5057.

KONG, W., KIM, C., GOFF, V. H., ZHANG, D. & PATERSON, A. H. 2015. Genetic analysis of rhizomatousness and its relationship with vegetative branching of recombinant inbred lines of Sorghum bicolor× S. propinquum. *American journal of botany,* 102**,** 718-724.

KONO, T. J., BROHAMMER, A. B., MCGAUGH, S. E. & HIRSCH, C. N. 2018. Tandem duplicate genes in maize are abundant and date to two distinct periods of time. *G3: Genes, Genomes, Genetics,* 8**,** 3049-3058.

KOZLOVA, L. V., NAZIPOVA, A. R., GORSHKOV, O. V., PETROVA, A. A. & GORSHKOVA, T. A. 2020. Elongating maize root: zone-specific combinations of polysaccharides from type I and type II primary cell walls. *Scientific reports,* 10**,** 1-20.

KUMAR, M., TOMAR, M., POTKULE, J., VERMA, R., PUNIA, S., MAHAPATRA, A., BELWAL, T., DAHUJA, A., JOSHI, S. & BERWAL, M. K. 2021. Advances in the plant protein extraction: Mechanism and recommendations. *Food Hydrocolloids,* 115**,** 106595.

KURTZ, S., PHILLIPPY, A., DELCHER, A. L., SMOOT, M., SHUMWAY, M., ANTONESCU, C. & SALZBERG, S. L. 2004. Versatile and open software for comparing large genomes. *Genome biology,* 5**,** 1-9.

LAI, X., BENDIX, C., YAN, L., ZHANG, Y., SCHNABLE, J. C. & HARMON, F. G. 2020. Interspecific analysis of diurnal gene regulation in panicoid grasses identifies known and novel regulatory motifs. *BMC genomics,* 21**,** 1-17.

LANDIS, D. A., GRATTON, C., JACKSON, R. D., GROSS, K. L., DUNCAN, D. S., LIANG, C., MEEHAN, T. D., ROBERTSON, B. A., SCHMIDT, T. M. & STAHLHEBER, K. A. 2018. Biomass and biofuel crop effects on biodiversity and ecosystem services in the North Central US. *Biomass and bioenergy,* 114**,** 18-29.

LAURIA, M., MOLINARI, F. & MOTTO, M. 2015. Genetic strategies to enhance plant biomass yield and quality-related traits for bio-renewable fuel and chemical productions. *Plants for the future.* InTech.

LAWRENCE, M., HUBER, W., PAGES, H., ABOYOUN, P., CARLSON, M., GENTLEMAN, R., MORGAN, M. T. & CAREY, V. J. 2013. Software for computing and annotating genomic ranges. *PLoS computational biology,* 9**,** e1003118.

LAWSON, D. J., DAVIES, N. M., HAWORTH, S., ASHRAF, B., HOWE, L., CRAWFORD, A., HEMANI, G., DAVEY SMITH, G. & TIMPSON, N. J. 2020. Is population structure in the genetic biobank era irrelevant, a challenge, or an opportunity? *Human Genetics,* 139**,** 23-41.

LAWSON, T., KRAMER, D. M. & RAINES, C. A. 2012. Improving yield by exploiting mechanisms underlying natural variation of photosynthesis. *Current Opinion in Biotechnology,* 23**,** 215-220.

LE GALL, H., PHILIPPE, F., DOMON, J.-M., GILLET, F., PELLOUX, J. & RAYON, C. 2015. Cell wall metabolism in response to abiotic stress. *Plants,* 4**,** 112-166.

LEDUC, S. D., ZHANG, X., CLARK, C. M. & IZAURRALDE, R. C. 2017. Cellulosic feedstock production on Conservation Reserve Program land: potential yields and environmental effects. *Gcb Bioenergy,* 9**,** 460-468.

LEE, J., CHIN, J. H., AHN, S. N. & KOH, H.-J. 2015. Brief history and perspectives on plant breeding. *Current Technologies in Plant Molecular Breeding.* Springer.

LEE, S., CHOI, S., JEON, D., KANG, Y. & KIM, C. 2020. Evolutionary impact of whole genome duplication in Poaceae family. *Journal of Crop Science and Biotechnology***,** 1-13.

LETUNIC, I. & BORK, P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic acids research,* 47**,** W256-W259.

LEVINE, U. Y., TEAL, T. K., ROBERTSON, G. P. & SCHMIDT, T. M. 2011. Agriculture's impact on microbial diversity and associated fluxes of carbon dioxide and methane. *The ISME journal,* 5**,** 1683.

LEWANDOWSKI, I. 2016. The role of perennial biomass crops in a growing bioeconomy. *Perennial Biomass Crops for a Resource-Constrained World.* Springer.

LEWANDOWSKI, I. & KICHERER, A. 1997. Combustion quality of biomass: practical relevance and experiments to modify the biomass quality of Miscanthus x giganteus. *European Journal of Agronomy,* 6**,** 163-177.

LI, A., WANG, R., LI, X., LIU, M., FAN, J., GUO, K., LUO, B., CHEN, T., FENG, S. & WANG, Y. 2016a. Proteomic profiling of cellulase-aid-extracted membrane proteins for functional identification of cellulose synthase complexes and their potential associated-components in cotton fibers. *Scientific reports,* 6**,** 1-12.

LI, A., XIA, T., XU, W., CHEN, T., LI, X., FAN, J., WANG, R., FENG, S., WANG, Y. & WANG, B. 2013. An integrative analysis of four CESA isoforms specific for fiber cellulose production between Gossypium hirsutum and Gossypium barbadense. *Planta,* 237**,** 1585-1597.

LI, M., PU, Y. & RAGAUSKAS, A. J. 2016b. Current understanding of the correlation of lignin structure with biomass recalcitrance. *Frontiers in chemistry,* 4**,** 45.

LI, M., XIONG, G., LI, R., CUI, J., TANG, D., ZHANG, B., PAULY, M., CHENG, Z. & ZHOU, Y. 2009. Rice cellulose synthase-like D4 is essential for normal cell-wall biosynthesis and plant growth. *The Plant Journal,* 60**,** 1055-1069.

LI, W., YANG, Z., YAO, J., LI, J., SONG, W. & YANG, X. 2018. Cellulose synthase-like D1 controls organ size in maize. *BMC plant biology,* 18**,** 1-15.

LI, X., CHAVES, A. M., DEES, D. C., MANSOORI, N., YUAN, K., SPEICHER, T. L., NORRIS, J. H., WALLACE, I. S., TRINDADE, L. M. & ROBERTS, A. W. 2022. Cellulose synthesis complexes are homo-oligomeric and hetero-oligomeric in Physcomitrium patens. *Plant Physiology,* 188**,** 2115-2130.

LI, X., SPEICHER, T. L., DEES, D. C., MANSOORI, N., MCMANUS, J. B., TIEN, M., TRINDADE, L. M., WALLACE, I. S. & ROBERTS, A. W. 2019. Convergent evolution of hetero-oligomeric cellulose synthesis complexes in mosses and seed plants. *The Plant Journal,* 99**,** 862-876.

LI, Y., QIAN, Q., ZHOU, Y., YAN, M., SUN, L., ZHANG, M., FU, Z., WANG, Y., HAN, B. & PANG, X. 2003. BRITTLE CULM1, which encodes a COBRA-like protein, affects the mechanical properties of rice plants. *The Plant Cell,* 15**,** 2020-2031.

LIEPMAN, A. H. & CAVALIER, D. 2012. The cellulose synthase-like A and cellulose synthase-like C families: recent advances and future perspectives. *Frontiers in plant science,* 3**,** 109.

LIEPMAN, A. H., WILKERSON, C. G. & KEEGSTRA, K. 2005. Expression of cellulose synthase-like (Csl) genes in insect cells reveals that CslA family members encode mannan synthases. *Proceedings of the National Academy of Sciences,* 102**,** 2221-2226.

LIERE, H., KIM, T. N., WERLING, B. P., MEEHAN, T. D., LANDIS, D. A. & GRATTON, C. 2015. Trophic cascades in agricultural landscapes: indirect effects of landscape composition on crop yield. *Ecological Applications,* 25**,** 652-661.

LIN, K., ZHANG, N., SEVERING, E. I., NIJVEEN, H., CHENG, F., VISSER, R. G., WANG, X., DE RIDDER, D. & BONNEMA, G. 2014. Beyond genomic variation-comparison and functional annotation of three Brassica rapagenomes: a turnip, a rapid cycling and a Chinese cabbage. *Bmc Genomics,* 15**,** 1-17.

LIN, Y.-R., SCHERTZ, K. F. & PATERSON, A. H. 1995. Comparative analysis of QTLs affecting plant height and maturity across the Poaceae, in reference to an interspecific sorghum population. *Genetics,* 141**,** 391-411.

LISCH, D. 2013. How important are transposons for plant evolution? *Nature Reviews Genetics,* 14**,** 49-61.

LITTLE, A., LAHNSTEIN, J., JEFFERY, D. W., KHOR, S. F., SCHWERDT, J. G., SHIRLEY, N. J., HOOI, M., XING, X., BURTON, R. A. & BULONE, V. 2019. A novel (1, 4)-β-linked glucoxylan is synthesized by members of the Cellulose synthase-like F gene family in land plants. *ACS central science,* 5**,** 73-84.

LITTLE, A., SCHWERDT, J. G., SHIRLEY, N. J., KHOR, S. F., NEUMANN, K., O'DONOVAN, L. A., LAHNSTEIN, J., COLLINS, H. M., HENDERSON, M. & FINCHER, G. B. 2018. Revised phylogeny of the cellulose synthase gene superfamily: insights into cell wall evolution. *Plant physiology,* 177**,** 1124-1141.

LIU, D., HUNT, M. & TSAI, I. J. 2018a. Inferring synteny between genome assemblies: a systematic evaluation. *BMC bioinformatics,* 19**,** 26.

LIU, H., WANG, X., WANG, G., CUI, P., WU, S., AI, C., HU, N., LI, A., HE, B. & SHAO, X. 2021. The nearly complete genome of Ginkgo biloba illuminates gymnosperm evolution. *Nature Plants,* 7**,** 748-756.

LIU, L., TUNICK, M., FISHMAN, M. L., HICKS, K. B., COOKE, P. H. & COFFIN, D. R. 2006. Pectin-based networks for non-food applications. ACS Publications.

LIU, Q., LUO, L. & ZHENG, L. 2018b. Lignins: Biosynthesis and biological functions in plants. *International journal of molecular sciences,* 19**,** 335.

LIU, X., WANG, Q., CHEN, P., SONG, F., GUAN, M., JIN, L., WANG, Y. & YANG, C. 2012. Four novel cellulose synthase (CESA) genes from birch (Betula platyphylla Suk.) involved in primary and secondary cell wall biosynthesis. *International journal of molecular sciences,* 13**,** 12195-12212.

LIU, Y., YOU, S., TAYLOR-TEEPLES, M., LI, W. L., SCHUETZ, M., BRADY, S. M. & DOUGLAS, C. J. 2014a. BEL1-LIKE HOMEODOMAIN6 and KNOTTED ARABIDOPSIS THALIANA7 interact and regulate secondary cell wall formation via repression of REVOLUTA. *The Plant Cell,* 26**,** 4843-4861.

LIU, Y., ZHANG, X., MIAO, J., HUANG, L., FRAZIER, T. & ZHAO, B. 2014b. Evaluation of salinity tolerance and genetic diversity of thirty-three switchgrass (Panicum virgatum) populations. *BioEnergy Research,* 7**,** 1329-1342.

LOQUE, D., SCHELLER, H. V. & PAULY, M. 2015. Engineering of plant cell walls for enhanced biofuel production. *Current opinion in plant biology,* 25**,** 151-161.

LU, F., LIPKA, A. E., GLAUBITZ, J., ELSHIRE, R., CHERNEY, J. H., CASLER, M. D., BUCKLER, E. S. & COSTICH, D. E. 2013. Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. *PLoS genetics,* 9**,** e1003215.

LU, J., SHI, Y., LI, W., CHEN, S., WANG, Y., HE, X. & YIN, X. 2019. RcPAL, a key gene in lignin biosynthesis in Ricinus communis L. *BMC plant biology,* 19**,** 1-11.

LUAN, W., LIU, Y., ZHANG, F., SONG, Y., WANG, Z., PENG, Y. & SUN, Z. 2011. OsCD1 encodes a putative member of the cellulose synthase-like D sub-family and is essential for rice plant architecture and growth. *Plant biotechnology journal,* 9**,** 513-524.

LÜBBERSTEDT, T., MELCHINGER, A. E., SCHÖN, C. C., UTZ, H. F. & KLEIN, D. 1997. QTL mapping in testcrosses of European flint lines of maize: I. Comparison of different testers for forage yield traits. *Crop science,* 37**,** 921-931.

LUDOVISI, R., TAURO, F., SALVATI, R., KHOURY, S., MUGNOZZA SCARASCIA, G. & HARFOUCHE, A. 2017. UAV-based thermal imaging for high-throughput field phenotyping of black poplar response to drought. *Frontiers in plant science,* 8**,** 1681.

LYE, Z. N. & PURUGGANAN, M. D. 2019. Copy number variation in domestication. *Trends in Plant Science,* 24**,** 352-365.

MAMMADOV, J., BUYYARAPU, R., GUTTIKONDA, S. K., PARLIAMENT, K., ABDURAKHMONOV, I. Y. & KUMPATLA, S. P. 2018. Wild relatives of maize, rice, cotton, and soybean: treasure troves for tolerance to biotic and abiotic stresses. *Frontiers in plant science,* 9**,** 886.

MANDROU, E., DENIS, M., PLOMION, C., SALIN, F., MORTIER, F. & GION, J.-M. 2014. Nucleotide diversity in lignification genes and QTNs for lignin quality in a multi-parental population of Eucalyptus urophylla. *Tree genetics & genomes,* 10**,** 1281-1290.

MARÇAIS, G., DELCHER, A. L., PHILLIPPY, A. M., COSTON, R., SALZBERG, S. L. & ZIMIN, A. 2018. MUMmer4: A fast and versatile genome alignment system. *PLoS computational biology,* 14**,** e1005944.

MARRON, N., BASTIEN, C., SABATTI, M., TAYLOR, G. & CEULEMANS, R. 2006. Plasticity of growth and sylleptic branchiness in two poplar families grown at three sites across Europe. *Tree physiology,* 26**,** 935-946.

MARROT, L., LEFEUVRE, A., PONTOIRE, B., BOURMAUD, A. & BALEY, C. 2013. Analysis of the hemp fiber mechanical properties and their scattering (Fedora 17). *Industrial Crops and Products,* 51**,** 317-327.

MCCORMICK, K. & KAUTTO, N. 2013. The bioeconomy in Europe: An overview. *Sustainability,* 5**,** 2589-2608.

MCISAAC, G. F., DAVID, M. B. & MITCHELL, C. A. 2010. Miscanthus and switchgrass production in central Illinois: impacts on hydrology and inorganic nitrogen leaching. *Journal of environmental quality,* 39**,** 1790-1799.

MCKOWN, A. D., KLÁPŠTĚ, J., GUY, R. D., GERALDES, A., PORTH, I., HANNEMANN, J., FRIEDMANN, M., MUCHERO, W., TUSKAN, G. A. & EHLTING, J. 2014. Genome-wide association implicates numerous genes underlying ecological trait variation in natural populations of Populus trichocarpa. *New Phytologist,* 203**,** 535-553.

MEADOWS, D. H., MEADOWS, D. L., RANDERS, J. & BEHRENS, W. W. 1972. *The limits to growth*, Universe Books.

MEEHAN, T. D., HURLBERT, A. H. & GRATTON, C. 2010. Bird communities in future bioenergy landscapes of the Upper Midwest. *Proceedings of the National Academy of Sciences***,** 201008475.

MEENTS, M. J., WATANABE, Y. & SAMUELS, A. L. 2018. The cell biology of secondary cell wall biosynthesis. *Annals of Botany,* 121**,** 1107-1125.

MEHMOOD, M. A., IBRAHIM, M., RASHID, U., NAWAZ, M., ALI, S., HUSSAIN, A. & GULL, M. 2017. Biomass production for bioenergy using marginal lands. *Sustainable Production and Consumption,* 9**,** 3-21.

MEYFROIDT, P., SCHIERHORN, F., PRISHCHEPOV, A. V., MÜLLER, D. & KUEMMERLE, T. 2016. Drivers, constraints and trade-offs associated with recultivating abandoned cropland in Russia, Ukraine and Kazakhstan. *Global environmental change,* 37**,** 1-15.

MICKELBART, M. V., HASEGAWA, P. M. & BAILEY-SERRES, J. 2015. Genetic mechanisms of abiotic stress tolerance that translate to crop yield stability. *Nature Reviews Genetics,* 16**,** 237.

MISTRY, J., FINN, R. D., EDDY, S. R., BATEMAN, A. & PUNTA, M. 2013. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic acids research,* 41**,** e121-e121.

MITCHELL, D. 2008. A note on rising food prices. *World bank policy research working paper*.

MITROS, T., SESSION, A. M., JAMES, B. T., WU, G. A., BELAFFIF, M. B., CLARK, L. V., SHU, S., DONG, H., BARLING, A. & HOLMES, J. R. 2020. Genome biology of the paleotetraploid perennial biomass crop Miscanthus. *Nature communications,* 11**,** 1-11.

MOHAMMED, Y. S. A., TAHIR, I. S. A., KAMAL, N. M., ELTAYEB, A. E., ALI, A. M. & TSUJIMOTO, H. 2014. Impact of wheat-Leymus racemosus added chromosomes on wheat adaptation and tolerance to heat stress. *Breeding science,* 63**,** 450-460.

MOL, A. P. 2010. Environmental authorities and biofuel controversies. *Environmental politics,* 19**,** 61-79.

MOLINARI, H. B., PELLNY, T. K., FREEMAN, J., SHEWRY, P. R. & MITCHELL, R. A. 2013. Grass cell wall feruloylation: distribution of bound ferulate and candidate gene expression in Brachypodium distachyon. *Frontiers in Plant Science,* 4**,** 50.

MONTEITH, J. 1972. Solar radiation and productivity in tropical ecosystems. *Journal of applied ecology,* 9**,** 747-766.

MONTEITH, J. L. 1977. Climate and the efficiency of crop production in Britain. *Phil. Trans. R. Soc. Lond. B,* 281**,** 277-294.

MORTIMER, J. C., MILES, G. P., BROWN, D. M., ZHANG, Z., SEGURA, M. P., WEIMAR, T., YU, X., SEFFEN, K. A., STEPHENS, E. & TURNER, S. R. 2010. Absence of branches from xylan in Arabidopsis gux mutants reveals potential for simplification of lignocellulosic biomass. *Proceedings of the National Academy of Sciences,* 107**,** 17409-17414.

MUELLER, K. E., TILMAN, D., FORNARA, D. A. & HOBBIE, S. E. 2013. Root depth distribution and the diversity–productivity relationship in a long-term grassland experiment. *Ecology,* 94**,** 787-793.

MULLET, J. E. 2017. High-biomass C4 grasses—Filling the yield gap. *Plant Science,* 261**,** 10-17.

MURRAY, S. C., ROONEY, W. L., MITCHELL, S. E., SHARMA, A., KLEIN, P. E., MULLET, J. E. & KRESOVICH, S. 2008. Genetic improvement of sorghum as a biofuel feedstock: II. QTL for stem and leaf structural carbohydrates. *Crop Science,* 48**,** 2180-2193.

MYBURG, A. A., GRATTAPAGLIA, D., TUSKAN, G. A., HELLSTEN, U., HAYES, R. D., GRIMWOOD, J., JENKINS, J., LINDQUIST, E., TICE, H. & BAUER, D. 2014. The genome of Eucalyptus grandis. *Nature,* 510**,** 356.

NADEZHDINA, N., DAVID, T. S., DAVID, J. S., NADEZHDIN, V., CERMAK, J., GEBAUER, R. & STOKES, A. 2012. Root structure: in situ studies through sap flow research. *Measuring Roots.* Springer.

NAKANO, Y., YAMAGUCHI, M., ENDO, H., REJAB, N. A. & OHTANI, M. 2015. NAC-MYB-based transcriptional regulation of secondary cell wall biosynthesis in land plants. *Frontiers in plant science,* 6**,** 288.

NAKHAMCHIK, A., ZHAO, Z., PROVART, N. J., SHIU, S.-H., KEATLEY, S. K., CAMERON, R. K. & GORING, D. R. 2004. A comprehensive expression analysis of the Arabidopsis proline-rich extensin-like receptor kinase gene family using bioinformatic and experimental approaches. *Plant and Cell Physiology,* 45**,** 1875-1881.

NARASIMHAMOORTHY, B., SAHA, M., SWALLER, T. & BOUTON, J. 2008. Genetic diversity in switchgrass collections assessed by EST-SSR markers. *Bioenergy Research,* 1**,** 136.

NATIONAL RESEARCH COUNCIL 2008. *Water implications of biofuels production in the United States*, National Academies Press.

NCBI RESOURCE COORDINATORS 2018. Database resources of the national center for biotechnology information. *Nucleic acids research,* 46**,** D8-D13.

NGARA, R. & NDIMBA, B. 2014. Model plant systems in salinity and drought stress proteomics studies: a perspective on Arabidopsis and Sorghum. *Plant Biology,* 16**,** 1029-1032.

NIAZIAN, M. & NIEDBAŁA, G. 2020. Machine learning for plant breeding and biotechnology. *Agriculture,* 10**,** 436.

NIJSEN, M., SMEETS, E., STEHFEST, E. & VUUREN, D. P. 2012. An evaluation of the global potential of bioenergy production on degraded lands. *Gcb Bioenergy,* 4, 130-147.

NORRIS, J. H., LI, X., HUANG, S., VAN DE MEENE, A. M., TRAN, M. L., KILLEAVY, E., CHAVES, A. M., MALLON, B., MERCURE, D. & TAN, H.-T. 2017. Functional specialization of cellulose synthase isoforms in a moss shows parallels with seed plants. *Plant Physiology,* 175, 210-222.

NOVAES, E., KIRST, M., CHIANG, V., WINTER-SEDEROFF, H. & SEDEROFF, R. 2010. Lignin and biomass: a negative correlation for wood formation and lignin content in trees. *Plant physiology,* 154, 555-561.

NUNEZ-LOPEZ, L., AGUIRRE-CRUZ, A., BARRERA-FIGUEROA, B. E. & PENA-CASTRO, J. M. 2015. Improvement of enzymatic saccharification yield in Arabidopsis thaliana by ectopic expression of the rice SUB1A-1 transcription factor. *PeerJ,* 3, e817.

OECD 2017. Biofuels. *OECD-FAO Agricultural Outlook 2017-2026.* Paris: OECD Publishing.

OHL, S., HEDRICK, S. A., CHORY, J. & LAMB, C. J. 1990. Functional properties of a phenylalanine ammonia-lyase promoter from Arabidopsis. *The Plant Cell,* 2, 837-848.

ÖHMAN, D., DEMEDTS, B., KUMAR, M., GERBER, L., GORZSÁS, A., GOEMINNE, G., HEDENSTRÖM, M., ELLIS, B., BOERJAN, W. & SUNDBERG, B. 2013. MYB 103 is required for FERULATE-5-HYDROXYLASE expression and syringyl lignin biosynthesis in A rabidopsis stems. *The Plant Journal,* 73, 63-76.

ORLOVIC, S., GUZINA, V., KRSTIC, B. & MERKULOV, L. 1998. Genetic variability in anatomical, physiological and growth characteristics of hybrid poplar (Populus x euramericana Dode (Guinier)) and eastern cottonwood (Populus deltoides Bartr.) clones. *Silvae genetica,* 47, 183-189.

PAN, Q., XU, Y., LI, K., PENG, Y., ZHAN, W., LI, W., LI, L. & YAN, J. 2017. The genetic basis of plant architecture in 10 maize recombinant inbred line populations. *Plant physiology*, pp. 00709.2017.

PANCALDI, F. & TRINDADE, L. M. 2020. Marginal lands to grow novel bio-based crops: A plant breeding perspective. *Frontiers in plant science,* 11, 227.

PANCALDI, F., VAN LOO, E. N., SCHRANZ, M. E. & TRINDADE, L. M. 2022a. Genomic Architecture and Evolution of the Cellulose synthase Gene Superfamily as Revealed by Phylogenomic Analysis. *Frontiers in plant science,* 13.

PANCALDI, F., VLEGELS, D., RIJKEN, H., VAN LOO, E. N. & TRINDADE, L. M. 2022b. Detection and Analysis of Syntenic Quantitative Trait Loci Controlling Cell Wall Quality in Angiosperms. *Frontiers in plant science,* 13.

PANCHY, N., LEHTI-SHIU, M. & SHIU, S.-H. 2016. Evolution of gene duplication in plants. *Plant physiology,* 171, 2294-2316.

PANG, C. H., LESTER, E. & WU, T. 2018. Influence of lignocellulose and plant cell walls on biomass char morphology and combustion reactivity. *Biomass and Bioenergy,* 119, 480-491.

PARADIS, E. & SCHLIEP, K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics,* 35, 526-528.

PARENTI, A., LAMBERTINI, C. & MONTI, A. 2018. Areas with Natural Constraints to Agriculture: Possibilities and Limitations for The Cultivation of Switchgrass (Panicum Virgatum L.) and Giant Reed (Arundo Donax L.) in Europe. *Land Allocation for Biomass Crops.* Springer.

PASSARGE, E., HORSTHEMKE, B. & FARBER, R. A. 1999. Incorrect use of the term synteny. *Nature genetics,* 23, 387-387.

PATERSON, A. H., SCHERTZ, K. F., LIN, Y.-R., LIU, S.-C. & CHANG, Y.-L. 1995. The weediness of wild plants: molecular analysis of genes influencing dispersal and persistence of johnsongrass, Sorghum halepense (L.) Pers. *Proceedings of the National Academy of Sciences,* 92, 6127-6131.

PAULY, M. & KEEGSTRA, K. 2010. Plant cell wall polymers as precursors for biofuels. *Current opinion in plant biology,* 13, 304-311.

PAWAR, P. M. A., RATKE, C., BALASUBRAMANIAN, V. K., CHONG, S. L., GANDLA, M. L., ADRIASOLA, M., SPARRMAN, T., HEDENSTRÖM, M., SZWAJ, K. & DERBA-MACELUCH, M. 2017. Downregulation of RWA genes in hybrid aspen affects xylan acetylation and wood saccharification. *New Phytologist,* 214**,** 1491-1505.

PE'ER, G., BONN, A., BRUELHEIDE, H., DIEKER, P., EISENHAUER, N., FEINDT, P. H., HAGEDORN, G., HANSJÜRGENS, B., HERZON, I. & LOMBA, Â. 2020. Action needed for the EU Common Agricultural Policy to address sustainability challenges. *People and Nature,* 2**,** 305-316.

PEAR, J. R., KAWAGOE, Y., SCHRECKENGOST, W. E., DELMER, D. P. & STALKER, D. M. 1996. Higher plants contain homologs of the bacterial celA genes encoding the catalytic subunit of cellulose synthase. *Proceedings of the National Academy of Sciences,* 93**,** 12637-12642.

PEETERS, A., LEFEBVRE, O., BALOGH, L., BARBERI, P., BATELLO, C., BELLON, S., GAIFAMI, T., GKISAKIS, V., LANA, M. & MIGLIORINI, P. 2020. A Green Deal for implementing agroecological systems: reforming the common agricultural policy of the European Union.

PEÑA-CASTRO, J. M., VAN ZANTEN, M., LEE, S. C., PATEL, M. R., VOESENEK, L. A., FUKAO, T. & BAILEY-SERRES, J. 2011. Expression of rice SUB1A and SUB1C transcription factors in Arabidopsis uncovers flowering inhibition as a submergence tolerance mechanism. *The plant journal,* 67**,** 434-446.

PENA, M. J., KULKARNI, A. R., BACKE, J., BOYD, M., O'NEILL, M. A. & YORK, W. S. 2016. Structural diversity of xylans in the cell walls of monocots. *Planta,* 244**,** 589-606.

PENG, X., PANG, H., ABBAS, M., YAN, X., DAI, X., LI, Y. & LI, Q. 2019. Characterization of Cellulose synthase-like D (CSLD) family revealed the involvement of PtrCslD5 in root hair formation in Populus trichocarpa. *Scientific reports,* 9**,** 1-9.

PENNING, B. W., HUNTER III, C. T., TAYENGWA, R., EVELAND, A. L., DUGARD, C. K., OLEK, A. T., VERMERRIS, W., KOCH, K. E., MCCARTY, D. R. & DAVIS, M. F. 2009. Genetic resources for maize cell wall biology. *Plant physiology,* 151**,** 1703-1728.

PENNING, B. W., MCCANN, M. C. & CARPITA, N. C. 2019. Evolution of the cell wall gene families of grasses. *Frontiers in plant science,* 10**,** 1205.

PÉREZ-SUÁREZ, M., CASTELLANO, M. J., KOLKA, R., ASBJORNSEN, H. & HELMERS, M. 2014. Nitrogen and carbon dynamics in prairie vegetation strips across topographical gradients in mixed Central Iowa agroecosystems. *Agriculture, ecosystems & environment,* 188**,** 1-11.

PETERSON, G. M. & GALBRAITH, J. 1932. The concept of marginal land. *Journal of Farm Economics,* 14**,** 295-310.

PETIT, J., SALENTIJN, E. M., PAULO, M.-J., DENNEBOOM, C., VAN LOO, E. N. & TRINDADE, L. M. 2020a. Elucidating the genetic architecture of fiber quality in hemp (Cannabis sativa L.) using a genome-wide association study. *Frontiers in genetics***,** 1101.

PETIT, J., SALENTIJN, E. M., PAULO, M.-J., THOUMINOT, C., VAN DINTER, B. J., MAGAGNINI, G., GUSOVIUS, H.-J., TANG, K., AMADUCCI, S. & WANG, S. 2020b. Genetic variability of morphological, flowering, and biomass quality traits in hemp (Cannabis sativa L.). *Frontiers in plant science,* 11**,** 102.

PETTOLINO, F. A., WALSH, C., FINCHER, G. B. & BACIC, A. 2012. Determining the polysaccharide composition of plant cell walls. *Nature protocols,* 7**,** 1590-1607.

PFLIEGER, S., LEFEBVRE, V. & CAUSSE, M. 2001. The candidate gene approach in plant genetics: a review. *Molecular breeding,* 7**,** 275-291.

PICART-PICOLO, A., GROB, S., PICAULT, N., FRANEK, M., LLAURO, C., HALTER, T., MAIER, T. R., JOBET, E., DESCOMBIN, J. & ZHANG, P. 2020. Large tandem duplications affect gene expression, 3D organization, and plant–pathogen response. *Genome research,* 30**,** 1583-1592.

PIERRET, A., MAEGHT, J.-L., CLÉMENT, C., MONTOROI, J.-P., HARTMANN, C. & GONKHAMDEE, S. 2016. Understanding deep roots and their functions in ecosystems: an advocacy for more unconventional research. *Annals of botany,* 118**,** 621-635.

PINGALI, P. L. 2012. Green revolution: impacts, limits, and the path ahead. *Proceedings of the National Academy of Sciences,* 109**,** 12302-12308.

PIOTROWSKI, S., CARUS, M. & ESSEL, R. 2015. Global bioeconomy in the conflict between biomass supply and demand. *Industrial Biotechnology,* 11**,** 308-315.

POOVAIAH, C. R., MAZAREI, M., DECKER, S. R., TURNER, G. B., SYKES, R. W., DAVIS, M. F. & STEWART, C. N. 2015. Transgenic switchgrass (Panicum virgatum L.) biomass is increased by overexpression of switchgrass sucrose synthase (PvSUS1). *Biotechnology journal,* 10**,** 552-563.

POPPER, Z. A. 2008. Evolution and diversity of green plant cell walls. *Current opinion in plant biology,* 11**,** 286-292.

PORTH, I., KLAPŠTE, J., SKYBA, O., HANNEMANN, J., MCKOWN, A. D., GUY, R. D., DIFAZIO, S. P., MUCHERO, W., RANJAN, P. & TUSKAN, G. A. 2013. Genome-wide association mapping for wood characteristics in Populus identifies an array of candidate single nucleotide polymorphisms. *New Phytologist,* 200**,** 710-726.

POST, W. M., NICHOLS, J. A., WANG, D., WEST, T. O., BANDARU, V. & IZAURRALDE, R. C. 2013. Marginal lands: concept, assessment and management. *Journal of Agricultural Science,* 5**,** 129.

PREISNER, M., KULMA, A., ZEBROWSKI, J., DYMIŃSKA, L., HANUZA, J., ARENDT, M., STARZYCKI, M. & SZOPA, J. 2014. Manipulating cinnamyl alcohol dehydrogenase (CAD) expression in flax affects fibre composition and properties. *BMC plant biology,* 14**,** 1-18.

PRIEFER, C., JÖRISSEN, J. & FRÖR, O. 2017. Pathways to shape the bioeconomy. *Resources,* 6**,** 10.

PURUGGANAN, M. D. & JACKSON, S. A. 2021. Advancing crop genomics from lab to field. *Nature Genetics,* 53**,** 595-601.

QIAO, X., LI, Q., YIN, H., QI, K., LI, L., WANG, R., ZHANG, S. & PATERSON, A. H. 2019. Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. *Genome biology,* 20**,** 1-23.

QIN, W., YIN, Q., CHEN, J., ZHAO, X., YUE, F., HE, J., YANG, L., LIU, L., ZENG, Q. & LU, F. 2020. The class II KNOX transcription factors KNAT3 and KNAT7 synergistically regulate monolignol biosynthesis in Arabidopsis. *Journal of experimental botany,* 71**,** 5469-5483.

QIN, Z., GRISCOM, B., HUANG, Y., YUAN, W., CHEN, X., DONG, W., LI, T., SANDERMAN, J., SMITH, P. & WANG, F. 2021. Delayed impact of natural climate solutions. *Global Change Biology,* 27**,** 215-217.

QUINN, L. D., STRAKER, K. C., GUO, J., KIM, S., THAPA, S., KLING, G., LEE, D. & VOIGT, T. B. 2015. Stress-tolerant feedstocks for sustainable bioenergy production on marginal land. *BioEnergy Research,* 8**,** 1081-1100.

RAE, A., ROBINSON, K., STREET, N. & TAYLOR, G. 2004. Morphological and physiological traits influencing biomass productivity in short-rotation coppice poplar. *Canadian Journal of Forest Research,* 34**,** 1488-1498.

RAE, A. M., PINEL, M. P., BASTIEN, C., SABATTI, M., STREET, N. R., TUCKER, J., DIXON, C., MARRON, N., DILLEN, S. Y. & TAYLOR, G. 2008. QTL for yield in bioenergy Populus: identifying G× E interactions from growth at three contrasting sites. *Tree Genetics & Genomes,* 4**,** 97-112.

RAFALSKI, J. A. 2010. Association genetics in crop improvement. *Current opinion in plant biology,* 13**,** 174-180.

RANADE, S. A. & YADAV, H. 2014. Universal molecular markers for plant breeding and genetics analysis. *Journal of Plant Biochemistry & Physiology*.

RANIK, M. & MYBURG, A. A. 2006. Six new cellulose synthase genes from Eucalyptus are associated with primary and secondary cell wall biosynthesis. *Tree physiology,* 26**,** 545-556.

RANJAN, P., YIN, T., ZHANG, X., KALLURI, U. C., YANG, X., JAWDY, S. & TUSKAN, G. A. 2010. Bioinformatics-based identification of candidate genes from QTLs associated with cell wall traits in Populus. *BioEnergy Research,* 3**,** 172-182.

RAO, X. & DIXON, R. A. 2018. Current Models for Transcriptional Regulation of Secondary Cell Wall Biosynthesis in Grasses. *Frontiers in plant science,* 9**,** 399.

REECK, G. R., DE HAËN, C., TELLER, D. C., DOOLITTLE, R. F., FITCH, W. M., DICKERSON, R. E., CHAMBON, P., MCLACHLAN, A. D., MARGOLIASH, E. & JUKES, T. H. 1987. "Homology" in proteins and nucleic acids: a terminology muddle and a way out of it. *Cell,* 50**,** 667.

REEVES, T. G. & CASSADAY, K. 2002. History and past achievements of plant breeding. *Australian Journal of Agricultural Research,* 53**,** 851-863.

REINHARDT, D. & KUHLEMEIER, C. 2002. Plant architecture. *EMBO reports,* 3**,** 846-851.

RENAULT, H., WERCK-REICHHART, D. & WENG, J.-K. 2019. Harnessing lignin evolution for biotechnological applications. *Current opinion in biotechnology,* 56**,** 105-111.

RENWICK, J. H. 1971. The mapping of human chromosomes. *Annual review of genetics,* 5**,** 81-120.

RIBEIRO, B. E. 2013. Beyond commonplace biofuels: Social aspects of ethanol. *Energy Policy,* 57**,** 355-362.

RICHMOND, T. A. & SOMERVILLE, C. R. 2000. The cellulose synthase superfamily. *Plant physiology,* 124**,** 495-498.

RITTER, K. B., JORDAN, D. R., CHAPMAN, S. C., GODWIN, I. D., MACE, E. S. & MCINTYRE, C. L. 2008. Identification of QTL for sugar-related traits in a sweet× grain sorghum (Sorghum bicolor L. Moench) recombinant inbred population. *Molecular Breeding,* 22**,** 367-384.

ROBERTSON, B. A., DORAN, P. J., LOOMIS, E. R., ROBERTSON, J. R. & SCHEMSKE, D. W. 2011a. Avian use of perennial biomass feedstocks as post-breeding and migratory stopover habitat. *PloS one,* 6**,** e16941.

ROBERTSON, B. A., DORAN, P. J., LOOMIS, L. R., ROBERTSON, J. R. & SCHEMSKE, D. W. 2011b. Perennial biomass feedstocks enhance avian diversity. *Gcb Bioenergy,* 3**,** 235-246.

ROBERTSON, G. P., HAMILTON, S. K., BARHAM, B. L., DALE, B. E., IZAURRALDE, R. C., JACKSON, R. D., LANDIS, D. A., SWINTON, S. M., THELEN, K. D. & TIEDJE, J. M. 2017. Cellulosic biofuel contributions to a sustainable energy future: Choices and outcomes. *Science,* 356**,** eaal2324.

ROBERTSON, G. P., HAMILTON, S. K., DEL GROSSO, S. J. & PARTON, W. J. 2011c. The biogeochemistry of bioenergy landscapes: carbon, nitrogen, and water considerations. *Ecological Applications,* 21**,** 1055-1067.

ROBERTSON, G. P., PAUL, E. A. & HARWOOD, R. R. 2000. Greenhouse gases in intensive agriculture: contributions of individual gases to the radiative forcing of the atmosphere. *Science,* 289**,** 1922-1925.

ROBSON, P., JENSEN, E., HAWKINS, S., WHITE, S. R., KENOBI, K., CLIFTON-BROWN, J., DONNISON, I. & FARRAR, K. 2013. Accelerating the domestication of a bioenergy crop: identifying and modelling morphological targets for sustainable yield increase in Miscanthus. *Journal of Experimental Botany,* 64**,** 4143-4155.

ROBSON, P., MOS, M., CLIFTON-BROWN, J. & DONNISON, I. 2012. Phenotypic variation in senescence in Miscanthus: towards optimising biomass quality and quantity. *Bioenergy Research,* 5**,** 95-105.

ROMAN, V. J., DEN TOOM, L. A., GAMIZ, C. C., VAN DER PIJL, N., VISSER, R. G., VAN LOO, E. N. & VAN DER LINDEN, C. G. 2020. Differential responses to salt stress in ion dynamics, growth and seed yield of European quinoa varieties. *Environmental and Experimental Botany,* 177**,** 104146.

ROSVALL, M., AXELSSON, D. & BERGSTROM, C. T. 2009. The map equation. *The European Physical Journal Special Topics,* 178**,** 13-23.

ROSVALL, M. & BERGSTROM, C. T. 2007. An information-theoretic framework for resolving community structure in complex networks. *Proceedings of the national academy of sciences,* 104**,** 7327-7331.

ROY, S. J., NEGRÃO, S. & TESTER, M. 2014. Salt resistant crop plants. *Current opinion in Biotechnology,* 26**,** 115-124.

SAELIM, L., AKIYOSHI, N., TAN, T. T., IHARA, A., YAMAGUCHI, M., HIRANO, K., MATSUOKA, M., DEMURA, T. & OHTANI, M. 2019. Arabidopsis Group IIId ERF proteins positively regulate primary cell wall-type CESA genes. *Journal of plant research,* 132**,** 117-129.

SALENTIJN, E. M., PEREIRA, A., ANGENENT, G. C., VAN DER LINDEN, C. G., KRENS, F., SMULDERS, M. J. & VOSMAN, B. 2007. Plant translational genomics: from model species to crops. *Molecular Breeding,* 20**,** 1-13.

SALENTIJN, E. M., ZHANG, Q., AMADUCCI, S., YANG, M. & TRINDADE, L. M. 2015. New developments in fiber hemp (Cannabis sativa L.) breeding. *Industrial crops and products,* 68**,** 32-41.

SAMSON, R. & GIROUARD, P. 1998. Bioenergy opportunities from agriculture. *Resource Efficient Agricultural Production-Canada***,** 1-4.

SANGHERA, G. S., WANI, S. H., HUSSAIN, W. & SINGH, N. 2011. Engineering cold stress tolerance in crop plants. *Current genomics,* 12**,** 30.

SARI, Y. W., SYAFITRI, U., SANDERS, J. P. & BRUINS, M. E. 2015. How biomass composition determines protein extractability. *Industrial Crops and Products,* 70**,** 125-133.

SARKAR, P., BOSNEAGA, E. & AUER, M. 2009. Plant cell walls throughout evolution: towards a molecular understanding of their design principles. *Journal of Experimental Botany,* 60**,** 3615-3635.

SARTONI, R., ZEGADA-LIZARAZU, W. & MONTI, A. 2015. A new compartmentalised rhizotron system for root phenotyping. *Italian Journal of Agronomy,* 10**,** 53-58.

SATTLER, S. E., FUNNELL-HARRIS, D. L. & PEDERSEN, J. F. 2010. Brown midrib mutations and their importance to the utilization of maize, sorghum, and pearl millet lignocellulosic tissues. *Plant Science,* 178**,** 229-238.

SCAGLIONE, D., PINOSIO, S., MARRONI, F., DI CENTA, E., FORNASIERO, A., MAGRIS, G., SCALABRIN, S., CATTONARO, F., TAYLOR, G. & MORGANTE, M. 2019. Single primer enrichment technology as a tool for massive genotyping: a benchmark on black poplar and maize. *Annals of botany,* 124**,** 543-551.

SCHMIDT, T., FERNANDO, A. L., MONTI, A. & RETTENMAIER, N. 2015. Life cycle assessment of bioenergy and bio-based products from perennial grasses cultivated on marginal land in the Mediterranean region. *BioEnergy Research,* 8**,** 1548-1561.

SCHULER, M. L., MANTEGAZZA, O. & WEBER, A. P. 2016. Engineering C4 photosynthesis into C3 chassis in the synthetic biology age. *The Plant Journal,* 87**,** 51-65.

SCHULZE, E. D., KÖRNER, C., LAW, B. E., HABERL, H. & LUYSSAERT, S. 2012. Large-scale bioenergy from additional harvest of forest biomass is neither sustainable nor greenhouse gas neutral. *Gcb Bioenergy,* 4**,** 611-616.

SCHWERDT, J. G., MACKENZIE, K., WRIGHT, F., OEHME, D., WAGNER, J. M., HARVEY, A. J., SHIRLEY, N. J., BURTON, R. A., SCHREIBER, M. & HALPIN, C. 2015. Evolutionary dynamics of the cellulose synthase gene superfamily in grasses. *Plant physiology,* 168**,** 968-983.

SEARCHINGER, T., HEIMLICH, R., HOUGHTON, R. A., DONG, F., ELOBEID, A., FABIOSA, J., TOKGOZ, S., HAYES, D. & YU, T.-H. 2008. Use of US croplands for biofuels increases greenhouse gases through emissions from land-use change. *Science,* 319**,** 1238-1240.

SEPPEY, M., MANNI, M. & ZDOBNOV, E. M. 2019. BUSCO: assessing genome assembly and annotation completeness. *Gene prediction.* Springer.

SERBA, D. D., DAVERDIN, G., BOUTON, J. H., DEVOS, K. M., BRUMMER, E. C. & SAHA, M. C. 2015. Quantitative trait loci (QTL) underlying biomass yield and plant height in switchgrass. *BioEnergy research,* 8**,** 307-324.

SHAKOOR, N., NORTHRUP, D., MURRAY, S. & MOCKLER, T. C. 2019. Big data driven agriculture: big data analytics in plant breeding, genomics, and the use of remote sensing technologies to advance crop productivity. *The Plant Phenome Journal,* 2**,** 1-8.

SHI, W.-Y., DU, Y.-T., MA, J., MIN, D.-H., JIN, L.-G., CHEN, J., CHEN, M., ZHOU, Y.-B., MA, Y.-Z. & XU, Z.-S. 2018. The WRKY transcription factor GmWRKY12 confers drought and salt tolerance in soybean. *International journal of molecular sciences,* 19**,** 4087.

SHORTALL, O. 2013. "Marginal land" for energy crops: Exploring definitions and embedded assumptions. *Energy Policy,* 62**,** 19-27.

SINDHU, A., LANGEWISCH, T., OLEK, A., MULTANI, D. S., MCCANN, M. C., VERMERRIS, W., CARPITA, N. C. & JOHAL, G. 2007. Maize Brittle stalk2 encodes a COBRA-like protein expressed in early organ development but required for tissue flexibility at maturity. *Plant Physiology,* 145**,** 1444-1459.

SIXTO, H., GRAU, J., ALBA, N. & ALIA, R. 2005. Response to sodium chloride in different species and clones of genus Populus L. *Forestry,* 78**,** 93-104.

SMALIYCHUK, A., MÜLLER, D., PRISHCHEPOV, A. V., LEVERS, C., KRUHLOV, I. & KUEMMERLE, T. 2016. Recultivation of abandoned agricultural lands in Ukraine: Patterns and drivers. *Global Environmental Change,* 38**,** 70-81.

SMEETS, E. M., FAAIJ, A. P., LEWANDOWSKI, I. M. & TURKENBURG, W. C. 2007. A bottom-up assessment and review of global bio-energy potentials to 2050. *Progress in Energy and combustion science,* 33**,** 56-106.

SMITH, P. J., WANG, H.-T., YORK, W. S., PEÑA, M. J. & URBANOWICZ, B. R. 2017. Designer biomass for next-generation biorefineries: leveraging recent insights into xylan structure and biosynthesis. *Biotechnology for biofuels,* 10**,** 1-14.

SOERGEL, B., KRIEGLER, E., WEINDL, I., RAUNER, S., DIRNAICHNER, A., RUHE, C., HOFMANN, M., BAUER, N., BERTRAM, C. & BODIRSKY, B. L. 2021. A sustainable development pathway for climate action within the UN 2030 Agenda. *Nature Climate Change,* 11**,** 656-664.

SOLDATOS, P. 2015. Economic aspects of bioenergy production from perennial grasses in marginal lands of South Europe. *BioEnergy Research,* 8**,** 1562-1573.

SOMERVILLE, C., BAUER, S., BRININSTOOL, G., FACETTE, M., HAMANN, T., MILNE, J., OSBORNE, E., PAREDEZ, A., PERSSON, S. & RAAB, T. 2004. Toward a systems approach to understanding plant cell walls. *Science,* 306**,** 2206-2211.

SONG, J.-M., GUAN, Z., HU, J., GUO, C., YANG, Z., WANG, S., LIU, D., WANG, B., LU, S. & ZHOU, R. 2020. Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. *Nature Plants,* 6**,** 34-45.

SONG, Q., ZHANG, G. & ZHU, X.-G. 2013. Optimal crop canopy architecture to maximise canopy photosynthetic CO2 uptake under elevated CO2–a theoretical study using a mechanistic model of canopy photosynthesis. *Functional Plant Biology,* 40**,** 108-124.

SONG, X., XU, L., YU, J., TIAN, P., HU, X., WANG, Q. & PAN, Y. 2019. Genome-wide characterization of the cellulose synthase gene superfamily in Solanum lycopersicum. *Gene,* 688**,** 71-83.

SØRENSEN, I., DOMOZYCH, D. & WILLATS, W. G. 2010. How have plant cell walls evolved? *Plant physiology,* 153**,** 366-372.

SPECK, T. & BURGERT, I. 2011. Plant stems: functional design and mechanics. *Annual review of materials research,* 41**,** 169-193.

STAMATAKIS, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics,* 30**,** 1312-1313.

STELKENS, R. & SEEHAUSEN, O. 2009. Genetic distance between species predicts novel trait expression in their hybrids. *Evolution: International Journal of Organic Evolution,* 63**,** 884-897.

STEWART, J. J., AKIYAMA, T., CHAPPLE, C., RALPH, J. & MANSFIELD, S. D. 2009. The effects on lignin structure of overexpression of ferulate 5-hydroxylase in hybrid poplar1. *Plant physiology,* 150**,** 621-635.

STRÖMBERG, C. A. 2011. Evolution of grasses and grassland ecosystems. *Annual review of Earth and planetary sciences,* 39**,** 517-544.

TADELE, Z. 2019. Orphan crops: their importance and the urgency of improvement. *Planta,* 250**,** 677-694.

TAI, T. H. & TANKSLEY, S. D. 1990. A rapid and inexpensive method for isolation of total DNA from dehydrated plant tissue. *Plant Molecular Biology Reporter,* 8**,** 297-303.

TAMASLOUKHT, B., WONG QUAI LAM, M. S.-J., MARTINEZ, Y., TOZO, K., BARBIER, O., JOURDA, C., JAUNEAU, A., BORDERIES, G., BALZERGUE, S. & RENOU, J.-P. 2011. Characterization of a cinnamoyl-CoA reductase 1 (CCR1) mutant in maize: effects on lignification, fibre development, and global gene expression. *Journal of experimental botany,* 62**,** 3837-3848.

TANG, H., BOWERS, J. E., WANG, X., MING, R., ALAM, M. & PATERSON, A. H. 2008. Synteny and collinearity in plant genomes. *Science,* 320**,** 486-488.

TANG, S., LIANG, H., YAN, D., ZHAO, Y., HAN, X., CARLSON, J. E., XIA, X. & YIN, W. 2013. Populus euphratica: the transcriptomic response to drought stress. *Plant molecular biology,* 83**,** 539-557.

TAYLOR-TEEPLES, M., LIN, L., DE LUCAS, M., TURCO, G., TOAL, T., GAUDINIER, A., YOUNG, N., TRABUCCO, G., VELING, M. & LAMOTHE, R. 2015. An Arabidopsis gene regulatory network for secondary cell wall synthesis. *Nature,* 517**,** 571-575.

TAYLOR, N. G., GARDINER, J. C., WHITEMAN, R. & TURNER, S. R. 2004. Cellulose synthesis in the Arabidopsis secondary cell wall. *Cellulose,* 11**,** 329-338.

TEICHMANN, T. & MUHR, M. 2015. Shaping plant architecture. *Frontiers in plant science,* 6**,** 233.

TEUFEL, F., ALMAGRO ARMENTEROS, J. J., JOHANSEN, A. R., GÍSLASON, M. H., PIHL, S. I., TSIRIGOS, K. D., WINTHER, O., BRUNAK, S., VON HEIJNE, G. & NIELSEN, H. 2022. SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nature biotechnology***,** 1-3.

THAVAMANIKUMAR, S., SOUTHERTON, S., SOUTHERON, R., BRAWNER, J. & THUMMA, B. 2018. Eucalypt MAS: Implementation of marker-assisted selection in Australia's major plantation eucalypts. Melbourne: Gondwana Genomics Pty Ltd.

THUMMA, B. R., SOUTHERTON, S. G., BELL, J. C., OWEN, J. V., HENERY, M. L. & MORAN, G. F. 2010. Quantitative trait locus (QTL) analysis of wood quality traits in Eucalyptus nitens. *Tree Genetics & Genomes,* 6**,** 305-317.

THUMULURI, V., ALMAGRO ARMENTEROS, J. J., JOHANSEN, A. R., NIELSEN, H. & WINTHER, O. 2022. DeepLoc 2.0: multi-label subcellular localization prediction using protein language models. *Nucleic Acids Research*.

TIAN, F., BRADBURY, P. J., BROWN, P. J., HUNG, H., SUN, Q., FLINT-GARCIA, S., ROCHEFORD, T. R., MCMULLEN, M. D., HOLLAND, J. B. & BUCKLER, E. S. 2011. Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nature genetics,* 43**,** 159.

TILMAN, D., BALZER, C., HILL, J. & BEFORT, B. L. 2011. Global food demand and the sustainable intensification of agriculture. *Proceedings of the national academy of sciences,* 108**,** 20260-20264.

TILMAN, D., HILL, J. & LEHMAN, C. 2006. Carbon-negative biofuels from low-input high-diversity grassland biomass. *Science,* 314**,** 1598-1600.

TILMAN, D., SOCOLOW, R., FOLEY, J. A., HILL, J., LARSON, E., LYND, L., PACALA, S., REILLY, J., SEARCHINGER, T. & SOMERVILLE, C. 2009. Beneficial biofuels—the food, energy, and environment trilemma. *Science,* 325**,** 270-271.

TOPP, C. N., IYER-PASCUZZI, A. S., ANDERSON, J. T., LEE, C.-R., ZUREK, P. R., SYMONOVA, O., ZHENG, Y., BUCKSCH, A., MILEYKO, Y. & GALKOVSKYI, T. 2013. 3D phenotyping and quantitative trait locus mapping identify core regions of the rice genome controlling root architecture. *Proceedings of the National Academy of Sciences,* 110**,** E1695-E1704.

TORRES, A. F., NOORDAM-BOOT, C. M., DOLSTRA, O., VAN DER WEIJDE, T., COMBES, E., DUFOUR, P., VLASWINKEL, L., VISSER, R. G. & TRINDADE, L. M. 2015a. Cell wall diversity in forage maize: genetic complexity and bioenergy potential. *BioEnergy Research,* 8**,** 187-202.

TORRES, A. F., VAN DER WEIJDE, T., DOLSTRA, O., VISSER, R. G. & TRINDADE, L. M. 2013. Effect of maize biomass composition on the optimization of dilute-acid pretreatments and enzymatic saccharification. *Bioenergy Research,* 6**,** 1038-1051.

TORRES, A. F., VISSER, R. G. & TRINDADE, L. M. 2015b. Bioethanol from maize cell walls: genes, molecular tools, and breeding prospects. *Gcb Bioenergy,* 7**,** 591-607.

TRINDADE, L., DOLSTRA, O., VAN LOO, E. R. & VISSER, R. 2010. Plant breeding and its role in a biobased economy. *The Biobased Economy.* Routledge.

TSIAFOULI, M. A., THÉBAULT, E., SGARDELIS, S. P., DE RUITER, P. C., VAN DER PUTTEN, W. H., BIRKHOFER, K., HEMERIK, L., DE VRIES, F. T., BARDGETT, R. D. & BRADY, M. V. 2015. Intensive agriculture reduces soil biodiversity across Europe. *Global change biology,* 21**,** 973-985.

TUOMINEN, L. K., JOHNSON, V. E. & TSAI, C.-J. 2011. Differential phylogenetic expansions in BAHD acyltransferases across five angiosperm taxa and evidence of divergent expression among Populus paralogues. *BMC genomics,* 12**,** 1-17.

TURNER, S. R. & SOMERVILLE, C. R. 1997. Collapsed xylem phenotype of Arabidopsis identifies mutants deficient in cellulose deposition in the secondary cell wall. *The plant cell,* 9**,** 689-701.

TUSKAN, G. A., DIFAZIO, S., JANSSON, S., BOHLMANN, J., GRIGORIEV, I., HELLSTEN, U., PUTNAM, N., RALPH, S., ROMBAUTS, S. & SALAMOV, A. 2006. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *science,* 313**,** 1596-1604.

UNIPROT CONSORTIUM 2018. UniProt: the universal protein knowledgebase. *Nucleic acids research,* 46**,** 2699.

VALENTINE, J., CLIFTON-BROWN, J., HASTINGS, A., ROBSON, P., ALLISON, G. & SMITH, P. 2012. Food vs. fuel: the use of land for lignocellulosic 'next generation'energy crops that minimize competition with primary food production. *Gcb Bioenergy,* 4**,** 1-19.

VAN BEL, M., DIELS, T., VANCAESTER, E., KREFT, L., BOTZKI, A., VAN DE PEER, Y., COPPENS, F. & VANDEPOELE, K. 2017. PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. *Nucleic acids research,* 46**,** D1190-D1196.

VAN DE PEER, Y., MIZRACHI, E. & MARCHAL, K. 2017. The evolutionary significance of polyploidy. *Nature Reviews Genetics,* 18**,** 411-424.

VAN DER CRUIJSEN, K., AL HASSAN, M., VAN ERVEN, G., DOLSTRA, O. & TRINDADE, L. M. 2021. Breeding Targets to Improve Biomass Quality in Miscanthus. *Molecules,* 26**,** 254.

VAN DER WEIJDE, T., ALVIM KAMEI, C. L., TORRES, A. F., VERMERRIS, W., DOLSTRA, O., VISSER, R. G. & TRINDADE, L. M. 2013. The potential of C4 grasses for cellulosic biofuel production. *Front Plant Sci,* 4**,** 107.

VAN DER WEIJDE, T., HUXLEY, L. M., HAWKINS, S., SEMBIRING, E. H., FARRAR, K., DOLSTRA, O., VISSER, R. G. & TRINDADE, L. M. 2017a. Impact of drought stress on growth and quality of miscanthus for biofuel production. *Gcb Bioenergy,* 9**,** 770-782.

VAN DER WEIJDE, T., KAMEI, C. L. A., SEVERING, E. I., TORRES, A. F., GOMEZ, L. D., DOLSTRA, O., MALIEPAARD, C. A., MCQUEEN-MASON, S. J., VISSER, R. G. & TRINDADE, L. M. 2017b. Genetic complexity of miscanthus cell wall composition and biomass quality for biofuels. *BMC genomics,* 18**,** 406.

VAN DER WEIJDE, T., KIESEL, A., IQBAL, Y., MUYLLE, H., DOLSTRA, O., VISSER, R. G., LEWANDOWSKI, I. & TRINDADE, L. M. 2017c. Evaluation of Miscanthus sinensis biomass quality as feedstock for conversion into different bioenergy products. *Gcb Bioenergy,* 9**,** 176-190.

VAN DER WERF, H. M., VAN DER VEEN, J. H., BOUMA, A. & TEN CATE, M. 1994. Quality of hemp (Cannabis sativa L.) stems as a raw material for paper. *Industrial crops and products,* 2**,** 219-227.

VAN ESBROECK, G., HUSSEY, M. & SANDERSON, M. 1997. Leaf appearance rate and final leaf number of switchgrass cultivars. *Crop Science,* 37**,** 864-870.

VAN ROOIJEN, R., AARTS, M. G. & HARBINSON, J. 2015. Natural genetic variation for acclimation of photosynthetic light use efficiency to growth irradiance in Arabidopsis. *Plant Physiology,* 167**,** 1412-1429.

VAN ROSSUM, B.-J., KRUIJER, W., VAN EEUWIJK, F., BOER, M., MALOSETTI, M., BUSTOS-KORTS, D., MILLET, E., PAULO, J., VEROUDEN, M. & WEHRENS, R. 2020. Package 'statgenGWAS'. *R package version 1.0. 7*.

VANDENBRINK, J. P., HILTEN, R. N., DAS, K., PATERSON, A. H. & FELTUS, F. A. 2012. Analysis of crystallinity index and hydrolysis rates in the bioenergy crop Sorghum bicolor. *Bioenergy Research,* 5**,** 387-397.

VANRADEN, P. M. 2008. Efficient methods to compute genomic predictions. *Journal of dairy science,* 91**,** 4414-4423.

VERHERTBRUGGEN, Y., YIN, L., OIKAWA, A. & SCHELLER, H. V. 2011. Mannan synthase activity in the CSLD family. *Plant signaling & behavior,* 6**,** 1620-1623.

VERMA, S. R. & DWIVEDI, U. 2014. Lignin genetic engineering for improvement of wood quality: Applications in paper and textile industries, fodder and bioenergy production. *South African Journal of Botany,* 91**,** 107-125.

VOGEL, J. 2008. Unique aspects of the grass cell wall. *Current opinion in plant biology,* 11**,** 301-307.

VOINICIUC, C., GÜNL, M., SCHMIDT, M. H.-W. & USADEL, B. 2015. Highly branched xylan made by IRREGULAR XYLEM14 and MUCILAGE-RELATED21 links mucilage to Arabidopsis seeds. *Plant physiology,* 169**,** 2481-2495.

VON COSSEL, M., LEWANDOWSKI, I., ELBERSEN, B., STARITSKY, I., VAN EUPEN, M., IQBAL, Y., MANTEL, S., SCORDIA, D., TESTA, G. & COSENTINO, S. L. 2019. Marginal agricultural land low-input systems for biomass production. *Energies,* 12**,** 3123.

WAGNER, A., DONALDSON, L. & RALPH, J. 2012. Lignification and lignin manipulations in conifers. *Advances in botanical research.* Elsevier.

WANG, H., AVCI, U., NAKASHIMA, J., HAHN, M. G., CHEN, F. & DIXON, R. A. 2010a. Mutation of WRKY transcription factors initiates pith secondary wall formation and increases stem biomass in dicotyledonous plants. *Proceedings of the National Academy of Sciences,* 107**,** 22338-22343.

WANG, L., GUO, K., LI, Y., TU, Y., HU, H., WANG, B., CUI, X. & PENG, L. 2010b. Expression profiling and integrative analysis of the CESA/CSL superfamily in rice. *BMC plant biology,* 10**,** 1-16.

WANG, M., YUAN, D., GAO, W., LI, Y., TAN, J. & ZHANG, X. 2013a. A comparative genome analysis of PME and PMEI families reveals the evolution of pectin metabolism in plant cell walls. *PloS one,* 8**,** e72082.

WANG, S., YIN, Y., MA, Q., TANG, X., HAO, D. & XU, Y. 2012a. Genome-scale identification of cell-wall related genes in Arabidopsis based on co-expression network analysis. *BMC plant biology,* 12**,** 138.

WANG, X., CNOPS, G., VANDERHAEGHEN, R., DE BLOCK, S., VAN MONTAGU, M. & VAN LIJSEBETTENS, M. 2001. AtCSLD3, a cellulose synthase-like gene important for root hair growth in Arabidopsis. *Plant Physiology,* 126**,** 575-586.

WANG, X., WANG, J., JIN, D., GUO, H., LEE, T.-H., LIU, T. & PATERSON, A. H. 2015. Genome alignment spanning major Poaceae lineages reveals heterogeneous evolutionary rates and alters inferred dates for key evolutionary events. *Molecular plant,* 8**,** 885-898.

WANG, Y.-H., ACHARYA, A., BURRELL, A. M., KLEIN, R. R., KLEIN, P. E. & HASENSTEIN, K. H. 2013b. Mapping and candidate genes associated with saccharification yield in sorghum. *Genome,* 56**,** 659-665.

WANG, Y., FAN, C., HU, H., LI, Y., SUN, D., WANG, Y. & PENG, L. 2016. Genetic modification of plant cell walls to enhance biomass yield and biofuel production in bioenergy crops. *Biotechnology advances,* 34**,** 997-1017.

WANG, Y. & LI, J. 2006. Genes controlling plant architecture. *Current Opinion in Biotechnology,* 17**,** 123-129.

WANG, Y., TANG, H., DEBARRY, J. D., TAN, X., LI, J., WANG, X., LEE, T.-H., JIN, H., MARLER, B. & GUO, H. 2012b. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic acids research,* 40**,** e49-e49.

WANG, Z., WANG, M., LIU, L. & MENG, F. 2013c. Physiological and proteomic responses of diploid and tetraploid black locust (Robinia pseudoacacia L.) subjected to salt stress. *International journal of molecular sciences,* 14**,** 20299-20325.

WARD JR, K. 1950. Crystallinity of cellulose and its significance for the fiber properties. *Textile Research Journal,* 20**,** 363-372.

WCED 1987. Report of the World Commission on Environment and Development: Our common future. Oxford.

WENG, J. K. & CHAPPLE, C. 2010. The origin and evolution of lignin biosynthesis. *New Phytologist,* 187**,** 273-285.

WENG, J. K., MO, H. & CHAPPLE, C. 2010. Over-expression of F5H in COMT-deficient Arabidopsis leads to enrichment of an unusual lignin and disruption of pollen wall formation. *The plant journal,* 64**,** 898-911.

WERLING, B. P., DICKSON, T. L., ISAACS, R., GAINES, H., GRATTON, C., GROSS, K. L., LIERE, H., MALMSTROM, C. M., MEEHAN, T. D., RUAN, L. L., ROBERTSON, B. A., ROBERTSON, G. P., SCHMIDT, T. M., SCHROTENBOER, A. C., TEAL, T. K., WILSON, J. K. & LANDIS, D. A. 2014. Perennial grasslands enhance biodiversity and multiple ecosystem services in bioenergy landscapes. *Proceedings of the National Academy of Sciences of the United States of America,* 111**,** 1652-1657.

WERLING, B. P., MEEHAN, T. D., ROBERTSON, B. A., GRATTON, C. & LANDIS, D. A. 2011. Biocontrol potential varies with changes in biofuel–crop plant communities and landscape perenniality. *Gcb Bioenergy,* 3**,** 347-359.

WHEELER, T. J. & EDDY, S. R. 2013. nhmmer: DNA homology search with profile HMMs. *Bioinformatics,* 29**,** 2487-2489.

WU, L., ZHANG, M., ZHANG, R., YU, H., WANG, H., LI, J., WANG, Y., HU, Z., WANG, Y. & LUO, Z. 2021. Down-regulation of OsMYB103L distinctively alters beta-1, 4-glucan polymerization and cellulose microfibers assembly for enhanced biomass enzymatic saccharification in rice. *Biotechnology for biofuels,* 14**,** 1-15.

WU, X. Y., KUAI, B. K., JIA, J. Z. & JING, H. C. 2012. Regulation of Leaf Senescence and Crop Genetic Improvement F. *Journal of integrative plant biology,* 54**,** 936-952.

WULLSCHLEGER, S. D. 1993. Biochemical limitations to carbon assimilation in C3 plants—a retrospective analysis of the A/Ci curves from 109 species. *Journal of Experimental Botany,* 44**,** 907-920.

XIA, X. 2013. What is comparative genomics? *Comparative Genomics.* Springer.

XIE, G. & PENG, L. 2011. Genetic engineering of energy crops: a strategy for biofuel production in china free access. *Journal of Integrative Plant Biology,* 53**,** 143-150.

XU, K., XU, X., FUKAO, T., CANLAS, P., MAGHIRANG-RODRIGUEZ, R., HEUER, S., ISMAIL, A. M., BAILEY-SERRES, J., RONALD, P. C. & MACKILL, D. J. 2006. Sub1A is an ethylene-response-factor-like gene that confers submergence tolerance to rice. *Nature,* 442**,** 705.

XU, Y., ZHANG, X., LI, H., ZHENG, H., ZHANG, J., OLSEN, M. S., VARSHNEY, R. K., PRASANNA, B. M. & QIAN, Q. 2022. Smart breeding driven by big data, artificial intelligence and integrated genomic-enviromic prediction. *Molecular Plant*.

XU, Z., PU, X., GAO, R., DEMURTAS, O. C., FLECK, S. J., RICHTER, M., HE, C., JI, A., SUN, W. & KONG, J. 2020. Tandem gene duplications drive divergent evolution of caffeine and crocin biosynthetic pathways in plants. *BMC biology,* 18**,** 1-14.

XU, Z., ZHANG, D., HU, J., ZHOU, X., YE, X., REICHEL, K. L., STEWART, N. R., SYRENNE, R. D., YANG, X. & GAO, P. Comparative genome analysis of lignin biosynthesis gene families across the plant kingdom. BMC bioinformatics, 2009. BioMed Central, S3.

YADAV, S. K. 2010. Cold stress tolerance mechanisms in plants. A review. *Agronomy for sustainable development,* 30**,** 515-527.

YADAV, U. P., AYRE, B. G. & BUSH, D. R. 2015. Transgenic approaches to altering carbon and nitrogen partitioning in whole plants: assessing the potential to improve crop yields and nutritional quality. *Frontiers in Plant Science,* 6**,** 275.

YAMAGUCHI, M., MITSUDA, N., OHTANI, M., OHME-TAKAGI, M., KATO, K. & DEMURA, T. 2011. VASCULAR-RELATED NAC-DOMAIN 7 directly regulates the expression of a broad range of genes for xylem vessel formation. *The Plant Journal,* 66**,** 579-590.

YAN, J., HU, Z., PU, Y., BRUMMER, E. C. & RAGAUSKAS, A. J. 2010. Chemical compositions of four switchgrass populations. *biomass and bioenergy,* 34**,** 48-53.

YANG, L., ZHAO, X., YANG, F., FAN, D., JIANG, Y. & LUO, K. 2016a. PtrWRKY19, a novel WRKY transcription factor, contributes to the regulation of pith secondary wall formation in Populus trichocarpa. *Scientific reports,* 6**,** 1-12.

YANG, W., FENG, H., ZHANG, X., ZHANG, J., DOONAN, J. H., BATCHELOR, W. D., XIONG, L. & YAN, J. 2020. Crop phenomics and high-throughput phenotyping: past decades, current challenges, and future perspectives. *Molecular Plant,* 13**,** 187-214.

YANG, X., WANG, X., JI, L., YI, Z., FU, C., RAN, J., HU, R. & ZHOU, G. 2015. Overexpression of a Miscanthus lutarioriparius NAC gene MlNAC5 confers enhanced drought and cold tolerance in Arabidopsis. *Plant cell reports,* 34**,** 943-958.

YANG, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution,* 24**,** 1586-1591.

YANG, Z., ALGESHEIMER, R. & TESSONE, C. J. 2016b. A comparative analysis of community detection algorithms on artificial networks. *Scientific reports,* 6**,** 30750.

YIN, H., CHEN, C. J., YANG, J., WESTON, D. J., CHEN, J.-G., MUCHERO, W., YE, N., TSCHAPLINSKI, T. J., WULLSCHLEGER, S. D. & CHENG, Z.-M. 2014. Functional genomics of drought tolerance in bioenergy crops. *Critical Reviews in Plant Sciences,* 33**,** 205-224.

YIN, Y., HUANG, J. & XU, Y. 2009. The cellulose synthase superfamily in fully sequenced plants and algae. *BMC plant biology,* 9**,** 1-14.

YOGENDRA, K. N., SARKAR, K., KAGE, U. & KUSHALAPPA, A. C. 2017. Potato NAC43 and MYB8 mediated transcriptional regulation of secondary cell wall biosynthesis to contain Phytophthora infestans infection. *Plant Molecular Biology Reporter,* 35**,** 519-533.

YOKOYAMA, R. 2020. A genomic perspective on the evolutionary diversity of the plant cell wall. *Plants,* 9**,** 1195.

YOKOYAMA, R. & NISHITANI, K. 2004. Genomic basis for cell-wall diversity in plants. A comparative approach to gene families in rice and Arabidopsis. *Plant and Cell Physiology,* 45**,** 1111-1121.

YONG, W., LINK, B., O'MALLEY, R., TEWARI, J., HUNTER, C. T., LU, C.-A., LI, X., BLEECKER, A. B., KOCH, K. E. & MCCANN, M. C. 2005. Genomics of plant cell wall biogenesis. *Planta,* 221**,** 747-751.

YORK, W. S. & O'NEILL, M. A. 2008. Biochemical control of xylan biosynthesis—which end is up? *Current opinion in plant biology,* 11**,** 258-265.

YOSHIDA, K., SAKAMOTO, S., KAWAI, T., KOBAYASHI, Y., SATO, K., ICHINOSE, Y., YAOI, K., AKIYOSHI-ENDO, M., SATO, H. & TAKAMIZO, T. 2013. Engineering the Oryza sativa cell wall with rice NAC transcription factors regulating secondary wall formation. *Frontiers in plant science,* 4**,** 383.

YU, Y., HU, R., WANG, H., CAO, Y., HE, G., FU, C. & ZHOU, G. 2013. MlWRKY12, a novel Miscanthus transcription factor, participates in pith secondary cell wall formation and promotes flowering. *Plant science,* 212**,** 1-9.

YUAN, Y., TENG, Q., ZHONG, R. & YE, Z.-H. 2013. The Arabidopsis DUF231 domain-containing protein ESK1 mediates 2-O-and 3-O-acetylation of xylosyl residues in xylan. *Plant and Cell Physiology,* 54**,** 1186-1199.

ZABOTINA, O. A. 2012. Xyloglucan and its biosynthesis. *Frontiers in plant science,* 3**,** 134.

ZEGADA-LIZARAZU, W., ELBERSEN, H. W., COSENTINO, S. L., ZATTA, A., ALEXOPOULOU, E. & MONTI, A. 2010. Agronomic aspects of future energy crops in Europe. *Biofuels Bioproducts & Biorefining-Biofpr,* 4**,** 674-691.

ZERI, M., HUSSAIN, M. Z., ANDERSON-TEIXEIRA, K. J., DELUCIA, E. & BERNACCHI, C. J. 2013. Water use efficiency of perennial and annual bioenergy crops in central Illinois. *Journal of Geophysical Research: Biogeosciences,* 118**,** 581-589.

ZHANG, B. & ZHOU, Y. 2011. Rice brittleness mutants: a way to open the 'Black Box'of monocot cell wall biosynthesis free access. *Journal of integrative plant biology,* 53**,** 136-142.

ZHANG, D., ZHANG, Z. & YANG, K. 2006. QTL analysis of growth and wood chemical content traits in an interspecific backcross family of white poplar (Populus tomentosa× P. bolleana)× P. tomentosa. *Canadian Journal of Forest Research,* 36**,** 2015-2023.

ZHANG, J., XIE, M., TUSKAN, G. A., MUCHERO, W. & CHEN, J.-G. 2018. Recent advances in the transcriptional regulation of secondary cell wall biosynthesis in the woody plants. *Frontiers in plant science,* 9.

ZHANG, L., WU, S., CHANG, X., WANG, X., ZHAO, Y., XIA, Y., TRIGIANO, R. N., JIAO, Y. & CHEN, F. 2020. The ancient wave of polyploidization events in flowering plants and their facilitated adaptation to environmental stress. *Plant, cell & environment,* 43**,** 2847-2856.

ZHANG, Z. S., Y-B; LUO, X-F 2010. Effects of Water Stress on Biomass and Photosynthetic Characteristics of Tetraploid Black Locust (Robinia pseudoacacia L.) Clones. *For Res,* 23.

ZHAO, H., LI, Q., HE, J., YU, J., YANG, J., LIU, C. & PENG, J. 2014. Genotypic variation of cell wall composition and its conversion efficiency in Miscanthus sinensis, a potential biomass feedstock crop in China. *Gcb Bioenergy,* 6**,** 768-776.

ZHAO, T., HOLMER, R., DE BRUIJN, S., ANGENENT, G. C., VAN DEN BURG, H. A. & SCHRANZ, M. E. 2017. Phylogenomic synteny network analysis of MADS-box transcription factor genes reveals lineage-specific transpositions, ancient tandem duplications, and deep positional conservation. *The Plant Cell,* 29**,** 1278-1292.

ZHAO, T. & SCHRANZ, M. E. 2017. Network approaches for plant phylogenomic synteny analysis. *Current opinion in plant biology,* 36**,** 129-134.

ZHAO, T. & SCHRANZ, M. E. 2019. Network-based microsynteny analysis identifies major differences and genomic outliers in mammalian and angiosperm genomes. *Proceedings of the National Academy of Sciences,* 116**,** 2165-2174.

ZHAO, X., YANG, X., PEI, S., HE, G., WANG, X., TANG, Q., JIA, C., LU, Y., HU, R. & ZHOU, G. 2016. The Miscanthus NAC transcription factor MlNAC9 enhances abiotic stress tolerance in transgenic Arabidopsis. *Gene,* 586**,** 158-169.

ZHAO, X., ZHANG, L. & LIU, D. 2012. Biomass recalcitrance. Part I: the chemical compositions and physical structures affecting the enzymatic hydrolysis of lignocellulose. *Biofuels, Bioproducts and Biorefining,* 6**,** 465-482.

ZHONG, R., BURK, D. H., MORRISON III, W. H. & YE, Z.-H. 2004. FRAGILE FIBER3, an Arabidopsis gene encoding a type II inositol polyphosphate 5-phosphatase, is required for secondary wall synthesis and actin organization in fiber cells. *The Plant Cell,* 16**,** 3242-3259.

ZHONG, R., CUI, D. & YE, Z. H. 2019. Secondary cell wall biosynthesis. *New Phytologist,* 221**,** 1703-1723.

ZHONG, R., LEE, C. & YE, Z.-H. 2010. Evolutionary conservation of the transcriptional network regulating secondary cell wall biosynthesis. *Trends in plant science,* 15**,** 625-632.

ZHONG, R., LEE, C., ZHOU, J., MCCARTHY, R. L. & YE, Z.-H. 2008. A battery of transcription factors involved in the regulation of secondary cell wall biosynthesis in Arabidopsis. *The Plant Cell,* 20**,** 2763-2782.

ZHONG, R. & YE, Z.-H. 2015. Secondary cell walls: biosynthesis, patterned deposition and transcriptional regulation. *Plant and cell physiology,* 56**,** 195-214.

ZHOU, J., ZHONG, R. & YE, Z.-H. 2014. Arabidopsis NAC domain proteins, VND1 to VND5, are transcriptional regulators of secondary wall biosynthesis in vessels. *PloS one,* 9**,** e105726.

ZHU, C., GORE, M., BUCKLER, E. S. & YU, J. 2008. Status and prospects of association mapping in plants. *The plant genome,* 1.

ZHU, X.-G., CHANG, T.-G., SONG, Q.-F., FINNAN, J., BARTH, S., MÅRTENSSON, L.-M. & JONES, M. 2016. A Systems approach guiding future biomass crop development on marginal land. *Perennial Biomass Crops for a Resource-Constrained World.* Springer.

ZHU, X.-G., LONG, S. P. & ORT, D. R. 2010. Improving photosynthetic efficiency for greater yield. *Annual review of plant biology,* 61**,** 235-261.

ZOGHBI-RODRÍGUEZ, N. M., GAMBOA-TUZ, S. D., PEREIRA-SANTANA, A., RODRÍGUEZ-ZAPATA, L. C., SÁNCHEZ-TEYER, L. F. & ECHEVARRÍA-MACHADO, I. 2021. Phylogenomic and Microsynteny Analysis Provides Evidence of Genome Arrangements of High-Affinity Nitrate Transporter Gene Families of Plants. *International Journal of Molecular Sciences,* 22**,** 13036.

ZOMMERE, G., VIĻUMSONE, A., KALNIŅA, D., SOLIŽENKO, R. & STRAMKALE, V. 2013. Comparative analysis of fiber structure and cellulose contents in flax and hemp fibres. *Materials Science. Textile and Clothing Technology,* 8**,** 96-104.

ZUO, J., NIU, Q.-W., NISHIZAWA, N., WU, Y., KOST, B. & CHUA, N.-H. 2000. KORRIGAN, an Arabidopsis endo-1, 4-β-glucanase, localizes to the cell plate by polarized targeting and is essential for cytokinesis. *The Plant Cell,* 12**,** 1137-1152.

Summary

Plant cell walls are complex biological structures that surround every plant cell and are mostly composed of polysaccharides, lignin and structural proteins. Their molecular composition varies considerably among plants, making cell walls exceptional models to study the mechanisms underlying the evolution of complex plant traits. In parallel, nearly all cell wall components have valuable industrial applications. Hence, the development of tools to efficiently improve the molecular properties of cell walls toward target industrial utilizations is pivotal to establish a global bio-based economy. This is particularly true for a range of novel, under-domesticated biomass crops that could be grown on marginal lands to produce biomass without competing for land and crop components with food crops. Given these premises, this thesis used large-scale genomic analyses and bioinformatic tools to (i) study the genomic determinants of cell wall evolution and diversification in plants at unprecedented scale, and (ii) develop tools to efficiently improve cell wall quality, and potentially other complex plant traits, in under-domesticated crops.

The fundamental research on cell wall evolution starts in **Chapter 2**, with a phylogenomic analysis of the *CELLULOSE SYNTHASE* gene superfamily (*CesA* and *Csl* genes). These genes are models for cell wall evolution and largely determine the cellulose and hemicellulose amounts in cell walls. Their analysis in 242 plant genomes revealed that bryophytes and lycophytes only have *CesA* genes homolog of the angiosperm *CesA* isoforms of primary cell walls (i.e. cell walls deposited during cell extension, and lacking lignin; as distinct from secondary cell walls, which are deposited after cell growth ceases and contain large amounts of lignin). Moreover, *CesA* diversification into multiple primary- and secondary-specific cell wall isoforms was gradual, with ferns being the first organisms with secondary *CesA* genes. *CesA* diversification was accompanied by a diversification of *CesA* genomic contexts, which favoured gene sub-functionalization, as shown by a set of cotton secondary *CesA* displaying perturbation of genomic contexts and exceptionally active during cotton fiber deposition. Moreover, parallelisms between diversification of genomic gene contexts and gene sub-functionalization was found also in the *CslD* genes.

The role of genomic dynamics in cell wall diversification was further studied in **Chapter 3**, where multiple genomic properties were analysed for 150 cell wall gene families over 169 genomes. A specific focus was warranted on investigating the relationship between the diversification of Poaceae (type II) and eudicot (type I) cell walls, and the genomic basis of this differentiation. Remarkably, cell wall genes displayed a marked differentiation of copy number, synteny, phylogenetic dynamics,

and occurrence of tandem clusters between Poaceae and eudicots. This differentiation was particularly marked for pathways showing different functionality in type I and type II cell walls. For example, the genes within the *BEL1-like HOMEODOMAIN 6* regulatory pathway induce and repress secondary cell wall synthesis in Poaceae and eudicots, respectively, and extensively showed differential synteny, copy number, and selection pressures between Poaceae and eudicots. Similar patterns were found for several other genes involved in hemicellulose and lignin metabolism, highlighting the relevant role of genomics dynamics at the basis of cell wall diversification.

Beyond their fundamental importance, the genomic properties of cell wall genes can be used to develop tools to breed novel, under-domesticated, biomass crops for biomass production on marginal lands (MALs). This aspect was conceptualized in **Chapter 4**. First, by pointing out the opportunities for restoring the environmental and economic value of MALs through the cultivation of mixtures of locally-adapted perennial lignocellulosic crops. Second, by analysing how biomass production systems on MALs avoid competition with food crops. Third, by highlighting that the economic viability of this prospect largely depends on biomass yield and quality, mostly meant as cell wall composition tailored to industrial uses. Since most crops suitable for MALs are under-domesticated, tools for quick and cost-effective breeding of a complex trait as cell wall quality in under-domesticated crops that are not object of wealthy breeding projects are needed. The translation of genetic information from model biomass species to target crops offers an opportunity in this sense.

The latter objective was experimentally attained in **Chapter 5**, where the syntenic conservation of >600 quantitative trait loci (QTLs) controlling cell wall quality in several well-studied biomass species was studied over 151 plant genomes. QTL synteny was analysed with a network approach, and a pipeline to perform a double decomposition of the synteny network into groups of correlated syntenic gene communities based on initial QTL profiling was developed. Such groups of correlated communities are syntenic QTLs (SQTLs); genomic regions conserved across multiple species and spanning (part of) initial QTL(s) in at least one species. SQTLs were effectively able to project important cell wall genes retained in initial biomass QTLs from model species to other crops. Moreover, they allowed for filtering of specific homologous isoforms based on exact positional gene conservation across species (i.e. synteny), maximizing the likelihood of their functional conservation. Finally, the SQTL detection pipeline is suited for high-throughput analyses, and can be applied to other quantitative plant traits than cell wall quality.

The research on cell wall SQTLs was continued in **Chapter 6**, were their genomic variability was assessed by comparing multiple genome assemblies from different intra-specific accessions of six diverse angiosperm species. This research revealed

extensive intra-specific allelic variability of SQTL regions, which was often causative of changes of protein sequences and/or structures between accessions. These changes were observed for important cell wall genes, as (homologs of) the *BRITTLE CULM-like 8* rice locus. Moreover, SQTLs were also found to significantly co-localize with sets of independent QTLs mapped in miscanthus and switchgrass. Altogether, these results highlight the significance of SQTLs for the breeding practice: they can be used to predict relevant trait loci in target crops without the necessity of constructing mapping populations. Moreover, they can point out allelic diversity at loci whose involvement in the genetics of target traits was demonstrated in other species.

Overall, the research of this thesis produced novel, relevant results on the evolution of plant cell walls, pointing out the importance of genomic reshuffling and evolution of (novel) genomic contexts across species and/or genes to drive cell wall diversification. Moreover, hypothesis on the early evolution of *CesA* genes and cell wall were also formulated in light of the results achieved and literature knowledge. Next to cell wall evolutionary genomics, the genomic properties of cell wall genes were used to develop tools – SQTLs – which can shorten pre-breeding cycles in under-domesticated crops, especially if coupled with recent technologies for e.g. targeted (re-)sequencing. These tools are a valuable asset to breed novel biomass crops for marginal lands, but they can be applied to other traits, too. As current agricultural challenges like climate change impose the quick adaptation of current plant varieties or the development of novel crops by taking advantage of the wealth of under-studied or under-domesticated plant species, the work of this thesis is a precious step to make this prospect more feasible.

Riassunto

Le pareti cellulari vegetali sono complesse strutture biologiche che circondano ogni cellula in una pianta e sono composte principalmente da polisaccaridi, lignina e proteine strutturali. La loro composizione molecolare varia considerevolmente all'interno del regno vegetale, il che rende le pareti cellulari eccezionali modelli per studiare i meccanismi alla base dell'evoluzione di caratteri complessi nelle piante. Inoltre, quasi tutti i componenti molecolari delle pareti cellulari possono essere utilizzati in applicazioni industriali. Pertanto, lo sviluppo di strategie e tecnologie per cambiare la composizione delle pareti cellulari verso specifiche applicazioni industriali è fondamentale per favorire lo sviluppo di una bio-economia. Quest'aspetto è particolarmente importante per una serie di nuove colture da biomassa che non sono ancora domesticate ma potrebbero essere coltivate in aree marginali, per produrre biomassa per catene di valore *green* senza competere con colture alimentari. Alla luce di queste premesse, questa tesi ha utilizzato estensive analisi genomiche e bioinformatiche per (i) studiare la base genomica dell'evoluzione e diversificazione delle pareti cellulari vegetali su vasta scala e (ii) sviluppare strumenti per assistere il miglioramente genetico della composizione molecolare delle pareti cellulari e, potenzialmente, di altri caratteri quantitativi delle piante, specialmente in nuove colture da biomassa funzionali ad una bio-economia.

La ricerca sull'evoluzione delle pareti cellulari vegetali inizia nel **Capitolo 2**, con un'analisi filogenomica della super-famiglia genica *CELLULOSE SYNTHASE* (geni *CesA* e *Csl*). Questi geni sono modelli per lo studio evolutivo delle pareti cellulari e sono alla base della sintesi di cellulosa ed emicellulosa nelle pareti cellulari. La loro analisi in 242 genomi di altrettante specie vegetali ha rivelato che i geni *CesA* delle specie briofite e licofite – che sono tra le piante terrestri piu' antiche in scala evolutiva – sono tutti omologhi dei geni *CesA* delle specie angiosperme operanti nelle pareti cellulari primarie (depositate durante l'estensione cellulare e prive di lignina, a differenza delle pareti cellulari secondarie, che vengono depositate dopo che la crescita cellulare cessa e contengono grandi quantità di lignina). Inoltre, la diversificazione dei geni *CesA* in gruppi operanti specificatamente nelle pareti cellulari primarie o secondarie è stata graduale, con la comparsa dei primi geni *CesA* specifici per le pareti cellulari secondarie nelle piante tracheofite (felci). La diversificazione funzionale dei geni *CesA* è andata in parallelo a una diversificazione dei contesti genomici dei geni stessi, come dimostrato da un insieme di geni *CesA* operanti nelle pareti cellulari secondarie delle piante di cotone che mostrano cambiamenti del loro contesto genomico rispetto ad altre piante angiosperme e che mostrano un'eccezionale attività durante la sintesi delle fibre di cotone. Inoltre, parallelismi tra la diversificazione dei contesti genomici dei geni e la loro diversificazione funzionale sono stati osservati anche nei geni *CslD*.

Il ruolo delle dinamiche genomiche nella diversificazione delle pareti cellulari vegetali è stato ulteriormente studiato nel **Capitolo 3**, dove sono state analizzate diverse proprietà genomiche di 150 famiglie geniche coinvolte nella sintesi delle pareti cellulari in 169 genomi di altrettante specie vegetali. La ricerca e' stata incentrata nello studiare la relazione tra la diversificazione delle pareti cellulari delle piante Poaceae (tipo II) ed eudicotiledoni (tipo I), cosi' come la base genomica di questa differenziazione. La maggior parte dei geni analizzati ha mostrato una marcata differenziazione del numero di copie geniche, della sintenia, delle dinamiche filogenetiche e dell'occorrenza di geni organizzati in *tandem clusters* tra Poaceae ed eudicotiledoni. Questa differenziazione è risultata particolarmente marcata per i geni coinvolti in vie metaboliche che mostrano diversità funzionale nelle pareti cellulari di tipo I e di tipo II. Ad esempio, i geni appartenenti alla via metabolica dipendente dal gene *BEL1-like HOMEODOMAIN 6* rispettivamente inducono e reprimono la sintesi della parete cellulare secondaria nelle Poaceae e nelle eudicotiledoni. Questi geni hanno rivelato marcate differenze nella loro organizzazione sintenica, nel numero di copie e nelle pressioni selettive tra Poaceae ed eudicotiledoni. Risultati simili sono stati trovati per diversi altri geni coinvolti nel metabolismo dell'emicellulosa e della lignina, evidenziando come diverse dinamiche genomiche hanno probabilmente avuto un ruolo determinante nel forgiare la variabilita' qualitativa e quantitativa delle pareti cellulari che si può osservare tra diverse specie vegetali.

Oltre alla loro importanza dal punto di vista puramente scientifico, le proprietà genomiche dei geni alla base delle pareti cellulari vegetali possono essere utilizzate per sviluppare strumenti che coadiuvano il miglioramento genetico di nuove colture da biomassa adattate ad aree marginali (AM). Questo aspetto è stato concettualizzato nel **Capitolo 4**. In primo luogo, evidenziando le opportunità offerte dalla messa a coltura di specie da biomassa perenni adattate alle condizioni contingenti delle AM per ripristinare il valore ambientale ed economico delle AM stesse. In secondo luogo, mostrando come la letteratura scientifica dimostri che la coltivazione di colture da biomassa su AM evitano la competizione con le colture alimentari. Infine, sottolineando come la fattibilità economica di questa prospettiva dipende in gran parte dalla resa e dalla qualità della biomassa, principalmente intesa come composizione della parete cellulare adatta agli utilizzi industriali a valle della catena economica. Poiché la maggior parte delle piante adattate alle AM non sono attualmente domesticate, sono necessari strumenti tecnologici che rendano possibile il miglioramento genetico rapido ed economico di caratteri complessi – come la composizione molecolare delle pareti cellulari – in piante orfane (mai state oggetto di un sistematico processo di miglioramento genetico). In questa prospettiva, un'opportunità è rappresentata dalla trasposizione delle informazioni sull'organizazzione genetica di tratti complessi tra specie diverse, e in particolare tra specie modello di ricerca genetica e nuove colture con poco studiate.

Quest’ultimo aspetto è stato oggetto di ricerca nel **Capitolo 5**, dove è stato studiato il livello di conservazione sintenica in oltre 151 genomi vegetali di ~600 loci quantitativi (QTL) che controllano la composizione delle pareti cellulari vegetali in diverse specie da biomassa che sono modello di ricerca. La sintenia dei QTL (ovvero la conservazione del tipo e dell’ordine dei geni al loro interno tra genomi diversi) è stata analizzata con una metodologia di *network analysis*, in cui un iniziale network rappresentante le relazioni sinteniche tra i geni dei QTL nei 151 genomi analizzati è stato decomposto in gruppi di comunità geniche sinteniche correlate in base alla profilazione dei QTL iniziali contenuti in ogni comunità. Tali gruppi di comunità correlate formano “QTL sintenici” (SQTL); regioni genomiche conservate tra diverse specie vegetali e che co-localizzano con QTL noti in almeno una specie. L’analisi dei SQTL ottenuti ha mostrato come essi siano in grado di identificare importanti geni coinvolti nella sintesi delle pareti cellulari da specie modello ad altre colture, anche al livello di specifiche copie geniche isoforme. Infine, la pipeline per il mappaggio di SQTL si adatta ad applicazioni *high-throughput* e può essere potenzialmente applicata a qualsiasi tratto quantitativo complesso.

La ricerca sui SQTL alla base delle pareti cellulari è continuata nel **Capitolo 6**, dove la loro variabilità genomica intra-specifica è stata analizzata in sei diverse specie di piante angiosperme. Questa ricerca ha rivelato come i SQTL presentino un’estesa variabilità allelica intra-specifica, spesso causativa di cambiamenti delle sequenze e/o delle strutture proteiche tra diversi individui. Questi cambiamenti sono stati osservati per importanti geni collegati alle proprietà delle pareti cellulari, come omologhi del locus *BRITTLE CULM-like 8* del riso. Inoltre, i SQTL co-localizzano significativamente con insiemi di QTL indipendenti che sono stati mappati in Miscanto e Panìco. Nel complesso, questi risultati evidenziano l'importanza dei SQTL in un’ottica di miglioramento genetico: possono essere utilizzati per prevedere loci genomici associati a caratteri di interesse in nuove colture senza la necessità di costruire popolazioni ad-hoc per il mappaggio genetico. Inoltre, i SQTL possono evidenziare la diversità allelica in loci la cui importanza per caratteri di interesse è stata dimostrata in altre specie.

Complessivamente, la ricerca di questa tesi ha prodotto risultati innovativi e rilevanti sull’evoluzione delle pareti cellulari vegetali, evidenziando l’importanza di diverse dinamiche genomiche nella diversificazione delle pareti cellulari avvenuta durante l’evoluzione delle piante terrestri. Inoltre, sono state formulate ipotesi sugli specifici percorsi evolutivi dei geni *CesA* e delle pareti cellulari alla luce dei risultati ottenuti. Accanto alla genomica evolutiva delle pareti cellulari, le proprietà genomiche dei geni alla base della loro sintesi sono state utilizzate per sviluppare strumenti – SQTL – che possono accorciare i cicli di miglioramento genetico in nuove colture non domesticate, specialmente se abbinati a tecnologie come il (re-)sequencing di specifici segmenti genomici. Questi strumenti sono un prezioso *asset* per la coltivazione di nuove colture da biomassa in AM, ma possono essere applicati anche ad altre caratteri e piante. Poiché le attuali sfide che interessano il settore agricolo, come il cambiamento climatico, impongono l'adattamento rapido delle attuali varietà vegetali o lo sviluppo di nuove colture sfruttando la biodiversità e variabilità genetica proveniente da specie vegetali poco studiate o non domesticate, il lavoro di questa tesi rappresenta un prezioso passo avanti per rendere questa prospettiva più facilmente percorribile.

Acknowledgements

When thinking back to the past four years, my mind goes immediately to all the people that contributed in different ways to complete this thesis and to turn this period of my life into the amazing experience it has been. Here, I would like to thank all these people, as without their touch into my PhD trajectory, this journey would have certainly been much less educative and enjoyable.

In making acknowledgements, I have no doubts to start with my first promotor, **Luisa Trindade**. Luisa, you gave me the possibility to realize my dream since the day I came to Wageningen University in 2016 for my MSc studies: obtaining a PhD in plant genetics and breeding, possibly by working on a topic related to the bio-based economy. You excellently supervised me during my whole journey at Plant Breeding and the BBE group, and I really feel I could not find a better supervisor. You gave me freedom to be independent in my research, but you always kept a critical eye on my work to offer me right and precious scientific advise at the proper moments. You offered me complete support for my personal and professional growth, and you pushed me to attain my best whenever needed. Finally, you always found a spot in your busy agenda to quickly correct all my manuscripts and to find the room for precious and pleasant meetings. For all these reasons, I want to deeply thank you, and I am extremly happy that we can continue to work together in the coming years!

The second person I would like to thank is my second promotor, **Eric Schranz**. Eric, you became an official member of my PhD committee along the way, but nevertheless you represented a fundamental guiding light for the whole PhD journey. I feel it would have been extremely hard to properly shape the research of my project without our discussions. Moreover, your cheerful personality, combined with being an extraordinary scientist, represent a model for me on how to approach life and work in academia. I feel extremely grateful for having found you along my way, and I hope we can stay in touch for our works in the future!

The third person I would like to thank from my PhD committee is my co-promotor, **Robert van Loo**. Robert, you are a brilliant scientist, and your scientific knowledge and approach to research impressed me since the times of my MSc thesis. Every minute spent by discussing with you about my research was an intense learning moment. As such, I want to deeply thank you for having represented a pillar of knowledge during the past years as my daily supervisor. You have always been available to discuss research plans and results until the finest detail. Moreover, it has been really nice to share with you pleasant moments at project meetings and BBE activities. For all this, thank you so much!

Besides the official PhD committee, I want to acknowledge other people for their direct contribution to the content of the thesis. One of these is **Joao Paulo**, who I would like to thank for the important discussions we had about the miscanthus GWAS throughout the past years, and for having always been available for pleasant chats about my thesis. Moreover, I would like to acknowledge the roles of **Kasper van der Cruijsen** and **Mohamad Al Hassan**: you are both co-authors of Chapter 6, which could not be there without the hard work you did in phenotyping the miscanthus collection along the past years, and the discussions we made about GWAS data and results. In addition, you have been really great friends within the BBE team: thank you for all the nice time spent together! Furthermore, a thank you to **Oene Dolstra**, who several years ago designed and established the miscanthus collection used in Chapter 6, and whose approach to plant breeding and science is reason of admiration for myself. Finally, I would like to thank all the BSc and MSc students that picked a topic within my PhD project for their theses. **Hugo**, **Dennis**, **Steven**, **Dimitris**, and **Sylwia**: you all contributed significantly to my research and my personal growth, and I am grateful of having had the chance of being your thesis supervisor in the past year. To Hugo especially, I am proud that you managed to start a PhD in the BBE group as well, and I am happy that we spent nice time as colleagues and friends after you finalized your MSc. I am sure you will do a great job in your PhD research!

Besides the people that were somehow directly involved in the research of this thesis, many other people from the Plant Breeding department gave an essential touch to my PhD trajectory with research tips, help on general issues, nice conversations, or great friendships. In this regard, a first, big thank you goes to all the past and present people of the BBE team which I was lucky to encounter in my journey. Starting from the "old days", thank you **Andres Torres**, **Jordi Petit**, **Behzad Rashidi**, **Viviana Jaramillo**, and **Agata Gulisano.** I think I could not have been luckier than finding people like you when I first joined the BBE team for my MSc thesis. I can only smile when thinking back to all the time spent together, and I am extremely grateful for all the positive energy, happiness, and fun you put in my PhD journey. Moreover, many of you helped me in my research with nice discussions and critical feedback on my results, so thank you for this! My mind is full of nice memories also when thinking to all the other BBE members. In particular, I want to send a very big thank you to my fellow BBE PhDs **Kasper van der Cruijsen**, **Mites Kleuter**, **Hugo Rijken**, **Antonio Lippolis**, **Sanne Put**, **Mathijs Peters** and **Ornela Bocova**: you are all responsible for the amazing environment I always encountered in our group, and I am happy I could have shared a lot of fun moments and nice discussions with all of you. Furthermore, thank you for the several critical confrontations on our research topics, which have definitely contributed in some ways to the work of this thesis. I want to give a special acknowledgement to **Kasper** and **Mites**, my paranimphs and two great friends of mine: thank you for having accepted to accompany me during the last steps of the PhD as paranymphs, and for all the fun we had playing squash and drinking beers along the years! I also want to thank **Annemarie Dechesne**, **Dianka Dees**, and **Elma Salentijn**, from which I feel I learned a lot despite not being directly involved in my PhD project, and with which I also had nice chats along the years. Outside the BBE team, I want to thank **Jillis Grubben** and **Antonino Crucitti** for their positive contribution to my life in and outside the department and, for Antonino, to the delicious food I was lucky to taste here in the Netherlands, directly from Calabria! Moreover, I want to thank the colleagues with which I shared the office along the years: **Petra Coenradi**, **Karin Burger** and, later, **Giorgio Tumino**. You are all nice people, and your pleasant company enriched the time spent at the department with a lot of enjoyable moments. I would also like to thank the secretaries of our department, **Nicole Trefflich**, **Danielle van der Wee**, and **Letty Dijker**, for all the precious and friendly help they gave me on basically any type of burocratic issues I encountered. Finally, I would like to thank other fellow PhDs from plant breeding which I was lucky to meet and spend some nice time with along the years: **Jorge**, **Marcella**, **Alejandro**, **Corentin**, **Daniel**, **Eleni**, **William**, **Miguel**, **Lampros.** Especially to Jorge: thank you for the many squash matches played together, also with Mites, Kasper, Hugo, **Romanos**, and before also **Jarst**! And to the whole squash team: a special thank for the non-competitive attitude we always had with each other during squash!

Outside university, a special thank you goes to **Laura Terzi.** From the first days of the MSc in plant breeding, until now that we all still live in the Netherlands, thank you immensely for the unconditional friendship we built, full of countless fun moments along the years! Our dinners, coffees, shopping and travels have filled the past six years and a half of indelible memories. I am looking forward to have some rest after the PhD in the Salentu this summer. A similar story goes for **Mirko Zucchini.** Our friendship is even longer, since the walls of the kindergarten of our village in Italy, and I am extremely happy that we were able to keep it strong and healthy until now. All our conversations and reciprocal visits during the last years certainly helped me to go on straigth and with a smile along my PhD journey, and also to grow as a person. We spent a lot of nice time together when you came to the Netherlands for your Erasmus! Thank you very much for all this! To **Beatrice Baldi** and **Sebastiano Zanini**: I also want to share a big thank you with you! From the exams of the bachelor, to the memorable dinners with Seba during the MSc and beginning of PhD, to all our Skype calls and sharing of stress, hopes, and fun of our lives as (PhD) students, your unconditional friendships have been a certainty to count on in my life during the last years. To **Nicoletta Bertolin**, thank you as well for the nice friendship we built here in the Netherlands and all the great moments spent together during the last years. I look back to all those with big smiles, and you also certainly helped me to take my PhD journey more relaxed. Finally, a big thank you to also all the other friends that made the life of the last years much funnier: **Brian**, **Jacopo**, **Veronica**, **Chiara**, **Valentina**, **Daniele**, **Alessandro**, **Martina** and **Tommaso**.

Per concludere, vorrei rivolgere un ringraziamento speciale a una serie di persone più intime della mia vita, che hanno avuto un ruolo primario durante il mio PhD. Le prime sono i miei genitori. In particolare mia **mamma**, che mi ha sempre sostenuto incondizionatamente nei miei studi e nella vita, e senza cui non potrei aver concretamente realizzato i miei piani e sogni, incluso questo dottorato. Ma anche mio **papà**, che pure mancando da ormai più di dieci anni è estremamente presente in tutto il percorso che mi ha portato fino al completamento di questa tesi, così come nella mia vita di tutti i giorni. Grazie veramente di cuore a entrambi. Poi il resto della mia famiglia – allargata – dai miei carissimi **zii e zie**, ai **genitori e parenti di Carolina**, che hanno tutti riempito di bellissimi e indelebili ricordi tutti gli ultimi anni, e mi e ci hanno sempre fatto forza nella mia e nostra vita qui in Olanda. Infine, **Carolina**. È impossibile riuscire a condensare in poche righe tutta la gratitudine e amore che provo per averti avuto al mio fianco negli ultimi 10 anni e mezzo, specialmente durante il PhD. Non solo mi hai sostenuto e compreso nei momenti difficili e in ogni scelta fatta, ma sei stata anche fonte di ispirazione per me stesso, grazie ai tuoi talenti personali e professionali. In più, hai avuto la forza e costanza di prendere iniziativa e sostenerci per fare tanti grandi passi nella nostra vita insieme. Per tutto questo, sono certo nel dire che gli ultimi anni non sarebbero potuti essere quelli che sono stati senza te al mio fianco, incluso questa tesi. Grazie veramente dal profondo del mio cuore, e che il nostro amore continui e accresca nel tempo, insieme a noi!

Arnhem, April 2023

About the author

Francesco Pancaldi was born on the 11th June 1994 in Bentivoglio, a small town in the province of Bologna, in Italy. In 2013 he started a Bachelor programme in Agricultural Sciences at the University of Bologna, from which he graduated *cum laude* in 2016. During his Bachelor, he completed a thesis in the group of Prof. Andrea Monti, during which he found deep interest for the concept of bio-based economy and the opportunities for agriculture within it. Moreover, he discovered a passion for genetic and breeding during the genetic course led by Prof. Silvio Salvi and Prof. Roberto Tuberosa. For these reasons, he decided to move to Wageningen in 2016 to start a Master programme in Plant Sciences, with specialization Plant Breeding and Genetic Resources. His MSc studies were completed *cum laude* in 2018. For his MSc thesis, he decided to work on a project entailing the genetic and biochemical characterization of EMS mutant lines of the industrial oilseed crop *Crambe abyssinica*, within the Bio-Based Economy Group at the Laboratory of Plant Breeding, led by Prof. Dr. Luisa Trindade, and supervised by Dr. Robert van Loo. After the completion of the MSc thesis, he had the opportunity to start working on the EU project MAGIC for a minor thesis, aiming at studying the genomic conservation of a set of miscanthus QTLs across diverse grass species. This project was then turned into the core topic of his PhD, which started in 2018 and whose results are described in this thesis. During the second year of PhD, he was granted participation to the School of Politics hold by the former Italian Prime Minister, Enrico Letta, which was a great opportunity to work on another great passion of him, politics and contigent socio-economic issues of our times. As of March, 2023, Francesco will continue to work at the Laboratory of Plant Breeding of Wageningen UR as post-doctoral researcher.

List of publications

Pancaldi, F. and Trindade, L.M., 2020. Marginal lands to grow novel bio-based crops: a plant breeding perspective. *Frontiers in Plant Science*, *11*, 227.

Pancaldi, F., van Loo, E.N., Schranz, M.E. and Trindade, L.M., 2022. Genomic Architecture and Evolution of the *Cellulose synthase* Gene Superfamily as Revealed by Phylogenomic Analysis. *Frontiers in Plant Science*, *13*.

Pancaldi, F., Vlegels, D., Rijken, H., van Loo, E.N. and Trindade, L.M., 2022. Detection and analysis of syntenic quantitative trait loci controlling cell wall quality in angiosperms. *Frontiers in Plant Science*, *13*.

Pancaldi, F., van Loo, E.N., Senio, S., Al-Hassan, M., van der Cruijsen, K., Paulo, M.J., Dolstra, O., Schranz, M.E. and Trindade, L.M., 2023. Syntenic cell wall QTLs as versatile breeding tools: intraspecific allelic variability and predictability of biomass quality loci in target plant species. *Plants*, *12(4)*, 779.

Pancaldi, F., Schranz, M.E., van Loo, E.N. and Trindade, L.M., 2023. Highly differentiated genomic properties underpin the different cell walls of Poaceae and eudicots. *Plant Physiology*, accepted.

Education Statement of the Graduate School Experimental Plant Sciences

Issued to: **Francesco Pancaldi**
Date: **12 June 2023**
Group: **Laboratory of Plant Breeding - Biobased Economy Group**
University: **Wageningen University & Research**

1) Start-Up Phase	*date*
► **First presentation of your project**	
Plant cell walls: from evolutionary genomics to novel breeding tools', Biobased Economy Group meeting	2 Apr 2019
► **Writing or rewriting a project proposal**	
Plant cell walls: from evolutionary genomics to novel breeding tools	8 Feb 2019
► **Writing a review or book chapter**	
Pancaldi, F., Trindade, L. M. Marginal lands to grow novel bio-based crops: a plant breeding perspective. Frontiers in Plant Science (2020). doi.org/10.3389/fpls.2020.00227	3 Mar 2020
► **MSc courses**	
Subtotal Start-Up Phase	13.5 credits*

2) Scientific Exposure	*date*
► **EPS PhD student days**	
EPS PhD days 'Get2Gether', Soest, NL	11-12 Feb 2019
EPS PhD days 'Get2Gether', Soest, NL	10-11 Feb 2020
EPS PhD days 'Get2Gether', online	1-2 Feb 2021
► **EPS theme symposia**	
EPS Theme 3 symposium 'Metabolism and Adaptation', Nijmegen, NL	21 Oct 2019
EPS Theme 3 symposium 'Metabolism and Adaptation', online	30 Oct 2020
EPS Theme 3 symposium 'Metabolism and Adaptation', online	5 Nov 2021
EPS Theme 4 symposium 'Genome Biology', online	11 Dec 2020
EPS Theme 4 symposium 'Genome Biology', online	17 Jan 2022
► **Lunteren Days and other national platforms**	
Annual meeting 'Experimental Plant Sciences', Lunteren, NL	8-9 Apr 2019
Annual meeting 'Experimental Plant Sciences', online	12-13 Apr 2021
Annual meeting 'Experimental Plant Sciences', Lunteren, NL	11-12 Apr 2022
► **Seminars (series), workshops and symposia**	
Seminar: Dr. Daniel Bopp, Genetic and evolution of sex determination system in the common housefly	25 Oct 2018
Seminar: Dr. Diana Santelia, Rewiring starch metabolism for plant environmental adaptation	1 Nov 2018
Seminar: Dr. Andrew Simkin, Metabolic engineering to enhance photosynthesis and increase crop yield	21 Mar 2019
Seminar: Dr. José Miguel Mulet, Yeast as a tool to improve plant tolerance against drought stress	17 May 2019
Seminar: Dr. Maheshi Dassanavake, Multi-Ion salt stress adaptation explored using extremophyte genomics	20 May 2019
Seminar: Dr. Ralph Panstruga, Phenotypic and molecular characterization of partially mlo-virulent isolates of the barley powdery mildew pathogen (Blumeria graminis f.sp. hordei)	23 May 2019
Seminar: Dr. Jaime Prohens, Introgression breeding from wild species for crops adaptation to climate change	23 May 2019
Seminar: Dr. Monika Doblin, Designing walls for a sustainable future	5 Jul 2019
Seminar: Dr. Roeland Voorrips, SNP haplotyping in polyploids - a genetic Sudoku	1 Oct 2019
Seminar: Dr. Jacob Weiner, Applying evolutionary theory to improve plant production	25 Feb 2020
Seminar: Molecules Webinar: Recent Advances in Carbohydrate-Active Enzymes	23 Jun 2021
Workshop: 'Breeding for diversity - opportunities and challenges', Wageningen, NL	30 Oct 2019
Workshop: Snakemake workshop (Dr. Johannes Köster), online	13 May 2020
Symposium: 'Seed systems for the future', Wageningen, NL	26 Jan 2023
► **Seminar plus**	
► **International symposia and congresses**	
Conference: 'CRISPRcon - Conversations on Science, Society, and the future of Gene Editing', Wageningen, NL	20-21 Jun 2019
Conference: 'MAGIC Consortium Meeting 2019', Catania, IT	17-18 Jul 2019

Conference: 'MAGIC Consortium Meeting 2021', online	30 Jun 2021
Conference: '32nd Annual Meeting of the Association for the Advancement of Industrial Crops (AAIC)', online	5-8 Sep 2021
Conference: 'Final MAGIC Consortium Meeting', Lisbon, PT	1-3 Dec 2021
Symposium: 'Plant breeding: perspectives from academia and industry', online	24 Apr 2020
Congress: 'Expo Miscanthus 2022', Hoofddorp, NL	28 Sep 2022
► **Presentations**	
Talk: 'Plant cell walls - from evolutionary genomics to novel breeding tools', MAGIC Consortium Meeting 2019, Catania, IT	17 Jul 2019
Talk: 'Phylogenomic analysis of the CesA gene superfamily: novel insights into cell wall evolution', Plant Breeding Monday Seminars, Wageningen, NL	1 Mar 2021
Talk: 'Development of genetic tools for rapid improvement of orphan biomass crops for marginal lands', 32nd Annual Meeting AAIC, online	6 Sep 2021
Talk: 'Growing novel biomass crops on marginal lands - challenges for Plant Breeding', 32nd Annual Meeting AAIC, online	8 Sep 2021
Talk: 'Syntenic QTLs for biomass quality: novel tools to speed up breeding in orphan biomass crops', Final MAGIC Consortium Meeting, Lisbon, PT	1 Dec 2021
Talk: 'A highly differentiated genomic landscape underlies the different cell walls of Poaceae and dicots', Plant Breeding Monday Seminars, Wageningen, NL	3 Oct 2022
Poster: 'Phylogenomic analysis of the CesA superfamily', Annual Meeting 'Experimental Plant Science' 2021, online	12-13 Apr 2021
► **IAB interview**	
► **Excursions**	
EPS company visit: Averis Seeds	7 Jun 2019
Subtotal Scientific Exposure	17.7 credits*

3) In-Depth Studies	*date*
► **Advanced scientific courses & workshops**	
PE&RC/SENSE course: 'Bayesian Statistics', Wageningen, NL	15-16 Oct 2018
SLU/WUR course: 'Plant Breeding and Biotechnology', Wageningen, NL	11-13 Jun 2019
VIB course: 'Bulk RNA-seq: from counts to differential expression', online	23 Jun 2020
EPS/VIB course: 'Gentle hands-on introduction to Python programming', online	2-3 Jul 2020
VIB course: 'Python for downstream data analysis', online	21-22 Sep 2020
PE&RC/WIMEK course: 'R and Big data', Wageningen, NL	14-15 Oct 2021
EPS/VIB course: 'PLAZA - Functional plant bioinformatics', online	25-26 Oct 2021
PE&RC course: 'Generalized Linear Models', Wageningen, NL	15-17 Jun 2022
WIAS/PE&RC course: 'Statistics for Data Science', Wageningen, NL	6-14 Dec 2022
► **Journal club**	
Member of Plant Breeding PhD Literature Discussion Club	2018-2022
► **Individual research training**	
Subtotal In-Depth Studies	8.0 credits*

4) Personal Development	*date*
► **General skill training courses**	
WGS course: 'Brain Training', Wageningen, NL	17 Apr 2019
EPS course: 'EPS Introduction Course', Wageningen, NL	29 Oct 2019
WGS course: 'Reviewing a Scientific Paper', Wageningen, NL	19 Nov 2019
WGS course: 'Project and Time Management', online	15 Sep - 27 Oct 2020
WGS course: 'Supervising BSc & MSc thesis students', Wageningen, NL	24-25 Sep 2020
WGS course: 'Adobe InDesign Essential Training', online	9-10 Nov 2020
WGS course: 'Scientific Writing', online	15 Feb - 6 Apr 2021
► **Organisation of meetings, PhD courses or outreach activities**	
► **Membership of EPS PhD Council**	
Subtotal Personal Development	5.2 credits*

TOTAL NUMBER OF CREDIT POINTS*	**44.4**

Herewith the Graduate School declares that the PhD candidate has complied with the educational requirements set by the Educational Committee of EPS with a minimum total of 30 ECTS credits.

** A credit represents a normative study load of 28 hours of study.*

The research described in this thesis was funded by the European Union’s Horizon 2020 research and innovation programme, under grant agreement 727698 for the MAGIC (Marginal Lands for Growing Industrial Crops) project.

Financial support from Wageningen University for printing this thesis is gratefully acknowledged.

Cover concept and design: Francesco Pancaldi

Printed by: ProefschriftMaken.nl