Journal of Plant Research (2021) 134:1253–1264
https://doi.org/10.1007/s10265-021-01343-y

REGULAR PAPER – GENETICS/DEVELOPMENTAL BIOLOGY

Transcriptome analysis of North American sweet birch (*Betula lenta*) revealed a higher expression of genes involved in the biosynthesis of secondary metabolites than European silver birch (*B. pendula*)

Kiran Singewar[1,2] · Birgit Kersten[2] · Christian R. Moschner[1] · Eberhard Hartung[1] · Matthias Fladung[2]

Received: 14 July 2021 / Accepted: 31 August 2021 / Published online: 9 September 2021

Abstract

The North American *Betula lenta* L. (sweet birch) has been used for medicinal reasons for centuries by native Americans. Although sophisticated technologies have rapidly been developed, a large information gap has been observed regarding genetic regulators of medicinally important compounds in sweet birch. Very little is known on the different genes involved in secondary metabolic biosynthesis in sweet birch. To gain a deeper insight into genetic factors, we performed a transcriptome analysis of each three biological samples from different independent trees of sweet and European silver birch (*B. pendula* Roth). This allowed us to precisely quantify the transcripts of about 24,000 expressed genes including 29 prominent candidate genes putatively involved in the biosynthesis of secondary metabolites like terpenoids, and aromatic benzoic acids. A total number of 597 genes were differentially expressed between *B. lenta* and *B. pendula*, while 264 and 210 genes showed upregulation in the bark and leaf of *B. lenta*, respectively. Moreover, we identified 39 transcriptional regulatory elements, involved in secondary metabolite biosynthesis, upregulated in *B. lenta*. Our study demonstrated the potential of RNA sequencing to identify candidate genes interacting in secondary metabolite biosynthesis in sweet birch. The candidate genes identified in this study could be subjected to genetic engineering to functionally characterize them in sweet birch. This knowledge can be beneficial to the increase of therapeutically important compounds.

Keywords Secondary metabolites · Silver birch · Sweet birch · Transcription regulators · Transcriptomics

Introduction

The genus *Betula* (birch)—one of the dominating woody plant species of the Northern Hemisphere—incorporates diverse species with a wide range of morphological, genetic, and physiological variations. Despite their conventional economic benefits, many species are of medicinal and pharmacological importance (Ebeling et al. 2014; Rastogi et al. 2015; Yin et al. 2012, 2017). Above all, sweet birch (*Betula lenta* L.), also known as black and cherry birch has elite importance in ancient therapeutics (COSEWIC 2006; Sharik and Burton 1971; Stephens and Waggoner 1970). A forest study suggests sweet birch is ample in Massachusetts, Connecticut, New York, and Pennsylvania (Fernald et al. 1958; Lamson 1990). Sweet birch was introduced into the landscape in 1759 (Leak 1965; Lorimer 1980). The medicinal use of *B. lenta* by the native Americans is well documented (Gilmore 1933). Birch sap can be combined with corn and fermented to make beer (Suryawanshi 2000). Previously it was the only source of wintergreen oil extraction since it has the aroma of wintergreen emanatingfrom crushed bark and leaves (Ashburner and McAllister 2013; Singewar et al. 2020a, b). Extensive harvesting of sweet birch caused to become it endangered until the 1950–1970s (Leak 1965).

Sweet birch is a diploid and deciduous woody plant species with 28 numbers of chromosomes (2n) (Ashburner and McAllister 2013; Wang et al. 2016). It is closely related to *B. alleghaniensis*, the yellow birch (Sharik and Burton 1971). Crossing between *B. lenta* and *B. alleghaniensis*, and successful production of hybrids has been recorded in the past with low vigor and seed germination rates in F1 offspring

✉ Matthias Fladung
matthias.fladung@thuenen.de

[1] Institute of Agricultural Process Engineering, Christian-Albrechts University of Kiel, Max-Eyth- Str. 6, 24118 Kiel, Germany

[2] Thuenen-Institute of Forest Genetics, Sieker Landstraße 2, 22927 Grosshansdorf, Germany

(Sharik and Burton 1971). *B. jackii* is a natural hybrid of *B. lenta* and *B. pumila* which occurred at the Arnold Arboretum (Jack 1895).

Among the many other taxonomical distributions of the genus *Betula* (Furlow 1990; Winkler 1904), De Jong (1993) classified *B. lenta* into the subgenus *Betulenta*. Further, in the most recent resolved classification, *B. lenta* is subdivided in the section *Lentae* of the subgenus *Aspera.* (Ashburner and McAllister 2013; Wang et al. 2016). Various phylogenetic analyses demonstrated that the subgenus *Betulenta* to be among the oldest (Bina et al. 2016; De Jong 1993). Also, many evolutionary and population studies have been conducted in the genus *Betula* (Bina et al. 2016; Li et al. 2005; Singewar et al. 2020a; Wang et al. 2016). The previous (Bina et al. 2016) and most recent network analysis (Singewar et al. 2020a) suggest that the diploid *B. lenta* is one of the ancestors of the genus *Betula*. Similar network analysis also supports the study where *B. lenta* including *B. lenta* var. *uber* forms the oldest clade (Bina et al. 2016).

Many species of the genus *Betula* are polyploid, the chromosome count differs from 2n = 2x = 28 to 2n = 12x = 168 and ranges from diploid to dodecaphonic (Ashburner and McAllister 2013). The de novo genome of European silver birch (*Betula pendula* Roth) has been sequenced which is a diploid organism with 28 chromosomes (2n) having a haploid genome size of about 440 megabase pairs (Salojärvi et al. 2017). According to our knowledge, *B. pendula* and *B. platyphylla* (Chen et al. 2021) are the only reference genomes available publicly, and the only genome of *B. pendula* available on the genome browser (https://genomevolution.org/coge/GenomeInfo.pl?gid=35080) in the genus *Betula*. According to a PubMed database of NCBI (https://pubmed.ncbi.nlm.nih.gov/) survey, a large gap has been observed in the genetics and genomics of sweet birch (Singewar 2020; Zoladeski and Hayes 2013).

Knowingly or unknowingly, humans have been ignoring the medicinal importance of the sweet birch. The genetic factors involved in the secondary metabolite biosynthesis that makes *B. lenta* a pharmacologically important forest tree, remain unknown. Considering the species conservation and pharmacological importance of sweet birch, it represents a relevant target for genetic and genomic studies. Here, we profiled transcriptomes of *B. lenta* leaf and bark separately and performed a comparative analysis with *B. pendula*. Transcriptome analysis showed about 24,000 expressed genes including 29 prominent candidate genes putatively involved in the biosynthesis of secondary metabolites including terpenoids, aroma, and benzoic compounds. Moreover, 39 transcriptional regulatory elements involved in secondary metabolite biosynthesis were upregulated in *B. lenta*.

Materials and methods

Plant material and growth conditions

Betula lenta and *Betula pendula* were selected for RNA sequencing out of 29 species based on previous molecular genetic studies (Singewar 2020; Singewar et al. 2020a, b). RNA sequencing of leaf and bark tissues of both species was performed as a basis for comparative gene expression analysis to understand the tissue-specific gene expression. In April 2017, seeds were germinated in normal soil and a natural environment without any fertilizer in the poly-house at the Institute of Agricultural Process Engineering, Kiel University, Germany. Plantlet cultivation was carried out in a glasshouse at the Thuenen-Institute of Forest Genetics, Grosshansdorf, Germany, under identical conditions for all plantlets. In July 2019, leaf and bark tissues were harvested from the three different biological replicates of each species originated from three different mother trees, representing three different genotypes per species for RNA sequencing, which resulted in 12 samples.

RNA extraction and sequencing

Total RNA was extracted from leaves and bark of *B. lenta* and *B. pendula*, described in Table 1 using the CTAB protocol of Dumolin et al. (1995). Three different leaves and three sections of bark per biological replicate were collected. To remove DNA contaminations, the Invitrogen Ambion Turbo DNA-free Kit (Fisher Scientific GmbH, Schwerte, Germany) was used following the manufacturer's instructions. The quantity of RNA was determined with a Nanodrop 1000 spectrophotometer (Thermo Fisher

Table. 1 Birch individuals used in this study

Species name	Source of material	Distribution	2n	Subgenus	Section
Betula lenta	BG Giessen, Germany	North America	2n	*Aspera*	*Lentae*
Betula pendula	Reinke Baumschulen, Germany	Europe and East Asia	2n	*Betula*	*Betula*

The table describes the source of the material and geographical distribution of the birch species. The ploidy condition and taxonomical position were decided according to Wang et al. (2016) and Ashburner and McAllister (2013)

BG Botanical garden

Scientific, Wilmington, USA), A260/A280 readings > 2.0, and A260/A230 > 1.9). The RNA degradation and potential contamination were determined using agarose gel electrophoresis. Finally, the quality was measured with the Bioanalyzer Agilent 2100 (Agilent Technologies, Waldbronn, Germany). Samples matching the criteria (Table S1) were sent to Novogene Bioinformatics Technology Co., Ltd. (Hong Kong). For each species, three samples per biological replicate with the best quality values were accumulated. During the library preparation, the RNA is reverse transcribed to complementary DNA (cDNA). Sequencing was completed on the Illumina HiSeq 4000 platform to create 150 base pairs paired-end reads (on average 65 million read pairs per sample, Table S2).

Raw data filtering and bioinformatics workflow

The annotated reference genome of *B. pendula* (Salojarvi et al. 2017) was used as a reference for bioinformatic analysis of the RNA-seq data of all 12 samples. The Fastq files of the raw reads were trimmed and filtered using Trimmomatic v0.35 (Bolger et al. 2014) following parameters: ILLUMINACLIP: <fastaWithAdapters> :2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW: 4:20 MINLEN: 50. Specific adapter sequences are given in Table S3. We mapped and quantified the filtered reads *versus* the *B. pendula* reference genome (version 1.4c) using the TopHat2 (Kim et al. 2013) with the mismatch = 2 parameter. Only filtered reads were used to analyze the mapping status of RNA-seq data to the reference genome. Further, the median number of reads mapped to the genome inside the window was calculated and transformed to the log2value. Gene expression level was measured by transcript abundance and estimated by counting the reads that mapped to the annotated genes or exons of the reference genome. The mapping results in BAM format were visualized in the integrative genomics viewer (IGV).

Gene expression level was measured by transcript abundance. The greater the abundance, the higher is the gene expression level. The Fragments Per Kilobase Million (FPKM) method was used to compare the gene expression levels of different genes which considers the effects of sequencing depth and gene length (Trapnell et al. 2010). HTSeq software was used to analyze the gene expression levels, using the union mode. The FPKM values > 1 are set as the threshold for determining whether the gene is expressed or not. Further, Bioconductor's package DESeq (Wang et al. 2010) was used to normalize the read counts using the negative binomial distribution model and to determine differentially expressed gene screening (Robinson and Oshlack 2010). Differentially expressed genes were defined as having an adjusted p-value < 0.01 and an absolute log2 fold change > 1.5.

GO and KEGG pathway network enrichment analyses

GO enrichment analysis was performed by GO seq (Young et al. 2010), which is based on Wallenius non-central hypergeometric distribution. All genes of the *B. pendula* reference genome with assigned GO terms were downloaded as a reference set. Differentially expressed genes (DEGs) were identified in the different comparisons and DEGs were used as test sets. GO terms with at least two differentially expressed genes in the test set were considered for the enrichment analyses. Further, Bioconductor's topGO tool was used for testing GO terms while accounting for the topology. GO terms with a corrected p-value of $p < 0.05$ (hypergeometric statistical test with Bonferroni multiple comparison corrections) were selected as significantly enriched in the test set. To identify the putative biological function of genes, KEGG (Kyoto Encyclopedia of Genes and Genomes) curated database for functional orthologs (KO) was used.

Transcription factor analysis

According to a previous study (Yang et al. 2012), various transcription factors (TF) from different families participate in the secondary metabolite regulation. Here, we screened the putative TFs in the differentially expressed genes and considered only those genes for further analysis which were enriched in GO and KEGG pathways. Furthermore, an additional survey was carried out to consider all possible TFs. In that, amino acid sequences of previously functionally studied TFs were used as BLAST queries (Table S4) to perform BLASTp search (1E−5) versus all annotated proteins in the *B. pendula* genome (https://genomevolution.org/). Eventually, the resulting hits were screened in the differentially expressed as well as GO and KEGG enriched genes. Significantly upregulated genes in *B. lenta* were selected for comparative analysis.

Results

Differentially expressed genes

The overall distribution of differentially expressed genes was inferred through the volcano plots with the following thresholds, padj < 0.01 and absolute log2 fold change > 1.5. A total number of 597 genes were differentially expressed between *B lenta* and *B. pendula*. Around 123 and 474 genes showed up- and down-regulation, respectively. A total number of 877 genes were differentially expressed between the bark tissue of *B. lenta* and *B. pendula*, respectively. Around 264 and 613 genes showed up- and down-regulation, respectively. Similarly, 210 and 807 genes were up- and down-regulated, respectively in *B. lenta* leaves when compared to *B. pendula* ones, respectively.

The co-expression of genes in the different comparisons

Gene expression level is measured by transcript abundance. The greater the abundance, the higher is the gene expression level. In our RNA-seq analysis, the gene expression level is estimated by counting the reads that map to genes or exons (Table S5). The read count is not only proportional to the actual gene expression level but is also proportional to the gene length and the sequencing depth. For the gene expression levels estimated from different genes and experiments to be comparable, the FPKM is used (Table S6).

A total number of 1532 or 1115 genes were expressed in *B. pendula* only or *B. lenta* only, respectively (expressed genes with FPKM values > 1). In total, 12,817 genes were co-expressed in the analyzed tissues of both species under the experimental conditions. The number of genes expressed in the bark and leaf of *B. pendula* only or *B. lenta* only were 1041 or 1544, and 2277 or 743, respectively (Fig. 1).

GO enrichment analysis of differentially expressed genes

A GO enrichment bar chart is used to illustrate functional GO annotation of differentially expressed genes in *B. lenta* versus *B. pendula* (Fig. 2). The most frequent 30 GO terms comprising all GO terms assigned to more than two differentially expressed genes are shown and significantly enriched GO terms are marked (with an asterisk). The GO term ‘catalytic activity’ showed the highest frequency in the set of differentially expressed genes. Significant enrichment of genes potentially involved in the biological processes of “cell redox homeostasis” and “homeostatic process” with predicted molecular functions of “oxidoreductase activity” and “iron ion binding” among others was obvious in the set of differentially expressed genes in *B. lenta* versus *B. pendula*.

Further, the directed acyclic graph (DAG) was used to show the results of GO enrichment of DEGs (Fig. 3). The branches represent the containment relationships, and the range of functions gets smaller and smaller from top to bottom. Generally, the top ten of GO enrichment results is selected as the master nodes in a DAG, showing the associated GO terms together via the containment relationship, and the degree of colors represent the extent of enrichment. In this study, DAG figures of biological process and molecular function are drawn. Here, the DAG figure of molecular function represents differentially expressed genes in *B. lenta* (Fig. 3). DAG figures of biological process are provided in supplementary data (Fig. S1).

KEGG enrichment analysis

In the scatter plot, the enrichment degree of KEGG was measured through the rich factor, Q value, and gene counts enriched by this pathway. The top 20 most significantly enriched pathways are chosen in KEGG scatter plot (Fig. 4). Genes involved in the biosynthesis of secondary metabolites were shown to be most expressed in KEGG analysis. Additionally, genes involved in plant-pathogen interactions, flavonoid biosynthesis, phenylpropanoid biosynthesis, and cyanoamino acid metabolism were shown to be enriched (Fig. 4).

Identification of genes involved in secondary metabolites and aromatic compounds

Most of the differentially expressed genes were from the GO term, catalytic activity, and from the KEGG pathway enrichment analysis, biosynthesis of secondary metabolites in the total transcriptome data. All the differentially expressed genes in *B. lenta* that were assigned to enriched GO terms and/or KEGG pathways were further analyzed. A total number of 11 and 18 genes involved in secondary

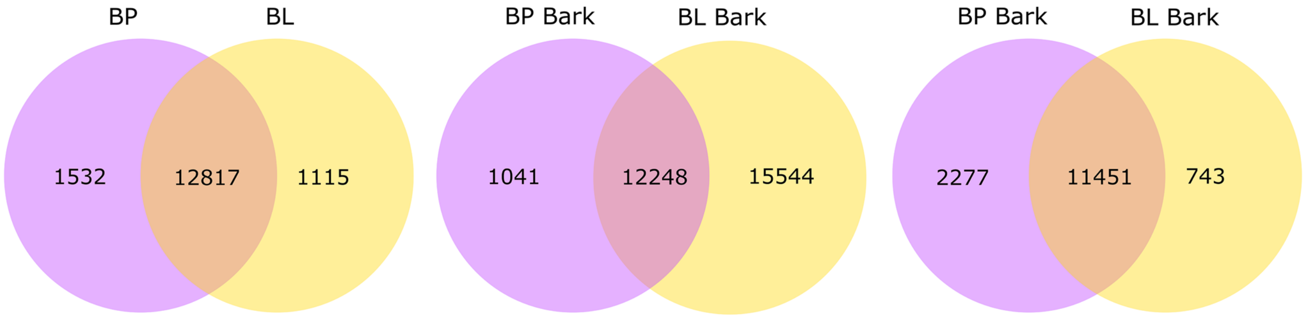

Fig. 1 Graphical representation of a total number of co-expressed genes. The sum of the numbers in each circle is the total number of genes expressed within a sample, and the overlap represents the genes expressed in common between samples

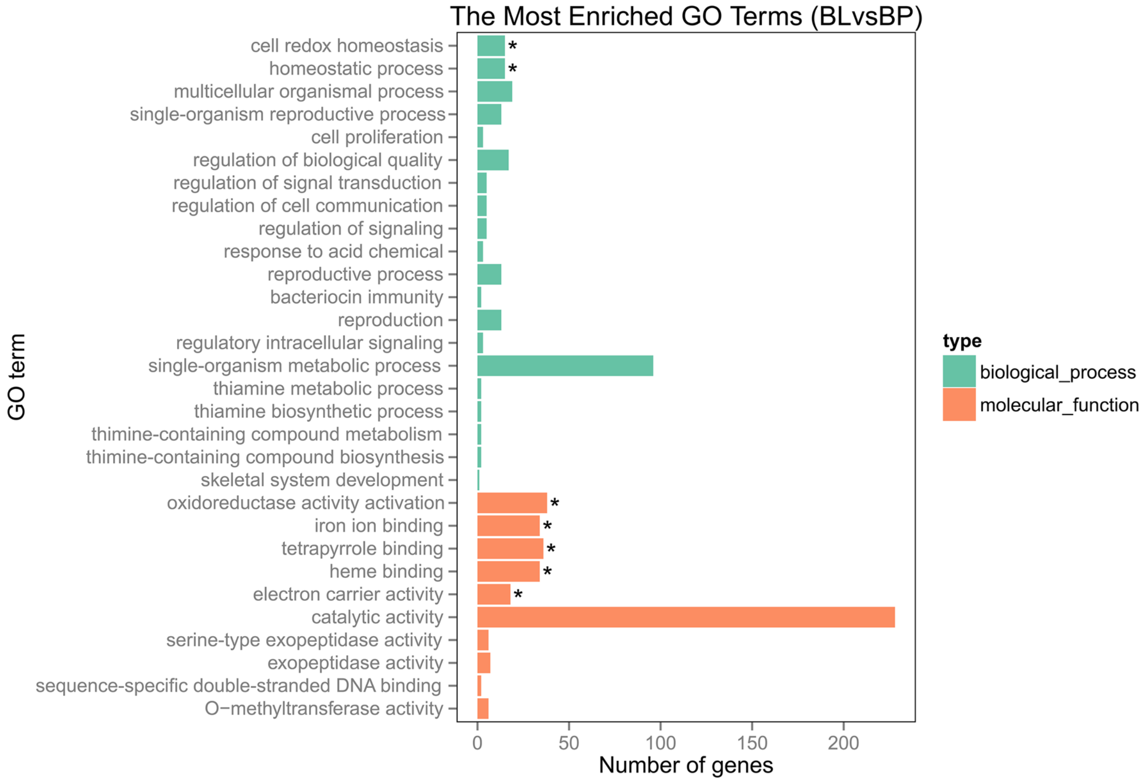

Fig. 2 GO annotation of differentially expressed genes between *B. lenta* and *B. pendula* and GO enrichment analysis. The most frequent 30 GO terms with at least 2 assigned differentially expressed genes are shown. The x-axis is the number of differentially expressed genes assigned to the respective GO term that is shown at the y-axis. In the GO enrichment analysis, all genes of the *B. pendula* reference genome were used as a reference set. Most of the differentially expressed genes were categorized in 'biological process' and 'molecular function'. Different colors are used for biological process and molecular function, in which the enriched GO terms are marked by "*"

metabolism and aromatic compounds were upregulated in *B. lenta*, respectively (Figs. 5 and 6). Almost all the genes showed higher expression in the bark tissues compared to the leaf. Gene identifiers and their designated annotations were collected (Table S7).

Identification of transcription factors (TF) involved in secondary metabolism

In plants, transcription factors (TFs) play a vital role in gene regulation that could also cause metabolic variability through interaction with the promoter region of a gene. The amino acid sequences of previously described transcription factors (Table S4) were used to identify the TF in *B. pendula* and *B. lenta*. All transcription factors identified among the *B. pendula* gene models which are in the list of upregulated DEGs in *B. lenta* were considered for further analysis. 39 putative *B. lenta* transcription factors belonging to nine TF families were identified (Fig. 7). Nine families of TF that are supposed to be involved in secondary metabolic biosynthesis including AP2-ERF, bHLH, bZIP, DOF, Zinc-finger, MYB, NAC, SPL9 and, WRKY, were detected. Every TF showed upregulated expression in leaf and bark tissue of *B. lenta* compared to *B. pendula*. Among them, *AP2-ERF* and *Zinc-finger* were the most abundant TF family (7 genes each), followed by *NAC* (6 genes), *WRKY* (5 genes), *MYB* and *DOF* (4 genes each), *SPL9* (3 genes), *bHLH* (2 genes) and *bZIP* (1 gene).

The *APETALA2/ethylene response factor* (*AP2/ERF*) family are the TFs that could be characterized by their DNA-binding AP2 domain. The domain consists of around 60 conserved amino acid residues (Mizoi et al. 2012). Further, zinc-finger TFs functions in gene regulation in stably transformed plants. The target of the *zinc-finger* TF is the *Arabidopsis thaliana APETALA3* (*AP3*) gene, involved in floral organ identification (Xuen et al. 2002). The *NAC* TF families were reported to be a regulator of camalexin, which is a plant secondary metabolite. A study has shown that *WRKY* TF is also taking part in the secondary metabolite volatile terpene biosynthesis (Skibbe et al. 2008). The expression of DNA-binding-with-one-finger (DOF) TF is studied to be in response to pathogens and the phytohormone jasmonic acid as a part of regulatory networks (Skirycz et al. 2006). The vital

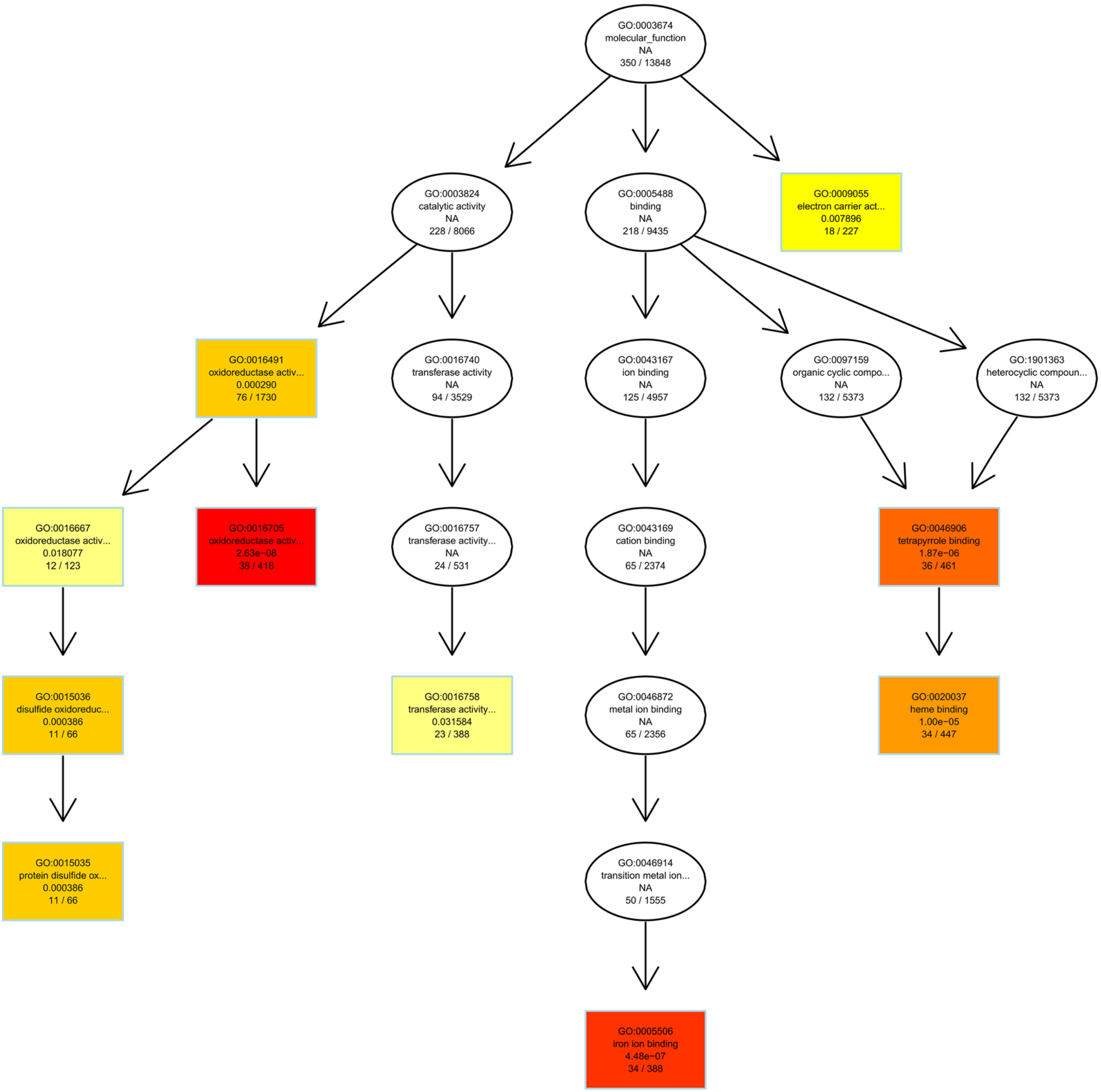

Fig. 3 The GO terms of the GO main class "molecular function" enriched in the set of differentially expressed genes in *B. lenta* versus *B. pendula.* Each node represents a GO term and significantly enriched GO terms are boxed. The darker the color is, the lower is the p-value and the higher is the enrichment level of the GO term in the enrichment test. The name and corrected p-value of each term are present on the node

function of *MYB* TF is to regulate secondary metabolites including terpenoids (Galis et al. 2006; Yin et al. 2017). The *SPL9* (*SQUAMOSA Promoter Binding Protein-Like*) participates in a broad range of developmental processes like anthocyanin accumulation which is important for pigments (Jin-Ying et al. 2011). Pathogen defense and stress signaling is regulated by TF *bZIP* (Jakoby et al. 2002). Functional genomics investigation of these TFs can allow us to understand the regulatory networks involved in the secondary metabolite biosynthesis in *B. lenta*.

Discussion

Betula lenta is extensively known as a traditional medicinal plant (Ebeling et al. 2014; Rastogi et al. 2015; Yin et al. 2012, 2017). According to authors knowledge and online research, *B. lenta* has largely been ignored by the modern generation in the past century. According to initial olfactory and monographic evidence (Ashburner and McAllister 2013; Singewar et al. 2020a, b), it contains

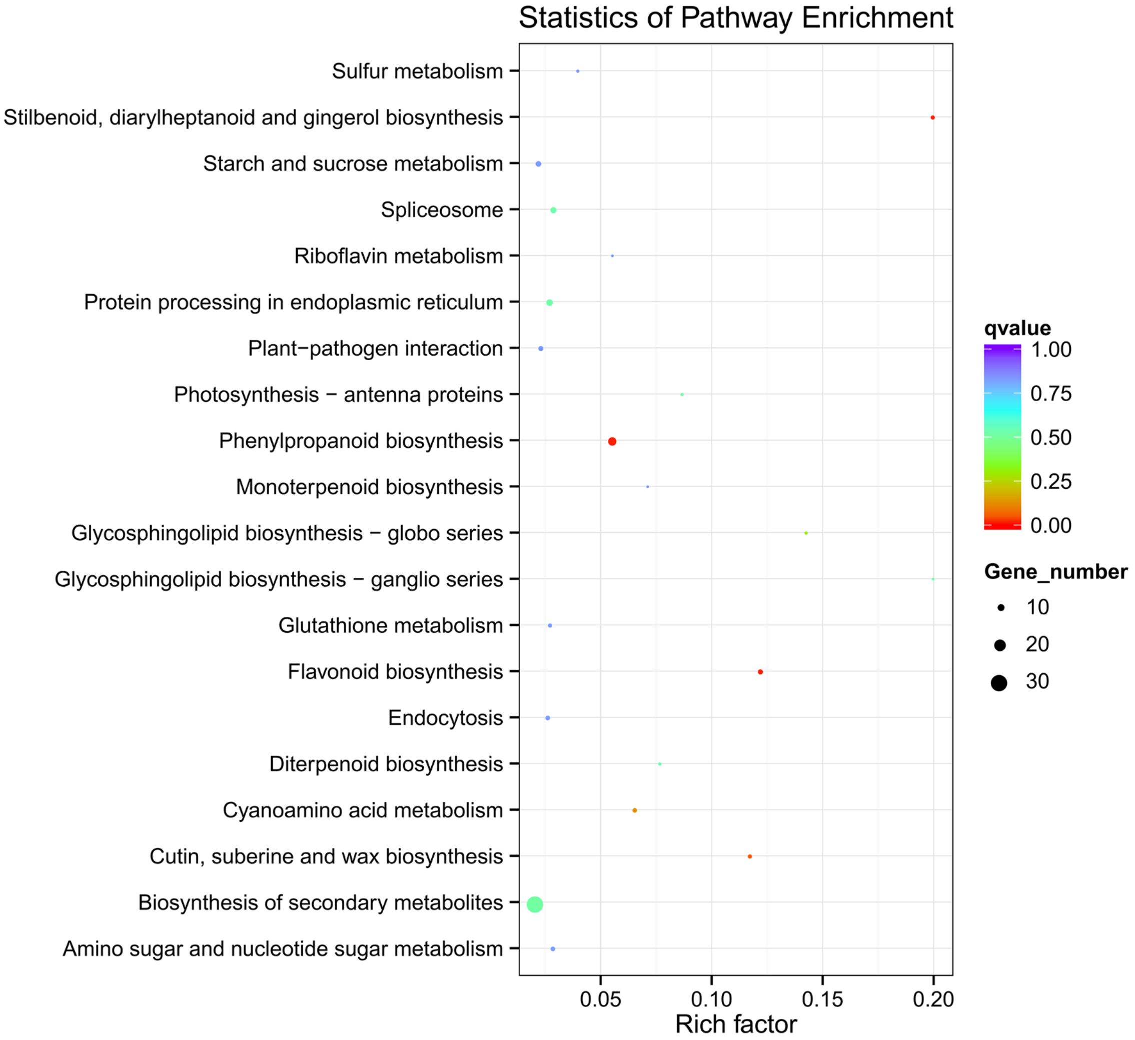

Fig. 4 The y-axis shows the name of the pathway, and the x-axis shows the rich factor. Dot size represents the number of different genes, and the color indicates the q-value The color of the circle represents the q-value; a smaller q-value indicates higher reliability for the significance of differential expression of genes in this pathway

abundant secondary metabolites including aromatic compounds and response regulators. Fortunately, the reference genomes of *B. pendula* and *B. platyphylla* are available (Chen et al. 2021; Salojärvi et al. 2017). Furthermore, *B. pendula* is the only species in the genus *Betula* whose genome is available on the web genome browser (https://genomevolution.org/), providing an opportunity to explore other therapeutically important and largely ignored species of the genus. Although the cost of sequencing technology has reduced drastically which is a good sign for more scientific research, a recent survey on available genome sequences of the species within the genus *Betula* in the NCBI database clearly shows the scarcity of knowledge within the subgenus *Aspera* (Singewar 2020; Zoladeski and Hayes 2013). We performed the transcriptome analysis of *B. lenta* in comparison to *B. pendula* to reduce the information gap.

Here, we performed high-throughput transcriptome sequencing for bark and leaf tissues of *B. lenta*. We utilized three different biological replicates of each species, originated from three different seeds each, representing three different genotypes per species. The comparative analysis of these divergent *Betula* genotypes resulted in the novel insights of the least studied woody plant species, *B. lenta* in comparison to *B. pendula*. A total number of 24,554 genes were identified to be expressed in the experimental dataset considering 28,153 genes annotated in the reference genome. When considering the set of upregulated DEGs in the *B. lenta* and *B. pendula* comparative analysis, several significantly enriched GO terms assigned to the GO main

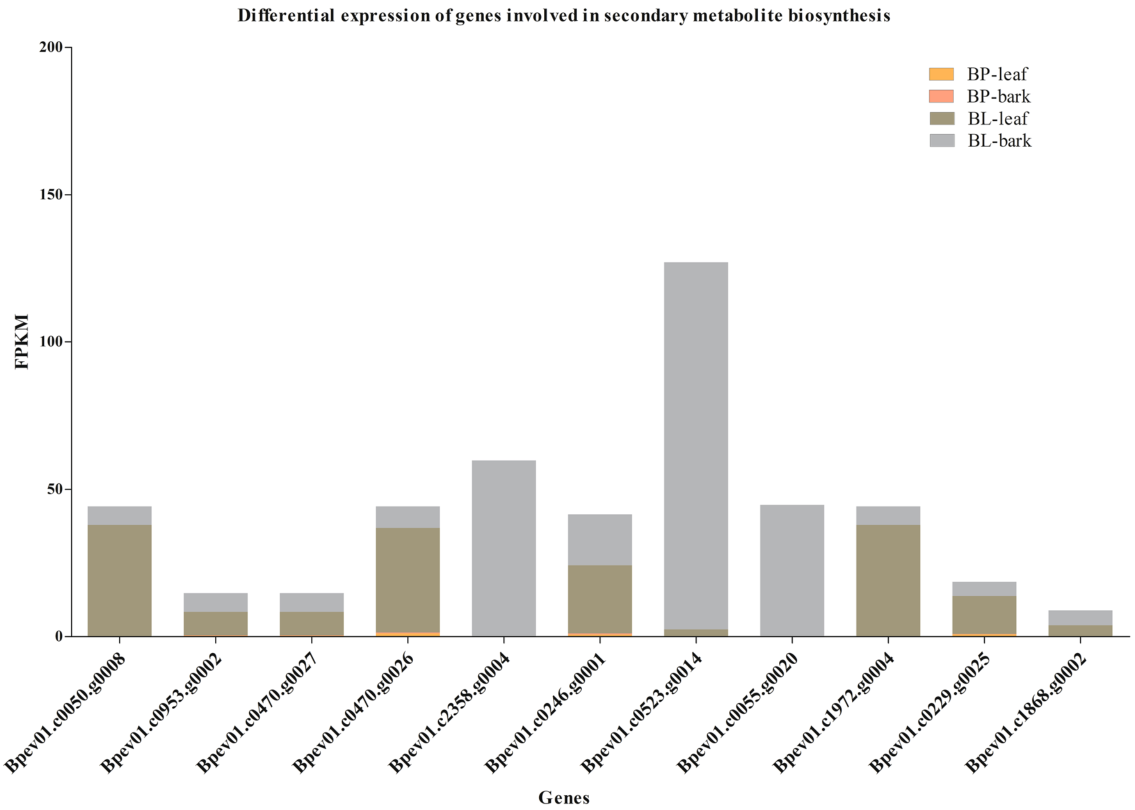

Fig. 5 Putative genes involved in secondary metabolite biosynthesis. A total number of 11 genes possibly involved in secondary metabolite biosynthesis were identified. All the genes showed upregulated expression in the bark and leaf tissue of *B. lenta* when compared t*o B. pendula*

classes like biological processes and molecular functions were identified (Fig. 2). Within the molecular function category, the vast majority of genes were related to catalytic activity, following the abundant secondary metabolites produced in *B. lenta*. Further, 18 genes were upregulated in *B. lenta* as assigned to the enriched GO term 'aromatic compounds biosynthesis process' in the main class biological processes (Fig. 6).

Through mapping the genes onto the KEGG pathways, 11 genes were discovered to be assigned to the enriched pathway 'biosynthesis of secondary metabolites' (Fig. 4). Among them, two genes have a putative function in shikimate dehydrogenase, four in beta-glucosidase, and five each involved in pyruvate dehydrogenase (PDH), cytochrome methyltransferase, gibberellin 2 oxidase, bioactive compounds, and disease resistance, respectively in *B. lenta* (Figs. 5 and 6). Genes from shikimate dehydrogenase have drawn great attention from researchers due to their special aromatic amino acid biosynthesis (Tzin and Galili 2018).

The identified 11 and 18 genes involved in secondary metabolism and biosynthesis of aromatic compounds were respectively, further subjected to NCBI BLAST search for their putative function confirmation (Table S7). The identification and annotation have provided the origin for analyzing specific pathways in *B. lenta*. These genes were involved in shikimate dehydrogenase, terpenoid backbone biosynthesis, and aromatic compounds including those encoding important proteins and regulatory elements (Gilchrist and Kosuge 1980; Tzin and Galili 2018). Unfortunately, the experimental validation of these novel genes was not part of the current study. Therefore, further functional characterization of these new and innovative findings will improve our understanding of the molecular mechanisms underlying aromatic amino acid and terpenoid biosynthesis that are important for therapeutics. Functional characterization of these genes will be vital to confirm the activity as well as use in therapeutics.

In addition, our study examined the differentially expressed genes between the bark and leaf based on FPKM values (Fig. 8). The results suggested that several genes were uniquely expressed in either bark or leaf tissues and many genes were expressed at different levels. Further study on these DEGs combined with metabolomes will enable us to more clearly understand the biosynthetic process of secondary metabolites.

TFs affect the metabolic flux by regulating gene expression. In this work, a total of 39 TFs specifically upregulated in *B. lenta* were identified, including bHLH (2), AP2/ERF (7), MYB (4), DOF (4), NAC (6), SPL9 (3), and WRKY

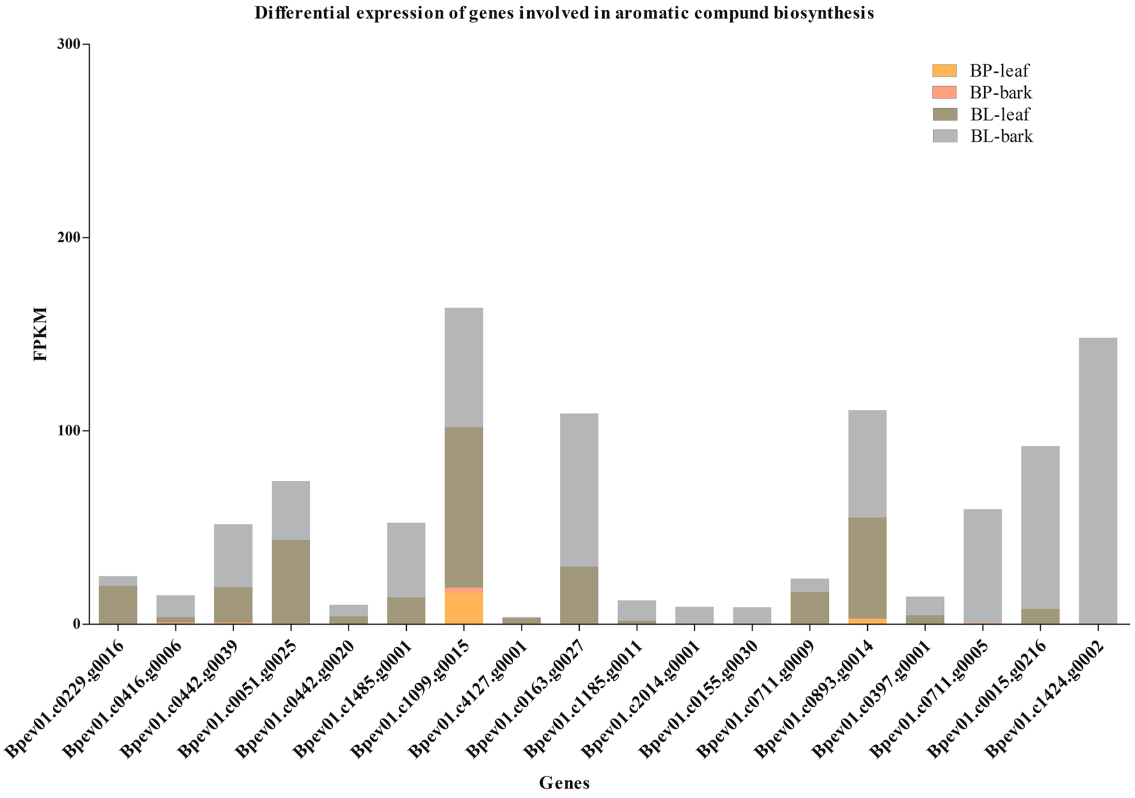

Fig. 6 Putative genes involved in aromatic compound biosynthesis. A total number of 18 genes involved in aromatic compound biosynthesis were identified. All the genes showed upregulation in bark and leaf tissues of *B. lenta* while low or no expression in *B. pendula*

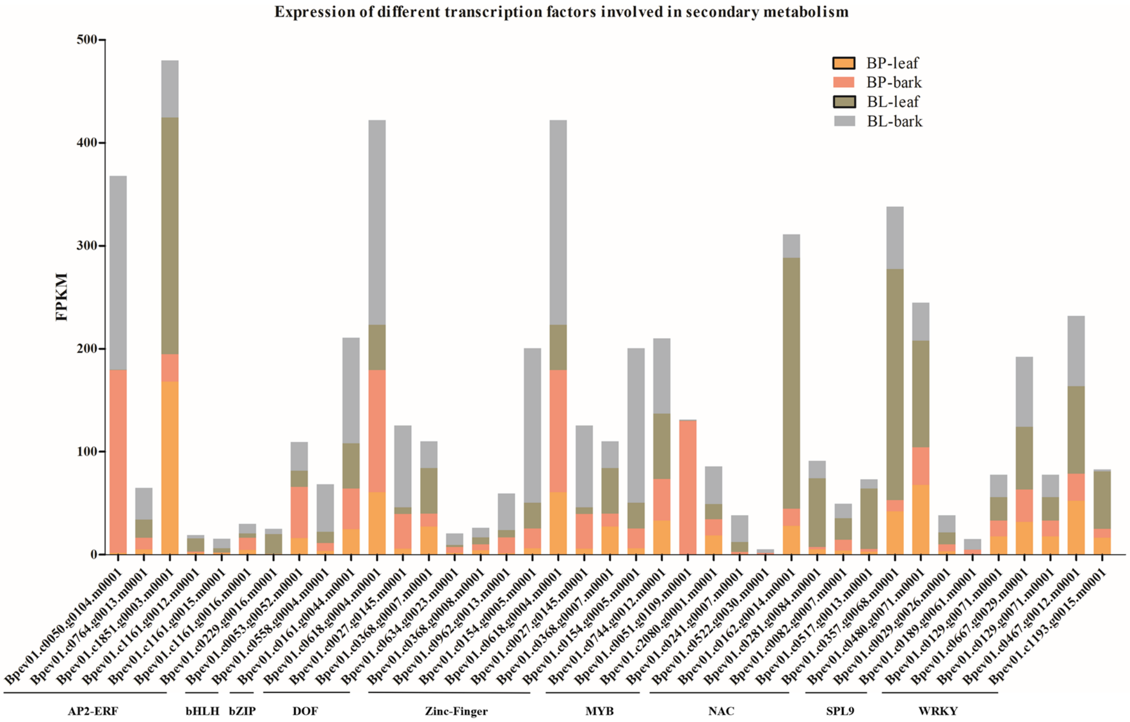

Fig. 7 Different transcription factors involved in secondary metabolism: A total number of 39 transcription factors identified showed upregulation in *B. lenta* versus *B. pendula*. All the identified genes are present on the x-axis, while transcript abundance is displayed on the y-axis

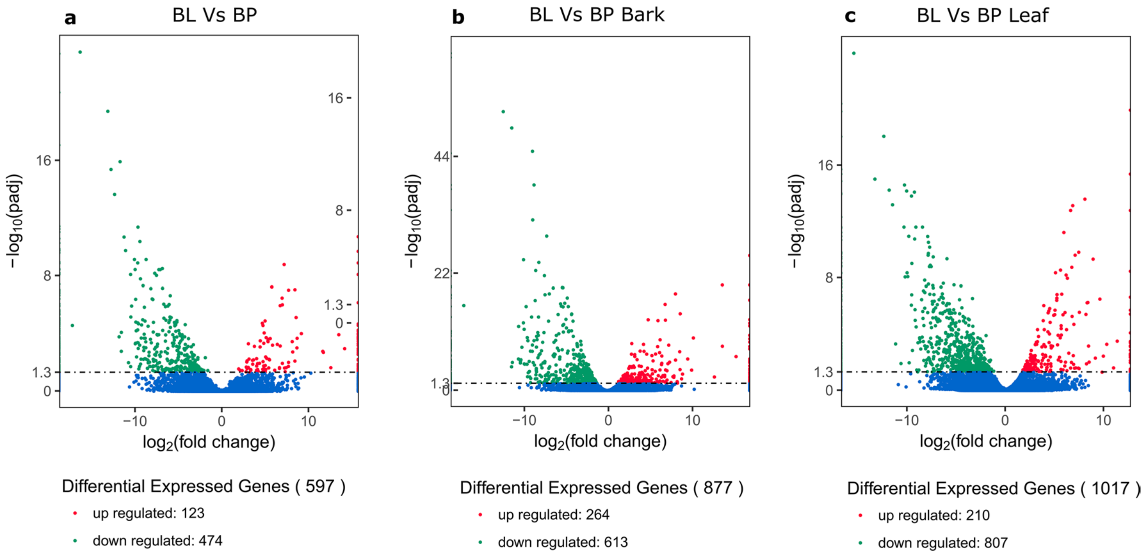

Fig. 8 Volcano plot showing differentially expressed genes in three different experimental comparisons. The x-axis shows the fold$_2$ (fold change) in gene expression between different samples, and the y-axis shows the statistical significance of the differences. Three different comparative analyses were carried out showed in **a–c** (BL: *B. lenta*; BP: *B. pendula*). Significantly up- and down-regulated genes in both species (adjusted $p < 0.01$ and absolute log2 fold change > 1.5) are highlighted in red and green, respectively. The dashed line indicates the p-value significance threshold. Blue dots represent genes that did not express differently in the comparison shown

(5) families (Fig. 7). Many studies have shown that these TFs play noteworthy roles in modulating the biosynthesis of secondary metabolites. For example, the bHLH transcription factor, AabHLH1 in *Artemisia annua*, constructively controls the biosynthesis of artemisinin (Ji et al. 2014). The AP2/ERF members of ORCA2 and ORCA3 in *Catharanthus roseus* hold together with the promoter of strictosidine synthases (STR) to tune the terpenoid indole alkaloid metabolism (van der Fits et al. 2001). It is significant to subject all these TF for the gene-editing methods to detect the TFs for regulating different aromatic amino acid and terpenoid biosynthesis in *B. lenta*. Further research on these TFs will be beneficial for the alteration of these metabolic pathways and eventually for escalating the production of secondary metabolites with medicinal value in *B. lenta*.

Conclusions and outlook

In this study, we presented the importance of transcriptome sequencing of the therapeutically important forest tree species *B. lenta*. A total number of 24,000 expressed genes were listed and several differentially expressed genes were annotated by the GO and KEGG database, referring to different plant metabolic pathways and biosynthesis processes. The identified candidate genes by KEGG and GO database are vital to understanding important regulators of secondary metabolite biosynthesis. Here, we also focused on searching for candidate genes involved in aromatic compound biosynthesis, in that 29 genes were involved in this bioprocess. The transcriptome information presented in our study also revealed that various genes are involved in the biosynthetic pathways of phenylpropanoid, flavonoids. Additionally, our study identified several transcription factors related to the biosynthesis of secondary metabolites in *B. lenta*. Taken together, the transcriptome data generated in our study allowed for discovering novel genes involved in specific secondary metabolic pathways and provide the basis for improving the yields of valuable metabolites in plants by metabolic engineering. Moreover, it is also highly valuable to pave the way for functional and comparative genomic studies of this promising medicinal plant *B. lenta* in the future. Unfortunately, biological validation of the unique genes was not feasible in the frame of this study. Experimental affirmation of the study is the next step and planned in the second part of the study.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10265-021-01343-y.

Acknowledgements We are extremely grateful to Dr. Niels Andreas Mueller for assisting with the data handling and manuscript review, and to all the staff of the Thuenen-Institute of Forest Genetics, Grosshansdorf, Germany, for their support, including Katrin Groppe, Susanne Jelkmann, and Vivian Kuhlenkamp for technical assistance, and Dr. Hilke Schroeder and Dr. Hans Hoenicka for helpful discussions on the plant secondary metabolic processes, and thanks are also to Ana Montalvao (Thuenen-Institute of Forest Genetics, Grosshansdorf, Germany) and Avneesh Kumar (Plant Breeding Institute, University of Kiel, Kiel, Germany) for assistance in data handling. Many thanks to the staff from the Institute for Agricultural Process Engineering, Christian-Albrechts University of Kiel, Germany. Thanks are due to Dina Fuehrmann (Thuenen Institute, Braunschweig, Germany) for English language editing.

Author contributions MF contributed to the experimental outline, data validation, and correcting of the manuscript. KB contributed to data investigation, bioinformatics analysis, and manuscript corrections. CRM has contributed to sample collection and writing the project proposal. EH guided the experiments and was also responsible for English language editing. Initial data and bioinformatics analysis, laboratory work, writing the first draft of the manuscript were completed by KS.

Funding Open Access funding enabled and organized by Projekt DEAL. The project was funded by the "Energiewende und Umweltinnovationen" state program for the economy; Ministry for Energy Transition, Agriculture, Environment, Nature and Digitalization of Schleswig-Holstein, Germany. [Project number: LPW-L/1.2/24] Principal Investigator: Dr. Christian R. Moschner.

Data availability The RNA-seq data set has been deposited in NCBI as BioProject under the number PRJNA756395.

Declarations

Conflict of interest All authors of the research article have no conflicts of interest to disclose.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

Ashburner K, McAllister HA (2013) The genus *Betula*: a taxonomic revision of birches. Kew Publishing, London

Bina H, Yousefzadeh H, Ali SS, Esmailpour M (2016) Phylogenetic relationships, molecular taxonomy, biogeography of *Betula*, with emphasis on the phylogenetic position of Iranian populations. Tree Genet Genomes 12:84. https://doi.org/10.1007/s11295-018-1228-2

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

Chen S, Wang Y, Yu L et al (2021) Genome sequence and evolution of *Betula platyphylla*. Hortic Res 8:37. https://doi.org/10.1038/s41438-021-00481-7

COSEWIC (2006) Assessment and status report on the cherry birch *Betula lenta* in Canada. Committee on the Status of Endangered Wildlife in Canada (COSEWIC). https://www.canada.ca/en/environment-climate-change/services/species-risk-public-registry/cosewic-assessments-status-reports/cherry-birch.html. Accessed 5 May 2021

De Jong PC (1993) An introduction to *Betula*: its morphology, evolution, classification and distribution with a survey of recent work. In: Proceedings of the IDS *Betula* symposium, 2–4 October 1992. International Dendrology Society, Richmond

Dumolin S, Demesure B, Petit RJ (1995) Inheritance of chloroplast and mitochondrial genomes in pedunculate oak investigated with an efficient PCR method. Theor Appl Genet 91:1253–1256. https://doi.org/10.1007/BF00220937

Ebeling S, Naumann K, Pollok S, Wardecki T, Vidal-Y-Sy S, Nascimento JM, Boerries M, Schmidt G, Brandner JM, Merfort I (2014) From a traditional medicinal plant to a rational drug: understanding the clinically proven wound healing efficacy of birch bark extract. PLoS One. https://doi.org/10.1371/journal.pone.0086147

Fernald ML, Kinsey AC, Rollins RC (1958) Edible wild plants of eastern North America. Harper and Bros, New York

Furlow JJ (1990) The genera of Betulaceae in the southeastern United States. J Arnold Arboretum 71:1–67. https://doi.org/10.5962/bhl.part.24925

Galis I, Simek P, Narisawa T, Sasaki M, Horiguchi T, Fukuda H, Matsuoka K (2006) A novel R2R3 MYB transcription factor NtMYBJS1 is a methyl jasmonate-dependent regulator of phenylpropanoid-conjugate biosynthesis in tobacco. Plant J 46:573–592. https://doi.org/10.1111/j.1365-313X.2006.02719.x

Gilchrist D, Kosuge T (1980) Aromatic amino acid biosynthesis and its regulation. In: Miflin BN (ed) The biochemistry of plants, vol&nbsp;5, vol 8. Academic Press, New York, pp 507–531

Gilmore MR (1933) Some Chippewa uses of plants. Mich Acad Sci Arts Lett 17:119–232

Guan X, Stege J, Kim M, Dahmani Z, Fan N, Heifetz P, Barbas CF 3rd, Briggs SP (2002) Heritable endogenous gene regulation in plants with designed polydactyl zinc finger transcription factors. Proc Natl Acad Sci USA 99:13296–13301. doi:https://doi.org/10.1073/pnas.192412899

Jack JG (1895) Hybrid birches. Garden For 8:243–244

Jakoby M, Weisshaar B, Dröge-Laser W, Vicente-Carbajosa J, Tiedemann J, Kroj T, Parcy F, bZIP Research Group (2002) bZIP transcription factors in *Arabidopsis*. Trends Plant Sci 7:106–111. https://doi.org/10.1016/s1360-1385(01)02223-3

Ji Y, Xiao J, Shen Y, Ma D, Li Z, Pu G, Li X, Huang L, Liu B, Ye H, Wang H (2014) Cloning and characterization of AabHLH1, a bHLH transcription factor that positively regulates artemisinin biosynthesis in *Artemisia annua*. Plant Cell Physiol 55:1592–1604. https://doi.org/10.1093/pcp/pcu090

Jin YG, Felipe FF, Chang JL, Weigel D, Wang JW (2011) Negative regulation of anthocyanin biosynthesis in *Arabidopsis* by a miR156-targeted SPL transcription factor. Plant Cell 23:1512–1522. https://doi.org/10.1105/tpc.111.084525

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley, Salzberg SL (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14:R36. https://doi.org/10.1186/gb-2013-14-4-r36

Lamson NI (1990) *Betula lenta* L. Sweet Birch. In: Burns RM, Honkala B (eds) Silvics of North America: hardwoods vol&nbsp;2. Agriculture handbook 654. U.S. Department of Agriculture, Washington, pp 148–152

Leak WB, Fowells HA (1965) Sweet birch (*Betula lenta* L.). Silvics of forest trees of the United States, vol 271. U.S. Department of Agriculture, Washington, pp 99–109

Li JH, Shoup S, Chen ZD (2005) Phylogenetics of *Betula* (Betulaceae) inferred from sequences of nuclear ribosomal DNA. Rhodora. https://doi.org/10.3119/04-14.1

Lorimer CG (1980) Age structure and disturbance history of a southern Appalachian virgin forest. Ecology 6:1160–1184

Mizoi J, Shinozaki K, Yamaguchi SK (2012) AP2/ERF family transcription factors in plant abiotic stress responses. Biochim Biophys Acta 1819:86–96. https://doi.org/10.1016/j.bbagrm.2011.08.004

Rastogi S, Pandey MM, Kumar Singh RA (2015) Medicinal plants of the genus *Betula*-traditional uses and a phytochemical-pharmacological review. J Ethnopharmacol 159:62–83. https://doi.org/10.1016/j.jep.2014.11.010

Robinson MD, Oshlack AA (2010) Scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 11:R25. https://doi.org/10.1186/gb-2010-11-3-r25

Salojärvi J, Smolander OP, Nieminen K et al (2017) Genome sequencing and population genomic analyses provide insights into the adaptive landscape of silver birch. Nat Genet 49:904–912. https://doi.org/10.1038/ng.3862

Sharik TL, Burton VB (1971) Hybridization in *Betula alleghaniensis*&nbsp;Britt. and *B. lenta*&nbsp;L.: a comparative analysis of controlled crosses. For Sci 17:415–424

Singewar K (2020) Phylogenetic relationships, marker analysis, and investigation of genes mediating high and low methyl salicylate biosynthesis in different birch species (*Betula* L., Betulaceae). Dissertation. urn:nbn:de:gbv:8:3-2021-00129-6. Accessed 17 May 2021

Singewar K, Moschner CR, Hartung E, Fladung M (2020a) Identification and analysis of key genes involved in methyl salicylate biosynthesis in different birch species. PLoS One 15:e0240246. https://doi.org/10.1371/journal.pone.0240246

Singewar K, Moschner CR, Hartung E, Fladung M (2020b) Species determination and phylogenetic relationships of the genus *Betula* inferred from multiple chloroplast and nuclear regions reveal the high methyl salicylate-producing ability of the ancestor. Trees doi. https://doi.org/10.1007/s00468-020-01984-x

Skibbe M, Qu N, Galis I, Baldwin IT (2008) Induced plant defenses in the natural environment: *Nicotiana attenuata* WRKY3 and WRKY6 coordinate responses to herbivory. Plant Cell 20:1984–2000. https://doi.org/10.1105/tpc.108.058594

Skirycz A, Reichelt M, Burow M, Birkemeyer C, Rolcik J, Kopka J, Zanor MI, Gershenzon J, Strnad M, Szopa J, Mueller-Roeber B, Witt I (2006) DOF transcription factor AtDof1.1 (OBP2) is part of a regulatory network controlling glucosinolate biosynthesis in *Arabidopsis*. Plant J 47:10–24. https://doi.org/10.1111/j.1365-313X.2006.02767.x

Stephens GR, Waggoner PE (1970) The forests anticipated from 40 years of natural transitions in mixed hardwoods: Bulletin 707. Connecticut Agricultural Experiment Station, New Haven, p 58

Suryawanshi DG (2000) An ancient writing material: Birch-Bark and its need of conservation. Restaurator 21(1):1–8. https://doi.org/10.1515/REST.2000.1

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28:511–515. https://doi.org/10.1038/nbt.1621

Tzin V, Galili G (2018) The Biosynthetic Pathways for Shikimate and Aromatic Amino Acids in *Arabidopsis thaliana*. Arabidopsis Book 8:e0132. https://doi.org/10.1199/tab.0132

van der Fits L, Memelink J (2001) The jasmonate-inducible AP2/ERF-domain transcription factor ORCA3 activates gene expression via interaction with a jasmonate-responsive promoter element. Plant J 25:43–53. doi:https://doi.org/10.1111/j.1365-313X.2001.00932.x

Wang L, Feng Z, Wang X, Wang X, Zhang X (2010) DEGseq: identify differentially expressed genes from RNA-seq data. Bioinformatics 26:136–8. https://doi.org/10.1093/bioinformatics/btp612

Wang N, McAllister HA, Bartlett PR, Buggs RJ (2016) Molecular phylogeny and genome size evolution of the genus *Betula* (Betulaceae). Ann Bot 117:1023–1035. https://doi.org/10.1093/aob/mcw048

Winkler H (1904) Betulaceae. In: Engler A (ed) Das Pflanzenreich, vol 19. W. Engelmann, Leipzig Heft, pp 1–49

Yang CQ, Fang X, Wu XM, Mao YB, Wang LJ, Chen XY (2012) Transcriptional regulation of plant secondary metabolism. J Integr Plant Biol 54:703–712. https://doi.org/10.1111/j.1744-7909.2012.01161.x

Yin J, Ren CL, Zhan YG, Li CX, Xiao JL, Qiu W, Li XY, Peng HM (2012) Distribution and expression characteristics of triterpenoids and OSC genes in white birch (*Betula platyphylla* suk.). Mol Biol Rep 39:2321–2328. https://doi.org/10.1007/s11033-011-0982-0

Yin J, Li X, Zhan Y, Li Y, Qu Z, Sun L, Wang S, Yang J, Xiao J (2017) Cloning and expression of BpMYC4 and BpbHLH9 genes and the role of BpbHLH9 in triterpenoid synthesis in birch. BMC Plant Biol 17:214. https://doi.org/10.1186/s12870-017-1150-z

Young MD, Wakefield MJ, Smyth GK, Oshlack A (2010) Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol 11:R14. https://doi.org/10.1186/gb-2010-11-2-r14

Zoladeski C, Hayes K (2013) Recovery strategy for the Cherry Birch (*Betula lenta*) in Ontario. Ontario Recovery Strategy Series. Prepared for the Ontario Ministry of Natural Resources, Peterborough, Ontario. vi + 12 pp. https://files.ontario.ca/environment-and-energy/species-at-risk/stdprod_075574.pdf. Accessed 5 June 2021