scientific reports

Check for updates

OPEN

Independent evolution of betulin biosynthesis in *Inonotus obliquus*

Omid Safronov[1,2,3]✉, Güleycan Lutfullahoglu Bal[2], Nina Sipari[4], Maya Wilkens[5], Pezhman Safdari[1], Olli-Pekka Smolander[2], Pia K. Laine[2], Jenna Lihavainen[6], Niko Silvan[7], Sitaram Rajaraman[1], Lars G. Paulin[2], Teemu H. Teeri[8], Petri Auvinen[2], Tytti Sarjala[7], Kirk Overmyer[1], Uwe Richter[3,9]✉ & Jarkko Salojärvi[1,10,11]✉

Chaga mushroom (*Inonotus obliquus*) is a fungal species in the family Hymenochaetaceae (Basidiomycota) and the causative agent of white rot decay in *Betula* species. We assembled a high-quality 50.7 Mbp genome from PacBio sequencing and identified a lineage-specific whole genome duplication event approximately 1.3 million years ago, which has contributed to a major increase in biochemical diversity in the species through preferential retention of cytochrome P450 superfamily members. Secondary metabolism has further evolved through small-scale segmental duplications, such as tandem duplications within fungal biosynthetic gene clusters. Metabolomic fingerprinting confirmed increased complexity in terpene biosynthesis chemistry compared to related species that lacked the duplication event. This metabolic diversity may have arisen from co-evolution with the primary host species, which evolved high betulin content in its bark 4–8 million years ago.

Inonotus obliquus, or, chaga mushroom, is a fungal species from the family Hymenochaetaceae (Basidiomycota) and it is distributed across the boreal forest zone of the Northern Hemisphere. It causes aggressive white rot disease mainly among *Betula* family members[1], but upon suitable conditions can also infect other tree species, such as oaks, poplars, ashes, and maples[2]. White rot refers to diseases that primarily degrade lignin (having darker color), leaving the light-colored cellulose intact. The infection starts when chaga spores obtain access to the stem hardwood through cracked or wounded bark. At later infection stages, chaga appears as a sterile conk, a solid charcoal-black mass on the surface of bark[1]. This sterile conk has been used in traditional medicine of many cultures, and extensive research on its bioactive chemicals suggests promise for pharmacological, medicinal, and industrial applications[3–7].

The triterpenoids, betulin (BE) and betulinic acid (BA), are highly abundant in the bark of all birch family members; they collectively comprise 30–60% of total bark composition, depending on the species and the tissue type[8–10]. Both BE and BA have industrial applications[11] and function as therapeutic substances in oncology as well as in fungal, bacterial, and viral infections[12–14]. In plants, their biosynthesis starts with squalene, a product of the mevalonate pathway, and involves two enzymatic steps where squalene is first converted to lupeol via lupeol synthase (LUS) and then to BE by lupeol monooxygenase. The latter enzyme is a member of the large multifunctional family of cytochrome P450 (P450) monooxygenases, more specifically subfamily 716 (CYP716). In birch, the same enzyme is likely responsible for oxidizing BE into BA. BE biosynthesis occurs across a wide taxonomic range of plants, from Malvales[15], Fagales, Rosales[16], Fabales[17], Vitales[18], and Asterales[19] to Arecales[20,21], suggesting either ancestral origin or convergent evolution. A comparative genomic analysis of bark tissue in silver birch (*Betula pendula*) and black alder (*Alnus glutinosa*) revealed birch-specific evolution of the mevalonate pathway (MVA), where a tandem duplication of lupeol synthase colocalized with lupeol 28-monooxygenase was suggested as the reason for the high production of betulinate compounds in birch phellem[10]. Remarkably, betulinate compounds have also been found in diverse fungi, spanning the orders

[1]Organismal and Evolutionary Biology Research Program, Faculty of Biological and Environmental Sciences, and Viikki Plant Science Centre, University of Helsinki, 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, Helsinki, Finland. [4]Viikki Metabolomics Unit, Faculty of Biological and Environmental Sciences, University of Helsinki, Helsinki, Finland. [5]Institute of Molecular Biology GmbH, Ackermannweg 4, 55128 Mainz, Germany. [6]Department of Plant Physiology, Umeå Plant Science Centre, Umeå University, 90187 Umeå, Sweden. [7]Natural Resources Institute Finland (Luke) Horticulture Technologies, Helsinki, Finland. [8]Department of Agricultural Sciences, Viikki Plant Science Centre, University of Helsinki, Helsinki, Finland. [9]Wellcome Centre for Mitochondrial Research, Biosciences Institute, Newcastle University, Newcastle, UK. [10]School of Biological Sciences, Nanyang Technological University, Singapore, Singapore. [11]Singapore Centre for Environmental Life Sciences Engineering, Nanyang Technological University, Singapore, Singapore. ✉email: omid.safronov@helsinki.fi; uwe.richter@helsinki.fi; jarkko@ntu.edu.sg

Eurotiales[22], Hymenochaetales[23], and Polyporales[24]. The birch pathogens chaga and *Fomitopsis betulina* are examples of BE and BA producing species[23,24]. The origin and evolution of betulinate biosynthesis in fungi are not known, and no members of CYP716 gene family have been identified in fungi. The trait may have been gained by either convergent evolution or horizontal gene transfer (HGT) of cytochrome P450 monooxygenase enzymes from plants. Plant CYP716s are multifunctional enzymes that produce a range of related triterpenoid compounds, such as betulinate, oleanolate, and ursolate[25]. The diversification and distribution of the CYP716 family has been studied in eudicots[26], but the existence of this family, or other possible enzyme families, responsible for BE production in fungi remains unexplored.

Cytochrome P450 monooxygenase enzymes are among the oldest and largest gene families, ubiquitous across both prokaryotes and eukaryotes[27]. They act in detoxification of foreign compounds and have important roles in secondary metabolism related to environmental adaptation. The low sequence similarity, high functional diversity, and enzymatic promiscuity among P450 monooxygenase enzymes limit functional predictions. P450s are classified into families and subfamilies based on protein sequence similarity; sequences with identity > 40% are assigned into families and sequences with > 55% similarity into subfamilies. Novel candidates that are more divergent than confirmed P450 families form new candidate families. Remarkably, over 800 different CYP families have been identified in fungi alone[28].

Global climate change is expected to extend the range of numerous tree pathogens. Warmer and more arid summers will increase drought stress, making forest species more vulnerable to diseases. Increased prevalence of fungal diseases diminish the ecological services fulfilled by trees, including their ability to act as carbon sinks and mitigate climate change. Therefore, it is crucial to enhance the understanding of host–pathogen interactions. In this study, we sequenced and assembled the chaga genome using PacBio long-read sequencing with the aim of identifying the origins of betulin biosynthesis in this species. Comparative genomics analyses and untargeted metabolic fingerprinting showed expanded triterpenoid secondary metabolism in the species and linked it with a whole genome duplication event followed by tandem expansions of P450 genes. A candidate P450 monooxygenase enzyme from chaga with the highest sequence similarity to the CYP716 family was identified and tested, demonstrating no betulin biosynthesis activity. Further analyses of highly conserved protein domains of P450s excluded the possibility of horizontal gene transfer from host species, suggesting that betulin biosynthesis has evolved independently in fungi.

Results and discussion

Genome assembly and annotation

PacBio sequencing of *I. obliquus* strain from Merikarvia, Finland yielded 4.82 Gb of data (96 × coverage) with N50 read length of 9,200 base pairs (bp). Falcon assembly resulted in a 41.1 Mbp genome, consisting of 301 primary contigs with an N50 value of 516 kbp (Table S1). Altogether, 90.7% of universally conserved single-copy genes were complete in the assembly (BUSCO v3.0 with fungi database[29]), with 21.4% of genes being duplicated. The quality of the assembly was comparable to the previously published assembly[30] (Table S1, Fig S1), with a size similar to other species from Hymenochaetales (Table S2).

RNA-seq from the reference strain was employed to support gene model prediction, yielding 13,778 gene models with 91.3% BUSCO completeness, with 31% of BUSCO genes being duplicated; the annotation quality was hence considerably improved over previous version[30] (87.9%; Table S1). Additionally, a mitochondrial assembly was generated (118 kbp; Table S3) consisting of 29 tRNAs, 32 coding sequences, and 3 rRNAs. Compared to existing assemblies for *I. obliquus*[30] and *I. hispidus*[31], our assembly was ~ 3–7 Mb larger and contained an additional set of 1200–1474 gene models (Table S1). The increase likely comes from the large body of RNA-seq evidence employed here, yielding more accurate information on expressed transcripts and allowing the training of species-specific gene predictors (Table S1).

Transposable elements (TEs) play a significant role in genome plasticity and evolution. The classification and characterization of genes in proximity to TEs are of general interest, particularly in pathogenic organisms[32], as active TEs may contribute to effector diversity[33,34]. The TEs contribute also to genome sizes, as a comparative study on wood-decaying fungi in Polyporales revealed a positive correlation with TE content[35]. However, in contrast to the closely related *Fomitiporia mediterranea* (with 42.27% of the genome consisting of repeat elements), the chaga TE content did not fully explain its large genome size, with only 26% being repeat sequences. While 14.2% of the elements remained unclassified, retrotransposon insertions covered 8.4% of the genome, and more DNA transposon elements (1.18%) were identified compared to the related *F. betulina* and *F. mediterranea* (Table S4). Genes flanking the predicted TEs contained domains associated with transposition; for example, gene clusters between two TEs from the same DNA transposon class were enriched for gene models involved in transmembrane transport, protein dimerization, transposition, as well as DNA binding and recombination (Table S5); the enrichment of the last two categories suggests that some of the predicted gene models may have transpositional origin or that they are novel unidentified transposable elements.

The palette of effector-like proteins and carbohydrate-active enzymes is concordant with the pathogenicity of the fungus

Mechanisms of *I. obliquus* pathogenicity and host interaction are unknown. Secreted proteins (SPs), especially cysteine-rich short SPs (CSSPs), are candidate effector proteins, which have essential roles in pathogen-host interactions. Altogether, 1052 short (< 1 bp) open reading frames (ORFs) (7.6% of all gene models) were predicted as CSSPs, with a minimal set of known homologs (Table S6). The predicted CSSPs were scattered across 128 contigs. Most of them were likely species-specific, as homology searches with other species were successful for only 110 CSSPs (Table S7). Twenty-one CSSPs overlapped with at least one class of TEs, of which 18 were unclassified. A total of 988 CSSPs co-localized between two TEs of the same TE class, in support of the hypothesis that TEs may actively contribute to CSSP evolution and diversification[36].

Carbohydrate-active enzymes (CAZymes) have biological roles in the metabolism of diverse carbohydrates, such as glycogen, trehalose, and glycoconjugates[37]. Out of 415 predicted CAZymes (Table S8), altogether 184 were classified as glycoside hydrolases (GHs) and 36 as carbohydrate-binding modules (CBMs). The overall number of glycosyl transferases (GTs) and carbohydrate esterases (Ces) were 99 and 19, respectively. Finally, 10 enzymes were assigned to polysaccharide lyases (PLs). Overall, there were less CAZyme families than in *Trametes trogii*, a lignin-degrading fungus within Polyporales[38], but the number was still expanded compared with *Russula griseocarnosa* from the related order Russulales[39], an ectomycorrhizal symbiont that has lost genes involved in degrading plant polysaccharides and lignin[40]. Analysis of RNA-seq data from *I. obliquus* grown on *B. pendula* wood dust and publicly available data[41] found 214 CAZymes to have elevated expression levels (Table S8) in at least one of the experimental conditions. Most of the differentially expressed (DE) CAZymes belonged to the GH and GT categories (Table S8) in support of their role in disassembling plant cell wall[42].

Gene families associated with secondary metabolism were expanded in chaga

To determine the taxonomic placement of *I. obliquus*, we analyzed the proteomes of 15 representative fungi from different orders within Basidiomycota and five outgroup species (Table S9). Proteins were first clustered into computational gene families (orthogroups) using Orthofinder[43], and phylogeny was then estimated from 4040 single-copy genes and rooted to Ascomycete species. The result illustrates the known taxonomy among the 20 fungal species (Fig. 1) and the split of Hymenochaetaceae family occurs at the expected phylogenetic position[44–46]. The obtained phylogeny grouped *Rickenella mellea* together with the Russulales representative; while whole-genome analysis can produce more accurate phylogenies than a small number of markers (including ITS), more representative genomes are needed to test the suggested placement.

Altogether, 167 orthologous gene clusters consisting of 2925 genes were significantly expanded in chaga, in comparison with the other 19 fungal species (adjusted p-value < 0.05, χ^2 test; Table S10). The expanded gene families were significantly enriched (Fisher's exact test, false discovery rate adjusted p-value < 0.05) for 19 GO terms that were associated with the mevalonate pathway, oxidative stress responses, transposition, protein dimerization activity, as well as heme and WD40-repeat domain binding (Table S10).

Recent whole genome duplication in chaga shows preferential retention of P450 genes

Self-self syntenic alignment of *I. obliquus* identified a set of 162 syntenic blocks of lengths 5–18 genes, consisting of a total of 1112 genes (Table S11). The distribution of synonymous mutations (Ks) between the syntenic duplicates (syntelogs) displayed a peak, as would be expected from a large-scale duplication event (Fig. 2). The syntelogs were scattered across the assembly, suggesting that the duplication had occurred on the entire genome and was therefore a whole genome duplication (WGD) event. Based on the Ks peak observed for the syntenic alignment against *F. mediterranea,* the event occurred after the split of the two species. Assuming that the species divergence occurred approximately 112 million years ago during the Triassic period[47], the WGD would have occurred relatively recently, at 1.3 million years ago (Mya). This is considerably more recent than the evolution of white-barked birches, which occurred approximately 8.4 Mya (even when including confidence intervals 3.6–

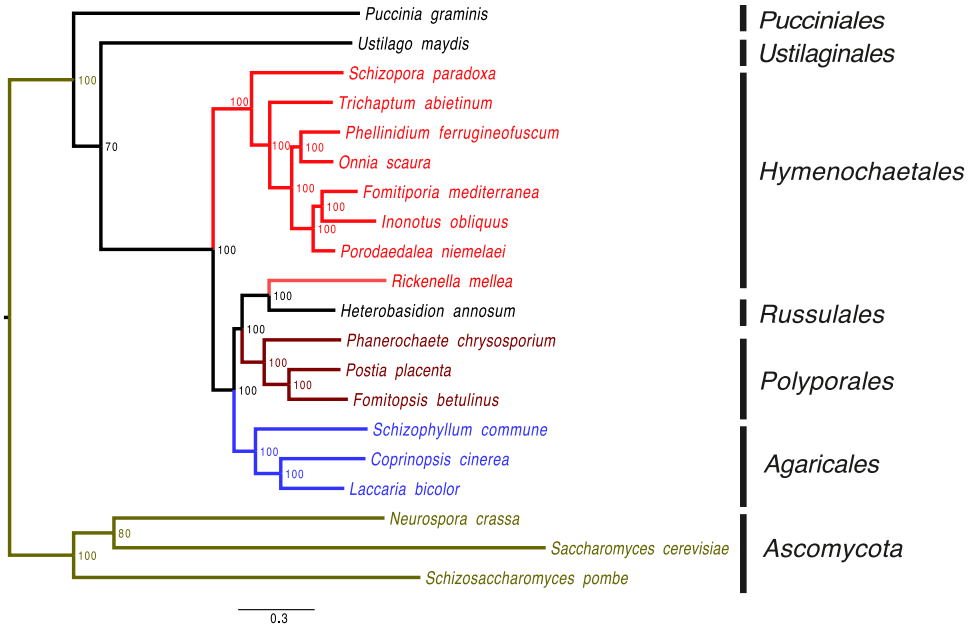

Fig. 1. Phylogenetic tree of three *Ascomycetes* species and 17 *Basidiomycetes.* Phylogeny was estimated from 4040 single-copy genes present in all species and rooted to *Ascomycota* species. *Ascomycetes* are indicated with green and grouped by phylum. *Basidiomycetes* orders are indicated with distinct colors and grouped by taxonomic order. Bootstrap support values report the level of confidence. The phylogenetic tree was rooted to *Ascomycetes* clade.

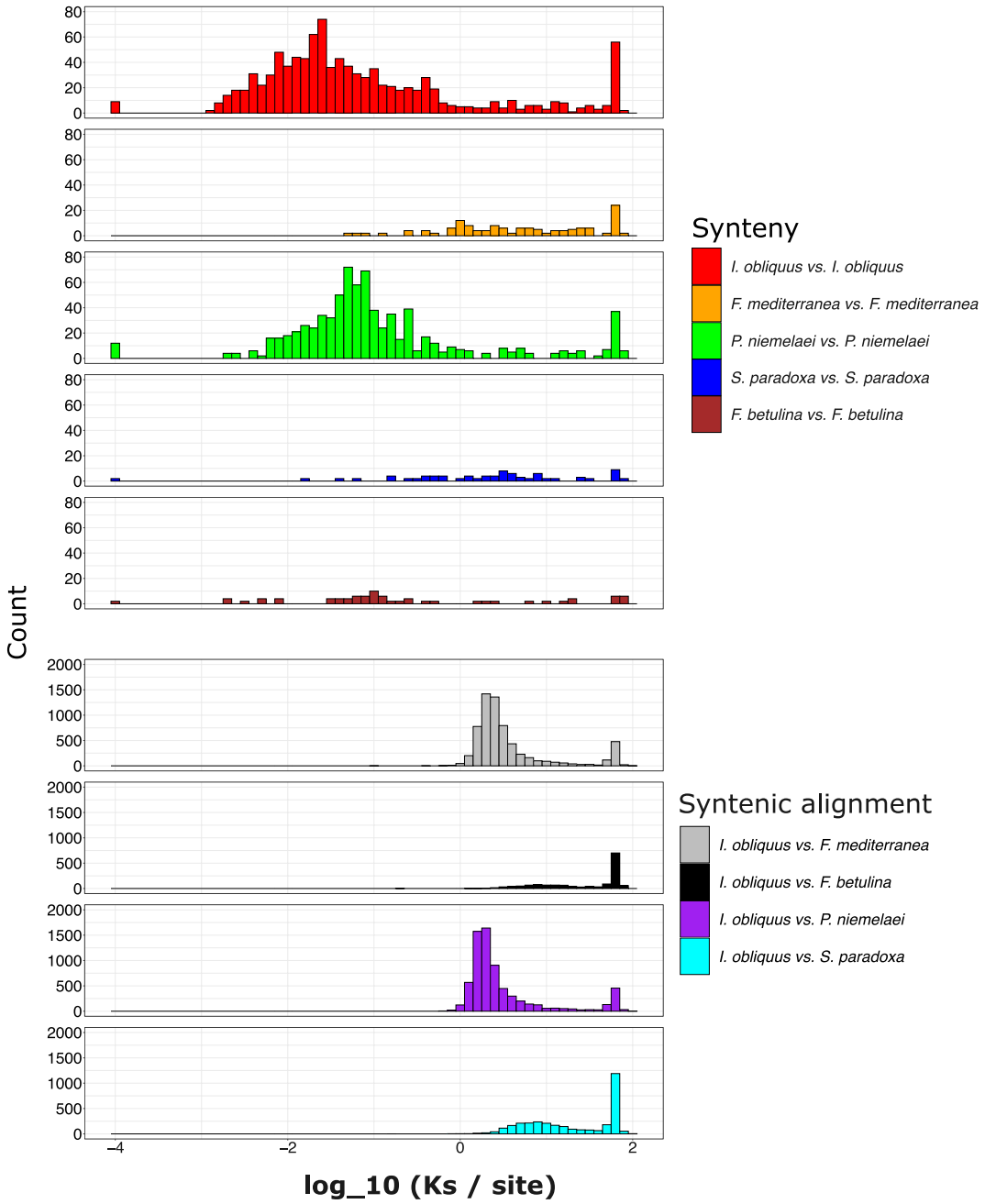

Fig. 2. Estimation of whole-genome duplication events. Histograms of log10 of the number of synonymous (Ks) substitutions in syntelogs identified from self-self alignments for *I. obliquus*, *F. mediterranea*, *P. niemelaei*, and *S. paradoxa* (top five panels) and interspecies alignments. The X-axis shows log10-scaled values of synonymous substitutions per syntelogous gene pairs (Ks values). Y-axis shows absolute counts.

13.6 Mya)[48]. Therefore, the evolution of triterpene biosynthesis pathways in chaga may have been driven by the evolution in the host plant.

Syntelogs were significantly enriched for GO processes related to terpene synthesis and cell division (Fig. 3, Table S12). We also observed a second independent lineage-specific WGD in *Porodaedalea niemelaei*. In this case, the 848 gene retained duplicates were enriched for GOs related to secondary metabolism, such as oxidoreductase activities and monooxygenase activities (Table S12). While other members of this genus are pine and spruce pathogens, *P. niemelaei* is a specialist on *Larix sibirica*[49], a species known for its highly decay-resistant wood[50].

The preferential retention of enzymatic gene families following WGDs in these fungi is remarkably different from other kingdoms such as plants where gene families in regulatory roles have been found retained, in concordance with dosage-balance hypothesis[51]. While ancestral WGDs are relatively uncommon in fungi[52], two recent WGDs among the five analyzed fungi suggest that these events are not necessarily rare and that they may be linked with environmental adaptation; see also[52].

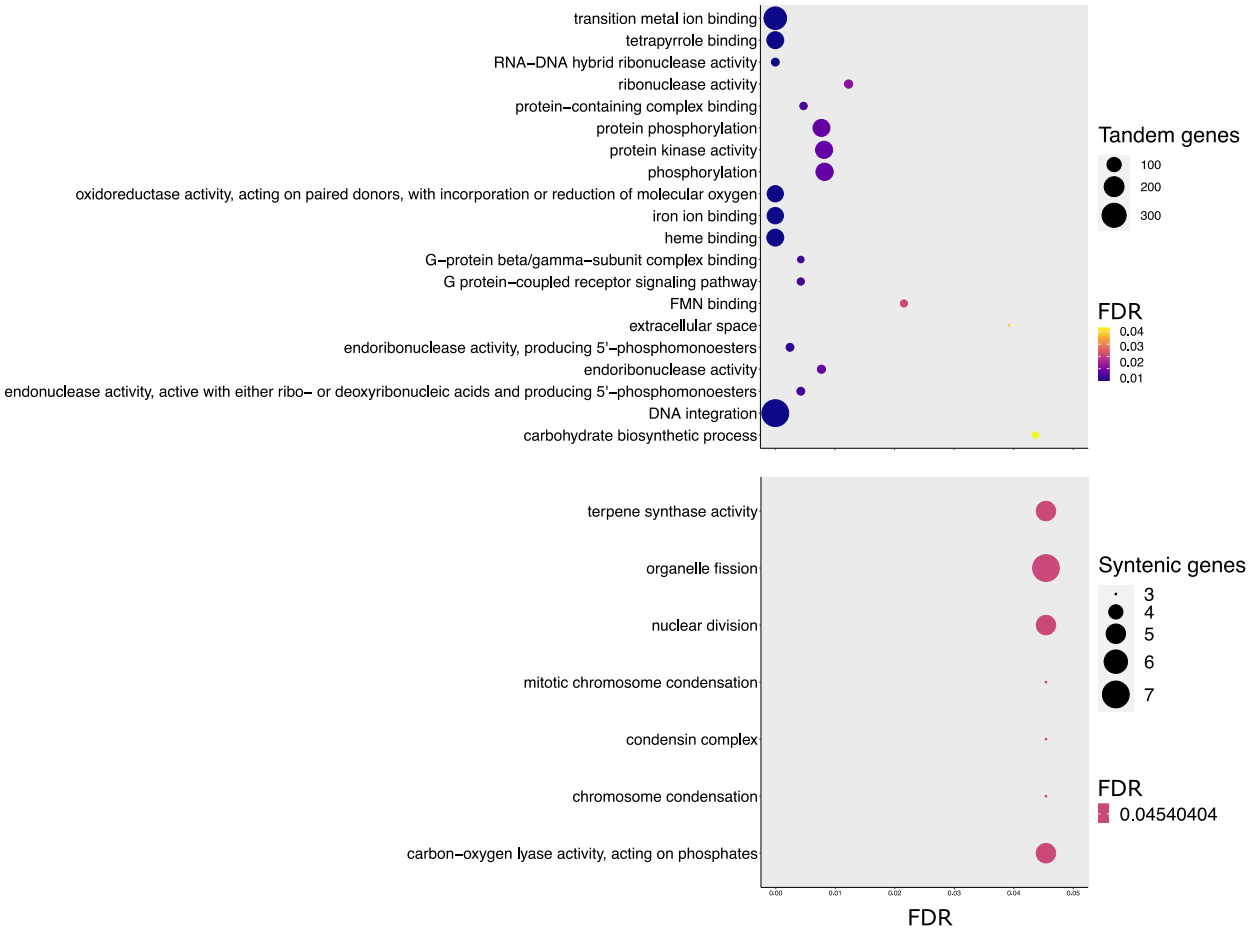

Fig. 3. Bubble plot of gene ontology (GO) enrichments of genes originating from tandem duplications (top panel) and whole-genome duplication events (bottom panel) in *I. obliquus*. The X-axis and the bubble color correspond to the adjusted *p*-value (FDR; see the "FDR" color key). The bubble size indicates the number of genes belonging to each category (see legend for size guide).

Tandemly duplicated genes reflect the short-term adaptation in the species and in general are associated with environmental responses[53]. In *I. obliquus*, we identified a high number of genes (6200) originating from tandem duplications. These genes were enriched for carbohydrate biosynthesis, heme binding, oxidoreductase activity, tetrapyrrole binding, and DNA transposition (Table S12). Overall, the different analyses showed complementary results; the expanded gene families in chaga showed a large significant overlap with tandemly duplicated genes but also with genes originating from WGDs ($p = 2.2e{-}16$; Fisher's exact test) (Fig. 4).

Altogether, the analyses on genome evolution suggested that secondary metabolism has been under positive selection in chaga, both in terms of genes retained after WGD and in subsequent tandem duplication events, both contributing to an expansion of P450 genes (Fig. 5). The members of P450 gene family have key roles in many biochemical modifications of secondary metabolites. A total of 172 P450s were predicted in chaga, suggesting a complex biochemical diversity in chaga secondary metabolism. Most of the P450 monooxygenases belong to the CYP620 family (68 members), followed by the CYP4, CYP512, and CYP505 families (Fig. 5). CYP620s are involved in terpenoid biosynthesis[39,54] and CYP4s in fatty acid metabolism[55], while the CYP512 family has been hypothesized to have catalytic activities towards steroidal-like compounds, primarily testosterone[56]. CYP505 is a fungal-specific family, possibly with a role in the biosynthesis of squalene-type triterpenoids[57,58]. A high proportion of P450 gene models (79 out of the total 172) were tandemly duplicated, many of which were members of the CYP620 family.

A GO enrichment analysis of genes flanking P450s (two upstream and two downstream) revealed enrichment of biological functions related to oxidoreductase activity, heme binding, transmembrane transporter activities, and tetrapyrrole binding (Table S12). This co-localization of genes near P450s prompted identification of biosynthetic clusters in the *I. obliquus* genome using antiSMASH fungi[59]. We identified 23 biosynthetic clusters in 17 contigs within the genome (15 terpene synthase, 3 polyketide synthase, and 5 non-ribosomal peptide synthetase clusters). The clusters were significantly enriched for tandemly duplicated genes (Fisher's exact test, $p = 6.34e{-}76$), suggesting that the mechanism has played a major role in biochemical diversification. Furthermore, the clusters were enriched for the P450 gene family ($p = 0.032$), highlighting their central enzymatic role in

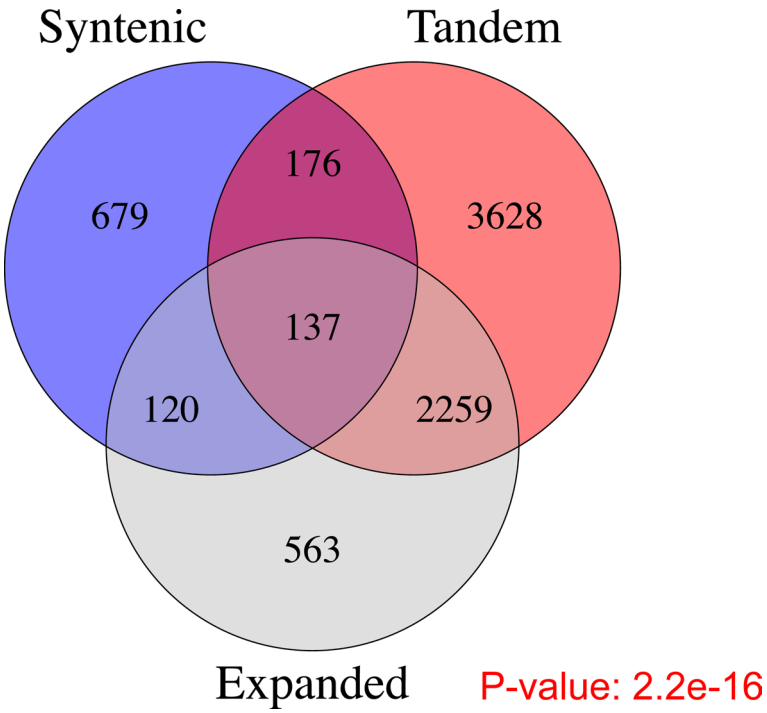

Fig. 4. Venn diagram of expanded gene families ("Expanded"), genes originating from whole-genome duplication ("Syntenic"), and tandemly duplicated ("Tandem") genes in *I. obliquus* genome. Each category is highlighted with a distinct color and labeled according to the category. *P*-value reports the *p*-value of Fisher's exact test, testing the statistical significance of overlaps.

secondary metabolism (Table S13). Altogether, these results highlight that both WGD and tandem duplication events contributed to the expansion of P450 gene families and to the evolution of biosynthetic clusters.

Metabolomic fingerprinting confirms complex terpenoid metabolism

Considering that genome evolutionary processes had amplified secondary metabolism gene counts in chaga, we tested whether this has resulted in a more diversified secondary metabolism. Comparative terpenoid metabolomic fingerprinting of five strains of *I. obliquus* and *F. mediterranea* revealed distinct differences (Table S14). Together, the chaga strains had 546 mass spectrum peaks, and only 135 of these were present in *F. mediterranea* (Fig. 6). In addition, pairwise comparisons of each chaga strain and *F. mediterranea* found 178 metabolomic features among *I. obliquus* strains with significantly higher abundance (Fig S2, Table S15). Among the different strains, Merikarvia had a distinct metabolomic fingerprint and did not group with the others (Fig. 6) in principal coordinate analysis, implicating that genotypic variation also contributes to biochemical diversity in chaga (Fig. 6). Although metabolomic fingerprinting does not accurately identify the underlying metabolites, the analysis suggested the presence of terpene, lupeol, and BE derivatives based on the predicted carbon and hydrogen contents. Many of the peaks were predicted to have molecular formulae with 30-, 31-, and 28-carbon backbones, similar to triterpenoid compounds such as lupeol, BE, and BA. The Merikarvia strain demonstrated greater BE and BA contents than other *I. obliquus* strains (Fig S2, Table S15), and while there were no mass spectra peaks that matched the BE standard or BA in *F. mediterranea*, a significant quantity of lupeol-like substances was discovered (Fig S2).

Further HPLC quantification of BE and BA in six species of *Betula* and three strains of chaga showed that chaga had a higher concentration of BA compared with BE (Fig. 7), while the relation was opposite in the *Betula* species. These results indicate that BE is predominantly in a different biochemical form in chaga compared to its host species, highlighting a difference in BE metabolism.

Betulin metabolism in chaga was not acquired through horizontal transfer

We next queried the chaga genome for candidate genes responsible for BE biosynthesis. The CYP716 family has been characterized as plant-specific; accordingly, no members of this family were predicted. We proceeded by identifying candidate genes in chaga and *F. betulina* based on sequence homology to 13 monooxygenase enzymes from eight plant species confirmed to produce betulinate compounds. A distinct divergence of fungal clades from the plant species was apparent (Fig S3). This result is consistent in both protein- and DNA-based gene trees and sequence homology, providing no support for horizontal transfer events (Fig S3, Fig. 8).

The gene model from *I. obliquus* with the highest BLAST sequence similarity to the birch CYP716 was a gene with ID c000016F_g277, a member of CYP505 family. Our analysis of CYP505 suggested the family to be specific to fungi, as we did not find homologs in any other kingdoms (Fig. 8). Upon closer inspection, the coding sequence length of c000016F_g277 was 3297 bp, approximately twice the length of most other P450s predicted in

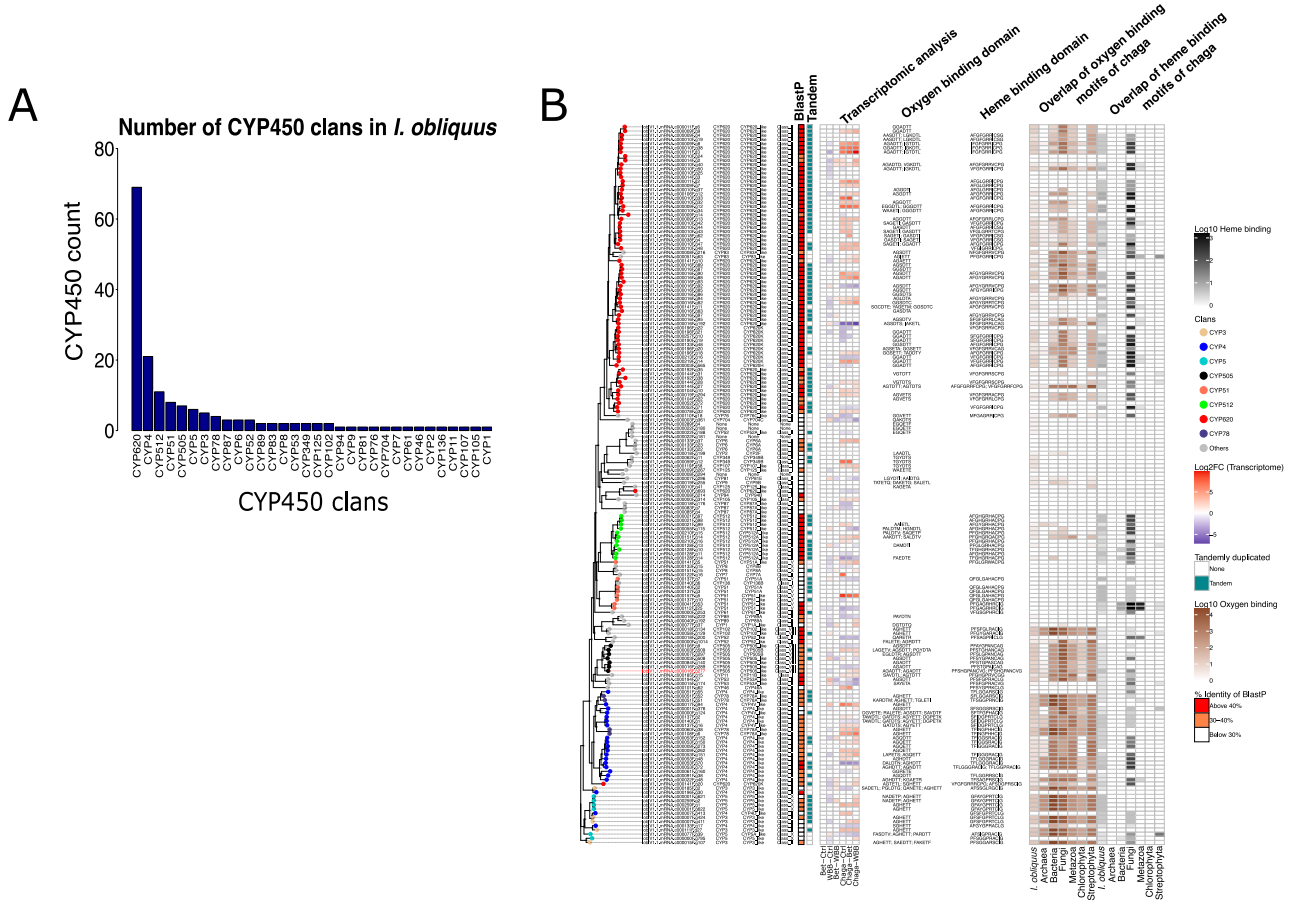

Fig. 5. Barplot of gene counts and the gene tree of cytochrome P450s predicted in *I. obliquus*. (**A**) Barplot shows the counts of P450 monooxygenase enzymes per family in *I. obliquus* genome. (**B**) Gene tree of P450 monooxygenase enzymes. Each leaf is highlighted with a point colored according to the major family of the enzyme. The sequences are labeled with gene ID, P450 families, family, and class. The heatmaps to the right of the tree indicate BLAST similarity, whether the gene is tandemly duplicated, differential expression in different RNA-seq experiments, and the motif present in conserved oxygen- and heme-binding domains. The last two columns show the $\log_{10}$-scaled count of chaga oxygen- and heme-binding domains and their overlap with motifs found in different kingdoms and taxa.

I. obliquus. The longer length was due to the presence of two characteristic P450 domains in the encoded protein, including both the oxygen-binding and heme-binding domains. These results suggest this to be a novel protein created by a fusion of two genes; the predicted gene model sequence was subsequently confirmed with PCR.

We analyzed the evolution of the fused CYP505 gene by performing microsynteny analyses against four Hymenochaetales species, *F. betulina*, and *S. paradoxa*. A homolog of the putative fusion gene with similar structure (gene_7933, family CYP505) was present in *F. mediterranea* (Fig S4), whereas other Hymenochaetales species did not possess a gene with such organization. This suggests that the fusion gene may have arisen from a non-homologous recombination event in the common ancestor of *F. mediterranea* and *I. obliquus* and after the divergence of *P. niemelaei*, in which the two fused P450s were still found separate. A dotplot comparison of the fusion gene models (Fig S5) showed the second gene model to be highly conserved between the two species, whereas the first gene model showed a greater level of sequence divergence. Here, the gene model evolution may have involved small-scale deletion events, since the first 200 bases were deleted from the beginning of the *F. mediterranea* gene model, whereas in *I. obliquus* there was a deletion in 3' end.

As fungal CYPs have been found to be multifunctional and enzymatically promiscuous[60], we tested the ability for betulin production. Our previous work suggested a candidate CYP716 gene (Bpev01.c0219.g0021) responsible for BE biosynthesis in *B. pendula*[10], and here we constructed a single insert expression vector for this gene (pRS424::CYP716). In addition, we also constructed a double insert vector including a lupeol synthase from *B. pendula*[10] in the second cloning site (pRS424::LUS-CYP716); this was then separately transferred into yeast. Heterologous expression of the single insert vector (when spiked with 98% lupeol) yielded a higher concentration of BE than the expression of the double-insert vector; these differences can be explained by the low amount of lupeol production from the double-insert vector. Within each sample, yeast cells exhibited higher concentrations of BE compared with growth media (Fig. 9). Our mass-spectrometry analysis of heterologous expression of CYP716 from silver birch thus confirms the function of the candidate gene in BE production.

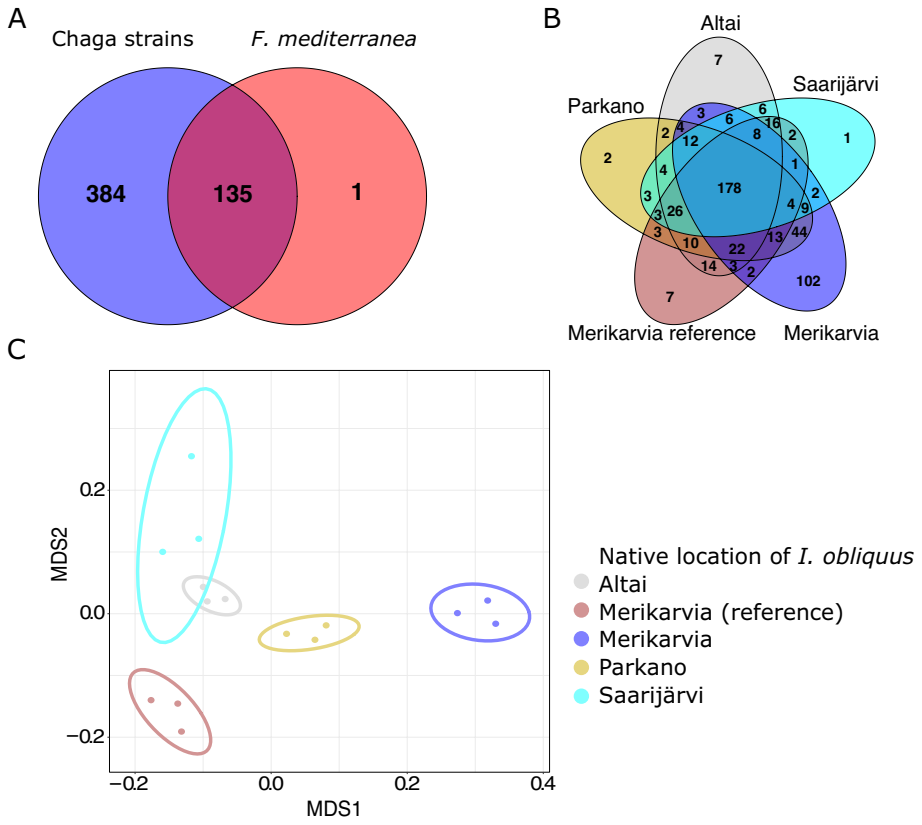

Fig. 6. Venn diagrams of UPLC-MS mass spectrum for metabolomic fingerprints. (**A**) Venn diagram illustrating the overlap between peaks found in pooled mass spectra from five strains of *I. obliquus* and one *F. mediterranea*. (**B**) Venn diagram illustrating the overlap of the mass spectrum peaks detected from five strains of *I. obliquus*. (**C**) Multidimensional scaling (MDS) plot of metabolite abundances of five strains of *I. obliquus*.

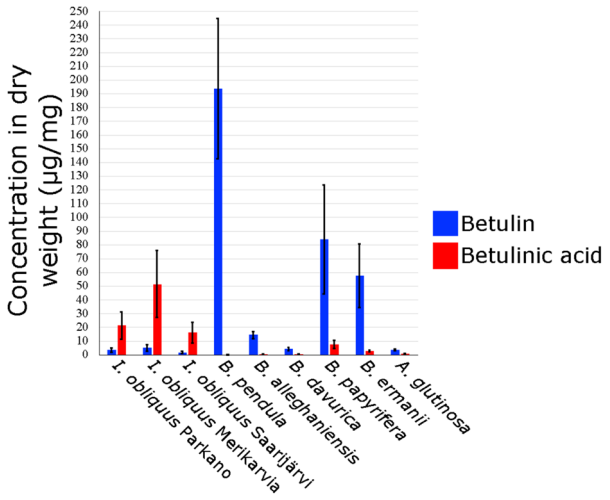

Fig. 7. Concentrations of BE and BA in three strains of chaga and six *Betula* species assessed using HPLC–MS.

Using the birch gene as a positive control, the c000016F.g277 gene (CYP505 family) from chaga was tested for BE synthesis activity. However, only weak production of BE compounds was detected (Fig. 9).

Previous studies using a yeast expression model have revealed a CYP505 family enzyme to be involved in the production of squalene-type triterpenoids[57,58], which further supports CYP505 as a possible candidate for BE biosynthesis. If that is the case, considering that 73% of deposited protein sequences annotated as CYP505

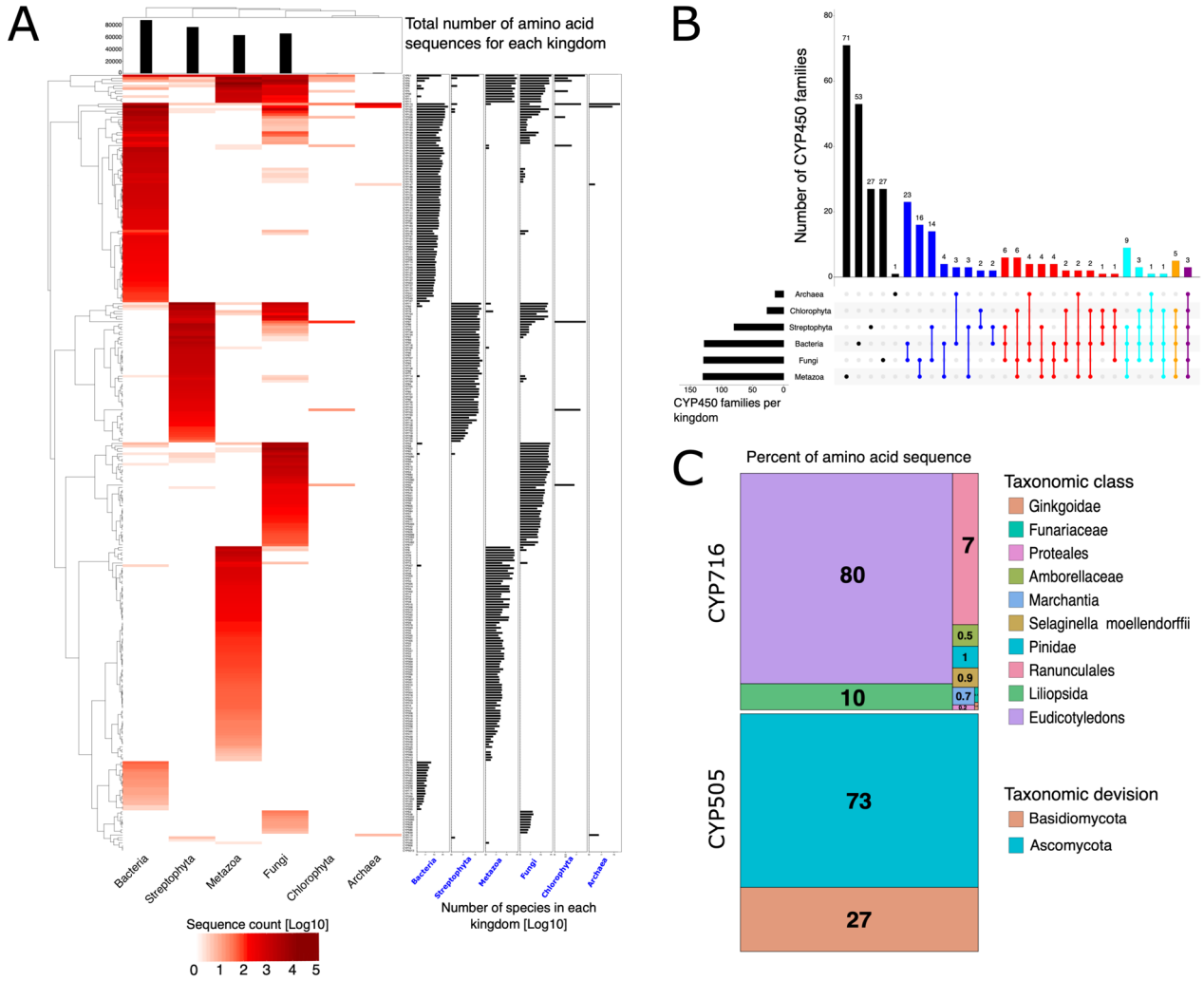

Fig. 8. Summary of all predicted cytochrome P450 (pfam ID: PF00067) amino acid sequences. (**A**) Heatmap of P450 families for different kingdoms and taxa. Rows indicate P450 families and columns indicate kingdoms and taxa. The color palette illustrates the log10-scaled protein sequence counts of P450 families in each of the kingdoms and two relevant divisions (see the color key below). Barplot above the columns illustrates the total number of proteins annotated to the P450 families in each of the kingdoms and divisions. (**B**) Upset plot of P450 families in different kingdoms and phyla. (**C**) Treemaps illustrating the percent of proteins annotated as CYP716 in different classes of streptophyta and CYP505 from the phyla of *Ascomycota* and *Basidiomycota*. Rectangle area reflects the percent of protein sequences predicted in the different taxa.

family members have been identified from *Ascomycete* species (Fig. 8C), it is likely that more *Ascomycete* species harboring betulinate biosynthesis pathways will be identified.

Gene expression illustrates high activity of the majority of P450 genes

The evolution of the P450 family in chaga, as observed from gene family expansions, an increased number of biosynthetic gene clusters, and novel gene structures such as the emergence of a novel fusion gene from the CYP505 family, suggests consistent positive selection for increased biochemical diversity. One possible explanation for the selection pressure is increased BE biosynthesis in the host species, which may have been driving the diversification of secondary metabolism in chaga. In addition to the putative evolution of CYP505 family members to metabolize BE, we also identified a major tandem expansion of CYP620 family genes. Since tandem duplications are associated with rapid evolutionary adaptation to changing environmental conditions, this expansion suggests that this family may also have a role in metabolizing BE.

To identify novel candidate BE biosynthesis genes, we reanalyzed transcriptomic profiling data from different culture media with and without BE and media with birch bark residues[41], and included one transcriptomic dataset where the media contained wood dust from *B. pendula*. Altogether, 119 P450 genes were significantly upregulated in at least one experiment (Fig. 5). Three key enzymes along the mevalonate pathway were also differentially expressed (Fig. 10, Table S18), and a pair of tandemly duplicated lupeol synthases had the highest expression level in comparison with other lupeol synthases. In agreement with earlier results[41], chaga samples

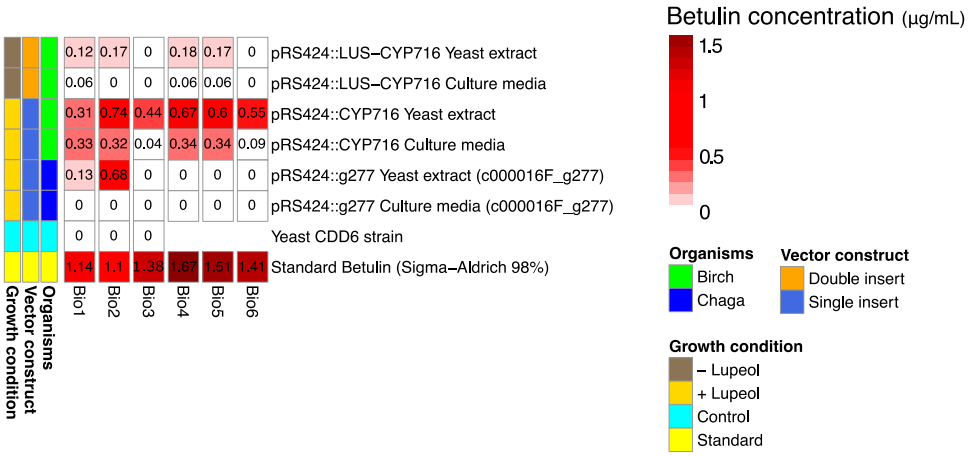

Fig. 9. Heatmap of BE biosynthesis concentration in yeast as a heterologous host. Columns of the heatmap indicate different biological replicates (Bio1-6). Rows indicate different expression vectors, such as the single vector construct of CYP716, double-insert vector construct of lupeol synthase and CYP716 genes, standard betulin, and CDD6 yeast strain as control. Species source for lupeol synthase and CYP716 was silver birch (*B. pendula*). c000016_g277 gene was extracted from chaga. Vector constructs are divided to double inserts (orange, lupeol synthase and CYP716) and single inserts (royal blue). The color palette of the heatmap illustrates the concentration (µg/mL) of BE found in yeast cells (Yeast extract) and yeast growth media (Culture media), with white color as minimum (zero) and dark red as maximum concentration of BE. Heatmap is also annotated according to the growth condition; brown (- Lupeol) is yeast growth culture without added lupeol (metabolic precursor to be potentially modified by CYP716 and CYP505 genes), gold (+ Lupeol) is yeast growth culture spiked with lupeol, cyan is CDD6 yeast strain as control, and yellow is BE standard (98% purity) obtained commercially.

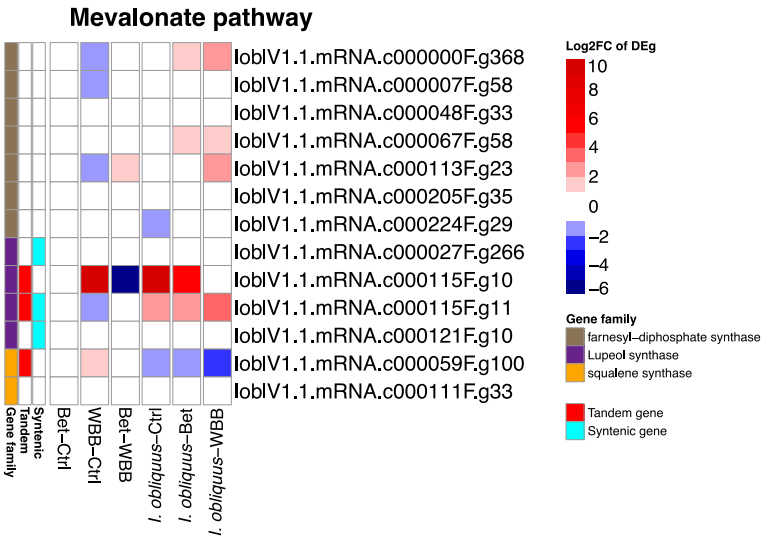

Fig. 10. Heatmap of differentially expressed (DE) genes homologous to terpenoid biosynthesis enzymes that act on substrates from the mevalonate pathway (MVA). Three main enzymes are highlighted with distinct colors. The color palette of the heatmap corresponds to the log2 fold-change (log2FC) of the DE genes. The gene families are indicated by duplication origins, syntenic for genes originating from whole-genome duplication events, and tandem for tandemly duplicated genes.

grown on a full block of *B. pendula* wood dust exhibited a higher number and elevated expression levels compared with samples grown in culture media spiked with birch bark residues or BE. These results suggest that the activation of the full enzymatic palette requires the presence of the host wood tissue. Therefore, it is possible that several P450 enzymes contribute to betulinate metabolism.

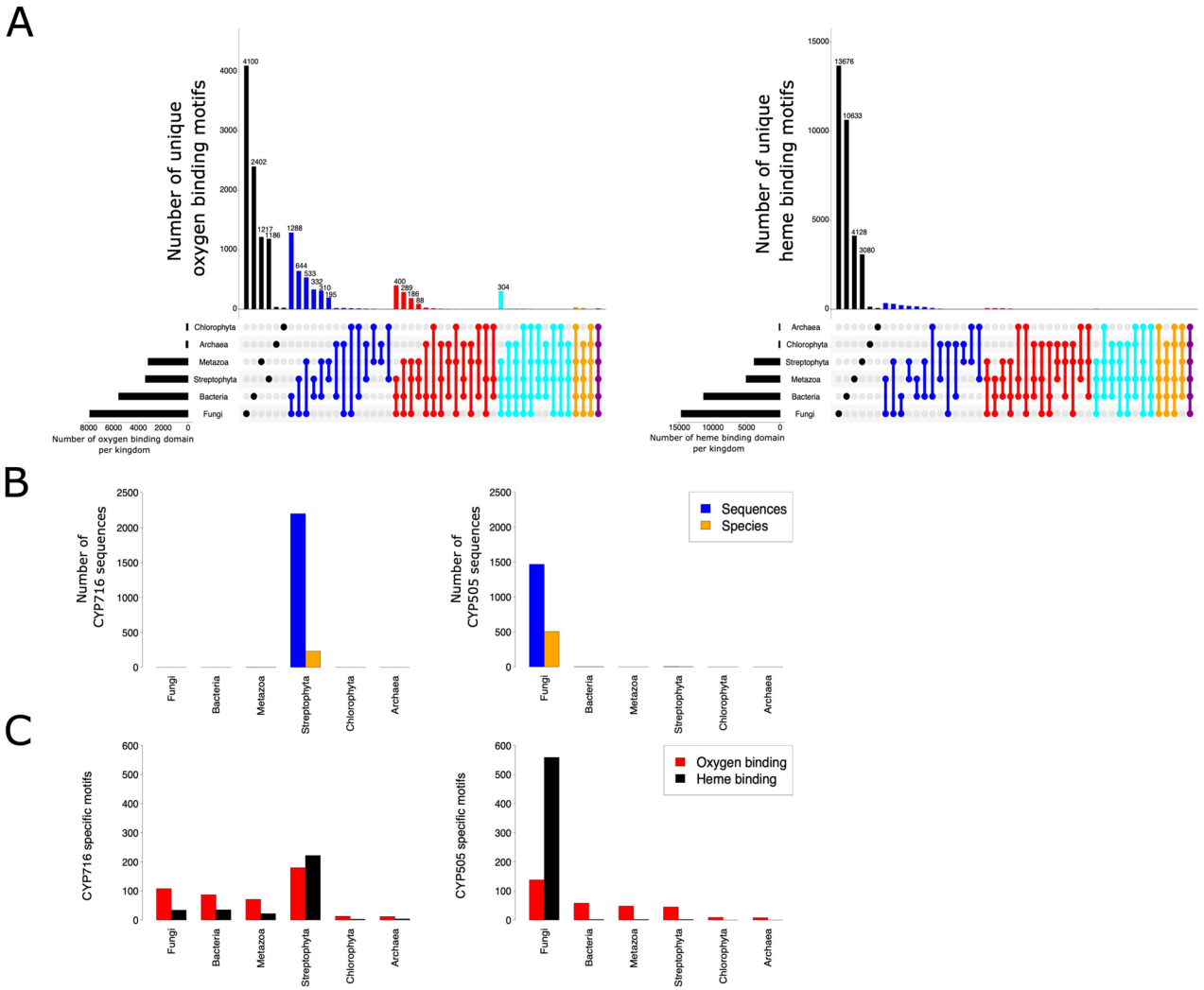

Fig. 11. Detailed analysis of CYP716 and CYP505 monooxygenase families. (**A**) Upset plot of oxygen- and heme-binding domains in different kingdoms and taxa. (**B**) Total number of proteins that were annotated as CYP716 and CYP505 and the total number of species containing CYP716 and CYP505 families in different kingdoms and phyla. (**C**) Numbers of oxygen- and heme-binding domains found from CYP716 and CYP505 across different kingdoms and taxa.

Kingdom-wide evolutionary analysis of cytochrome P450 genes shows no evidence of horizontal transfer in chaga

To study the evolutionary origin of BE biosynthesis in fungi, namely if it has been transferred from a plant host species or evolved independently, we performed a global analysis of P450 families across all kingdoms of life by annotating all 337,360 predicted P450s in NCBI GenBank using BioCatNet (CYPED v6.0) database[61]. Three P450 families (such as CYP174, CYP51, CYP704) were common to all six kingdoms and phyla, and compared with other kingdoms, fungi and bacteria had a higher number of shared families (Fig. 8, Table S19). However, many families were found unique to their respective kingdoms and phyla (Fig. 8, Table S19), and in particular, CYP716 was confirmed as plant-specific (Fig. 8, Table S19). We identified altogether 2200 sequences from 230 species of Streptophyta with high similarity to CYP716 (with 75.6% median sequence identity) (Fig. 11). Nearly 87% of the annotated CYP716 sequences were from eudicot species and 10% from monocots (Fig. 8, Table S19). No CYP716 was found in fungal species; the best matching 18 sequences from 14 species had only 26.4% median sequence similarity to CYP716 family representatives, which is much lower than the minimum 40% similarity threshold for family-level assignment. Similarly, assignments to the CYP505 (median percent identity: 35.6%) family consisted only of sequences from fungal species (altogether 506 species) (Fig. 11).

Since the production of BE and BA has been confirmed in two pathogens of *Betulaceae* (*I. obliquus* and *F. betulina*), it is still possible that the enzymes responsible for betulinate production were introduced into the common ancestor of these pathogens, and the transfer event is not identifiable due to rapid molecular evolution in fungi, which may have led to highly diverged protein sequences. Due to their functional importance, the active domains in P450 (such as oxygen- and heme-binding domains and the ERR triad domain) (Table S16, Table S17, Table S20) are highly conserved[62]. To conclusively exclude the possibility of horizontal transfer, we

analyzed the conserved oxygen and heme binding motifs in all predicted P450 genes. These motifs were found to be highly conserved across kingdoms and phyla, suggesting them to be good markers for horizontal transfer. The motifs of the oxygen-binding domains in all CYP716 and those from CYP505 in fungi showed considerable overlaps between other kingdoms and phyla (Figs. 5, 11, Table S16), whereas the heme-binding motifs were mostly kingdom- and phyla-specific (Fig. 11, Table S17), suggesting that this pattern has a more reliable signature for identifying the evolutionary origins of P450 enzymes. The heme-binding domain has a critical role in substrate recognition, impacting the functionality of P450 monooxygenase[63], which is why purifying selection is required. In case of chaga, the motifs were specific to fungi and thus did not support transfer events from plants. Therefore, the P450 enzymes responsible for betulinate biosynthesis, in both *Basidiomycete* species (*I. obliquus* and *F. betulina*), are the result of convergent evolution. BE production has also been reported in other fungal species, such as marine *Paecilomyces* WE3-F[22](phylum *Ascomycete)*; since the species has deep evolutionary divergence and a very different habitat from the *Basidiomycete* species examined here, this suggests a second case of independent evolution of betulinate biosynthesis in fungal species. Moreover, comparative analysis did not identify any shared heme-binding motifs in CYP505s predicted in *I. obliquus* and *F. betulina*, whereas we observed 27 CYP620s from chaga have similar heme-binding motifs with 16 CYP620s from *F. betulina*.

Conclusion

We observed complex terpene metabolism in *I. obliquus* that has evolved both by preferential retention of genes following a WGD event and tandem duplications involving the P450 gene family. Our metabolomics fingerprint confirmed a highly diverse palette of terpenoid derivatives. Two alternative hypotheses for the evolution of betulinate biosynthesis in the fungus exist, namely a horizontal transfer event from the plant host to the pathogenic fungus with subsequent enzymatic diversification, or alternatively metabolic coevolution in chaga as a response to the host protective mechanisms. Sequence homology assessments combined with conserved motif analysis did not reveal evidence supporting horizontal gene transfer (HGT) between *B. pendula* and *I. obliquus*. The lack of significant sequence similarity beyond what is expected from conserved cytochrome P450 domains, along with the phylogenetic consistency of these genes within fungal lineages, strongly suggests that the enzymatic pathway to betulinate biosynthesis in *I. obliquus* evolved independently. These findings support a model of convergent evolution rather than HGT.

The relatively recent WGD in *I. obliquus* (~1.3 Mya), occurring well after the diversification of *Betula* species (~8.4 Mya), implies that the expansion and specialization of secondary metabolism genes, including cytochrome P450 monooxygenases implicated in triterpene biosynthesis, may have been driven by long-standing ecological and biochemical interactions with its host. This evolutionary timeline is consistent with convergent metabolic adaptation, where host-derived compounds (e.g., betulin) acted as selective forces for the retention and functional innovation of paralogous genes post-WGD. As birch trees developed increasingly potent triterpenoid-based defenses, chaga adapted through the expansion and specialization of secondary metabolic pathways likely facilitated by a recent whole genome duplication. This co-evolutionary feedback loop illustrates the perpetual molecular arms race characteristic of host–pathogen interactions, where both partners must continually evolve to maintain ecological advantages such as competitive dominance of chaga against other opportunistic birch pathogens. The evolutionary trajectory of triterpene biosynthesis in *I. obliquus* may thus represent a textbook example of Red Queen hypothesis[64].

Further experimental work is needed to identify the specific enzyme(s) responsible for BE biosynthesis from the pool of 176 candidate genes in *I. obliquus*. Based on our mass spectrometry analysis, we propose that the P450 enzyme responsible for producing betulinate compounds in chaga may have evolved distinct biochemical characteristics, through positive selection events, compared to the analogous plant enzyme. This hypothesis is supported by the higher concentrations of betulinic acid (BA) found in the chaga sample. These findings could have significant biotechnological applications, as BA is a highly desirable pharmaceutical compound.

We additionally detected a bloom of genes encoding CYP620 family enzymes, which may also confer BE biosynthesis activity. The low sequence similarity and high functional diversity among P450 enzymes make functional predictions difficult and require experimental confirmations, and the high number of duplicates (71) in chaga will make the task challenging. Thus, the integration of multi-omics data, such as metabolomics and transcriptomics data, and eventual enzymatic confirmations in vitro, are necessary to correctly identify the wide range of secondary metabolites and the enzymes associated in their biosynthesis in different species.

Materials and methods

Sample collection

Four *Inonotus obliquus* strains (Table S14) were isolated from different regions of Finland (Merikarvia, Parkano, and Saarijärvi municipalities). One strain was isolated from Altai Mountain in Russia. The strain Pakuri-i from Merikarvia was selected for whole genome sequencing (location 61°58′38.6"N 21°44′43.1"E). In addition, we also obtained a strain of *Fomitiporia mediterranea* as an outgroup to chaga (Mycobank: MB384943). All samples were cultivated on Hagem agar overlayed by a cellophane membrane. Composition of Hagem media is presented in (Table S14).

Isolation of chaga mushrooms from host trees was performed by cutting a piece of the conk (Fig S6), which was then laid on an agar plate after a short H_2O_2 bath. The samples were re-cultured repeatedly and sequenced for internal transcribed spacer 1 (ITS1) [TCCGTAGGTGAACCTGCGG] and ITS4 [TCCTCCGCTTATTGAT ATGC] regions to confirm the species assignment of the *I. obliquus* isolate.

RNA isolation, sequencing, and de novo assembly of transcriptome

The method from Chang et al. [65] was used to isolate total RNA from *I. obliquus*. Briefly, *I. obliquus* was inoculated and grown on autoclaved wood dust from a clone of *B. pendula* (SB1, 12-year-old tree, 167 cm^2 disk, dry weight 200 g) sequenced for *B. pendula* reference genome[66]. A total of 150 mg of ground sample (mortar and pestle and liquid N_2) was transferred on ice for 30 s, and 500 µl of pre-warmed (65–68 °C) extraction buffer (2% CTAB, 2% PVP K-30, 100 mM Tris–HCl [pH 8.0], 25 mM EDTA, 2 M NaCl, 200 µl β-met/10 ml extraction buffer) was added and vortexed vigorously. Extraction was performed three times with chloroform:isoamyl alcohol (24:1) by centrifugation at 200–300 rpm for 15 min followed by 15 min at 10 000 rpm. Then, 1/4 volume 10 M LiCl was added and left to precipitate on ice overnight. The overnight sample was centrifuged at 10 000 rpm for 20–30 min at 4 °C. The resulting pellet was dissolved in 500 µl of pre-warmed (65 °C) sodium dodecyl sulfate-Tris–HCl-EDTA (SSTE) buffer and extracted once (or several times, if necessary) with chloroform:isoamyl alcohol (24:1). The mixture was precipitated by adding 2 volumes of absolute EtOH incubated at −20 °C overnight and centrifuged at 13 000 rpm for 20–30 min at 4 °C. The precipitate was washed with 70% EtOH, after which the pellet was dried and dissolved in 10–30 µl RNase-free water. RNase inhibitor was then added.

TruSeq stranded mRNA kit was used to construct the RNA-seq library. cDNA was synthesized from 5 µl of total RNA extracted from the reference *I. obliquus* plate using random hexamers. DNA polymerase I and dUTP nucleotides were used to synthesize the second strand of cDNA. Double-stranded cDNA was then purified and ends were repaired. Library preparation was continued by A-tailing and ligation of Y-adaptors containing indexes from the kit. The fragments were amplified using PCR followed by purification steps using AMPure XP. Sequencing was performed on a HiScan SQ platform (paired-end 88 bp + 74 bp).

The raw paired-end RNA-seq data were controlled for quality using FastQC v0.11.2[67]. Trimmomatic v0.33[68] was used in pair-end mode to remove the adapters, barcodes, low-quality bases from both ends of each sequence, and reads shorter than 25 bp (LEADING:20, TRAILING:20, MINLEN:25, -phred33). After removal of duplicate sequences, the unpaired sequences were mapped to *I. obliquus* reference genome using Tophat2 v2.1.2[69] for junction discoveries (-i:10, and –coverage-search); paired end reads were mapped separately (Tophat2; -i:10, and –coverage-search). The aligned reads were separated according to their orientation on reference genome to forward and reverse strands, which were then aligned individually by Trinity v2.1.1[70] using –genome_guided_bam and –genome_guided_max_intron: 1000 options for de novo transcriptome assemblies. The forward and reverse de novo transcriptome assemblies were combined. Duplicated assemblies were removed using GenomeTools v1.5.1[71] using sequniq option. The unique de novo transcriptome assemblies were clustered by using CD-HIT v4.6[72] and aligned to the *I. obliquus* reference genome by Program to Assemble Spliced Alignments (PASA) v2.2.0[73].

Processing of the publicly available RNAseq data[41] was performed in an analogous manner. Both data sets were mapped to *I. obliquus* gene models using kallisto quant v0.44.0[74]. The orphan reads and pair-end reads (separated during preprocessing by trimmomatic) were mapped separately by using kallisto v.0.46.0 quant single (options: –single, -l 200, -s 20, -b 4000) and pair-end (option: -b 4000) modes, respectively. The raw count table from Kallisto was imported to R (for both single and pair-end count tables) using tximport package v1.18.0[75] with default options. The single and pair-end counts were summed together to form a single count table for each data set. Differential gene expression analysis was conducted using DESeq2 package[76]. The final tables for differentially expressed genes (DEg) were filtered based on the false discovery rate adjusted *p*-value threshold of 0.05 (p-adj. ≤ 0.05).

DNA isolation, genome assembly, and annotation

A modified version of the method described in Lodhi et al.[77] was used for DNA extraction from *I. obliquus* strains. A maximum of 0.5 g of material was ground in liquid nitrogen. The ground sample was transferred into ice-cold NaCl-Tris–EDTA (STE) buffer (1.4 M NaCl, 0 mM EDTA, 100 mM Tris–HCl pH 8.0) and centrifuged at 8000 rpm for 5 min at 4 °C. STE buffer was discarded and 10 ml of pre-warmed (60 °C) cetyltrimethyl ammonium bromide (CTAB) buffer (20 mM EDTA, 100 mM Tris–HCl pH 8.0, 1.4 M NaCl, 2.0% CTAB, 1.0% PVP 40, and 2% β-MeOH) was added to the pellet. The mixture was then vortexed and incubated for 30–60 min at 60 °C and cooled to room temperature. A 24:1 chloroform:isoamyl alcohol (IAA) mixture was added for extraction followed by centrifugation at 10 000 rpm for 15 min at room temperature. The supernatant was collected into a new tube and mixed with 2 × CTAB buffer, vortexed, and incubated for 30–60 min at 60 °C. The chloroform:IAA extraction step was repeated 2–3 times, followed by addition of 2 × volume of cold (−20 °C) absolute EtOH to the supernatant. The EtOH mixture was stored overnight at 4 °C. The mixture then was centrifuged at 10 000 rpm for 15 min at 4 °C. The DNA pellet was washed with absolute EtOH (−20 °C) and air dried. The sample was treated with RNase followed by chloroform:IAA extraction. The sample was then precipitated with EtOH, air dried, dissolved in DNase/RNase free water, and stored at -80 °C.

The genome of *I. obliquus* was sequenced with a Pacific Biosciences PacBio RSII instrument using P6-C4 chemistry. Eight SMRTcells were used to sequence the sample with movie time of 240 min. The number of obtained sequences was 712 759, which totaled up to 4.82 Gb of data with read length N50 of 9200 bp. Hierarchical Genome Assembly Process (HGAP) V3 implemented in SMRT Analysis package (v2.3.0) was first used to generate an initial de novo genome assembly with default parameters. The mitochondrial genome contig was separated from the chromosomal contigs and circularized manually using GAP4 v2.0.0.b11[78]. The obtained mitochondrial sequence length of 118 085 bp and > 4000 × sequencing coverage was polished using SMRT Analysis RS Resequencing protocol with Quiver consensus algorithm. The FALCON assembly program[79] was then used to generate the final de novo genome assembly with seed read length of 10 000 bp. The obtained contig sequences were polished using SMRT Analysis RS Resequencing protocol with Quiver consensus algorithm with approximately 75 × coverage. BUSCO v3.0[29] (Fungi datasets, -m geno, -long) was used to quantify the completeness of the genome.

Repeat analysis of the contigs was performed according to the guidelines of RepeatModeler and RepeatMasker v4.0.7 (http://www.repeatmasker.org/). To predict the gene models, the following multiple evidence tracks from different platforms were obtained: ab initio gene predictors based on Hidden Markov Models (HMMs), spliced transcript evidence from RNA-seq, and orthologous proteins from closely related fungal species. HMM-based models, such as AUGUSTUS v3.3.2[80] and GeneMark-ES v4.33[81] (–fungus mode, and –evidence: de novo transcriptome assembly) were used for ab initio gene predictions. In addition, BRAKER2 v2.1.5[82] (options: –fungus, –rounds = 100, and –bam) was run for ab initio gene predictions. To identify the ORFs within the genome, getorf (EMBOSS v6.6.0) program[83] was used (-find:1, and –maxsize: 5000). The ORFs were then queried against NR database by DIAMOND v0.9.24[84] (blastp, –more-sensitive) and filtered for similarity (sequence identity ≥ 75, and score ≥ 300); the homologous sequences above the threshold were collected. Selected ORFs were used as the input for exonerate v2.46.2[85] (–model:protein2genome, –minintron:10, –maxintron:1000; –percent:65) to map the candidate ORFs to *I. obliquus* reference genome. Additionally, orthologous proteins from 13 fungal species (*Coprinopsis cinerea*, *Fomitiporia mediterranea*, *Heterobasidion annosum*, *Laccaria bicolor*, *Onnia scaura*, *Phanerochaete chrysosporium*, *Phellinus ferrugineofuscus*, *Porodaedalea niemelaei*, *Postia placenta*, *Puccinia graminis*, *Rickenella mellea*, *Schizopora paradoxa*, *Trichaptum abietinum*) were aligned against *I. obliquus* reference genome with exonerate v2.46.2 with options –model:protein2genome, –minintron:10, –maxintron:1000; –percent:65. In addition to orthologous proteins, the protein sequences discovered from BUSCO predictions were collected and aligned to the reference genome using exonerate with the same parameters as described above. All the evidence (ab initio gene models, spliced transcript alignments, spliced protein alignments, ORFs, and BUSCO) was combined to consensus, high-confidence gene models using EVidenceModeler v1.1.1[73]. This was followed by addition of untranslated regions (UTR) to the gene models by PASA v2.4.1[73].

The mitochondrial genome was assembled using modified Newbler assembler v2.8[86] and annotated as described previously[66] using Mitofy (http://dogma.ccbb.utexas.edu/mitofy/).

Interproscan v5.25–64.0[87] was used to assign the protein function to gene models. Additionally, Ensemble Enzyme Prediction (E2P2) v3.1[88] and antiSMASH v2.0 fungal version[59] were used to predict the metabolomic pathways.

Comparative genomic analyses

The proteomes of 20 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 using Orthofinder v2.3.3[43] run with default parameters, and using RAxML and MAFFT for obtaining protein trees (-T raxml -M msa -t 50). A phylogenetic tree was then estimated from 4400 protein sequences identified to be present as single copies. Alignment of protein sequences was performed using MAFFT (–auto –retree 10,000 –maxiterate 10,000) and tree was estimated using RAxML (Best tree and bootstrap trees: -N = 1000,-p = 20,000, -b = 20,000; bipartitioning: –bootstop-perms = 1000).

Synteny analyses

Synteny analysis of self-self alignment of *I. obliquus*, four other fungal species (*F. mediterranea*, *S. paradoxa*, *F. betulina*, and *P. niemelaei*) and alignments against previously published *I. obliquus* and *I. hispidus* assemblies were conducted using SynMap application in CoGe platform (https://genomevolution.org/coge/) and Quota Align algorithm with default parameters. The list of syntenic duplicates was obtained from DAGchainer[89]; tandem duplicates were obtained as part of the preprocessing pipeline.

Discovery of candidate effector proteins and carbohydrate active enzymes (CAZymes)

The Getorf function of EMBOSS v6.6.0[83] was used to discover the ORFs (-find:1, and –maxsize: 1000). All ORFs were analyzed by SignalP v5[9085] for the presence of signal peptide. Signal peptides were removed from predicted ORF sequences. Cysteine amino acids were counted for every sequence. An ORF sequence with three or more cysteine residues was considered a cysteine-rich short secreted protein (CSSP).

CAZymes were annotated during gene model annotation steps. To further classify these enzymes, the total proteome of *I. obliquus* was queried against dbCAN2 database[91] using DIAMOND v0.9.24[84] (blastp, –more-sensitive), and the best hit was selected (score ≥ 200, percentage identity ≥ 55) as a homologous sequence.

Annotation, motif discovery, and phylogenetics of cytochrome P450 monooxygenases

All proteins in Uniprot containing a PF00067 domain were assigned to the P450 monooxygenase family and downloaded. Proteins < 200 amino acids were removed from the dataset, resulting in 337 365 proteins. Blastp v2.10.1[92] was used to assign family and sub-family information to the sequences using BioCatNet (CYPED v6.0) database[61] as a reference. The sequence similarity threshold was set to ≥ 35% and the best hit to closest species (-max_target_seqs 1) was selected for each query sequence. The conserved protein domains of P450 monooxygenase enzymes, namely oxygen- ([A,G]G[A-Z][E,D]T[T,S]) and heme- ([A-Z]F[A-Z][A-Z]G[A-Z]R[A-Z]C[A-Z]G) binding, and ERR triad domains ([A-Z]E[A-Z][A-Z]R and PER[A-Z]) were collected from literature[27,60,93,94] and identified using a custom bash shell script.

A phylogenetic tree was first constructed for 13 Streptophyta CYP716 proteins that have been experimentally confirmed to produce betulinate compounds, including *Betula pendula*[66], *Betula platyphylla*[95], *Phoenix dactylifera*[96], *Medicago truncatula*[97], and *Vitis vinifera*[98]. P450 amino acid sequences with high sequence similarity (> 40%) in *I. obliquus* and *F. betulina* were then included in the analyses. A total of 77 sequences were

aligned using MUSCLE v3.8.31[99] (-maxiters: 1000). The aligned amino acid sequences were reverse translated by PAL2NAL v14.1[100]. Both amino acid and nucleotide sequences were used to estimate the phylogenetic trees using RaxmlHPC-HYBRID-AVX2 v8.2.12[101].

I. obliquus and *F. mediterranea* metabolomics

Five strains of *I. obliquus* and one *F. mediterranea* (three biological replicates for each) were grown in liquid Hagem media for 2 weeks. Submerged mycelia were washed with sterile milli-Q water three times and ground with liquid nitrogen. All samples (fresh weight approximately 500 mg, dry weight approximately 30 mg) were extracted twice with 1.0 ml of ethyl acetate (Merck) by vortexing for 15 min and followed by centrifugation at 15 000 rpm for 10 min according to Zhao et al.[102] at room temperature. The internal standards (ISTD) (10 µl, 10 µg/ml) testosterone and 4-methylumbelliferone (Sigma) were added to each sample in the first extraction step. The supernatant was evaporated to dryness with a MiVac Duo concentrator at 40 °C (GeneVac LtD, Ipswich, UK). The residue was re-solubilized in 100 µl ACN (Honeywell). A quality control (QC) sample was prepared by combining extracts from each sample line.

Triterpenoid profiling was performed using extracts with UPLC-PDA-QTOF/MS. The UPLC-MS system consisted of a Waters Acquity UPLC attached to an Acquity PDA detector and to a Waters Synapt G2 (HDMS) QTOF mass spectrometer (Waters, Milford, MA, USA). Separation of the analytes was performed in an Acquity BEH C18 (2.1 mm × 50 mm, 1.7 µm) column (Waters, Milford, MA, USA) at 40 °C. The autosampler temperature was set to 27 °C. The mobile phase consisted of water (A) and acetonitrile (B) both with 0.1% formic acid. Flow rate was 0.6 ml/min, and the injection volume set to 3 µl. The linear gradient started with 30% B and proceeded to 98% in 9 min, followed by 1 min at 98% B, giving a total run time of 10 min. ESI/MS detection was performed in positive sensitivity ion mode with capillary voltage 3.0 kV, cone voltage 30 V, desolvation gas 800 L/h, cone gas 20 L/h, desolvation temperature 320 °C, source temperature 120 °C, and extractor lens 3.00 V. MarkerLynx XS V4.1 software (Waters, Milford, MA, USA, https://www.waters.com/) was used for data processing. UV spectra, negative MS runs, and fragmentation patterns from MS^e runs of QC sample were used as additional tools for annotation of triterpenes, sterols, and phenolic compounds in *I. obliquus* samples.

Standard solutions of BE and BA (1.0 mg/ml) were prepared in ethyl acetate. Testosterone (100 µg/ml) standard was prepared in methanol. Working solutions (10 µg/ml, 100 ng/ml) were prepared by diluting the standard solution with acetonitrile. Optimization of the quantification method was performed with a mix of BE, BA, and testosterone standards (100 ng/ml). Standard mix (100 µl) was derivatized with PTSI. MS parameters were optimized by repeated injection of the sample.

Due to extremely low concentrations of BE and BA, extracts were derivatized with p-toluenesulfonyl isocyanate (PTSI)[103,104] to improve sensitivity. After metabolite profiling with UPLC-QTOF/MS, samples (90 µl) were derivatized for 3 min with 10 µl of 60% p-toluenesulfonyl isocyanate (PTSI) (Sigma) in acetonitrile[103,104]. The derivatization reaction was terminated with 50 µl of methanol (Merck) with 30 s vortexing, yielding the total volume of 150 µl.

The UPLC-MS/MS system consisted of an UPLC (ABSciex, Shimadzu) attached to ABSciex 6500 + QTRAP mass spectrometer with ESI source. An Acquity BEH C18 (2.1 mm × 50 mm, 1.7 µm) (Waters, Milford, MA, USA) column was used for separation of compounds. Column oven temperature was 40 °C. Autosampler temperature was set to 25 °C. Injection volume was 2 µl. The mobile phases were water (A) and acetonitrile (B) both with 0.1% of formic acid. The flow rate was 0.6 ml/min. The linear gradient started from 30% B and proceeded to 98% in 6.5 min, followed by 1.5 min at 98% B, giving a total run time of 8 min. The data was normalized to dry weight (DW) and to the peak area of internal standard. Analyst v1.6.3 software (ABSciex, https://sciex.com/) was used for data processing and quantification.

PTSI derivatization reagent generated betulin p-toluenesulfonyl carbamic diester, betulinic acid p-toluenesulfonyl carbamic ester, and testosterone toluenesulfonyl carbamic ester. Two MRM transitions were selected for each analyte, one for quantification and the other for qualification. The ratio between quantification (quan) and qualification (qual) transitions should stay stable among runs. The transitions were as follows: betulin MRM 835.3→620.3 quan $[M\text{-}PTSI\text{-}H_2O\text{-}H]^-$, 835.3→638.3 qual $[M\text{-}PTSI\text{-}H]^-$, betulinic acid MRM 652.3→455.2 quan $[M\text{-}PTSI\text{-}H]^-$ 652.3→437.2 qual $[M\text{-}PTSI\text{-}H_2O\text{-}H]^-$, and testosterone 484.2→287.2 quan $[PTSI\text{-}H]^-$, 484.2→269.2 qual $[PTSI\text{-}H_2O\text{-}H]^-$. ESI source temperature was set to 450 °C. ESI/MS/MS detection was performed in negative ion mode with ion spray (IS) voltage of −4000, curtain gas (CUR) 30, collision gas (CAD) at medium, entrance potential (EP) -10, declustering potential (DP) -60 (betulinic acid, testosterone) or -100 (betulin), collision energy (CE) -50, and collision cell exit potential (CXP) -10.

Cloning and mass spectrometry of lupeol synthase and CYP716 genes from *B. pendula*

Two major cloning constructs were designed for BE biosynthesis using pRS424 vector[105]. This version of pRS424 vector contains two cloning sites, one under control of the GAL1 promoter and the other under GAL10. The first vector was designed as a single insert construct of P450 monooxygenase enzymes from *B. pendula*, where we cloned CYP716 genes (pRS424::CYP716)[10] under control of the GAL1 promoter. The *B. pendula* (V5834) material was obtained from the experimental field of Viikki Campus (University of Helsinki). The second construct was the insertion of a lupeol synthase gene (Bpev01.c0219.g0020.m0001) under control of the GAL10 promoter of the previously generated single insert vector (pRS424::CYP716) to create a double insert vector (pRS424::LUS-CYP716). Both vectors were transformed into yeast (*Saccharomyces cerevisiae* [w303 background]). Transgenic yeasts were grown and induced according to Zhou et al.[95] with minor modifications. SD-TRP was used as the drop-out medium. For single insert vectors, we used 50 µg lupeol (98% purity from Cayman and dissolved in DMSO:EtOH [1:1]) in induced growth media. After 60 h of induction, the yeast growth media were centrifuged and both media and cell pellets were collected and sent for mass spectrometry.

Four samples (yeast cells [2 tubes] and cell culture media [2 tubes]) were analyzed with UPLC-QTRAP/MS (MRM). BE was extracted from the media twice with 1.0 ml ethyl acetate (Merck) for 60 min at RT and centrifuged at 15 000 rpm for 5 min according to Zhao et al.[102]. Testosterone was used as an internal standard (ISTD, 1.0 µl, 1.0 µg/ml). The cells were extracted in an analogous manner as media, but yeast cells were treated with 500 µl H_2O and 1000 µl chloroform twice and disrupted with 3 freeze/thaw cycles with ultra-sonication (15 min) prior to the extraction procedure.

The upper ethyl acetate layer was evaporated to dryness with a MiVac Duo concentrator at 40 °C (GeneVac Ltd., Ipswich, UK). The residue was resolubilized in 100 µl ACN. Due to extremely low concentrations, BE extracts had to be derivatized with p-toluenesulfonyl isocyanate (PTSI)[103,104] to improve sensitivity. Samples were derivatized at RT for 3 min with 10 µl of 60% PTSI (Sigma Aldrich) in ACN[103,104]. The derivatization reaction was terminated with 90 µl MeOH with 30 s vortexing, yielding a total volume of 200 µl. Immediately after PTSI derivatization, MRM analysis of BE was performed using UPLC-QTRAP/MS (ABSciex).

The UPLC-MS/MS system consisted of ABSciex UPLC attached to ABSciex 6500 + QTRAP mass spectrometer. The column for separation of analytes was an Acquity BEH C18 (2.1 mm × 50 mm, 1.7 µm) (Waters, Milford, MA, USA) at a temperature of 40 °C. The autosampler temperature was set to 25 °C. Injection volume was 10 µl. The chromatographic conditions were as described previously[104]. The mobile phase consisted of water with 0.1% of formic acid in H_2O (A) and acetonitrile (B) with a flow rate of 0.6 ml/min. The linear gradient started at 30% B and proceeded to 98% in 6.5 min, left at 98% B for 2 min, and switched back to initial conditions and left to stabilize, yielding a total analysis time of 10 min.

ESI source temperature was set to 450 °C. ESI/MS/MS detection was performed in negative ion mode with ion spray (IS) voltage of -4000, curtain gas (CUR) 30, collision gas (CAD) at medium, entrance potential (EP) -10, de-clustering potential (DP), -100 (betulin), collision energy (CE) -50, and collision cell exit potential (CXP) -10. Analyst v1.6.3 software (ABSciex, https://sciex.com/) was used for data processing. PTSI derivatization reagent generated betulin p-toluenesulfonyl carbamic diester (BTCD). The transitions for BE and ISTD (testosterone) were as follows: Betulin MRM 835.2→620.2 $[M\text{-}PTSI\text{-}H_2O\text{-}H]^-$, 835.2→638.3 $[M\text{-}PTSI\text{-}H]^-$ and 835.2→196.0 $[PTSI\text{-}H]^-$, and for testosterone MRM 484.2→287.2 $[M\text{-}PTSI\text{-}H]^-$ and MRM 484.2→269.2 $[M\text{-}PTSI\text{-}H_2O\text{-}H]^-$. The most intense transitions of MRM 835.2→620.2 (BE) and MRM 484.2→287.2 (ISTD, testosterone) were used.

Data availability

All sequencing data are available at NCBI under bioproject ID PRJNA1203915 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1203915/), including a biosample (https://www.ncbi.nlm.nih.gov/biosample/SAMN46010789) and RNAseq data (https://www.ncbi.nlm.nih.gov/sra/SRX27220539). The genome assembly and annotation are provided in figshare (https://doi.org/10.6084/m9.figshare.25556040.v1).

Received: 4 December 2024; Accepted: 2 June 2025
Published online: 01 July 2025

References

1. Blanchette, R. A. Progressive stages of discolortion and decay associated with the canker-rot fungus, *Inonotus obliquus*, in birch. *Phytopathology* **72**, 1272–1277 (1982).
2. Ryvarden, L. & Gilbertson, R.L. *European polypores. 1 : Abortiporus-Lindtneria*, (Oslo : Fungiflora, 1993).
3. Ma, L., Chen, H., Dong, P. & Lu, X. Anti-inflammatory and anticancer activities of extracts and compounds from the mushroom *Inonotus obliquus*. *Food Chem* **139**, 503–508 (2013).
4. Yan, Z. F. et al. Inhibitory and Acceleratory Effects of *Inonotus obliquus* on Tyrosinase Activity and Melanin Formation in B16 Melanoma Cells. *Evid Based Complement Alternat Med* **2014**, 259836 (2014).
5. Nagajyothi, P. C., Sreekanth, T. V., Lee, J. I. & Lee, K. D. Mycosynthesis: antibacterial, antioxidant and antiproliferative activities of silver nanoparticles synthesized from *Inonotus obliquus* (Chaga mushroom) extract. *J Photochem Photobiol B* **130**, 299–304 (2014).
6. Song, F. Q., Liu, Y., Kong, X. S., Chang, W. & Song, G. Progress on understanding the anticancer mechanisms of medicinal mushroom: *Inonotus obliquus*. *Asian Pac J Cancer Prev* **14**, 1571–1578 (2013).
7. Ern, P.T.Y., Yin, Q.T., Shin, Y.F. & and Yin, A.C.Y. Therapeutic properties of *Inonotus obliquus* (Chaga mushroom): A review. *Mycology* **15**, 144–161 (2024).
8. Holonec, L., Ranga, F., Crainic, D., Truța, A. & Socaciu, C. Evaluation of Betulin and Betulinic Acid Content in Birch Bark from Different Forestry Areas of Western Carpathians. *Notulae Botanicae Horti Agrobotanici Cluj-Napoca* **40**(2012).
9. P. Kovalenko, L. *et al.* Antiallergenic activity of birch bark dry extract with at least 70% betulin content. *Pharm Chem J* **43**, 110–114 (2009).
10. Alonso-Serra, J. et al. Tissue-specific study across the stem reveals the chemistry and transcriptome dynamics of birch bark. *New Phytol* **222**, 1816–1831 (2019).
11. Šiman, P. et al. Effective Method of Purification of Betulin from Birch Bark: The Importance of Its Purity for Scientific and Medicinal Use. *PLoS ONE* **11**, e0154933 (2016).
12. Shai, L.J., McGaw, L.J., Aderogba, M.A., Mdee, L.K. & Eloff, J.N. Four pentacyclic triterpenoids with antifungal and antibacterial activity from *Curtisia dentata* (Burm.f) C.A. Sm. leaves. *J Ethnopharmacol* **119**, 238–44 (2008).
13. Salin, O. et al. Inhibitory effect of the natural product betulin and its derivatives against the intracellular bacterium *Chlamydia pneumoniae*. *Biochem Pharmacol* **80**, 1141–1151 (2010).
14. Gong, Y. et al. The synergistic effects of betulin with acyclovir against herpes simplex viruses. *Antiviral Res* **64**, 127–130 (2004).
15. Zhang, H.J. *et al.* Natural anti-HIV agents. Part IV. Anti-HIV constituents from *Vatica cinerea*. *J Nat Prod* **66**, 263–8 (2003).
16. Andre, C. M. et al. Unusual Immuno-Modulatory Triterpene-Caffeates in the Skins of Russeted Varieties of Apples and Pears. *J. Agric. Food Chem.* **61**, 2773–2779 (2013).
17. Wu, J., Niu, Y., Bakur, A., Li, H. & Chen, Q. Cell-Free Production of Pentacyclic Triterpenoid Compound Betulinic Acid from Betulin by the Engineered *Saccharomyces cerevisiae*. *Molecules* **22**(2017).
18. Fukushima, E. O. et al. CYP716A subfamily members are multifunctional oxidases in triterpenoid biosynthesis. *Plant Cell Physiol* **52**, 2050–2061 (2011).
19. Siddiqui, S. A. et al. A novel triterpenoid 16-hydroxy betulinic acid isolated from *Mikania cordata* attributes multi-faced pharmacological activities. *Saudi J Biol Sci* **26**, 554–562 (2019).

20. Khelil, R., Jardé, E., Cabello-Hurtado, F., Ould-el-Hadj Khelil, A. & Esnault, M.-A. Structure and composition of the wax of the date palm, *Phoenix dactylifera* L., from the septentrional Sahara. *Scientia Horticulturae* **201**, 238–246 (2016).
21. Koolen, H. H. F. et al. Triterpenes and flavonoids from the roots of *Mauritia flexuosa*. *Rev Bras Farmacogn* **22**, 189–192 (2012).
22. Khouloud Barakat, M.S. Bioactive Betulin produced by marine Paecilomyces WE3-F. *J Appl Pharm Sci*, 034–040 (2016).
23. Yin, Y., Cui, Y. & Ding, H. Optimization of betulin extraction process from *Inonotus obliquus* with pulsed electric fields. *Innov Food Sci Emerg Technol* **9**, 306–310 (2008).
24. Alresly, Z. et al. Bioactive Triterpenes from the Fungus *Piptoporus betulinus*. *Rec Nat Prod* **10**, 103–108 (2015).
25. Suzuki, H. et al. Comparative analysis of CYP716A subfamily enzymes for the heterologous production of C-28 oxidized triterpenoids in transgenic yeast. *Plant Biotechnol (Tokyo)* **35**, 131–139 (2018).
26. Miettinen, K. et al. The ancient CYP716 family is a major contributor to the diversification of eudicot triterpenoid biosynthesis. *Nat Commun* **8**, 14153 (2017).
27. Sezutsu, H., Le Goff, G. & Feyereisen, R. Origins of P450 diversity. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **368**, 20120428–20120428 (2013).
28. Nelson, D.R. Cytochrome P450 diversity in the tree of life. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **1866**, 141–154 (2018).
29. Waterhouse, R.M. *et al.* BUSCO applications from quality assessments to gene prediction and phylogenomics. (2017).
30. Duan, Y. et al. Genome sequencing of *Inonotus obliquus* reveals insights into candidate genes involved in secondary metabolite biosynthesis. *BMC Genomics* **23**, 314 (2022).
31. Zhang, R.-q. *et al.* Genomic and Metabolomic Analyses of the Medicinal Fungus *Inonotus hispidus* for Its Metabolite's Biosynthesis and Medicinal Application. in *Journal of Fungi* Vol. 8 (2022).
32. Faino, L. et al. Transposons passively and actively contribute to evolution of the two-speed genome of a fungal pathogen. *Genome Res* **26**, 1091–1100 (2016).
33. Ali, S. et al. An immunity-triggering effector from the Barley smut fungus *Ustilago hordei* resides in an Ustilaginaceae-specific cluster bearing signs of transposable element-assisted evolution. *PLoS Pathog* **10**, e1004223 (2014).
34. Kang, S., Lebrun, M. H., Farrall, L. & Valent, B. Gain of virulence caused by insertion of a Pot3 transposon in a *Magnaporthe grisea* avirulence gene. *Mol Plant Microbe Interact* **14**, 671–674 (2001).
35. Hage, H. *et al.* Gene family expansions and transcriptome signatures uncover fungal adaptations to wood decay. *Environ Microbiol* (2021).
36. Torres, D.E., Thomma, B.P.H.J. & Seidl, M.F. Transposable Elements Contribute to Genome Dynamics and Gene Expression Variation in the Fungal Plant Pathogen *Verticillium dahliae*. *Genome Biology and Evolution* **13**(2021).
37. Cantarel, B. L. et al. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res* **37**, D233–D238 (2009).
38. Liu, Y. et al. Lignin degradation potential and draft genome sequence of *Trametes trogii* S0301. *Biotechnol Biofuels* **12**, 256 (2019).
39. Yu, F., Song, J., Liang, J., Wang, S. & Lu, J. Whole genome sequencing and genome annotation of the wild edible mushroom. *Russula griseocarnosa. Genomics* **112**, 603–614 (2020).
40. Liu, Y. et al. Whole genome sequencing of an edible and medicinal mushroom, *Russula griseocarnosa*, and its association with mycorrhizal characteristics. *Gene* **808**, 145996 (2022).
41. Fradj, N. *et al.* RNA-Seq de Novo Assembly and Differential Transcriptome Analysis of Chaga (Inonotus obliquus) Cultured with Different Betulin Sources and the Regulation of Genes Involved in Terpenoid Biosynthesis. *Int J Mol Sci* **20**(2019).
42. Kfoury, B., Rodrigues, W. F. C., Kim, S.-J., Brandizzi, F. & Del-Bem, L.-E. Multiple horizontal gene transfer events have shaped plant glycosyl hydrolase diversity and function. *New Phytol.* **242**, 809–824 (2024).
43. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015).
44. Hibbett, D. S. & Thorn, R. G. Basidiomycota: Homobasidiomycetes. In *Systematics and evolution* (eds McLaughlin, D. J. et al.) 121–168 (Springer Berlin Heidelberg, Berlin, 2001).
45. Matheny, P. B. et al. Contributions of *rpb2* and *tef1* to the phylogeny of mushrooms and allies (*Basidiomycota*, Fungi). *Mol Phylogenet Evol* **43**, 430–451 (2007).
46. Zhao, R.-L. et al. A six-gene phylogenetic overview of *Basidiomycota* and allied phyla with estimated divergence times of higher taxa and a phyloproteomics perspective. *Fungal Divers* **84**, 43–74 (2017).
47. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: A resource for timelines, TimeTrees, and divergence times. *Mol Biol Evol* **34**, 1812–1819 (2017).
48. Yang, Z. et al. Plastome phylogenomics provide new perspective into the phylogeny and evolution of *Betulaceae* (Fagales). *BMC Plant Biol.* **22**, 611 (2022).
49. Fischer, M. *Porodaedalea* (*Phellinus pini* group, Basidiomycetes) in Europe: A new species on *Larix sibirica*. *P. niemelaei. Karstenia* **40**, 43–48 (2000).
50. Mohamed, J., Michael, P. & Nasko, T. Natural durability of selected larch and scots pine heartwoods in laboratory and field tests. *Int. Biodeterior. Biodegrad* **91**, 88–96 (2014).
51. Birchler, J. A. & Veitia, R. A. The gene balance hypothesis: Implications for gene regulation, quantitative traits and evolution. *New Phytol.* **186**, 54–62 (2010).
52. Corrochano, L. M. et al. Expansion of signal transduction pathways in fungi by extensive genome duplication. *Curr Biol* **26**, 1577–1584 (2016).
53. Panchy, N., Lehti-Shiu, M. & Shiu, S.-H. Evolution of gene duplication in plants. *Plant Physiol.* **171**, 2294–2316 (2016).
54. Yap, H.-Y.Y. et al. The genome of the tiger milk mushroom, *Lignosus rhinocerotis*, provides insights into the genetic basis of its medicinal properties. *BMC Genom* **15**, 635 (2014).
55. Jarrar, Y. B. & Lee, S.-J. Molecular functionality of cytochrome P450 4 (CYP4) genetic polymorphisms and their clinical implications. *Int. J. Mol. Sci.* **20**(17), 4274 (2019).
56. Ide, M., Ichinose, H. & Wariishi, H. Molecular identification and functional characterization of cytochrome P450 monooxygenases from the brown-rot basidiomycete *Postia placenta*. *Arch. Microbiol.* **194**, 243–253 (2012).
57. Song, X. et al. Biosynthesis of squalene-type triterpenoids in *Saccharomyces cerevisiae* by expression of CYP505D13 from *Ganoderma lucidum*. *Bioresourc Bioprocess* **6**, 19 (2019).
58. Fang, Y., Luo, M., Song, X., Shen, Y. & Xiao, H. Improving the production of squalene-type triterpenoid 2,3;22,23-squalene dioxide by optimizing the expression of CYP505D13 in *Saccharomyces cerevisiae*. *J. Biosci. Bioeng.* **130**, 265–271 (2020).
59. Blin, K. et al. antiSMASH 2.0: A versatile platform for genome mining of secondary metabolite producers. *Nucl. Acids Res.* **41**(W1), W204–W212 (2013).
60. Durairaj, P., Hur, J.-S. & Yun, H. Versatile biocatalysis of fungal cytochrome P450 monooxygenases. *Microb. Cell Fact.* **15**, 125 (2016).
61. Buchholz, P. C. et al. BioCatNet: A database system for the integration of enzyme sequences and biocatalytic experiments. *ChemBioChem* **17**, 2093–2098 (2016).
62. Chen, W. et al. Fungal cytochrome p450 monooxygenases: Their distribution, structure, functions, family expansion, and evolutionary origin. *Genome Biol Evol* **6**, 1620–1634 (2014).
63. Otey, C. R. et al. Functional evolution and structural conservation in chimeric cytochromes p450: Calibrating a structure-guided approach. *Chem Biol* **11**, 309–318 (2004).

64. Van Valen, L. Molecular evolution as predicted by natural selection. *J. Mol. Evol.* **3**, 89–101 (1974).
65. Chang, S., Puryear, J. & Cairney, J. A simple and efficient method for isolating RNA from pine trees. *Plant Mol Biol Rep* **11**, 113–116 (1993).
66. Salojärvi, J. et al. Genome sequencing and population genomic analyses provide insights into the adaptive landscape of silver birch. *Nat Genet* **49**, 904 (2017).
67. Andrews, S. FastQC: A quality control tool for high throughput sequence data. in http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
68. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
69. Kim, D. et al. TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**, R36 (2013).
70. Grabherr, M. G. et al. Trinity: Reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat. Biotechnol.* **29**, 644–652 (2011).
71. Gremme, G., Steinbiss, S. & Kurtz, S. GenomeTools: A comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **10**, 645–656 (2013).
72. Godzik, A. & Li, W. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
73. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol* **9**, R7–R7 (2008).
74. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**, 525–527 (2016).
75. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: Transcript-level estimates improve gene-level inferences. *F1000Res* **4**, 1521 (2015).
76. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).
77. Lodhi, M., Ye, G.-N., Weeden, N. & Reisch, B. A simple and efficient method for DNA extraction from grapevine cultivars and *Vitis* species. *Plant Mol. Biol. Rep.* **12**, 6–13 (1994).
78. Bonfield, J. K., Smith, K. & Staden, R. A new DNA sequence assembly program. *Nucleic Acids Res* **23**, 4992–4999 (1995).
79. Chin, C. S. et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods* **13**, 1050–1054 (2016).
80. Stanke, M. & Morgenstern, B. AUGUSTUS: A web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucl. Acids Res.* **33**, W465–W467 (2005).
81. Besemer, J. & Borodovsky, M. GeneMark: Web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucl. Acids Res.* **33**, W451–W454 (2005).
82. Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M. & Stanke, M. BRAKER1: Unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* **32**, 767–769 (2015).
83. Rice, P., Longden, I. & Bleasby, A. EMBOSS: The European molecular biology open software suite. *Trends Genet* **16**, 276–277 (2000).
84. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat Meth* **12**, 59 (2014).
85. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinform* **6**, 31 (2005).
86. Margulies, M. et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
87. Quevillon, E. et al. InterProScan: Protein domains identifier. *Nucleic Acids Res* **33**, W116–W120 (2005).
88. Schlapfer, P. et al. Genome-wide prediction of metabolic enzymes, pathways, and gene clusters in plants. *Plant Physiol* **173**, 2041–2059 (2017).
89. Haas, B. J., Delcher, A. L., Wortman, J. R. & Salzberg, S. L. DAGchainer: A tool for mining segmental genome duplications and synteny. *Bioinformatics* **20**, 3643–3646 (2004).
90. Almagro Armenteros, J. J. et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat Biotechnol* **37**, 420–423 (2019).
91. Zhang, H. et al. dbCAN2: A meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res* **46**, W95–W101 (2018).
92. Mount, D. W. Using the basic local alignment search tool (BLAST). *CSH Protoc.* **2007**(7), pdb.top17 (2007).
93. Werck-Reichhart, D., Bak, S. & Paquette, S. Cytochromes p450. *Arabidopsis Book* **1**, e0028 (2002).
94. Deng, J., Carbone, I. & Dean, R. A. The evolutionary history of cytochrome P450 genes in four filamentous ascomycetes. *BMC Evol Biol* **7**, 30 (2007).
95. Zhou, C., Li, J., Li, C. & Zhang, Y. Improvement of betulinic acid biosynthesis in yeast employing multiple strategies. *BMC Biotechnol.* **16**, 59 (2016).
96. Al-Mssallem, I. S. et al. Genome sequence of the date palm *Phoenix dactylifera* L. *Nat. Commun.* (2013).
97. Tang, H. et al. An improved genome release (version Mt4.0) for the model legume *Medicago truncatula*. *BMC Genom* **15**, 312 (2014).
98. The French–Italian Public Consortium for Grapevine Genome, C. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463 (2007).
99. Edgar, R. C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
100. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609–W612 (2006).
101. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
102. Zhao, G., Yan, W. & Cao, D. Simultaneous determination of betulin and betulinic acid in white birch bark using RP-HPLC. *J. Pharm. Biomed. Anal.* **43**, 959–962 (2007).
103. Zuo, M., Gao, M. J., Liu, Z., Cai, L. & Duan, G. L. p-Toluenesulfonyl isocyanate as a novel derivatization reagent to enhance the electrospray ionization and its application in the determination of two stereo isomers of 3-hydroxyl-7-methyl-norethynodrel in plasma. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* **814**, 331–337 (2005).
104. Hu, Z. et al. Development and validation of an LC-ESI/MS/MS method with precolumn derivatization for the determination of betulin in rat plasma. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* **939**, 38–44 (2013).
105. Burgers, P. M. Overexpression of multisubunit replication factors in yeast. *Methods* **18**, 349–355 (1999).

Acknowledgements

We thank Brendan Battersby for assistance with yeast cloning, Cory D. Dunn who provided the yeast strain, and Peter M.J. Burgers, Ville O. Paavilainen, and Juho Kellosalo for providing the expression vector. We also acknowledge the computational infrastructure of CSC IT Center for Science, Finland. J.S would like to acknowledge funding from the University of Helsinki three-year grant, Academy of Finland (decisions 318288 and 319947),

TreeBio - Research Council of Finland Centre of Excellence (decision 346140), and the Nanyang Technological University start-up grant. We also commemorate Prof. Jaakko Kangasjärvi (1960-2024), whose intellectual and funding contributions were invaluable to this work. We also appreciate all the hard work by Dr. Derek Ho from Medical and Scientific Writer ScriboMedica Ltd and the language center/services of the University of Helsinki (https://www.helsinki.fi/en/language-centre), for the language revision of this manuscript.

Author contributions

O.S and J.S conceived and designed the project. Funding acquisition was performed by J.S and J.K. O.S collected the DNA and RNA samples. O.S and J.S managed and coordinated all bioinformatics activities. O-P.S, L.G.P, and P.A performed RNA and DNA library construction, sequencing, and genome assembly. O.S, S.R, and P.S performed genome annotation. O.S analyzed the RNA sequencing data, including de novo assembly of RNAseq. O.S and J.S performed comparative genomics analyses. T.S and N.S were involved in field research for sample isolation. O.S and M.W grew and collected the samples for mass spectrometry. G.L.B, B.B, M.W, and O.S performed cloning and expression of P450 and Lupeol synthase enzymes. N.S and J.L performed mass spectrometry, including sample pretreatment, method development, UPLC-HDMS analysis, metabolite identification, and data interpretation. K.O was involved in language revision. O.S and J.S contributed to data interpretation. O.S and J.S wrote the original manuscript with input from K.O, U.R, N.S, O-P.S, T.T, and other coauthors.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-05414-1.

Correspondence and requests for materials should be addressed to O.S., U.R. or J.S.