Running title: Poales genomes revealed grass metabolic innovations

Title:

Genomes of Poaceae sisters reveal key metabolic innovations preceding the emergence of grasses

Authors:
Yuri Takeda-Kimura[1,2], Bethany Moore[1,a], Samuel Holden[3], Sontosh K. Deb[4], Matt Barrett[5,c], David Lorence[6], Marcos V. V. de Oliveira[1], Jane Grimwood[7], Melissa Williams[7], Lori Beth Boston[7], Jerry Jenkins[7], Christopher Plott[7], Shengqiang Shu[8], Kerrie Barry[8], David M. Goodstein[8], Jeremy Schmutz[7,8], Matthew J. Moscou[3,9], Michael R. McKain[4], James H. Leebens-Mack[10], Hiroshi A. Maeda[1]

[1]Department of Botany, University of Wisconsin-Madison; Madison, WI, 53706, USA
[2]Faculty of Agriculture, Yamagata University; Tsuruoka, Yamagata, 997-0037, Japan
[3]The Sainsbury Laboratory, University of Cambridge; Norwich Research Park, Norwich, NR4 7UH, United Kingdom
[4]Department of Biological Sciences, University of Alabama, Tuscaloosa, AL, 35487, USA
[5]Australian Tropical Herbarium, James Cook University, Douglas, Townsville, QLD 4811, Australia
[6]National Tropical Botanical Garden, Kalaheo, Hawaii, 96741, USA
[7]Genome Sequencing Center, HudsonAlpha Institute for Biotechnology, Huntsville, Alabama 35806, USA
[8]US Department of Energy Joint Genome Institute, Berkeley, California 94720, USA
[9]USDA-ARS, Cereal Disease Laboratory, St. Paul, MN, 55108, USA
[10]Department of Plant Biology, University of Georgia, Athens, GA, USA

Current addresses:
[a]Morgridge Institute for Research, Madison, WI, 53715, USA
[b]Faculty of Land and Food Systems, The University of British Columbia, Vancouver, BC, Canada
[c]Australian Tropical Herbarium, Sir Robert Norman Building, James Cook University Cairns Campus, McGregor Road, Smithfield, Qld 4878, Australia

Corresponding authors:
James H. Leebens-Mack (jleebensmack@uga.edu), Hiroshi A. Maeda (maeda2@wisc.edu)

SUMMARY

The grass family (Poaceae, Poales) holds immense economic and ecological significance, exhibiting unique metabolic traits, including dual starch and lignin biosynthetic pathways. To investigate when and how the metabolic innovations known in grasses evolved, we sequenced the genomes of a non-core grass, *Pharus latifolius*, the non-grass graminids, *Joinvillea ascendens* and *Ecdeiocolea monostachya*, representing the sister clade to Poaceae, and *Typha latifolia*, representing the sister clade to the remaining Poales. The *rho* whole genome duplication (ρWGD) in the ancestral lineage for all grasses contributed to the gene family expansions underlying cytosolic starch biosynthesis, whereas an earlier tandem duplication of *phenylalanine ammonia lyase* (*PAL*) gave rise to *phenylalanine/tyrosine ammonia lyase* (*PTAL*) responsible for the dual lignin biosynthesis. Two mutations were sufficient to expand ancestral PAL function into PTAL. The integrated genomic and biochemical analyses of grass relatives in Poales revealed the evolutionary and molecular basis of key metabolic innovations of grasses.

KEYWORDS

metabolic innovations, gene duplication, genome multiplication, neofunctionalization, lignin, phenylalanine tyrosine ammonia lyase, Poales, Poaceae, Ecdeiocoleaceae, Joinvilleaceae

Joinvillea ascendens (Photo by David H. Lorence)

INTRODUCTION

The grass family, Poaceae, emerged around 100 million years ago (Mya) within the monocot order Poales[1]. Poaceae contains 780 genera and more than 11,000 species (**Figure 1A**)[1–3], acting as foundational primary producers for diverse ecosystems across the globe. Economically, grasses are the most important crops and include cereals, such as rice, wheat, and maize, which have starch-rich endosperms and provide more than 50% of calories worldwide[4]. Grasses also contain critical industrial crops, such as sugarcane and sorghum, which grow rapidly to produce large biomass for bioenergy production. Wheat and barley were the first plants to be domesticated in the Fertile Crescent ~12,000 years ago[5], maize and rice were domesticated ~9,000 years ago in what is now Mexico and the lower Yangtze Valley, respectively[6,7]. The evolutionary history and genetic basis of grass domestications have been extensively studied[8–11], revealing a "domestication syndrome" including selection for reduced pre-harvest dispersal (non-shattering) and increased grain size[12].

Grasses exhibit a series of unique morphological, physiological, and metabolic traits. For example, grasses possess the unique inflorescence, the ear, which is a domestication target for loss of shattering[13–15]. Grasses also exhibit striking evolutionary innovations in primary metabolism resulting in unique chemical composition and properties that contribute to their adaptation in the environment and agriculture. While starch is typically produced in the plastids, grasses can synthesize starch in both the plastids and the cytosol, leading to the starch-rich endosperm of grass seeds, such as grains and kernels[16]. Additionally, grasses have two entry pathways to synthesize lignin (**Figure 1B**), the major phenolic polymer of secondary cell walls, which represents up to 30% of grass dry mass[17] and is essential for mechanical strength, water transpiration, and physical barrier[18]. Lignin provides renewable aromatic feedstocks for high-value materials and fuels[19–21], many derived from bioenergy grasses (e.g., switchgrass, sorghum). Lignin is typically synthesized from the aromatic amino acid L-phenylalanine (Phe) via the committed enzyme, Phe ammonia-lyase (PAL, **Figure 1B**). Notably, grasses uniquely possess an additional entry pathway to produce lignin from L-tyrosine (Tyr) with Tyr ammonia-lyase (TAL) activity mediated by the bifunctional Phe/Tyr ammonia-lyase (PTAL) enzyme[22,23]. The PTAL pathway, found in a broader range of grass species[24–28], contributes to nearly half of lignin biosynthesis in grasses[22,23], potentially facilitating the rapid growth of many grasses while depositing substantial amounts of lignin in their scattered vascular bundles (**Figure 1B**). However, when and how these unique metabolic enzymes and pathways evolved is unknown.

Recent advancement in genome sequencing has permitted tracing whole genome duplications (WGD), gene family evolution, and gene innovations[29–34]. Several key features in the evolutionary history of monocot genomes include a set of three rearrangements in the plastid genome[35,36], numerous polyploidy events[37–39] such as the *rho* WGD (ρWGD) that occurred on the stem leading to the Poaceae (**Figure 1A**)[3,33,39]. The ρWGD facilitated the evolution of complex traits and new species[40,41] and has also been linked to proliferation of MADS-box genes and the morphological flexibility and success of the Poaceae[14]. The ρWGD may have also produced gene duplicates for metabolic processes such as cytosolic production of starch in the endosperm of grasses[42].

The majority of sequenced genomes in the Poales are found in the Poaceae, especially cereal crops due to their agricultural significance[43–46]. The genome of a non-core grass, *Streptochaeta angustifolia* Soderstr. (subfamily Anomochlooideae) was recently sequenced and revealed potential contribution of ρWGD and subsequent gene losses to the pseudo-spikelet morphology of *S. angustifolia*[15]. Grasses have many derived functional traits that are hypothesized to have contributed to their evolutionary success and ecological dominance[47,48], but aside from some

aspects of inflorescence architecture[49,50] and grain yield[51], the molecular basis and evolution of these features are poorly understood. The sister clade to Poaceae includes two families, Ecdeiocoleaceae (N=3) and Joinvilleaceae (N=4)[3,52]. Their reference-quality genome assemblies are currently unavailable but would enhance the utility of comparative genomics and facilitate identification of genome and molecular features associated with grass-specific functional traits (**Figure 1A**).

To study the evolutionary history of grasses and their unique metabolic traits, here we generated high quality genomes for a non-core Poaceae grass species, *Pharus latifolius* L., two non-grass graminids, *Joinvillea ascendens* Gaudich. ex Brongn. & Gris and *Ecdeiocolea monostachya* F.Muell., representing the sister clade to the grasses, and the cattail species *Typha latifolia* L. representing the sister lineage to all other families in the order Poales[3,52,53]. The Poaceae and its sister clade diverged approximately 100 Mya, prior to ρWGD, and Typhaceae and the ancestor of the remainder of Poales diverged approximately 120 Mya (**Figure 1A**)[53,54]. Our integrated genomic and biochemical analyses of grass relatives in the order Poales unveiled evolutionary history and molecular basis of key metabolic innovations of grasses. The study highlights the critical roles of both whole-genome and tandem gene duplications underlying these evolutionary innovations and provides a gene editing target that can introduce the dual lignin entry pathway in non-grass plants. The Poales genomes obtained here offer valuable resources for dissecting diverse and complex grass traits uniquely evolved in this critical plant family.

RESULTS

Genomes of *Joinvillea ascendens*, *Ecdeiocolea monostachya*, *Pharus latifolius*, and *Typha latifolia* provide critical resources to study grass evolution

Comparative genomic analyses including species within the grass family (Poaceae) and its sister clade including Joinvilleaceae and Ecdeiocoleaceae (**Figure 1A**)[3,52,53] are crucial for understanding the evolution of grass functional traits. *Ecdeiocolea monostachya* is a wild perennial native to western Australia and one of three species within the Ecdeiocoleaceae (**Figure 1A**). Sampling was performed on an accession of *E. monostachya* from Western Australia and, using flow cytometry, estimated between 38 to 42 chromosomes and haploid genome size of 980 Mb (4C of ~4.0 pg), consistent with an outcrossing species with an estimated chromosome number (2N) of 38[55]. The genome assembly was performed using HiFiAsm using PacBio HiFi long-reads. Gene models were predicted using *ab initio*, homology, and evidence-based gene models using RNAseq from sheath, root and flower tissue. The final haploid reduced assembly of a heterozygous, diploid *E. monostachya* isolate EM_001 encompasses 897 Mb over 1,114 scaffolds with N50 of 1.7 Mb and a total of 27,801 genes, with 97.9% of the Embryophyte BUSCO genes (**Table 1**).

Additionally, we sequenced genomes of *Pharus latifolius*, which belongs to the Poaceae subfamily Pharoideae, a small lineage diverging after the Anomochloideae and before the Puelioideae and BOP-PACMAD lineages, *Joinvillea ascendens,* together with *E. monostachya* representing the sister clade to Poaceae, and *Typha latifolia*, representing the sister lineage to the remainder of Poales (**Figure 1A**). These three genomes were assembled using HiFiAsm, which integrates PacBio HiFi and Hi-C data, and polished using RACON[56]. Scaffold orientation, ordering, and joining were based on Hi-C contact mapping (See Methods). Evidence-based gene annotations

were performed using Illumina RNAseq and PacBio IsoSeq reads. The *P. latifolius* genome was sequenced to a depth of 44x coverage yielding a chromosome level assembly spanning 1.12 Gb over 198 scaffolds. Gene models were inferred for 32,144 protein-coding genes, including 98.4% of the Embryophyte BUSCO genes. As compared to the *P. latifolius* genome released by Ma et al.[57], the use of PacBio HiFi reads generated a reduced number of contigs (222 vs. 535) and slightly larger final genome size (1.12 Gb vs. 1.00 Gb).

Joinvillea ascendens is native to the Hawaiian Islands and Oceania and one of four species within Joinvilleaceae. The genome of *J. ascendens* was sequenced to 125x coverage yielding a scaffold assembly with 18 chromosomes with a genome size of 1.21 Gb with N50 of 63.7 Mb and 29,122 protein coding genes (**Table 1**), with 98% of genes found within the Embryophyte BUSCO gene list. The *T. latifolia* genome was sequenced to depths of 74.70x coverage resulting in a chromosomal assembly including 215 Mb spanning the expected 15 chromosomes and one additional 54 Kb scaffold (**Table 1**). A total of 22,107 protein-coding gene models were characterized in the *T. latifolia* genome including 99.7% of the Embryophyte BUSCO genes.

A comparison of the genome structure across Poales was conducted using the GENESPACE toolkit[58], which combines sequence similarity with gene order to robustly estimate orthology. Synteny patterns were consistent with the previously estimated timing of the ρWGD after divergence of Poaceae and its sister clade (**Figure 1A**). At the same time, a riverine plot reveals rearrangements of chromosomal blocks over the 120 My history of Poales (**Figure 1C**).

To understand gene family divergence and conservation across Poales, we used 39 species in our phylogenomic analyses with complete genomes representing the basal angiosperms (ANA grade), Magnoliids, eudicots, and monocots with taxa focused in Poales (**Table S1**). After filtering for a minimum of 20 taxa, we used 10,255 orthogroups for phylogenetic reconstruction. The Phylogenetic Placement of Polyploidy Using Genomes (PUG[3,52,53]) analysis estimated a total of 2,493,258 putative paralog pairs from the 10,255 orthogroup trees across all taxa. A total of 330,476 unique nodes across gene trees were evaluated using these inferred paralog pairs (**Table S1)**. In all, 24,517 putative paralog pairs provided support for duplication events that fit the PUG's search criteria of the species tree. After filtering for nodes in the gene trees with a bootstrap support value (BSV) of 80 or better, 21,105 unique duplications were retained for whole genome multiplication (WGM) assessment (**Figure 1D**). We evaluated the relationship between gene duplications on specific branches and previously characterized WGM events. The highest number of gene duplications of 3,200 was found on the branch leading to the last common ancestor of wheat (*Triticum aestivum*) and barley (*Hordeum vulgare*). The *rho* event—previously placed prior to the diversification of grasses[3,15,59]—was supported with 3,004 gene duplications placed on the branch leading to the last common ancestor of all grasses (**Figure 1D; Table S2)**. We also found 1,122, 972, and 2,196 gene duplications associated with the *sigma* WGM event[60], the *tau* WGM event[3,61], and the maize allopolyploid event[60,62], respectively (**Table S2**). Additionally, 868 and 936 unique gene duplications were mapped to before the divergence of the BOP-PACMAD clades and the PACMAD clade, respectively (**Table S2**), likely representing either examples of homoeologous exchange between *rho* duplications[63] or possibly other types of duplication. Finally, we see evidence of the *gamma* event (1,212 duplications) prior to the divergence of *Vitis* from other Pentapetalae[64,65], as well as 803 unique duplications that map to the branch prior to the divergence of *Aquilegia* from other eudicots (**Table S2**). Together the genomes of *E. monostachya*, *J. ascendens*, *P. latifolius,* and *T. latifolia* establish the precise timing of the *rho* WGM event and provide critical resources to study the evolution of grasses and Poales.

Duplication of genes involved in starch and fatty acid biosynthesis during grass evolution

Starch is the major non-structural carbohydrate in plants and stores carbon and energy for maintaining metabolism and growth[66,67]. Grasses have a high starch reserve, especially in their endosperm, where starch biosynthesis takes place in the cytosol[16,68], besides plastids where starch is stored and typically synthesized[69] (**Figure 2A**). The cytosolic starch biosynthesis can utilize a relatively abundant cytosolic pool of ATPs in non-photosynthetic plastids, without importing ATPs into the plastids for ADP-glucose pyrophosphorylase (AGPase)-catalyzed reaction[16]. The starch-rich grass seeds support rapid seedling establishment[47] and also provide the major source of calories consumed by humans (e.g., cereal grains and kernels). However, when and how this unique metabolic trait of grasses evolved is not fully understood.

To investigate the origin of the cytosolic starch biosynthesis, we constructed phylogenetic trees of three genes required for the cytosolic starch biosynthesis—AGPase large and small subunits (LSU and SSU, respectively) and ADP glucose transporter (**Figure 2A**). Grasses have four types of AGPase LSUs with the duplication of Type 3 AGPase LSUs that likely generated Type 2 with cytosolic AGPase LSUs[16]. Consistent with this notion, our phylogeny showed two distinct, well-supported Type 3 and Type 2 AGPase LSU clades both having *S. angustifolia* and *P. latifolius* sequences, while *J. ascendens* and *E. monostachya* sequences were in their outgroup (**Figure 2B; Figure S1A**), suggesting that the duplication event occurred within grasses. Similarly, AGPase SSUs duplicated within grasses, giving rise to Type 1 AGPase SSUs (**Figure 2C; Figure S1B**), which is dual localized but reside in the cytosol in the endosperm, whereas Type 2 AGPase SSUs are targeted to the plastids in leaves[70].

ADP-glucose generated by the cytosolic AGPase is then transported to the plastids by the plastidial ADP glucose transporter (**Figure 2A**)[16]. We detected gene duplication within the common ancestor of all grasses (**Figure 2D**), giving rise to the plastidial ADP glucose transporter within the Type 2 Plastidial Adenine Nucleotide Transporters (PANTs), which belong to the Mitochondrial Carrier protein Family (MCF)[71]. The long branch leading to the grass-specific plastidial ADP glucose transporter suggests potential positive selection for neofunctionalization. Notably, *J. ascendens* and all grasses, including *S. angustifolia* and *P. latifolius*, had both Type 1 and Type 2 PANTs (**Figure 2D**), suggesting that there was an earlier duplication event at a common ancestor of grasses and non-grass graminids, giving rise to the Type 2 PANTs.

Grasses also have a unique biochemical feature for fatty acid biosynthesis. Unlike other plants having heteromeric prokaryotic-type and homomeric eukaryotic-type acetyl-CoA carboxylase (ACCase) in the cytosol and plastids, respectively, grasses have homomeric eukaryotic-type ACCase in both cytosol and plastids[72,73]. This makes grasses sensitive to certain herbicides that inhibit homomeric but not heteromeric ACCase[72,74]. Homomeric ACCase in both compartments may also reduce the efficiency of fatty acid synthesis in contrast to efficient starch synthesis of grasses in two compartments (**Figure 2A**). Interestingly, the homomeric eukaryotic-type ACCase duplicated within non-grass graminids, giving rise to the plastidic homomeric eukaryotic-type ACCase in grasses as well as *J. ascendens* and *E. monostachya* (**Figure 2E**). These results together revealed the evolutionary history of key metabolic genes underlying the unique metabolic traits of grasses, some of which have already appeared *before* the emergence of grasses and ρWGD.

PTAL evolved within a common ancestor of grasses and non-grass graminid Joinvillea, before ρWGD and the appearance of grasses.

Utilizing these novel genomic resources, we further investigated the evolutionary history of the

dual lignin entry pathways, uniquely found in grasses (**Figure 1B**)[22–24]. We first conducted synteny analysis of the genomic region having the *PAL*/*PTAL* gene(s) using multiple genomes of representative grasses, non-core grasses, non-grass graminids and Poales. The *J. ascendens* genome had two copies of *PAL/PTAL* homologs (Joasc.05G060400.1, Joasc.05G060500.1), which were located in tandem within the same syntenic block at the chromosome 5 (**Figure 3A**). The corresponding syntenic block in *A. comosus* also had one of four *PAL/PTAL* homologs. Within grasses, the *PAL/PTAL* syntenic block was further duplicated, likely due to the ρWGD, and *PAL* genes in one of the two duplicated blocks were expanded in core grasses (**Figure 3A**). One *PAL/PTAL* copy within these syntenic blocks in grasses encodes previously characterized or predicted PTAL[23], except for *B. distachyon* having only one *PTAL*, likely due to a loss of one *PTAL* duplicate. Based on *Ks* distributions from the synteny analysis (**Figure S2**), the average *Ks* value of these *PAL/PTAL* syntenic blocks were always higher than the *Ks* value of the largest peak with the lowest *Ks* value, which in grasses represents the ρWGD[3]. The results support that the tandemly duplicated *PAL/PTAL* homologs in *J. ascendens* emerged before ρWGD and hence the divergence of non-grass graminids and Poaceae. The subsequent ρWGD further contributed to the expansion of *PAL* and *PTAL* genes within grasses.

To further examine the functionality of *PAL/PTAL* homologs, Maximum Likelihood (ML) gene trees were constructed for monocot *PALs* and *PTALs* using amino acid and coding DNA sequences. *PAL* homologs from *J. ascendens*, and *E. monostachya* formed a grade leading to a clade including all *PTAL* orthologs sampled from *J. ascendens*, *E. monostachya* and all grasses, whereas all grass *PAL* homologs formed a clade comprised of distinct ρWGD homeolog clades (**Figure 3B; Figure S3, S4**). Mapping the gene tree to the species tree using TreeSolve[75], however, resulted in 25,000 optimized trees forming a consensus tree with 100% support for one of *J. ascendens* and *E. monostachya* genes (Joasc.05G060500.1 and Emoptg000374l_1G000600.1) to be part of the PAL clade and the other genes (Joasc.05G060400.1, Emoptg000374l_1G000630.1) to be part of the PTAL clade (**Figure S5A**). Duplication-Transfer-Loss (DTL) reconciliation from TreeSolve showed that the optimally resolved gene tree has a lower reconciliation cost than the original gene tree (389 vs. 448) and hence is a more probable tree given the species tree. Notably, non-grass graminids, *J. ascendens* and *E. monostachya,* and non-core grasses, *S. angustifolia* and *P. latifolius*, had PTAL candidates that formed a grade leading to the PTAL clade of core grasses. Together with the synteny data (**Figure 3A**), these results suggest that monocot PAL enzymes diverged at a common ancestor of the non-grass graminids and subsequently PTAL enzymes emerged through neofunctionalization of a PAL enzyme (**Figure 3B; Figure S3**).

The His140 residue, located at the substrate binding pocket, was previously shown to be critical for the Tyr substrate recognition in the bacterial TAL enzyme[76]. We noted that the corresponding His140 residues were present in all grass PTAL enzymes, including both of the *S. angustifolia* enzymes (strangu_020769−RA and strangu_019386−RA), as well as one of each enzyme from *J. ascendens* (Joasc.05G060400.1) and *E. monostachya* (Emoptg000374l_1G000630.1), suggesting that these proteins are bifunctional PTAL enzymes. To test this hypothesis experimentally, we cloned, expressed, and purified the recombinant enzymes of PAL/PTAL orthologs from *S. angustifolia*, *J. ascendens*, and *E. monostachya* as well as PALs and PTALs from *Sorghum bicolor* (SbPAL and SbPTAL) and *Brachypodium distachyon* (BdPAL and BdPTAL) as positive controls[23,77]. These purified enzymes were first mixed with the substrate Phe or Tyr at 1 mM and the production of cinnamic acid (CA) or *p*-coumaric acid (*p*CA) were analyzed by HPLC for detection of PAL and TAL activity, respectively. All enzymes efficiently converted Phe into CA (**Figure 3C**), which was well above those of boiled enzyme or

no substrate controls (**Figure S6**). Efficient production of *p*CA from Tyr was observed in the reaction mixtures containing SbPTAL, BdPTAL, strangu_020769−RA, strangu_019386−RA, Emoptg000374l_1G000630.1, and Joasc.05G060400.1 (**Figure 3C**). Joasc.05G060500.1, Emoptg000374l_1G000600.1, BdPAL, and SbPAL had detectable TAL activity as compared to controls but much lower (50-fold) than PTALs above (**Figure 3C**). These results showed that the PAL/PTAL orthologs with His140 are bifunctional PTAL enzymes having TAL activity, in addition to PAL activity. Therefore, we tentatively named both of *S. angustiifolia* enzymes and one of *E. monostachya*, and *J. ascendens* enzymes with His140 as SaPTAL-a, SaPTAL-b, EmoPTAL, and JaPTAL, and another one of *E. monostachya*, and *J. ascendens* enzymes with Phe140 as EmoPAL and JaPAL (**Figure 3B**).

To further quantify the TAL and PAL activity of these PTAL (SbPTAL, BdPTAL, SaPTAL-a, SaPTAL-b, EmoPTAL, and JaPTAL) and PAL (JaPAL, EmoPAL, BdPAL, and SbPAL) enzymes, we determined their enzyme kinetics parameters using variable concentrations of Tyr and Phe substrates, respectively (**Figure 3D**; **Figure S6**; **Table S3**). The apparent K_m of PTAL enzymes for Tyr were 11 to 19 μM, whereas those of PAL enzymes were much higher (3449-6211 μM) and hence inefficient in utilizing Tyr. The k_{cat} values of the TAL activity were approximately twice higher for PTALs (0.04 to 0.09 s^{-1}) than for PALs (0.02 to 0.04 s^{-1}, **Figure 3E**; **Table S2**). Consequently, the catalytic efficiency (k_{cat}/K_m) of TAL activity for PTAL enzymes (2.55 to 7.96 $s^{-1}\mu M^{-1}$) were much higher (485-fold in average) than those for PAL enzymes (0.01 $s^{-1}\mu M^{-1}$) (**Figure 3E; Table S3**). These quantitative data further support that *S. angustiifolia*, *E. monostachya*, and *J. ascendens* have at least one enzyme having strong TAL activity. We also found that the PAL activity (k_{cat}/K_m) of JaPTAL (6.8 $s^{-1}\mu M^{-1}$) was much lower than that of JaPAL (78.8 $s^{-1}\mu M^{-1}$) with significant lower k_{cat} (0.5 s^{-1} vs. 1.9 s^{-1}) and higher K_m (66 μM vs. 24 μM) values (**Figure 3E; Table S3**). Moreover, grass PTAL enzymes showed higher K_m toward Phe (150-227 μM) than graminid PTALs (64-66 μM) (**Table S3**), leading to ~3.1-fold higher TAL/PAL activity ratio for PTALs from grasses than non-grass graminids (**Figure S7A**). These experimental data demonstrates that the bifunctional PTAL enzymes emerged within the common ancestor of grasses and non-grass graminids (e.g., *J. ascendens*), and later gained higher TAL/PAL ratio in grasses.

Ile112 and His140 are critical for acquisition of the TAL activity in graminid PTALs.

To experimentally test the role of His140 for the acquisition of TAL activity, site-directed mutagenesis was conducted on the PAL and PTAL enzymes of grasses and non-grass graminids characterized above. For the PAL enzymes, the residue corresponding to Phe140 was converted to His to generate $JaPAL^{F140H}$, $EmoPAL^{F134H}$, $BdPAL^{F137H}$, and $SbPAL^{F135H}$ mutant enzymes. Compared to their corresponding wild-type enzymes, all of these PAL mutants had increased overall TAL activity (9.7-fold in k_{cat}/K_m average) with significant reduction of K_m values towards Tyr (**Table S3**). The reciprocal mutants of PTAL enzymes—$SbPTAL^{H123F}$, $BdPTAL^{H123F}$, $SaPTAL\text{-}a^{H118F}$, $SaPTAL\text{-}b^{H126F}$, $EmoPTAL^{H127F}$, and $JaPTAL^{H125F}$— also exhibited decreased TAL activity (0.01-fold k_{cat}/K_m in average) with significant increase in K_m towards Tyr (**Figure S7B; Table S3**). These results support the role of His140 as a critical residue for the recognition of Tyr substrate in PTAL enzymes, consistent with prior studies[76–78].

The introduction of His140 itself was, however, not *sufficient* to convert PALs into PTALs. The catalytic efficiency of TAL activity (k_{cat}/K_m) of PAL^{F140H} mutants was still much lower (~10% in average) than that of wild-type PTALs, because of the much higher K_m values (222-450 μM) than wild-type PTALs (11-19 μM) (**Figure S7B; Table S3**). Furthermore, $PTAL^{H140F}$ mutants still

showed higher TAL activity than that of wild-type PALs (e.g., K_m value of 531-765 vs. 3448-6211 μM). Thus, unlike the bacterial TAL enzyme[70], other residues besides the His140 are required for the acquisition of strong TAL activity in PTALs of grasses and closely-related non-grass graminids.

To identify the residues critical for the transition of PAL to PTAL in this plant lineage, we first conducted a positive selection analysis using *dN/dS* (nonsynonymous to synonymous substitution rate) using PAML[79,80]. We compared the genes in the PTAL vs. PAL clade of the graminids and found a total of 30 positively selected sites between the two clades with a probability >0.7 (**Table S4**). We next conducted a phylogeny-guided sequence comparison[81] by utilizing the phylogenetic distribution of the functional PAL and PTAL enzymes (**Figure 3**). In the amino acid alignment of monocot PALs and PTALs (**Figure 4A, Figure S8**), we identified 16 residues, besides His140, that are highly conserved in PTAL enzymes, all of which were found to be under positive selection (**Table S4**). Eight residues (magenta in **Figure 4A**) were highly conserved within PAL and PTAL groups but distinct between these two groups; the additional 8 residues (purple in **Figure 4A**) were highly conserved only among PTALs and variable in different PALs (**Figure 4A; Table S4**). To investigate the position of these residues in their protein structures, a homology model of JaPAL was generated using the well-characterized parsley PAL structure as a template (PDB:6F6T[82]). Most of these 16 residues were found to be located near the active site, except for a few peripheral residues highlighted in purple (**Figure 4B**).

To investigate the potential role of these residues in TAL activity, two JaPAL mutant enzymes were generated: one with the 8 magenta residues and the other with both magenta and purple (total 16) residues being simultaneously substituted from PAL- to PTAL-type (**Table S4**), in addition to the F140H mutation ($JaPAL^{F140H_MUT8}$ and $JaPAL^{F140H_MUT16}$, respectively). The enzyme kinetics analyses showed that the apparent K_m value of $JaPAL^{F140H_MUT8}$ (17.9 μM) became significantly improved than that of the $JaPAL^{F140H}$ single mutant (222.7 μM) and approached closely to that of wild-type JaPTAL (10.9 μM) with the similar k_{cat} (**Figure 4C; Table S3**). The PAL activity of $JaPAL^{F140H_MUT8}$ showed 2-fold higher K_m for Phe than wild-type JaPTAL with comparable k_{cat} values (**Figure 4C**). The $JaPAL^{F140H_MUT16}$ mutant also showed significantly improved K_m (42.2 μM) for TAL activity than $JaPAL^{F140H}$ (and JaPAL wild-type) but to a lesser extent than $JaPAL^{F140H_MUT8}$ (17.9 μM, **Figure 4C**). Thus, these results demonstrate that some of these 8 magenta residues, besides His140, are involved in the efficient TAL activity of non-grass graminid PTALs.

To determine which of the eight magenta residues are essential in the conversion of PAL to PTAL enzyme, each one of these residues on $JaPAL^{F140H_MUT8}$ was mutated back to the PAL type, one by one, and their effects on their catalytic efficiency were determined. The substitution of seven out of eight residues had no to minor impacts on the overall TAL and PAL activity (**Figure 4D**). In contrast, when the I112S substitution–one of the positively selected residues in PTALs–was introduced to $JaPAL^{F140H_MUT8}$ to make $JaPAL^{F140H_MUT8_I112S}$, both TAL and PAL activities were significantly decreased because of the drastic increase of K_m and slight reduction of k_{cat} (**Figure 4D; Table 3**). Based on the homology model structures of JaPAL and JaPTAL proteins generated with parsley PAL and sorghum PTAL protein models, respectively, as a template (**Figure 4E**), the Ser/Ile112 is not in direct contact with the substrate but is located next to the Tyr113/98 (for PAL/PTAL), a critical proton acceptor for the catalysis[77,83]. Therefore, the Ile112 residue of the PTAL enzyme near the active site appears to be crucial for the TAL activity.

F140H and S112I substitutions are sufficient to change plant PALs into bifunctional PTALs.

To further test the potential role of the Ile112 residue in gaining TAL activity, the reciprocal S112I

mutation was introduced to the JaPAL^{F140H} single mutant to generate the JaPAL^{F140H_S112I} double mutant. For the comparison, the single residue corresponding to Ser112 was converted to Ile to generate the JaPAL^{S112I} single mutant as well. While k_{cat} was not drastically affected by these mutations, K_m of JaPAL^{F140H_S112I} for Tyr (17.5 μM) became significantly lower than those of the wild-type JaPAL (4859 μM) and the single mutants JaPAL^{F140H} and JaPAL^{S112I} (223 μM and 354 μM, respectively) and reached to the level of wild-type JaPTAL (11.0 μM) (**Figure 4F**). Therefore, Ile112 is essential for gaining the efficient TAL activity, and the simultaneous introduction of the Ile112 and His140 residues can alter monofunctional PAL into bifunctional PTAL.

To test if the two amino acid substitutions equivalent to F140H and S112I can confer TAL activity in distantly-related PALs, we introduced these mutations to the recombinant protein of Arabidopsis PAL1 having high PAL activity with very weak TAL activity[76,84] (**Table S3**). The AtPAL1^{F144H_S116I} mutant showed dramatic reduction in its K_m towards Tyr (20.2 μM) as compared to wild-type AtPAL1 (3070 μM) and its single mutants, AtPAL1^{F144H} and AtPAL1^{S116I} (314 μM and 515 μM, respectively). Overall, the kinetic behavior of the AtPAL1^{F144H_S116I} and JaPAL^{H140F_I112S} double mutants were similar (**Figure 4F, G, Table S3**). Thus, the introduction of two key residues Ile112 and His140, identified from the PTAL evolutionary analysis in Poales, can alter monofunctional PALs into bifunctional PTALs even in a distantly-related plant PAL enzyme.

DISCUSSION

Determining the genetic and molecular basis underlying the evolution of lineage-specific traits remains challenging especially at deep evolutionary nodes. Here we showed that increasing the density of sequenced genomes across phylogenetic taxa, especially at the key sister lineages (i.e., the novel genomes of Poales sisters to grasses, *J. ascendens* and *E. monostachya*), provides powerful means to trace evolutionary innovations of grasses. Our phylogenomic analysis suggests that some genes involved in the cytosolic starch biosynthetic pathway emerged through the ρWGD, although a preceding duplication into PANT Type 1 and 2 might have potentiated the evolution of ADP-glucose transport across the plastid membrane (**Figure 2E**). Strikingly, the dual lignin entry pathway catalyzed by bifunctional PTAL enzyme, which was believed to be specific to grasses[23], actually emerged before the evolution of grasses at the common ancestor of Joinvilleaceae, Ecdeiocoleaceae, and Poaceae. Our synteny analysis supports that the PTAL evolved through a tandem duplication event, rather than ρWGD event, followed by neofunctionalization (**Figure 3**). The tandemly duplicated PAL/PTAL block was further duplicated through ρWGD, leading to an additional PTAL copy in many grasses and further expansion of PALs through more recent tandem duplications (**Figure 3A**). Many plant families, including grasses, radiated after a WGD following a lull in diversification known as the lag-time hypothesis[65,85,86]. WGD-associate diversification has been attributed to sub/neofunctionalization of duplicates leading to morphological and functional innovations[86,87]. The PAL/PTAL tandem array detected in this study illustrates that tandemly duplicated gene families can neofunctionalize and expand by subsequent WGD.

Precise determination of the timing of the PTAL evolution, which predated grass evolution, also revealed the underlying molecular basis of this key metabolic innovation. A novel residue Ile112 responsible for the transition from PAL to PTAL is located at the 'inner mobile loop' (**Figure 4E**), which is identified as a crucial domain for the substrate binding and catalysis based on the X-ray structure analyses[83,88]. The structural modifications occurring within this inner mobile

loop are likely to influence the conformation of the substrate binding pocket, thereby leading to distinct catalytic activities between graminid PALs and PTALs. In microbes, altering the one residue, His/Phe140, is sufficient to switch between TAL and PAL activity[76,78]. This is likely because bacterial and fungal PALs and TALs possess Ile112 or Val112, while Ser112 was introduced during the evolution of angiosperm PALs (**Figure S8, S9, Table S5**) and likely prevents the introduction of TAL activity in angiosperm PALs by the single His/Phe140 substitution. Indeed, simultaneously introducing these two mutations converted both monocot and dicot PAL into PTAL enzymes (**Figure 4F, G**). Thus, the phylogeny-guided structure-function analysis unraveled the molecular basis of PTAL evolution and also provided novel plant genome editing targets to introduce the second entry pathway of lignin and phenylpropanoid biosynthesis.

The acquisition of PTAL offers several advantages, as PTAL allows for dual entry pathways to lignin biosynthesis and provides flexibility and redundancy in the production of lignin precursors. Additionally, this TAL pathway can bypass the P450 enzyme cinnamate 4-hydroxylase (C4H) and hence is considered to be more efficient than the canonical PAL-C4H pathway (**Figure 1B**)[89,90]. A recent study showed that grasses have altered regulation of the upstream aromatic amino acid biosynthetic pathway, resulting in efficient synthesis of both Phe and Tyr[91]. Together with the enhanced precursor supply, the dual entry pathway via PTAL likely enabled efficient synthesis of phenylpropanoids in fast-growing grasses that still accumulate substantial amounts of lignin (up to 30% of dry weight[21,92,17]) in the scattered vascular bundles of their culms and leaf sheaths. Leveraging insights from grass evolution, the introduction of a second entry pathway of lignin biosynthesis via the Ile112 and His140 residues, along with the engineering of the upstream Phe and Tyr biosynthesis[91,93,94], presents a promising strategy to enhance the production of diverse phenylpropanoid compounds in various plants.

Genomes of closely related species in sister lineages provide critical resources to dissect the genetic and molecular basis of unique traits that evolved in a target organism. For example, the chimpanzee and bonobo genomes have accelerated our understanding of the evolution of human traits[95,96]. This study demonstrated that integrating genomic and biochemical analysis of Poales that are close-relative to core-grasses can unveil the evolutionary origin and genetic and molecular basis of key metabolic innovations of grasses. These novel Poales genomes will similarly offer valuable resources to dissect complex traits uniquely evolved in grasses and their relatives, uncovering the genetic, biochemical, and molecular basis of the evolution of this most prominent plant family with immense economic and ecological significance.

ACKNOWLEDGEMENTS

This work was supported by the U.S. National Science Foundation (NSF) Plant Genome Research Program (IOS-1836824) to H.A.M.. Y.T-K. was partly supported by the Oversea research fellow of the Japan Society for the Promotion of Science (JSPS). Funding for the *Ecdeiocolea monostachya* sequencing includes United Kingdom Research and Innovation-Biotechnology and Biological Sciences Research Council (BBSRC) Norwich Research Park Doctoral Training Partnership (grant no. BB/M011216/1 to S.H.) and Institute Strategic Programme (grant no. BB/P012574/1 to M.J.M. and BBS/E/J/000PR9795 to M.J.M.), Gatsby Charitable Foundation (M.J.M.), and United States Department of Agriculture-Agricultural Research Service CRIS #5062-21220-025-000D (M.J.M.). The sequencing, assembly and annotation of *Typha latifolia*, *Joinvillea ascendens*, and *Pharas latifolius* genomes was conducted by the U.S. Department of Energy Joint Genome Institute (https://ror.org/04xm1d337; proposal: 10.46936/10.25585/60001405), a DOE Office of Science User Facility, supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231. We greatly appreciate valuable discussions with Drs. Thomas J. Givnish for the selection of monocot species, Drs. Barbara Briggs and Elizabeth Kellogg for the *Ecdeiocolea monostachya* genome sequencing and analysis, and Drs. Jaime Barros-Rios and Richard Dixon for grass PTAL functions. We thank Dr. Jorge El-Azaz for his assistance with the biochemical analysis of PALs/PTALs, Sarah Friedrich from the UW Botany Media Studio for illustration, and Cara Steekstra from the UW-Madison Botany Greenhouse for the growth of *Joinvillea ascendens.*

AUTHOR CONTRIBUTIONS

Conceptualization: H.A.M., J.H.L-M, M.R.M., M.J.M., J.S.
Investigation: Y.T-K., B.M., S.H., M.B., M.O., D.L., J.G., M.W., L.B.B., K.B.
Analysis: Y.T-K., B.M., M.R.M., S.K.D., M.J.M., S.H., J.J., C.P., S.S., K.B., D.M.G., J.S., H.A.M.
Project administration: H.A.M., J.H.L-M, M.R.M., M.J.M., J.S.
Writing – original draft: Y.T-K., H.A.M., J.H.L-M., S.H., M.J.M., J.J.
Writing – review & editing: Y.T-K., B.M., H.A.M., J.H.L-M., S.H., M.J.M., M.R.M., J.J., J.S.

DECLARATION OF INTERESTS STATEMENT

Y.T-K, B.M., and H.A.M. have a pending patent application related to the two mutations that can convert PAL enzymes into PTAL enzymes. The other authors declare that they have no competing interests related to this work.

FIGURES AND TABLES

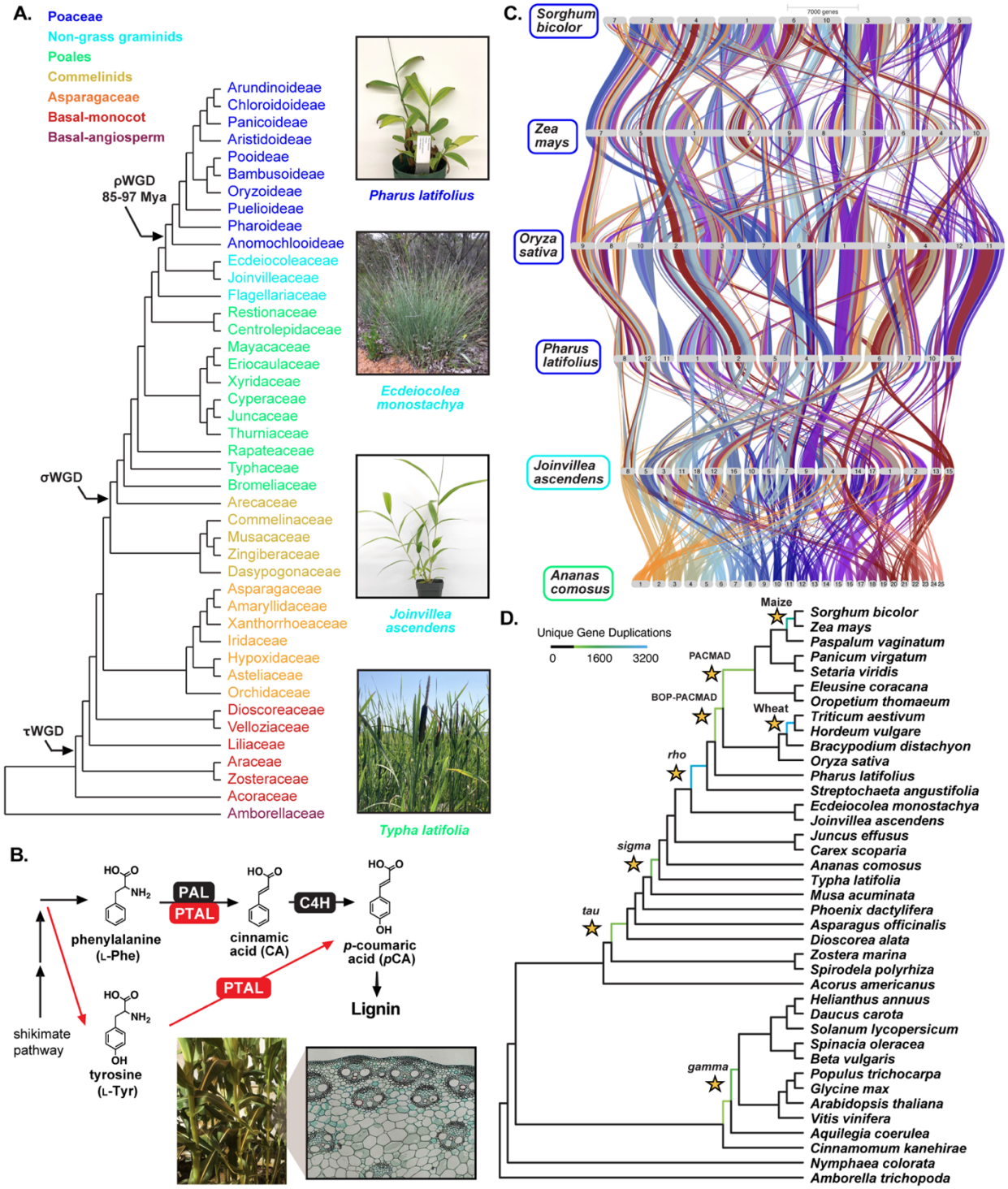

Figure 1: Sequencing genomes of Poaceae sisters to trace the evolutionary history of grass metabolic innovations.

A. The family tree of monocot plants showing three ancient whole genome duplication (WGD) events and the images of three plant species sister to core grasses. **B**. Grasses have a unique metabolic trait, the dual entry pathways to synthesize lignin and phenylpropanoids from both L-

phenylalanine (Phe) and L-tyrosine (Tyr) due to the presence of Phe/Tyr ammonia lyase (PTAL). A stem cross section of *Zea mays* showing scattered vascular bundles (Image by Michael Clayton). **C.** The whole genome comparisons of gene order (synteny) across representative Poaceae and Poales. **D.** Mapping of gene duplications counted in gene trees using the Phylogenetic Placement of Polyploidy Using Genomes (PUG) toolkit highlights ρWGD in the last common ancestor of all grasses (Poaceae), as well as sigma (σ) and tau (τ) WGMs.

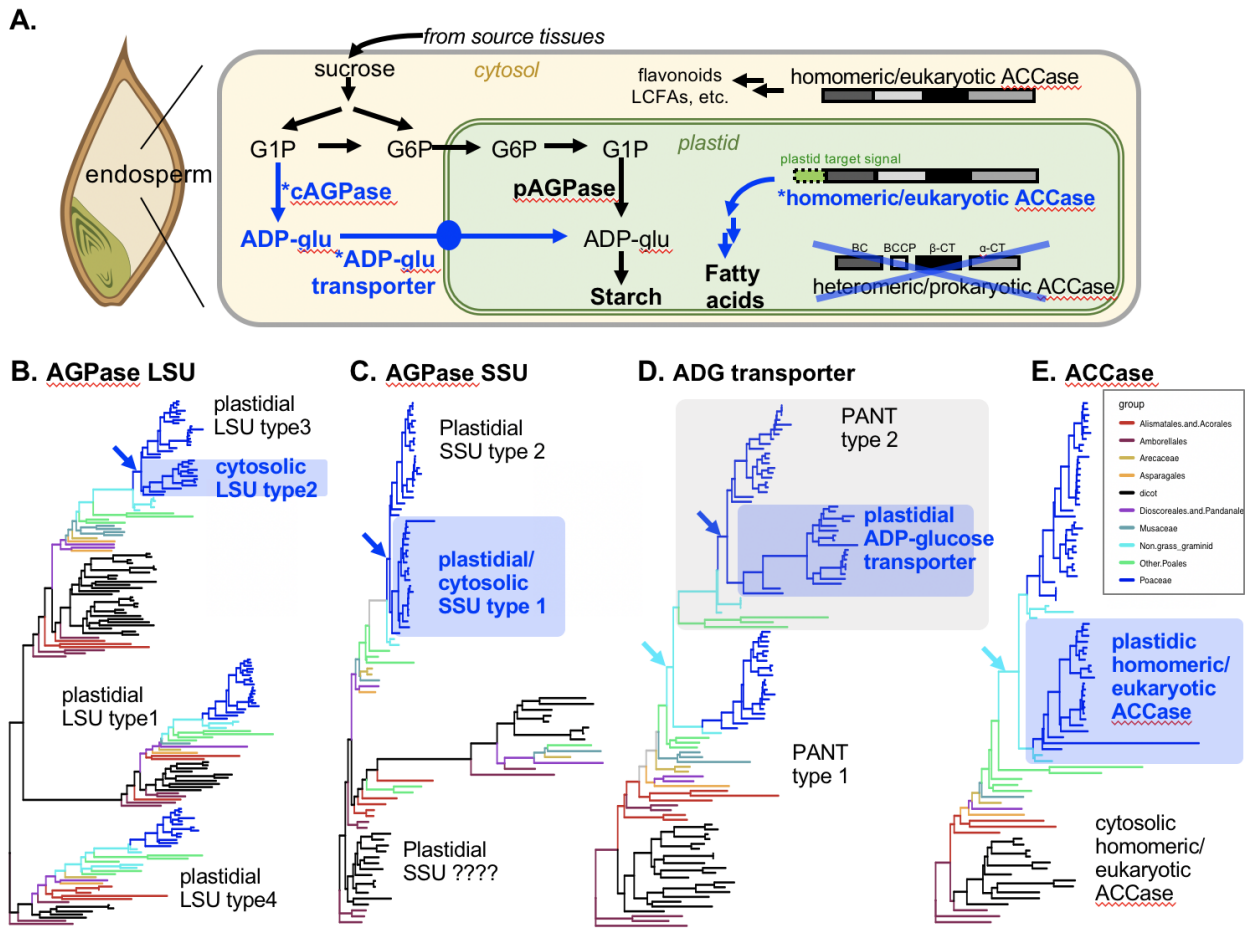

Figure 2. Alteration of starch and fatty acid biosynthesis during the grass evolution.

A. In the endosperm of grasses, starch can be synthesized in both cytosol and plastids due to the presence of cytosolic ADP-glucose pyrophosphorylase (AGPase) and ADP-glucose (ADG) transporter. Grasses also have unique fatty acid biosynthetic enzymes, homomeric eukaryotic-type acetyl-CoA carboxylase (ACCase) in the plastids, instead of heteromeric prokaryotic-type ACCase typically found in other plant plastids. **B-E.** The phylogenetic trees of AGPase large subunit (LSU, B), small subunit (SSU, C), ADG transporter (D), and ACCase (E) of grasses, Poales, as well as other monocots and angiosperms. Blue and gray boxes denote clades specific to Poaceae and non-grass graminids that likely contributed to the evolution of their unique starch and fatty acid biosynthetic pathways. Blue and sky-blue arrows indicate the underlying duplication events that occurred after and before the emergence of grasses, respectively. The original trees with branch supports and labels are shown in **Figure S1**.

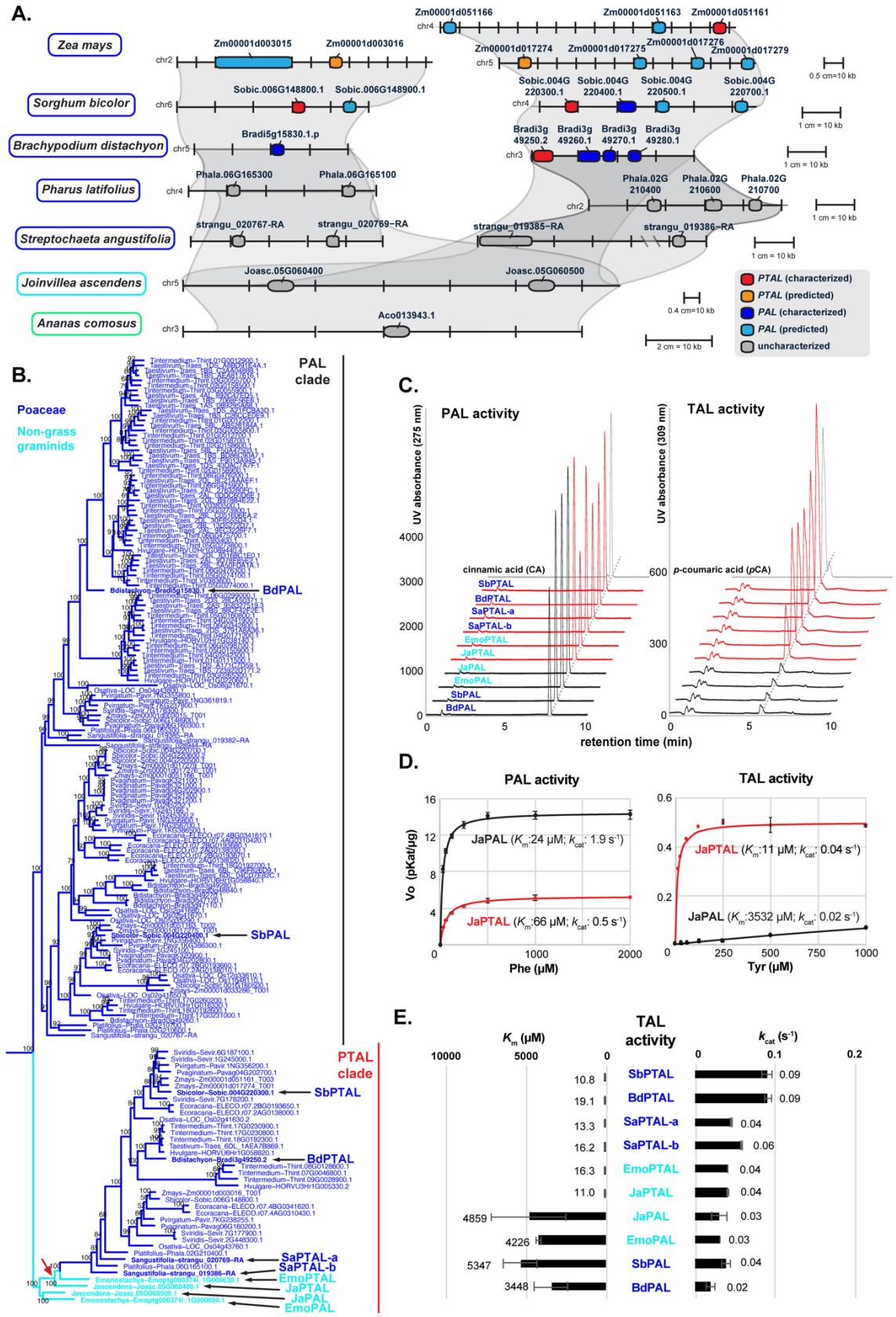
A.
Zea mays
Sorghum bicolor
Brachypodium distachyon
Pharus latifolius
Streptochaete angustifolia
Joinvillea ascendens
Ananas comosus
Zm00001d003015
Zm00001d003016
Zm00001d051166
Zm00001d051163
Zm00001d051161
Zm00001d017274
Zm00001d017275
Zm00001d017276
Zm00001d017279
Sobic.006G148800.1
Sobic.006G148900.1
Sobic.004G 220300.1
Sobic.004G 220400.1
Sobic.004G 220500.1
Sobic.004G 220700.1
Bradi5g15830.1.p
Bradi3g 49250.2
Bradi3g 49260.1
Bradi3g 49270.1
Bradi3g 49280.1
Phala.06G165300
Phala.06G165100
Phala.02G 210400
Phala.02G 210600
Phala.02G 210700
strangu_020767-RA
strangu_020769-RA
strangu_019385-RA
strangu_019386-RA
Joasc.05G060400
Joasc.05G060500
Aco013943.1
0.5 cm=10 kb
1 cm = 10 kb
0.4 cm=10 kb
2 cm = 10 kb
PTAL (characterized)
PTAL (predicted)
PAL (characterized)
PAL (predicted)
uncharacterized
B.
PAL clade
Poaceae
Non-grass graminids
BdPAL
SbPAL
PTAL clade
SbPTAL
BdPTAL
SaPTAL-a
SaPTAL-b
EmoPTAL
JaPTAL
JaPAL
EmoPAL
C.
PAL activity
TAL activity
UV absorbance (275 nm)
UV absorbance (309 nm)
cinnamic acid (CA)
p-coumaric acid (pCA)
SbPTAL
BdPTAL
SaPTAL-a
SaPTAL-b
EmoPTAL
JaPTAL
JaPAL
EmoPAL
SbPAL
BdPAL
retention time (min)
D.
PAL activity
TAL activity
Vo (pKat/µg)
JaPAL (Km:24 µM; kcat: 1.9 s-1)
JaPTAL (Km:66 µM; kcat: 0.5 s-1)
JaPTAL (Km:11 µM; kcat: 0.04 s-1)
JaPAL (Km:3532 µM; kcat: 0.02 s-1)
Phe (µM)
Tyr (µM)
E.
TAL activity
Km (µM)
kcat (s-1)
SbPTAL 10.8 0.09
BdPTAL 19.1 0.09
SaPTAL-a 13.3 0.04
SaPTAL-b 16.2 0.06
EmoPTAL 16.3 0.04
JaPTAL 11.0 0.04
JaPAL 4859 0.03
EmoPAL 4226 0.03
SbPAL 5347 0.04
BdPAL 3448 0.02

Figure 3: The PTAL enzyme evolved before the emergence of grasses through tandem duplication predating ρWGD.

A. *PAL/PTAL* gene synteny is shown for *Ananas comosus, Joinvillea ascendens, Streptochaeta angustifolia, Pharus latifolius, Brachypodium distachyon, Sorghum bicolor,* and *Zea mays*. The *PAL* gene is duplicated and neofunctionalized to PTAL in the ancestor of *J. ascendens*. The *PAL*/*PTAL* pair is then duplicated within grasses, likely due to the ρWGD, and again in *Z. mays*, likely due to the *Z. mays* WGD. Figure shows gains and losses of *PAL*/*PTAL* in synteny across Poales. **B**. Phylogenetic tree of *PAL*/*PTAL* genes in monocots focusing on Poales species. *PAL*/*PTAL* homologs characterized in this study are highlighted with black arrows. The red arrow indicates the node used for the positive selection analysis. The full-scaled phylogenetic tree is shown in **Figure S3**. **C.** HPLC trace of PAL and TAL activity assay for PTALs/PALs from *S. bicolor* (SbPTAL and SbPAL), *B. distachion* (BdPTAL and BdPAL), *S. angustifolia* (SaPTAL-a and SaPTAL-b), *J. ascendens* (JaPTAL and JaPAL) and *Ecdeiocolea monostachya* (EmoPTAL and EmoPAL). See **Figure S6** for additional traces of their control reactions. **D.** Michaelis-Menten curves of TAL and PAL assay for JaPTAL and JaPAL. Data are means ± SD. **E.** K_m and k_{cat} of TAL activity for PTALs/PALs. Data are means ± SD.

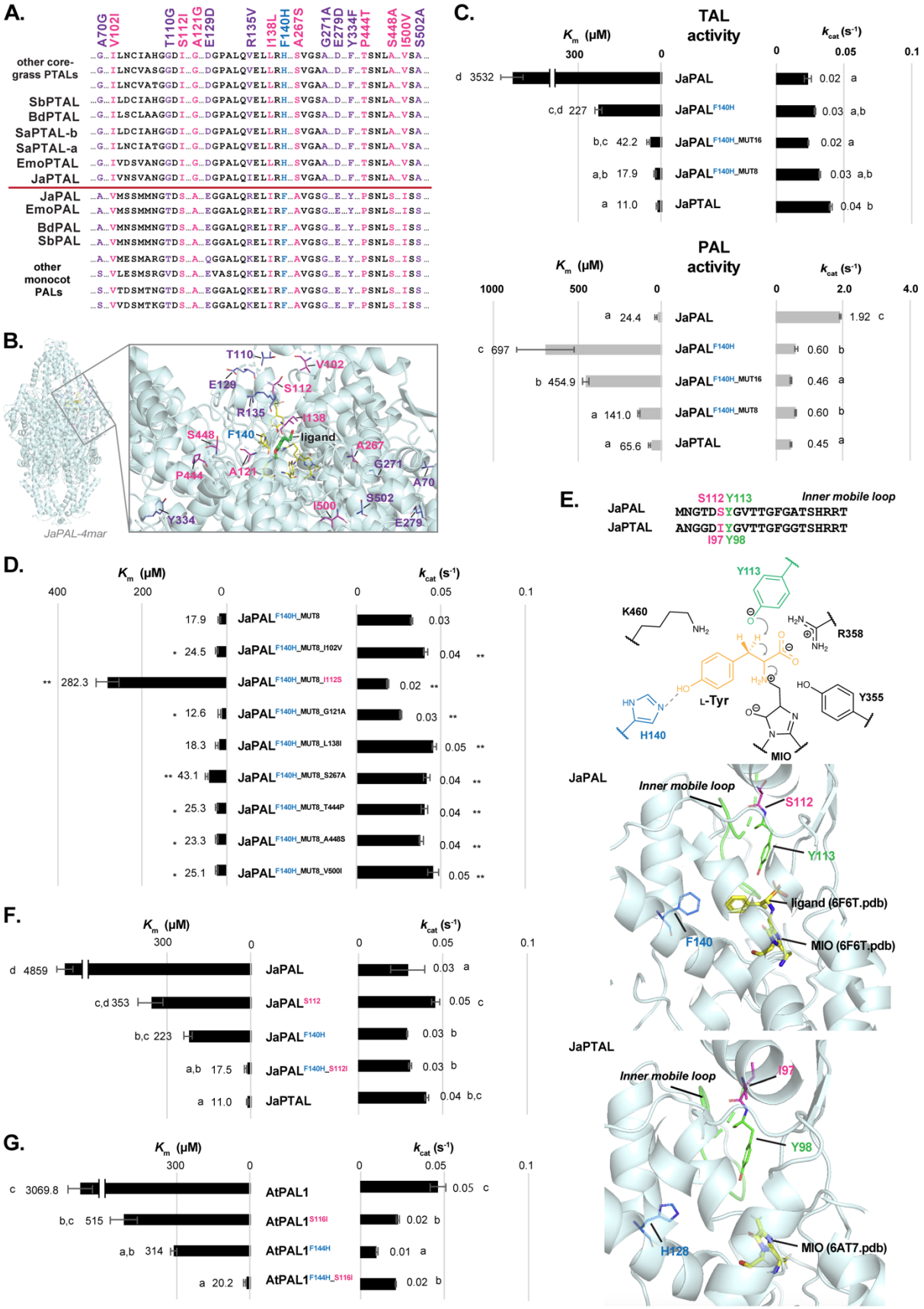
A.
A70G V102I T110G S112I A121G E129D R135V I138L F140H A267S G271A E279D Y334F P444T S448A I500V S502A
other core-grass PTALs
SbPTAL
BdPTAL
SaPTAL-b
SaPTAL-a
EmoPTAL
JaPTAL
JaPAL
EmoPAL
BdPAL
SbPAL
other monocot PALs
B.
JaPAL-4mar
T110
V402
E129
S112
R135
I138
S448
F140
ligand
A267
G271
P444
A121
A70
S502
I500
Y334
E279
C.
TAL activity
Km (μM)
kcat (s-1)
JaPAL
JaPAL F140H
JaPAL F140H_MUT16
JaPAL F140H_MUT8
JaPTAL
PAL activity
D.
JaPAL F140H_MUT8
JaPAL F140H_MUT8_I102V
JaPAL F140H_MUT8_I112S
JaPAL F140H_MUT8_G121A
JaPAL F140H_MUT8_L138I
JaPAL F140H_MUT8_S267A
JaPAL F140H_MUT8_T444P
JaPAL F140H_MUT8_A448S
JaPAL F140H_MUT8_V500I
E.
JaPAL MNGTDSYGVTTGFGATSHRRT
JaPTAL ANGGDIYGVTTGFGGTSHRRT
S112 Y113
I97 Y98
Inner mobile loop
K460
R358
Y113
L-Tyr
Y355
H140
MIO
JaPAL
Inner mobile loop
S112
Y113
ligand (6F6T.pdb)
F140
MIO (6F6T.pdb)
JaPTAL
Inner mobile loop
I97
Y98
H128
MIO (6AT7.pdb)
F.
JaPAL
JaPAL S112
JaPAL F140H
JaPAL F140H_S112I
JaPTAL
G.
AtPAL1
AtPAL1 S116I
AtPAL1 F144H
AtPAL1 F144H_S116I

Figure 4: Two mutations, F140H and S112I, convert PAL into PTAL enzymes.

A. Partial amino acid alignment, highlighting residue His/Phe140 that has been reported to be critical for the recognition of substrate phenylalanine and tyrosine (cyan), residues highly conserved uniquely in either PTALs and PALs (magenta), and the additional 8 residues were highly conserved among PTALs but not among PALs (purple). The full-scaled alignment is shown in **Figure S8**. **B.** A protein structure model of JaPAL highlighting the selected 16 candidate amino acid residues identified by the phylogeny-guided alignment. Structural information of the substrate ligand and MIO cofactor are from the template 6F6T.pdb. The residues important for the enzyme reaction as well as the substrate ligand are displayed in yellow. **C.** K_m and k_{cat} of TAL and PAL activity for JaPAL, JaPAL with site directed mutation at His140 site ($JaPAL^{F140H}$) and additional mutations ($JaPAL^{F140H_MUT8}$ and $JaPAL^{F140H_MUT16}$), and JaPTAL. **D.** K_m and k_{cat} of TAL activity for $JaPAL^{F140H_MUT8}$ enzymes with mutation on 8 of each residue back to PAL type. **E.** A potential mechanism of TAL reaction[77] and hypothetical roles of the two residues His140 and Ile112 for the catalysis of the PTAL enzyme. The residue Ser/Ile112 is located next to the Tyr113 residue critical for the catalysis, and these residues are in the 'inner mobile loop' shown in light green cartoon in the models, which has been suggested to be critical for the substrate binding and catalysis. **F.** K_m and k_{cat} of TAL activity for JaPAL enzyme with site-directed mutation at His140 site ($JaPAL^{F140H}$) and Ile112 site ($JaPAL^{S112I}$). **G.** K_m and k_{cat} of TAL activity for Arabidopsis AtPAL1 enzyme with site-directed mutation at His140 site ($AtPAL1^{F144H}$) and Ile112 site ($AtPAL1^{S116I}$). Data are means ± SD. Different letters indicate a significant difference between treatment conditions (ANOVA with post hoc Tukey–Kramer test or Kruskal-Wallis nonparametric test, $p < 0.05$).

Table 1. Summary statistics of the genome assembly of *J. ascendens*, *E. monostachya*, *P. latifolius,* and *T. latifolia.*

Feature	*J. ascendens*	*E. monostachya*	*P. latifolius*	*T. latifolia*
Total length	1.21 Gb	897 Mb	1.12 Gb	215 Mb
Number of scaffolds	684	1,114	198	16
N50	63.7 Mb	1.7 Mb	97.0 Mb	14.6 Mb
Longest scaffold	89.4 Mb	12.4 Mb	115.8 Mb	24.1 Mb
Mean scaffold length	1.8 Mb	805 kb	5.6 Mb	13.5 Mb
BUSCO genes	98%	97.9%	98.4%	99.7%

SUPPLEMENTAL DATA

The following materials are available in the online version of this article.

Figure S1. Full-scaled gene trees for genes involved in starch and fatty acid biosynthesis

Figure S2. Distributions of synonymous substitution rates (*K*s) across synthetic genes between Poales species

Figure S3. Amino acid based phylogenetic tree of *PAL/PTAL* genes in monocot

Figure S4. Coding sequence based phylogenetic tree of *PAL/PTAL* genes in monocot

Figure S5. TreeSolve analysis of the PAL/PTAL trees

Figure S6. HPLC chromatogram of PAL and TAL reaction product from PALs/PTALs

Figure S7. Kinetics parameter of PAL/TAL activity

Figure S8. Alignment of PTAL and PAL proteins from monocot

Figure S9. Alignment of PAL homologs across all domains of life, excluding plants

Figure S10. Phylogenetic tree of PAL homologs in all domain of life, suggesting the Ser112 residue was acquired during the land plant evolution

Figure S11. Identification and sampling site of *E. monostachya* accession E001.

Figure S12. Distribution of k-mers from *E. monostachya* short read sequencing relative to genome.

Figure S13. Distribution of k-mers from *E. monostachya* short read sequencing relative to genome.

Table S1. List of species used in the phylogenetic tree analysis and total number of putative paralogs per species

Table S2. Total counts of unique gene duplications mapped to placement of known WGD events on the species tree and to portions of the tree without expected WGD events

Table S3. Kinetics parameters of recombinant PTAL and PAL proteins

Table S4. Residues potentially involved in the transition from PAL to PTAL

Table S5. List of *PAL* genes used for building the phylogenetic tree of *PAL* family genes from all domains of life

Table S6. Genomic libraries included in the *Pharus latifolius* genome assembly and their respective assembled sequence coverage levels in the final release

Table S7. PACBIO CCS library statistics for the libraries included in the *Pharus latifolius* genome assembly and their respective assembled sequence coverage levels

Table S8. Summary statistics of the initial output of the RACON polished HiFiAsm+HIC assembly of the *Pharus latifolius* genome

Table S9. Final summary assembly statistics for the version 1 chromosome scale assembly of *Pharus latifolius*

Table S10. Genomic libraries included in the *Typha latifolia* genome assembly and their respective assembled sequence coverage levels in the final release

Table S11. PACBIO CCS library statistics for the libraries included in the *Typha latifolia* genome assembly and their respective assembled sequence coverage levels

Table S12. Summary statistics of the initial output of the HAP1 RACON polished HiFiAsm+HIC assembly of the *Typha latifolia* genome

Table S13. Summary statistics of the initial output of the HAP2 RACON polished HiFiAsm+HIC assembly of the *Typha latifolia* genome

Table S14. Final summary assembly statistics for the version 1 HAP1 chromosome scale assembly of *Pharus latifolius*

Table S15. Final summary assembly statistics for the version 1 HAP2 chromosome scale assembly of *Pharus latifolius*

Table S16. Primers used in this study

Data S1. The complete peptide sequences of PALs and PTALs used for the phylogenetic tree analysis.

Data S2. The complete CDS sequences of PALs and PTALs used for the dN/dS analysis.

MATERIALS AND METHODS

Key resources table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, peptides, and recombinant proteins		
JaPTAL	Phytozome v13	Joasc.05G060400
JaPAL	Phytozome v13	Joasc.05G060500
SbPTAL	Phytozome v13	SbiRTX430.04G229900.1
SbPAL	Phytozome v13	SbiRTX430.04G230000.1
BdPTAL	Phytozome v13	Bradi3g49250.2
BdPAL	Phytozome v13	Bradi5g15830.1
AtPAL1	Phytozome v13	AT2G37040.1
Deposited data		
Genome of *Joinvillea ascendens*	Joint Genome Institute (JGI) Open Green Genome Initiative (OGG) https://phytozome-next.jgi.doe.gov/ogg	https://phytozome-next.jgi.doe.gov/ogg/info/Jascendens_v1_1
Genome of *Ecdeiocolea monostachya* accession E001	DDBJ/ENA/GenBank	BioProject codes PRJNA894727 (Illumina/Oxford Nanopore Technology hybrid assembly) and PRJNA1179411 (PacBio HiFi assembly)
Genome of *Pharus latifolius*	Joint Genome Institute (JGI) Open Green Genome Initiative (OGG) https://phytozome-next.jgi.doe.gov/ogg	https://phytozome-next.jgi.doe.gov/info/Platifolius_v1_1
Genome of *Typha latifolia*	Joint Genome Institute (JGI) Open Green Genome Initiative (OGG) https://phytozome-next.jgi.doe.gov/ogg	https://phytozome-next.jgi.doe.gov/info/Tlatifoliavar_2019_6_v1_1
Experimental models:		
Joinvillea ascendens subsp. glabra	The National Tropical Garden (Kalaheo, HI), from seed collected in New Caledonia.	Accession No. 800379
Ecdeiocolea monostachya	Western Australia near Jurien Bay	Accession No. E001

Pharus latifolius	Missouri Botanical Garden (St. Louis, MO)	Accession No. 1993-0885-2
Typha latifolia	Penn State Arboretum (University Park, PA)	Accession No. CWD-2019.6
Software and algorithms		
ape 5.0	Paradis and Schliep[97]	http://cran.r-project.org/package=ape
Augustus	Keller et al.[98]	https://github.com/Gaius-Augustus/Augustus
BLAST	Camacho et al.[99]	https://www.ncbi.nlm.nih.gov
BLAT	Kent [100]	http://genome.ucsc.edu/cgi-bin/hgBlat
BUSCO	Manni et al.[101]	https://busco.ezlab.org/
bwa-mem	Li [102]	https://github.com/lh3/bwa
bwa-mem2	Li [102]	https://github.com/bwa-mem2/bwa-mem2
COGE	Lyons and Freeling[103]	https://genomevolution.org/coge/
Cufflinks/Cuffmerge	Trapnell et al.[104]	http://cole-trapnell-lab.github.io/cufflinks/
EVM	Haas et al.[105]	https://github.com/EVidenceModeler/EVidenceModeler/wiki
findGSE	Sun et al.[106]	https://github.com/schneebergerlab/findGSE
GATK	Van der Auwera and O'Connor [107]	https://gatk.broadinstitute.org
GeMoMa	Keilwagen et al.[108]	https://github.com/Jstacs/Jstacs
GeneMarkS-T	Tang et al. [109]	https://genemark.bme.gatech.edu/
GENESPACE	Lovell et al. [58]	https://github.com/jtlovell/GENESPACE
GenomeScope	Vurture et al.[110]	http://genomescope.org/
GSNAP	Wu et al. [111]	http://research-pub.gene.com/gmap/
HiFiAsm	Cheng et al. [112,113]	https://github.com/chhylp123/hifiasm

hisat2	Kim et al. [114]	http://daehwankimlab.github.io/hisat2/
IQ-TREE 2	Minh et al. [115]	http://www.iqtree.org
jellyfish	Marcais and Kingsford[116]	https://github.com/gmarcais/Jellyfish
Juicer	Durand et al. [117]	https://github.com/aidenlab/juicer
KmerGenie	Chikhi and Medvedev[118]	http://kmergenie.bx.psu.edu/
k-mer analysis toolkit	Mapleson et al. [119]	https://github.com/TGAC/KAT
kraken2	Wood et al.[120]	https://github.com/DerrickWood/kraken2
MAFFT	Katoh and Standley [109]	https://mafft.cbrc.jp/
MAKER	Campbell et al. [121]	https://github.com/Yandell-Lab/maker
MaSuRCA	Zimin et al. [122]	https://github.com/alekseyzimin/masurca
MECAT	Xiao et al. [123]	https://thub.com/xiaochuanle/MECAT
ModelTest-NG	Darriba et al. [124]	https://github.com/ddarriba/mgiodeltest
OrthoFinder	Emms and Kelly [125]	https://github.com/davidemms/OrthoFinder
PAL2NAL	Suyama et al. [126]	https://www.bork.embl.de/pal2nal/
PAML	Yang [80]	https://github.com/abacus-gene/paml
PASA	Haas et al.[127]	https://github.com/PASApipeline/PASApipeline/
PERTRAN	Shu et al.[128]	N/A
PUG	McKain et al. [3]	https://github.com/mrmckain/PUG
purge_dups	Guan et al. [129]	https://github.com/dfguan/purge_dups
R	R Development Core Team, 2020	www.R-project.org/
RACON	Vaser et al. [56]	https://github.com/isovic/racon

RAxML-NG	Kozlov et al.[130]	https://github.com/amkozlov/raxml-ng
RepeatModeler	Smit and Hubley	https://github.com/Dfam-consortium/RepeatModeler
samtools	Danece et al. [131]	https://github.com/samtools/samtools
Solver	Microsoft Office Excel	N/A
SNAP	Korf [132]	https://github.com/KorfLab/SNAP
StringTie	Pertea et al.[133]	https://github.com/gpertea/stringtie
SWISS-MODEL	Waterhouse et al. [134]	https://swissmodel.expasy.org/
TransDecoder	N/A	https://github.com/TransDecoder/TransDecoder
TreeSolve	Kordi et al. [75]	https://compbio.engr.uconn.edu/software/TreeSolve/
Trimmomatic	Bolger et al.[135]	http://www.usadellab.org/cms/?page=trimmomatic
Trinity	Grabherr et al.[136]	https://github.com/trinityrnaseq/trinityrnaseq/wiki
3D-DNA	Dudchenko et al.[137]	https://github.com/aidenlab/3d-dna

Experimental models and Method details

Genome sequencing of Joinvillea ascendens

To obtain genome sequence *J. ascendens*, in May 2017, we first harvested drupes, fruits containing seeds (the seed batch of NTBG #800379), from the *J. ascendens* subsp. glabra plant grown at the National Tropical Garden (Kalaheo, HI, Accession No. 800379, herbarium voucher Lorence 7657, PTBG) from seed originally collected in New Caledonia. These drupes were sent to University of Wisconsin-Madison, physically removed for fruits using a laser blade, and the seeds were soaked with 5% liquid smoke (Wright's Liquid Smoke, Hickory) for 24 hours[138], rinsed in distilled water, and then planted on three parts perlite to one part potting mix. After two months in July 2017, one out of >100 planted seeds germinated. After 6 months of the germination in January 2018, newly developed young leaves were sprayed with clean distilled water, and ~100 mg fresh tissues were harvested and ground in liquid nitrogen and used for genomic DNA isolation using the DNeasy Plant Mini Kit (Qiagen).

The isolated DNA was sent to and sequenced at JGI using PACBIO platform at 125x coverage

(average read length of 10,007). Main assembly was performed and polished using MECAT[123]. Misjoins in the assembly were identified using Hi-C data. A total of 27 misjoins were identified in the polished assembly. Scaffolds were then oriented, ordered, and joined together using Hi-C scaffolding. Significant telomeric sequence was properly oriented in the assembly. A total of 186 joins were applied to the broken assembly to form the final assembly consisting of 18 chromosomes (2N), with a total of 96.5% of the assembled sequence contained in the chromosomes. Hi-C reads were then aligned to the joined release. The alignments were converted to contact map to quality control check on the order/orientation of contigs in the chromosomes. Seven discrepancies were corrected. Finally, care was taken to ensure that telomere sequence was properly oriented in the chromosomes, and the resulting sequence was screened for retained vector and/or contaminants. Adjacent alternative haplotypes were identified on the joined contig set. Alternative haplotype regions were collapsed using the longest common substring between the two haplotypes. A total of 67 adjacent alternative haplotypes were collapsed. Chromosomes were numbered from largest to smallest, and the p-arm of each chromosome was oriented to the 5' end. Additionally, homozygous SNPs and INDELs were corrected in the release sequence using 40x of Illumina reads (2x150, 400bp insert).

To aid in annotation by MAKER[121], ~2B pairs of 2X150 stranded paired-end Illumina RNA sequencing was performed on various tissues: the leaf tip, leaf base, upper internodes, base internodes, and root tissues harvested into liquid nitrogen at University of Wisconsin-Madison, as well as the inflorescence tissues harvested and immediately cooled by dry ice at the National Tropical Garden. transcript assemblies were generated from 2 × 150 paired-end Illumina RNA-seq reads using PERTRAN[128], which conducts genome-guided transcriptome short read assembly via GSNAP[111] and builds splice alignment graphs after alignment validation, realignment and correction. The highest-scoring predictions for each locus were selected using multiple positive factors including EST and protein support, and one negative factor: overlap with repeats. Improvement includes adding UTRs, splicing correction, and adding alternative transcripts. The transcripts were selected if the Cscore was larger than or equal to 0.5 and protein coverage larger than or equal to 0.5, or if it had EST coverage, but the CDS overlapping with repeats was less than 20%. For gene models whose CDS overlaps with repeats for more that 20%, its Cscore had to be at least 0.9 and homology coverage at least 70% to be selected. The selected gene models were subject to Pfam analysis and gene models whose protein was more than 30% in Pfam Transposable elements (TE) domains were removed. Repetitive DNA elements were identified de novo with RepeatModeler.

Genome sequencing and analysis of Ecdeiocolea monostachya

E. monostachya accession E001 was isolated at flowering stage on 2018 Sep 12 (**Figure S11**). A genome size estimate of *E. monostachya* accession E001 was 2.0 pg (2C) with 38 or 42 chromosomes (2N). Analysis of three additional accessions (E002, E010, and E014) had similar estimates of genome size and chromosome number. Hanson and colleagues (2005) previous estimated chromosome number at 38 and genome size of 1.98 pg (2C). Therefore, *E. monostachya* is estimated to have a haploid genome size of 1.94 Gb.

For genome assembly, genomic DNA was collected from sheath tissue of *E. monostachya* accession E001 and sequenced using paired-end Illumina short-read and PacBio HiFi long-read sequencing. The estimated haploid genome size using GenomeScope[110] ranged from 734.0 Mb (k=17) to 839.2 (k=24) and heterozygosity estimate of 2.33% (k=31) to 2.65% (k=17). findGSE[106] and KmerGenie[118] predicted diploid genome sizes of 1.47 Gb (k=24) and 1.31 Gb (k=107).

A total of 16.84 Gb of PacBio HiFi data was generated with an approximate depth of 21x and average read length of 14.4 kb. Genome assembly was performed using HiFiAsm (v0.16)[112,113] using parameters '-I = 2' and '-n=4'. Contigs derived from heterozygous regions were removed using purge_dups (v.1.2.5)[129] using default parameters. The final draft haploid genome assembly was 897 Mbp over 1,114 scaffolds (**Table 1**). The majority of the genome (>50%) spans contigs greater than 1.7 Mb. Evaluation of *k*-mer composition of the resulting assembly relative to raw Illumina reads found that the genome was partially phased (**Figure S12**).

An integrated gene annotation approach using de novo prediction, homology search, and evidence-based (RNAseq) assembly was used to annotate protein-coding genes in the genome. Prediction of *de novo* gene models was performed with Augustus (v2.4)[98] and SNAP (v2006-07-28)[132]. The homolog-based approach involved GeMoMa (v1.7)[108] using reference gene model from Poales species (*Brachypodium distachyon*, *Setaria italica*, *Oryza sativa*, and *Joinvillea ascendens*). Evidence-based gene prediction used paired-end Illumina RNAseq derived from sheath, root, and flower (immature embryos) sampled from *E. monostachya* accession E001. Spliced alignment using hisat2[114] found that the majority of RNAseq reads (89.5-91.0%) aligned to the genome with non-aligning RNAseq reads (9.1-9.6%) associated with bacterial, human, Achaea, or viral contamination based on kraken2[120] classification. Reads were mapped using hisat (v2.0.4)[114] and assembled by Stringtie (v1.2.3)[133]. Genes were predicted from assembled transcripts using GeneMarkS-T (v5.1)[109]. Gene prediction from Trinity (v2.11)[136] was performed with PASA (v2.0.2)[127]. EVM (v1.1.1)[105] was used to merge gene models from these multiple approaches and was updated by PASA. Analysis of the benchmark universal single copy orthologs (BUSCO)[101] using the predicted transcripts found 94.6% complete, 2.9% fragmented, and 2.5% missing genes. The majority of BUSCO genes were not duplicated (7.0%) reflecting the purging of redundant contigs. The final set of protein encoding genes included 27,801 gene models.

Previous work has shown that the GC content of exons in Poaceae genomes has a distinct bimodal distribution[3,139]. Evaluation of the GC content of *E. monostachya* found monomodal distribution that is skewed towards high GC content (**Figure S13**). This suggests that a subset of genes have a higher GC content than expected through a neutral process.

For sampling and DNA/RNA extraction, a survey was performed to identify *E. monostachya* in Western Australia near Jurien Bay (**Figure S11**). For *E. monostachya* accession E001, samples of sheath, root, and flower were sampled and frozen on dry ice. Genomic DNA of *E. monostachya* was extracted using sheath tissue with a CTAB-based extraction protocol[140]. Total RNA was extracted from sheath, root, and flower tissue using a Trizol-based extraction protocol. For flower tissue, developing seeds were extracted from whole flowers in liquid nitrogen.

For short read and long-read sequencing, Illumina and Oxford Nanopore Technology (ONT) library construction and sequencing using *E. monostachya* accession E001 genomic DNA was performed by Novogene (Beijing, China). Illumina paired end libraries were generated using inserts of 250 bp and 350 bp and sequenced using Illumina HiSeq. In total, 123.2 Gb of short-read Illumina sequencing data was generated. In total, 12.8 million ONT reads were generated that ranged in length from 500 bp to 849 kb with a median length of 4.6 kb and total size of 69.7 Gb.

Hybrid assembly of the *E. monostachya* accession E001 genome was carried out using MaSuRCA (v3.3.0) using default parameters on an Amazon AWS x1.32xlarge instance (128 CPUs, 1,952 GB RAM) running SUSE Linux with raw Illumina and ONT reads. BUSCO genes were identified using BUSCO (v3.1.0)[101] with default parameters and Embryophyta (ODB9) database.

For k-mer analysis, Illumina reads were trimmed using Trimmomatic (v0.36)[135] using the parameters 2:30:10 for clipping reads based on TruSeq3 paired end adapters, leading and trailing

of 5 bp, sliding window of 4:10, and a minimum length of 36 bp. Initial *k*-mer analysis of Illumina short-read data to determine coverage was performed using jellyfish (v1.1.12) with k of 17, 24, 27, and 31 bp. Genome size was estimated using KmerGenie[118] (v1.7048), findGSE[106], and GenomeScope[110] (http://genomescope.org/). The *kat comp* command in the *k*-mer analysis tookit (v2.4.1) using default parameters (k=27) was used to generate the stacked histogram to determine the proportion and frequency of *k*-mers in short-read sequencing data relative to the hybrid genome assembly.

To determine potential contamination in *E. monostachya* RNAseq samples, kraken2 (v2.0.8-beta)[120] was used to classify the kingdom origin of individual paired reads using default parameters and libraries: archaea, bacteria, plasmid, viral, human, fungi, plant, protozoa, UniVec_core, and nt. The kraken2 database was generated on Nov 27th, 2019.

For gene annotation, RNAseq reads were trimmed using Trimmomatic (v0.36)[135] using the same parameters for genomic DNA reads. Spliced alignments to the hybrid genome were performed using hisat2 (v2.2.1)[114] with the following parameters: maximum intron length of 20 kb, k of 1, no soft clipping, and alignments tailored for Cufflinks. SAM files were converted into sorted BAM files using samtools (v1.11). Cufflinks/Cuffmerge (v2.2.1) pipeline using default parameters was used to identify evidence-based gene models. TransDecoder (v5.5.0) was used with default parameters to identify open reading frames (LongORFs) and predicted proteins (Predict) using Pfam. Completeness of conserved orthologs was performed using BUSCO (v3)[101] with the embryophyta_odb9 lineage and default parameters.

Genome sequencing of Pharus latifolius

We sequenced *Pharus latifolius* (accession 1993-0885-2 from the Missouri Botanical Garden; herbarium voucher McKain 320) using a whole genome shotgun sequencing strategy and standard sequencing protocols. This specimen was originally collected from near Napo, Ecuador on February 7, 1992 by MacDougal & Lalumondier (collection number 4792). Since that time, it has been maintained and propagated in the climatron at the Missouri Botanical Garden. Sequencing reads were collected using Illumina and PACBIO platforms. Illumina and PACBIO reads were sequenced at the HudsonAlpha Institute in Huntsville, Alabama using the Illumina NovoSeq 6000 platform and the SEQUEL II platform, respectively. One 400bp insert 1x250 Illumina fragment library (170.65x) was sequenced along with one 2x150 HiC library (76.53x) (**Table S6**). Prior to assembly, Illumina fragment reads were screened for phix contamination. Reads composed of >95% simple sequences were removed. Illumina reads <50 bp after trimming for adapter and quality (q<20) were removed. The final Illumina read set consists of 1,901,151,696 reads for a total high-quality base pair yield of 247.18x. For the PACBIO sequencing, a total raw sequence yield of 39.2 Gb, with a total coverage of 35.08x (**Tables S6, S7**).

For genome assembly and construction of pseudomolecule chromosomes, the version 1.0 assembly was generated by assembling the 2,058,145 PACBIO CCS reads (35.08x) using the HiFiAsm assembler[112,113] and subsequently polished using RACON[56]. This produced an initial assembly consisting of 3,267 scaffolds (3,267 contigs), with a contig N50 of 31.6 Mb, and a total genome size of 1,234.5 Mb (**Table S8**).

Hi-C Illumina reads from *Pharus latifolius* (var. MRM_UAlabama_McKain320), were separately aligned to the contigs with Juicer[117], and chromosome scale scaffolding was performed with 3D-DNA[137]. No misjoins were identified in the assembly, and the contigs were then oriented, ordered, and joined together into 12 chromosomes using the HiC data. A total of 47 joins were applied to the assembly. Each chromosome join is padded with 10,000 Ns. Contigs terminating in

significant telomeric sequence were identified using the $(TTTAGGG)_n$ repeat, and care was taken to make sure that they were properly oriented in the production assembly. The remaining scaffolds were screened against bacterial proteins, organelle sequences, GenBank nr and removed if found to be a contaminant. After forming the chromosomes, it was observed that some small (<20Kb) redundant sequences were present on adjacent contig ends within chromosomes. To resolve this issue, adjacent contig ends were aligned to one another using BLAT[100], and duplicate sequences were collapsed to close the gap between them. A total of 23 adjacent contig pairs were collapsed in the assembly.

Finally, homozygous SNPs and INDELs were corrected in the V1 release using ~40x of Illumina reads (2x150, 400bp insert) by aligning the reads using bwa-mem[102] and identifying homozygous SNPs and INDELs with the GATK's UnifiedGenotyper tool[107]. A total of 291 homozygous SNPs and 12,285 homozygous INDELs were corrected in the V1 release. The final version 1.0 release contained 1,117.9 Mb of sequence, consisting of 222 contigs with a contig N50 of 57.1 Mb and a total of 99% of assembled bases in chromosomes (**Table S9**).

Completeness of the euchromatic portion of the version 1.0 assembly was assessed by aligning the existing IsoSeq reads to the V1 release. The aim of this analysis is to obtain a measure of completeness of the assembly, rather than a comprehensive examination of gene space. The IsoSeq alignments indicate that 99.68% of the reads are aligned to the version 1 release.

For screening and final assembly releases, scaffolds that were not anchored in a chromosome were classified into bins depending on sequence content. Contamination was identified using blastn against the NCBI non-redundant nucleotide collection (NR/NT) and blastx using a set of known microbial proteins. Additional scaffolds were classified in the version 1 release as repetitive (>95% masked with 24mers that occur more than 4 times in the chromosomes) (815 scaffolds, 30.7 Mb), redundant (>95% masked with 24mers that occur 2 or more times in all scaffolds) (2,220 scaffolds, 85.5 Mb), prokaryote (9 scaffolds, 414.5 Kb), and mitochondria (1 scaffolds, 41.3 Kb).

Genome sequencing of Typha latifolia

The accession of *Typha latifolia*, CWD-2019.6, was collected from the Penn State Arboretum and vouchered by Claude dePamphilis in 2019. We sequenced *Typha latifolia* (var. 2019.6) using a whole genome shotgun sequencing strategy and standard sequencing protocols. Sequencing reads were collected using Illumina and PACBIO platforms. Illumina and PACBIO reads were sequenced at the HudsonAlpha Institute in Huntsville, Alabama. Illumina reads were sequenced using the Illumina NovoSeq6000 platform, and the PACBIO reads were sequenced using the REVIO platform. One 400 bp insert 2x150 Illumina fragment library (261.16x per haplotype) was sequenced along with one 2x150 OmniC library (230.01x per haplotype, **Table S10**). Prior to assembly, Illumina fragment reads were screened for phix contamination. Reads composed of >95% simple sequences were removed. Illumina reads <50 bp after trimming for adapter and quality (q<20) were removed. The final read set consists of 755,162,004 reads for a total of 491.17x of high quality Illumina bases. For the PACBIO sequencing, a total raw sequence yield of 22.4 Gb, with a total coverage of 104.23x per haplotype (**Table S11**).

For genome assembly and construction of pseudomolecule chromosomes, the version 1.0 assemblies were generated by assembling the 1,066,852 PACBIO CCS reads (104.23x per haplotype) using the HiFiAsm assembler[112,113] and subsequently polished using RACON[115]. This produced initial assemblies of both haplotypes. The HAP1 assembly consisted of 410 scaffolds (410 contigs), with a contig N50 of 11.7 Mb, and a total genome size of 235.5 Mb (**Table S12**).

The HAP2 assembly consisted of 421 scaffolds (421 contigs), with a contig N50 of 9.5 Mb, and a total genome size of 233.9 Mb (**Table S13**).

Hi-C Illumina reads from *Typha latifolia* (var. 2019.6), were separately aligned to the HAP1 and HAP2 contig sets with Juicer[117], and chromosome scale scaffolding was performed with 3D-DNA[123]. No misjoins were identified in either the HAP1 or HAP2 assemblies. The contigs were then oriented, ordered, and joined together into 15 chromosomes per haplotype using the HiC data. A total of 9 joins were applied to the HAP1 assembly, and 14 joins for the HAP2 assembly. Chromosomes were numbered by size from largest to smallest with the p-arm to the left. Each chromosome join is padded with 10,000 Ns. Contigs terminating in significant telomeric sequence were identified using the $(TTTAGGG)_n$ repeat, and care was taken to make sure that they were properly oriented in the production assembly. The remaining scaffolds were screened against bacterial proteins, organelle sequences, GenBank nr and removed if found to be a contaminant. After forming the chromosomes, it was observed that some small (<20 Kb) redundant sequences were present on adjacent contig ends within chromosomes. To resolve this issue, adjacent contig ends were aligned to one another using BLAT[100], and duplicate sequences were collapsed to close the gap between them. A total of 2 adjacent contig pairs were collapsed in the HAP1 assembly and 1 in the HAP2 assembly.

Finally, homozygous SNPs and INDELs were corrected in the HAP1 and HAP2 releases using ~60x of Illumina reads (2x150, 400 bp insert) by aligning the reads using bwa-mem[102] and identifying homozygous SNPs and INDELs with the GATK's UnifiedGenotyper tool[107]. A total of 107 homozygous SNPs and 6,519 homozygous INDELs were corrected in the HAP1 release, while a total of 121 homozygous SNPs and 7,009 homozygous INDELs were corrected in the HAP2 release. The final version 1.0 HAP1 release contained 215.3 Mb of sequence, consisting of 30 contigs with a contig N50 of 9.9 Mb and a total of 99.975% of assembled bases integrated into chromosomes (**Table S14**). The final version 1.0 HAP2 release contained 214.7 Mb of sequence, consisting of 20 contigs with a contig N50 of 13.5 Mb and a total of 100% of assembled bases in chromosomes (**Table S15**).

Completeness of the euchromatic portion of the version 1.0 assemblies was assessed using rnaSEQ reads. The aim of this analysis is to obtain a measure of completeness of the assembly, rather than a comprehensive examination of gene space. The reads were aligned to the assembly using bwa-mem2[102]. The screened alignments indicate that 99.82% of the ranaSEQ reads are aligned to the V1 HAP1 release, and 99.83% aligned to the V1 HAP2 release.

Gene tree reconstruction

We identified putative orthogroups for gene models of 39 species (**Table S1**) using OrthoFinder (v.2.5.4)[125] under default settings. Following McKain et al.[3], we filtered the orthogroups to have a minimum of 20 taxa. Amino acids for each orthogroup were aligned using MAFFT (v.7.490)[109] with the "auto" option for algorithm selection. The PAL/PTAL tree and other gene trees were reconstructed on these filtered alignments using IQ-TREE 2 (v.2.2.6)[115] allowing for the optimal model and 1000 bootstrap replicates to be estimated for each alignment. For building a coding sequence tree for the PAL/PTAL orthogroup, we used the aligned amino acid sequences to create a codon alignment of nucleotide sequences using default parameters of PAL2NAL (v.14)[126]. Codon alignments were filtered so that the length of a gene sequence had to be at least 50% of the final alignment length and sequences were not permitted to have more than 30% gaps in the alignment[3]. Then, the tree was generated using RAxML-NG[117] with the evolutionary model TVM+G4 determined by ModelTest-NG[124]. To reconcile the PAL/PTAL gene tree with the

species tree, we used TreeSolve[75] to generate optimal trees. For input, we used the species tree generated from OrthoFinder and generated an amino acid tree for PAL/PTAL genes from these species using RAxML-NG with the evolutionary model JTT+I+G4+F. We also used 100 bootstrap trees generated from RAxML-NG as input.

Identification and placement of polyploidy events in the species phylogeny

A total of 10,055 multicopy gene trees were analyzed to identify potential polyploidy events and place them on the species tree obtained from Timilsena et al.[52] using the PUG software and algorithm[3]. The "estimate_paralogs" option was selected in PUG allowing identification of all possible paralogs within a single gene tree by identifying all gene model pairs from the same taxon. Each multicopy gene tree was re-rooted to a preferred list of outgroups (*Amborella*, *Nymphaea*, *Cinnamon*, *Aquilegia*, *Vitis*, *Arabidopsis*, *Glycine*, *Populus*, *Beta*, *Spinacia*, *Solanum*, *Helianthus*, and *Acorus* in that preferred order. PUG attempts to use the farthest outgroup, as designated by the user, and works down the list until a sequence from that taxon is found. That sequence is then set as the outgroup for the gene tree. PUG identifies the most recent common ancestor node in the gene tree for each potential paralog pair, and the taxa composition of the subtree with the MRCA as the root is used to locate the equivalent node in the species tree. A duplication event placed on the species phylogeny was considered valid if two criteria were met: 1) the taxa above the node matched those in the gene tree and 2) at least one sister taxon to the queried node was found in both the species and gene trees. PUG does not consider potential duplications on tip branches of the species phylogeny, so species-specific duplications are ignored. PUG was run for all gene trees and putative paralogs across all species to identify all potential WGD events. Assessment for putative WGD events was limited to support from orthogroup tree nodes that have a bootstrap support value of 80 or more though we also report values for BSV 50 in the supplemental materials. Unique duplications were mapped on the species tree using the PUG_Figure_Maker.R script (https://github.com/mrmckain/PUG) on R v.4.3.3 and ape v.5.7-1[97]. All branches with duplication counts less than 25% of the largest count on a single branch (3,200) were ignored when mapping.

Synteny analyses

The GENESPACE analysis was conducted using v.1.2.3[55], default parameters, and genome assemblies and annotations for *Ananas cosmosus*, *Joinvillea ascendens*, *Pharus latifolius*, *Oryza sativa*, *Sorghum bicolor*, and *Zea mays* (**Table S1)** as representatives of the Poales with well-scaffolded genomes.

Syntenic analyses of *PAL*/*PTAL* genes were performed using COGE[103] across selected Poales species with the following options: DAGchainer: relative gene order, maximum distance between 2 matches: 20 genes, minimum number of aligned pairs: 5 genes, merged syntenic blocks with quota align, window size 100 genes, use all genes in the target genome, calculate synonymous substitution rates, and tandem duplication distance is 10. Synteny was calculated between the following species: *B. distachyon* (version 556, 3.0) and *J. ascendens* (version 1.1); *B. distachyon* (version 556, 3.0) and *Z. mays* (version PH207 UMN 1.0); *B. distachyon* (version 556, 3.0) and *P. latifolius* (version 1.0); *B. distachyon* (version 556, 3.0) and *S. bicolor* (version 454, 3.0.1); *P. latifolius* (version 1.0) and *S. angustifolia* (version 1.1); *S. angustifolia* (version 1.1) and *J. ascendens* (version 1.1); *J. ascendens* (version 1.1) and *A. comosus* (version 3); *J. ascendens* (version 1.1) and *P. latifolius* (version 1.0).

Identification of the residues involved in the transition from PAL to PTAL

To find positively selected residues, we selected the node in the coding sequence tree of PAL/PTAL containing all PTAL genes including JaPTAL, but excluding JaPAL (red arrow in **Figure 4B**). PAML v4.9 software[80] was then used to determine the positively selected residues of this clade.

Cloning of PAL and PTAL candidates with and without mutations

The coding sequences of PAL and PTAL candidate enzymes from *S. bicolor*, *B. distachyon*, *S. angustifolia* and *J. ascendens* were first amplified by nested PCR using the corresponding cDNA and gene specific primer with PrimeSTAR MAX DNA polymerase (Takara Bio). The PCR fragments were gel-purified and then used as a PCR template for the second PCR with additional primers with in-fusion tag. The obtained PCR products were inserted into the pET28a vector at EcoRI and NdeI sites using the in-fusion HD Cloning Kit (Takara Bio) according to the manufacturer's instruction. The obtained plasmid vectors were submitted for the sequence analysis and confirmed that the coding sequences were matched with the ones in the database. BdPTAL, EmoPTAL, EmoPAL, JaPAL$^{F140H\text{-}MUT8}$, and JaPAL$^{F140H\text{-}MUT16}$ were gene synthesized and cloned into the pET28a vector (SynbioTech). For site-directed mutagenesis, the 1:100 diluted plasmid and mutagenesis primers were used for the PCR amplification with PrimeSTAR MAX DNA polymerase (Takara Bio). The primers used in this study are listed in **Table S16**.

Recombinant protein expression and purification

For the recombinant protein expression, the cloned pET28a vectors were transformed into Rosetta-2 (DE3) *E. coli* and cultured in 3 ml of TB medium containing kanamycin (50 µg/ml), chloramphenicol (34 µg/ml), and 0.1 % glucose at 37 °C and 200 rpm. After the overnight cultivation, 500 µl of pre-culture solution were added to 50 ml TB medium containing the same antibiotics and further cultured at 27 °C and 200 rpm until the OD600 reached 0.5-0.7. After the bacterial cultures were cooled down on ice, isopropyl β-D-1-thiogalactopyranoside (IPTG, 0.5 mM of final concentration) incubated at 22 °C under the constant shaking at 200 rpm. After 24 h, the cultures were harvested by centrifugation (5000 g, 5 min, 4 °C) and the pellets were frozen at -30 °C. The pellets were thawed and resuspended with the lysis buffer containing 50 mM sodium phosphate buffer (pH 8.0), 300 mM NaCl, 10% glycerol, and 0.25 mg of lysozyme. After 30 min incubation on ice, the suspension was sonicated three times for 20 sec and the supernatant was recovered after the centrifugation (12500 g, 20 min, 4 °C). The supernatants were added to the new tube containing 100 µl of Ni-NTA beads (Millipore) and incubate the mixture at 25 °C for 30 min under constant inversion. After the unbound proteins were washed away for three times with washing buffer containing 50 mM sodium phosphate buffer (pH 8.0), 300 mM NaCl, 10% glycerol, and 10 mM imidazole, target proteins were eluted with elution buffer containing 50 mM sodium phosphate buffer (pH 8.0), 300 mM NaCl, 10% glycerol, and 300 mM imidazole. The collected enzyme solutions were desalted by using sephadex G-50 column (GE Healthcare). The protein concentration was determined using the BioRad protein assay dye (BioRad). The purity was confirmed to be >90% by SDS-PAGE and calculated by using ImageJ software.

PAL and TAL enzyme assays

All the substrate solutions were prepared with 0.01N NaOH to increase the solubility of L-Tyr. The mixture containing 100 mM Tris-HCl (pH 8.5), 1% glycerol and the purified enzyme in a total

volume of 50 µl was preincubate for 3 min at 30 °C. PAL and TAL reactions were started by addition of the 50 µl of 1 mM substrate (L-Phe or L-Tyr, respectively) and incubated at 30 °C for 20 min unless otherwise noted. The reaction was terminated by addition of 6N AcOH (10 µl).

The reaction products were analyzed using HPLC (1200 Infinitely Series - Infinitely better, Agilent Technologies) to directly detect the final products of PAL and TAL assay, cinnamic acid and *p*-coumaric acid, respectively. Analytical conditions were as follows: column, Neptune T3 C18 column (3 µm, 2.1 × 150 mm, ES industries); solvent system, solvent A (water including 0.1 % [v/v] formic acid) and solvent B (acetonitrile including 0.1 % [v/v] formic acid); gradient program: 99% A/1% B at 0 min, 99% A/1% B at 4.5 min, 95% A/5% B at 7.5 min, 85% A/15% B at 12 min, 75% A/25% B at 16.5 min, 70% A/30% B at 21 min, 5% A/95% B at 23 min, 5% A/95% B at 26 min, 99% A/5% B at 26.5 min, and 99% A/5% B at 30 min; flow rate: 0.3 mL/min; DAD: 275 nm for cinnamic acid, 309 nm for *p*-coumaric acid.

The kinetics parameters of the recombinant enzymes were determined using HPLC. The mixture contains 100 mM Tris-HCl (pH 8.5), 1% glycerol and purified enzymes (0.15 µg for PAL assay and 1 µg for TAL assay) in a 50 µl total volume were preincubate for 3 min at 30 °C. PAL and TAL reactions were started by addition of 50 µl substrate solution prepared with the concentration range of 0-4 mM for L-Phe and 0-2 mM for L-Tyr. After 10 min and 20 min incubation for PAL and TAL assay, respectively, at 30 °C, the reaction was terminated by addition of 6N AcOH (10 µl). Analytical conditions were as follows: column, Atlantis T3 C18 column (3 µm, 2.1 × 150 mm, Waters); solvent system, solvent A (water including 0.1 % [v/v] formic acid) and solvent B (acetonitrile including 0.1 % [v/v] formic acid); gradient program: 85% A/15% B at 0 min, 85% A/15% B at 1 min, 70% A/30% B at 3 min, 15% A/95% B at 6.5 min, 15% A/95% B at 7.5 min, 85% A/15% B at 8.5 min, and 85% A/15% B at 10 min; flow rate: 0.4 mL/min; DAD: 275 nm for cinnamic acid, 309 nm for *p*-coumaric acid. The products were quantified on the basis of the calibration curves generated with the authentic standards. Non-linear hyperbolic regression analyses were conducted by using the Excel Solver tool to calculate K_m and k_{cat} values.

Protein modeling analysis

The structures of JaPAL and JaPTAL were generated by SWISS-MODEL[134] using homo-tetrameric PAL structure from parsley 6F6T.pdb[82] and homo-dimeric PTAL structure from sorghum 6AT7.pdb[73], respectively, as the templates. The sequence identity against each template were 77.3 % and 80.5 % for JaPAL and JaPTAL, respectively.

Quantification and statistical analysis

Data of the PAL/PTAL enzyme assay are displayed as mean ± SD. Statistical significance was determined by the statistical tests indicated at the figure legend. Student's *t*-test and ANOVA with post hoc Tukey–Kramer test or Kruskal-Wallis nonparametric test were conducted using Microsoft Excel and R, respectively.

RESOURCE AVAILABILITY

Material availability

The PAL and PTAL gene constructs generated in this study are available upon request.

Data and code availability

The raw PacBio and Illumina gDNA sequencing and RNAseq data for *Ecdeiocolea monostachya* is deposited in the NCBI database under BioProject code PRJNA894727. This *Ecdeiocolea monostachya* Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under BioProject codes PRJNA894727 (Illumina/Oxford Nanopore Technology hybrid assembly) and PRJNA1179411 (PacBio HiFi assembly; this version used in this paper). Code not previously published for the high-throughput production of cleaned alignments for phylogenetic reconstruction is available here: https://github.com/mrmckain/Polyploid_Genomics_Utilities. The rest of genomes available in Phytozome as listed in the Resource Table. All other data reported in this paper will be shared by the lead contacts upon request.

Lead contact

- Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Hiroshi A. Maeda <maeda2@wisc.edu>.
- Further information of the genomes of *Joinvillea ascendens*, *Pharus latifolius*, and *Typha latifolia* should be directed to and will be fulfilled by the lead contact, James H Leebens-Mack <jleebensmack@uga.edu>.
- Further information of *Ecdeiocolea monostachya* genome should be directed to and will be fulfilled by the lead contact, Matthew J. Moscou <matthew.moscou@usda.gov>.

REFERENCES

1. Huang, W., Zhang, L., Columbus, J.T., Hu, Y., Zhao, Y., Tang, L., Guo, Z., Chen, W., McKain, M., Bartlett, M., et al. (2022). A well-supported nuclear phylogeny of Poaceae and implications for the evolution of C4 photosynthesis. Mol. Plant *15*, 755–777. https://doi.org/10.1016/j.molp.2022.01.015.
2. Givnish, T.J., Ames, M., McNeal, J.R., McKain, M.R., Steele, P.R., dePamphilis, C.W., Graham, S.W., Pires, J.C., Stevenson, D.W., Zomlefer, W.B., et al. (2010). Assembling the tree of the monocotyledons: plastome sequence phylogeny and evolution of Poales. Ann. Missouri Bot. Gard. 97, 584–616. https://doi.org/10.3417/2010023.
3. McKain, M.R., Tang, H., McNeal, J.R., Ayyampalayam, S., Davis, J.I., dePamphilis, C.W., Givnish, T.J., Pires, J.C., Stevenson, D.Wm., and Leebens-Mack, J.H. (2016). A phylogenomic assessment of ancient polyploidy and genome evolution across the Poales. Genome Biol. Evol. 8, 1150–1164. https://doi.org/10.1093/gbe/evw060.
4. *Websites*: ESS: 9.3 World Calories – Total. https://www.fao.org/economic/the-statistics-division-ess/chartroom-and-factoids/chartroom/93-world-calories-total/en/.
5. Salamini, F., Ozkan, H., Brandolini, A., Schäfer-Pregl, R., and Martin, W. (2002). Genetics and geography of wild cereal domestication in the near east. Nat. Rev. Genet. *3*, 429–441. https://doi.org/10.1038/nrg817.
6. Gutaker, R.M., Groen, S.C., Bellis, E.S., Choi, J.Y., Pires, I.S., Bocinsky, R.K., Slayton, E.R., Wilkins, O., Castillo, C.C., Negrão, S., et al. (2020). Genomic history and ecology of the geographic spread of rice. Nat. Plants *6*, 492–502. https://doi.org/10.1038/s41477-020-0659-6.
7. van Heerwaarden, J., Doebley, J., Briggs, W.H., Glaubitz, J.C., Goodman, M.M., de Jesus Sanchez Gonzalez, J., and Ross-Ibarra, J. (2011). Genetic signals of origin, spread, and introgression in a large sample of maize landraces. Proc. Natl. Acad. Sci. U. S. A. *108*, 1088–1092. https://doi.org/10.1073/pnas.1013011108.
8. Cunniff, J., Wilkinson, S., Charles, M., Jones, G., Rees, M., and Osborne, C.P. (2014). Functional traits differ between cereal crop progenitors and other wild grasses gathered in the Neolithic fertile crescent. PLoS One *9*, e87586. https://doi.org/10.1371/journal.pone.0087586.
9. Glémin, S., and Bataillon, T. (2009). A comparative view of the evolution of grasses under domestication. New Phytol. *183*, 273–290. https://doi.org/10.1111/j.1469-8137.2009.02884.x.
10. Wang, R.L., Stec, A., Hey, J., Lukens, L., and Doebley, J. (1999). The limits of selection during maize domestication. Nature *398*, 236–239. https://doi.org/10.1038/18435.
11. Konishi, S., Izawa, T., Lin, S.Y., Ebana, K., Fukuta, Y., Sasaki, T., and Yano, M. (2006). An SNP caused loss of seed shattering during rice domestication. Science *312*, 1392–1396. https://doi.org/10.1126/science.1126410.
12. Fuller, D.Q., Denham, T., Arroyo-Kalin, M., Lucas, L., Stevens, C.J., Qin, L., Allaby, R.G., and Purugganan, M.D. (2014). Convergent evolution and parallelism in plant domestication revealed by an expanding archaeological record. Proc. Natl. Acad. Sci. U. S. A. *111*, 6147–6152. https://doi.org/10.1073/pnas.1308937110.

13. Lin, Z., Li, X., Shannon, L.M., Yeh, C.-T., Wang, M.L., Bai, G., Peng, Z., Li, J., Trick, H.N., Clemente, T.E., et al. (2012). Parallel domestication of the Shattering1 genes in cereals. Nat. Genet. *44*, 720–724. https://doi.org/10.1038/ng.2281.

14. Wang, Y., Bi, X., and Zhong, J. (2022). Revisiting the origin and identity specification of the spikelet: A structural innovation in grasses (Poaceae). Plant Physiol. *190*, 60–71. https://doi.org/10.1093/plphys/kiac257.

15. Seetharam, A.S., Yu, Y., Bélanger, S., Clark, L.G., Meyers, B.C., Kellogg, E.A., and Hufford, M.B. (2021). The Genome and the Evolution of the Grasses. Front. Plant Sci. *12*, 710383. https://doi.org/10.3389/fpls.2021.710383.

16. Comparot-Moss, S., and Denyer, K. (2009). The evolution of the starch biosynthetic pathway in cereals and other grasses. J. Exp. Bot. *60*, 2481–2492. https://doi.org/10.1093/jxb/erp141.

17. Bhatia, R., Gallagher, J.A., Gomez, L.D., and Bosch, M. (2017). Genetic engineering of grass cell wall polysaccharides for biorefining. Plant Biotechnol. J. *15*, 1071–1092. https://doi.org/10.1111/pbi.12764.

18. Boerjan, W., Ralph, J., and Baucher, M. (2003). Lignin Biosynthesis. Annual Review of Plant Biology *54*, 519–546. https://doi.org/10.1146/annurev.arplant.54.031902.134938.

19. Ragauskas, A.J., Beckham, G.T., Biddy, M.J., Chandra, R., Chen, F., Davis, M.F., Davison, B.H., Dixon, R.A., Gilna, P., Keller, M., et al. (2014). Lignin Valorization: Improving Lignin Processing in the Biorefinery. Science *344*, 1246843–1246843. https://doi.org/10.1126/science.1246843.

20. Rinaldi, R., Jastrzebski, R., Clough, M.T., Ralph, J., Kennema, M., Bruijnincx, P.C.A., and Weckhuysen, B.M. (2016). Paving the Way for Lignin Valorisation: Recent Advances in Bioengineering, Biorefining and Catalysis. Angew. Chem. Int. Ed. 55, 8164–8215 https://doi.org/10.1002/anie.201510351.

21. Umezawa, T. (2018). Lignin modification *in planta* for valorization. Phytochem. Rev., 1–23. https://doi.org/10.1007/s11101-017-9545-x.

22. Rosler, J., Krekel, F., Amrhein, N., and Schmid, J. (1997). Maize phenylalanine ammonia-lyase has tyrosine ammonia-lyase activity. Plant Physiol. *113*, 175–179.

23. Barros, J., Serrani-Yarce, J.C., Chen, F., Baxter, D., Venables, B.J., and Dixon, R.A. (2016). Role of bifunctional ammonia-lyase in grass cell wall biosynthesis. Nat. Plants *2*, 16050. https://doi.org/10.1038/nplants.2016.50.

24. Neish, A.C. (1961). Formation of *m*- and *p*-coumaric acids by enzymatic deamination of the corresponding isomers of tyrosine. Phytochemistry *1*, 1–24. https://doi.org/10.1016/S0031-9422(00)82806-X.

25. Young, M.R., and Neish, A.C. (1966). Properties of the ammonia-lyases deaminating phenylalanine and related compounds in *Triticum aestivum* and *Pteridium aquilinum*. Phytochemistry *5*, 1121–1132. https://doi.org/10.1016/S0031-9422(00)86105-1.

26. Higuchi, T., and Shimada, M. (1969). Metabolism of phenylalanine and tyrosine during lignification of bamboos. Phytochemistry *8*, 1185–1192. https://doi.org/10.1016/S0031-9422(00)85555-7.

27. Havir, E.A., Reid, P.D., and Marsh, H.V., Jr. (1971). L-Phenylalanine ammonia-lyase (Maize): evidence for a common catalytic site for L-phenylalanine and L-tyrosine 1. Plant

Physiol. *48*, 130–136. https://doi.org/10.1104/pp.48.2.130.

28. Jangaard, N.O. (1974). The characterization of phenylalanine ammonia-lyase from several plant species. Phytochemistry *13*, 1765–1768. https://doi.org/10.1016/0031-9422(74)85086-7.
29. Van de Peer, Y., Ashman, T.-L., Soltis, P.S., and Soltis, D.E. (2021). Polyploidy: an evolutionary and ecological force in stressful times. Plant Cell *33*, 11–26. https://doi.org/10.1093/plcell/koaa015.
30. Almeida-Silva, F., and Van de Peer, Y. (2023). Whole-genome duplications and the long-term evolution of gene regulatory networks in angiosperms. Mol. Biol. Evol. *40*. https://doi.org/10.1093/molbev/msad141.
31. Walden, N., German, D.A., Wolf, E.M., Kiefer, M., Rigault, P., Huang, X.-C., Kiefer, C., Schmickl, R., Franzke, A., Neuffer, B., et al. (2020). Nested whole-genome duplications coincide with diversification and high morphological disparity in Brassicaceae. Nat. Commun. *11*, 3795. https://doi.org/10.1038/s41467-020-17605-7.
32. Moharana, K.C., and Venancio, T.M. (2020). Polyploidization events shaped the transcription factor repertoires in legumes (Fabaceae). Plant J. *103*, 726–741. https://doi.org/10.1111/tpj.14765.
33. Zhang, T., Huang, W., Zhang, L., Li, D.-Z., Qi, J., and Ma, H. (2024). Phylogenomic profiles of whole-genome duplications in Poaceae and landscape of differential duplicate retention and losses among major Poaceae lineages. Nat. Commun. *15*, 3305. https://doi.org/10.1038/s41467-024-47428-9.
34. Clark, J.W., and Donoghue, P.C.J. (2018). Whole-genome duplication and plant macroevolution. Trends Plant Sci. *23*, 933–945. https://doi.org/10.1016/j.tplants.2018.07.006.
35. Jacobowitz, J.R., Doyle, W.C., and Weng, J.-K. (2019). PRX9 and PRX40 are extensin peroxidases essential for maintaining tapetum and microspore cell wall integrity during Arabidopsis anther development. Plant Cell *31*, 848–861. https://doi.org/10.1105/tpc.18.00907.
36. Michelangeli, F.A., Davis, J.I., and Stevenson, D.Wm. (2003). Phylogenetic relationships among Poaceae and related families as inferred from morphology, inversions in the plastid genome, and sequence data from the mitochondrial and plastid genomes. Am. J. Bot. *90*, 93–106. https://doi.org/10.3732/ajb.90.1.93.
37. Paterson, A.H., Bowers, J.E., and Chapman, B.A. (2004). Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proc. Natl. Acad. Sci. U.S.A. *101*, 9903–9908. https://doi.org/10.1073/pnas.0307901101.
38. Yu, J., Wang, J., Lin, W., Li, S., Li, H., Zhou, J., Ni, P., Dong, W., Hu, S., Zeng, C., et al. (2005). The genomes of *Oryza sativa*: a history of duplications. PLOS Biol. *3*, e38. https://doi.org/10.1371/journal.pbio.0030038.
39. Tang, H., Bowers, J.E., Wang, X., and Paterson, A.H. (2010). Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. Proc. Natl. Acad. Sci. U. S. A. *107*, 472–477. https://doi.org/10.1073/pnas.0908007107.
40. *An entire book:* Ohno, S. (1970) Evolution by Gene Duplication. (Springer)

41. Vanneste, K., Baele, G., Maere, S., and Van de Peer, Y. (2014). Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary. Genome Res. *24*, 1334–1347. https://doi.org/10.1101/gr.168997.113.

42. Wu, Y., Zhu, Z., Ma, L., and Chen, M. (2008). The preferential retention of starch synthesis genes reveals the impact of whole-genome duplication on grass evolution. Mol. Biol. Evol. *25*, 1003–1006. https://doi.org/10.1093/molbev/msn052.

43. Goff, S.A., Ricke, D., Lan, T.-H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., et al. (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). Science *296*, 92–100. https://doi.org/10.1126/science.1068275.

44. Yu, J. (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). Science *296*, 79–92. https://doi.org/10.1126/science.1068037.

45. Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. Science *326*, 1112–1115. https://doi.org/10.1126/science.1178534.

46. International Wheat Genome Sequencing Consortium (IWGSC) (2014). A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. Science *345*, 1251788. https://doi.org/10.1126/science.1251788.

47. Linder, H.P., Lehmann, C.E.R., Archibald, S., Osborne, C.P., and Richardson, D.M. (2018). Global grass (Poaceae) success underpinned by traits facilitating colonization, persistence and habitat transformation. Biol. Rev. Camb. Philos. Soc. *93*, 1125–1144. https://doi.org/10.1111/brv.12388.

48. Hodkinson, T.R. (2018). Evolution and taxonomy of the grasses (Poaceae): A model family for the study of species-rich groups. Annu. Plant Rev., 1–39. https://doi.org/10.1002/9781119312994.apr0622.

49. Whipple, C.J. (2017). Grass inflorescence architecture and evolution: the origin of novel signaling centers. New Phytol. *216*, 367–372. https://doi.org/10.1111/nph.14538.

50. Kellogg, E.A. (2022). Genetic control of branching patterns in grass inflorescences. Plant Cell *34*, 2518–2533. https://doi.org/10.1093/plcell/koac080.

51. McSteen, P., and Kellogg, E.A. (2022). Molecular, cellular, and developmental foundations of grass diversity. Science *377*, 599–602. https://doi.org/10.1126/science.abo5035.

52. Timilsena, P.R., Wafula, E.K., Barrett, C.F., Ayyampalayam, S., McNeal, J.R., Rentsch, J.D., McKain, M.R., Heyduk, K., Harkess, A., Villegente, M., et al. (2022). Phylogenomic resolution of order- and family-level monocot relationships using 602 single-copy nuclear genes and 1375 BUSCO genes. Front. Plant Sci. *13*, 876779. https://doi.org/10.3389/fpls.2022.876779.

53. Elliott, T.L., Spalink, D., Larridon, I., Zuntini, A.R., Escudero, M., Hackel, J., Barrett, R.L., Martín-Bravo, S., Márquez-Corro, J.I., Granados Mendoza, C., et al. (2024). Global analysis of Poales diversification - parallel evolution in space and time into open and closed habitats. New Phytol. *242*, 727–743. https://doi.org/10.1111/nph.19421.

54. Givnish, T.J., Zuluaga, A., Spalink, D., Soto Gomez, M., Lam, V.K.Y., Saarela, J.M., Sass, C., Iles, W.J.D., de Sousa, D.J.L., Leebens-Mack, J., et al. (2018). Monocot plastid phylogenomics, timeline, net rates of species diversification, the power of multi-gene

analyses, and a functional model for the origin of monocots. Am. J. Bot. *105*, 1888–1910. https://doi.org/10.1002/ajb2.1178.

55. Hanson, L., Boyd, A., Johnson, M.A.T., and Bennett, M.D. (2005). First nuclear DNA C-values for 18 eudicot families. Ann. Bot. *96*, 1315–1320. https://doi.org/10.1093/aob/mci283.
56. Vaser, R., Sović, I., Nagarajan, N., and Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. *27*, 737–746. https://doi.org/10.1101/gr.214270.116.
57. Ma, P.-F., Liu, Y.-L., Jin, G.-H., Liu, J.-X., Wu, H., He, J., Guo, Z.-H., and Li, D.-Z. (2021). The Pharus latifolius genome bridges the gap of early grass evolution. Plant Cell *33*, 846–864. https://doi.org/10.1093/plcell/koab015.
58. Lovell, J.T., Sreedasyam, A., Schranz, M.E., Wilson, M., Carlson, J.W., Harkess, A., Emms, D., Goodstein, D.M., and Schmutz, J. (2022). GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. Elife *11*. https://doi.org/10.7554/eLife.78526.
59. Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M., and Paterson, A.H. (2008). Synteny and collinearity in plant genomes. Science *320*, 486–488. https://doi.org/10.1126/science.1153917.
60. Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A., et al. (2009). The *Sorghum bicolor* genome and the diversification of grasses. Nature *457*, 551–556. https://doi.org/10.1038/nature07723.
61. Jiao, Y., Li, J., Tang, H., and Paterson, A.H. (2014). Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in monocots. Plant Cell *26*, 2792–2802. https://doi.org/10.1105/tpc.114.127597.
62. *Article on a preprint server or other repository:* McKain, M.R., Estep, M.C., Pasquet, R., Layton, D.J., Díaz, D.M.V., Zhong, J., Hodge, J.G., Malcomber, S.T., Chipabika, G., Pallangyo, B., et al. (2018). Ancestry of the two subgenomes of maize. Preprint at bioRxiv, https://doi.org/10.1101/352351 https://doi.org/10.1101/352351.
63. Deb, S.K., Edger, P.P., Pires, J.C., and McKain, M.R. (2023). Patterns, mechanisms, and consequences of homoeologous exchange in allopolyploid angiosperms: a genomic and epigenomic perspective. New Phytol. *238*, 2284–2304. https://doi.org/10.1111/nph.18927.
64. Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature *449*, 463–467. https://doi.org/10.1038/nature06148.
65. One Thousand Plant Transcriptomes Initiative (2019). One thousand plant transcriptomes and the phylogenomics of green plants. Nature *574*, 679–685. https://doi.org/10.1038/s41586-019-1693-2.
66. Zeeman, S.C., Kossmann, J., and Smith, A.M. (2010). Starch: its metabolism, evolution, and biotechnological modification in plants. Annu. Rev. Plant Biol. *61*, 209–234. https://doi.org/10.1146/annurev-arplant-042809-112301.
67. MacNeill, G.J., Mehrpouyan, S., Minow, M.A.A., Patterson, J.A., Tetlow, I.J., and Emes, M.J. (2017). Starch as a source, starch as a sink: the bifunctional role of starch in carbon

allocation. J. Exp. Bot. *68*, 4433–4453. https://doi.org/10.1093/jxb/erx291.

68. Beckles, D.M., Smith, A.M., and ap Rees, T. (2001). A cytosolic ADP-glucose pyrophosphorylase is a feature of graminaceous endosperms, but not of other starch-storing organs. Plant Physiol. *125*, 818–827. https://doi.org/10.1104/pp.125.2.818.
69. Pfister, B., and Zeeman, S.C. (2016). Formation of starch in plant cells. Cell. Mol. Life Sci. *73*, 2781–2807. https://doi.org/10.1007/s00018-016-2250-x.
70. Huang, B., Hennen-Bierwagen, T.A., and Myers, A.M. (2014). Functions of multiple genes encoding ADP-glucose pyrophosphorylase subunits in maize endosperm, embryo, and leaf. Plant Physiol. *164*, 596–611. https://doi.org/10.1104/pp.113.231605.
71. Patron, N.J., Greber, B., Fahy, B.F., Laurie, D.A., Parker, M.L., and Denyer, K. (2004). The lys5 mutations of barley reveal the nature and importance of plastidial ADP-glc transporters for starch synthesis in cereal endosperm. Plant Physiol. *135*, 2088–2097. https://doi.org/10.1104/pp.104.045203.
72. Konishi, T., and Sasaki, Y. (1994). Compartmentalization of two forms of acetyl-CoA carboxylase in plants and the origin of their tolerance toward herbicides. Proc. Natl. Acad. Sci. U.S.A. *91*, 3598–3601. https://doi.org/10.1073/pnas.91.9.3598.
73. SASAKI, Y., and NAGANO, Y. (2004). Plant acetyl-CoA carboxylase: structure, biosynthesis, regulation, and gene manipulation for plant breeding. Biosci. Biotechnol. Biochem. *68*, 1175–1184. https://doi.org/10.1271/bbb.68.1175.
74. Délye, C., Zhang, X.-Q., Michel, S., Matéjicek, A., and Powles, S.B. (2005). Molecular bases for sensitivity to acetyl-coenzyme A carboxylase inhibitors in black-grass. Plant Physiol. *137*, 794–806. https://doi.org/10.1104/pp.104.046144.
75. *Article in a book*: Kordi, M., and Bansal, M.S. (2020). TreeSolve: rapid error-correction of microbial gene trees. In Algorithms for Computational Biology, C. Martín-Vide, M. A. Vega-Rodríguez, and T. Wheeler, eds. (Springer International Publishing), pp. 125–139. https://doi.org/10.1007/978-3-030-42266-0_10.
76. Watts, K.T., Lee, P.C., and Schmidt-Dannert, C. (2006). Biosynthesis of plant-specific stilbene polyketides in metabolically engineered Escherichia coli. BMC biotechnol. *6*, 22.
77. Jun, S.-Y., Sattler, S.A., Cortez, G.S., Vermerris, W., Sattler, S.E., and Kang, C. (2018). Biochemical and structural analysis of substrate specificity of a phenylalanine ammonia-lyase. Plant Physiol. *176*, 1452–1468. https://doi.org/10.1104/pp.17.01608.
78. Louie, G.V., Bowman, M.E., Moffitt, M.C., Baiga, T.J., Moore, B.S., and Noel, J.P. (2006). Structural determinants and modulation of substrate specificity in phenylalanine-tyrosine ammonia-lyases. Chem. Biol. *13*, 1327–1338. https://doi.org/10.1016/j.chembiol.2006.11.011.
79. Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. *13*, 555–556. https://doi.org/10.1093/bioinformatics/13.5.555.
80. Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. *24*, 1586–1591. https://doi.org/10.1093/molbev/msm088.
81. Maeda, H.A. (2019). Harnessing evolutionary diversification of primary metabolism for plant synthetic biology. J. Biol. Chem. *294*, 16549–16566.

https://doi.org/10.1074/jbc.REV119.006132.

82. Bata, Z., Molnár, Z., Madaras, E., Molnár, B., Sánta-Bell, E., Varga, A., Leveles, I., Qian, R., Hammerschmidt, F., Paizs, C., et al. (2021). Substrate tunnel engineering aided by X-ray crystallography and functional dynamics swaps the function of MIO-enzymes. ACS Catal. *11*, 4538–4549. https://doi.org/10.1021/acscatal.1c00266.

83. Röther, D., Poppe, L., Morlock, G., Viergutz, S., and Rétey, J. (2002). An active site homology model of phenylalanine ammonia-lyase from *P. crispum*. Eur. J. Biochem. *269*, 3065–3075. https://doi.org/10.1046/j.1432-1033.2002.02984.x.

84. Cochrane, F.C., Davin, L.B., and Lewis, N.G. (2004). The Arabidopsis phenylalanine ammonia lyase gene family: kinetic characterization of the four PAL isoforms. Phytochemistry *65*, 1557–1564. https://doi.org/10.1016/j.phytochem.2004.05.006.

85. Barker, M.S., Arrigo, N., Baniaga, A.E., Li, Z., and Levin, D.A. (2016). On the relative abundance of autopolyploids and allopolyploids. New Phytol. *210*, 391–398. https://doi.org/10.1111/nph.13698.

86. Schranz, M.E., Mohammadin, S., and Edger, P.P. (2012). Ancient whole genome duplications, novelty and diversification: the WGD radiation lag-time model. Curr. Opin. Plant Biol. *15*, 147–153. https://doi.org/10.1016/j.pbi.2012.03.011.

87. Preston, J.C., Christensen, A., Malcomber, S.T., and Kellogg, E.A. (2009). MADS-box gene expression and implications for developmental origins of the grass spikelet. Am. J. Bot. *96*, 1419–1429. https://doi.org/10.3732/ajb.0900062.

88. Barros, J., and Dixon, R.A. (2020). Plant phenylalanine/tyrosine ammonia-lyases. Trends Plant Sci. *25*, 66-79. https://doi.org/10.1016/j.tplants.2019.09.011.

89. Maeda, H.A. (2016). Lignin biosynthesis: tyrosine shortcut in grasses. Nat. Plants *2*, 1–2. https://doi.org/10.1038/nplants.2016.80.

90. Amthor, J.S. (2003). Efficiency of lignin biosynthesis: a quantitative analysis. Ann. Bot. *91*, 673–695. https://doi.org/10.1093/aob/mcg073.

91. El-Azaz, J., Moore, B., Takeda-Kimura, Y., Yokoyama, R., Wijesingha Ahchige, M., Chen, X., Schneider, M., and Maeda, H.A. (2023). Coordinated regulation of the entry and exit steps of aromatic amino acid biosynthesis supports the dual lignin pathway in grasses. Nat. Commun. *14*, 7242. https://doi.org/10.1038/s41467-023-42587-7.

92. Tye, Y.Y., Lee, K.T., Wan Abdullah, W.N., and Leh, C.P. (2016). The world availability of non-wood lignocellulosic biomass for the production of cellulosic ethanol and potential pretreatments for the enhancement of enzymatic saccharification. Renew. Sustain. Energy Rev. *60*, 155–172. https://doi.org/10.1016/j.rser.2016.01.072.

93. Lopez-Nieves, S., El-Azaz, J., Men, Y., Holland, C.K., Feng, T., Brockington, S.F., Jez, J.M., and Maeda, H.A. (2022). Two independently evolved natural mutations additively deregulate TyrA enzymes and boost tyrosine production in planta. Plant J. *109*, 844–855. https://doi.org/10.1111/tpj.15597.

94. Yokoyama, R., de Oliveira, M.V.V., Takeda-Kimura, Y., Ishihara, H., Alseekh, S., Arrivault, S., Kukshal, V., Jez, J.M., Stitt, M., Fernie, A.R., et al. (2022). Point mutations that boost aromatic amino acid production and CO_2 assimilation in plants. Sci. Advances *8*, eabo3416. https://doi.org/10.1126/sciadv.abo3416.

95. Chimpanzee Sequencing and Analysis Consortium (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. Nature *437*, 69–87. https://doi.org/10.1038/nature04072.

96. Prüfer, K., Munch, K., Hellmann, I., Akagi, K., Miller, J.R., Walenz, B., Koren, S., Sutton, G., Kodira, C., Winer, R., et al. (2012). The bonobo genome compared with the chimpanzee and human genomes. Nature *486*, 527–531. https://doi.org/10.1038/nature11128.

97. Paradis, E., and Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics *35*, 526–528. https://doi.org/10.1093/bioinformatics/bty633.

98. Keller, O., Kollmar, M., Stanke, M., and Waack, S. (2011). A novel hybrid gene prediction method employing protein multiple sequence alignments. Bioinformatics *27*, 757–763. https://doi.org/10.1093/bioinformatics/btr010.

99. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421. https://doi.org/10.1186/1471-2105-10-421.

100. Kent, W.J. (2002). BLAT—The BLAST-Like Alignment Tool. Genome Res. *12*, 656–664. https://doi.org/10.1101/gr.229202.

101. Manni, M., Berkeley, M.R., Seppey, M., Simão, F.A., and Zdobnov, E.M. (2021). BUSCO Update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol. Biol. Evol. *38*, 4647–4654. https://doi.org/10.1093/molbev/msab199.

102. *Article on a preprint server or other repository:* Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at arXiv, https://doi.org/10.48550/arXiv.1303.3997.

103. Lyons, E., and Freeling, M. (2008). How to usefully compare homologous plant genes and chromosomes as DNA sequences. Plant J. *53*, 661–673. https://doi.org/10.1111/j.1365-313X.2007.03326.x.

104. Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protocol *7*, 562–578. https://doi.org/10.1038/nprot.2012.016.

105. Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R., and Wortman, J.R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. Genome Biol. *9*, R7. https://doi.org/10.1186/gb-2008-9-1-r7.

106. Sun, H., Ding, J., Piednoël, M., and Schneeberger, K. (2018). findGSE: estimating genome size variation within human and Arabidopsis using *k*-mer frequencies. Bioinformatics *34*, 550–557. https://doi.org/10.1093/bioinformatics/btx637.

107. *An entire book:* Auwera, G. van der, and O'Connor, B.D. (2020). Genomics in the Cloud: Using Docker, GATK, and WDL in Terra, van der Auwera, G.A. and O'Connor. B.D., eds. (O'Reilly). https://www.oreilly.com/library/view/genomics-in-the/9781491975183/

108. *Article in a book*: Keilwagen, J., Hartung, F., and Grau, J. (2019). GeMoMa: homology-based

gene prediction utilizing intron position conservation and RNA-seq data. In Gene Prediction: Methods and Protocols, M. Kollmar, ed. (Springer), pp. 161–177. https://doi.org/10.1007/978-1-4939-9173-0_9.

109. Tang, S., Lomsadze, A., and Borodovsky, M. (2015). Identification of protein coding regions in RNA transcripts. Nucleic Acids Research *43*, e78. https://doi.org/10.1093/nar/gkv227.

110. Vurture, G.W., Sedlazeck, F.J., Nattestad, M., Underwood, C.J., Fang, H., Gurtowski, J., and Schatz, M.C. (2017). GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics *33*, 2202–2204. https://doi.org/10.1093/bioinformatics/btx153.

111. Wu, T.D., and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics *26*, 873–881. https://doi.org/10.1093/bioinformatics/btq057.

112. Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., and Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat. Methods *18*, 170–175. https://doi.org/10.1038/s41592-020-01056-5.

113. Cheng, H., Jarvis, E.D., Fedrigo, O., Koepfli, K.-P., Urban, L., Gemmell, N.J., and Li, H. (2022). Haplotype-resolved assembly of diploid genomes without parental data. Nat. Biotechnol *40*, 1332–1335. https://doi.org/10.1038/s41587-022-01261-x.

114. Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol *37*, 907–915. https://doi.org/10.1038/s41587-019-0201-4.

115. Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. *37*, 1530–1534. https://doi.org/10.1093/molbev/msaa015.

116. Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics *27*, 764–770. https://doi.org/10.1093/bioinformatics/btr011.

117. Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S.P., Huntley, M.H., Lander, E.S., and Aiden, E.L. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cells *3*, 95–98. https://doi.org/10.1016/j.cels.2016.07.002.

118. *Article on a preprint server or other repository:* Chikhi, R., and Medvedev, P. (2013). Informed and automated k-mer size selection for genome assembly. arXiv.org. https://arxiv.org/abs/1304.5665v1.

119. Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J., and Clavijo, B.J. (2017). KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. Bioinformatics *33*, 574–576. https://doi.org/10.1093/bioinformatics/btw663.

120. Wood, D.E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken. Genome Biol. *20*, 257. https://doi.org/10.1186/s13059-019-1891-0.

121. Campbell, M.S., Holt, C., Moore, B., and Yandell, M. (2014). Genome annotation and curation using MAKER and MAKER-P. Curr. Protoc. Bioinformatics *48*, 4.11.1-4.11.39. https://doi.org/10.1002/0471250953.bi0411s48.

122. Zimin, A.V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S.L., and Yorke, J.A. (2013). The

MaSuRCA genome assembler. Bioinformatics *29*, 2669–2677. https://doi.org/10.1093/bioinformatics/btt476.

123. Xiao, C.-L., Chen, Y., Xie, S.-Q., Chen, K.-N., Wang, Y., Han, Y., Luo, F., and Xie, Z. (2017). MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. Nat. Methods *14*, 1072–1074. https://doi.org/10.1038/nmeth.4432.

124. Darriba, D., Posada, D., Kozlov, A.M., Stamatakis, A., Morel, B., and Flouri, T. (2020). ModelTest-NG: A new and scalable tool for the selection of DNA and protein evolutionary models. Mol. Biol. Evol. *37*, 291–294. https://doi.org/10.1093/molbev/msz189.

125. Emms, D.M., and Kelly, S. (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. *16*, 157. https://doi.org/10.1186/s13059-015-0721-2.

126. Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucl. Acids Res. *34*, W609–W612. https://doi.org/10.1093/nar/gkl315.

127. Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith Jr, R.K., Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D., et al. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucl. Acids Res. *31*, 5654–5666. https://doi.org/10.1093/nar/gkg770.

128. *Websites*: Shu, S., Goodstein, D., and Rokhsar, D. (2013). PERTRAN: Genome-guided RNA-seq Read Assembler. https://www.osti.gov/biblio/1241180

129. Guan, D., McCarthy, S.A., Wood, J., Howe, K., Wang, Y., and Durbin, R. (2020). Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics *36*, 2896–2898. https://doi.org/10.1093/bioinformatics/btaa025.

130. Kozlov, A.M., Darriba, D., Flouri, T., Morel, B., and Stamatakis, A. (2019). RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics *35*, 4453–4455. https://doi.org/10.1093/bioinformatics/btz305.

131. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., et al. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008. https://doi.org/10.1093/gigascience/giab008.

132. Korf, I. (2004). Gene finding in novel genomes. BMC Bioinformatics *5*, 59. https://doi.org/10.1186/1471-2105-5-59.

133. Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol *33*, 290–295. https://doi.org/10.1038/nbt.3122.

134. Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., de Beer, T.A.P., Rempfer, C., Bordoli, L., et al. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. Nucl. Acids Res. *46*, W296–W303. https://doi.org/10.1093/nar/gky427.

135. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

136. Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis,

X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. *29*, 644–652. https://doi.org/10.1038/nbt.1883.

137. Dudchenko, O., Batra, S.S., Omer, A.D., Nyquist, S.K., Hoeger, M., Durand, N.C., Shamim, M.S., Machol, I., Lander, E.S., Aiden, A.P., et al. (2017). De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science *356*, 92–95. https://doi.org/10.1126/science.aal3327.

138. Sebastian, J., Wong, M.K., Tang, E., and Dinneny, J.R. (2014). Methods to promote germination of dormant *Setaria viridis* seeds. PLOS ONE *9*, e95109. https://doi.org/10.1371/journal.pone.0095109.

139. Salinas, J., Matassi, G., Montero, L.M., and Bernardi, G. (1988). Compositional compartmentalization and compositional patterns in the nuclear genomes of plants. Nucl. Acids Res. *16*, 4269–4285. https://doi.org/10.1093/nar/16.10.4269.

140. Stewart, C.N., and Via, L.E. (1993). A rapid CTAB DNA isolation technique useful for RAPD fingerprinting and other PCR applications. Biotechniques *14*, 748–750.