Convergent evolution of mevalonate pathway in *Inonotus obliquus* and *Betula pendula*.

Omid Safronov[1,2,3], Guleycan Lutfullahoglu Bal[2], Nina Sipari[4], Maya Wilkens[5], Pezhman Safdari[1], Olli-Pekka Smolander[6], Jenna Lihavainen[7], Niko Silvan[8], Sitaram Rajaraman[1], Pia K. Laine[2], Lars G Paulin[2], Petri Auvinen[2], Tytti Sarjala[8], Kirk Overmyer[1], Jaakko Kangasjärvi[1], Brendan Battersby[2,3], Uwe Richter[2,9,*], Jarkko Salojärvi[1,10,*].

[1]Organismal and Evolutionary Biology Research Program, Faculty of Biological and Environmental Sciences, and Viikki Plant Science Centre, University of Helsinki, Finland.

[2]Institute of Biotechnology, HiLIFE, University of Helsinki, Helsinki, Finland.

[3]Molecular and Integrative Biosciences Research Program, Faculty of Biological and Environmental Sciences, and Viikki Plant Science Centre, University of Helsinki, Finland.

[4]Viikki Metabolomics Unit, Faculty of Biological and Environmental Sciences, University of Helsinki, Helsinki Finland.

[5]Current address: Quantitative Proteomics, Institute of Molecular Biology, Mainz, Germany.

[6]Department of Chemistry and Biotechnology, Tallinn University of Technology, Tallinn, Estonia.

[7]Department of Plant Physiology, Umeå Plant Science Centre, Umeå University, Umeå, Sweden.

[8]Biomass Characterization and Properties Group, Production Systems Unit, Natural Resources Institute Finland, Helsinki, Finland.

[9]Wellcome Centre for Mitochondrial Research, Biosciences Institute, Newcastle University, UK.

[10]School of Biological Sciences, Nanyang Technological University, Singapore, Singapore.

*Corresponding authors: uwe.richter@helsinki.fi, jarkko@ntu.edu.sg

Abstract

Inonotus obliquus, Chaga mushroom, is a fungal species from *Hymenochaetaceae* family (*Basidiomycota*) which has been widely used for traditional medicine in Europe and Asia. Here, chaga genome was sequenced using Pacbio sequencing into a 50.7Mbp assembly consisting of 301 primary contigs with an N50 value of 375 kbp. Genome evolution analyses revealed a lineage-specific whole genome duplication event and an expansion of Cytochrome P450 superfamily. Fungal biosynthetic clusters were enriched for tandemly duplicated genes, suggesting that biosynthetic pathway evolution has proceeded through small-scale duplications. Metabolomic fingerprinting

34 confirmed a highly complex terpene biosynthesis chemistry when compared against related fungal

35 species lacking the genome duplication event.

Introduction

37 *Inonotus obliquus*, Chaga mushroom, is a fungal species from *Hymenochaetaceae* family

38 (*Basidiomycota*) distributed across the boreal forest zone in the Northern hemisphere. It causes

39 aggressive white rot disease mainly among *Betula* family members (Blanchette, 1982), but upon

40 suitable conditions it can infect also other tree species such as oaks, poplars, ashes and maples

41 (Ryvarden & Gilbertson, 1993). White rot disease is the result of lignin degradation (having darker

42 color) while the light-coloured cellulose is left intact. The infection starts when *I. obliquus* spores get

43 access to the hardwood of the stem through an opening or wounded bark. At the later stages of the

44 infection, *I. obliquus* appears as a sterile conk, a solid charcoal-black mass on the surface of bark

45 (Blanchette, 1982). The sterile conk has been used in traditional medicine in many cultures. A large

46 body of research on biochemical compounds extracted from the conk suggests that the species may

47 have a wide range of pharmaceutical, medicinal, and industrial applications (Ma, Chen, Dong, & Lu,

48 2013; Nagajyothi, Sreekanth, Lee, & Lee, 2014; Song, Liu, Kong, Chang, & Song, 2013; Yan et al.,

49 2014).

50 Betulin (BE) and betulinic acid are highly abundant triterpenoids in the bark of all birch family

51 members involved in protection against fungi, bacteria and viruses; they collectively form 30-60% of

52 total tissue composition, depending on the species and the tissue type (Holonec, Ranga, Crainic,

53 Truța, & Socaciu, 2012; P. Kovalenko et al., 2009; Safronov et al., 2019). Both betulinate compounds

54 are being studied for industrial applications (Šiman et al., 2016), and as therapeutic substances in

55 oncology (Król, Kiełbus, Rivero-Müller, & Stepulak, 2015) and infectious diseases (fungal, bacterial,

56 and viral infections) (Gong et al., 2004; Salin et al., 2010; Shai, McGaw, Aderogba, Mdee, & Eloff,

57 2008). In plants, the biosynthesis of betulinate compounds starts with squalene, a product of

58 mevalonate pathway, and involves two enzymatic steps where squalene is first converted to lupeol

59 via lupeol synthase and then to betulinate by lupeol monooxygenase, an enzyme which is a member

60 of the large family of cytochrome P450 monooxygenases, more specifically subfamily 716 (CYP716).

61 Betulin biosynthesis is found across a wide taxonomic range in plants, from *Malvales* (H. J. Zhang et

62 al., 2003), *Fagales* (Safronov et al., 2019), *Rosales* (Andre et al., 2013; S. Zhao et al., 2015), *Fabales*

63 (Wu, Niu, Bakur, Li, & Chen, 2017), *Vitales* (Fukushima et al., 2011), and *Asterales* (Siddiqui et al.,

64 2019) to *Arecales* (Khelil, Jardé, Cabello-Hurtado, Ould-el-Hadj Khelil, & Esnault, 2016; Koolen et al.,

65 2012), suggesting either ancestral origin or convergent evolution. A comparative genomic analysis

66 of the bark tissue in silver birch (*Betula pendula*) and grey alder (*Alnus glutinosa*) revealed birch-

67 specific evolution of mevalonate pathway (MVA), where a tandem duplication of lupeol synthase

68 colocalized with lupeol 28-monooxygenase was suggested as the reason for increased production of

69 betulinate compounds in birch phellem (Safronov et al., 2019). Interestingly, in addition to plants,

70 betulinate compounds have also been identified in diverse range of fungal species from *Eurotiales*

71 (Khouloud Barakat, 2016), *Hymenochaetales* (Yin, Cui, & Ding, 2008), and *Polyporales* (Alresly et al.,

72 2015) families, even though no members of CYP716 gene family have yet been identified or

73 characterized in fungi. Birch fungal pathogens *Inonotus obliquus* and *Fomitopsis betulina* are such

74 examples of fungal species that produce BE and BA compounds (Alresly et al., 2015; Yin et al., 2008),

75 even though the compounds are anti-fungal by nature. The evolution of betulinate biosynthesis in

76 these two fungal species is not known, but one can hypothesize it to be the result of either

77 convergent evolution or horizontal gene transfer (HGT) of the responsible cytochrome P450

78 monooxygenase enzymes from the host species. There exists a recent study on the diversification

79 and distribution of CYP716 enzyme in eudicots (Miettinen et al., 2017), but no studies of this enzyme

80 in fungal species have been carried out, and the enzymes underlying betulin production in the fungal

81 species, known to produce betulinate compounds, have not yet been identified.

82 The CYP450 monooxygenase enzymes are among the oldest and largest gene families,

83 encompassing both prokaryotic and eukaryotic organisms (Sezutsu, Le Goff, & Feyereisen, 2013).

84 They act as key enzymes for detoxification of toxic compounds, and they have an important function

85 in secondary metabolism related to adaptation to environmental conditions. The low sequence

86 similarity, high functional diversity and enzymatic promiscuity among CYP450 monooxygenase

87 enzymes makes functional predictions difficult. The CYP450s are generally classified into families

88 and subfamilies based on sequence similarity; the sequences with identity >40% are assigned into

89 families and sequences with >55% similarity into their own subfamilies; novel candidates with lower

90 identity to the set of identified CYP450s form new candidate families. Based on these criteria, so far

91 over 800 different CYP families have been identified (Lepesheva et al., 2008).

92 In this study we sequenced and assembled an *I. obliquus* genome from an isolate from Merikarvia

93 region in Finland. The genome was annotated using *ab initio* gene model prediction and spliced

94 transcript data obtained from total RNA sequencing. We carried out comparative genomic and

95 gene family expansion analysis among 16 *Basidiomycete* and 3 *Ascomycetes* species together with

96 *I. obliquus* genome and studied the untargeted terpenoid metabolic fingerprints (using UPLC-

97 QTOF/MS) in five strains of *I. obliquus* and one *Fomitiporia mediterranea* strain, focusing on the

98 quantification of betulin (BE) and betulinic acid (BA) abundances across the samples. To confirm

99 our functional predictions we cloned the candidate lupeol synthase from *B. pendula* and CYP450

100 monooxygenase enzymes from both *B. pendula* and *I. obliquus* and tested their ability to produce

101 betulin compounds.

Materials and methods

Sample collection

104 Four *Inonotus obliquus* strains (Supp. table 1) were collected and isolated from different regions

105 in Finland and one from Altai mountains in Russia. The strain from Merikarvia was selected for whole

106 genome sequencing (location 61°58'38.6"N 21°44'43.1"E). In addition, we also obtained a strain of

107 *Fomitiporia mediterranea* as an outgroup to chaga (Mycobank: MB384943). All samples were

108 cultivated on Hagem agar overlayed by a cellophane membrane.

109 The isolation of chaga mushrooms from the host trees was done by cutting a piece of the conk (Supp.

110 Fig 1), which was then laid on agar plate after short H_2O_2 bath. The samples were re-cultured

111 repeatedly and sequenced for internal transcribed spacer 1 (ITS1) [TCCGTAGGTGAACCTGCGG] and

112 ITS4 [TCCTCCGCTTATTGATATGC] regions confirm the species assignment of *I. obliquus* isolate.

RNA isolation, sequencing, and *de novo* assembly of transcriptome

114 To isolate the total RNA from *I. obliquus,* the method from Chang *et al.* (Chang, Puryear, & Cairney,

115 1993) was used. Briefly, the *I. obliquus* was inoculated and grown on autoclaved wood dust from a

116 clone of *B. pendula* (12 years old tree, 167 cm^2 disk, dry weight of 200 grams) sequenced for *B.*

117 *pendula* reference genome (Salojärvi et al., 2017). A total of 150 milligrams of ground sample

118 (mortar and pestle, and liquid N_2) was transferred on ice for 30 seconds, and 500 µl of pre-warmed

119 (+65-68°C) extraction buffer (2% CTAB, 2% PVP K-30, 100 mM Tris-HCl [pH 8.0], 25 mM EDTA, 2 M

120 NaCl, and 200 µl β-MeOH/10 ml of extraction buffer) was added and vortexed vigorously. Extraction

121 was carried out three times with chloroform:isoamyl alcohol (24:1) by spinning at 200-300 rpm for

122 15 minutes, and then centrifuging at 10 000 rpm 15 minutes. Then, 1/4 volume 10 M LiCl was added

123 and left to precipitate on ice overnight. The overnight sample was centrifuged with 10000 rpm for

124 20-30 minutes at +4°C, and the resulting pellet was dissolved in 500 µl of pre-warmed (+65°C)

125 sodium dodecyl sulfate–Tris-HCl–EDTA (SSTE) buffer, and extracted once (or several times, if

126 necessary) with chloroform:isoamyl alcohol (24:1). The mixture was precipitated by adding 2

127 volumes of absolute EtOH (place at -20°C overnight), and centrifuged at 13 000 rpm, for 20-30

128 minutes at +4°C. the precipitate was washed with 70% EtOH, after which the pellet was dried, and

129 then dissolved in 10-30 µl RNase-free water, and RNase inhibitor was added.

130 TruSeq stranded mRNA kit was used to construct the RNA-seq library. The cDNA was

131 synthesized from 5 µl of total RNA extracted from reference *I. obliquus* plate using random

132 hexamers. DNA polymerase I and dUTP nucleotides were used to synthesize the second strand of

133　cDNA. Then, double stranded cDNA were purified, and ends were repaired. Library preparation

134　was continued by A-tailing, and ligation of Y-adaptors containing indexes from the kit. The

135　fragments were amplified using polymerase chain reaction (PCR), followed by purification steps

136　using AMPure XP. The sequencing was carried out in HiScan SQ platform (paired-end 88 bp + 74

137　bp).

138　　The raw paired end RNA-seq data were controlled for quality using FastQC v0.11.2 (Andrews).

139　Trimmomatic v0.33 (Bolger, Lohse, & Usadel, 2014) was used in pair-end mode to remove the

140　adapters, barcodes, low quality bases from both ends of each sequence, and reads shorter than 25

141　base pairs (LEADING:20, TRAILING:20, MINLEN:25, -phred33). After the removal of duplicate

142　sequences, the unpaired sequences were mapped to *I. obliquus* reference genome using Tophat2

143　(Kim et al., 2013) for junction discoveries (-i:10, and --coverage-search); paired end reads were

144　mapped separately (Tophat2; -i:10, and --coverage-search). The aligned reads were separated

145　according to their orientation on reference genome to forward and reverse strands, which were then

146　aligned　individually　by　Trinity　v2.1.1,　using　--genome_guided_bam,　and　--

147　genome_guided_max_intron: 1 000 options (Grabherr et al., 2011) for *de novo* transcriptome

148　assemblies. The forward and reverse *de novo* transcriptome assemblies were combined, and

149　duplicated assemblies were removed using GenomeTools v1.5.1, using sequniq option (Gremme,

150　Steinbiss, & Kurtz, 2013). The unique *de novo* transcriptome assemblies were clustered by using CD-

151　HIT v4.6 (Godzik & Li, 2006) and aligned to *I. obliquus* reference genome by Program to Assemble

152　Spliced Alignments (PASA v2.2.0) (Brian J. Haas et al., 2008).

153　　The processing of the publicly available RNAseq data (Fradj et al., 2019) was carried out in a

154　similar manner. Both data sets were mapped to *I. obliquus* gene models using kallisto quant v0.44.0

155　(Bray, Pimentel, Melsted, & Pachter, 2016). The orphan reads and pair-end reads (separated during

156　preprocessing by trimmomatic) were mapped separately by using kallisto quant single (options: --

157　single, -l 200, -s 20, -b 4000) and pair-end (option: -b 4000) modes, respectively. The raw count table

158　from Kallisto was imported to R (for both single and pair-end count tables) using tximport package

159　v1.18.0 with default options (Soneson, Love, & Robinson, 2015). The single and pair-end counts were

160　summed together to form a single count table for each data set. Differential gene expression analysis

161　was conducted using DESeq2 (Love, Huber, & Anders, 2014). The final tables for differentially

162　expressed genes (DEg) were filtered based on the false discovery rate adjusted p-value threshold of

163　0.05 (p-adj. ≤ 0.05).

164 **DNA isolation, genome assembly and annotation**

165   Modified version of Lodhi *et al.* (Lodhi, Ye, Weeden, & I. Reisch, 1994) was used for DNA

166 extraction from *I. obliquus* strains. Maximum 0.5 g of material was ground in liquid N_2. The ground

167 sample was transferred into ice cold Sodium chloride-Tris-EDTA (STE) buffer (1,4 M NaCl, 0 mM

168 EDTA, 100 mM Tris-HCl pH 8.0), and centrifuged for 5 minutes at 8 000 rpm and +4°C. STE buffer

169 was discarded, and 10 ml of pre-warmed (60°C) cetyltrimethyl ammonium bromide (CTAB) buffer

170 (1 liter CTAB: 20 mM EDTA, 100 mM Tris-HCl pH 8.0, 1.4 M NaCl, 2.0% CTAB, 1.0% PVP 40, and 2%

171 β-MeOH [50µl]) was added to pellet. Subsequently, the mixture was vortexed and incubated for 30-

172 60 min at 60°C and cooled to the room temperature. Chloroform: isoamyl alcohol (IAA) (24:1 ration)

173 mixture was added for extraction (centrifugation: 15 minutes, 10000 rpm at room temperature).

174 The supernatant was collected to a new tube and mixed with 2X CTAB buffer, which was then

175 vortexed and incubated for 30-60 min at 60°C. The chloroform:IAA extraction step was repeated 2-

176 3 times, followed by adding of 2X volume of cold (-20°C) absolute ethanol (EtOH) to supernatant.

177 The EtOH mixture was stored for overnight at +4°C. The mixture then was centrifuged for 15

178 minutes, at 10000 rpm and 4°C. DNA pellet was washed with absolute EtOH (-20°C) and air dried.

179 The sample was treated for the RNA (RNase A), followed by chloroform:IAA extraction, EtOH

180 precipitation, air drying of the DNA pellet, dissolving in DNase/RNase free water, and storing at -

181 80°C.

182

183   The genome of the *Inonotus obliquus* was sequenced with Pacific Biosciences PacBio RSII

184 instrument using P6-C4 chemistry. Eight SMRTcells were used for sequencing the sample with movie

185 time of 240 minutes. The number of obtained sequences was 712,759 which totaled up to 4.82 Gb

186 of data with read length N50 of 9200 bp. At first, hierarchical Genome Assembly Process (HGAP) V3

187 implemented in SMRT Analysis package (v2.3.0) was used to generate an initial *de novo* genome

188 assembly with default parameters. Mitochondrial genome contig was separated from the

189 chromosomal contigs and circularized manually using GAP4 program (Bonfield, Smith, & Staden,

190 1995). Obtained mitochondrial sequence in length of 118 085 bp and > 4000X sequencing coverage

191 was polished using SMRT Analysis RS Resequencing protocol with Quiver consensus algorithm.

192 Second, the FALCON assembly program (Chin et al., 2016) was used to generate the final *de novo*

193 genome assembly with seed read length of 10 000 bp. Obtained contig sequences were polished

194 using SMRT Analysis RS Resequencing protocol with Quiver consensus algorithm with approximately

195 75x coverage. To quantify completeness of the genome, BUSCO (v3.0, Fungi datasets, -m geno, –

196 long) (Waterhouse et al., 2017) was used.

197   Repeat analysis of the contigs was carried out according to the guidelines of RepeatModeler and

198   RepeatMasker (http://www.repeatmasker.org/, v 4.0.7). To predict the gene models, multiple

199   evidence tracks from different platforms were obtained: *ab initio* gene predictors based on Hidden

200   Markov Models (HMMs), spliced transcript evidence from RNA-seq, and orthologous proteins

201   from closely related fungal species. HMM-based models such as AUGUSTUS (v3.3.2) (Stanke &

202   Morgenstern, 2005), and GeneMark-ES (version 4.33; --fungus mode, and –evidence: *de novo*

203   transcriptome assembly) (Besemer & Borodovsky, 2005) were used for *ab initio* gene predictions.

204   In addition, BRAKER2 (Hoff, Lange, Lomsadze, Borodovsky, & Stanke, 2015) (options: --fungus, --

205   rounds=100, and --bam) was run for *ab initio* gene predictions. To identify the open reading frames

206   (ORFs) within the genome, getorf (EMBOSS v6.6.0.0) program (Rice, Longden, & Bleasby, 2000) was

207   used (-find:1, and –maxsize: 5000). The ORFs were then queried against NR database by DIAMOND

208   (v0.9.24, blastp, --more-sensitive) (Buchfink, Xie, & Huson, 2014) and filtered for similarity

209   (sequence identity ≥ 75, and score ≥ 300); the homologous sequences above the threshold were

210   collected. Selected ORFs were used as the input for exonerate (v2.46.2, --model:protein2genome, -

211   -minintron:10, --maxintron:1000; --percent:65) (Slater & Birney, 2005) to map the candidate ORFs

212   to *I. obliquus* reference genome. Additionally, orthologous proteins from 13 fungal species

213   (*Coprinopsis cinerea, Fomitiporia mediterranea, Heterobasidion annosum, Laccaria bicolor, Onnia*

214   *scaura, Phanerochaete chrysosporium, Phellinus ferrugineofuscus, Porodaedalea niemelaei, Postia*

215   *placenta, Puccinia graminis, Rickenella mellea, Schizopora paradoxa, Trichaptum abietinum*) were

216   aligned against *I. obliquus* reference genome with exonerate (v2.46.2, --model:protein2genome, --

217   minintron:10, --maxintron:1000; --percent:65) (Slater & Birney, 2005). In addition to orthologous

218   proteins, the protein sequences discovered from BUSCO predictions were collected and aligned to

219   reference genome by exonerate as well using the same parameters as given above (Slater & Birney,

220   2005). All the evidence (*ab initio* gene models, spliced transcript alignments, spliced protein

221   alignments, ORFs, and BUSCO) was combined to consensus, high-confidence gene models, using

222   EVidenceModeler (v1.1.1). This was followed by the addition of untranslated regions (UTR) to the

223   gene models by PASA (Brian J. Haas et al., 2008).

224   Mitochondrial genome was also assembled and annotated as described previously (Salojärvi et

225   al., 2017), resulting in 29 tRNAs, 32 coding sequences, and 3 rRNAs.

226   Interproscan (v5.25-64.0) (Quevillon et al., 2005) was used to assign the protein function to gene

227   models. Additionally, Ensemble Enzyme Prediction (E2P2, v3.1) (Schlapfer et al., 2017) and

228   antiSMASH (v2.0) fungal version (Blin et al., 2013) were used to predict the metabolomic pathways.

Comparative genomic analyses

The proteomes of twenty fungal species *Laccaria bicolor, Coprinopsis cinerea, Schizophyllum commune, Fomitiporia mediterranea, Inonotus obliquus, Onnia scaura, Phellinidium ferrugineofuscum, Porodaedalea niemelaei, Trichaptum abietinum, Rickenella mellea, Schizopora paradoxa, Fomitopsis betulina, Postia placenta, Phanerochaete chrysosporium, Puccinia graminis, Heterobasidion annosum, Ustilago maydis, Saccharomyces cerevisiae, Schizosaccharomyces pombe,* and *Neurospora crassa* from Ascomycete and Basidiomycete clades were downloaded from MycoCosm (https://mycocosm.jgi.doe.gov) and included for gene family analysis by Orthofinder (Emms & Kelly, 2015) (v2.3.3), run with default parameters.

Synteny analyses

Synteny analysis of self-self alignment of *I. obliquus*, and four other fungal species, namely *F. mediterranea*, *S. paradoxa*, *F. betulina*, and *P. niemelaei,* were conducted using SynMap application in CoGe platform (https://genomevolution.org/coge/), using Quota Align algorithm with default parameters. The list of syntenic duplicates were obtained from DAGchainer (B. J. Haas, Delcher, Wortman, & Salzberg, 2004); tandem duplicates were obtained as part of the preprocessing pipeline.

Discovery of secreted proteins and carbohydrate active enzymes (CAZymes)

Getorf function of EMBOSS (v6.6.0.0) (Rice et al., 2000) was used to discover the ORFs (-find:1, and –maxsize: 1000). All the ORFs were analyzed by signalp (v5) (Almagro Armenteros et al., 2019) for the presence of signal peptide. Signal peptides were removed from predicted ORF sequences, and the cysteine amino acids were counted for every sequence. ORF sequence with three cysteine residues was predicted as a possible secreted protein (SP).

CAZymes are annotated during gene model annotation steps. In order to further classify these enzymes, total proteome of *I. obliquus* was queried against dbCAN2 database using DIAMOND (v0.9.24, blastp, --more-sensitive) (Buchfink et al., 2014; H. Zhang et al., 2018), and the best hit was selected (score $\geq$ 200, percentage identity $\geq$ 55) as a homologous sequence.

Gene tree of cytochrome P450 monooxygenases

A phylogenetic tree was constructed for 15 CYP716 gene models from Streptophyta species which have been confirmed to produce betulinate compounds [*Betula pendula* (Salojärvi et al., 2017), *Betula platyphylla*, *Phoenix dactylifera* (Al-Mssallem et al., 2013), *Medicago truncatula* (Tang et al., 2014), and *Vitis vinifera* (The French–Italian Public Consortium for Grapevine Genome et al., 2007)], after which CYP450 genes with high sequence similarity (< 40%) in *I. obliquus* and *F. betulina* were added to the tree. A total of 77 sequences were subjected to multiple sequence alignment (MSA)

262   using MUSCLE (v3.8.31, -maxiters: 1000) (Edgar, 2004). The amino acid sequences were reverse

263   translated by PAL2NAL (Suyama, Torrents, & Bork, 2006), and both amino acid and nucleotide

264   sequences were used to construct the phylogenetic trees by RaxmlHPC-HYBRID-AVX2 (v8.2.12)

265   (Stamatakis, 2014).

266   **Mass spectrometry, sample preparation and derivatization of *I. obliquus* and *F.***
267   ***mediterranea* metabolites**

268   Five strains of *I. obliquus* and one *F. mediterranea* (three biological replicates for each) were

269   grown in liquid Hagem media for two weeks. Submerged myceliums were washed with sterile milli-

270   Q water three times, and grinded with liquid nitrogen. All samples (fresh wight ~500mg, dry weight

271   ~30mg) were extracted twice with 1.0 ml of ethyl acetate (Merck) by vortexing for 15 min and

272   centrifuged for 10 min in 15000 rpm according to Cao et al. (G. Zhao, Yan, & Cao, 2007) at room

273   temperature. Internal standards (ISTD) (10 µl, 10 µg/ml), testosterone and 4-methylumbelliferone

274   (Sigma), were added to each sample in the first extraction step. The supernatant was evaporated to

275   dryness with MiVac Duo concentrator +40˚C (GeneVac LtD, Ipswich, UK) and the residue was re-

276   solubilized in 100 µl ACN (Honeywell). Quality control (QC) sample was prepared by combining

277   extracts from each sample line.

278   The triterpenoid profiling was executed from the extracts with UPLC-PDA-QTOF/MS. The UPLC-

279   MS system consisted of a Waters Acquity UPLC attached to a Acquity PDA-detector and to a Waters

280   Synapt G2 (HDMS) QTOF mass spectrometer (Waters, Milford, MA, USA). The separation of the

281   analytes was executed in Acquity BEH C18 (2.1mm x 50mm, 1.7µm) column (Waters, Milford, MA,

282   USA) with the temperature of +40˚C. The autosampler temperature was set to +27˚C. The mobile

283   phase consisted of water (A) and acetonitrile (B) both with 0.1% formic acid and the flow rate was

284   0.6 ml/min. The injection volume was 3 µl. The linear gradient started with 30% B and proceeded to

285   98% in 9 min, followed by 1 min at 98% B, giving a total run time of 10 min. ESI/MS detection was

286   performed in positive sensitivity ion mode with capillary voltage 3.0 kV, cone voltage 30 V,

287   desolvation gas 800 L/h, cone gas 20 L/h, desolvation temperature 320 °C, source temperature 120

288   °C and extractor lens 3.00 V. The MarkerLynx software (Waters, Milford, MA, USA) was used for data

289   processing. UV spectra, negative MS-runs and fragmentation patterns from MS[e] runs of QC sample

290   were used as additional tools for annotation of triterpenes, sterols and phenolic compounds in

291   *Inonotus obliquus* samples.

292   The standard solutions of betulin and betulinic acid (1.0 mg/ml) were prepared in ethyl acetate

293   and testesterone (100 µg/ml) standard was prepared in methanol. Working solutions (10 µg/ml, 100

294   ng/ml) were prepared by diluting the standard solution with acetonitrile. The optimization of

295   quantification method was executed with a mix of betulin, betulinic acid and testosterone standards

296 (100 ng/ml). Standard mix (100 µl) was derivatized with PTSI, and MS parameters were optimized

297 by repeated injection of the sample.

298   Due to very low concentration of betulin and betulinic acid, extracts were derivatized with p-

299 toluenesulfonyl isocyanate (PTSI) (Hu et al., 2013; Zuo, Gao, Liu, Cai, & Duan, 2005) to improve

300 sensitivity. After metabolite profiling with UPLC-QTOF/MS, samples (90 µl) were derivatized for 3

301 min with 10 µl of 60% p-toluenesulfonyl isocyanate (PTSI) (Sigma) in acetonitrile (Hu et al., 2013;

302 Zuo et al., 2005). The derivatization reaction was terminated with 50 µl of methanol (Merck) with

303 30 s vortex mixing, giving the total volume of 150 µl.

304   The UPLC-MS/MS system consisted of a UPLC (ABSciex, Shimadzu) attached to ABSciex 6500+

305 QTRAP mass spectrometer with ESI source. The Acquity BEH C18 (2.1mm x 50mm, 1.7µm) (Waters,

306 Milford, MA, USA) column was used for the separation of compounds, and column oven

307 temperature was +40˚C. The autosampler temperature was set to +25˚C. Injection volume was 2 µl.

308 The mobile phases were water (A) and acetonitrile (B) both with 0.1% of formic acid and the flow

309 rate was 0.6 ml/min. The linear gradient started from 30% B and proceeded to 98% in 6.5 min,

310 followed by 1.5 min at 98% B, giving a total run time of 8 min. The data was normalized to dry weight

311 (DW) and to the peak area of internal standard. The Analyst software (ABSciex) was used for data

312 processing and quantification.

313   PTSI derivatization reagent generated betulin p-toluenesulfonyl carbamic diester, betulinic acid

314 p-toluenesulfonyl carbamic ester, and testosterone toluenesulfonyl carbamic ester. Two MRM

315 transitions were selected for each analyte, one for quantification and other for qualification. The

316 ratio between quantification (quan) and qualification (qual) transitions should stay stable among

317 runs. The transitions were as follows: betulin MRM $835.3 \rightarrow 620.3$ quan [M-PTSI-H_2O-H]$^-$, $835.3 \rightarrow$

318 638.3 qual [M-PTSI-H]$^-$, betulinic acid MRM $652.3 \rightarrow 455.2$ quan [M-PTSI-H]$^-$ $652.3 \rightarrow 437.2$ qual [M-

319 PTSI-H_2O-H]$^-$, and testosterone $484.2 \rightarrow 287.2$ quan [PTSI-H]$^-$, $484.2 \rightarrow 269.2$ qual [PTSI-H_2O-H]$^-$. ESI

320 source temperature was set to 450 ˚C. ESI/MS/MS detection was performed in negative ion mode

321 with ion spray (IS) voltage of -4000, curtain gas (CUR) 30, collision gas (CAD) at medium, entrance

322 potential (EP) -10, declustering potential (DP) -60 (betulinic acid, testosterone) or -100 (betulin),

323 collision energy (CE) -50, collision cell exit potential (CXP) -10.

Cloning and mass spectrometry of lupeol synthase and CYP450 monooxygenase enzymes

325   Two major cloning constructs were designed for betulin biosynthesis using pRS424 vector

326 (Burgers, 1999). This version of pRS424 vector contained two multiple cloning sites (MCS), one under

327 GAL1 promoter and the second under GAL10. First, single constructs of CYP450 monooxygenase

328 enzymes were isolated from both *B. pendula* (pRS424::CYP716), as well as four homologous CYP450

329   monooxygenases from *I. obliquus* (pRS424::CYP450). The second construct was a double insertion

330   (pRS424::LUS-CYP716) of lupeol synthase (Bpev01.c0219.g0020.m0001 under GAL10 promoter) and

331   CYP450 monooxygenase (Bpev01.c0219.g0021.m0001, under GAL1 promoter) enzymes isolated

332   from *B. pendula* in pRS424 vector. All the vector constructs were transformed to yeast strain

333   (*Saccharomyces cerevisiae* [w303 background]). The transgenic yeasts were grown and induced

334   according to Zhou et al. (Zhou, Li, Li, & Zhang, 2016), with minor changes. SD-TRP was us as the drop

335   out medium. For single inserted vectors, we used 50um lupeol (dissolved in DMSO:EtOh [1:1]) in

336   induced growth media. After 60 hours of induction, the yeast growth media were centrifuged, and

337   both media and the cell pallets were collected and sent for mass spectrometry.

338   Total of eight samples (yeast cells (4 tubes), and cell culture media (4 tubes)) were analyzed with

339   UPLC-QTRAP/MS (MRM). Three triterpenoids (betulin [BE]) were extracted first from the media

340   twice with 1.0 ml ethyl acetate (Merck) for 60 min in RT and centrifuged for 5 min in 15 000 rpm

341   according to Cao et al. (G. Zhao et al., 2007). Testosterone was used as an internal standard (ISTD,

342   1.0 µl, 1.0 µg/ml). The cells were extracted in a similar manner as media, but yeast cells with 500 µl

343   H_2O and 1000 µl chloroform twice and were disrupted with freeze/thaw cycle (3 cycles) with ultra-

344   sonication (15min) prior to extraction procedure.

345   The upper ethyl acetate was evaporated to dryness with MiVac Duo concentrator +40˚C

346   (GeneVac Ltd., Ipswich, UK). The residue was re-solubilized in 100 µl ACN. Due to very low

347   concentration of lupeol, betulin and betulinic acid, extracts had to be derivatized with p-

348   toluenesulfonyl isocyanate (PTSI) (Hu et al., 2013; Zuo et al., 2005) to improve sensitivity. Samples

349   were derivatized in RT for 3 min with 10 µl of 60% p-toluenesulfonyl isocyanate (PTSI) (Sigma

350   Aldrich) in ACN (Hu et al., 2013; Zuo et al., 2005). The derivatization reaction was terminated with

351   90 µl of MeOH with 30 s vortex mixing, giving the total volume of 200µl. Immediately after PTSI-

352   derivatization, the MRM analysis of lupeol, betulin and betulinic acid was executed with UPLC-

353   QTRAP/MS (ABSciex).

354   The UPLC-MS/MS system consisted of ABSciex UPLC attached to ABSciex 6500+ QTRAP mass

355   spectrometer. The separation of the analytes column was Acquity BEH C18 (2.1mm x 50mm, 1.7µm)

356   (Waters, Milford, MA, USA), with the temperature of +40˚C. The autosampler temperature was set

357   to +25˚C. The injection volume was 10µl. The chromatographic conditions were executed as

358   described previously at Hua *et al.* (Hu et al., 2013). The mobile phase consisted of water with 0.1%

359   of formic acid in H_2O (A) and acetonitrile (B) with a flow rate of 0.6 ml/min. The linear gradient

360   started 30% B and proceeded to 98% in 6.5 min, left in 98% B for 2 min, and switched back to initial

361   conditions and left to stabilize, giving a total analysis time of 10 min.

362 ESI source temperature was set to 450˚C. ESI/MS/MS detection was performed in negative ion

363 mode with ion spray (IS) voltage of -4000, curtain gas (CUR) 30, collision gas (CAD) at medium,

364 entrance potential (EP) -10, de-clustering potential (DP) -60 (betulinic acid, testosterone) or -100

365 (betulin), collision energy (CE) -50, collision cell exit potential (CXP) -10. The Analyst software

366 (ABSciex) was used for data processing. PTSI derivatization reagent generated betulin p-

367 toluenesulfonyl carbamic diester (BTCD). The transitions for betulin and ISTD (testosterone) was as

368 follows: Betulin (BE) MRM 835.2 → 620.2 [M-PTSI-H_2O-H]$^-$, 835.2 → 638.3 [M-PTSI-H]$^-$ and 835.2 →

369 196.0 [PTSI-H]$^-$, and for testosterone MRM 484.2 → 287.2 [M-PTSI-H]$^-$ and MRM 484.2 → 269.2 [M-

370 PTSI-H_2O-H]$^-$. The most intense transitions of MRM 835.2 → 620.2 (betulin), and MRM 484.2 →

371 287.2 (ISTD, testosterone) were used.

372 ## Results and discussion

373 ### Nuclear and mitochondrial genome assemblies and annotations

374 Pacbio sequencing of *I. obliquus* strain from Merikarvia yielded 4.82 Gb of data (96x coverage)

375 with N50 read length of 9,200 bp. Falcon assembly resulted in a 41.1 million bp genome, consisting

376 of 301 primary contigs with an N50 value of 516 kilobases. Overall, the genome size of *I. obliquus*

377 was comparable with other species from *Hymenochaetales* (Supp. table 1).

378 Genome annotation of primary assembly yielded 13,778 gene models with 91.7% of universally

379 conserved single-copy genes being present (BUSCO v3.0, fungi database) (Waterhouse et al., 2017).

380 To support gene model prediction, RNA-seq was carried out from total RNA extracted from *I.

381 obliquus* reference strain sample grown on wood dust. Genome-guided *de novo* assembly of RNAseq

382 data showed 91.3% BUSCO completeness with 31% duplicated genes (Supp. table 1). Altogether

383 70.8% of *I. obliquus* genome consisted of coding sequence, with 53.1% in exons. Mean intron, exon,

384 CDS and gene lengths were 89, 284, 1447, and 2113 base pairs, respectively (Supp. table 1).

385 Mitochondrial genome was also assembled and annotated as described previously (Salojärvi et al.,

386 2017), resulting in 29 tRNAs, 32 coding sequences, and 3 rRNAs.

387 Transposable elements (TEs) have been suggested to play a major role in genome plasticity and

388 evolution. Thus, the classification and characterization of genes in close proximity of TEs are of

389 general interest, especially in the case of pathogenic organisms (Faino et al., 2016). In *I. obliquus*,

390 the total genome repeat content was found to be 26%, with 14.24% of repeats being unclassified.

391 The percentage of retrotransposon elements was 8.37%, and DNA transposon elements 1.2%. In

392 contrast to retrotransposon elements, *I. obliquus* genome contained higher amounts of DNA

393 transposon elements compared to the related *F. betulina* and *F. mediterranea* (Supp. table 1). Unlike

394 *F. mediterranea* (42.27% repeat sequences), TE content of *I. obliquus* did not fully explain its large

395 genome size (26% of repeat sequences) (Hage et al., 2021). Gene models flanking the upstream and

396 downstream of TEs contained mainly transposition elements, and gene clusters between two

397 transposable elements from the same DNA transposon class suggested the enrichment of gene

398 models involved in transmembrane transporter (GO:0003677), protein dimerization (GO:0046915),

399 transposition (GO:0046983 and GO:0006310), and DNA biding and recombination (GO:0006313 and

400 GO:0032196) (Supp. table 1) (Ali et al., 2014; Kang, Lebrun, Farrall, & Valent, 2001). The enrichment

401 of the last two categories suggests that some of the predicted gene models may be unidentified

402 transposable elements, and they are organized as clusters in the genome.

Secreted proteins in *I. obliquus*

404 The exact mechanisms of *I. obliquus* pathogenicity and its modes of interaction with the host are

405 not known, but characterization of secreted proteins is the first step to shed more light on the

406 mechanisms involving the initial penetration of plant defences by effector proteins. Secreted

407 proteins (SPs) are known for their essential role in pathogen-host interactions. Altogether 1052

408 open reading frames (ORFs), 7.6% of all gene models, were predicted as possible secreted proteins,

409 with minimal known homologs (Supp. table 2). The SPs were scattered across 128 contigs. Most of

410 the ORFs were likely species-specific, since homology searches with known secreted proteins from

411 other species were successful for only 110 of the ORFs (Supp. table 2). Twenty-one ORFs overlapped

412 at least with one class of TEs, having eighteen unclassified categories. Total of 988 ORFs were co-

413 localized between two TEs of the same TE class, suggesting a role for TEs in SP evolution and

414 diversification.

Carbohydrate-active enzymes

416 The palette of carbohydrate-active enzymes (CAZymes) present in the genome dictate to a large

417 extent the modes of substrate utilization by the fungus (Eastwood et al., 2011; Navarro et al., 2021).

418 We identified 466 candidate genes classified as carbohydrate-active enzymes (CAZymes), known for

419 their biological roles in anabolism and catabolism of different carbohydrates such as glycogen,

420 trehalose, and glycoconjugates (Supp. table 5). Altogether 211 enzymes were classified as glycoside

421 hydrolases (GHs) and 43 were categorized as carbohydrate binding modules (CBMs) whereas the

422 overall number of glycosyltransferases (GTs) and carbohydrate esterases (CEs) were found to be 110

423 and 23, respectively. Finally, 10 enzymes were assigned to polysaccharide lyases (PLs) (Supp. table

424 5). Overall the CAZyme palette was similar to other lignin-degrading fungi (Liu et al., 2019). RNAseq

425 analysis of *I. obliquus* grown on *B. pendula* wood dust and publicly available RNAseq data (Fradj et

426 al., 2019) identified in 214 (out of 466) CAZymes with positive Log_2FC (Supp. table 5) in at least one

427 of the experimental conditions. Majority of DE CAZymes belonged to glycoside hydrolases and

428 glycosyltransferases categories.

Phylogenomics and expanded gene families in *I. obliquus* genome

430 In order to estimate a taxonomic placement for *I. obliquus*, we collected the proteomes of fifteen

431 representative fungal species from different orders among *Basidiomycota*: all sequenced species

432 within *Hymenochaetales* (*Fomitiporia mediterranea, Inonotus obliquus, Onnia scaura, Phellinidium*

433 *ferrugineofuscum, Porodaedalea niemelaei, Trichaptum abietinum, Rickenella mellea, Schizopora*

434 *paradoxa),* and representatives of *Russulales (Heterobasidion annosum), Polyporales (Fomitopsis*

435 *betulina, Postia placenta, Phanerochaete chrysosporium)* and Agaricales (*Laccaria bicolor,*

436 *Coprinopsis cinerea, Schizophyllum commune)*. To root the taxonomy we added five outgroup

437 species, including two representatives of the other major classes in *Basidiomycota*: *Pucciniales*

438 *(Puccinia graminis)* and *Ustilaginales (Ustilago maydis*), as well as three model *Ascomycota* species

439 from *Saccharomycetales (Saccharomyces cerevisiae), Schizosaccharomycetales*

440 *(Schizosaccharomyces pombe),* and *Sordariales (Neurospora crassa)*. We next clustered the full

441 proteomes into gene families (orthogroups) using Orthofinder and identified single copy

442 orthogroups. Phylogeny estimation was carried out using 4040 single copy gene groups and rooted

443 to *Ascomycota* species. The resulting tree illustrates the known taxonomy among the 20 fungal

444 species (Figure 1; Supp. table 3) and the split of *Hymenochaetaceae* family occurs at the expected

445 phylogenetic position (Hibbett & Thorn, 2001; Matheny et al., 2007; R.-L. Zhao et al., 2017) (Figure

446 1). Interestingly *R. mellea* was placed together with the *Russulales* representative, further analysis

447 is however beyond the scope of the present work.

448 To look for gene family evolution we then identified gene families that were expanded in *I.*

449 *obliquus*. Altogether 167 orthologous gene clusters were significantly expanded in comparison to

450 the other nineteen fungal species (chi-squared test; Supp. table 4**Error! Reference source not**

451 **found.**). The expanded gene families were enriched for 23 GO terms such as terpene synthase

452 (GO:0010333), oxidoreductase (GO:0016684), and hydrolase (GO:0016788) activities. In addition,

453 GOs related to oxidative stress responses (GO:0006979), transposition (GO:0015074, GO:0006313),

454 and protein dimerization activity (GO:0046983) were also expanded (Supp. table 4); most of these

455 categories involve members of cytochrome P450 gene family.

Genome evolution in *I. obliquus*

457 The high-quality whole genome assembly allowed us to gain further insight into the gene family

458 evolution by synteny analyses using self-self alignments. The analysis suggested a recent whole

459   genome duplication (WGD) event in *I. obliquus*. Based on the synonymous mutation (Ks) spectrum

460   the event occurred after the split from *F. mediterranea* (approximately 112 million years ago during

461   the Triassic period; (Kumar, Stecher, Suleski, & Hedges, 2017)). Similarly, an independent lineage-

462   specific WGD was observed also in *P. niemelaei* (Figure 2).

463   Synteny analysis identified a total of 1,112 genes originating from the whole genome duplication

464   event, whereas a considerably higher amount, 6,200 genes, were identified in tandem duplications

465   (Supp. table 3). The tandemly duplicated genes reflect the shorter-term adaptation in the species,

466   and in general have been found to be associated with environmental responses (Panchy, Lehti-Shiu,

467   & Shiu, 2016). In *I. obliquus*, the tandemly duplicated genes were enriched for carbohydrate

468   biosynthesis, heme binding, oxidoreductase activity, tetrapyrrole binding, and DNA transposition. In

469   contrast, syntenic regions harbored genes related to biological pathways such as terpene synthesis

470   and cell cycle (Figure 3, Supp. table 6). The overlaps between tandemly duplicated, syntenic genes

471   and expanded gene families were significant (p-value=6.1094e-11, Fisher exact test) (Figure 4).

472   Altogether, the genome evolution analyses highlight the significance of tandem duplication events

473   in adaptation of the *I. obliquus* to different ecological niches and the central role of secondary

474   metabolism and particularly the expansion of CYP450 gene family by small scale duplication events

475   (Figure 5).

476   The members of CYP450 family have critical roles in fungal metabolism and adaptation to specific

477   ecological niches. Altogether 172 CYP450 monooxygenases were predicted in chaga, suggesting a

478   complex biochemical diversity in chaga metabolism. Division into clans and families revealed that

479   most of the enzymes belonged to clan CYP620 (69 members), followed by CYP4 and CYP512 clans

480   (Figure 5). CYP620 is shown to be involved in terpenoid synthases (Yap et al., 2014; Yu, Song, Liang,

481   Wang, & Lu, 2020). CYP4 is studied predominantly in phylum of *Arthropoda*, and shown to be

482   involved in biosynthesis of endogenous compounds (Zhu, Moural, Shah, & Palli, 2013). Finally,

483   CYP512 clan has been hypothesised to have catalytic activities towards steroidal-like compounds,

484   primarily testosterone (Ide, Ichinose, & Wariishi, 2012).

485   A high proportion CYP450 gene models, 79 out of the total of 172, were tandemly duplicated, and

486   many were members of CYP620 clan. GO enrichment analysis of two genes upstream and two genes

487   in downstream of all CYP450 monooxygenase enzymes suggested the enrichment of biological

488   functions related to oxidoreductase activity, heme binding, transmembrane transporter activities,

489   and tetrapyrrole binding (Supp. table 6). These results suggest tandem duplications of CYP450s, and

490   additionally the colocalization with cytochrome P450 reductase (CPR) partners, facilitates their

491   functional divergence (Ebrecht et al., 2019).

492   The observed colocalization of genes related to CYP450s suggested the presence of biosynthetic

493   clusters in the *I. obliquus* genome. We therefore sought biosynthetic gene clusters by antiSMASH

494   (Blin et al., 2013), identifying altogether 24 clusters in 17 contigs: 15 terpene synthase, 3 polyketide

495   synthase, and 4 non-ribosomal peptide synthetase clusters. The clusters were significantly enriched

496   for tandemly duplicated genes (Fisher exact test, p.value=6.34E-76), suggesting that tandem

497   duplications are a dominant process in their diversification. Furthermore, the clusters were enriched

498   for CYP450 gene family (p-value: 0.03161975), highlighting their central enzymatic role in secondary

499   metabolism (Supp. table 6).

Metabolomics fingerprinting of terpenoid compounds in five *I. obliquus* strains and *F. mediterranea*

502   Since chaga showed a significant expansion of CYP450 genes and a considerable number of

503   biosynthetic clusters, we next carried out metabolic fingerprinting of *I. obliquus* to study whether

504   the secondary metabolism was indeed diversified in chaga compared to *F. mediterranea*. Terpenoid

505   fingerprints in five strains of *I. obliquus* were distinctly different from *F. mediterranea* (Supp. table

506   1). Altogether the chaga strains showed 546 mass spectrum peaks, and only 135 of them were

507   shared with *F. mediterranea* (Figure 6 A). In addition, pairwise comparisons of each chaga strain and

508   *F. mediterranea* found 178 metabolomic features among *I. obliquus* strains with significantly higher

509   abundance (Supp. Fig 2, Supp. table 7). Many of the peaks were predicted to have molecular

510   formulae with 30, 31, and 28 carbon backbones, similar to lupeol, betulin and betulinic acid. Among

511   *I. obliquus* strains, Merikarvia had distinct metabolomic fingerprints and clustered more distant from

512   other strains (Figure 6 B, C) in principal coordinate analysis. This suggests that genotypic variation

513   plays a role in metabolic diversity (Figure 6 C). With regards to betulinate compounds, Merikarvia

514   strain had higher abundance of betulin and betulinic acid production in comparison to other strains

515   of *I. obliquus*. There were no peaks which resembled the standard for betulin or betulinic acid (98%,

516   Sigma-aldrich) in *F. mediterranea*, but a significant quantity of lupeol-like substance was discovered

517   (Supp. Fig 2).

Functional analysis of lupeol synthase and CYP450 monooxygenase

519   Even though metabolic fingerprinting does not identify the underlying metabolites, the analysis

520   suggested the presence of terpene and lupeol as well as betulin derivatives based on the predicted

521   carbon backbones. Intraspecies quantification of betulin and betulinic acid (Using HPL) among six

522   species of betula and three strains of chaga showed a higher concentration of betulinic acid

523   compared to betulin in chaga strains. In contrast, the opposite result was observed in six species of

524 Betula, where the concentration of betulin was consistently higher in comparison to betulinic acid

525 (Figure 7).

526 The high quality gene model predictions allowed us to look for the candidate enzymes

527 responsible for the betulin biosynthesis in chaga. Since no members of CYP716 family were not

528 predicted in chaga we identified four best candidates based on homology analysis of CYP450s to

529 known CYP716 family members from plant species, such as *B. pendula*. Yeast expression system has

530 been used successfully for cloning CYP450 monooxygenase enzyme from *B. platyphylla* (Zhou et al.,

531 2016). To functionally validate our candidate genes, we first constructed single insert expression

532 vector for birch CYP716 enzyme (pRS424::CYP716). In addition, we also constructed a double insert

533 vector where lupeol synthase and lupeol monooxygenase from *B. pendula* (Safronov et al., 2019)

534 were inserted into two multiple cloning sites of a vector (pRS424::LUS-CYP716), which was then

535 transferred into yeast expression system. Similar to single insert vector for CYP450 monooxygenase

536 enzyme from *B. pendula*, we isolated four CYP450s from *I. obliquus* and cloned them to engineer

537 single inserted constructs. The single constructs were grown in media which was spiked with

538 standard lupeol compound (98%, Cayman) as the precursor. In contrast to single insert vectors,

539 double insert vector from *B. pendula* (pRS424::LUS-CYP716) expressed lower concentration of

540 betulin compared to single insert vector (pRS424::CYP716) from *B. pendula*. These differences might

541 be explained by the lower initial amount of available precursor compound, lupeol, for CYP716

542 monooxygenase (Figure 8). In addition to *B. pendula* constructs, all four candidate CYP450

543 monooxygenases from *I. obliquus* (pRS424::CYP450) showed some degree of betulin production

544 when compared to standard betulin (98%, Sigma-Aldrich) spectrum. Among the four candidates, the

545 enzyme with gene ID c000016F_g277 (clan CYP505) had the highest amount of betulin production.

546 Interestingly, the cDNA length of this enzyme was 3297 bp, almost twice the length of the other

547 three candidate CYP450 monooxygenase homologs from *I. obliquus*. Upon close examination the

548 amino acid and nucleotide sequences of c000016F_g277 resemble a chimeric isoform of two CYP450

549 monooxygenase enzymes (Figure 8). In general, our study showed that yeast cell fractions contained

550 higher concentration of betulin compared to the culture media fractions (Figure 8), thus confirming

551 the function of the inserted enzymes.

552 To study the evolution of potential homologs and orthologs for the four candidate genes from *I.*

553 *obliquus*, we carried out microsynteny analysis for the cloned CYP450 monooxygenase enzymes

554 against four *Hymenochaetales* and *F. betulina* species. Orthologous one-to-one relationship with

555 other fungal species was confirmed for c000112F.g25 (clan CYP51) and c000041F.g53 (clan CYP51)

556 enzymes (Supp. Fig 3-C,D), whereas microsynteny analysis of c000000F.g253 (clan CYP61) enzyme

557 found a cluster of homologous genes in 5' and 3' of the c000000F.g253 in *I. obliquus* genome (Supp.

Fig 3-B). The microsynteny of the chimeric c000016F.g277 linked to a putative ortholog in *F. mediterranea* with similar organization (gene_7933, clan CYP505) (Supp. Fig 3**Error! Reference source not found.**-A), whereas in other *Hymenochaetales* the syntenic analysis identified two separate CYP450s. This suggests that the fusion gene has arisen from a non-homologous recombination event in the common ancestor of *F. mediterranea* and *I. obliquus*, and after divergence of *P. niemelaei* where the CYP450s were still found separate. Both c000016F.g277 and gene_7933 contain two heme-binding domains, but gene_7933 from *F. mediterranea* has three oxygen-binding domains (with AGADTT/GGDDTG motifs) instead of two in *I. obliquus* (AGADTT). Therefore the fusion may have occurred also independently in chaga and *F. mediterranea* (Supp. table 8), or then involved a loss of the third oxygen-binding domain in chaga.

Evolution of conserved domains and phylogeny reconstruction of cytochrome P450 monooxygenase

Since betulin and betulinic acid are antifungal substances and they are produced in the main natural host of *I. obliquus*, it is possible that the enzymes have been introduced into chaga or its ancestor via horizontal gene transfer, either directly from the host species or then via another species cohabiting with chaga. However, a phylogenetic tree of a set of monooxygenase enzymes (77 enzymes) from *I. obliquus* and *F. betulina* with sequence similarity to plant CYP716, as well as CYP716 enzymes of eight plant species known to produce betulinate compounds, shows a distinct divergence of fungal clades from the plant species. This result is consistent both in protein and DNA based phylogenetic trees, providing no evidence of gene transfer events (Figure 9).

Gene expression analysis

Total of 119 (out of 172) monooxygenase enzymes were significantly expressed with positive log_2fold change (log_2FC) values in at least one of the DEg comparisons (Figure 5) and three key enzymes involved in mevalonate pathways were among this set (**Error! Reference source not found.**). We also observed a pair of tandemly duplicated lupeol synthase enzymes to have the highest expression levels. The expression profiles of the chaga samples grown on *B. pendula* wood dust were stronger than the samples grown in culture media from (Fradj et al., 2019). When inspecting the expression profiles for genes with positive log_2FC , enrichments were found for WD40-repeat binding (50 genes out of 294), melanin biosynthesis (65 genes out of 157), aquaporin (20 genes), lipases and peptidases (Supp. table 9).

Conclusion

588

589   We observed genome evolution leading towards complex terpene biosynthesis in *I. obliquus*,

590   both in genes originating from whole genome duplication events as well as tandem duplications

591   within the CYP450 gene family. It is possible that the whole genome duplication event is associated

592   with the initial expansion of terpenoid biosynthesis capacity in *I. obliquus*, since no such expansion

593   was observed in the related species *F. mediterranea*. In contrast to eg plants, the number of whole

594   genome duplication events in fungal kingdom has been low (Albertin & Marullo, 2012), but this may

595   be due to faster genome evolution in fungi, making the WGDs difficult to identify (Campbell, Ganley,

596   Gabaldón, & Cox, 2016). The CYP450 superfamily is associated with many reactions in secondary

597   metabolism, and through metabolomics fingerprinting we confirmed that the fungus indeed

598   produces a rich palette of terpenoid derivatives. However, we found no evidence of a horizontal

599   gene transfer event between *B. pendula* and *I. obliquus*, and the identified candidate lupeol

600   monooxygenases in *I. obliquus* were members of a different CYP505 clan with low sequence

601   similarity to their birch counterpart enzymes. Therefore CYP450 monooxygenases enzymes

602   responsible for betulinate biosynthesis in the two species most likely result from convergent

603   evolution.

Author's contributions

604

605   O.S and J.S conceived and designed the project. Funding acquisition is carried by J.S and J.K. O.S

606   collected the DNA and RNA samples. O.S and J.S managed and coordinated all bioinformatics

607   activities. O-P.S, L.G.P, and P.A did RNA and DNA library construction and sequencing and

608   participated in genome assembly. O.S did the genome and functional annotation. S.R and P.S

609   participated in genome annotation. O.S analyzed the RNA sequencing data including de novo

610   assembly of RNAseq. O.S did comparative genomics analyses. T.S and N.S were involved in field

611   research for sample isolations. O.S and M.W grown and collected the samples for mass

612   spectrometry. G.L.B, B.B, M.W, and O.S were sequenced were involved in cloning and expression of

613   CYP450 and Lupeol synthase enzymes. N.S and J.L did mass spectrometry, including sample

614   pretreatment, method development, UPLC-HDMS analysis, metabolite identification and data

615   interpretation, and O.S and J.S contributed to data interpretation. O.S and J.S wrote the original

616   manuscript with input from O-P.S , U.R, K.O.

617

624

References

626 Al-Mssallem, I. S., Hu, S., Zhang, X., Lin, Q., Liu, W., Tan, J., . . . Yu, J. (2013). Genome sequence of
627   the date palm *Phoenix dactylifera* L. from Nat Commun
628 Albertin, W., & Marullo, P. (2012). Polyploidy in fungi: evolution after whole-genome duplication.
629   *Proceedings of the Royal Society B: Biological Sciences, 279*(1738), 2497-2509.
630   doi:10.1098/rspb.2012.0434
631 Ali, S., Laurie, J. D., Linning, R., Cervantes-Chavez, J. A., Gaudet, D., & Bakkeren, G. (2014). An
632   immunity-triggering effector from the Barley smut fungus Ustilago hordei resides in an
633   Ustilaginaceae-specific cluster bearing signs of transposable element-assisted evolution.
634   *PLoS Pathog, 10*(7), e1004223. doi:10.1371/journal.ppat.1004223
635 Almagro Armenteros, J. J., Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S., .
636   . . Nielsen, H. (2019). SignalP 5.0 improves signal peptide predictions using deep neural
637   networks. *Nature Biotechnology, 37*(4), 420-423. doi:10.1038/s41587-019-0036-z
638 Alresly, Z., Lindequist, U., Lalk, M., Porzel, A., Arnold, N., & Wessjohann, L. A. (2015). Bioactive
639   Triterpenes from the Fungus *Piptoporus betulinus*. *Rec Nat Prod, 10*, 103-108.
640 Andre, C. M., Larsen, L., Burgess, E. J., Jensen, D. J., Cooney, J. M., Evers, D., . . . Laing, W. A. (2013).
641   Unusual immuno-modulatory triterpene-caffeates in the skins of russeted varieties of
642   apples and pears. *J Agric Food Chem, 61*(11), 2773-2779. doi:10.1021/jf305190e
643 Andrews, S. FastQC A Quality Control tool for High Throughput Sequence Data.
644   *http://www.bioinformatics.babraham.ac.uk/projects/fastqc/*. Retrieved from
645   http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
646 Besemer, J., & Borodovsky, M. (2005). GeneMark: web software for gene finding in prokaryotes,
647   eukaryotes and viruses. *Nuc acids res, 33*. doi:10.1093/nar/gki487
648 Blanchette, R. A. (1982). Progressive stages of discolortion and decay associated with the canker-
649   rot fungus, Inonotus obliquus, in birch. *Phytopathology., 72*(10), 1272-1277.
650   doi:10.1094/phyto-72-1272
651 Blin, K., Medema, M. H., Kazempour, D., Fischbach, M. A., Breitling, R., Takano, E., & Weber, T.
652   (2013). antiSMASH 2.0--a versatile platform for genome mining of secondary metabolite
653   producers. *Nucleic Acids Res, 41*(Web Server issue), W204-212. doi:10.1093/nar/gkt449
654 Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina
655   sequence data. *Bioinformatics, 30*(15), 2114-2120. doi:10.1093/bioinformatics/btu170
656 Bonfield, J. K., Smith, K., & Staden, R. (1995). A new DNA sequence assembly program. *Nucleic
657   Acids Res, 23*(24), 4992-4999.
658 Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq
659   quantification. *Nat Biotechnol, 34*(5), 525-527. doi:10.1038/nbt.3519
660 Buchfink, B., Xie, C., & Huson, D. H. (2014). Fast and sensitive protein alignment using DIAMOND.
661   *Nat Meth, 12*, 59. doi:10.1038/nmeth.3176
662 https://www.nature.com/articles/nmeth.3176#supplementary-information
663 Burgers, P. M. (1999). Overexpression of multisubunit replication factors in yeast. *Methods, 18*(3),
664   349-355. doi:10.1006/meth.1999.0796
665 Campbell, M. A., Ganley, A. R. D., Gabaldón, T., & Cox, M. P. (2016). The Case of the Missing
666   Ancient Fungal Polyploids. *The American Naturalist, 188*(6), 602-614. doi:10.1086/688763
667 Chang, S., Puryear, J., & Cairney, J. (1993). A Simple and Efficient Method for Isolating RNA from
668   Pine Trees. *Plant Mol Biol Rep, 11*(2), 113-116. doi:10.1007/BF02670468

669   Chin, C. S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., . . . Schatz, M. C.
670   (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nat*
671   *Methods, 13*(12), 1050-1054. doi:10.1038/nmeth.4035
672   Eastwood, D. C., Floudas, D., Binder, M., Majcherczyk, A., Schneider, P., Aerts, A., . . . Watkinson
673   Sarah, C. (2011). The Plant Cell Wall–Decomposing Machinery Underlies the Functional
674   Diversity of Forest Fungi. *Science, 333*(6043), 762-765. doi:10.1126/science.1205411
675   Ebrecht, A. C., van der Bergh, N., Harrison, S. T. L., Smit, M. S., Sewell, B. T., & Opperman, D. J.
676   (2019). Biochemical and structural insights into the cytochrome P450 reductase from
677   Candida tropicalis. *Scientific Reports, 9*(1), 20088. doi:10.1038/s41598-019-56516-6
678   Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high
679   throughput. *Nucleic Acids Res, 32*(5), 1792-1797. doi:10.1093/nar/gkh340
680   Emms, D. M., & Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome
681   comparisons dramatically improves orthogroup inference accuracy. *Genome Biology, 16*(1),
682   157. doi:10.1186/s13059-015-0721-2
683   Faino, L., Seidl, M. F., Shi-Kunne, X., Pauper, M., van den Berg, G. C., Wittenberg, A. H., & Thomma,
684   B. P. (2016). Transposons passively and actively contribute to evolution of the two-speed
685   genome of a fungal pathogen. *Genome Res, 26*(8), 1091-1100. doi:10.1101/gr.204974.116
686   Fradj, N., Goncalves Dos Santos, K. C., de Montigny, N., Awwad, F., Boumghar, Y., Germain, H., &
687   Desgagne-Penix, I. (2019). RNA-Seq de Novo Assembly and Differential Transcriptome
688   Analysis of Chaga (Inonotus obliquus) Cultured with Different Betulin Sources and the
689   Regulation of Genes Involved in Terpenoid Biosynthesis. *Int J Mol Sci, 20*(18).
690   doi:10.3390/ijms20184334
691   Fukushima, E. O., Seki, H., Ohyama, K., Ono, E., Umemoto, N., Mizutani, M., . . . Muranaka, T.
692   (2011). CYP716A subfamily members are multifunctional oxidases in triterpenoid
693   biosynthesis. *Plant Cell Physiol, 52*(12), 2050-2061. doi:10.1093/pcp/pcr146
694   Godzik, A., & Li, W. (2006). Cd-hit: a fast program for clustering and comparing large sets of
695   protein or nucleotide sequences. *Bioinformatics, 22*(13), 1658-1659.
696   doi:10.1093/bioinformatics/btl158
697   Gong, Y., Raj, K. M., Luscombe, C. A., Gadawski, I., Tam, T., Chu, J., . . . Sacks, S. L. (2004). The
698   synergistic effects of betulin with acyclovir against herpes simplex viruses. *Antiviral Res,*
699   *64*(2), 127-130. doi:10.1016/j.antiviral.2004.05.006
700   Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., . . . Regev, A. (2011).
701   Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data.
702   *Nat Biotechnol, 29*(7), 644-652. doi:10.1038/nbt.1883
703   Gremme, G., Steinbiss, S., & Kurtz, S. (2013). GenomeTools: a comprehensive software library for
704   efficient processing of structured genome annotations. *IEEE/ACM Trans Comput Biol*
705   *Bioinform, 10*(3), 645-656. doi:10.1109/TCBB.2013.68
706   Haas, B. J., Delcher, A. L., Wortman, J. R., & Salzberg, S. L. (2004). DAGchainer: a tool for mining
707   segmental genome duplications and synteny. *Bioinformatics, 20*(18), 3643-3646.
708   doi:10.1093/bioinformatics/bth397
709   Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., . . . Wortman, J. R. (2008).
710   Automated eukaryotic gene structure annotation using EVidenceModeler and the Program
711   to Assemble Spliced Alignments. *Genome Biol, 9*(1), R7-R7. doi:10.1186/gb-2008-9-1-r7
712   Hage, H., Miyauchi, S., Viragh, M., Drula, E., Min, B., Chaduli, D., . . . Rosso, M. N. (2021). Gene
713   family expansions and transcriptome signatures uncover fungal adaptations to wood
714   decay. *Environ Microbiol*. doi:10.1111/1462-2920.15423
715   Hibbett, D. S., & Thorn, R. G. (2001). Basidiomycota: Homobasidiomycetes. In D. J. McLaughlin, E.
716   G. McLaughlin, & P. A. Lemke (Eds.), *Systematics and Evolution* (pp. 121-168). Berlin,
717   Heidelberg: Springer Berlin Heidelberg.

718   Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M., & Stanke, M. (2015). BRAKER1: Unsupervised
719   RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics,*
720   *32*(5), 767-769. doi:10.1093/bioinformatics/btv661

721   Holonec, L., Ranga, F., Crainic, D., Truța, A., & Socaciu, C. (2012). Evaluation of Betulin and
722   Betulinic Acid Content in Birch Bark from Different Forestry Areas of Western Carpathians.
723   *Notulae Botanicae Horti Agrobotanici Cluj-Napoca, 40.* doi:10.15835/nbha4027967

724   Hu, Z., Guo, N., Wang, Z., Liu, Y., Wang, Y., Ding, W., . . . Yan, X. (2013). Development and
725   validation of an LC-ESI/MS/MS method with precolumn derivatization for the
726   determination of betulin in rat plasma. *J Chromatogr B Analyt Technol Biomed Life Sci, 939*,
727   38-44. doi:10.1016/j.jchromb.2013.09.005

728   Ide, M., Ichinose, H., & Wariishi, H. (2012). Molecular identification and functional characterization
729   of cytochrome P450 monooxygenases from the brown-rot basidiomycete Postia placenta.
730   *Archives of Microbiology, 194*(4), 243-253. doi:10.1007/s00203-011-0753-2

731   Kang, S., Lebrun, M. H., Farrall, L., & Valent, B. (2001). Gain of virulence caused by insertion of a
732   Pot3 transposon in a Magnaporthe grisea avirulence gene. *Mol Plant Microbe Interact,*
733   *14*(5), 671-674. doi:10.1094/MPMI.2001.14.5.671

734   Khelil, R., Jardé, E., Cabello-Hurtado, F., Ould-el-Hadj Khelil, A., & Esnault, M.-A. (2016). Structure
735   and composition of the wax of the date palm, Phoenix dactylifera L., from the septentrional
736   Sahara. *Scientia Horticulturae, 201*, 238-246.
737   doi:https://doi.org/10.1016/j.scienta.2016.02.012

738   Khouloud Barakat, M. S. (2016). Bioactive Betulin produced by marine Paecilomyces WE3-F. *J Appl*
739   *Pharm Sci*(Volume: 6, Issue: 3), 034-040. Retrieved from
740   http://japsonline.com/abstract.php?article_id=1799

741   Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: accurate
742   alignment of transcriptomes in the presence of insertions, deletions and gene fusions.
743   *Genome Biol, 14*(4), R36. doi:10.1186/gb-2013-14-4-r36

744   Koolen, H. H. F., Soares, E. R., Silva, F. M. A. d., Souza, A. Q. L. d., Rodrigues Filho, E., & Souza, A. D.
745   L. d. (2012). Triterpenes and flavonoids from the roots of Mauritia flexuosa. *Rev Bras*
746   *Farmacogn, 22*, 189-192. Retrieved from
747   http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-
748   695X2012000100028&nrm=iso

749   Król, S. K., Kiełbus, M., Rivero-Müller, A., & Stepulak, A. (2015). Comprehensive Review on Betulin
750   as a Potent Anticancer Agent. *BioMed Research International, 2015*, 11.
751   doi:10.1155/2015/584189

752   Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). TimeTree: A Resource for Timelines,
753   Timetrees, and Divergence Times. *Mol Biol Evol, 34*(7), 1812-1819.
754   doi:10.1093/molbev/msx116

755   Lepesheva, G. I., Hargrove, T. Y., Kleshchenko, Y., Nes, W. D., Villalta, F., & Waterman, M. R.
756   (2008). CYP51: A major drug target in the cytochrome P450 superfamily. *Lipids, 43*(12),
757   1117-1125. doi:10.1007/s11745-008-3225-y

758   Liu, Y., Wu, Y., Zhang, Y., Yang, X., Yang, E., Xu, H., . . . Yan, J. (2019). Lignin degradation potential
759   and draft genome sequence of Trametes trogii S0301. *Biotechnol Biofuels, 12*, 256.
760   doi:10.1186/s13068-019-1596-3

761   Lodhi, M., Ye, G.-N., Weeden, N., & I. Reisch, B. (1994). A simple and efficient method for DNA
762   extraction from grapevine cultivars andVitis species. *Plant Mol Biol Rep, 12*(1), 6-13.
763   doi:10.1007/BF02668658

764   Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion
765   for RNA-seq data with DESeq2. *Genome Biol, 15*(12), 550. doi:10.1186/s13059-014-0550-8

766 Ma, L., Chen, H., Dong, P., & Lu, X. (2013). Anti-inflammatory and anticancer activities of extracts
767   and compounds from the mushroom Inonotus obliquus. *Food Chem, 139*(1-4), 503-508.
768   doi:10.1016/j.foodchem.2013.01.030
769 Matheny, P. B., Wang, Z., Binder, M., Curtis, J. M., Lim, Y. W., Nilsson, R. H., . . . Hibbett, D. S.
770   (2007). Contributions of rpb2 and tef1 to the phylogeny of mushrooms and allies
771   (Basidiomycota, Fungi). *Mol Phylogenet Evol, 43*(2), 430-451.
772   doi:10.1016/j.ympev.2006.08.024
773 Miettinen, K., Pollier, J., Buyst, D., Arendt, P., Csuk, R., Sommerwerk, S., . . . Goossens, A. (2017).
774   The ancient CYP716 family is a major contributor to the diversification of eudicot
775   triterpenoid biosynthesis. *Nat Commun, 8*, 14153.
776 Nagajyothi, P. C., Sreekanth, T. V., Lee, J. I., & Lee, K. D. (2014). Mycosynthesis: antibacterial,
777   antioxidant and antiproliferative activities of silver nanoparticles synthesized from
778   Inonotus obliquus (Chaga mushroom) extract. *J Photochem Photobiol B, 130*, 299-304.
779   doi:10.1016/j.jphotobiol.2013.11.022
780 Navarro, D., Chaduli, D., Taussac, S., Lesage-Meessen, L., Grisel, S., Haon, M., . . . Favel, A. (2021).
781   Large-scale phenotyping of 1,000 fungal strains for the degradation of non-natural,
782   industrial compounds. *Communications Biology, 4*(1), 871. doi:10.1038/s42003-021-02401-
783   w
784 P. Kovalenko, L., Shipaeva, E., V. Balakshin, V., A. Presnova, G., N. Chistyakov, A., Klodt, P., . . .
785   Durnev, A. (2009). Antiallergenic activity of birch bark dry extract with at least 70% betulin
786   content. *Pharm Chem J, 43*, 110-114. doi:10.1007/s11094-009-0242-y
787 Panchy, N., Lehti-Shiu, M., & Shiu, S.-H. (2016). Evolution of Gene Duplication in Plants. *Plant
788   Physiology, 171*(4), 2294-2316. doi:10.1104/pp.16.00523
789 Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., & Lopez, R. (2005).
790   InterProScan: protein domains identifier. *Nucleic Acids Res, 33*(Web Server issue), W116-
791   120. doi:10.1093/nar/gki442
792 Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open
793   Software Suite. *Trends Genet, 16*(6), 276-277. Retrieved from
794   https://www.ncbi.nlm.nih.gov/pubmed/10827456
795 Ryvarden, L., & Gilbertson, R. L. (1993). *European polypores. 1 : Abortiporus-Lindtneria*: Oslo :
796   Fungiflora.
797 Safronov, O., Alonso-Serra, J., Lim, K. J., Fraser-Miller, S. J., Blokhina, O. B., Campilho, A., . . .
798   Salojarvi, J. (2019). Tissue-specific study across the stem reveals the chemistry and
799   transcriptome dynamics of birch bark. *New Phytol, 222*(4), 1816-1831.
800   doi:10.1111/nph.15725
801 Salin, O., Alakurtti, S., Pohjala, L., Siiskonen, A., Maass, V., Maass, M., . . . Vuorela, P. (2010).
802   Inhibitory effect of the natural product betulin and its derivatives against the intracellular
803   bacterium *Chlamydia pneumoniae. Biochem Pharmacol, 80*(8), 1141-1151.
804   doi:10.1016/j.bcp.2010.06.051
805 Salojärvi, J., Smolander, O.-P., Nieminen, K., Rajaraman, S., Safronov, O., Safdari, P., . . .
806   Kangasjärvi, J. (2017). Genome sequencing and population genomic analyses provide
807   insights into the adaptive landscape of silver birch. *Nat Genet, 49*, 904.
808   doi:10.1038/ng.3862
809 Schlapfer, P., Zhang, P., Wang, C., Kim, T., Banf, M., Chae, L., . . . Rhee, S. Y. (2017). Genome-Wide
810   Prediction of Metabolic Enzymes, Pathways, and Gene Clusters in Plants. *Plant Physiol,
811   173*(4), 2041-2059. doi:10.1104/pp.16.01942
812 Sezutsu, H., Le Goff, G., & Feyereisen, R. (2013). Origins of P450 diversity. *Philosophical
813   transactions of the Royal Society of London. Series B, Biological sciences, 368*(1612),
814   20120428-20120428. doi:10.1098/rstb.2012.0428

815   Shai, L. J., McGaw, L. J., Aderogba, M. A., Mdee, L. K., & Eloff, J. N. (2008). Four pentacyclic
816   triterpenoids with antifungal and antibacterial activity from *Curtisia dentata* (Burm.f) C.A.
817   Sm. leaves. *J Ethnopharmacol, 119*(2), 238-244. doi:10.1016/j.jep.2008.06.036

818   Siddiqui, S. A., Rahman, A., Rahman, M. O., Akbar, M. A., Ali, M. A., Al-Hemaid, F. M. A., . . . Farah,
819   M. A. (2019). A novel triterpenoid 16-hydroxy betulinic acid isolated from Mikania cordata
820   attributes multi-faced pharmacological activities. *Saudi J Biol Sci, 26*(3), 554-562.
821   doi:10.1016/j.sjbs.2018.03.002

822   Šiman, P., Filipová, A., Tichá, A., Niang, M., Bezrouk, A., & Havelek, R. (2016). Effective Method of
823   Purification of Betulin from Birch Bark: The Importance of Its Purity for Scientific and
824   Medicinal Use. *PLOS ONE, 11*(5), e0154933. doi:10.1371/journal.pone.0154933

825   Slater, G. S. C., & Birney, E. (2005). Automated generation of heuristics for biological sequence
826   comparison. *BMC Bioinformatics, 6*(1), 31. doi:10.1186/1471-2105-6-31

827   Soneson, C., Love, M. I., & Robinson, M. D. (2015). Differential analyses for RNA-seq: transcript-
828   level estimates improve gene-level inferences. *F1000Res, 4*, 1521.
829   doi:10.12688/f1000research.7563.2

830   Song, F. Q., Liu, Y., Kong, X. S., Chang, W., & Song, G. (2013). Progress on understanding the
831   anticancer mechanisms of medicinal mushroom: inonotus obliquus. *Asian Pac J Cancer
832   Prev, 14*(3), 1571-1578. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/23679238

833   Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
834   phylogenies. *Bioinformatics, 30*(9), 1312-1313. doi:10.1093/bioinformatics/btu033

835   Stanke, M., & Morgenstern, B. (2005). AUGUSTUS: a web server for gene prediction in eukaryotes
836   that allows user-defined constraints. *Nuc acids res, 33*. doi:10.1093/nar/gki458

837   Suyama, M., Torrents, D., & Bork, P. (2006). PAL2NAL: robust conversion of protein sequence
838   alignments into the corresponding codon alignments. *Nucleic Acids Res, 34*(Web Server
839   issue), W609-612. doi:10.1093/nar/gkl315

840   Tang, H., Krishnakumar, V., Bidwell, S., Rosen, B., Chan, A., Zhou, S., . . . Town, C. D. (2014). An
841   improved genome release (version Mt4.0) for the model legume Medicago truncatula.
842   *BMC Genomics, 15*, 312. doi:10.1186/1471-2164-15-312

843   The French–Italian Public Consortium for Grapevine Genome, C., Jaillon, O., Aury, J.-M., Noel, B.,
844   Policriti, A., Clepet, C., . . . Wincker, P. (2007). The grapevine genome sequence suggests
845   ancestral hexaploidization in major angiosperm phyla. *Nature, 449*, 463.
846   doi:10.1038/nature06148
847   https://www.nature.com/articles/nature06148#supplementary-information

848   Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., . . .
849   Zdobnov, E. M. J. M. B. E. (2017). BUSCO applications from quality assessments to gene
850   prediction and phylogenomics. doi:10.1093/molbev/msx319

851   Wu, J., Niu, Y., Bakur, A., Li, H., & Chen, Q. (2017). Cell-Free Production of Pentacyclic Triterpenoid
852   Compound Betulinic Acid from Betulin by the Engineered Saccharomyces cerevisiae.
853   *Molecules, 22*(7). doi:10.3390/molecules22071075

854   Yan, Z. F., Yang, Y., Tian, F. H., Mao, X. X., Li, Y., & Li, C. T. (2014). Inhibitory and Acceleratory
855   Effects of Inonotus obliquus on Tyrosinase Activity and Melanin Formation in B16
856   Melanoma Cells. *Evid Based Complement Alternat Med, 2014*, 259836.
857   doi:10.1155/2014/259836

858   Yap, H.-Y. Y., Chooi, Y.-H., Firdaus-Raih, M., Fung, S.-Y., Ng, S.-T., Tan, C.-S., & Tan, N.-H. (2014).
859   The genome of the Tiger Milk mushroom, Lignosus rhinocerotis, provides insights into the
860   genetic basis of its medicinal properties. *BMC Genomics, 15*(1), 635. doi:10.1186/1471-
861   2164-15-635

862 Yin, Y., Cui, Y., & Ding, H. (2008). Optimization of betulin extraction process from Inonotus
863   Obliquus with pulsed electric fields. *Innov Food Sci Emerg Technol, 9*(3), 306-310.
864   doi:https://doi.org/10.1016/j.ifset.2007.07.010
865 Yu, F., Song, J., Liang, J., Wang, S., & Lu, J. (2020). Whole genome sequencing and genome
866   annotation of the wild edible mushroom, Russula griseocarnosa. *Genomics, 112*(1), 603-
867   614. doi:https://doi.org/10.1016/j.ygeno.2019.04.012
868 Zhang, H., Yohe, T., Huang, L., Entwistle, S., Wu, P., Yang, Z., . . . Yin, Y. (2018). dbCAN2: a meta
869   server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res, 46*(W1),
870   W95-W101. doi:10.1093/nar/gky418
871 Zhang, H. J., Tan, G. T., Hoang, V. D., Hung, N. V., Cuong, N. M., Soejarto, D. D., . . . Fong, H. H.
872   (2003). Natural anti-HIV agents. Part IV. Anti-HIV constituents from Vatica cinerea. *J Nat
873   Prod, 66*(2), 263-268. doi:10.1021/np020379y
874 Zhao, G., Yan, W., & Cao, D. (2007). Simultaneous determination of betulin and betulinic acid in
875   white birch bark using RP-HPLC. *J Pharm Biomed Anal, 43*(3), 959-962.
876   doi:10.1016/j.jpba.2006.09.026
877 Zhao, R.-L., Li, G.-J., Sánchez-Ramírez, S., Stata, M., Yang, Z.-L., Wu, G., . . . Hyde, K. D. (2017). A six-
878   gene phylogenetic overview of Basidiomycota and allied phyla with estimated divergence
879   times of higher taxa and a phyloproteomics perspective. *Fungal Diversity, 84*(1), 43-74.
880   doi:10.1007/s13225-017-0381-5
881 Zhao, S., Park, C. H., Li, X., Kim, Y. B., Yang, J., Sung, G. B., . . . Park, S. U. (2015). Accumulation of
882   Rutin and Betulinic Acid and Expression of Phenylpropanoid and Triterpenoid Biosynthetic
883   Genes in Mulberry (Morus alba L.). *J Agric Food Chem, 63*(38), 8622-8630.
884   doi:10.1021/acs.jafc.5b03221
885 Zhou, C., Li, J., Li, C., & Zhang, Y. (2016). Improvement of betulinic acid biosynthesis in yeast
886   employing multiple strategies. *BMC Biotechnology, 16*(1), 59. doi:10.1186/s12896-016-
887   0290-9
888 Zhu, F., Moural, T. W., Shah, K., & Palli, S. R. (2013). Integrated analysis of cytochrome P450 gene
889   superfamily in the red flour beetle, Tribolium castaneum. *BMC Genomics, 14*(1), 174.
890   doi:10.1186/1471-2164-14-174
891 Zuo, M., Gao, M. J., Liu, Z., Cai, L., & Duan, G. L. (2005). p-Toluenesulfonyl isocyanate as a novel
892   derivatization reagent to enhance the electrospray ionization and its application in the
893   determination of two stereo isomers of 3-hydroxyl-7-methyl-norethynodrel in plasma. *J
894   Chromatogr B Analyt Technol Biomed Life Sci, 814*(2), 331-337.
895   doi:10.1016/j.jchromb.2004.10.054

Figures:

896
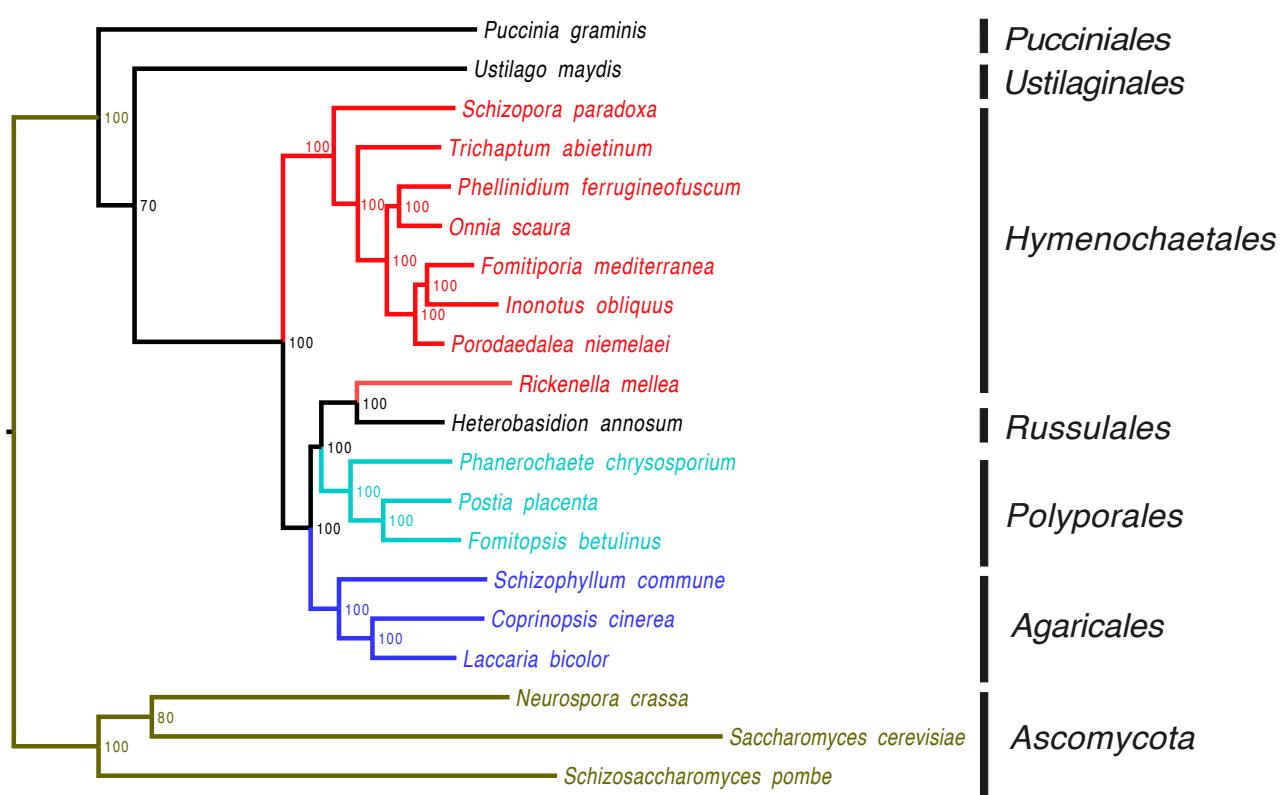
Figure 1. Phylogenetic tree of three *Ascomycetes* and 17 *Basidiomycetes* species. *Ascomycetes* are highlighted with green colour and grouped by Phylum. *Basidiomycetes* species are grouped by taxonomic order. Bootstrap values are plotted next to the nodes illustrating the level of the confidence of the split, and the phylogenetic tree was rooted to *Ascomycetes* clade.

897
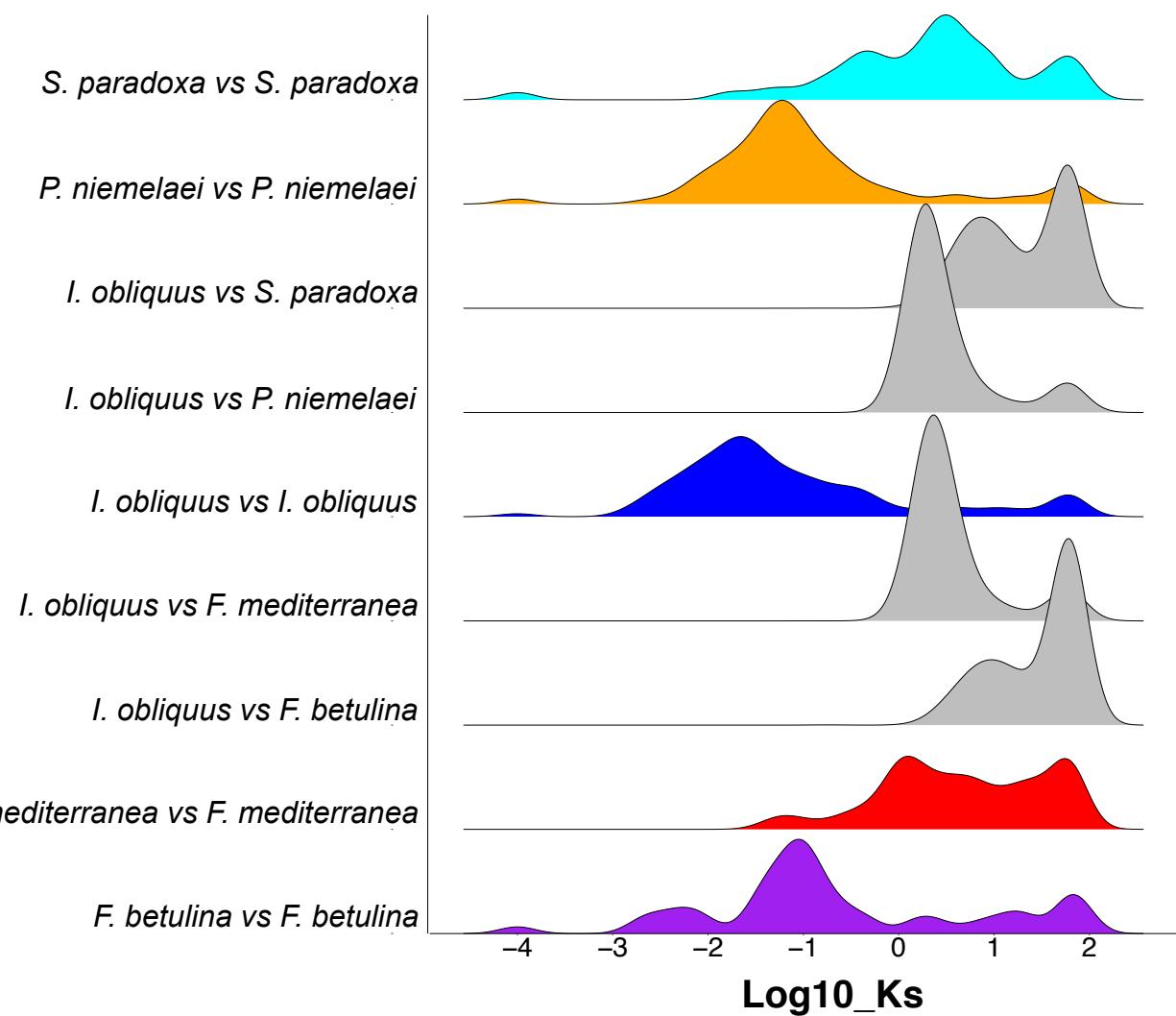
Figure 2. Density plots of the number of synonymous (Ks) substitutions in syntelogs identified from syntenic alignments of *I. obliquus*, *F. mediterranea*, *P. niemelaei*, and *S. paradoxa*. X-axis is displayed as log10 of synonymous substitutions per synonymous site (Ks).
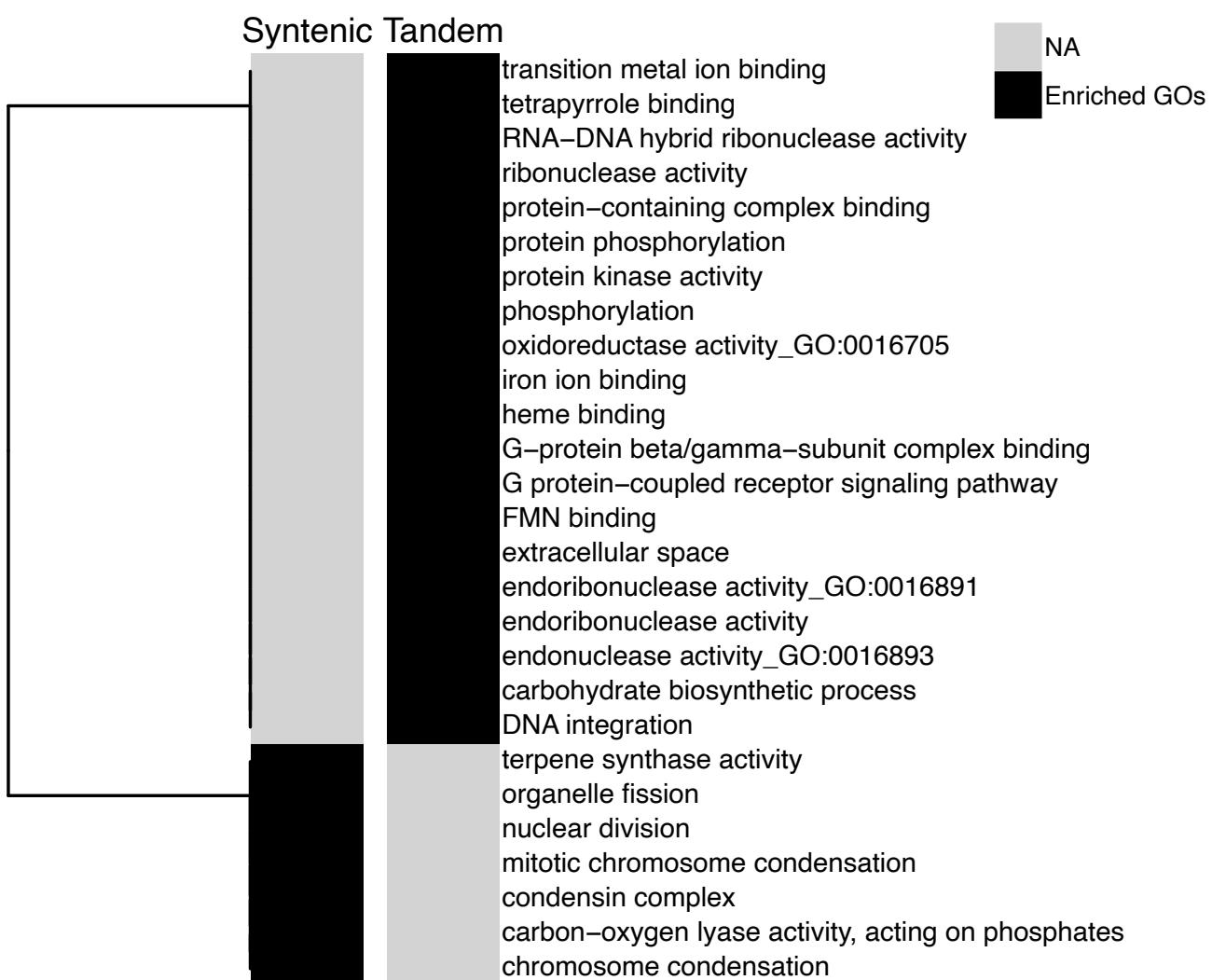
Figure 3. Gene ontology (GO) enrichment of genes identified in syntenic and tandemly duplicated regions in *I. obliquus* and four other species. Separate heatmaps are shown for the syntenic and tandemly duplicated regions, black color shows significantly enriched GOs and grey illustrates non-significant enrichments (NS). GOs are clustered using the Euclidean distance and hierarchical clustering (method: complete).

Expanded

Tandem   Syntenic
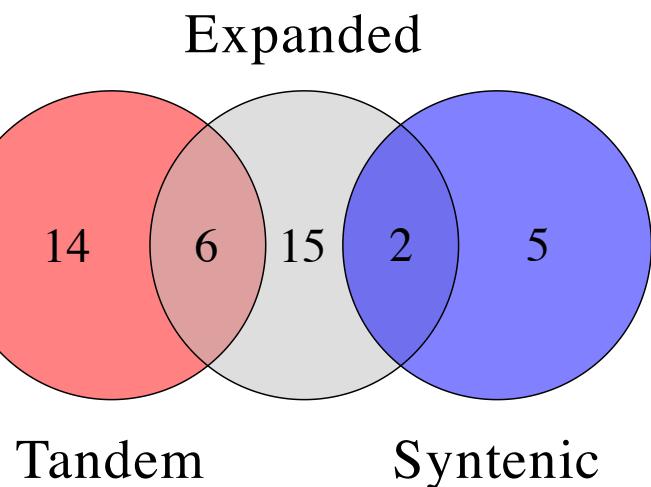
p.value: 6.1094e-11

Figure 4: Venn diagram of GO enrichment analysis for expanded gene families and tandemly duplicated genes from *I. obliquus* genome. Each category is highlighted and labeled by specific color. P -value was calculated with Fisher exact test assessing the the statistical significance of the overlap. Tandem: tandemly duplication genes; Expanded: expanded gene families; syntenic: syntenic self-self alignment of *I. obliquus* resulting in the set of genes originating from whole genome duplication event.

908
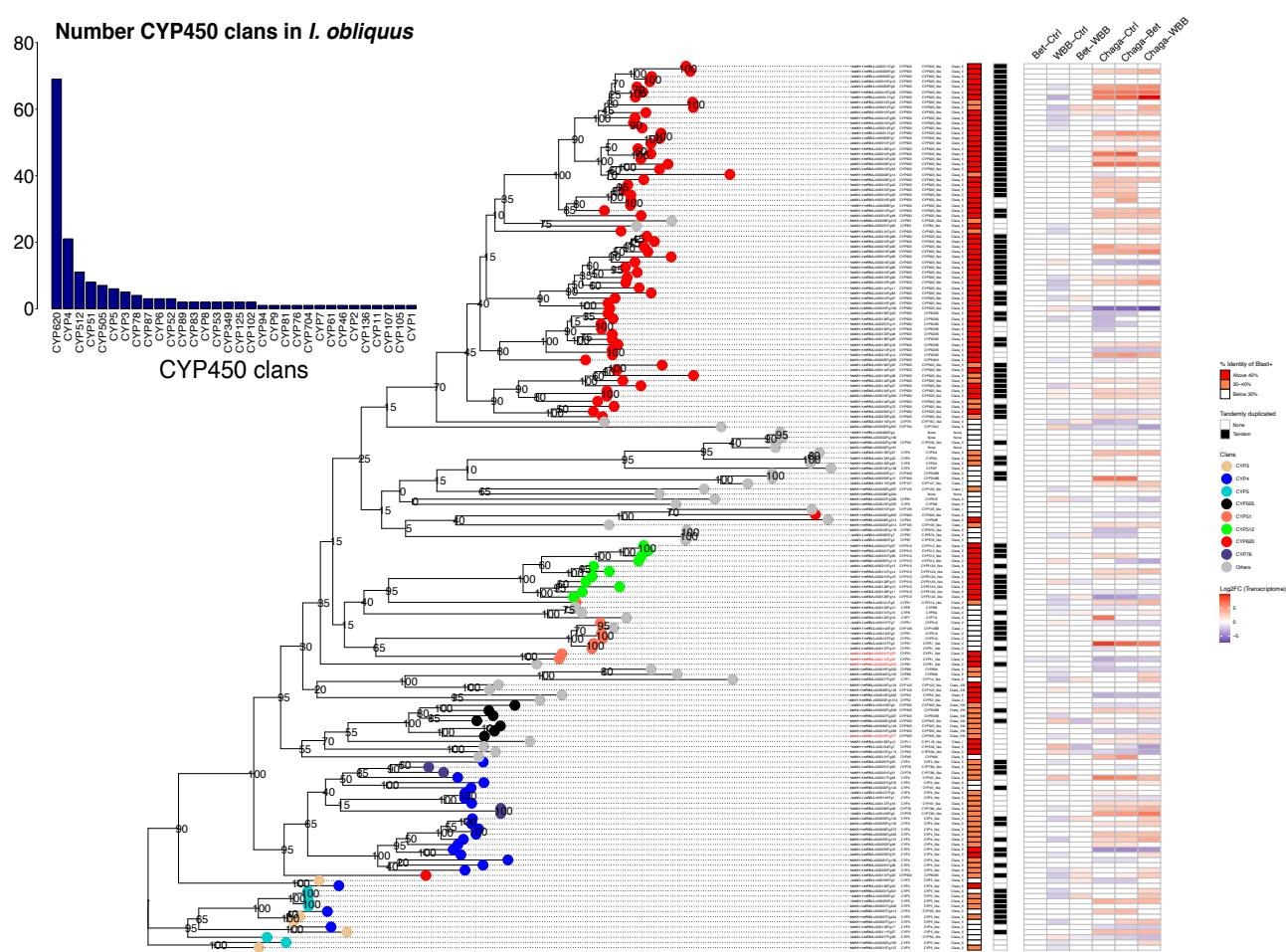
909

Figure 5. A gene tree of CYP450s predicted in *I. obliquus*. The clades are highlighted according to the major clan of the enzyme and tandemly duplicated genes are indicated with black squares; the color scheme is shown in the color key to the right of the plots. The branches are labeled by gene ID, cytochrome P450 clans, family, and class. The heatmaps illustrate BLAST similarities (red), tandemly duplicated (black/white), and differential expression of the genes. Barplot in the inset shows the number of CYP450 clans in *I. obliquus*.
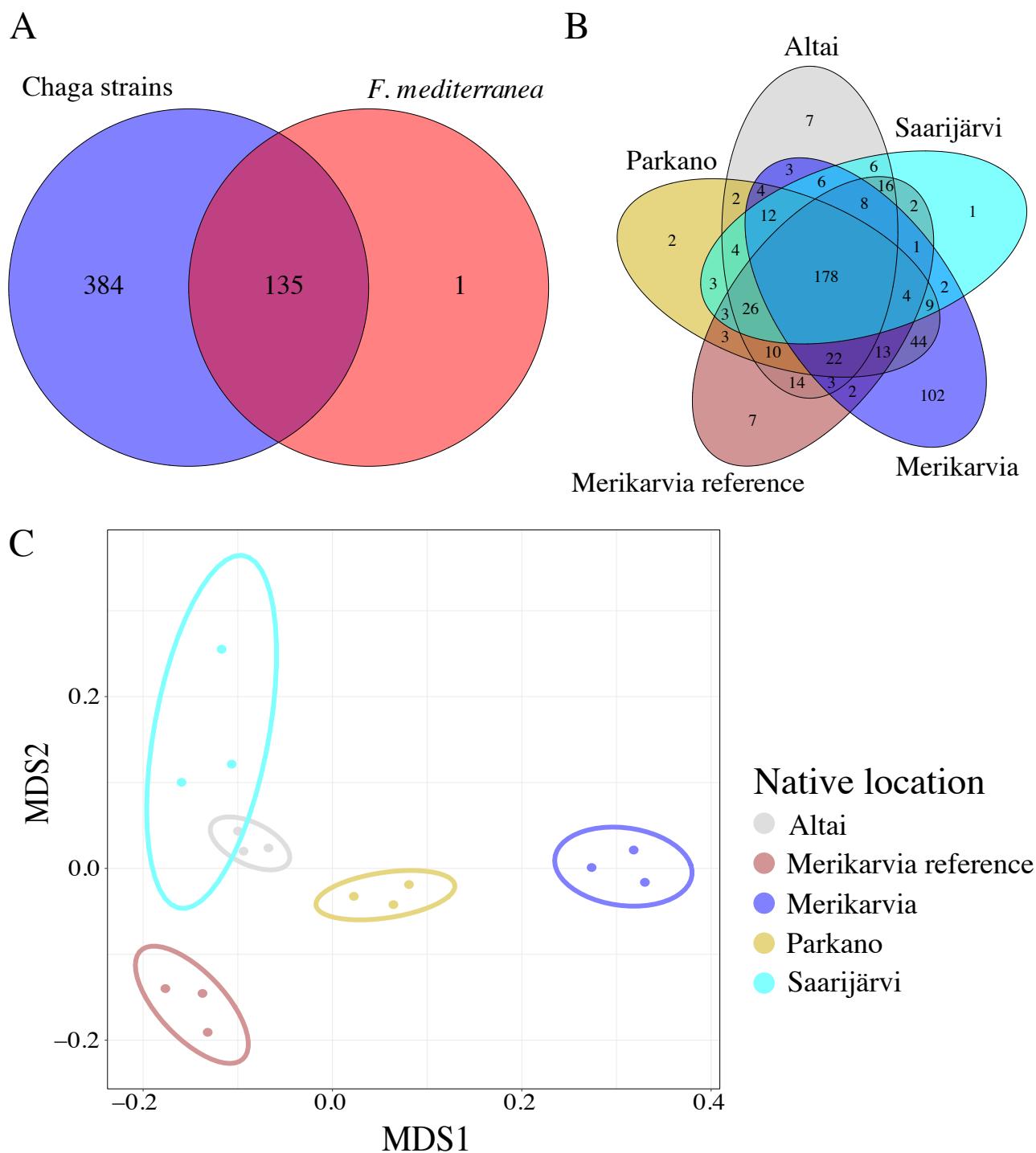
Figure 6: Venn diagrams of UPLC-MS mass spectrum for metabolomic fingerprints. A) Pooled mass spectrums from five strains of *I. obliquus* and one *F. mediterranea,* B) mass spectrums of five strains of *I. obliquus*, C) multidimensional scaling (MDS) plot of metabolomics abundant of five strains of *I. obliquus*. Panels B and C use the same color coding, explained in panel C legend. .
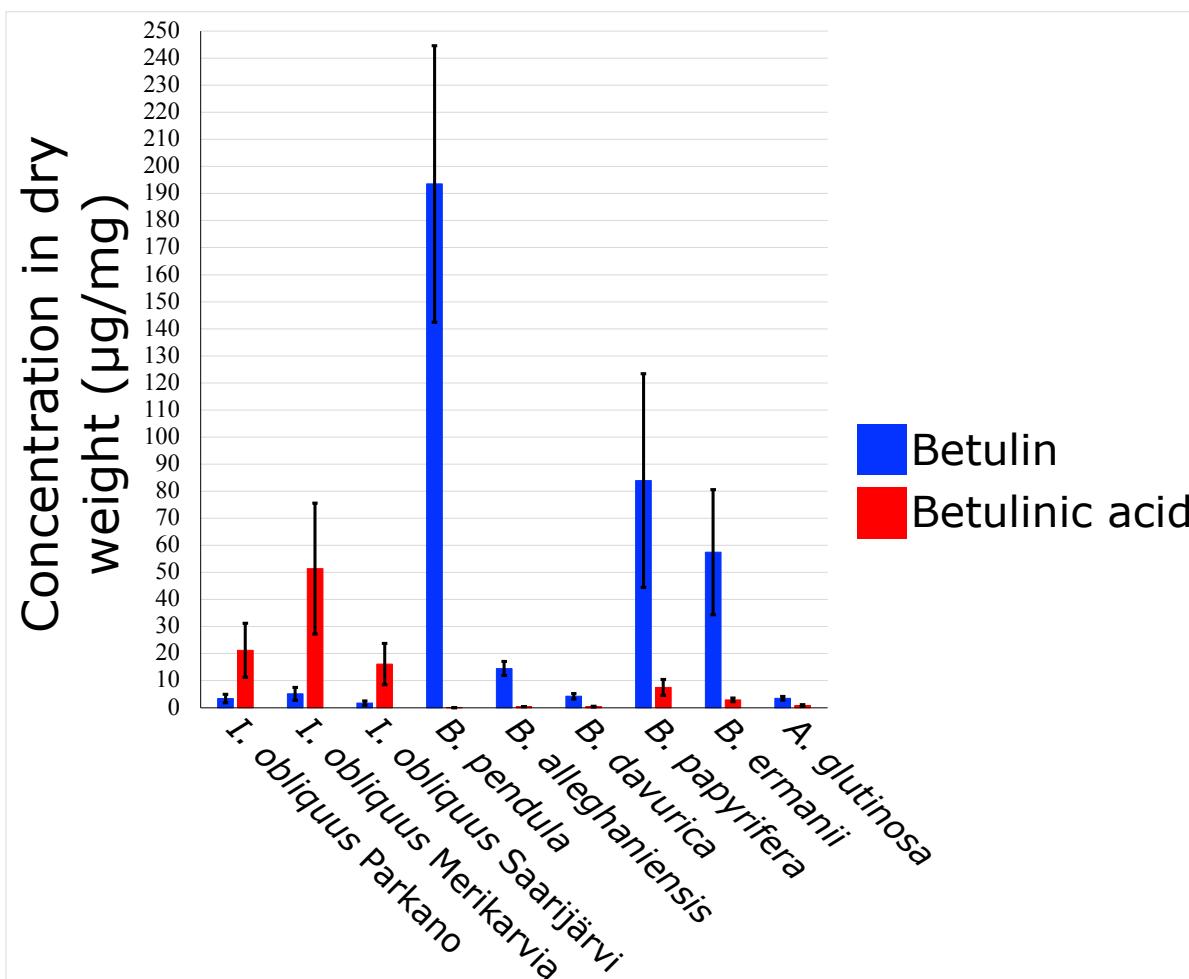
Figure 7: Concentration of betulin and betulinic acid in three strains of chaga and six betula species by HPLC-MS. Betulin and betulinic acid are highlighted in red and blue.
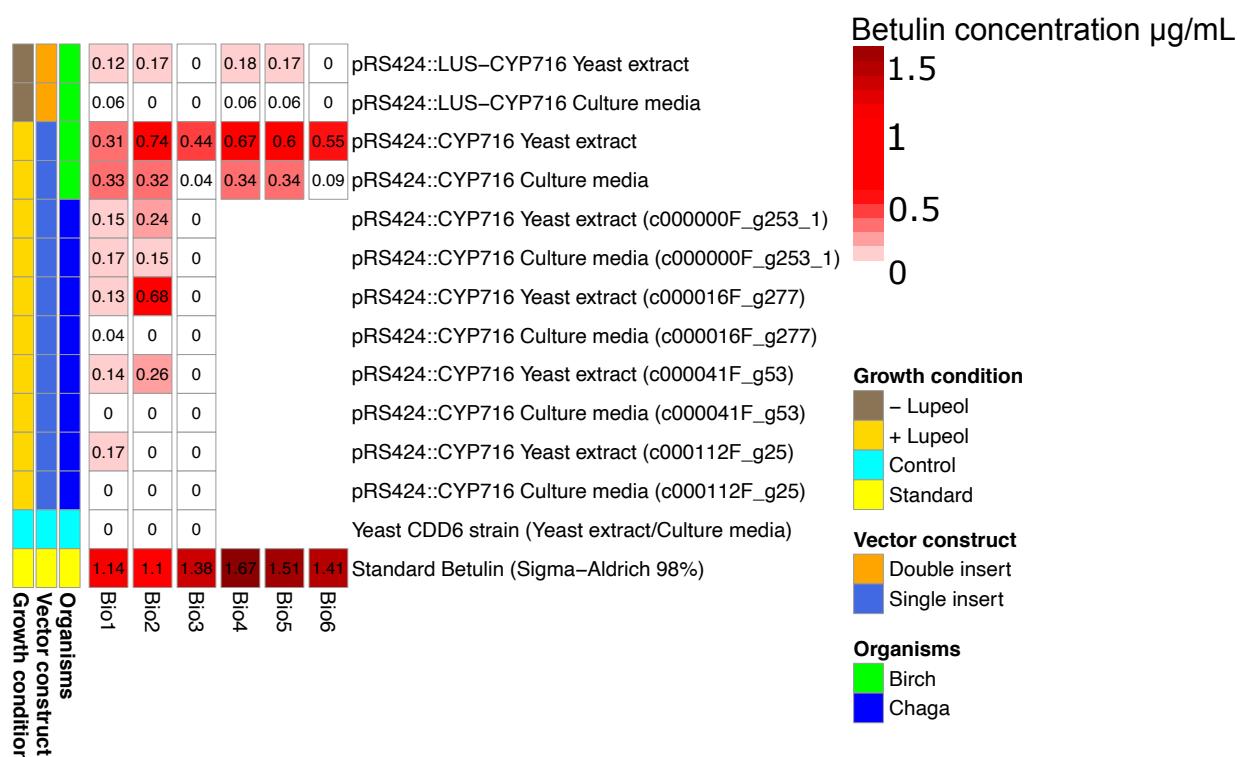
Figure 8: Heatmap illustrating the amount of betulin synthesized in transgenic yeast. The columns illustrate the biological replicates, and rows the vector constructs for lupeol synthase and CYP716 genes. The biological source for lupeol synthase was birch (green, *B. pendula*), and the biological sources for CYP716 enzymes were from birch and chaga mushroom (dark blue, *I. obliquus*). Vector constructs were divided to double insert (orange, lupeol synthase and CYP716 from birch in a single vector), and single inserts (royal blue, only CYP716 from birch and chaga). The color palette of the heatmap illustrates the concentrations (μg/mL) of betulin found in yeast cell (Yeast extract) and yeast growth media (Culture media), with white color assigned to minimum and dark red to maximum concentration of betulin. The second heatmap illustrates the growth conditions: brown (- Lupeol) is yeast growth media without standard lupeol (precursor for CYP716 gene), gold (+ Lupeol) yeast growth media with standard lupeol, and cyan CDD6 yeast strain used as control, and finally yellow is lupeol standard (98% purity).
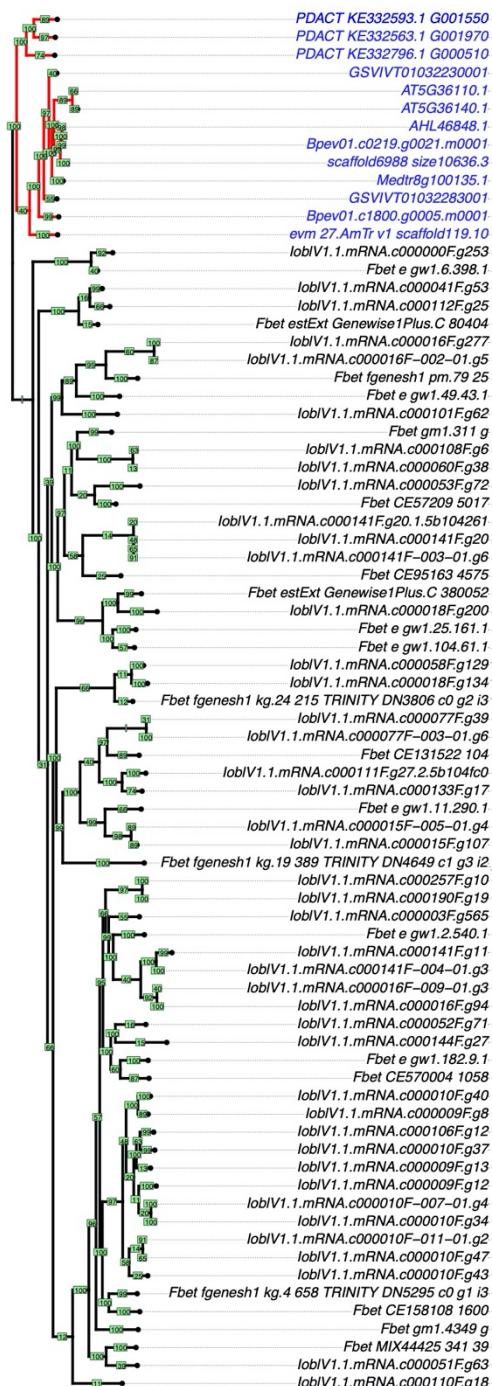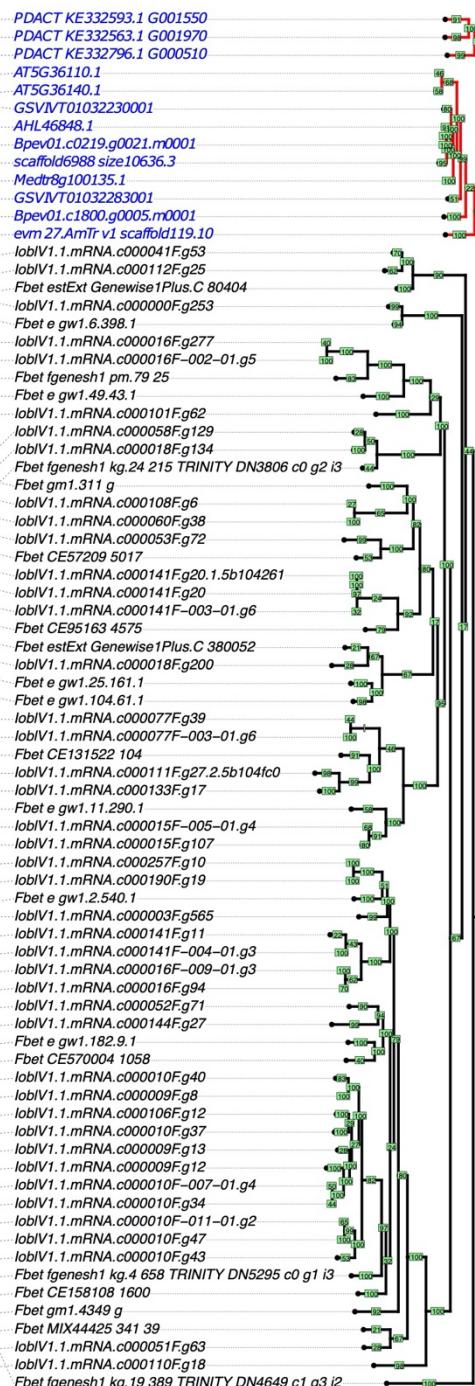
Figure 9: Phylogenetic reconstruction of 70 monooxygenase enzymes from *I. obliquus*, *F. betulina*, and 8 plant species. Trees are labeled according to the type of the sequence used in multiple sequence alignment (amino acid or nucleotide) and rooted to plant species. Cophylo (from phytools) function is used in order to rotate the branches to match the tips and the labels. Bootstrap values (green rectangular) illustrate the level of the confidence, and plant species are highlighted in blue.
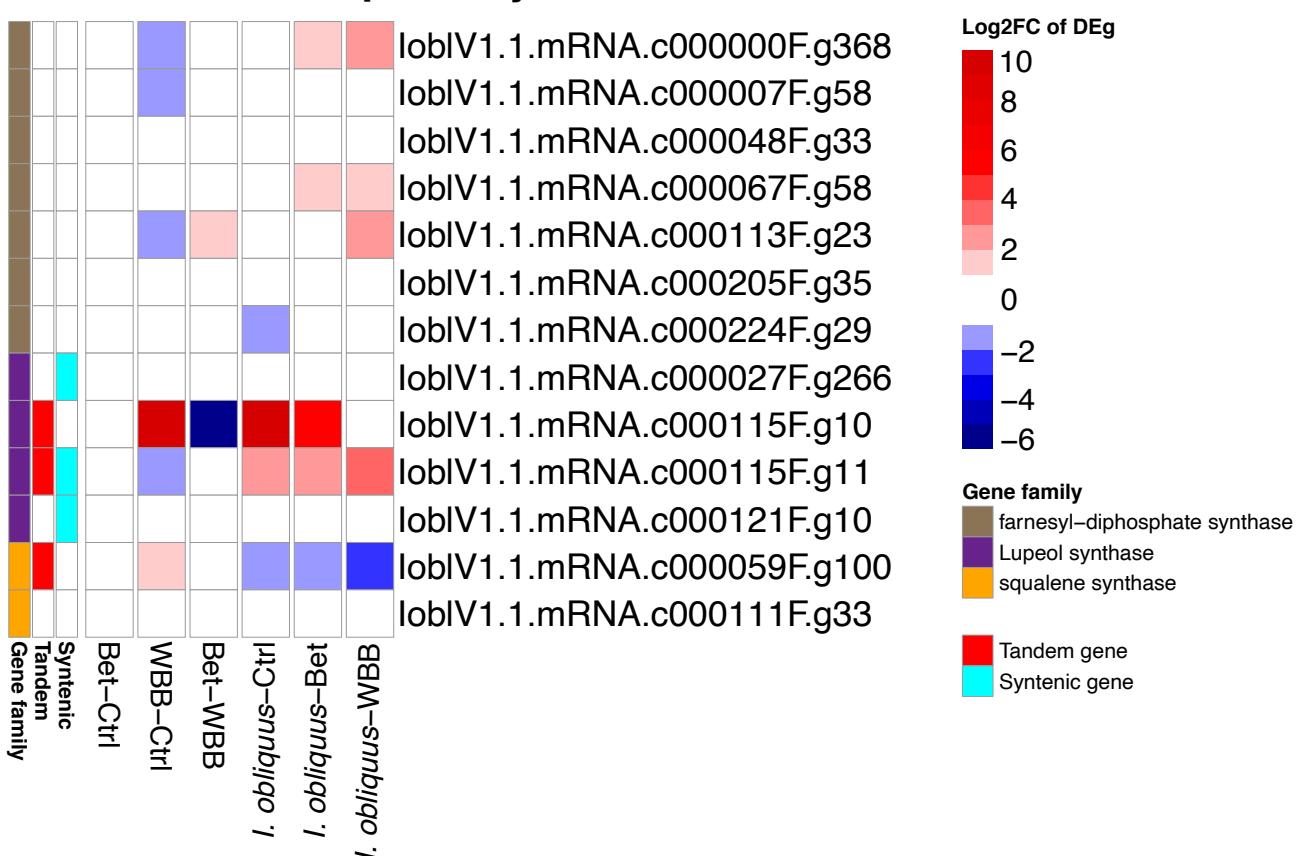
Figure 10: Heatmap of differentially expressed (DE) genes annotated: A) CYP450 monooxygenase enzymes, and B) key enzymes in mevalonate pathways (MVA). The larger heatmap illustrates the $\log_2$ fold-changes ($\log_2$FC) of DE genes, the color is proportional to differential expression. The smaller heatmap illustratres the duplication origins of the gene, either syntenic originating from whole genome duplication or tandem originating from segmental duplication. Genes are clustered using the Euclidean distance and hierarchical clustering (method: complete). Gene ID highlighted in blue was selected for cloning.
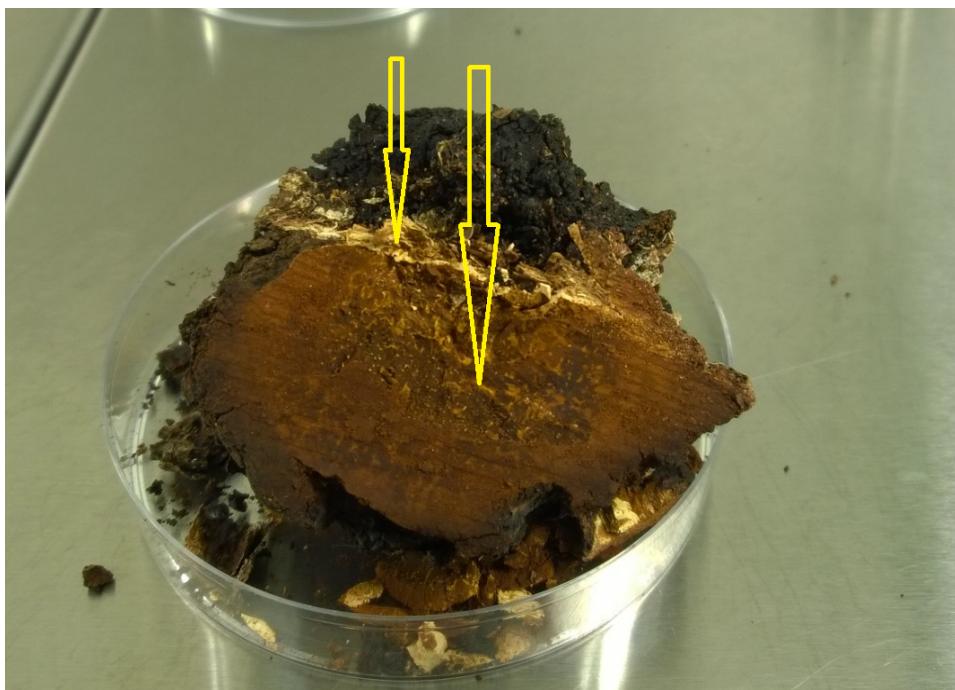
Supp. Fig 1: Fragment of *Inonotus obliquus* conk collected from Merikarvia, Finland [N62.00°, E24.74°]. Yellow arrows point at sites where samples were collected for inoculation of culture media. There might be contamination of birch bark at the site where the smaller arrow is pointing. The larger arrow points to the center of the conk where more pure sample was collected.
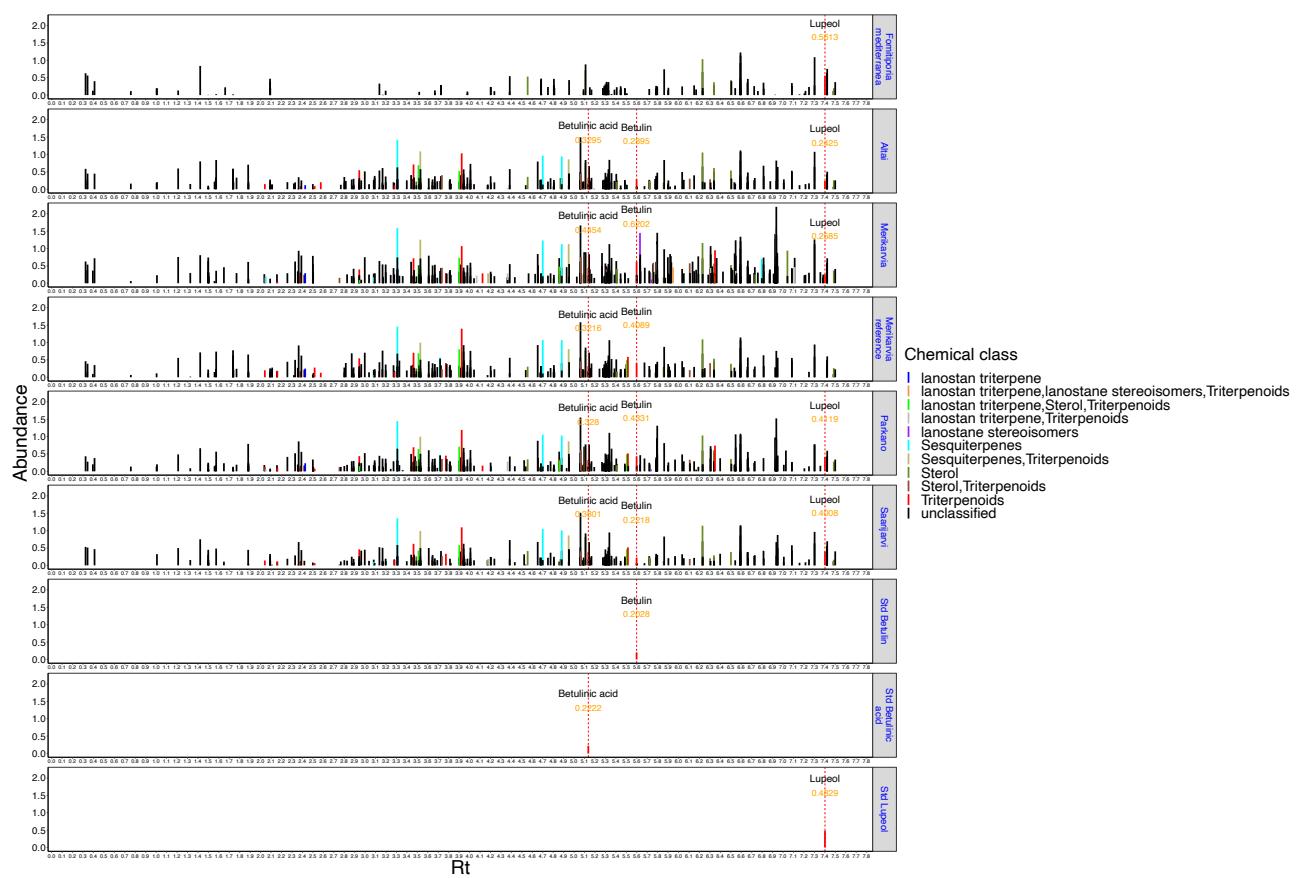
Supp. Fig 2: Relative abundances of secondary metabolites among five strains of *I. obliquus*, one strain of *F. mediterranea*, and standard (98%) lupeol, betulin, and betulinic acid. X-axis is the retention time, and Y-axis the relative abundance (ggplot, scales='free_x'). Vertical red lines show the standard retention times for lupeol, betulin, and betulinic acid, labeled with the name of the compound and its relative abundance. Mass spectra are classified to known chemical class. Std: standard (98%).

910
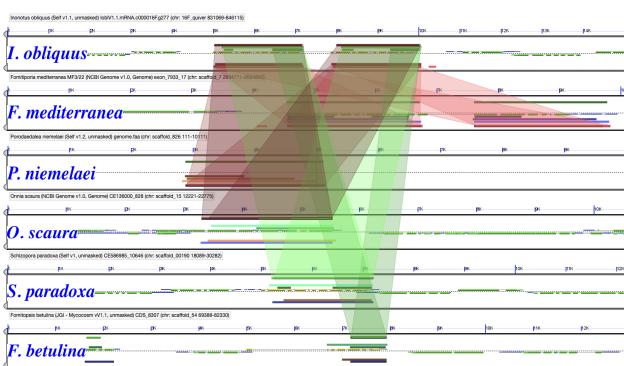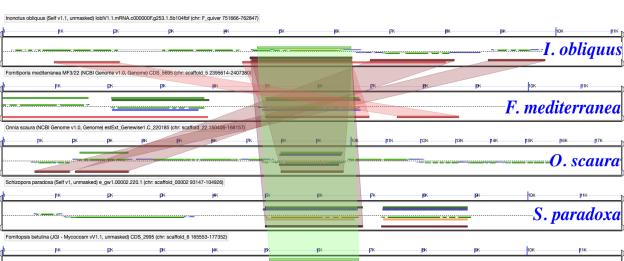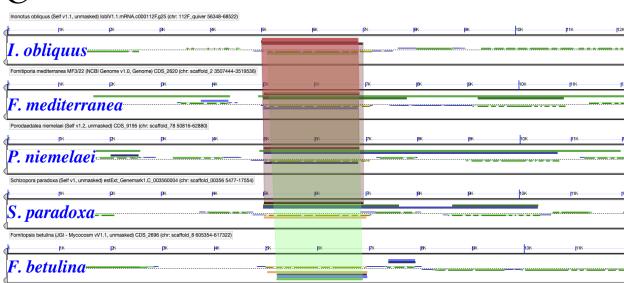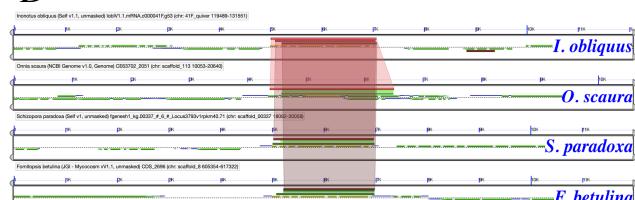
Supp. Fig 3: Microsynteny analysis of four cloned genes from *I. obliquus* and four *Hymenochaetales* species and *Fomitopsis betulina*. A) gene ID : c000016F.g277, **B)** gene ID : c000000F.g253, **C)** gene ID : c000112F.g25, and **D)** gene ID : c000041F.g53.

Supp. table 1: Statistics of I obliquus genome assembly and annotation, genome size of orthologous species, repeat masking statistics, and the GO enrichment of genes adjacent to DNA transposable elements.

Supp. table 2: List of putative secreted proteins from I obliquus.

Supp. table 3: List of genes from I obliquus which are associated to syntenic or tandemly duplicated regions of the genome.

Supp. table 4: List of expanded gene families and their GO enrichment in *I. obliquus*.

Supp. table 5: List of homologous C*AZymes from I. obliquus.*

Supp. table 6: List of syntenic and tandemly duplicated genes, and gene ontology (GO) enrichment in these genomic regions.

Supp. table 7: List of metabolomics fingerprints from five strains of *I. obliquus*, and one strain of *F. mediterranea*.

Supp. table 8: List of putative oxygen, heme, ERR-Triad domains across multiple kingdoms.

Supp. table 9: List of differentially expressed gene families.