Identification of new leaf intrinsic yield genes using cross-species network analysis in plants

Pasquale Luca Curci[1,2,3], Jie Zhang[1,2+], Niklas Mähler[4+], Carolin Seyfferth[1,2,4], Chanaka Mannapperuma[4], Tim Diels[1,2], Tom Van Hautegem[1,2], David Jonsen[6], Nathaniel Street[4], Torgeir R. Hvidsten[4,7], Magnus Hertzberg[6], Ove Nilsson[5], Dirk Inze[1,2], Hilde Nelissen[1,2], Klaas Vandepoele[1,2,8*]

(1) Department of Plant Biotechnology and Bioinformatics, Ghent University, Technologiepark 71, 9052 Ghent, Belgium

(2) VIB Center for Plant Systems Biology, Technologiepark 71, 9052 Ghent, Belgium

(3) Institute of Biosciences and Bioresources, National Research Council (CNR), Via Amendola 165/A, 70126 Bari, Italy

(4) Umea Plant Science Centre (UPSC), Department of Plant Physiology, Umeå University, 90187 Umeå, Sweden.

(5) Umea Plant Science Centre (UPSC), Department of Forest Genetics and Plant Physiology, Swedish University of Agricultural Sciences, 90183 Umeå, Sweden.

(6) SweTree Technologies AB, Skogsmarksgränd 7, SE-907 36 Umeå, Sweden

(7) Faculty of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, 1432 Ås, Norway

(8) Bioinformatics Institute Ghent, Ghent University, Technologiepark 71, 9052 Ghent, Belgium

+ both authors contributed equally

* corresponding authors: Klaas.Vandepoele@psb.vib-ugent.be

Short title: Identification of new intrinsic yield genes in plants

The author(s) responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) is: Klaas Vandepoele (Klaas.Vandepoele@psb.vib-ugent.be).

Abstract

Plant leaves differ in their size, form and structure, and the processes of cell division and cell expansion contribute to this diversity. Leaf transcriptional networks covering cell division and cell expansion in *Arabidopsis thaliana*, maize (*Zea mays*) and aspen (*Populus tremula*) were compared to identify candidate genes that are conserved in plant growth and ultimately have the potential to increase biomass (intrinsic yield, IY). Our approach revealed that genes showing strongly conserved co-expression were mainly involved in fundamental leaf developmental processes such as photosynthesis, translation, and cell proliferation. Next, known intrinsic yield genes (IYGs) together with cross-species conserved networks were used to predict novel potential Arabidopsis leaf IYGs. Using an in-depth literature screening, 34 out of 100 top predicted IYGs were confirmed to affect leaf phenotype if mutated or overexpressed and thus represent novel potential IYGs. Globally, these new IYGs were involved in processes mostly covering cell cycle, plant defense responses, gibberellin, auxin and brassinosteroid signaling. Application of loss-of-function lines and phenotypic characterization confirmed two newly predicted IYGs to be involved in leaf growth (*NPF6.4* and *LATE MERISTEM IDENTITY2*). In conclusion, the presented network approach offers an integrative cross-species strategy to identify new yield genes and to accelerate plant breeding.

Introduction

New plant organs are formed and then grow continuously throughout development. Upon adverse conditions, growth adjustments are among the first plant responses, rendering growth regulation an important yield component (Gray and Brady, 2016; Nowicka, 2019). The growth of plants involves complex mechanisms controlling processes from the cellular to the whole-organism level. Cell division and cell expansion are the two major processes regulating leaf growth and previous research has shown that largely similar cellular and molecular pathways govern these fundamental processes in dicots and monocots (Anastasiou et al., 2007; Nelissen et al., 2016). Numerous genes have been identified that when mutated or ectopically expressed increase organ size, such as leaf size, in plants. These so-called 'intrinsic yield genes' (IYGs) are part of functional modules of genes/proteins that govern sub-processes that constitute organ growth. Many of these genes are functionally conserved across plant species and some of these genes promote, when mutated or ectopically expressed, organ growth in both dicots and monocots. Notable examples are genes encoding CYP78A, ARGOS, rate limiting GA biosynthesis enzymes, BRI1, ANGUSTIFOLIA3 and GROWTH-REGULATING FACTORS (Powell and Lenhard, 2012; Vercruysse et al., 2020).

However, the complex and highly dynamic nature of the regulatory networks controlling such complex traits makes the identification of new growth regulatory genes challenging (Baxter, 2020). Moreover, duplication events across the plant kingdom have caused a general enlargement of gene families and, with it, plant- and tissue-specific functional specialization (Jones and Vandepoele, 2020). Gene orthology information is essential to transfer functional annotations from model plants with high quality annotations (e.g. *Arabidopsis thaliana*) to other species. Functional annotations derived from experimental evidence can be used to identify relevant orthologs and drive gene function discovery in crops (Lee et al., 2015, 2019). This approach is not straightforward, mainly for two reasons: first, the orthology approach normally leads to the identification of complex (one-to-one, one-to-many and many-to-many) orthology relationships (Movahedi et al., 2011; van Bel et al., 2012); second, for genes with multiple orthologs, it has been observed that the closest ortholog in terms of protein sequence similarity is often not the closest ortholog in terms of regulation, indicating that identifying functionally conserved orthologs is challenging (Patel et al., 2012; Netotea et al., 2014).

Biological networks offer the means to study the complex organization of gene interactions. Densely connected network clusters form gene modules, defined as groups of linked genes with similar expression profiles (i.e. co-expressed genes), which also tend to be co-regulated and functionally related (Heyndrickx and Vandepoele, 2012; Klie et al., 2012). Although transferring network links from better annotated species to crops is the most intuitive approach and has proven to be helpful (Ficklin and Feltus, 2011; Obertello et al., 2015), it has been shown that only ~20-40% of the co-expression links are conserved in pairwise comparison of *Arabidopsis thaliana* (Arabidopsis), *Populus*, and *Oryza sativa* (Netotea et al., 2014). On the other hand, it has been shown that using gene modules that are conserved across species can increase the amount of biological knowledge transferred from one species to another (Mutwil et al., 2011; Heyndrickx and Vandepoele, 2012; Cheng et al., 2021). Such conserved gene modules mirror biological processes conserved across species, meaning that the orthologous genes present in these modules are involved in the same process and potentially perform the same function (Ruprecht et al., 2011). Significantly conserved cross-species modules (with many shared orthologs) can be used to transfer gene function annotations and analyze expression conservation for paralogs involved in complex many-to-many orthology relationships. A guilt-by-association approach can also then be used to infer functions of unknown genes from the functions of co-expressed annotated genes (Wolfe et al., 2005; Lee et al., 2010; De Smet and Marchal, 2010; Klie et al., 2012; Rhee and Mutwil, 2014).

Here, we focused on high-resolution leaf transcriptomes covering cell proliferation and expansion in three plant species: two dicotyledonous plants, the annual plant Arabidopsis and the perennial plant *Populus tremula* (aspen), and one monocotyledonous plant, *Zea mays* (maize). We constructed aggregated co-expression networks for each species, performed cross-species comparison of these leaf development transcriptional networks and used conserved gene neighborhoods to predict and validate new IYGs in Arabidopsis with the potential to increase organ growth.

Results

Network construction and gene neighborhood conservation analysis

To perform network construction based on gene expression information, expression compendia were built for Arabidopsis, maize and aspen that contained a minimum of 24 leaf samples (Fig. 1, step 1; Table S1; Supplemental Methods). These expression compendia all include developmental stages with active cell proliferation and cell expansion. The Arabidopsis expression compendium was composed of three main developmental phases: cell proliferation, cell expansion and the transition between these two phases. For maize, the developmental expression compendium included a newly-generated high-resolution dataset and covers cell proliferation, cell expansion and mature phases of development (Supplemental Methods). For aspen, samples covered the developmental stages ranging from the very youngest leaf primordia to fully expanded and mature leaves. In total, expression data covered 20,313 genes for Arabidopsis, 29,383 genes for maize, and 35,309 genes for aspen (Dataset 1).

The network construction was performed for each species with Seidr, a toolkit to perform multiple gene network inferences and combine their results into a unified meta-network (Schiffthaler et al., 2018). For each network inference algorithm included, a fully connected weighted gene network was constructed. These were in turn aggregated into a weighted meta-network (Fig. 1, step 2). When applying a weight threshold, the network density was defined as the ratio between the number of links with a weight higher than this threshold and the number of links in the weighted network. To dissect the network structure, several thresholds were used to subset the meta-networks into more stringent density subnetworks (DSs). For each species meta-network, five DSs were obtained ranging from DS1 (top 0.1% links) with an average of 358,455 links, to DS5 (top 10% links) with an average of 35,845,512 links (Fig. 1, step 3), with higher densities corresponding to a higher number of neighbors for each gene in the network (Fig. S1).

To identify genes showing conserved co-expression in different species, a gene neighborhood conservation analysis was performed using each DS and the information on the orthology relationships between Arabidopsis, maize and aspen genes (Fig. 1, step 4a). The network neighborhood of a gene is represented by all genes connected to it, at a given threshold. This concept was used to identify conserved “triplets” (Dataset 2), each containing three orthologous genes across Arabidopsis, maize and aspen with statistically significant overlaps of their gene

network neighborhoods. In an example triplet (Fig. 1, step 4a), a specific Arabidopsis gene *A1*, will have an ortholog *Z1* in maize and another ortholog *P1* in aspen and these three genes will have a significant overlap of their gene network neighborhoods. Due to the complex orthology relationships that exist in plants, each gene can belong to one or multiple triplets as it can have one or more orthologs. For example, an Arabidopsis gene with only one ortholog in maize and aspen, assuming they have significant overlap of their gene network neighborhoods, will belong to one triplet. In contrast, another Arabidopsis gene with two orthologs in maize and three in aspen, assuming they also all have significant overlaps of their gene network neighborhoods, will belong to six triplets. We refer to the set of unique genes that are part of triplets as "triplet genes". Next, the conserved gene neighborhoods were used to dissect the complex network structures of these plants and to functionally harness the orthology relationships. The cross-species networks are available in an interactive web application (https://beta-complex.plantgenie.org).

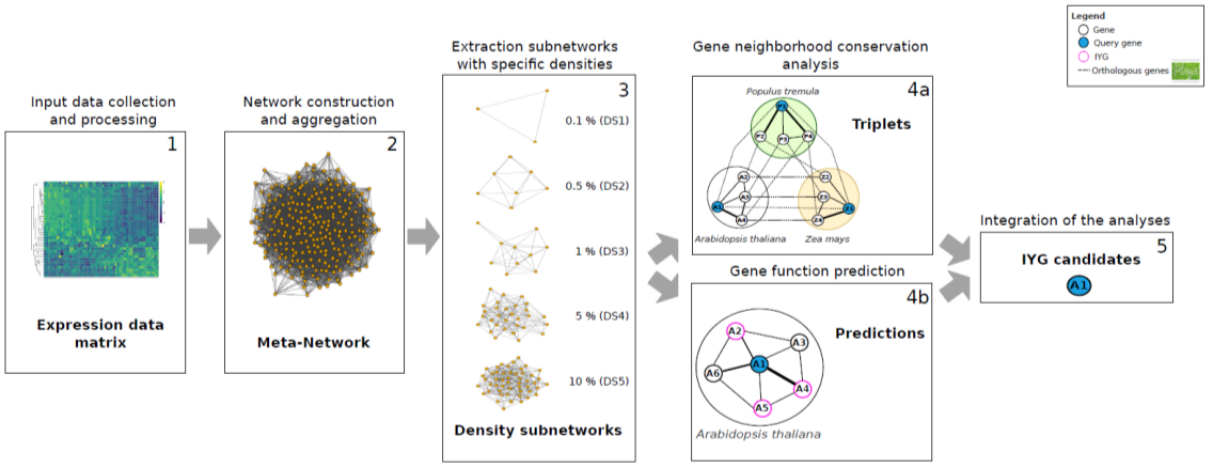


Figure 1. **Outline of the cross-species network approach to identify new intrinsic yield gene candidates.** For Arabidopsis, maize and aspen, the expression data (step 1) is used as input to construct a fully connected meta-network per species (step 2). Subsequently each meta-network is split into five density subnetworks (DSs) by applying specific density cutoffs (step 3). These DSs are the input for two different analyses: they are used first as input to compute cross-species gene neighborhood conservation (step 4a). Secondly, they are used to predict new functions via guilt-by-association (step 4b). This leads to gene function annotations of query genes (blue circles) based on prior knowledge on IYGs (purple circles). Edge thickness defines in which subnetwork the interaction is conserved (line thickness represents the DS and ranges from 1, the most stringent DS represented by the thickest line, to 5, the least stringent DS represented by the thinnest line). Finally, the results of these two analyses (steps 4a and 4b) are integrated to obtain a list of IYG candidates (step 5).

164

165 **Delineation of conserved intrinsic yield genes contributing to leaf growth**

166 Since the output of cell proliferation and expansion are strongly contributing to leaf size, we
167 hypothesized that the generated triplets were an excellent source to extract orthologs potentially
168 affecting plant leaf growth (intrinsic yield genes (IYGs)). An IYG is defined as a gene that when
169 inactivated or ectopically expressed increases organ size, here leaf size, by stimulating cell
170 proliferation (and thus higher cell number, as in the case of GRF (GROWTH-REGULATING
171 FACTOR) and GIF (GRF-INTERACTING FACTOR) proteins (Lee et al., 2009)) and/or cell
172 expansion (as in the case of *ZHD5* (*ZINC-FINGER HOMEODOMAIN 5*) (Hong et al., 2011)).
173 We generated a list of known IYGs from all three plant species ("primary-IYGs") composed of
174 71 primary-IYGs from Arabidopsis, 71 from aspen and eight from maize. This list of genes was
175 obtained by collecting scientific literature and by large-scale phenotypic screenings of mutant
176 and over-expression of Arabidopsis, maize, and aspen (Table S2).

177 We then used the triplets to transfer IYGs from maize and aspen to Arabidopsis ("translated-
178 IYGs"). Briefly, primary-IYGs from maize and aspen, also identified as triplet genes, were used
179 to extract Arabidopsis orthologs with conserved co-expression. The primary-IYGs and
180 translated-IYGs were finally merged and filtered for high expression variation in the Arabidopsis
181 expression compendium to retain only those active during either cell proliferation or cell
182 expansion. The resulting set, named "expression-supported IYGs" (Table S2, Fig. S2), was
183 composed of 82 IYGs, including 24 Arabidopsis primary-IYGs and 58 translated-IYGs (*GRF2*
184 and *GA20OX1* (*GIBBERELLIN 20-OXIDASE 1*) were shared between primary-IYG and
185 translated-IYG sets). According to their expression profile in Arabidopsis, 35 expression-
186 supported IYGs showed maximal expression during cell proliferation, including several
187 proliferation marker genes like GROWTH-REGULATING FACTORs (e.g. *GRF1*, *GRF2*,
188 *GRF3*), AINTEGUMENTA (*ANT* (Mizukami and Fischer, 2000) and *KLUH* (Anastasiou et al.,
189 2007)), and 47 expression-supported IYGs had increased expression during cell expansion, such
190 as *GA20Ox1* (Barboza et al., 2013) and *BR ENHANCED EXPRESSION 2* (*BEE2* (Friedrichsen et
191 al., 2002)).

192 The 82 expression-supported IYGs (from here on simply referred to as "IYGs") represented our
193 "guide" set of yield-related genes, obtained by the integration of prior knowledge on plant

194 growth and the cross-species gene neighborhood conservation approach, to identify new
195 candidate IYGs.

196

197 **Functional analysis of cross-species conserved networks underlying leaf cell proliferation**
198 **and expansion**

199 To explore cross-species conserved genes that function during cell proliferation and expansion,
200 we performed a Gene Ontology (GO (Ashburner et al., 2000)) functional enrichment analysis of
201 the Arabidopsis triplet genes from each DS across two sets: (1) all triplet genes (All) and (2) the
202 subset of triplet genes including the 82 IYGs and their co-expressed triplet genes (IYG-related
203 triplet genes) (Fig. 2). The total number of triplets ranged from 1,739 (DS1) to 243,645 (DS5)
204 (Fig. 2A; Dataset 2). To assess the significance of these numbers, a permutation approach was
205 employed where the orthology relationships were randomized 500 times and the number of
206 triplets obtained from each permutation was recorded. The number of triplets observed were
207 highly significant with not a single permutation for any DS exceeding the number of triplets
208 observed in the non-permuted data (p-value<0.002). The number of unique Arabidopsis triplet
209 genes ranged from 211 (DS1) to 6,526 (DS5) indicating that less sparse networks tend to have
210 more genes and more conserved gene neighborhoods (Fig. 2A). Interestingly, IYGs and their
211 network neighbors on average made up 71% of the triplet genes across the five DSs, suggesting
212 that plant growth-related gene networks are well conserved during leaf development across plant
213 species. For simplicity, from here on we will refer to triplet genes at a specific DS as, for
214 example at DS1, as "genes conserved at DS1". The functional enrichment (Fig. 2B) showed that
215 triplet genes from the most stringent subnetwork (DS1) were enriched for fundamental biological
216 processes during leaf development, including photosynthesis (e.g. glucose metabolic process,
217 response to light and carbon fixation), translation (e.g. large and small ribosomal subunits) and
218 cell proliferation (e.g. positive regulation of cell cycle, chromatin organization, microtubule-
219 based movement). Processes such as cell division and cell cycle regulation were significantly
220 enriched for genes conserved at DS2 and DS3, including genes coding for cyclins (type A, B, D
221 and P), cyclin dependent kinases (*CDK*) and their subunits (*CKS*), and other genes involved in
222 the spindle formation (i.e. *MICROTUBULE-ASSOCIATED PROTEINS (MAP)65-4 and -5*). Cell
223 expansion-related processes were identified among genes conserved at DS3 and included genes

coding for expansins (EXP) and xyloglucan endotransglucosylases/hydrolases (XTH). Genes conserved at the two least stringent subnetworks (DS4 and DS5) were enriched for GO terms related to cell wall organization (e.g. lignan biosynthesis, pectin degradation, lignin metabolism), defense response to biotic and abiotic stresses (e.g. defense response to oomycetes, response to salt stress and heat stress), and transmembrane transport and hormone signaling (e.g. response to auxin, ethylene and brassinosteroid). The category "regulation of transcription" was enriched for genes conserved at DS3, DS4, and DS5. IYGs were significantly over-represented at DSs starting from DS2, indicating that IYGs have highly conserved gene network neighborhoods. Most of the IYGs (87%) were conserved in one or more DSs (mainly DS4) (Fig. 2C).

Among the IYGs conserved at DS2, 32% were transcription factors (TFs), including regulators of cell cycle (e.g. *AINTEGUMENTA*) and cell elongation such as *BEE2* and its homolog *HBI1* (Fig. S3). These results suggest a conserved role of these TFs in leaf development across the three plant species. Genes involved in hormone-mediated transcriptional regulation (*INDOLEACETIC ACID-INDUCED PROTEIN (IAA)3*, *IAA14*, *IAA30*, and *AUXIN RESISTANT* (*AUX*)*1*) were also detected. Cell growth regulators, including the GRF family, were found conserved and among them, *GRF2* was conserved at DS2. Literature information on differentially expressed gene (DEG) sets from perturbation experiments was also included in the functional enrichment analyses for several primary-IYGs. In particular, genes up- and down-regulated in *SAMBA* loss-of-function mutants (Eloy et al., 2012) and *JAW (JAGGED AND WAVY)* overexpression lines (Gonzalez et al., 2010) were significantly enriched in the IYG-related set (Fig. 2B). Whereas *SAMBA* plays a key role in organ size control (seeds, leaves and roots), transgenic overexpression lines of *JAW* showed enlarged leaves and an increased cell number, indicative of prolonged cell proliferation (Gonzalez et al., 2010; Eloy et al., 2012). An additional functional enrichment analysis was performed focusing on TF families to identify their cross-species conservation level. In particular, genes conserved from DS2 to DS (Fig. S4) were significantly enriched for the ETHYLENE RESPONSE FACTOR (ERF) family (q-value < 0.01), which has a recognized role in plant growth (Dubois et al., 2018). At DS3, among others, MYB and WRKY TF families, known to be involved in developmental processes, appeared strongly conserved. At the least stringent DSs (DS4 and DS5) we could observe other conserved TF families like DOF (regulating the transcriptional machinery in plant cells), MIKC-MADS (involved in floral development) and NAC (with functions in plant growth, development and

stress responses) (Lehti-Shiu et al., 2017). For TFs conserved at DS2, a significant enrichment was observed for the CONSTANS-like TF-family when considering IY-related triplet genes and included *BBX3*, *BBX4, BBX14* and *BBX16*. A number of BBX proteins have been linked with photomorphogenesis, neighborhood detection, and photoperiodic regulation of flowering (Vaishak et al., 2019).

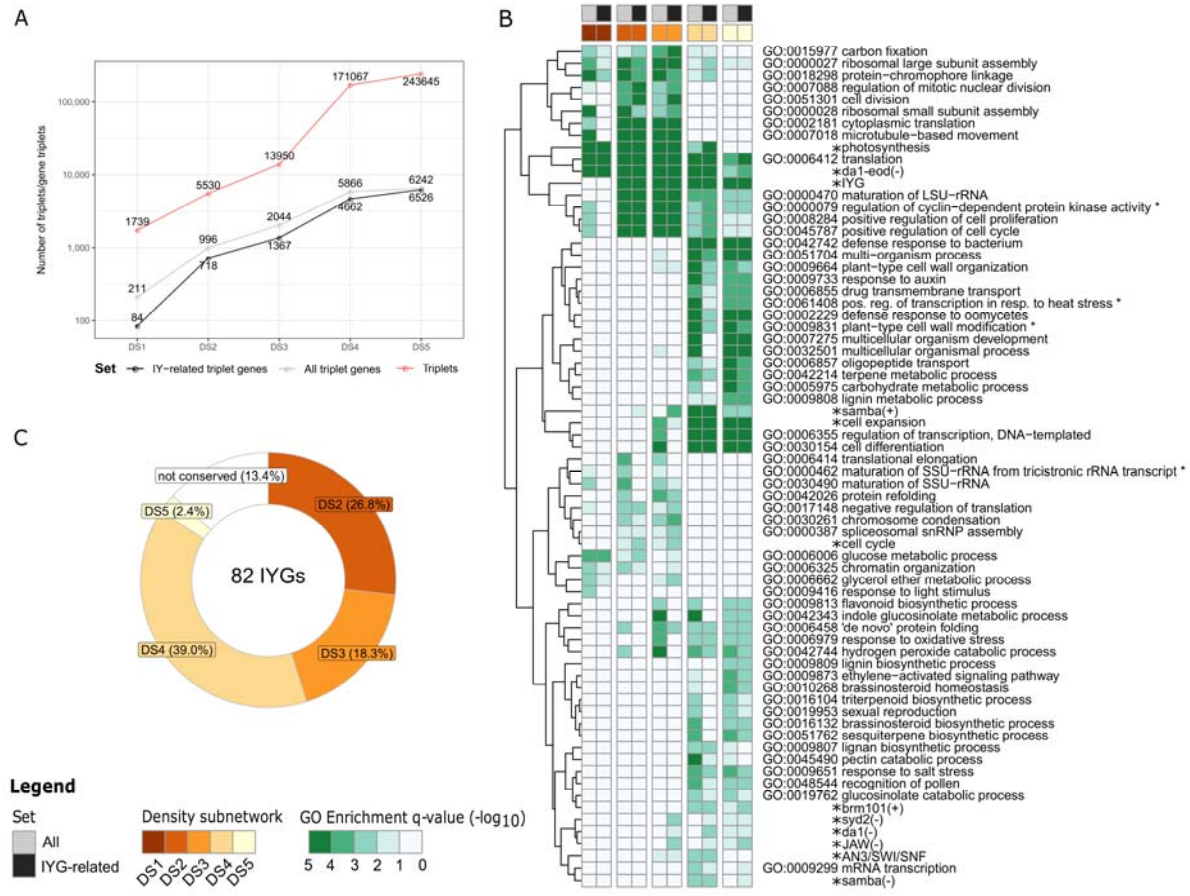


Figure 2. **Triplets and their functional enrichments in cross-species conserved leaf networks**. (A) The number of triplet genes showing cross-species neighborhood conservation is plotted for all density subnetworks (DSs). (B) The functional over-representation of biological processes of interest is summarized for two sets: all triplet genes (All) and for IYGs and their network neighbor (IYG-related) triplet genes, subset of all triplet genes, in each DS. (C) Overview of IYGs with (and without) cross-species neighborhood conservation in different DSs.

Network-based prediction of novel IYGs

Apart from analyzing the conservation level of known IYGs, we subsequently investigated if new IYGs could be identified. To obtain high quality IYG predictions, a combined strategy was adopted to leverage the known IYGs and the gene neighborhood conservation analysis through a guilt-by-association (GBA) approach. The GBA principle states that genes with related function tend to be protein interaction partners or share features such as expression patterns or close network neighborhood (Oliver Stephen, 2000). First, gene function prediction through GBA was performed, where guide IYGs were used for network exploration and gene function discovery (Fig. 1, step 4b). New gene functions were assigned through functional enrichment in the Arabidopsis networks, at different DSs. As a result, genes that were part of network neighborhoods significantly enriched for known IYGs were classified as newly predicted IYGs, and a GBA score was assigned to quantify the reliability of the predicted IYGs (Methods). Secondly, the new predictions were filtered for those already identified as triplet genes (Fig. 1, step 5). These filtered predictions, forming the predicted IYG set, were labelled with their species names if they were already known as IYGs (primary or translated-IYG) or with "new" if they were not (Table S4). This approach led to 2206 IYG predictions, of which 66 were known IYGs. For the latter, 11 were uniquely from the Arabidopsis IYG primary set, 53 uniquely from the aspen translated-IYGs, and the remaining two were shared among species. From DS1 to DS5, the subsets of IYG predictions covered 175, 496, 421, 891 and 223 genes, respectively (circle size in Fig. 3B) (Table S4). The expression profiles enabled identification of two major clusters, with 1013 genes peaking at proliferation and 1193 at expansion phase (Table S4).

To evaluate the reliability of the predicted IYG set and its potential use for discovering genes with a significant effect on plant growth, the public phenotype database RARGE II (Akiyama et al., 2014), covering 17,808 genes and 35,594 lines, was screened obtaining a list of 391 Arabidopsis genes that, if mutated, caused a phenotype change in Arabidopsis leaf length, width and/or size (RARGE II leaf trait genes, Table S5). Subsequently, functional enrichment was applied to further investigate the predicted IYG set using these RARGE II leaf trait genes as well as differentially expressed gene (DEG) sets from published yield-related perturbation experiments (Fig. 3A). The RARGE II leaf trait gene set was significantly overrepresented only in DS1, indicating that in stringent networks there is a higher chance of detecting genes causing a visible phenotype if mutated (Fig. 3A). The gene set of downregulated genes in *da1-eod* double

mutants (Vanhaeren et al., 2017) was found enriched at DS1. These mutants showed significantly increased cell size and cell number as compared to control due to processes taking place before transition from cell proliferation to cell expansion delaying transition. The genes, part of this set, were mostly coding for chloroplast proteins, involved in chloroplast development (*LRGB*), light-harvesting (*LHCA2),* photosystems (*PSAD-2, OXYGEN EVOLVING COMPLEX SUBUNIT 33 KDA (OEC33))*, and stomatal movement and conversion of carbon dioxide and water (*CARBONIC ANHYDRASES* (*CA1*, *CA2*, *CA4*)). This suggests that genes active in processes marking the onset of photosynthesis are well conserved across monocots and dicots. At DS2 other sets of genes significantly overrepresented were those downregulated in plants overexpressing ANGUSTIFOLIA3 (*AN3*), and *JAW*, which carried larger leaves due to an enhanced cell proliferation (Vercruyssen et al., 2014; Gonzalez et al., 2010). In common across these two gene sets we found *AKT2,* which encodes for potassium inward channels that contribute to the osmotically driven water uptake for expansion, and *DMR6*, a defense-associated 2OG-Fe(II) oxygenase. Among known IYGs conserved at any of the five DSs, we found important genes acting at cell proliferation (e.g. *CYCD3*, *ANT*, *KLUH*, *GRF1*, *GRF3*, *GRF5*, and *GIF2*) and others acting at cell expansion phase, such as *GA3OX1*, *GA20OX1*, and *GA20OX6*, involved in the gibberellin biosynthetic pathway. Overall, cell proliferation genes were statistically enriched at DS2 and expansion genes at DS3 (Fig. 3A). When investigating the gene recovery for the RARGE II leaf trait genes (Fig. 3B), a clear trend was observed in phenotype recovery ranging from DS1, with higher recovery (~3 and ~4.3 fold compared to what is expected by chance for proliferation and expansion, respectively), to DS5, with almost no recovery. This result indicates that the most stringent density subnetwork is a valuable source to identify genes with a potential effect on leaf phenotype.

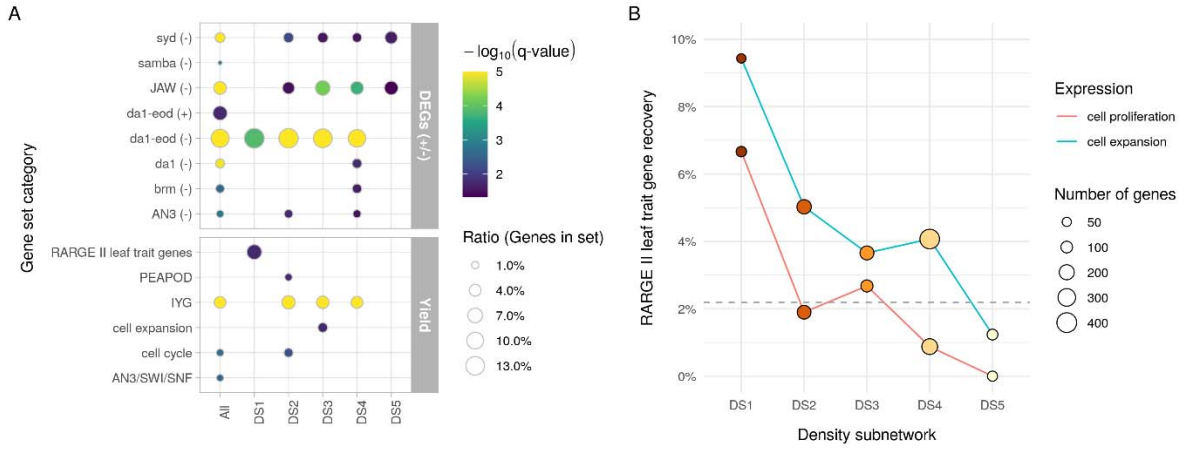


Figure 3. Functional enrichment of the intrinsic yield gene predictions. (A) Functional enrichment of the predictions split by density subnetwork. Gene sets in the upper panel were represented by upregulated (+) and downregulated (-) genes in overexpression lines (genes in upper-case) or mutant lines (genes in lower-case) collected from literature (Bezhani et al., 2007; Eloy et al., 2012; Vercruyssen et al., 2014; Vanhaeren et al., 2017). Gene sets in the lower panel were belonging to functional categories listed in Table S3. (B) Recovery of RARGE II leaf trait genes for each DS split in proliferation and expansion. The grey dashed line indicates the leaf-related phenotype gene recovery expected by chance (within the RARGE II dataset).

Validation of IYG predictions using literature and leaf phenotyping

To validate the assumption that the IYG predictions top ranked by GBA are more likely to show a phenotype, an in-depth literature analysis was performed to summarize the connection with yield pathways (Table S6) and to score known growth-related phenotypes for the top 100 IYG predictions (Table S7). For 61 of these 100 predicted genes, mutant lines and/or lines with ectopic expression were reported. For 34 out of the 61 genes (55.7%), obvious alterations to leaf size and shape as well as petiole length were reported when mutated or overexpressed, which included five Arabidopsis primary-IYGs, six translated-IYGs from aspen, and five paralogs to known IYGs (Fig. 4; Table S7). By screening the literature, these five IYG paralogs (*APETALA 2 (AP2)*, *GLABRA 1 (GL1)*, *CYCD3;3*, *AUXIN RESISTANT 2 (AXR2)*, *OBP1*) were reported to show a leaf phenotype when mutated or overexpressed (Table S7).

Functional analysis of the 34 genes with described leaf phenotypes revealed their involvement in several biological processes and pathways such as cell cycle regulation, hormone response, photosynthesis, carbon utilization and cell wall modification (Fig. 4). Importantly, we could find conserved relationships between five specific genes active in the expansion phase: CATIONIC

AMINO ACID TRANSPORTER *(CAT)2*, *THIOREDOXIN* X *(THX)*, BETA CARBONIC ANHYDRASE *(BCA)4*, *CA2*, and *PMDH2*. Among them, *CAT2* and *BCA4* were also high ranked by GBA score. For the proliferation cluster, we could observe strong relationships between *ANT*, *OBP1*, *GRF2*, *CYCD3;3*, *GL1*, *HTA8 (HISTONE H2A 8)*, and *AN3*. Among them, we identified TFs mainly involved in cell cycle process (*ANT*, *OBP1*, *GRF2*), cell wall (*GL1*), and hormone signaling pathways such as jasmonate (*GL1*), abscisic acid (*ANT*), and gibberellin (*GL1*). Twenty-seven of the 61 predictions with knock-down mutations and/or ectopic expression lines did not show a correlation with leaf growth, which may be partially due to the redundancy of large gene families or that the leaf phenotype was not explored in those studies. Additionally, three of these 23 genes have been reported to influence root or hypocotyl development, which may also contribute to overall plant growth and organ size.

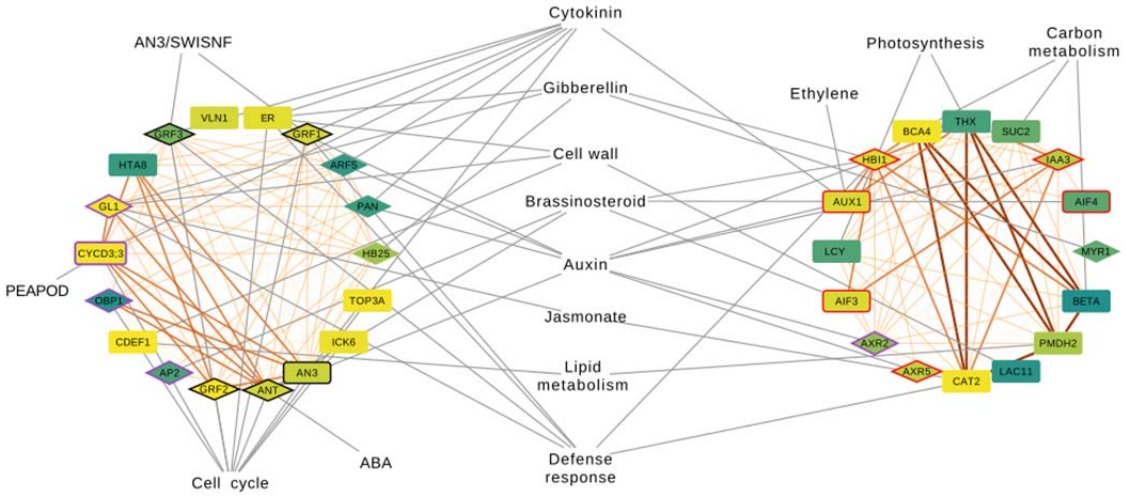


Figure 4. Gene-function network of the 34 phenotype-related genes out of the top 100 intrinsic yield gene predictions. Predictions are clustered by expression profile (proliferation on the left and expansion on the right). Node label colours from yellow (strong) to dark green (weak) represent the reliability of the gene prediction (GBA score). Node border colours indicate known IYGs from Arabidopsis (black), known IYGs from aspen (red), and Arabidopsis known IYG paralogs (violet). Diamonds represent transcription factors. Links from dark orange thick (DS1) to light orange thin (DS5) represent the density subnetwork where the genes were found connected. Genes are linked to the yield pathway they are related with (centered if connecting to both proliferation and expansion related genes) by grey links. Anti-correlation links (connecting proliferation with expansion genes) were removed for clarity.

373 To further validate the role of these new IYGs in leaf development, we collected the mutants of
374 nine genes among the 27 predicted IYGs which have not been reported with a leaf phenotype
375 (Table S8). Molecular identification of these mutants was conducted and a detailed analysis of
376 leaf growth in controlled long-day soil-grown conditions was made (Fig. S5). By following the
377 projected rosette area (PRA), compactness and stockiness of each mutant line over time, this
378 phenotypic characterization revealed that the mutants of two IYG candidate genes showed
379 altered rosette growth. The mutant lines of a nitrate transporter gene *NPF6.4/NRT1.3*, *sper3-1*
380 and *sper3-3*, both displayed decreased PRA compared with the wild-type plants (Fig. 5A). The
381 *sper3-1* harbored a mutation at a conserved glutamate *of NRT1.3,* while the T-DNA line *sper3-3*
382 was a knockout allele (Tong et al., 2016). The reduction in size of *sper3-3* was smaller and
383 occurred later in development compared with *sper3-1*. Before bolting (26 DAS), *sper3-1* and
384 *sper3-3* were 37.3% and 13.2% smaller, respectively, compared with the wild-type (Table S8).
385 Besides *NPF6.4*, the mutants of *LATE MERISTEM IDENTITY2* (*LMI2*) which has been reported
386 to be required for correct timing of the meristem identity transition (Pastore et al., 2011), also
387 showed altered rosette growth. In standard long-day conditions in soil, a significant reduction of
388 PRA was detected in *lmi2-1*, which displayed elevated *LMI2* expression in seedlings. By
389 contrast, the *lmi2-2* mutants in which the T-DNA insertion gave rise to a truncated non-
390 functional LMI2 protein, exhibited significantly increased PRA and were 13.5% larger than the
391 wild-type plants at 26 DAS (Fig. 5B and Table S8). Both *NPF6.4* and *LMI2* were highly ranked
392 by GBA (rank 18 and 20, respectively), which further implies that the predictions with a low
393 GBA score are more likely to show a leaf phenotype. Taken together, these experimentally
394 validated genes lend additional support to the potential of our predictions for plant growth.

395

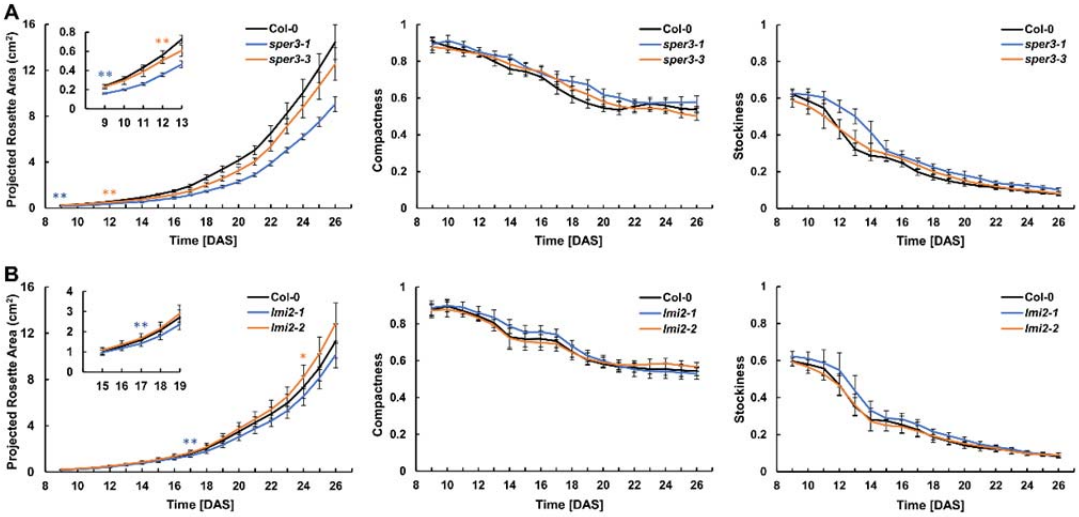


Figure 5. Mutants of IYG predicted genes *NPF6.4*/*NRT1.3* and *LMI2* showed altered rosette growth.

Dynamic growth analysis of projected rosette area, compactness and stockiness over time of wild-type Col-0 and the mutants of *NPF6.4*/*NRT1.3* (A) and *LMI2* (B) in soil. The asterisks represent the time points at which differences in the PRA become significant between the mutants and wild-type, as determined by Student's *t* test (*, $P<0.05$; **, $P<0.01$).

Discussion

In this study, we aimed to identify candidate genes that would be robust targets for altering leaf development using cross-species gene network analysis. To identify relevant context-specific gene interactions, it is highly recommended to focus the gene network analysis on a specific condition or context, rather than integrating multiple conditions (e.g. different stresses, growth conditions, development stages) (Pavlidis and Gillis, 2012; Liseron-Monfils and Ware, 2015; Serin et al., 2016). For this reason, expression datasets were generated and compiled capturing two main features of leaf growth: cell proliferation and cell expansion. These two processes are governed by similar cellular and molecular pathways across monocots and dicots (Nelissen et al., 2016), which inspired the selection of transcriptional datasets from two dicots (Arabidopsis and aspen) and one monocot (maize). The network construction was carried out integrating multiple inference methods to leverage the power and complementarity of different network inference algorithms (Marbach et al., 2012; Schiffthaler et al., 2018). To evaluate the strength of different biological signals in our network, the gene interactions obtained after applying different network density cutoffs (DS1-5) were studied. To identify functionally conserved genes across species, we relied on two main approaches: the guilt-by-association principle, which is frequently used for gene discovery, and network neighborhood conservation analysis, which detects significantly overlapping network neighborhoods across species to identify reliable functional orthologs (Movahedi et al., 2011; Netotea et al., 2014).

From the gene neighbourhood conservation analysis on five different density subnetworks, we observed that, with an increased network density, the number of genes with conserved network neighborhood also grew. This is expected and is probably due to a greater statistical power when comparing larger neighborhoods (Netotea et al., 2014). Overall, as previously observed (Vercruysse et al., 2020), the integration of different sequence-based orthology detection methods was important because of their complementarity, highlighting complex orthology relationships and evaluating the strength of the orthology support. Overall, 36% of the Arabidopsis genes (7,320 out of 20,313 genes present in the network) had conserved neighborhoods across Arabidopsis, aspen, and maize, in any of the five density subnetworks. This result is similar to what has been found across Arabidopsis, poplar and rice, although a different network construction pipeline was used there (Netotea et al., 2014).

From a plant breeding perspective, we were interested in focusing on cross-species conserved targets with experimental evidence in more than one species. *GA20-oxidase1* represents a well-known example of an IYG that is functionally conserved across monocots and dicots. This gene was confirmed in our analyses to be conserved at the network neighbourhood level. *GA20-oxidase1* is in fact a rate limiting enzyme for gibberellin growth hormone biosynthesis in Arabidopsis, aspen, maize and rice (Gonzalez et al., 2010; Nelissen et al., 2012; Qin et al., 2013; Eriksson et al., 2000). To validate the functional relevance of the predicted IYGs, we screened the top 100 IYG predictions and observed that, among the 34 Arabidopsis predicted genes with a known leaf phenotype in Arabidopsis, six were also already known to affect plant growth in aspen ("translated-IYGs"). Plants, in fact, develop tissues and organs through the activity of both primary and secondary meristems (vascular cambium) and there are overlapping regulatory mechanisms between them (Baucher et al., 2007). The six translated-IYGs were *AUX1*, *IAA3*/*SHY2*, *AXR5*, *AIF3*, *AIF4*, and *HBI1* and their expression in Arabidopsis was peaking at the cell expansion phase. The first three genes are auxin-related genes. Auxin is important for regulating root meristem growth and is crucial for root initiation and lateral root number. *AUX1* was translated from *aspen* Potra002054g16021 while *IAA3*/*SHY2* and *AXR5* were translated from aspen Potra000605g04596. For both these aspen genes, generated aspen RNAi lines exhibited an increase in stem diameter, an important indicator for tree biomass yield, connecting back to the underlying regulatory processes in the meristematic tissues (Table S2). *AUX1* is an auxin transport protein which regulates auxin distribution across source (young leaf) and sink organs (young roots) (Marchant et al., 2002). *IAA3*/*SHY2* is crucial for root meristem development in Arabidopsis, being the converging point of cytokinin and auxin regulatory circuit (Li et al., 2020). Arabidopsis mutants for *AUX1* and *IAA3*/*SHY2* showed alterations in number and size of lateral roots (Tian and Reed, 1999; Marchant et al., 2002) while *AXR5* is an auxin response factor and mutant plants for this gene are resistant to auxin and show alterations of root and shoot tropisms (Yang et al., 2004). Our network results and phenotypes in aspen and Arabidopsis indicate that these genes also play an import role in meristem growth in other organs apart from root. *HBI1*, *AIF3*, and *AIF4*, encode a tier of interacting bHLH transcription factors downstream of BR and regulate the cell elongation in leaf blade and petiole (Bai et al., 2013; Ikeda et al., 2013). *AIF3* and *AIF4* were translated from Potra004144g24626 while *HBI1* was translated from Potra186144g28414. These two aspen genes have been tested with an overexpression approach

in aspen trees showing even a bigger increase in stem diameter as compared with the auxin-related aspen genes Potra000605g04596 and Potra002054g16021 (Table S2). Arabidopsis mutants for these genes (*HBI1*, *AIF3*, and *AIF4*) have been linked with alteration of petiole length (Table S7).

LMI2 was a highly ranked IYG prediction. Importantly, *LMI2* (a MYB TF) is not a paralog of *LATE MERISTEM IDENTITY 1* (*LMI1,* a homeobox TF), also predicted here. Although *LMI1* and *LMI2* belong to different TF families, they both function downstream of LEAFY to regulate meristem transition (Pastore et al., 2011). *LMI1* was reported to regulate leaf growth in Arabidopsis and other species (Vlad et al., 2014; Andres et al., 2017; Li et al., 2021). Arabidopsis *LMI1* loss-of-function mutant showed decreased leaf serration and promoted tissue growth in stipules (Vuolo et al., 2018). The observed phenotype of mutated *LMI2* was related to an increase of the number of cauline leaves and secondary inflorescences (Pastore et al., 2011). Here, *LMI2* transgenic lines were subjected to phenotypic analysis, which demonstrated that a *LMI2* loss-of-function mutant showed increased rosette area. The neighbourhood conservation of both *LMI1* and *LMI2* suggests that it would be worthwhile to further explore their roles in leaf shape control across monocots and dicots.

Other known examples of functionally conserved predictions across monocots and dicots were GRFs (e.g. the highly ranked *GRF2*), which have a recognized role in leaf size regulation, and AN3/GIF1, a transcriptional co-activator protein (Nelissen et al., 2016). This was also testified by their network conservation in stringent density subnetworks (DS2). A second gene *GL1*, had its network neighbourhood conserved with GRMZM2G022686 from maize. This maize gene encodes for the MYB-related protein *Myb4*. This protein plays important roles in plant improved tolerance to cold and freezing in Arabidopsis and barley (Soltész et al., 2012), but no connections with improved yield have been observed for this gene. Arabidopsis *SUC2* showed conservation with GRMZM2G307561, a sucrose/H^+ symporter which remobilize sucrose out of the vacuole to the growing tissues. Mutants for this gene showed reduced growth and the accumulation of large quantities of sugar and starch in vegetative tissues in Arabidopsis (Srivastava et al., 2008), while in maize mutants, slower growth, smaller tassels and ears, and fewer kernels were observed (Leach et al., 2017). This gene is thus also important for growth, development, and yield across monocots and dicots.

A total of 11 primary-IYGs from Arabidopsis showed no network neighbourhood conservation. Lack of conservation might be the result of (1) missing orthologs in a target species or (2) different set of co-expressing genes across species, which in turn might be caused by different transcriptional control. One clear example of no conservation due to a lack of orthologs is PEAPOD 2 (*PPD2)*, which is a TIFY transcriptional regulator part of the PEAPOD (PPD) pathway. This pathway plays an important role in cell proliferation and, with its PPD/KIX/SAP module, is involved in leaf, flower, fruit, and seed development. This pathway is highly conserved among flowering plant species but absent in monocot grasses (Schneider et al., 2021). The reason for this absence might be found back in intrinsic differences between eudicots and grasses, being mainly lack of meristemoids and functional redundancy for the regulation of cell proliferation. Surprisingly, several non-grass monocot species such as *Musa acuminata* (banana) and *Elaeis guineensis* (oil palm), basal angiosperm *Amborella trichopoda* and lycophytes, carry PPD/KIX/SAP orthologs, although information about their functionality is missing (Schneider et al., 2021). Another gene with orthologs but lacking network neighborhood conservation was *AHK3*, a cytokinin receptor that controls cytokinin-mediated leaf longevity. This might be explained by knock-out experiments on *AHK* receptors showing contrasting effects on flowering time or floral development across Arabidopsis and rice (Burr et al., 2020). Another non-conserved IYG was *ZHD5* that regulates floral architecture and leaf development and is regulated by *MIF1 (MINI ZINC-FINGER 1)* (Hong et al., 2011), which also lacked network conservation. *ZHD5* regulation might thus be different across species. Similarly, FBX*92 (F-BOX PROTEIN92)* was not conserved, which might be explained by the opposite effects on leaf size shown by *ZmFBX92* and *AtFBX92* gain of function in Arabidopsis due to the presence of an F-box-associated domain in *AtFBX92*, lacking in *ZmFBX92*. *FBX92* orthologs might thus undergo different transcriptional regulation (Baute et al., 2017). *EPF1 (EPIDERMAL PATTERNING FACTOR 1)* was also a non-conserved IYG. This gene affects stomatal density and water use efficiency. Recent work suggested that, in monocots and dicots, *EPF1* orthologs probably have different temporal dynamics of gene expression in the stomatal lineage (Buckley et al., 2020), which might result in different co-expressors.

Based on the validation results of our IYG prediction pipeline, a correlation between network size and recovery of genes affecting leaf size was observed. In particular, the most stringent

Arabidopsis network showed a high recovery rate of leaf phenotype related genes, either considering RARGE II leaf trait genes or an in-depth literature analysis. With increasing network size, the recovery rate decreased. The network neighborhood conservation of genes in the most stringent networks involved different fundamental processes, suggesting their functional similarity across monocots and dicots. Not surprisingly, genes involved in cell cycle regulation and plant hormonal response were found, as both processes have a key role in leaf development. Several cell cycle regulators were predicted as IYGs, like the cyclin gene *CYCD3;3*, the *CDK* inhibitor *KRP3 (KIP-RELATED PROTEIN)*, and a DOF transcription factor gene *OBP1* (*OBF BINDING PROTEIN 1*) that controls cell cycle progression (Dewitte et al., 2007; Skirycz et al., 2008; Jun et al., 2013). The auxin-responsive transcription factor gene MONOPTEROS (*MP*) is crucial for leaf vascular development (Hardtke and Berleth, 1998), while the Aux/IAA gene that represses auxin signaling, *AXR2*, whose gain-of-function leads to strong inhibition of leaf growth (Mai et al., 2011), was also predicted. Besides auxin, brassinosteroid (BR) and gibberellin (GA) coordinately play key roles in regulating plant cell elongation. The other two predicted transcription factor genes, *HB25* (*HOMEOBOX PROTEIN 25*) and *MYR1*, which modulate bioactive GA biosynthesis, were also shown to have an effect on the petiole growth (Bueso et al., 2014). It is noteworthy that nearly half of all the 34 genes with leaf phenotype were transcription regulators, which highlights the importance of TF-mediated gene expression regulation during leaf development. In addition to hormone-related genes and TFs, genes related to photosynthesis are also important for leaf development. A carotenoid biosynthesis gene *LCY* and a chloroplast redox-regulating gene *THIOREDOXIN X* were predicted as IYG and have been shown to affect leaf size (Li et al., 2009; Pulido et al., 2010). Moreover, the cytoplasmic carbonic anhydrase genes *CA2* and *BCA4* were identified, consistent with the view that carbon utilization in leaves is closely linked to leaf area (DiMario et al., 2016). Cell wall modification is considered to be another important determinant of leaf development. The predicted candidate genes *LACCASE11 (LAC11)* and *CUTICLE DESTRUCTING FACTOR 1* (*CDEF1)*, encoding for a laccase that associates with the lignin deposition in cell wall and a cutinase essential for the degradation of cell wall components, respectively, are also involved in regulating leaf growth and morphology (Takahashi et al., 2010; Qin et al., 2013). Among Arabidopsis genes with a reported phenotype in the RARGE II loss-of-function dataset, *ACO2 (ACC OXIDASE 2)* led to increased leaf size, and AT3G43270, a member of Plant invertase/pectin methylesterase inhibitor superfamily, to

558 smaller leaves. IYGs translated from aspen led, through our integrative network approach, to the
559 prediction of *NITRATE TRANSPORTER 1.3* (*NPF6.4*/*NRT1.3)* as a new potential IYG. In our
560 experiments, we showed that this gene, when mutated, is altering leaf growth. It was previously
561 hypothesized that *NPF6.4*/*NRT1.3* may play a role in supplying nitrate to photosynthesizing cells
562 (Tong et al., 2016). This cross-species conserved gene would thus contribute to nitrogen
563 assimilation, that, closely interacting with carbon metabolism, sustains plant growth and
564 development (Nunes-Nesi et al., 2010).
565 In conclusion, the approach developed in this study fully exploits the potential of integrative
566 biology to translate and expand yield-related functional annotations in different plant species, as
567 such accelerating crop breeding.

568

569

Methods

Integration of developmental expression datasets and network construction

Transcriptomic datasets were obtained from a list of studies in Arabidopsis, maize and aspen covering samples from the main leaf developmental phases (Table S1, Supplemental Methods, Dataset 1). Details about processing of these samples were reported in TableS1. Maize data was mainly composed by a developmental compendium newly generated in this work (Supplemental Methods). The network inference was carried out with Seidr (Schiffthaler et al., 2018), which infers gene networks by using multiple inference algorithms and then aggregating them into a meta-network. This approach has been shown to strongly improve the accuracy of the results (Marbach et al., 2012). Each network was subset into five density subnetworks (DSs) using five different network density values. This procedure consisted in selecting the top 0.1, 0.5, 1, 5 and 10% top Seidr links in each species-specific network and generating five DSs (from the most stringent DS1 to the least stringent DS5).

Orthology and network neighborhood conservation

In order to compute cross-species gene network neighborhood conservation, orthology information between genes from Arabidopsis, maize and aspen was computed using the PLAZA comparative genomics platform (Van Bel et al., 2018). A custom version of this platform was built covering in total 15 eukaryotic species including *Arabidopsis thaliana* (TAIR10), *Eucalyptus grandis* (v2.0), *Populus trichocarpa* (v3.01), *Populus tremula* (v1.1), *Vitis vinifera* (12X March 2010 release), *Zea mays* (AGPv3.0), *Oryza sativa* ssp. *Japonica* (MSU RGAP 7), *Triticum aestivum* (TGACv1), *Amborella trichopoda* (Amborella v1.0), *Picea abies* (v1.0), *Pinus taeda* (v1.01), *Selaginella moellendorffii* (v1.0), *Physcomitrella patens* (v3.3), *Chlamydomonas reinhardtii* (v5.5) and *Micromonas commode* (v3.0). PLAZA allows identifying orthologs using different methods (evidences), corresponding to orthologous gene families inferred through sequence-based clustering with OrthoFinder (Emms and Kelly, 2015), phylogenetic trees, and multispecies Best-Hits-and-Inparalogs families (van Bel et al., 2012). The PLAZA orthology relationships were extracted and filtered retaining all orthologs having a requirement of 2/3 orthology evidences and, for those with 1/3 evidence and >25 orthologs, the ones corresponding to the best 25 blast hits (sorted by e-value) were retained. The generated orthology output was used for the following pipeline steps.

The generated DSs and the orthology information were used to compare the three species using a network neighborhood conservation analysis (ComPlEx analysis, as in Netotea et al.). In this analysis, the co-expression of a gene was considered conserved if its network neighborhood (i.e. all genes with a link to it) had a statistically significant ($q < 0.05$) overlap with the network neighborhood of its ortholog in the other species (Netotea et al., 2014). Here, the comparison was performed for all pairs of networks between the datasets of the three species, and the output of this analysis was collated to create "triplets". The triplets are sets of three orthologous genes–one per network/species–that have a significantly conserved network neighborhood in all three pairs of comparisons. Since the test is not commutative, the neighborhoods had to be significantly conserved in both directions of the test. To estimate the false discovery rate (FDR) of the detection of triplets, a permutation strategy was adopted. For 500 runs of ComPlEx, ortholog relationships were shuffled, keeping the relative number of orthologs per gene and per species, and then comparing the number of triplets computed from randomization with those resulting using the original (unshuffled) orthologs.

Functional information for gene function prediction

Gene Ontology (Ashburner et al., 2000) functional annotations for Arabidopsis, maize and aspen were retrieved from TAIR (download 25/12/2018), Gramene (AGPv3.30, http://bioinfo.cau.edu.cn/agriGO/download.php) and PlantGenIE (ftp://ftp.plantgenie.org/Data/PopGenIE/Populus_tremula/v1.1/annotation/), respectively, and filtered for the genes present in the corresponding species networks. We focused on biological processes (BP) and excluded the general GO BP terms with >= 1500 genes as well as GO terms with <= 10 genes to avoid biases towards very general (e.g. biological regulation) and specific terms. For each gene, all GO annotations were recursively propagated in order to include parental GO terms. To get a complete view on all relevant processes related to plant yield, information from literature was collected on intrinsic yield genes (IYGs). Experimentally validated genes in Arabidopsis, maize and aspen (primary-IYGs) were retrieved from public databases (Gonzalez et al., 2010; Beltramino et al., 2018). A second set of experimentally validated aspen genes were obtained by access to SweTree Technologies' private database that contains data from the large-scale testing of >1,000 genes and their yield-related properties, an effort where more than 1,500 recombinant DNA constructs were used to either introduce a new gene product or alter the level of an existing gene product by over-expression or RNA

interference in aspen trees, whose growth characteristics were then monitored in greenhouse and field experiments to provide extensive gene-to-yield data. The Arabidopsis IYG primary set was then enlarged with high quality IYG orthologs from maize and aspen using the triplets ("translated-IYGs") to obtain a combined IYG set. The combined set was finally filtered with genefilter package for Bioconductor (Gentleman et al., 2021) to remove genes with small expression variance (var.func=IQR, var.cutoff=0.8) and focus on genes active during proliferation or expansion phases of leaf development ("expression-supported IYGs", Table S2). Other information on functional categories (Vercruysse et al., 2020) (Table S3) and differentially expressed genes from relevant studies on plant development was also included in the functional enrichment analyses (Anastasiou et al., 2007; Gonzalez et al., 2010; Eloy et al., 2012; Vercruyssen et al., 2014).

The expression-supported IYG set was used to perform network-guided gene function prediction via a guilt-by-association (GBA) approach. This approach is based on the assumption that genes close to the input IYGs in the network are likely to have similar functions. The GBA approach was applied to attribute new functions based on GO enrichment in the modules of each DS yielding five sets of gene predictions. By this procedure, gene neighborhoods significantly enriched for known IYGs were functionally annotated (hypergeometric distribution). This allowed to predict new IYGs and estimate, for each of them, a corresponding FDR adjusted p-value (or q-value), which was renamed "GBA-score". The GBA-score is a confidence score that ranks genes low if they are connected with many IYGs in the network being a low score and indicator of strong enrichment. For an example IYG prediction (in one of any of the five DSs), the GBA-score from the five DSs was summarized taking the mean of the GBA-scores and setting the GBA-score to 0.05 for the DSs where the gene was not predicted. This yielded a list of IYG predictions that was then further filtered by only retaining those predictions having conserved neighborhood in at least one DS.

Validation of IYG predictions using large-scale phenotyping data

To perform a validation of the gene function predictions, the RARGE II (Akiyama et al., 2014) database was interrogated to retrieve a list of Arabidopsis genes that, when mutated, showed an increased or decreased length, width and size for rosette leaf, vascular leaf and cauline leaf (leaf trait genes). The IYG prediction list was split according to the DS conservation of the predictions

in five subsets. Each subset was analyzed for over-representation of specific biological process annotations using a functional enrichment analysis, using the hypergeometric distribution together with Benjamini-Hochberg (BH) correction for multiple testing (Benjamini and Hochberg, 1995). A similar approach was used to explore the top 100 predictions ranked by GBA-score. In this case a manual literature search was performed to retrieve all genes with a reported phenotype including information about the biological pathway the gene might be active in, and other public functional annotations.

Rosette growth phenotyping

The *Arabidopsis thaliana* ecotype Coumbia-0 (Col-0) was used as the wild-type in this study. The T-DNA insertion lines for At4g26530 (Salk_080758/*fba5-1*), At3g21670 (Salk_001553/*sper3-3*), At3g61250 (Salk_066767/*lmi2-1*, Salk_020792/*lmi2-2*), At4g25240 (Salk_113731), At1g63470 (Salk_123590/*ahl5*), At4g37980 (Salk_001773/*chr hpl*), At2g38530 (Salk_026257/*ltp2-1*), At4g28950 (Salk_019272), and At1g12240 (Salk_016136) were confirmed using PCR with a T-DNA primer and gene-specific primers (Lu et al., 2012; Zhao et al., 2013; Jacq et al., 2017; Tanaka et al., 2018; Pastore et al., 2011; Tong et al., 2016). All tested seeds were stratified in the darkness at 4 °C for 3 days and then sown on soil in the 7 cm wide square pots with a density of four seeds per pot. After 8 days in the growth room (with controlled temperature at 22 °C and light intensity 110 µmol m^{-2} s^{-1} in a 16 h/8 h cycle), the four seedlings were screened, leaving one seedling per pot, which most closely resembled the genotype average. The plants were imaged in a phenotyping platform (MIRGIS) with fixed cameras located directly above the plants, which images plants at the same time every day. These images were then processed to extract the rosette growth parameters of each plant. The mean PRA, compactness and stockiness values were calculated over time for each genotype.

Author Contributions

P.L.C, N.S, T.R.H, H.N and K.V designed the research; P.L.C, J.Z, N.M, C.S, C.M, T.D, and T.V.H performed research and data analysis; N.M, T.V.H, D.J, and M.H contributed new data and analytic tools, P.L.C, J.Z and K.V wrote the paper with input from all co-authors.

Accession Numbers

Sequence data from this article have been submitted to ENA (E-MTAB-11108).

Supplemental data

Supplemental Figure 1. Number of neighbors per gene at each density subnetwork in Arabidopsis.

Supplemental Figure 2. Expression patterns for the expression-supported intrinsic yield gene set in Arabidopsis.

Supplemental Figure 3. Expression supported IYGs with neighborhood conservation at each density level.

Supplemental Figure 4. Functional enrichment of cross-species conserved transcription factors (TF) grouped by TF family.

Supplemental Figure 5. Identification of T-DNA insertion lines.

Supplemental Table 1. Overview of the expression datasets used for the network computation

Supplemental Table 2. List of expression-supported intrinsic yield genes

Supplemental Table 3. Yield functional categories and genes in Arabidopsis

Supplemental Table 4. Intrinsic yield gene predictions

Supplemental Table 5. List of RARGE II leaf trait genes known to affect leaf phenotype if mutated

Supplemental Table 6. Top 100 intrinsic yield gene predictions annotated

Supplemental Table 7. In depth literature analysis for the top 100 intrinsic yield gene predictions

Supplemental Table 8. List of genes tested for leaf phenotype in this study

Supplemental Dataset 1. Expression datasets for Arabidopsis, maize, and aspen

Supplemental Data set 2. Triplets generated with ComPlEx

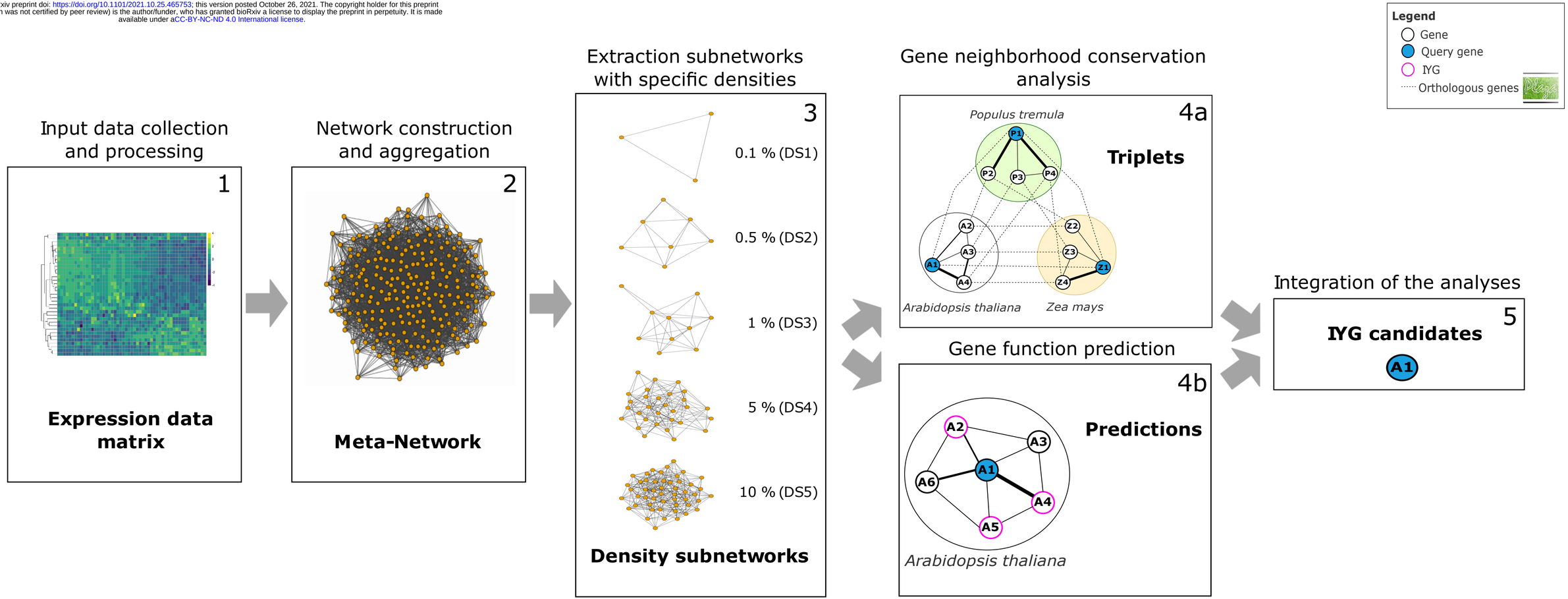
Supplemental Methods. Detailed methods for expression dataset retrieval, generation, and processing.

Acknowledgements

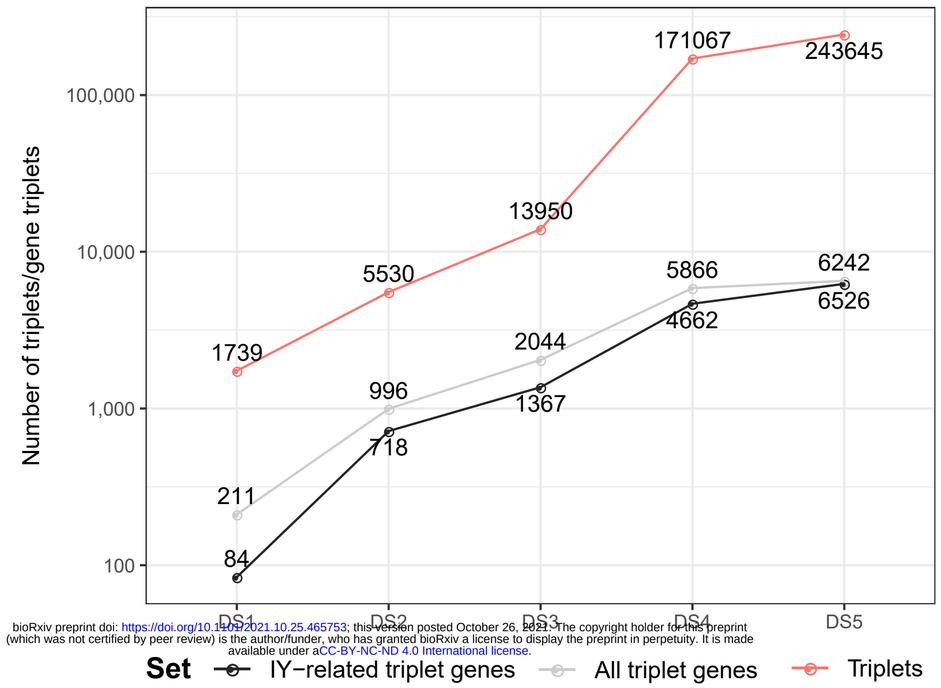
We thank Julie Pevernagie for her technical support.

Funding

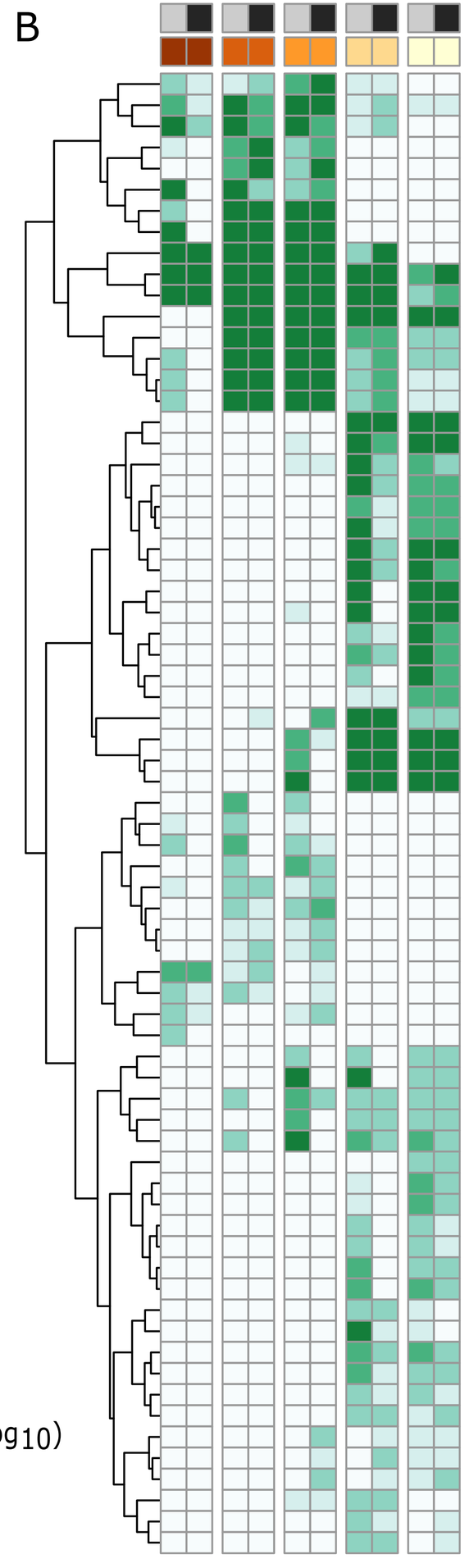
TRH, KV and NRS was partially funded by the Research Council of Norway (project number 287465).



A

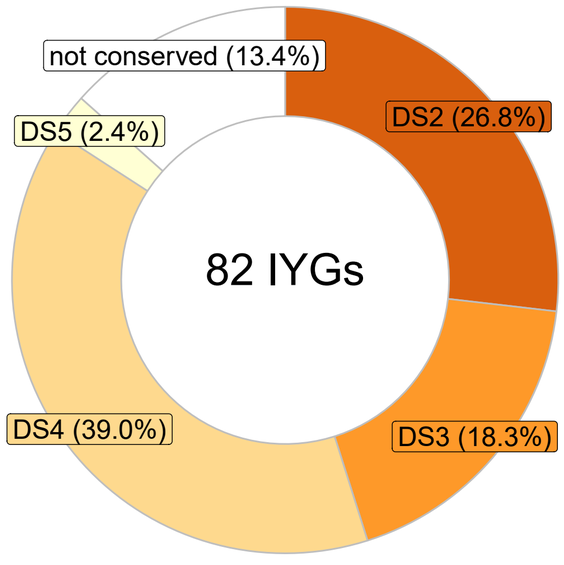


B



GO:0015977 carbon fixation
GO:0000027 ribosomal large subunit assembly
GO:0018298 protein−chromophore linkage
GO:0007088 regulation of mitotic nuclear division
GO:0051301 cell division
GO:0000028 ribosomal small subunit assembly
GO:0002181 cytoplasmic translation
GO:0007018 microtubule−based movement
∗photosynthesis
GO:0006412 translation
∗da1-eod(-)
∗IYG
GO:0000470 maturation of LSU−rRNA
GO:0000079 regulation of cyclin−dependent protein kinase activity *
GO:0008284 positive regulation of cell proliferation
GO:0045787 positive regulation of cell cycle
GO:0042742 defense response to bacterium
GO:0051704 multi−organism process
GO:0009664 plant−type cell wall organization
GO:0009733 response to auxin
GO:0006855 drug transmembrane transport
GO:0061408 pos. reg. of transcription in resp. to heat stress *
GO:0002229 defense response to oomycetes
GO:0009831 plant−type cell wall modification *
GO:0007275 multicellular organism development
GO:0032501 multicellular organismal process
GO:0006857 oligopeptide transport
GO:0042214 terpene metabolic process
GO:0005975 carbohydrate metabolic process
GO:0009808 lignin metabolic process
∗samba(+)
∗cell expansion
GO:0006355 regulation of transcription, DNA−templated
GO:0030154 cell differentiation
GO:0006414 translational elongation
GO:0000462 maturation of SSU−rRNA from tricistronic rRNA transcript *
GO:0030490 maturation of SSU−rRNA
GO:0042026 protein refolding
GO:0017148 negative regulation of translation
GO:0030261 chromosome condensation
GO:0000387 spliceosomal snRNP assembly
∗cell cycle
GO:0006006 glucose metabolic process
GO:0006325 chromatin organization
GO:0006662 glycerol ether metabolic process
GO:0009416 response to light stimulus
GO:0009813 flavonoid biosynthetic process
GO:0042343 indole glucosinolate metabolic process
GO:0006458 'de novo' protein folding
GO:0006979 response to oxidative stress
GO:0042744 hydrogen peroxide catabolic process
GO:0009809 lignin biosynthetic process
GO:0009873 ethylene−activated signaling pathway
GO:0010268 brassinosteroid homeostasis
GO:0016104 triterpenoid biosynthetic process
GO:0019953 sexual reproduction
GO:0016132 brassinosteroid biosynthetic process
GO:0051762 sesquiterpene biosynthetic process
GO:0009807 lignan biosynthetic process
GO:0045490 pectin catabolic process
GO:0009651 response to salt stress
GO:0048544 recognition of pollen
GO:0019762 glucosinolate catabolic process
∗brm101(+)
∗syd2(-)
∗da1(-)
∗JAW(-)
∗AN3/SWI/SNF
GO:0009299 mRNA transcription
∗samba(-)

C

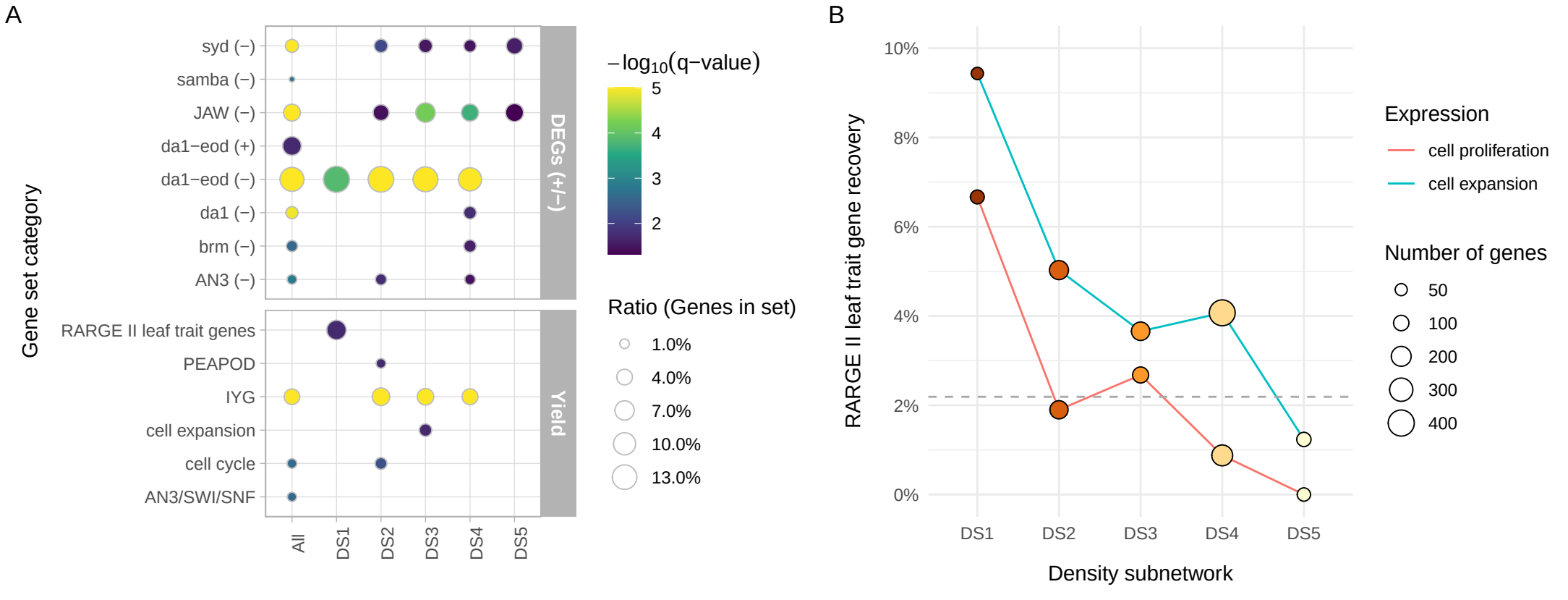


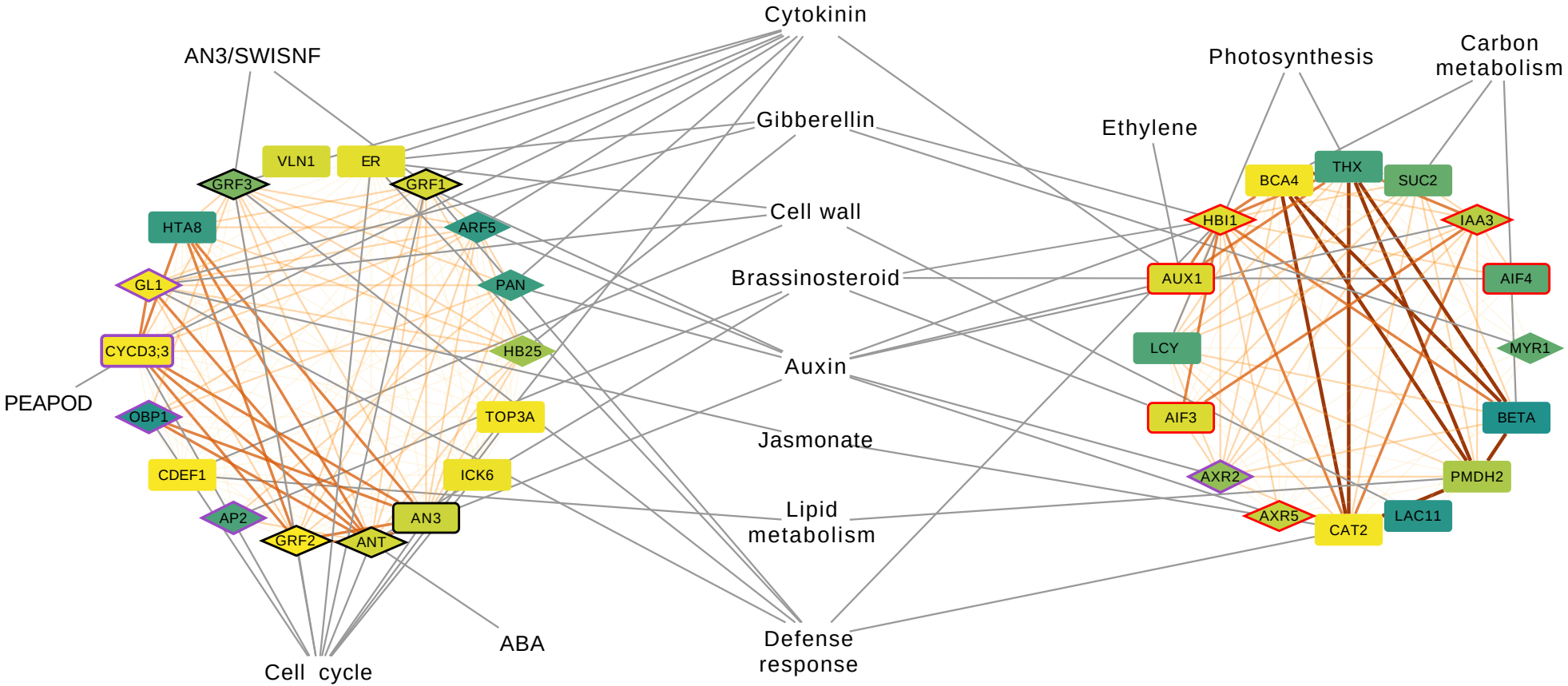
Legend

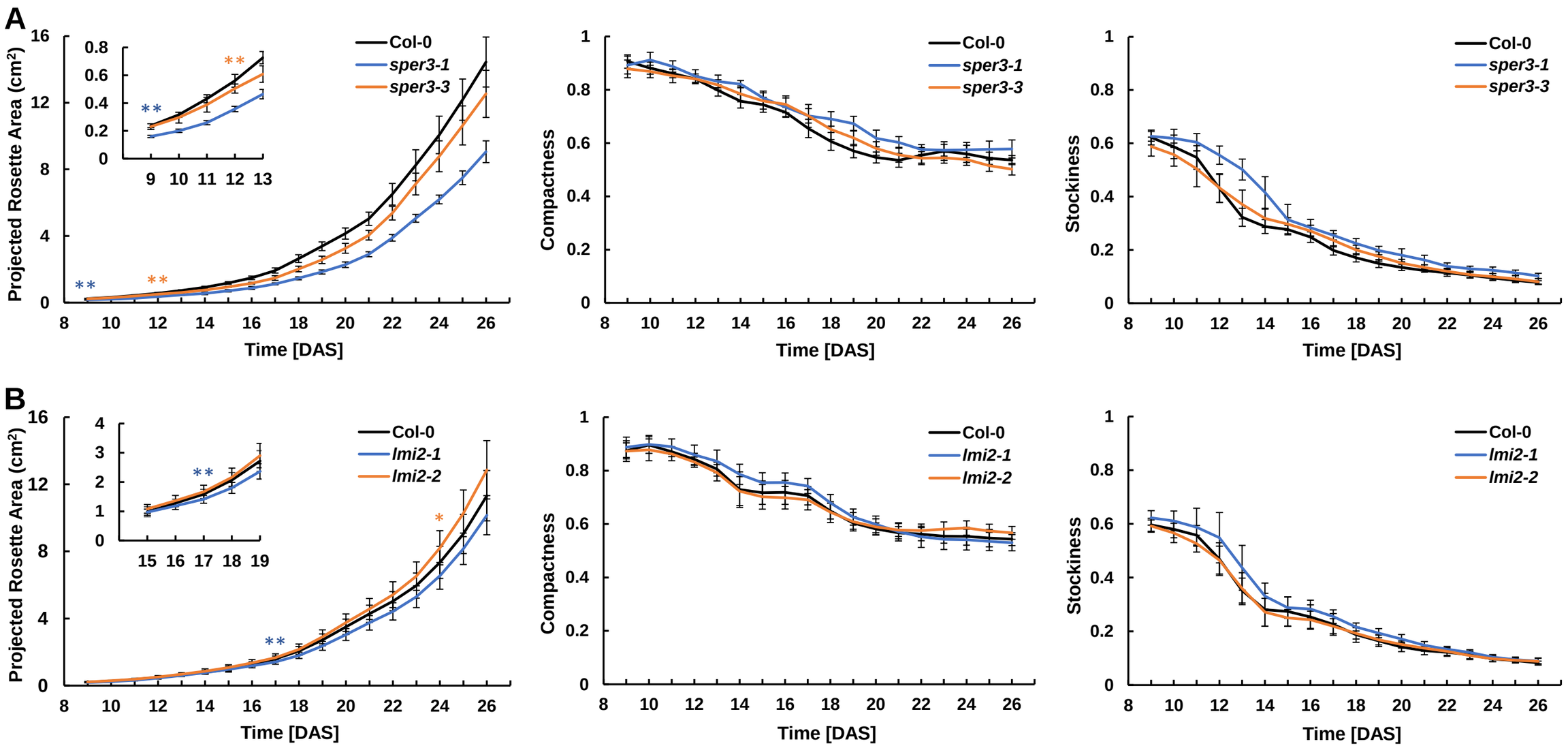
Set
All
IYG-related

Density subnetwork
DS1 DS2 DS3 DS4 DS5

GO Enrichment q-value ($-\log_{10}$)
5 4 3 2 1 0


A
Gene set category
syd (−)
samba (−)
JAW (−)
da1−eod (+)
da1−eod (−)
da1 (−)
brm (−)
AN3 (−)
DEGs (+/−)
RARGE II leaf trait genes
PEAPOD
IYG
cell expansion
cell cycle
AN3/SWI/SNF
Yield
All
DS1
DS2
DS3
DS4
DS5
−log10(q−value)
5
4
3
2
Ratio (Genes in set)
1.0%
4.0%
7.0%
10.0%
13.0%
B
RARGE II leaf trait gene recovery
10%
8%
6%
4%
2%
0%
DS1
DS2
DS3
DS4
DS5
Density subnetwork
Expression
cell proliferation
cell expansion
Number of genes
50
100
200
300
400


Cytokinin
AN3/SWISNF
Gibberellin
Photosynthesis
Carbon metabolism
Ethylene
VLN1
ER
GRF3
GRF1
Cell wall
HTA8
ARF5
BCA4
THX
SUC2
HBI1
IAA3
Brassinosteroid
GL1
PAN
AUX1
AIF4
CYCD3;3
HB25
LCY
MYR1
Auxin
PEAPOD
OBP1
TOP3A
AIF3
BETA
Jasmonate
CDEF1
ICK6
AXR2
PMDH2
AP2
AN3
AXR5
LAC11
Lipid metabolism
GRF2
ANT
CAT2
ABA
Cell cycle
Defense response


A
Projected Rosette Area (cm²)
Time [DAS]
Col-0
sper3-1
sper3-3
**
**
**
Compactness
Stockiness
B
lmi2-1
lmi2-2
*

Parsed Citations

Andriankaja, M., Dhondt, S., De Bodt, S., Vanhaeren, H., Coppens, F., De Milde, L., Mü hlenbock, P., Skirycz, A., Gonzalez, N., Beemster, G.T., and Inzé, D. (2012). Exit from Proliferation during Leaf Development in Arabidopsis thaliana: A Not-So-Gradual Process.

Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. 34: 525–527.
Google Scholar: Author Only Title Only Author and Title

Chotewutmontri, P. and Barkan, A. (2016). Dynamics of Chloroplast Translation during Chloroplast Differentiation in Maize. PLoS Genet. 12: 1–28.
Google Scholar: Author Only Title Only Author and Title

Dubois, M., Claeys, H., Van den Broeck, L., and Inz??, D. (2017). Time of day determines Arabidopsis transcriptome and growth dynamics under mild drought. Plant Cell Environ. 40: 180–189.
Google Scholar: Author Only Title Only Author and Title

Heldt, H-W. and Piechulla, B. (2021) Photosynthesis needs the consumption of water. Plant Biochemistry (Fifth edition): 191-216
Google Scholar: Author Only Title Only Author and Title

Leek, J.T., Scharpf, R.B., Bravo, H.C., Simcha, D., Langmead, B., Johnson, W.E., Geman, D., Baggerly, K., and Irizarry, R.A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. Nat. Rev. Genet. 11: 733–739.
Google Scholar: Author Only Title Only Author and Title

Mähler, N., Schiffthaler, B., Robinson, K.M., Terebieniec, B.K., Vučak, M., Mannapperuma, C., Bailey, M.E.S., Jansson, S., Hvidsten, T.R., and Street, N.R. (2020). Leaf shape in Populus tremula is a complex, omnigenic trait. Ecol. Evol. 10: 11922–11940.
Google Scholar: Author Only Title Only Author and Title

Nelissen, H. et al. (2018). The reduction in maize leaf growth under mild drought affects the transition between cell division and cell expansion and cannot be restored by elevated gibberellic acid levels. Plant Biotechnol. J. 16: 615–627.
Google Scholar: Author Only Title Only Author and Title

Schlüter, U. and Weber, A. (2019). Regulation and Evolution of C4 Photosynthesis. FASEB J. 33: 183–215.
Google Scholar: Author Only Title Only Author and Title

Schmittgen, T.D. and Livak, K.J. (2008). Analyzing real-time PCR data by the comparative CT method. Nat. Protoc. 3: 1101–1108.
Google Scholar: Author Only Title Only Author and Title

Skirycz, A., De Bodt, S., Obata, T., De Clercq, I., Claeys, H., De Rycke, R., Andriankaja, M., Van Aken, O., Van Breusegem, F., Fernie, A.R., and Inze, D. (2010). Developmental Stage Specificity and the Role of Mitochondrial Metabolism in the Response of Arabidopsis Leaves to Prolonged Mild Osmotic Stress. Plant Physiol. 152: 226–244.
Google Scholar: Author Only Title Only Author and Title

Skirycz, A., Claeys, H., De Bodt, S., Oikawa, A., Shinoda, S., Andriankaja, M., Maleux, K., Eloy, N.B., Coppens, F., Yoo, S.-D., Saito, K., and Inzé, D. (2011). Pause-and-Stop: The Effects of Osmotic Stress on Cell Proliferation during Early Leaf Development in Arabidopsis and a Role for Ethylene Signaling in Cell Cycle Arrest. Plant Cell 23: 1876–1888.
Google Scholar: Author Only Title Only Author and Title

Sun, X. et al. (2017). Altered expression of maize PLASTOCHRON1 enhances biomass and seed yield by extending cell division duration. Nat. Commun. 8: 14752.
Google Scholar: Author Only Title Only Author and Title

Thompson, J.A., Tan, J., and Greene, C.S. (2016). Cross-platform normalization of microarray and RNA-seq data for machine learning applications. PeerJ 4: e1621.
Google Scholar: Author Only Title Only Author and Title

Vaneechoutte, D. and Vandepoele, K. (2019). Curse: Building expression atlases and co-expression networks from public RNA-Seq data. Bioinformatics 35: 2880–2881.
Google Scholar: Author Only Title Only Author and Title

Wang, Y., Long, S.P., and Zhu, X.G. (2014). Elements required for an efficient NADP-malic enzyme type C4 photosynthesis. Plant Physiol. 164: 2231–2246.
Google Scholar: Author Only Title Only Author and Title

Andriankaja, M., Dhondt, S., De Bodt, S., Vanhaeren, H., Coppens, F., De Milde, L., Mü hlenbock, P., Skirycz, A., Gonzalez, N., Beemster, G.T., and Inzé, D. (2012). Exit from Proliferation during Leaf Development in Arabidopsis thaliana: A Not-So-Gradual Process.

Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. 34: 525–527.
Google Scholar: Author Only Title Only Author and Title

Chotewutmontri, P. and Barkan, A. (2016). Dynamics of Chloroplast Translation during Chloroplast Differentiation in Maize. PLoS Genet. 12: 1–28.
Google Scholar: Author Only Title Only Author and Title

Dubois, M., Claeys, H., Van den Broeck, L., and Inz??, D. (2017). Time of day determines Arabidopsis transcriptome and growth dynamics under mild drought. Plant Cell Environ. 40: 180–189.
Google Scholar: Author Only Title Only Author and Title

Heldt, H-W. and Piechulla, B. (2021) Photosynthesis needs the consumption of water. Plant Biochemistry (Fifth edition): 191-216
Google Scholar: Author Only Title Only Author and Title

Leek, J.T., Scharpf, R.B., Bravo, H.C., Simcha, D., Langmead, B., Johnson, W.E., Geman, D., Baggerly, K., and Irizarry, R.A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. Nat. Rev. Genet. 11: 733–739.
Google Scholar: Author Only Title Only Author and Title

Mähler, N., Schiffthaler, B., Robinson, K.M., Terebieniec, B.K., Vučak, M., Mannapperuma, C., Bailey, M.E.S., Jansson, S., Hvidsten, T.R., and Street, N.R. (2020). Leaf shape in Populus tremula is a complex, omnigenic trait. Ecol. Evol. 10: 11922–11940.
Google Scholar: Author Only Title Only Author and Title

Nelissen, H. et al. (2018). The reduction in maize leaf growth under mild drought affects the transition between cell division and cell expansion and cannot be restored by elevated gibberellic acid levels. Plant Biotechnol. J. 16: 615–627.
Google Scholar: Author Only Title Only Author and Title

Schlüter, U. and Weber, A. (2019). Regulation and Evolution of C4 Photosynthesis. FASEB J. 33: 183–215.
Google Scholar: Author Only Title Only Author and Title

Schmittgen, T.D. and Livak, K.J. (2008). Analyzing real-time PCR data by the comparative CT method. Nat. Protoc. 3: 1101–1108.
Google Scholar: Author Only Title Only Author and Title

Skirycz, A., De Bodt, S., Obata, T., De Clercq, I., Claeys, H., De Rycke, R., Andriankaja, M., Van Aken, O., Van Breusegem, F., Fernie, A.R., and Inze, D. (2010). Developmental Stage Specificity and the Role of Mitochondrial Metabolism in the Response of Arabidopsis Leaves to Prolonged Mild Osmotic Stress. Plant Physiol. 152: 226–244.
Google Scholar: Author Only Title Only Author and Title

Skirycz, A., Claeys, H., De Bodt, S., Oikawa, A., Shinoda, S., Andriankaja, M., Maleux, K., Eloy, N.B., Coppens, F., Yoo, S.-D., Saito, K., and Inzé, D. (2011). Pause-and-Stop: The Effects of Osmotic Stress on Cell Proliferation during Early Leaf Development in Arabidopsis and a Role for Ethylene Signaling in Cell Cycle Arrest. Plant Cell 23: 1876–1888.
Google Scholar: Author Only Title Only Author and Title

Sun, X. et al. (2017). Altered expression of maize PLASTOCHRON1 enhances biomass and seed yield by extending cell division duration. Nat. Commun. 8: 14752.
Google Scholar: Author Only Title Only Author and Title

Thompson, J.A., Tan, J., and Greene, C.S. (2016). Cross-platform normalization of microarray and RNA-seq data for machine learning applications. PeerJ 4: e1621.
Google Scholar: Author Only Title Only Author and Title

Vaneechoutte, D. and Vandepoele, K. (2019). Curse: Building expression atlases and co-expression networks from public RNA-Seq data. Bioinformatics 35: 2880–2881.
Google Scholar: Author Only Title Only Author and Title

Wang, Y., Long, S.P., and Zhu, X.G. (2014). Elements required for an efficient NADP-malic enzyme type C4 photosynthesis. Plant Physiol. 164: 2231–2246.
Google Scholar: Author Only Title Only Author and Title