bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.350058; this version posted October 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Research article

Title: Atlas of tissue-specific and tissue-preferential gene expression in ecologically and economically significant conifer *Pinus sylvestris*

Authors

Sandra Cervantes[1,2], Jaana Vuosku[1], Dorota Paczesniak[1], Tanja Pyhäjärvi*[1]

[1]Department of Ecology and Genetics, University of Oulu, Finland; [2]Biocenter Oulu, University of Oulu, Finland.

*Corresponding author

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.350058; this version posted October 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Abstract

Background: Despite their ecological and economical importance, conifers still have limited genomic resources, mainly due to the large size and complexity of their genomes. In addition, several of the available genomic resources lack complete structural and functional annotation. Transcriptomic resources have been commonly used to compensate for these deficiencies, though for most conifer species the currently available transcriptomes are limited to a small number of tissues, or capture only a fraction of the genes present in the genome.

Results: Here we provide an atlas of gene expression patterns for conifer *Pinus sylvestris* grown under natural conditions across five tissues: embryo, megagametophyte, needle, phloem, and vegetative bud. Compared to previous studies, we used a wider range of tissues and focused our analyses on the expression profiles of genes at tissue level. We provide comprehensive information of the per-tissue normalized expression level, and indication of tissue preferential upregulation or tissue preferential expression. We identified a total of 48,001 tissue preferentially upregulated and tissue specifically expressed genes, of which 28% have annotation in the Swiss-Prot database. The annotated genes were associated with a total of 84,498 GO terms, of which 1,834 had significant enrichment in different processes and functions, for example glyoxylate cycle in megagametophyte and defense response in needle. Even though most of the genes originating from the transcriptome do not have functional information in current biological databases, the tissue-specific patterns identified here provide valuable information about their potential functions for further studies.

Conclusions: The genes identified in this study will contribute to improve the annotation of the already available and forthcoming conifer genomes. This atlas of gene expression also provides ground to further the research in the areas of plant physiology, population genetics, and genomics in general. As we provide information on tissue specificity at both diploid and haploid life stages, our

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.350058; this version posted October 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

data will also contribute to the understanding of evolutionary rates of different tissue types and ploidy levels.

Keywords: Scots pine, RNA-seq, *Pinus sylvestris*, tissue-specific gene expression, conifer, transcriptomics, needle, phloem, megagametophyte, embryo, vegetative bud

Background

Conifers, a clade within the gymnosperms, represent a group of plants with significant economic and ecological relevance [1]. Several coniferous trees are among the most important sources of wood and timber, as for example *Pinus* and *Picea* [2, 3]. Conifers dominate boreal forests worldwide and can form large forested areas hosting a variety of ecosystems. Furthermore, conifer forests are one of the major ecosystem services providers and they are crucial for carbon sequestration [2, 4–6]. Despite their importance, genomic resources for conifers, and gymnosperms in general, lag behind in availability compared to angiosperms. Although several contributions have been made recently to fill this gap [7–11], conifer genome annotation remains a challenge, with both structural and functional annotations being far from perfect [12, 13]. Conifer genomics resources are limited due to the large size of their genomes, ranging from 8 to 70 Gbp [14] and to the large number of repetitive elements (approximately 80%) within them [7, 15, 16]. Proper and complete annotation of the conifer genomes has also been complicated by the presence of long introns [7, 13], which prevents the routine use of common annotation software. Moreover, analyses of ortholog genes across different species indicate that there are several gene groups which are unique to conifers or conifer species specific, with no well-defined homologs in any of the angiosperm plant models [7, 13, 16, 17].

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.350058; this version posted October 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Transcriptomic resources have been particularly important for research in conifers and other non-model species, as a strategy to compensate for the challenges associated with efficient genome assembly and annotation [12, 18]. As the biological functions can not be directly inferred from nucleotide sequences, reference transcriptomes and gene expression studies are useful in the identification and annotation of genes [13, 19–22]. Transcriptome information can also be used in conifers that lack reference genomes, as this information can be used in the design of reduced genome representation targets [23, 24]. In addition to this, RNA-seq analyses allow the identification of expression patterns and expression levels, which are essential components of evolutionary genomics studies. For example, selective constraints in genes can be inferred from their expression patterns, as both breadth and expression level are known determinants of evolutionary rates [25, 26]. Selective constraints are also expected to differ between haploid and diploid tissues which differ in the relative rate of expression, as tissue specificity and ploidy has potentially drastic effects on the dynamics of e.g. purifying selection [27].

Here we give a first glimpse of the expression patterns of tissue preferentially upregulated (PUR) and tissue specifically expressed genes across five tissues (embryo, megagametophyte, needle, phloem, and vegetative bud) of *Pinus sylvestris*. *P. sylvestris* is a widely distributed conifer of large economic and ecological importance in Northern Eurasia [28]. *P. sylvestris* is one of the main sources of timber and raw material for the pulp and paper industry in Europe and is a dominant species in boreal forests, with an estimated coverage area of 145 millions hectares [28]. *P. sylvestris* is also a suitable model to answer evolutionary and genetic questions, especially regarding gymnosperm reproductive biology, its evolution and genetic consequences. For example, in conifers the maternal nuclear haplotype of an embryo is identical to the megagametophyte's nuclear haplotype [29], which makes it possible to separate expression of paternal and maternal haplotypes and alleles in the embryo [30].

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.350058; this version posted October 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Despite its importance and potential, *P. sylvestris* still lacks a reference genome, and currently there are limited genomic resources for this species (see however [20, 31–35]). To date, the few transcriptomic studies of *P. sylvestris* have been based on a small number of tissue types such as needles or seed tissues [20, 32]. Identification of tissue preferentially upregulated and tissue specific genes is relevant because 1) understanding the different patterns of expression across different kinds of tissues can aid to elucidate the organization of transcriptomes [19]. 2) Knowing the different profiles of expression across tissues can set the ground for evolutionary analysis, as it is known from studies in mammals and angiosperms that the evolution of gene expression differs across tissues or organs [36, 37]. Ultimately this knowledge will help to gain a deeper understanding of the determinants and main factors that affect the rate of adaptive evolution and the dynamics at the genome level.

In this study we 1) provide a comparative transcriptomic resource for *P. sylvestris* describing the expression level in five different tissues, 2) identify genes that are tissue preferentially upregulated and tissue specifically expressed in each of the five tissues, 3) provide quantitative measures of tissue-specific expression for each gene per tissue combination, and 4) conduct gene ontology enrichment analysis for each tissue type. Our results are important for future studies in comparative conifer genomics, plant physiology, population genetic analyses, evolutionary genetic studies, further gene expression analyses, and aid in the annotation of present and forthcoming conifer genome sequences.

Results and discussion

Transcript quantification and abundance matrices construction

We mapped a total of 707,063,773 trimmed and adapter removed reads from five different tissues (embryo, megagametophyte, needle, phloem, and vegetative bud) and six biological replicates (six

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.350058; this version posted October 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

different genotypes) per tissue type to *P. sylvestris* TRINITYguided transcriptome [33]. On average 23,568,792 reads originated from each tissue, ranging from 29,591,629 reads for needle to 20,469,80 reads for phloem. On average 76% of the reads per replicate were successfully mapped to the reference (Table S1). After mapping 1,307,500 contigs had aligned reads at the isoform level. Of those, 120,040 contigs were removed from the downstream analyses as they were identified as contaminants (Data S1). The final set consisted of 1,187,460 contigs at isoform level and were used to construct raw counts and normalized matrices at gene level for downstream analyses (see Methods section). The total number of putative genes with expression signal in the gene level matrices was 715,398, much higher than the number of annotated genes in any conifer [7, 16, 38]. This magnitude, albeit probably an overestimate, is typical to transcriptome studies [21]. This is likely a result of single genes being present in multiple fragments, isoforms split into multiple genes, and different alleles originating from heterozygous material identified as separate genes during assembly and classification as genes by Trinity [33]. However, part of the genes originate from gene families and since clustering similar genes is possible in downstream analysis, we chose to err on the side of potentially over splitting the genes rather than imperfectly clustering similar transcripts as a single gene, as over clustering will inherently lead to loss of information. We believe that providing expression data with minimum clustering will be most versatile for later use of the transcriptome and expression data in genome annotations and other studies.

Quality assessment of biological replicates

As we used different genotypes as biological replicates, we first verified that the replicates clustered by tissue type and not by genotype, and checked for the presence of potential outliers in the dataset. We used the raw counts matrix data (Table S2), a principal component analysis (PCA) and a Pearson correlation to verify this. The PCA separated the tissue samples into five distinct clusters

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.350058; this version posted October 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

without any overlap, indicating that among-tissue variation is the main factor of among-sample variation (Figure 1). Hence, our approach captures the differentiating gene expression profiles of the five tissues. In the PCA, the seed-derived megagametophyte and embryo samples clustered closest to each other, suggesting similarity in their gene expression profiles. Also phloem and bud samples clustered close to each other, whereas needle samples showed the most unique gene expression profile. In the hierarchical clustering analysis, based on the correlations of gene expression profiles, the differences among tissues are relatively shallow. But, similarly to the PCA, all replicates are clustered according to their tissue type and not according to their genotypes, corroborating the PCA results (Figure S1).

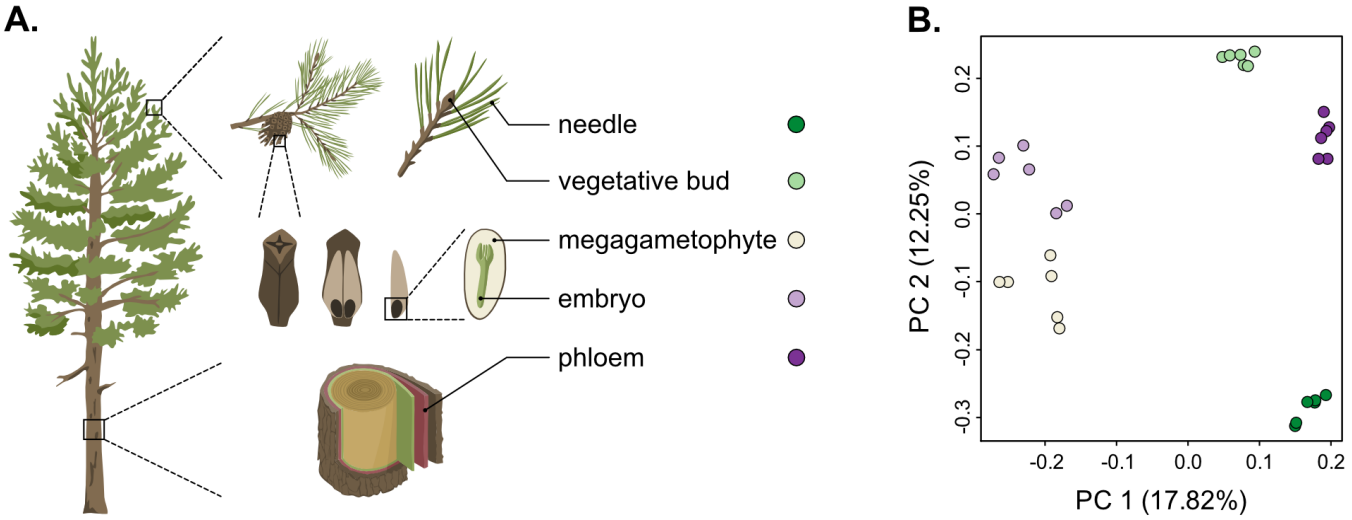

Figure 1. A. Schematic representation of the five tissues used in the transcriptome profiling of *Pinus sylvestris*: needle, vegetative bud, megagametophyte, embryo and phloem. B. Scatterplot of the first two axes of the principal component analysis (PCA). Tissue types are denoted by colors.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.350058; this version posted October 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Tissue preferentially upregulated and tissue-specific gene expression

We defined a gene as tissue PUR when there was a significant log fold change in the expression value compared to the other tissues. To identify tissue PUR we first did a differential expression (DE) analysis. For this we included all the genes in the raw count matrix (Table S2). We decided not to apply any minimum number of counts per gene as a filtering threshold to run the analysis, as we later applied a 5% false discovery rate (FDR) threshold for the identification of PUR genes. Out of the 715,398 genes initially included in the DE analysis, 198,413 genes had a maximum 5% FDR for differential expression and were further included in the analysis to identify PUR genes. We identified a total of 48,001 genes with tissue preferential expression, and out of the five tissues needle has the highest number of PUR genes (Table 1)

Quantification of tissue specificity allows a powerful statistical analysis of correlation between tissue-specific expression and e.g. evolutionary rate or other dependent or explanatory variables and factors. We identified the tissue specifically expressed genes by calculating the τ score per gene. The score ranges from zero to one, with a zero given to genes expressed in all tissues and one given to completely tissue specific genes. For this analysis we retained a set of 177,075 genes (Table S3) after applying the filtering criteria described in Methods. We considered a gene as tissue specifically expressed only if its τ=1. We identified a total of 3,899 genes with a tissue-specific pattern of expression, similarly the PUR analysis results, needle has the highest number of tissue-specific genes (Table 1). To obtain the annotation of the genes identified as tissue PUR and tissue specific, we retrieved the corresponding UniProtKB identifiers [39] from the Trinotate for the 715,398 putative genes in the TMM count matrix, out of which 97,435 (14 %) had a Swiss-Prot [40] protein match based on BLASTX [33]. Most of the Swiss-Prot annotations (67%) originated from *Arabidopsis thaliana* (65,214 genes). Other common annotation sources were *Nicotiana tabacum* (9,794; 10%) and *Oryza sativa* (8,946; 9%). Only 1663 genes (1.7%) had an annotation to other

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.350058; this version posted October 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Pinus species, of which 177 (10.6%) were hits to *P. sylvestris*, and 608 (36.5%) genes had Swiss-Prot annotation to *Picea*. Note that Swiss-Prot is a manually curated database that does not currently have a comprehensive set of annotated gymnosperm proteins and therefore the best matches are often obtained from the model plants such as *A. thaliana*. A proportion of our putative genes share the same gene identifier (annotation) (Table 1). This probably reflects the incomplete collapse of different isoforms in the assembled transcriptome used as reference, or the presence of gene families [13]. Also, a high number of the genes identified as PUR or tissue specific lack annotation altogether, which is not surprising as genes with higher tissue-specific expression have less conserved sequences and are less likely to find orthologs among other species [19, 41]. A summary of the 715,398 genes indicating their normalized expression level (TMM), τ score, tissue specificity status, PUR status, and annotation can be found in the Supplementary information (Table S4).

Cursory inspection of annotations of highly expressed tissue PUR and tissue-specific genes are congruent with some of the already known functions of the tissues. These results confirm that our analyses capture biologically meaningful characteristics of the tissues. For example in megagametophytes, enzymes related to seed storage lipid mobilization and germination were upregulated and specifically expressed. Similarly, in needles, several chlorophyll a-b binding proteins are upregulated. In embryo, multiple ribosomal proteins and other proteins indicating active protein synthesis were upregulated. In vegetative buds, expression of genes involved in defense against insect attack, like (-)-alpha-pinene synthase and dirigent [42] that take part in oleoresin synthesis, were highly expressed and specific to this tissue. In phloem, the two genes annotated as metallothionein-like protein EMB30, an aquaporin and a thioredoxin-like protein were highly expressed, similarly to *Quercus suber* phellem (cork) where metallothionein reacts to oxidative stress [43] or in *Pinus taeda* xylem where the same proteins were among the most highly expressed genes [44].

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.350058; this version posted October 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

tissue

10 most enriched/significant biological processes

PROCESSES

10 most highly expressed genes
in bold: preferentially upregulated genes

GENES

mega-gametophyte

PROCESSES

- GO:0006097 Glyoxylate cycle
- GO:0046487 Glyoxylate metabolic process
- GO:0044242 Cellular lipid catabolic process
- GO:0009109 Coenzyme catabolic process
- GO:0006099 Tricarboxylic acid cycle
- GO:0046356 Acetyl-CoA catabolic process
- GO:0009062 Fatty acid catabolic process
- GO:0019395 Fatty acid oxidation
- GO:0006635 Fatty acid beta-oxidation
- GO:0016042 Lipid catabolic process

GENES

Antimicrobial peptide 1
Metallothionein-like protein EMB30
Cysteine proteinase 15A
Antifungal protein ginkbilobin-2
Glycine-rich RNA-binding protein GRP1A
Malate dehydrogenase
Non-specific lipid-transfer protein 1
3-ketoacyl-CoA thiolase 2, peroxisomal
Isocitrate lyase
Catalase

vegetative bud

PROCESSES

- GO:0071555 Cell wall organization
- GO:0071554 Cell wall organization or biogenesis
- GO:0009698 Phenylpropanoid metabolic process
- GO:0009699 Phenylpropanoid biosynthetic process
- GO:0071669 Plant-type cell wall organization or biogenesis
- GO:0005976 Polysaccharide metabolic process
- GO:0070882 Cellular cell wall organization or biogenesis
- GO:0042546 Cell wall biogenesis
- GO:0044264 Cellular polysaccharide metabolic process
- GO:0010383 Cell wall polysaccharide metabolic process

GENES

Glycine-rich RNA-binding protein GRP1A
Chlorophyll a-b binding protein 16, chloroplastic
Metallothionein-like protein EMB30
Abscisic stress-ripening protein 1
Cold shock domain-containing protein 4
Cysteine proteinase 15A
Major pollen allergen Bet v 1-A
S-adenosylmethionine decarboxylase proenzyme
Non-specific lipid-transfer protein
High mobility group B protein 2

embryo

PROCESSES

- GO:0009451 RNA modification
- GO:0006412 Translation
- GO:0042254 Ribosome biogenesis
- GO:0022613 Ribonucleoprotein complex biogenesis
- GO:0006260 DNA replication
- GO:0016554 Cytidine to uridine editing
- GO:0006261 DNA-dependent DNA replication
- GO:0009813 Flavonoid biosynthetic process
- GO:0016553 Base conversion or substitution editing
- GO:0010467 Gene expression

GENES

Glycine-rich RNA-binding protein GRP1A
Cold shock domain-containing protein 4
60S ribosomal protein L36-1
Metallothionein-like protein EMB30
Translationally-controlled tumor protein homolog
Non-specific lipid-transfer protein
Chlorophyll a-b binding protein 1A, chloroplastic
Non-specific lipid-transfer protein 2
Ribulose bisphosphate carboxylase small chain, chloroplastic
60S acidic ribosomal protein P1

needle

PROCESSES

- GO:0015979 Photosynthesis
- GO:0042440 Pigment metabolic process
- GO:0006952 Defense response
- GO:0046148 Pigment biosynthetic process
- GO:0019684 Photosynthesis, light reaction
- GO:0009657 Plastid organization
- GO:0009658 Chloroplast organization
- GO:0009765 Photosynthesis, light harvesting
- GO:0009617 Response to bacterium
- GO:0019748 Secondary metabolic process

GENES

Metallothionein-like protein type 3
Ribulose bisphosphate carboxylase small chain, chloroplastic
Chlorophyll a-b binding protein 16, chloroplastic
Chlorophyll a-b binding protein type 2 member 1B, chloroplastic
Galactinol synthase 4
RuBisCO activase 1, chloroplastic
Beta carbonic anhydrase 4
Chlorophyll a-b binding protein type I, chloroplastic
Carbonic anhydrase, chloroplastic
Chlorophyll a-b binding protein 1A, chloroplastic

phloem

PROCESSES

- GO:0006952 Defense response
- GO:0050896 Response to stimulus
- GO:0051707 Response to other organism
- GO:0006950 Response to stress
- GO:0006468 Protein amino acid phosphorylation
- GO:0009607 Response to biotic stimulus
- GO:0042221 Response to chemical stimulus
- GO:0051704 Multi-organism process
- GO:0046777 Protein amino acid autophosphorylation
- GO:0070887 Cellular response to chemical stimulus

GENES

Metallothionein-like protein EMB30
Glycine-rich RNA-binding protein GRP1A
Non-specific lipid-transfer protein 4
Cysteine proteinase 15A
Translationally-controlled tumor protein homolog
17.8 kDa class I heat shock protein
Cold shock domain-containing protein 4
Catalase
Granule-bound starch synthase 1, chloroplastic/amyloplastic
Dormancy-associated protein 1

204

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.350058; this version posted October 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure 2. Ten most significantly enriched biological processes (with corresponding GO-term IDs) and ten most highly expressed annotated genes in each of the five tissues. Genes preferentially upregulated (PUR) in a given tissue are in bold.

Among the five tissues analysed, the needle had the highest number of genes with tissue-specific expression and embryo the lowest (Table 1). Except for two genes, one in megagametophyte and one in needle, all the genes with tissue-specific expression were also among the PUR genes. However, as tissue specificity does not require a high expression level, genes with τ score equal to one are not necessarily the most upregulated genes in their respective tissues. Comparison of our findings to other studies is not straightforward as there are very few transcriptomic studies in *P. sylvestris*. But in comparison to a previous study [20], where they focus on the comparison between megagametophyte and embryo tissues at different developmental stages, we identified less megagametophyte and embryo specifically expressed genes. One of the reasons for this difference could be that the identification of unique genes in the previous study [20] was based only on the comparison between embryo and megagametophyte tissues. As the identification of tissue specific genes is contingent to the number of tissues used for the analysis, it is expected that the higher the number of tissues used in the comparison, the lower the number of tissue specific genes that will be identified. In contrast, we found a higher number of tissue specific genes in embryo, bud, and needle compared to a previous study in conifers [19], where several tissue types were used. One notable difference between this [19] and ours was the higher number of tissue-specific genes for megagametophyte found in *P. glauca.* This analysis [19] found the highest number of unique genes in the megagametophyte in comparison to other tissues analyzed. The low number of megagametophyte specific genes identified in our study could be due to the use of mature embryos as starting material. Previous research suggests that the number of unique transcripts in the megagametophyte varies during the developmental stages of embryogenesis [20].

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.350058; this version posted October 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

One caveat of our analyses is that, unlike other studies, we did not use microdissection in order to obtain the tissue samples [22]. Hence, some of the "tissues" are a mix of tissue types. Needles, for example, include several tissues (phloem among them) [45], and mature embryos contain the shoot and root meristems as well as cotyledons [46]. In contrast, the mature megagametophyte is a quite uniform storage tissue consisting of cells packed with starch protein and lipids [47, 48]. Another limitation of the dataset is that it represents only one point in time and space, although gene expression is a dynamic process and quantitative and qualitative variations exist over spatial and temporal scales. Instead of sampling across several developmental stages or across a spatial gradient our dataset represents a wider set of tissues, which increases the power to identify tissue PUR and tissue specifically expressed genes. The added value of the dataset lies in the unexpected functions and connections discovered among biological pathways and genes with previously unidentified signals of tissue-specificity or up-regulation.

Table 1. Number of genes identified as tissue preferentially upregulated and tissue-specific in five *P. sylvestris* tissues. The percentage of unique UniProtKB identifiers is also shown.

	Tissue preferentially upregulated genes			Tissue specifically expressed genes		
	Total	Annotated	Unique (%)	Total	Annotated	Unique (%)
Bud	8225	2515	30.6	693	342	49.3
Embryo	10430	2820	27.0	498	206	41.3
Megagametophyte	7171	1515	21.1	679	220	32.4
Needle	13128	3993	30.4	1495	603	40.3
Phloem	9047	2603	28.7	534	202	37.8

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.350058; this version posted October 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Functional characterization of tissue preferentially upregulated and tissue-specific genes

GO enrichment analysis allows the identification of gene functions enriched with certain functional roles. The number of enriched functions was of the same magnitude across tissue types, ranging from 253 to 452 for PUR genes and from 58 to 169 for tissue-specific genes (Tables S5-S14). The total number of GO terms and the number of significant enriched terms per tissue are shown in Table 2, a summary of the most highly expressed genes per tissue, and the most significantly enriched biological processes is shown in Figure 2. Most of the genes (86%) with expression signals in our study lacked annotation from the Trinotate pipeline. Thus, they did not contribute to functional analysis or GO enrichment results.

The complete lists of gene identifiers and their corresponding GO terms per tissue and per each set of genes (Data S2-S11), along with tables with the results of the SEA showing each GO terms, its p-value, and FDR (Table S5-S14) are provided Supplementary information.

Table 2. Total number and number of significant GO terms and percentage of enriched terms in *P. sylvestris* tissues.

	Tissue preferentially upregulated genes			Tissue-specific genes		
	Total	Significant	Percentage (%)	Total	Significant	Percentage (%)
Bud	15681	452	2.9	2019	137	6.7
Embryo	17461	253	1.4	1178	75	6.4
Megagametophyte	9690	306	3.1	1363	111	8.1
Needle	25295	401	1.6	3818	169	4.3
Phloem	16371	422	2.6	1249	58	4.6

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.350058; this version posted October 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

In needles the significant GO terms reflected the exposure of trees to various stresses and interactions with other organisms, whereas in embryos, buds and the phloem significant GO terms were mainly connected to different development-related processes. In needles the enriched biological process GO terms among tissue-specific genes were immune response (GO:0006955) as well as response to stress (GO:0006950) and other organisms (GO:0051707) such as oomycetes (GO:0002229), bacteria (GO:0042742) and fungi (GO:0009817). Moreover, terpene synthase activity (GO:0010333), which may play a key role in the defense against herbivores [49], was an enriched molecular function among tissue-specific genes in needles, but also in embryos and vegetative buds. For example, reactive oxygen species (ROS) related biological processes (GO:0006800, GO:0042743, GO:0034614) and molecular functions (GO:0004601, GO:0004364) were enriched among the tissue-specific genes in embryos consistent with an active ROS protection in developing tissues. In the phloem, a special differentiation process, syncytium formation (GO:0006949), indicating the interconnection of phloem sieve elements to generate a transport route [50] was an enriched biological process among the tissue specific genes.

Megagametophyte-specific genes have crucial functions in seed germination and energy conversion

Gymnosperms are characterized by the haploid female gametophyte tissue, the megagametophyte, which surrounds the embryo in developing and mature seeds. The megagametophyte can be considered a functional homolog of the endosperm in angiosperms due to its role as a nourishing tissue [51, 52]. However, the megagametophyte develops from a haploid megaspore before the fertilization [46] and is therefore entirely maternally inherited unlike the diploid or triploid endosperms of biparental origin [53, 54]. To give an example of the potential uses of the dataset, we provide a more detailed description of the megagametophyte expression profile, but leave the in-depth analysis of the other tissues for later investigations.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.350058; this version posted October 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Among highly expressed and up-regulated genes in the megagametophyte were malate synthase (EC 2.3.3.9) and isocitrate lyase (EC 4.1.3.1) that are essential in glyoxylate cycle converting lipids into carbohydrates in seeds [55], as well as other glyoxysomal proteins like Acetyl-CoA acyltransferase (EC 2.3.1.16), ABC transporter and peroxisomal fatty acid beta-oxidation multifunctional protein AIM1 [56] . Seed storage related genes such as 2S seed storage-like protein, 11S globulin seed storage protein 2 and 13S globulin basic chain and some isocitrate lyase copies were completely megagametophyte-specific (τ=1). Antimicrobial and antifungal protein coding genes were the most highly expressed among annotated megagametophyte-upregulated genes.

The enriched GO terms of biological processes and molecular functions in the megagametophyte tissue-specific genes included seed germination and the mobilization of nutrient reserves. Nutrient reservoir activity (GO:0045735) indicated the mobilization of energy sources from the megagametophyte for seed germination and early seedling growth, as well as lipid catabolic processes (e.g. GO: 0016042, GO:0044242). Malate dehydrogenase activity (GO:0016615) and heme binding (GO:0020037), which mostly originated from the cytochrome P450 enzymes containing heme cofactors [57], reflected the resume of active metabolism. Also, response to ROS (GO:0034614) and antioxidant activity (GO:0016209) suggested active metabolism and signaling. ROS are natural by-products of metabolism and may be detrimental to seed viability because they can cause oxidative stress. However, in the seed ROS also work as signals which underpin the breaking of dormancy and provide protection against pathogens [58]. Megagametophyte cells showed responses to hormone stimulus (GO: 0032870) and the function of hormone-mediated signaling pathways (GO:0009755) including abscisic acid (GO:0009738), auxin (GO:0009734) and ethylene (GO:0009873) which also belong to the molecular networks regulating seed dormancy and germination [59–62]. Cellulose biosynthetic process (GO:0030244) and primary cell wall biogenesis (GO:0009833) suggest that cell walls in the megagametophyte may participate in water retention and give mechanical support to the germinating embryo [63]. Similarly to previous findings in *P.*

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.350058; this version posted October 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

sylvestris [20] megagametophytes, we found enrichment for processes involved in the response to chemical and endogenous stimuli (GO:0042221, GO:0071495). Merino et al. (2016) [20] suggested that the megagametophyte could also be involved in the regulation of the embryo development through the induction of signaling pathways triggered by sensing environmental signals in a similar way the angiosperms' endosperm does [64]. Altogether, our findings show that the megagametophyte is not just a reserve nutrition for the germinating embryo, but a metabolically active tissue contributing in multiple ways to seed germination and, thus, underline the importance of the haploid stage in *P. sylvestris* life cycle.

Several enzymes widely used in allozyme-based population genetic studies ([65] and references therein) such as aconitate hydratase (EC 4.2.1.3), malate dehydrogenase (EC 1.1.1.37) and aspartate aminotransferase (EC 2.6.1.1) were megagametophyte-specific and among the top 50 expressed genes in the tissue. As they may be more prone to natural selection against recessive deleterious variants when expressed at the haploid stage, early population genetic analyses may have bias in e.g. estimates of the overall genetic diversity based on these loci.

Conclusions

We provide a widely and interdisciplinary applicable genome-wide atlas of tissue-level transcription patterns based on RNA-seq for economically and ecologically significant coniferous tree *P. sylvestris*. Quantitative data and analysis of expression level, as well as breadth and tissue specificity are provided for 715,398 different putative genes. The mapping and bioinformatic analyses of gene expression are based on the most complete and high-quality reference transcriptome of *P. sylvestris* available to date [33]. Previous transcriptome studies of *P. sylvestris* have concentrated on a narrow set of tissues in each study such as wood [66], embryo [20], and needles [32, 67] or focused on a limited set of genes [68]. The present study allows comparison

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.350058; this version posted October 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

across a wide set of genes expressed in the above-ground parts of adult *P. sylvestris* trees growing in a natural forest.

In addition to genome sequence annotations, we foresee multiple potential uses for the dataset. Level and breadth of gene expression are known to be linked to the evolutionary rate and level of conservation (ref). By combining our data with similar data in other conifers or angiosperms it is possible to study the evolutionary conservation of expression patterns, or the differences in evolutionary rates across tissue-specific expression levels and gain a deeper understanding of the determinants and main factors affecting e.g. rate of adaptive evolution and dynamics at the genome level. The response of trees to a combination of different stresses is unique and cannot be directly extrapolated from studying only single stressors in experimental conditions [69]. The transcriptome resource for adult *P. sylvestris* trees growing under natural conditions, where they are simultaneously exposed to a number of different abiotic and biotic stresses as well as interactions with other organisms, provides a valuable tool also for physiological studies. Finally, un-annotated conifer genes with high expression or tissue specificity can open up whole new research avenues, independent of the previously available knowledge based on angiosperm model plants such as *A. thaliana* and *Populus*.

Methods

Plant material

During the growing season of 2016 (May 26th-27th), we sampled needles, phloem, and vegetative buds (called tissues for brevity, but see results and discussion section) from six non-related adult *Pinus sylvestris* trees growing at the Punkaharju Intensively Studied Site (ISS) [70] in Southern Finland, resulting in total of 30 samples (Table S15). The same plant material and sequenced

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.350058; this version posted October 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

libraries have been used previously to assemble multiple reference transcriptomes of *P. sylvestris [33]* and a more detailed description of the plant material and RNA extraction procedure is described by Ojeda et al. [33]. Needle, phloem, and bud samples were stored immediately in RNAlater (Thermo Fisher Scientific) or frozen *in situ* and transported to the storage on dry ice. Samples were stored at -80°C (samples in dry ice) or -20°C (samples in RNAlater) until RNA extraction. We obtained megagametophyte and embryo tissues by dissecting mature seeds collected from the same mother trees from which the vegetative tissues were obtained. Seeds were stored in the dark at 4°C until germination was induced by exposure to moisture and continuous light for 48 h.

RNA isolation, library preparation and sequencing

We extracted total RNA from needle, bud, and phloem using the Spectrum Plant Total RNA Kit (Protocol B, Sigma). Total RNA extraction was followed by mRNA capture with the NEBNext® Poly(A) mRNA Magnetic Isolation Module (New England Biolabs Inc.). For embryo and megagametophyte, mRNA was directly extracted from the whole tissues with Dynabeads mRNA Direct Micro Kit (Thermo Fisher Scientific) according to manufacturer's protocol, except for using 200 µl of lysis buffer. RNA concentration was quantified with Qubit RNA HS Assay kit (Thermo Fisher Scientific). We prepared a total of 30 libraries using the NEBNext Ultra Directional RNA Library Prep Kit for Illumina (New England Biolabs Inc.). We selected an insert size of 300 bp by using a fragmentation time of 5-12 min, followed by size selection with 40-45 µl / 20 µl AMPure XP beads (Agencourt). Libraries were single indexed with NEBNext Multiplex Oligos Set 1 for Illumina, and finally enriched with 12-15 PCR cycles. We quantified the libraries and visually checked the fragment size distribution before sequencing. We used paired-end (2 x 150 bp) and sequenced five

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.350058; this version posted October 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

pools of 6 to 12 libraries on an Illumina NextSeq 500 at the Biocenter Oulu Sequencing Centre (Oulu, Finland).

Transcript quantification

We used trimmed reads (BioProject PRJNA531617) as input for transcript quantification, using the $\text{Trinity}_{\text{guided}}$ [8] as a reference transcriptome (Data S12). We followed the Trinity Post-Transcriptome Assembly Downstream Analyses pipeline (Trinity v. 2.6.6) [71, 72] to generate quantification files at isoform level, and raw counts and normalized count matrices at putative gene level (hereafter referred as gene level matrices). We first obtained transcript abundance independently for each of the six individuals in each one of the five tissues. This was done by pseudo-aligning the RNA-seq reads to the transcriptome reference with Salmon 0.9.1 [73] using the --SS_lib_type (strand specific) and --trinity_mode options. The --trinity_mode option allowed the estimation of counts from isoforms to generate counts at a putative gene level during the count matrix generation step. Before any further analysis, we checked for the presence of possible contaminants by searching contigs that had hits to the keywords ‘alveolata’, ‘metazoa’, ‘fungi’, ‘bacteria’, and ‘archaea’. We search for exact matches to these keywords from the results of a translated blast (BLASTX) of the transcriptome annotation file [33, 39]. We then combined our list of putative contaminants with the contaminants and organelles contigs lists reported in Ojeda et al. [33], and excluded them from the isoform quantification files and the gene_trans_map. Contaminants were removed after the pseudo-aligning stage to avoid the false mapping of contaminant reads to non-contaminant contigs in the reference transcriptome.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.350058; this version posted October 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Abundance matrices construction

We built three count matrices with the Trinity pipeline (abundance_estimates_to_matrix.pl) at the gene level based on the cleaned independent transcript quantification. First, we generated a gene level raw counts matrix (Table S2), which was then used to construct a transcript per million length normalized gene count matrix (TPM escalated matrix) (Table S16). The TPM escalated matrix accounts for differences in isoform lengths that otherwise could inflate FDR due to differential transcript usage [74] . Finally, the TPM escalated matrix was used to construct a gene counts matrix normalized using the Trimmed Mean of M values (TMM) method (Table S17), which accounts for differences in the distribution of transcript expression that could lead to an increase in false positive rates, and decrease the power to detect truly differentially expressed genes [75]. Before doing the differential expression analyses and the estimation of tissue specificity, we evaluated the quality of our samples by doing a principal component analysis (PCA) and a Pearson correlation matrix using the gene raw count matrix, according to the Trinity QC samples and biological replicates pipeline [72]. The intention of these analyses was to look for the presence of batch effects or sample outliers, and to verify that biological replicates clustered within each tissue type and not among sampled individuals.

Differential expression analysis and identification of tissue preferentially upregulated genes.

Differentially expressed genes (DEG) and PUR genes were identified using the Trinity Differential Expression and Sample-Specific Expression pipelines [72, 76]. Briefly, we first identified DEG using the gene raw counts matrix with edgeR 3.28.0 [77, 78]. The differential expression analysis was based on pairwise comparisons of each of the 5 tissues (10 pairs), using the six samples per tissue as biological replicates.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.350058; this version posted October 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

For each pair of DEG identified we obtained their associated false discovery rate (FDR), and then we used this information combined with the normalized counts of the TMM matrix to identify the PUR genes in each of the five tissues. We obtained a normalized mean value of expression for each tissue by averaging and log 2 transforming the counts for each gene across the six replicates for each tissue on the TMM gene matrix. Each DEG with a maximum FDR of 0.05 for differential expression and with positive logFC of the log2 transformed gene counts in the TMM matrix was classified as PUR. A summary of pairwise expression differences between tissues based on the logFC of the log2 transformed gene counts in the TMM matrix is provided in Data S13.

Tissue-specific expression

As an alternative approach to quantitatively assess the tissue-specific expression of the genes we calculated the τ index based on the TMM gene counts matrix. The τ index ranges between 0 for widely expressed genes, and 1 for exclusively tissue-specific genes [79] . As the τ index considers tissue specificity independently of the level of expression, we set as “not expressed” genes with expression values < 1 from our TMM matrix in order to exclude genes with low support for true expression and low signal to noise ratio. To do this, we first log2 transformed the matrix in order to normalize the distribution of the expression values. We set all negative values in the matrix to zero, as this represented values < 1 before log2 transformation. We excluded contigs that had no expression values or that had expression in just one out of the 30 samples. Then, the τ index was computed separately for each gene across all tissues and replicates according to the following equation [79, 80] :

$$\tau = \frac{\sum_{i=1}^{N} =(1-X_i)}{N-1}, \; X_i = \frac{x_i}{max(x_i)} \text{ where max}(x_i)\; 1 \leq i \leq N$$

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.350058; this version posted October 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Where N represents the number of tissues, x_i is the mean expression in tissue *i* and X_i is the expression level in tissue *i* normalized by the maximum mean expression among all tissues [81].

Singular enrichment analysis

To further characterize the gene expression in the five tissues, we identified the biological pathways for both tissue-specific and tissue preferentially upregulated gene sets with independent singular enrichment analysis (SEA) [82, 83]. First, we retrieved the UniProt IDs corresponding to our putative genes from the blastx field from our reference annotation file [33]. Then we uploaded the list of UniProt IDs to the uniprot retrieve/ID mapping tool [84] and restricted the result to GO terms only. We repeated this procedure with the genes used as a background list for the SEA: all the contigs in the gene raw counts matrix for the PUR genes (Data S14), and all the contigs in the filtered TMM matrix in the case of the tissue-specific genes (Data S15).

Of the 715,398 putative genes in the raw counts matrix used for the differential expression analysis, 17,227 have a unique UniProt ID and represent 108,947 GO terms. The background list for the tissue-specific genes data set consisted of 177,075 contigs of which 14,079 have a unique annotation and represent 90,198 GO terms. For both data sets only uniquely annotated genes and their corresponding GO terms (Data S2-S11) were used for running the singular enrichment analyses to avoid inflating the number of GO terms falsely, and creating a bias in the analysis.

We used the GO terms along the UniProt IDs as input for the SEA using the agriGO platform [83, 85, 86]. We used the custom background list option, applied a hypergeometric test as statistical test method with a minimum of 5 mapping entries per term, and Hochberg FDR as multi-test adjustment method with a significance level of 0.05.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.350058; this version posted October 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

Clean reads corresponding to each of the five tissues used in the transcript quantification can be found in BioProject PRJNA531617. The Trinotate file used to obtain the gene identifiers for the tissue PUR and tissue-specific genes identified in this work is at https://doi.org/10.6084/m9.figshare.12865997.v1.

Supplementary information files:

Data S1 list of putative contaminant contigs removed from quantification files.txt

Data S12.Trinity_guided_with_contaminants.fasta

Data S13.DE_pairwise_sumary.txt

Table S2. raw counts matrix.txt

Table S3. Filtered TMM_Tau_final_matrix.txt

Table S4. CombinedTissueExpressionInfo.txt

Table S16.TPM.not_cross_norm.matrix.txt

Table S17.TMM.EXPR.matrix.txt

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.350058; this version posted October 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

from this work are available in the Figshare repository, https://doi.org/10.6084/m9.figshare.c.5181788.

Competing interest

The authors declare that they have no competing interests.

Funding

The Academy of Finland grants 287431, 293819, and 319313 to TP. Part of the work was carried out with the support of Biocenter Oulu to SC.

Authors' contributions

SC: Main responsibility of the statistical and bioinformatic analyses, main responsibility of writing the manuscript.

JV: Writing parts of the manuscript, editing and commenting the manuscript.

DP: Visualization of the data, editing and commenting the manuscript.

TP: Concept, laboratory work, acquisition of funding, writing and editing the manuscript.

Acknowledgments

The authors wish to acknowledge CSC – IT Center for Science, Finland, for computational resources.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.350058; this version posted October 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Bibliography

1. Farjon A. A natural history of conifers. 2008.

2. San-Miguel-Ayanz J, De Rigo D, Caudullo G, Durrant TH, Mauri A. European atlas of forest tree species. Publications Office of the European Union; 2016.

3. الكتاب السنوي للمنتجات الحرجية | 粮农组织林产品年鉴 | FAO Yearbook of Forest Products | Annuaire FAO des produits forestiers | Ежегодник лесной продукции ФАО | Anuario FAO de productos forestales 2018. 2020. doi:10.4060/cb0513m.

4. Bonan GB, Chapin FS, Thompson SL. Boreal forest and tundra ecosystems as components of the climate system. Clim Change. 1995;29:145–67.

5. Boonstra R, Andreassen HP, Boutin S, Hušek J, Ims RA, Krebs CJ, et al. Why Do the Boreal Forest Ecosystems of Northwestern Europe Differ from Those of Western North America? Bioscience. 2016;66:722–34.

6. DeAngelis DL. Boreal Forest. Encyclopedia of Ecology. 2008;:493–5. doi:10.1016/b978-008045405-4.00319-0.

7. Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin Y-C, Scofield DG, et al. The Norway spruce genome sequence and conifer genome evolution. Nature. 2013;497:579–84.

8. Stevens KA, Wegrzyn JL, Zimin A, Puiu D, Crepeau M, Cardeno C, et al. Sequence of the Sugar Pine Megagenome. Genetics. 2016;204:1613–26.

9. Mosca E, Cruz F, Gómez-Garrido J, Bianco L, Rellstab C, Brodbeck S, et al. A Reference Genome Sequence for the European Silver Fir (Abies alba Mill.): A Community-Generated Genomic Resource. G3 . 2019;9:2039–49.

10. Birol I, Raymond A, Jackman SD, Pleasance S, Coope R, Taylor GA, et al. Assembling the 20 Gb white spruce (Picea glauca) genome from whole-genome shotgun sequencing data. Bioinformatics. 2013;29:1492–7.

11. Zimin A, Stevens KA, Crepeau MW, Holtz-Morris A, Koriabine M, Marçais G, et al. Sequencing and assembly of the 22-gb loblolly pine genome. Genetics. 2014;196:875–90.

12. Cañas RA, Pascual MB, Fernando N, Ávila C, Cánovas FM. Resources for conifer functional genomics at the omics era. In: Advances in Botanical Research. Elsevier; 2019. p. 39–76.

13. Wegrzyn JL, Liechty JD, Stevens KA, Wu L-S, Loopstra CA, Vasquez-Gross HA, et al. Unique features of the loblolly pine (Pinus taeda L.) megagenome revealed through sequence annotation. Genetics. 2014;196:891–909.

14. Zonneveld BJM. Conifer genome sizes of 172 species, covering 64 of 67 genera, range from 8 to 72 picogram. Nord J Bot. 2012;30:490–502.

15. De La Torre AR, Piot A, Liu B, Wilhite B, Weiss M, Porth I. Functional and morphological evolution in gymnosperms: A portrait of implicated gene families. Evol Appl. 2020;13:210–27.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.350058; this version posted October 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

16. Neale DB, Wegrzyn JL, Stevens KA, Zimin AV, Puiu D, Crepeau MW, et al. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biol. 2014;15:R59.

17. Baker EAG, Wegrzyn JL, Sezen UU, Falk T, Maloney PE, Vogler DR, et al. Comparative Transcriptomics Among Four White Pine Species. G3 . 2018;8:1461–74.

18. Wegrzyn JL, Falk T, Grau E, Buehler S, Ramnath R, Herndon N. Cyberinfrastructure and resources to enable an integrative approach to studying forest trees. Evol Appl. 2020;13:228–41.

19. Raherison E, Rigault P, Caron S, Poulin P-L, Boyle B, Verta J-P, et al. Transcriptome profiling in conifers and the PiceaGenExpress database show patterns of diversification within gene families and interspecific conservation in vascular gene expression. BMC Genomics. 2012;13:434.

20. Merino I, Abrahamsson M, Sterck L, Craven-Bartle B, Canovas F, von Arnold S. Transcript profiling for early stages during embryo development in Scots pine. BMC Plant Biol. 2016;16:255.

21. Little SA, Boyes IG, Donaleshen K, von Aderkas P, Ehlting J. A transcriptomic resource for Douglas-fir seed development and analysis of transcription during late megagametophyte development. Plant Reprod. 2016;29:273–86.

22. Canas RA, Li Z, Pascual MB. The gene expression landscape of pine seedling tissues. The Plant. 2017. https://onlinelibrary.wiley.com/doi/abs/10.1111/tpj.13617.

23. Tyrmi JS, Vuosku J, Acosta JJ, Li Z, Sterck L, Cervera MT, et al. Genomics of Clinal Local Adaptation in Pinus sylvestris Under Continuous Environmental and Spatial Genetic Setting. G3: Genes|Genomes|Genetics. 2020;10:2683–96. doi:10.1534/g3.120.401285.

24. Rellstab C, Dauphin B, Zoller S, Brodbeck S, Gugerli F. Using transcriptome sequencing and pooled exome capture to study local adaptation in the giga-genome of Pinus cembra. Mol Ecol Resour. 2019;19:536–51.

25. Wright SI, Yau CBK, Looseley M, Meyers BC. Effects of gene expression on molecular evolution in Arabidopsis thaliana and Arabidopsis lyrata. Mol Biol Evol. 2004;21:1719–26.

26. Slotte T, Bataillon T, Hansen TT, St Onge K, Wright SI, Schierup MH. Genomic determinants of protein evolution and polymorphism in Arabidopsis. Genome Biol Evol. 2011;3:1210–9.

27. Otto SP, Scott MF, Immler S. Evolution of haploid selection in predominantly diploid organisms. Proc Natl Acad Sci U S A. 2015;112:15952–7.

28. Pyhäjärvi T, Kujala ST, Savolainen O. 275 years of forestry meets genomics in. Evol Appl. 2020;13:11–30.

29. Williams CG. Selfed embryo death in Pinus taeda: a phenotypic profile. New Phytol. 2008;178:210–22.

30. Verta JP, Landry CR, MacKay J. Dissection of expression-quantitative trait locus and allele specificity using a haploid/diploid plant system–insights into compensatory evolution of transcriptional …. New Phytol. 2016. https://nph.onlinelibrary.wiley.com/doi/abs/10.1111/nph.13888.

31. Höllbacher B, Schmitt AO, Hofer H, Ferreira F, Lackner P. Identification of Proteases and

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.350058; this version posted October 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Protease Inhibitors in Allergenic and Non-Allergenic Pollen. Int J Mol Sci. 2017;18. doi:10.3390/ijms18061199.

32. Wachowiak W, Trivedi U, Perry A, Cavers S. Comparative transcriptomics of a complex of four European pine species. BMC Genomics. 2015;16:234.

33. Ojeda DI, Mattila TM, Ruttink T, Kujala ST, Kärkkäinen K, Verta J-P, et al. Utilization of Tissue Ploidy Level Variation in de Novo Transcriptome Assembly of Pinus sylvestris. G3 . 2019;9:3409–21.

34. Perry A, Wachowiak W, Downing A, Talbot R, Cavers S. Development of a SNP array for population genomic studies in four European pine species. Mol Ecol Resour. 2020. doi:10.1111/1755-0998.13223.

35. Li Z, De La Torre AR, Sterck L, Cánovas FM, Avila C, Merino I, et al. Single-Copy Genes as Molecular Markers for Phylogenomic Studies in Seed Plants. Genome Biol Evol. 2017;9:1130–47.

36. Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, et al. The evolution of gene expression levels in mammalian organs. Nature. 2011;478:343–8.

37. Yang R, Wang X. Organ evolution in angiosperms driven by correlated divergences of gene sequences and expression patterns. Plant Cell. 2013;25:71–82.

38. Gonzalez-Ibeas D, Martinez-Garcia PJ, Famula RA, Delfino-Mix A, Stevens KA, Loopstra CA, et al. Assessing the Gene Content of the Megagenome: Sugar Pine (Pinus lambertiana). G3 . 2016;6:3787–802.

39. Ojeda D. Pinus sylvestris transcriptome annotation-Trinotate. Utilization of Tissue Ploidy Level Variation in de Novo Transcriptome Assembly of Pinus sylvestris. 2020. doi:10.6084/m9.figshare.12865997.v1.

40. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. 2000;28:45–8.

41. Lemos B, Meiklejohn CD, Cáceres M, Hartl DL. Rates of divergence in gene expression profiles of primates, mice, and flies: stabilizing selection and variability among functional categories. Evolution. 2005;59:126–37.

42. Ralph S, Park J-Y, Bohlmann J, Mansfield SD. Dirigent proteins in conifer defense: gene discovery, phylogeny, and differential wound- and insect-induced expression of a family of DIR and DIR-like genes in spruce (Picea spp.). Plant Mol Biol. 2006;60:21–40.

43. Mir G, Domènech J, Huguet G, Guo W-J, Goldsbrough P, Atrian S, et al. A plant type 2 metallothionein (MT) from cork tissue responds to oxidative stress. J Exp Bot. 2004;55:2483–93.

44. Lorenz WW, Dean JFD. SAGE profiling and demonstration of differential gene expression along the axial developmental gradient of lignifying xylem in loblolly pine (Pinus taeda). Tree Physiol. 2002;22:301–10.

45. Pongrac P, Baltrenaite E, Vavpetič P, Kelemen M, Kladnik A, Budič B, et al. Tissue-specific element profiles in Scots pine (Pinus sylvestris L.) needles. Trees. 2019;33:91–101.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.350058; this version posted October 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

46. Singh H. Embryology of Gymnosperms. 1978.

47. Vuosku J, Sutela S, Kestilä J, Jokela A, Sarjala T, Häggman H. Expression of catalase and retinoblastoma-related protein genes associates with cell death processes in Scots pine zygotic embryogenesis. BMC Plant Biol. 2015;15:88.

48. Simola LK. The Ultrastructure of Dry and Germinating Seeds of Pinus Sylvestris L. Societas pro Fauna et Flora Fennica; 1974.

49. Achotegui-Castells A, Llusià J, Hódar JA, Peñuelas J. Needle terpene concentrations and emissions of two coexisting subspecies of Scots pine attacked by the pine processionary moth (Thaumetopoea pityocampa). Acta Physiologiae Plantarum. 2013;35:3047–58. doi:10.1007/s11738-013-1337-3.

50. Geldner N. Making phloem--a near-death experience. Science. 2014;345:875–6. doi:10.1126/science.1258711.

51. King JE, Gifford DJ. Amino Acid Utilization in Seeds of Loblolly Pine during Germination and Early Seedling Growth (I. Arginine and Arginase Activity). Plant Physiol. 1997;113:1125–35.

52. Costa LM, Gutièrrez-Marcos JF, Dickinson HG. More than a yolk: the short life and complex times of the plant endosperm. Trends Plant Sci. 2004;9:507–14.

53. Williams JH, Friedman WE. Identification of diploid endosperm in an early angiosperm lineage. Nature. 2002;415:522–6. doi:10.1038/415522a.

54. Baroux C, Spillane C, Grossniklaus U. Evolutionary origins of the endosperm in flowering plants. Genome Biol. 2002;3:reviews1026.

55. Ching TM. Glyoxysomes in megagamethophyte of germinating ponderosa pine seeds. Plant Physiol. 1970;46:475–82.

56. Graham IA. Seed Storage Oil Mobilization. Annual Review of Plant Biology. 2008;59:115–42. doi:10.1146/annurev.arplant.59.032607.092938.

57. Xu J, Jun XU, Wang X-Y, Wang-zhen GUO. The cytochrome P450 superfamily: Key players in plant development and defense. Journal of Integrative Agriculture. 2015;14:1673–86. doi:10.1016/s2095-3119(14)60980-1.

58. Jeevan Kumar SP, Rajendra Prasad S, Banerjee R, Thammineni C. Seed birth to death: dual functions of reactive oxygen species in seed physiology. Ann Bot. 2015;116:663–8.

59. Seo M, Nambara E, Choi G, Yamaguchi S. Interaction of light and hormone signals in germinating seeds. Plant Mol Biol. 2009;69:463–72.

60. Guangwu Z, Xuwen J. Roles of Gibberellin and Auxin in Promoting Seed Germination and Seedling Vigor in Pinus massoniana. Forest Science. 2014;60:367–73. doi:10.5849/forsci.12-143.

61. Miransari M, Smith DL. Plant hormones and seed germination. Environmental and Experimental Botany. 2014;99:110–21. doi:10.1016/j.envexpbot.2013.11.005.

62. Shu K, Liu X-D, Xie Q, He Z-H. Two Faces of One Seed: Hormonal Regulation of Dormancy

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.350058; this version posted October 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

and Germination. Mol Plant. 2016;9:34–45.

63. Otegui MS. Endosperm Cell Walls: Formation, Composition, and Functions. Plant Cell Monographs. :159–77. doi:10.1007/7089_2007_113.

64. Yan D, Duermeyer L, Leoveanu C, Nambara E. The functions of the endosperm during seed germination. Plant Cell Physiol. 2014;55:1521–33.

65. Szmidt AE, Muona O. Linkage relationships of allozyme loci in Pinus sylvestris. Hereditas. 1989;111:91–7. doi:10.1111/j.1601-5223.1989.tb00382.x.

66. Paasela T, Lim K-J, Pietiäinen M, Teeri TH. The O-methyltransferase PMT2 mediates methylation of pinosylvin in Scots pine. New Phytol. 2017;214:1537–50.

67. Duarte GT, Volkova PY, Geras'kin SA. The response profile to chronic radiation exposure based on the transcriptome analysis of Scots pine from Chernobyl affected zone. Environ Pollut. 2019;250:618–26.

68. Guseva T, Biriukov V, Sadovsky M. Role of Homeobox Genes in the Development of Pinus Sylvestris. Bioinformatics and Biomedical Engineering. 2020;:429–37. doi:10.1007/978-3-030-45385-5_38.

69. Niinemets Ü. Responses of forest trees to single and multiple environmental stresses from seedlings to mature plants: Past stress history, stress interactions, tolerance and acclimation. Forest Ecology and Management. 2010;260:1623–39. doi:10.1016/j.foreco.2010.07.054.

70. EVOLTREE :: Intensive Study Sites. http://www.evoltree.eu/index.php/intensive-study-sites. Accessed 15 Oct 2020.

71. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013;8:1494–512.

72. trinityrnaseq. trinityrnaseq. https://github.com/trinityrnaseq/trinityrnaseq/wiki/. Accessed 15 Oct 2020.

73. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14:417–9.

74. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Res. 2015;4:1521.

75. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11:R25.

76. Bryant DM, Johnson K, DiTommaso T, Tickle T, Couger MB, Payzin-Dogru D, et al. A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors. Cell Rep. 2017;18:762–76.

77. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.350058; this version posted October 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

78. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 2012;40:4288–97.

79. Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, et al. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics. 2005;21:650–9.

80. Kryuchkova-Mostacci N, Robinson-Rechavi M. A benchmark of gene expression tissue-specificity metrics. Brief Bioinform. 2017;18:205–14.

81. severinEvo. tau.R. Github. https://github.com/severinEvo/gene_expression/blob/master/tau.R. Accessed 15 Oct 2020.

82. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37:1–13.

83. Du Z, Zhou X, Ling Y, Zhang Z, Su Z. agriGO: a GO analysis toolkit for the agricultural community. Nucleic Acids Res. 2010;38 Web Server issue:W64–70.

84. Retrieve/ID mapping. https://www.uniprot.org/uploadlists/. Accessed 15 Oct 2020.

85. Tian T, Liu Y, Yan H, You Q, Yi X, Du Z, et al. agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. Nucleic Acids Res. 2017;45:W122–9.

86. agriGO: GO Analysis Toolkit and Database for Agricultural Community. http://systemsbiology.cau.edu.cn/agriGOv2/index.php. Accessed 15 Oct 2020.

Supplementary information

Additional file 1: Table S1. Table including information about the number of reads mapped to each one of the tissues.

Additional file 2: Data S1. List of contigs identified as contaminants.

Additional file 3: Table S2. Table including the raw counts per genes across the five tissues.

Additional file 4: Figure S1. Heat map showing the correlation between the expression patterns of the five tissues. A=embryo, KS=vegetative bud, M=megagametophyte, N=needle, Ni=phloem.

Additional file 5: Table S3. Table including the TMM normalized and filtered counts used for the identification of genes with tissue-specific expression patterns. A=embryo, KS=vegetative bud, M=megagametophyte, N=needle, Ni=phloem.

Additional file 6: Table S4. Table including the level of expression, indication of tissue PUR or tissue-specific expression, and annotation for 715,398 genes across the five tissues. Column one

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.350058; this version posted October 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

indicates gene ID, columns two to six contain the TMM normalized gene counts per tissue, column seven indicates the gene tau score, columns eight to 12 indicate if in which tissue the gene is preferentially upregulated, column 13 indicates the UniProt ID, column 14 indicates the protein name, column 15 indicates the gene name, column 16 indicates the organism from where the annotation was obtained.

Additional file 7: Table S5. Table containing the results of the SEA for tissue PUR genes expressed in vegetative bud. Column two indicates the ontological process where P=biological processes, F=molecular function, C=cellular component.

Additional file 8: Table S6. Table containing the results of the SEA for tissue PUR genes expressed in embryo. Column two indicates the ontological process where P=biological processes, F=molecular function, C=cellular component.

Additional file 9: Table S7. Table containing the results of the SEA for tissue PUR genes expressed in megagametophyte. Column two indicates the ontological process where P=biological processes, F=molecular function, C=cellular component.

Additional file 10: Table S8. Table containing the results of the SEA for tissue PUR genes expressed in needle. Column two indicates the ontological process where P=biological processes, F=molecular function, C=cellular component.

Additional file 11: Table S9. Table containing the results of the SEA for tissue PUR genes expressed in phloem. Column two indicates the ontological process where P=biological processes, F=molecular function, C=cellular component.

Additional file 12: Table S10. Table containing the results of the SEA for genes with tissue-specific expression pattern in vegetative bud. Column two indicates the ontological process where P=biological processes, F=molecular function, C=cellular component.

Additional file 13: Table S11. Table containing the results of the SEA for genes with tissue-specific expression pattern in embryo. Column two indicates the ontological process where P=biological processes, F=molecular function, C=cellular component.

Additional file 14: Table S12. Table containing the results of the SEA for genes with tissue-specific expression pattern in megagametophyte. Column two indicates the ontological process where P=biological processes, F=molecular function, C=cellular component.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.350058; this version posted October 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Additional file 15: Table S13. Table containing the results of the SEA for genes with tissue-specific expression pattern in needle. Column two indicates the ontological process where P=biological processes, F=molecular function, C=cellular component.

Additional file 16: Table S14. Table containing the results of the SEA for genes with tissue-specific expression pattern in phloem. Column two indicates the ontological process where P=biological processes, F=molecular function, C=cellular component.

Additional file 17: Data S2. Genes identified in vegetative bud with PUR expression pattern and their associated GO terms.

Additional file 18:Data S3. Genes identified in embryo with PUR expression pattern and their associated GO terms.

Additional file 19: Data S4. Genes identified in megagametophyte with PUR expression pattern and their associated GO terms.

Additional file 20: Data S5. Genes identified in needle with PUR expression pattern and their associated GO terms.

Additional file 21: Data S6. Genes identified in phloem with PUR expression pattern and their associated GO terms.

Additional file 22: Data S7. Genes identified in vegetative bud with tissue-specific expression pattern and their associated GO terms.

Additional file 23: Data S8. Genes identified in embryo with tissue-specific expression pattern and their associated GO terms.

Additional file 24: Data S9. Genes identified in megagametophyte with tissue-specific expression pattern and their associated GO terms.

Additional file 25: Data S10. Genes identified in needle with tissue-specific expression pattern and their associated GO terms.

Additional file 26: Data S11. Genes identified in phloem with tissue-specific expression pattern and their associated GO terms.

Additional file 27: Table S15. Table with information about the geographical location of the trees used in this study.

Additional file 28: Data S12. $\text{Trinity}_{\text{guided}}$ transcriptome used as reference (contaminants included).

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.350058; this version posted October 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Additional file 29: Table S16. Table containing the TPM normalized gene counts. A=embryo, KS=vegetative bud, M=megagametophyte, N=needle, Ni=phloem

Additional file 30: Table S17. Table containing the TMM normalized unfiltered gene counts. A=embryo, KS=vegetative bud, M=megagametophyte, N=needle, Ni=phloem

Additional file 31: Data S13. Pairwise expression differences between tissues.

Additional file 32: Data S14. List of genes used as a background list for the SEA of PUR genes.

Additional file 33: Data S15. List of genes used as a background list for the SEA of tissue-specific genes.