Are there conserved biosynthetic genes in lichens? Genome-wide assessment of terpene biosynthetic genes suggests conserved evolution of the squalene synthase cluster

Garima Singh (garima.singh@unipd.it)
University of Padua
Anna Pasinato
University of Padua
Alejandra López-Chicheri Yriarte
University of Padua
David Pizarro
Complutense University of Madrid
Pradeep K. Divakar
Complutense University of Madrid
Imke Schmitt
Senckenberg Biodiversity and Climate Research Centre
Francesco Dal Grande
University of Padua

Research Article

Keywords: BiG-SCAPE, secondary metabolites, genome mining, ergosterol, antiSMASH, Lecanoromycetes

Posted Date: January 25th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3875510/v1

Additional Declarations: No competing interests reported.

Abstract

Lichen-forming fungi (LFF) are prolific producers of functionally and structurally diverse secondary metabolites, most of which are taxonomically exclusive delivering lineage-specific roles. But are there evolutionary conserved biosynthetic pathways in lichens? Based on the current evidence it seems there aren't any. This notion is derived from polyketide-derivatives as most biochemical investigations on lichens focused on PKSs whereas the other class of compounds remain mostly unexplored. Here we present first systematic identification and comparison of lichenized fungal terpene synthases which are ever-present in the LFF, however, they have never been systematically analyzed. We gathered 111 genomes of LFF representing 60 genera and 23 families and implemented genome mining and gene clustering approaches to identify and group the clusters into networks of similar biosynthetic clusters. Our large-scale analysis led to the identification of 733 terpene BGCs with varying degrees of mutual similarity, most of them were unique. Remarkably, we found two widely distributed and homologous networks – both squalene/phytoene clusters - putatively involved in sterol/ergosterol biosynthesis. These clusters contain the same core gene, but different sets of accessory genes. This indicates an early gene duplication event, followed by gene losses and gains of accessory genes, leading to different architecture of the two clusters. We show that sequence divergence, gene duplications, losses and rearrangement and are the major factors driving the evolution of this cluster across LFF. Our study shows lichenized fungi do have a core biosynthetic cluster, bringing the first indication that a biosynthetic gene may form an essential gene content in lichens.

Background

Lichen-forming fungi (LFF), a symbiotic association between a fungus and photosynthetic partner/s (algae/cyanobacteria) are prolific producers of structurally diverse secondary metabolites which play essential role in organism survival, defense and ecological interactions [1–3]. Although about 1000 lichen metabolites have been reported, most of the compounds have a narrow taxonomic distribution [4–7] and no metabolite or secondary metabolite pathway has been shown to be evolutionarily conserved in fungi. This notion however is mostly based on the study of polyketide-derivatives - the most well-studied lichen compounds in terms of chemistry and pathway – which indeed are either genus or species-specific (olivetoric acid, [8], e.g., grayanic acid, [9], usnic acid [10–12], gyrophoric acid [13]. In rare cases the PKS-derivatives are reported to be secreted by a few taxa or families (atranorin [14], anthraquinones in Teloschistales, [15]. An evolutionary conserved, cross-kingdom or cross-class presence of a gene would suggest an important functional role, shedding light on the significance of secondary metabolites in lichens. An in-depth search for conserved biosynthetic gene hasn't been done yet, mostly because of the scarcity of lichen genomic data. The increased taxonomic coverage of genomic databases in lichenized fungi provides a premise to explore the possibility of an evolutionary conserved BGC in lichens [16, 17].

Terpenes constitute an interesting class of biosynthetic compounds owing to their ecological and physiological importance in bacteria [18–20], non-lichenized model fungi, as well as in plants [21]. Furthermore, terpene derivatives have a high pharmaceutical and commercial importance [19, 22, 23] with

menthol and taxol being among the most renowned terpene-derivatives for their use in food flavorings/aroma [24, 25] and anticancer medication [26, 27], respectively. As for the terpene BGCs, like other BGC classes, most genes involved in terpene synthesis are also linearly organized [28–31]. A terpene cluster typically consists of a terpene cyclase (TPC) as the core enzyme which forms the hydrocarbon backbones of terpenoids, and of a few accessory enzymes which modify the backbone terpene or are involved in the regulation of terpene synthesis or the transportation of the final product [29, 30, 32]. The most common tailoring enzyme associated with terpene clusters are cytochrome P450 mono-oxygenases (CYP450), NAD(P)+, and flavin-dependent oxidoreductases. A recent study suggests that about 20% of the biosynthetic genes in LFF are terpenes [33]. Despite this, only little is known about lichen terpenes (ergosterol and squalene among a few studied [34–36]). To our knowledge all studies so far focused on terpene detection and its taxonomic significance for species identification [35, 37, 38], whereas widescale metagenomic investigation of terpene BGCs was never carried out leaving the diversity and evolution of this BGC class completely unknown.

In this study we set out to answer the following questions: 1) What is the diversity of BGCs linked to terpene biosynthesis in LFF? 2) Are there any evolutionarily conserved BGCs related to terpene biosynthesis? And 3) What are the major forces shaping the evolution of terpene BGCs in lichens.

Methods

Dataset, genome assembly and annotation

A total of 111 LFF were included in the study, 102 Lecanoromycetes, four Eurotiomycetes and five Dothidiomycetes (Supplementary material 1). Gene prediction and functional annotation were performed as according to Singh et al. [39]. Genome completeness was estimated using BUSCO [40] with the ascomycota dataset. Single copy BUSCOs were quality-filtered and compared to extract those present in most taxa (maximum one sample missing). The single copy BUSCOs were then concatenated and the concatenated sequences from all the taxa were aligned using MAFFT L-INS-I [41, 42]. Phylogenetic relationships were inferred from the alignment using maximum likelihood (ML) analysis as implemented in IQTree v1.5.5 [43] using standard model selection and 1,000 bootstrap replicates. The resulting tree was visualized using FigTree 1.3.1 [44] and annotated in iTOL [45].

Hunting for a conserved BGC: BGC identification and clustering using automated genome mining software

Core biosynthetic genes were predicted and annotated in 112 taxa using the automated genome mining pipeline AntiSMASH (antibiotics & SM Analysis Shell, v7.0 [46]) which identifies BGCs based on the sequence similarity of the concerned enzymes/genes. The BGCs identified via antiSMASH (PKSs, NRPSs, terpenes, RiPPs and hybrid BGCs) comprise varying degrees of core gene and cluster similarity (as evident from the closest MiBiG hit). To compare cluster sequence and architectural relationships between the BGCS identified via AntiSMASH and identify the homologous and cross taxa conserved BGC if any,

we used biosynthetic gene similarity clustering and prospecting engine' (BiG-SCAPE [47], https://git.wageningenur.nl/medema-group/BiG-SCAPE) which builds sequence similarity networks and groups BGCs into gene cluster families (GCFs) potentially coding for the same or a similar metabolite across a range of organisms. BiG-SCAPE groups similar BGCs into gene cluster families (GCFs) based on distance metrices was used to classify and group BGCs. We computed the BGC assignment into GCFs using a conservative approach – the raw distance cut-offs of 0.6. The analysis was done retaining singletons and with the PFAM database.

BiG-SCAPE clustered the similar BGCs into GCFs. Two or more GCFs potentially coding for structurally similar compounds are grouped into networks/clans. Of the clans obtained via BiG-SCAPE (PKSs, NRPSs, RiPPs, hybrids and terpene BGCs, the largest were obtained for terpenes, hereto being the most promising candidate for a conserved BGC across lichens. We therefore continued with in-depth analyses of the two terpene clans to hunt for conserved biosynthetic gene. Detailed Materials and methods for this section can be found in Supplementary material 2.

Core terpene BGCs of lichens – gene sequence conservation and cluster synteny

To display the synteny and homology of the core genes in the two terpene clans among taxa, we used Clinker [48]. The collinearity analysis was performed between the members of Clan1 and Clan2 and the resulting figure was adjusted for presentation using Inkscape. Although the core gene of was identified as a putative squalene synthase, based on a MiBiG hit, most of the accessory genes could not be identified, even when using BLAST search individually (all hits resulting in hypothetical proteins). We therefore performed a conserved domain search on the amino acid sequences of the genes upstream and downstream of the squalene synthase gene to predict their putative function.

TPS Classification and Motif/Domain Annotation

The core genes of the two clans were deduced based on MiBiG similarity (sequence similarity of individual genes) and were found to be squalene synthase clusters, involved in the synthesis of cholesterol/ergosterol. However, as the conserved terpene BGCs in lichens did not cluster with the *Aspergillus* squalene synthase cluster in BiG-SCAPE, this suggests lichen squalene synthase cluster may be evolutionary diverse than that of *Aspergillus* and non-lichenized fungi in general (MiBiG provides sequence similarity of individual genes whereas BiG-SCAPE deduces cluster synteny and similarity). We therefore inferred the sequence conservation, putative structure of the core terpene cluster in lichens and inferred its physiochemical properties to bring in additional evidence whether it could be a squalene synthase.

The bacterial, fungal, and plant terpene synthases have a pair of characteristic conserved metal-binding domains consisting – [D/N)DXX(D/E) or DDXXXE] located within 80–120 (bacterial and fungal terpene synthases). In addition, it is reported that there is a 30 aa region at the C terminus specific to fungi. We tested the presence of these regions in LFF putative squalene synthases. We aligned the putative

squalene synthase sequences of LFF (Clan1 – FAM02526, FAM00230, FAM03031, FAM03070, FAM6129 and FAM07143) to that of non-lichenized fungi (*Aspergillus flavus*, *A. niger*, and *Candida albicans*) and a protozoan squalene synthase from *Trypanosoma cruzi* to – 1) confirm if the C-terminal region, ~ 30aa long, previously reported from *Aspergillus* to be fungal specific is widespread in fungi, i.e. LFF and 2) identify the characteristic conserved metal-binding domains ([D/N)DXX(D/E) or DDXXXE] reportedly located within 80–120 (bacterial and fungal terpene synthases) and 3) to identify characteristic aa of lichen squalene synthase if any. The alignment comprises 590aa. For LFF squalene synthase we predicted the motifs and structure from the sequence of the gene using the programs mentioned below.

To test if the biosynthetically identified squalene synthase from LFF is indeed a transmembrane protein and to predict the transmembrane, intracellular and extracellular regions of the squalene synthase we implemented two programs Phyre2 [49] and DeepTMHMM. DeepTMHMM (http://heliquest.ipmc.cnrs.fr) predicts protein structure and membrane topology of both alpha-helical and beta-barrel transmembrane proteins using deep learning and computes the probability whether if the protein is situated inside or outside of the cell, and if it is a transmembrane protein (comprising a high percentage of hydrophobic amino acid residues). Phyre2 on the other hand predicts three-dimensional (3D) structure of the protein from the given amino acids using the alignment of hidden Markov models via Hhsearch1. We then checked the physiochemical properties (i.e., hydrophobicity etc.) of the entire sequence as we all as the region predicted to be transmembrane (22aa) using HeliQuest [50]. This program returns hydrophobicity ⟨H⟩ and hydrophobic moment ⟨μH⟩ values for the input sequences. A positive ⟨H⟩ value (e.g., > 1) denotes a hydrophobic helix, and a negative value indicates a hydrophilic helix. The transmembrane proteins have high hydrophobicity.

Results

Biosynthetic gene diversity and clustering

AntiSMASH predicted the presence of 5542 BGCs in 111 genomes (Supplementary material 1). BiG-SCAPE analyses clustered BGC of each class into networks (GCFs and clans) and those clusters without any similar biosynthetic gene with the dataset or in MiBiG were obtained as singletons. Although several networks were generated for each BGC class – PKSs, NRPSs, RiPPs, hybrids and terpenes, the largest networks were obtained for terpene BGCs, making them the most promising candidate for a conserved BGC across LFF.

In-depth analysis of the diversity of terpene BGCs

Of the total ~ 5000 BGCs predicted, 733 (13%) were identified as presumptive terpene synthases. Among these 680 were detected in Lecanoromycetes (6.6 ± 3.2 BGCs/taxon), 24 in Eurotiomycetes (6.75 ± 1.2 BGCS/taxon) and 29 in Dothidiomycetes (7.4 ± 2.07 BGCS/taxon) (Fig. 1A). Physciaceae constitutes most terpene BGCs rich family (14.5 ± 2.1 BGCs/taxon) (Fig. 1B)). The most terpene BGCs rich taxa were – *Parmelia*, Parmeliaceae (12.5 + 3.5 BGCs/taxon), *Bulbothrix* (13.5 + 0.5BGC/taxon) and *Evernia prunastri* (14 BGCs/taxon). None of the terpene BGCs detected in the dataset clustered with a terpene

BGC annotated in MiBiG (a public database of annotated biosynthetic genes) indicating structural and putatively functional novelty of the identified terpenes. This could be because MiBiG mostly constitutes plant terpene BGCs which are diterpenes whereas LFF are known to produce triterpenes and therefore the BGCs reported here maybe involved in the triterpenes synthesis.

Terpene clusters and BiG-SCAPE clustering

The 733 terpene BGCs were grouped into 445 GCFs based on the sequence similarity of core genes and cluster synteny thus indicating that these terpene synthases are highly structurally diverse (Fig. 1C). We detected both species-specific as well as highly conserved terpene BGCs. Overall, 377 were singletons, detected only once and unique to a species whereas others grouped into GCFs with two/more members. Some GCFs were rare and specific to a few taxa (e.g. 418 GCFs [94%] occur across three or fewer genomes).

Core terpene BGCs of lichens

We retrieved two big terpene biosynthetic gene networks/clans, each comprising a few GCFs (Fig. 2). Clan1 consisted of six GCFs and 83 BGCs and Clan2 of three GCFs and 68 BGCs (Fig. 2, 3A). Clan1 was reported from 94.6% taxa encompassing all families and genera included in the study (missing in six taxa, Fig. 2) whereas Clan2 was shared across 80% taxa (missing in 23 taxa, Fig. 2). Interestingly, in Eurotiomycetes and Dothidiomycetes the clans were detected only via local BLAST and not when using BiG-SCAPE indicating sequence and cluster divergence from that of LFF (, Fig. 2). Seventy five percent of the LFF contain both Clan1 and Clan2. Each taxon had at least one copy of the squalene synthase.

Evolutionary conservation of clans in related lichenized fungi and non-lichenized fungi

Although the corresponding cluster in LFF from Dothideomycetes and Eurotiomycetes do not group with clan2 in the BiG-SCAPE analyses, the member taxa form an independent network and the non-lichenized taxa from these classes did not group with the LFF. This cluster has the same gene composition but probably sequence divergence guides clustering of these taxa together rather than with lecanoromycetes.

In clan1, apart from the lichenized fungi from Lecanoromycetes, Eurotiomycetes and Dothideomycetes, four non-lichenized parasitic fungi belonging to Dothideomycetes grouped within. The non-lichenized clusters however were phylogenetically most distant, and shared low conservation to that of lichenized fungi. Based on this evidence we propose that clan2 is restricted to lichenized fungi whereas clan1 is conserved in LFF but also shared by some closely-related non-lichenized fungi.

Cluster composition of conserved clans

Clan1 consists of two genes conserved across LFF, i.e., a squalene synthase and a J domain-containing protein, whereas several other genes in Clan 1 are conserved within a family or genus, e.g., Abhydrolase/SLC5-6-like and BTB/POZ domain containing protein (Fig. 3A, 3B). The J domain-containing protein is located upstream of a terpene synthase. These accessory genes are usually involved in protein

(re)folding, trafficking, remodeling, disaggregation, and degradation. Other protein frequently found in the conserved clusters of Clan1 have conserved domains typically found in the proteins related to transport and signaling function, for instance protein kinases (cellular signaling), DNA translocase FtsK and Abhydrolase/SLC5-6-like (locatization and transport) and transcription regulation (The BTB/POZ domain containing protein).

Clan2 contains four genes conserved (90%) – squalene synthase, phytoene desaturase, RPE65 and Opsin (Fig. 3C). RPE65 is most probably involved in regulatory function being located upstream of the squalene synthase whereas phytoene desaturase and opsin are located downstream of the squalene synthase. Other genes occasionally present in the clan include protein kinase, abhydrolase, GPI2, helicase etc.

The core gene of both clans is the squalene synthase although that of Clan2 is slightly shorter than Clan1 (Clan1: 1000-1600bp, 330-550aa vs. Clan2: 900 nucleotide, 300 aa long) ((Fig. 3B, C). Interestingly, while the squalene synthase of a clan displayed high sequence similarity among themselves, both sequence and cluster composition were different from those of members of Clan2 (Fig. 3B, C).

TPS Classification and Motif/Domain Annotation

As the conserved squalene synthase did not group with its fungal or plant counterpart in MiBIG, we used the following indices to validate that the terpene synthase in the conserved clans is indeed a squalene synthase: sequence similarity, cluster synteny, conserved motifs, and other physiochemical features (hydrophobicity and polarity of the domains).

We found that the conserved, characteristic metal-binding domain of terpene synthases [D/N)DXX(D/E) was present in the BGCs of both conserved clans across LFF in the same region as in bacterial, fungal, and plant terpene synthases, i.e., between within 80–120aa. We identify this region to be DTI/VEDD. The most conserved region between fungi and non-lichenized fungi is about 50-100aa from the N terminus, whereas the C terminus region specific to fungi including the protein-protein interaction and transmembrane region is highly variable between the two.

We additionally inferred the presence of highly hydrophobic residues towards the C terminus which form the transmembrane domain of the protein, another typical feature of squalene synthases. DeepTMHMM predicted the presence of one transmembrane helix whereas Phyre2 also detected an additional helix towards the N terminus. However, as this transmembrane helix has slow probability and confidence, we propose that the LFF squalene synthase has only one transmembrane helix of about 22–25 aa towards the C terminus. The protein contains intracellular and extracellular regions of about 420–430 bp and 40-50aa aa before and after the transmembrane helix, respectively (Fig. 3A, B). Furthermore, the putative squalene synthase has a C-terminal region that is typical of fungal squalene synthases and absent in plants, fungi and humans. The *A. fumigatus* squalene synthases contains in fact a conserved hinge region of 26 aa before the transmembrane domain in LFF. We identified this region in LFF as well and found it to be highly conserved across the class, except for a few variations (Fig. 3).

Fungal squalene synthases are transmembrane proteins, having polar, intra and extracellular regions and a hydrophobic transmembrane helix. The predicted transmembrane region present towards the C-terminal region of LFF squalene synthase has similar properties ((% polar residues, hydrophobicity, and hydrophobic moment) to that of C-terminal helical domain of *Aspergillus fumigatus* squalene epoxidase, suggesting it to be synthesizing a transmembrane protein as in *Aspergillus fumigatus*. (Fig. 4A, B). The membrane spanning portion of the squalene synthase is also highly variable within LFF as was shown within each kingdom of life and retains only physiochemical properties typical of the region (high hydrophobicity), not the sequence similarity. The TM domain in LFF SQS is slightly more polar (22% polarity) than in the fungal and human SQSs (16% polarity) but less than that of *Trypanosoma cruzi* (11%, 16% and 33% polarity respectively) and may bind in a somewhat different membrane-associated manner.

Before the fungal-specific C domain, there is a hinge sequence of 26aa linking the catalytic and transmembrane domain which is highly conserved within each kingdom of life [51]. In fungi, it is involved in the assembly of ergosterol multi-enzyme complexes. This region is present and highly conserved also in LFF as well, expanding the current knowledge beyond the model fungi Saccharomycetes, Euascomycetes (*Candida albicans*) and Eurotiomycetes (*Aspergillus*). The kingdom specificity of this domain is particularly interesting and can be used to develop targeted antifungal and antiprotozoal therapeutics as kingdom specific therapeutics mimicking the hinge region would not interfere with the human sterol synthesis pathway.

Discussion

Diversity of BGCs linked to terpene biosynthesis

We found that the terpene BGCs are taxonomically widespread among LFF, and all lichenized fungi have at least one BGC (or core gene) for synthesizing terpenes (Fig. 1A). Interestingly, except for the two big networks, most BGCs were singletons bearing no similarity to other lichen terpenes or to those present in MiBiG database, indicating the presence of mostly lineage-specific terpene BGCs in LFF (Fig. 1C). This is not surprising as LFF largely produce structurally diverse and unique triterpenes. For example, the lichen family Physciaceae is widely known to produce unique and diverse triterpenes [52]. This is also the case for plant terpene-related BGCs, some of which are widely distributed (e.g., cholesterol synthase) while others have a more restricted, taxon-specific distribution (e.g., carotenoids, vitamins A and D, steroid hormones, pheromones, essential oil as camphor and defense-related metabolites; [53, 54]. Presence of rare, species- or lineage-specific terpenes is known from non-lichenized fungi as well. For instance, *Aspergillus ochraceus* produces several sesquiterpenoids which are relatively rare [55]. Similarly, the marine sediment-derived fungus *Eutypella scoparia* produces a 7-methyl oxidized 2-carene derivative [56], and the marine alga-related fungus *Cochliobolus lunatus* produces two dendryphiellins, both of which are rare terpenes [57]. The presence of numerous taxon-specific terpenes reflects the breadth of the functional potential and species-specific function of these genes in LFF. Given the wide ecological niches

of lichens it is probable that the lineage-specific terpenes play specific role in lichen biology, such as biotic and abiotic stress responses and defense, among others.

Are there conserved BGCs in lichens?

Two BGC networks are widely distributed and homologous across LFF (Fig. 2). Both networks contain a squalene/phytoene synthase as the core gene, putatively involved in sterol biosynthesis (Fig. 3B, C). Each taxon has at least one copy of squalene synthase, showing 100% conservation in LFF (Fig. 2). In fact, squalene synthase is conserved across kingdoms and has been reported from fungi, plants and animals including rats and humans [51, 58]. However, cluster organization differs among them, generating the diversity of sterols among organisms [51, 58]. For instance, the squalene cluster codes ergosterol in non-lichenized fungi, ß-carotenoids in plants and cholesterol in mammals/humans [59]. The squalene synthase is one of the starting enzymes in the sterol pathway (catalyzes the first step in sterol biosynthesis – condensation of two molecules of farnesyl diphosphate (FPP) to form squalene, the precursor of all steroids) and consecutive enzymes dictate the final protein produced [60, 61]. Ergosterol, for instance, differs from the predominant mammalian sterol, cholesterol, by the presence of a methyl group among other differences and hence requires a methyl transferases. Furthermore, despite being conserved, the LFF squalene synthase gene shows low sequence similarity to the corresponding gene from non-lichenized fungi. The gene conservation but high sequence divergence coupled with different downstream enzymes explains their involvement in the important metabolic pathway (carotenoids/cholesterol/ergosterol synthesis) but with slightly different, kingdom specific modifications.

Of the two conserved clans, Clan2 comprises LFF from Lecanoromycetes, Dothideomycetes, Eurotiomycetes as well as from Lichinomycetes whereas Clan1 has mostly LFF, but also a few non-lichenized, parasitic fungi belonging to Dothideomycetes (Fig. 3A, Supplementary material 2,3,4). Given the data, we propose that both Clan1 and Clan2 are conserved in LFF and Clan1 is also shared by parasitic non-lichenized fungi. Presence of an evolutionary conserved BGC indicates functional relevance of the cluster for the organism [62–64]. Non-lichenized fungi produce ergosterol via a complex process involving several enzymes as oxidases, methyltransferases, demethylases, desaturases, isomerases etc. Studies suggest that some lichens do produce ergosterol [36, 65–67] but neither of the conserved squalene cluster we report has the same cluster composition as reported from *Aspergillus* [68, 69] implying either that lichens produce a slightly different variant of ergosterol or that enzymes outside the cluster may participate in the ergosterol synthesis. Clan2 BGCs do comprise desaturase and oxidases required for ergosterol synthesis (Fig. 3C) but are not the same as reported from *Aspergillus* [68, 69]. Ergosterol or sterols in are involved in cell wall maintenance, regulation of membrane fluidity and structure [70]. Given the symbiotic nature of LFF it is expected that the membrane structural requirements to sustain fungal-algal cross-talks are different from that of non-lichenized fungi, which may explain the cluster difference between the two.

Our study overturns the notion that secondary metabolite clusters, unlike primary metabolite clusters, show narrow taxonomic distribution. The presence of an evolutionary conserved biosynthetic pathway in

lichens is certainly intriguing, considering that it was so far showed that lichen-fungal BGCs have a narrow, almost lineage-specific, taxonomic distribution [4, 7–9, 11, 13, 15, 71].

Gene duplication and squalene synthase isoforms

The presence of two isoforms of squalene synthase in lichens suggests a possible gene duplication of this gene at the base of Lecanoromycete diversification, followed by independent evolutionary pathway leading to different cluster architecture and most probably, its function. Copy number variation for squalene BGC is a common phenomenon and is an important factor in the evolution of these genes. For instance, single copy of this gene has been reported in rice [72], Japanese yew *Taxus cuspidate* [73] and petroleum plant *Euphorbia tirucalli* [74] whereas two copies have been reported from tobacco [75], Russian dandelion *Taraxacum koksaghyz* [76], licorice plant *Glycyrrhiza glabra* [77], barrel medic plant *Medicago truncatula* [78] and *Arabidopsis thaliana* [79]. In rare cases, more than two copies have been found, for instance three copies in the ginseng plant *Panax ginseng* [80].

Squalene synthase cluster in LFF

We found that both conserved clans comprise squalene synthase homologs as the core genes even though the accessory genes are different in these clans as well from the homologous cluster in non-lichenized fungi (Fig. 3B, C). In Clan2, apart from the squalene synthase, early carotenoid biosynthesis enzymes - phytoene desaturase, RPE65 and Opsin – are highly conserved among evolutionarily diverse lichens. Phytoene desaturase is present immediately downstream of squalene synthase. Studies show that desaturases desaturate squalene to make dehydrosqualene, and subsequently carotenoid pigments. Opsin, present further downstream of the phytoene desaturase, also constitutes one of the highly conserved genes of the cluster. Opsins are universal photoreceptors. It is proposed that they are involved in the production of carotenoid molecules with photoprotective or antioxidant molecules by acquiring promiscuous desaturases.

The accessory genes in the conserved squalene cluster were different from the homologous cluster in non-lichenized fungi. For instance, squalene cluster of fungi contains oxidative enzyme of the Cytochrome P450 monooxygenase (CYP450) family which are involved in functional modification and hence diversification of the terpenes. Interestingly, we did not find any cytochrome family genes in the putative squalene synthase clusters in LFF, but RPE65: Retinal which belongs to the carotenoid oxygenase family protein and may play similar oxidative function in this cluster in LFF. This indicates that the sterol pathway in LFF involves different genes than what reported from non-lichenized fungi.

Major factors shaping the evolution of conserved terpene BGCs in LFF

Although the two networks are mostly conserved in LFF, they differ in gene content and organization across taxa. In general, the gene order in prokaryotic as well as eukaryotic genomes tends to be poorly conserved through evolution [81–83]. However, certain groups of genes remain adjacent to each other in the genomes even over long evolutionary distances, which suggests that selection tends to preserve their

genomic co-localization [84]. Our comparative genomic analyses revealed a general conserved pattern of gene organization (Rhodanese-like domain containing protein synthase and Squalene/phytoene synthase next to each in clan 1 and RPE65, squalene synthase, phytoene desaturase and opsin in Clan2, (Fig. 3B, C)) but also several species- as well as genus-specific structural rearrangements in these conserved BGCs in LFF.

Certain genes of the cluster are restricted to a genus or even only a few or one species. For instance, helicase genes in the cluster are present only in two species. Similarly, a putative alpha/beta hydrolases gene is present only in two taxa and showed location and orientation difference between the species (Clan2, Fam121, (Fig. 3C)). Another example is the gene coding for protein kinases, present only in some taxa among the members of Clan2 (Fig. 3C). Interestingly, this gene is found in certain taxa of the Parmeliaceae between the terpene synthase and the phytoene desaturase gene, disrupting the otherwise conserved arrangement of these two genes next to each other. The sporadic presence of this gene in certain taxa indicates gene gain/loss as an essential process for the evolution of this cluster in Parmeliaceae. Gene duplications, gene losses and functional divergence of genes in the trichothecene [85] MGCs were linked to the structural diversity of the trichothecene toxins in plant pathogenic fungi. Similarly, Sterigmatocystin and aflatoxin MGCs are evolutionarily related but differ in gene content, order and orientation with aflatoxin producers having extra genes which facilitate the downstream steps for conversion of sterigmatocystin into different types of aflatoxins [86]. The accessory gene family combinations in the cluster further potentially increase the functional plasticity of these genes tailoring the product to an organism's specific needs.

Conclusions

This study dwells into the diversity and evolution of terpene biosynthetic gene clusters in lichens and delivers new perspectives on conservation of these pathways. First, we show that lichenized fungi have the potential to synthesize many different, species-specific terpenes, which could be explored for their unique bioactivity. Second, while terpene BGCs are predominantly taxon-specific, two clusters are highly conserved. Interestingly, both clusters had a squalene synthase as the core gene, but different sets of accessory genes indicating gene duplication, losses and gains are the major factors driving the evolution of this cluster. As opposed to the previous view of taxa-restricted distribution of lichen metabolites, for the first time we unveil the presence of biosynthetic gene clusters shared by most lichenized fungi. Third, we provide the putative structure of the various domains in the squalene synthase gene, which is particularly interesting as a target for cholesterol drug development.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

Genomes used in the study are available at ncbi (accession numbers in the supplementary material 1). For the genomes generated for this study the accession numbers will be filled as soon as available. The alignment and the phylogenetic tree are available at figshare link (XXXXX).

Competing interest

There are no conflicts to declare.

Funding

PKD and DP thank to Spanish Ministerio de Ciencia e Innovación, grant number PID2019-105312GB-I00 for financial support.

Author Contributions

Garima Singh, Anna Pasinato and Francesco Dal Grande contributed to the study conception and design. Material preparation and data collection was done by David Pizarro, Pradeep Kumar Divakar, Garima Singh, Anna Pasinato, and Alejandra López-Chicheri Yriarte, Francecso Dal Grande and Imke Schmitt. Analyses were performed by Garima Singh, Anna Pasinato, and Alejandra López-Chicheri Yriarte. The first draft of the manuscript was written by Garima Singh and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank Dr. Nuria Beltran (University of Padova, Italy) for help with violin plots. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

References

1. DePriest PT. Early molecular investigations of lichen-forming symbionts: 1986-2001*. Annual Review of Microbiology. 2004;58:273–301.

2. Honegger R. The Lichen Symbiosis—What is so Spectacular about it? The Lichenologist. 1998;30:193–212.
3. Calcott MJ, Ackerley DF, Knight A, Keyzers RA, Owen JG. Secondary metabolism in the lichen symbiosis. Chem Soc Rev. 2018;47:1730–60.
4. Singh G. Linking Lichen Metabolites to Genes: Emerging Concepts and Lessons from Molecular Biology and Metagenomics. Journal of Fungi. 2023;9:160.
5. Molnár K, Farkas E. Current results on biological activities of lichen secondary metabolites: a review. Z Naturforsch C J Biosci. 2010;65:157–73.
6. Boustie J, Grube M. Lichens—a promising source of bioactive secondary metabolites. Plant Genetic Resources. 2005;3:273–87.
7. Xu M, Heidmarsson S, Olafsdottir ES, Buonfiglio R, Kogej T, Omarsdottir S. Secondary metabolites from cetrarioid lichens: Chemotaxonomy, biological activities and pharmaceutical potential. Phytomedicine. 2016;23:441–59.
8. Singh G, Armaleo D, Dal Grande F, Schmitt I. Depside and Depsidone Synthesis in Lichenized Fungi Comes into Focus through a Genome-Wide Comparison of the Olivetoric Acid and Physodic Acid Chemotypes of Pseudevernia furfuracea. Biomolecules. 2021;11:1445.
9. Armaleo D, Sun X, Culberson C. Insights from the first putative biosynthetic gene cluster for a lichen depside and depsidone. Mycologia. 2011;103:741–54.
10. Abdel-Hameed M, Bertrand RL, Piercey-Normore MD, Sorensen JL. Putative identification of the usnic acid biosynthetic gene cluster by de novo whole-genome sequencing of a lichen-forming fungus. Fungal Biol. 2016;120:306–16.
11. Pizarro D, Divakar PK, Grewe F, Crespo A, Dal Grande F, Lumbsch HT. Genome-Wide Analysis of Biosynthetic Gene Cluster Reveals Correlated Gene Loss with Absence of Usnic Acid in Lichen-Forming Fungi. Genome Biol Evol. 2020;12:1858–68.
12. Egbert S, Hoffman JR, McMullin RT, Lendemer JC, Sorensen JL. Unraveling usnic acid: A comparison of biosynthetic gene clusters between two reindeer lichen (Cladonia rangiferina and C. uncialis). Fungal Biology. 2022. https://doi.org/10.1016/j.funbio.2022.08.007.
13. Singh G, Calchera A, Merges D, Valim H, Otte J, Schmitt I, et al. A Candidate Gene Cluster for the Bioactive Natural Product Gyrophoric Acid in Lichen-Forming Fungi. Microbiol Spectr. 2022;10:e0010922.
14. Kim W, Liu R, Woo S, Kang KB, Park H, Yu YH, et al. Linking a Gene Cluster to Atranorin, a Major Cortical Substance of Lichens, through Genetic Dereplication and Heterologous Expression. mBio. 2021;12:e0111121.
15. Llewellyn T, Nowell RW, Aptroot A, Temina M, Prescott TAK, Barraclough TG, et al. Metagenomics Shines Light on the Evolution of “Sunscreen” Pigment Metabolism in the Teloschistales (Lichen-Forming Ascomycota). Genome Biology and Evolution. 2023;15:evad002.
16. Bertrand RL, Abdel-Hameed M, Sorensen JL. Lichen Biosynthetic Gene Clusters. Part I. Genome Sequencing Reveals a Rich Biosynthetic Potential. J Nat Prod. 2018;81:723–31.

17. Roth MG, Westrick NM, Baldwin TT. Fungal biotechnology: From yesterday to tomorrow. Front Fungal Biol. 2023;4:1135263.
18. Yamada Y, Kuzuyama T, Komatsu M, Shin-ya K, Omura S, Cane DE. Terpene synthases are widely distributed in bacteria. 2015;112:857–62.
19. Rudolf JD, Alsup TA, Xu B, Li Z. Bacterial terpenome. Nat Prod Rep. 2021;38:905–80.
20. Avalos M, Garbeva P, Vader L, Wezel GP van, Dickschat JS, Ulanova D. Biosynthesis, evolution and ecology of microbial terpenoids. Nat Prod Rep. 2022;39:249–72.
21. Holopainen JK, Himanen SJ, Yuan JS, Chen F, Stewart CN. Ecological Functions of Terpenoids in Changing Climates. In: Ramawat KG, Mérillon J-M, editors. Natural Products: Phytochemistry, Botany and Metabolism of Alkaloids, Phenolics and Terpenes. Berlin, Heidelberg: Springer; 2013. p. 2913–40.
22. Del Prado-Audelo ML, Cortés H, Caballero-Florán IH, González-Torres M, Escutia-Guadarrama L, Bernal-Chávez SA, et al. Therapeutic Applications of Terpenes on Inflammatory Diseases. Frontiers in Pharmacology. 2021;12.
23. Wiart C, Kathirvalu G, Raju CS, Nissapatorn V, Rahmatullah M, Paul AK, et al. Antibacterial and Antifungal Terpenes from the Medicinal Angiosperms of Asia and the Pacific: Haystacks and Gold Needles. Molecules. 2023;28:3873.
24. Caputi L, Aprea E. Use of terpenoids as natural flavouring compounds in food industry. Recent Pat Food Nutr Agric. 2011;3:9–16.
25. Masyita A, Mustika Sari R, Dwi Astuti A, Yasir B, Rahma Rumata N, Emran TB, et al. Terpenes and terpenoids as main bioactive compounds of essential oils, their roles in human health and potential application as natural food preservatives. Food Chem X. 2022;13:100217.
26. Gallego-Jara J, Lozano-Terol G, Sola-Martínez RA, Cánovas-Díaz M, de Diego Puente T. A Compressive Review about Taxol®: History and Future Challenges. Molecules. 2020;25:5986.
27. Sousa-Pimenta M, Estevinho LM, Szopa A, Basit M, Khan K, Armaghan M, et al. Chemotherapeutic properties and side-effects associated with the clinical practice of terpene alkaloids: paclitaxel, docetaxel, and cabazitaxel. Frontiers in Pharmacology. 2023;14.
28. Lin H-C, Chooi Y-H, Dhingra S, Xu W, Calvo AM, Tang Y. The Fumagillin Biosynthetic Gene Cluster in Aspergillus fumigatus Encodes a Cryptic Terpene Cyclase Involved in the Formation of β-trans-Bergamotene. J Am Chem Soc. 2013;135:4616–9.
29. Jiang K, Du C, Huang L, Luo J, Liu T, Huang S. Phylotranscriptomics and evolution of key genes for terpene biosynthesis in Pinaceae. Frontiers in Plant Science. 2023;14.
30. Wilson K, de Rond T, Burkhardt I, Steele TS, Schäfer RJB, Podell S, et al. Terpene biosynthesis in marine sponge animals. Proceedings of the National Academy of Sciences. 2023;120:e2220934120.
31. Rabara RC, Kudithipudi C, Timko MP. Identification of Terpene-Related Biosynthetic Gene Clusters in Tobacco through Computational-Based Genomic, Transcriptomic, and Metabolic Analyses. Agronomy. 2023;13:1632.

32. Yuan Y, Cheng S, Bian G, Yan P, Ma Z, Dai W, et al. Efficient exploration of terpenoid biosynthetic gene clusters in filamentous fungi. Nat Catal. 2022;5:277–87.
33. Singh G, Dal Grande F, Schmitt I. Genome mining as a biotechnological tool for the discovery of novel biosynthetic genes in lichens. Frontiers in Fungal Biology. 2022;3.
34. Rao PS, Sarma KG, Seshadri TR. Chemical Components of the Lobaria Lichens from the Western Himalayas. Current Science. 1965;34:9–11.
35. Malíček J. A revision of the epiphytic species of the Lecanora subfusca group (Lecanoraceae, Ascomycota) in the Czech Republic. The Lichenologist. 2014;46:489–513.
36. Patnana DP, Biswal RP, Dandamudi RB, S C, Pandey M. Simple HPLC-DAD-based method for determination of ergosterol content in lichens and mushrooms. Journal of Liquid Chromatography & Related Technologies. 2021;44:229–34.
37. Yosioka I, Yamaki M, Kitagawa I. On the triterpenic constituents of a lichen, Parmelia entotheiochroa Hue.; zeorin, leucotylin, leucotylic acid, and five new related triterpenoids. Chem Pharm Bull (Tokyo). 1966;14:804–7.
38. Paukov A, Teptina A, Ermoshin A, Kruglova E, Shabardina L. The Role of Secondary Metabolites and Bark Chemistry in Shaping Diversity and Abundance of Epiphytic Lichens. Frontiers in Forests and Global Change. 2022;5.
39. Singh G, Calchera A, Schulz M, Drechsler M, Bode HB, Schmitt I, et al. Climate-specific biosynthetic gene clusters in populations of a lichen-forming fungus. Environmental Microbiology. 2021;23:4260–75.
40. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Molecular Biology and Evolution. 2021;38:4647–54.
41. Yamada KD, Tomii K, Katoh K. Application of the MAFFT sequence alignment program to large data-reexamination of the usefulness of chained guide trees. Bioinformatics. 2016;32:3246–51.
42. Rozewicki J, Li S, Amada KM, Standley DM, Katoh K. MAFFT-DASH: integrated protein sequence and structural alignment. Nucleic Acids Research. 2019;47:W5–10.
43. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Molecular Biology and Evolution. 2015;32:268–74.
44. Rambaut A. FigTree. 2010.
45. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Research. 2021;49:W293–6.
46. Blin K, Shaw S, Kloosterman AM, Charlop-Powers Z, van Wezel GP, Medema MH, et al. antiSMASH 6.0: improving cluster detection and comparison capabilities. Nucleic Acids Research. 2021;49:W29–35.

47. Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW, Kautsar SA, Tryon JH, Parkinson EI, et al. A computational framework to explore large-scale biosynthetic diversity. Nature Chemical Biology. 2020;16:60–8.
48. Gilchrist CLM, Chooi Y-H. clinker & clustermap.js: automatic generation of gene cluster comparison figures. Bioinformatics. 2021;37:2473–5.
49. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc. 2015;10:845–58.
50. Gautier R, Douguet D, Antonny B, Drin G. HELIQUEST: a web server to screen sequences with specific alpha-helical properties. Bioinformatics. 2008;24:2101–2.
51. Linscott KB, Niehaus TD, Zhuang X, Bell SA, Chappell J. Mapping a kingdom-specific functional domain of squalene synthase. Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids. 2016;1861:1049–57.
52. Culberson CF. Biogenetic Relationships of the Lichen Substances in the Framework of Systematics. The Bryologist. 1986;89:91–8.
53. Bohlmann J, Meyer-Gauen G, Croteau R. Plant terpenoid synthases: Molecular biology and phylogenetic analysis. Proceedings of the National Academy of Sciences. 1998;95:4126–33.
54. Pichersky E, Raguso RA. Why do plants produce so many terpenoid compounds? New Phytologist. 2018;220:692–702.
55. Tan Y, Yang B, Lin X, Luo X, Pang X, Tang L, et al. Nitrobenzoyl Sesquiterpenoids with Cytotoxic Activities from a Marine-Derived Aspergillus ochraceus Fungus. J Nat Prod. 2018;81:92–7.
56. Liu H-X, Zhang L, Chen Y-C, Sun Z-H, Pan Q-L, Li H-H, et al. Monoterpenes and sesquiterpenes from the marine sediment-derived fungus Eutypella scoparia FS46. J Asian Nat Prod Res. 2017;19:145–51.
57. Jiang M, Wu Z, Guo H, Liu L, Chen S. A Review of Terpenes from Marine-Derived Fungi: 2015–2019. Mar Drugs. 2020;18:321.
58. Robinson GW, Tsay YH, Kienzle BK, Smith-Monroy CA, Bishop RW. Conservation between human and fungal squalene synthetases: similarities in structure, function, and regulation. Mol Cell Biol. 1993;13:2706–17.
59. Baker MAB, Brown AJ. A detour to sterol synthesis. Nat Microbiol. 2019;4:214–5.
60. Alcazar-Fuoli L, Mellado E. Ergosterol biosynthesis in Aspergillus fumigatus: its relevance as an antifungal target and role in antifungal drug resistance. Frontiers in Microbiology. 2013;3.
61. Liu J-F, Xia J-J, Nie K-L, Wang F, Deng L. Outline of the biosynthesis and regulation of ergosterol in yeast. World J Microbiol Biotechnol. 2019;35:98.
62. Jordan IK, Rogozin IB, Wolf YI, Koonin EV. Essential Genes Are More Evolutionarily Conserved Than Are Nonessential Genes in Bacteria. Genome Res. 2002;12:962–8.
63. Shakhnovich BE, Koonin EV. Origins and impact of constraints in evolution of gene families. Genome Res. 2006;16:1529–36.

64. Luo H, Gao F, Lin Y. Evolutionary conservation analysis between the essential and nonessential genes in bacterial genomes. Sci Rep. 2015;5:13210.
65. Daselva EJ, Englund B. Occurrence of Tocopherol and Ergosterol in Swedish Lichens. The Lichenologist. 1974;6:96–9.
66. Sundberg B, Ekblad A, Näsholm T, Palmqvist K. Lichen respiration in relation to active time, temperature, nitrogen and ergosterol concentrations. Functional Ecology. 1999;13:119–25.
67. Chang W, Zhang M, Li Y, Li X, Gao Y, Xie Z, et al. Lichen endophyte derived pyridoxatin inactivates Candida growth by interfering with ergosterol biosynthesis. Biochimica et Biophysica Acta (BBA) - General Subjects. 2015;1850:1762–71.
68. Mellado E, Diaz-Guerra TM, Cuenca-Estrella M, Rodriguez-Tudela JL. Identification of Two Different 14-α Sterol Demethylase-Related Genes (cyp51A and cyp51B) in Aspergillus fumigatus and Other Aspergillus species. Journal of Clinical Microbiology. 2001;39:2431–8.
69. Alcazar-Fuoli L, Mellado E. Ergosterol biosynthesis in Aspergillus fumigatus: its relevance as an antifungal target and role in antifungal drug resistance. Front Microbiol. 2013;3:439.
70. Rodrigues ML. The Multifunctional Fungal Ergosterol. mBio. 2018;9:e01755-18.
71. Gerasimova JV, Beck A, Werth S, Resl P. High Diversity of Type I Polyketide Genes in Bacidia rubella as Revealed by the Comparative Analysis of 23 Lichen Genomes. J Fungi (Basel). 2022;8:449.
72. Hata S, Sanmiya K, Kouchi H, Matsuoka M, Yamamoto N, Izui K. cDNA cloning of squalene synthase genes from mono- and dicotyledonous plants, and expression of the gene in rice. Plant Cell Physiol. 1997;38:1409–13.
73. Huang Z, Jiang K, Pi Y, Hou R, Liao Z, Cao Y, et al. Molecular cloning and characterization of the yew gene encoding squalene synthase from Taxus cuspidata. J Biochem Mol Biol. 2007;40:625–35.
74. Uchida H, Yamashita H, Kajikawa M, Ohyama K, Nakayachi O, Sugiyama R, et al. Cloning and characterization of a squalene synthase gene from a petroleum plant, Euphorbia tirucalli L. Planta. 2009;229:1243–52.
75. Devarenne TP, Shin DH, Back K, Yin S, Chappell J. Molecular characterization of tobacco squalene synthase and regulation in response to fungal elicitor. Arch Biochem Biophys. 1998;349:205–15.
76. Unland K, Pütter KM, Vorwerk K, van Deenen N, Twyman RM, Prüfer D, et al. Functional characterization of squalene synthase and squalene epoxidase in Taraxacum koksaghyz. Plant Direct. 2018;2:e00063.
77. Hayashi H, Hirota A, Hiraoka N, Ikeshiro Y. Molecular cloning and characterization of two cDNAs for Glycyrrhiza glabra squalene synthase. Biol Pharm Bull. 1999;22:947–50.
78. Suzuki H, Achnine L, Xu R, Matsuda SPT, Dixon RA. A genomics approach to the early stages of triterpene saponin biosynthesis in Medicago truncatula. The Plant Journal. 2002;32:1033–48.
79. Busquets A, Keim V, Closa M, del Arco A, Boronat A, Arró M, et al. Arabidopsis thaliana contains a single gene encoding squalene synthase. Plant Mol Biol. 2008;67:25–36.

80. Kim T-D, Han J-Y, Huh GH, Choi Y-E. Expression and functional characterization of three squalene synthase genes associated with saponin biosynthesis in Panax ginseng. Plant Cell Physiol. 2011;52:125–37.
81. Suyama M, Bork P. Evolution of prokaryotic gene order: genome rearrangements in closely related species. Trends in Genetics. 2001;17:10–3.
82. Hurst LD, Pál C, Lercher MJ. The evolutionary dynamics of eukaryotic gene order. Nat Rev Genet. 2004;5:299–310.
83. Dávila López M, Martínez Guerra JJ, Samuelsson T. Analysis of gene order conservation in eukaryotes identifies transcriptionally and functionally linked genes. PLoS One. 2010;5:e10654.
84. Gherman A, Wang R, Avramopoulos D. Orientation, distance, regulation and function of neighbouring genes. Hum Genomics. 2009;3:143–56.
85. Proctor RH, McCormick SP, Kim H-S, Cardoza RE, Stanley AM, Lindo L, et al. Evolution of structural diversity of trichothecenes, a family of toxins produced by plant pathogenic and entomopathogenic fungi. PLoS Pathog. 2018;14:e1006946.
86. Shen L, Porée F, Gaslonde T, Lalucque H. Functional characterization of the sterigmatocystin secondary metabolite gene cluster in the filamentous fungus Podospora anserina : involvement in oxidative stress response , sexual development , pigmentation and interspecific competitions. https://doi.org/10.1111/1462-2920.14698.

Figures

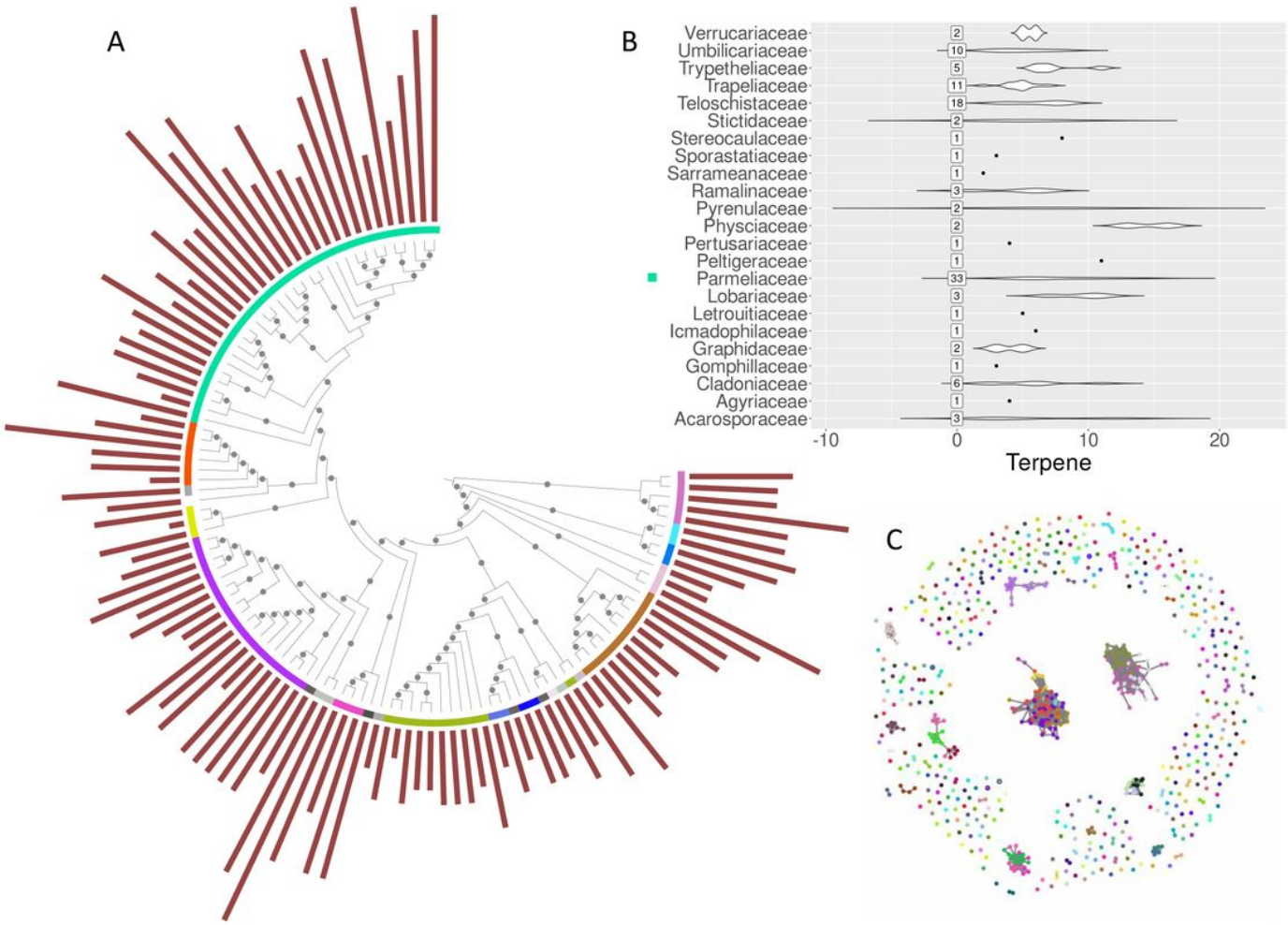

Figure 1

Distribution of terpene biosynthetic genes. A) The ML phylogenomic tree based on single copy BUSCOs of 112 taxa with black circles representing the >70% bootstrap support. Brown bars depict the number of terpene clusters present in each taxon. Although the number of terpenes varies across taxa, all lichens have terpene BGCs in their genomes. B) A violin plot showing the number of terpenes BGCs in each family. The numbers on the plot are the number of taxa of the family included in the study. Physiaceae are the most BGC rich family. C) BiG-SCAPE clustering (threshold 0.6) showing the clustering of BGCs into gene cluster families based on gene homology and cluster synteny. Each dot in the cluster represents a terpene BGC from a taxon and lines indicate the presence of a similar cluster in another taxa. Most terpene clusters ended up as individual dots with no connections, indicating that they are lineage specific and have a restricted taxonomic distribution. Two big terpene clans were detected, represented by the clustered dots connected by lines, suggesting the widespread and conserved presence of two terpene clans in LFF. These are hereafter referred as Clan1 and Clan2.

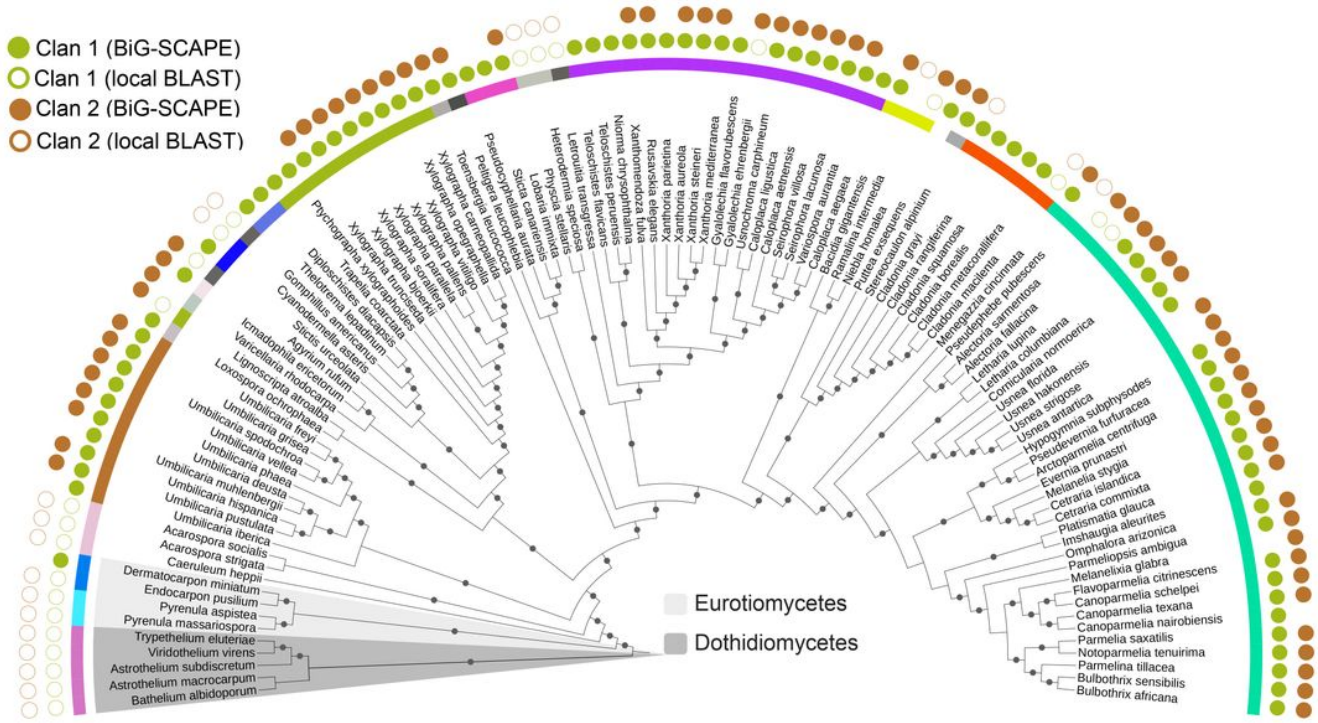

Figure 2

Phylogenetic distribution of conserved terpene BGC clans. A phylogenomic tree depicting the distribution of conserved terpene BGC clans in lichens as seen in figure 1C as well as in the outgroup taxa. Colored empty circles represent the taxa where the BGC/core gene was not detected by BiG-SCAPE but only by the local sequence similarity search using the sequence of the core gene. 75.67% (84 taxa) taxa contain both conserved clans. The core gene of both clans is a squalene synthase homolog, but different accessory genes, indicating a gene duplication of squalene synthase in lichens.

A

Clan	BIG-SCAPE Gene cluster families	#taxa	#Phylogenetic families	Phylogenetic families
Clan 1	FAM02526	35	11	Parmeliaceae, Cladoniaceae, Graphidaceae, Icmadophilaceae, Teloschistaceae, Lobariaceae, Sarrameanaceae, Sporastatiaceae, Umbilicariaceae, Verrucariaceae
	FAM00230	22	8	Parmeliaceae, Ramalinaceae, Peltigeraceae, Lobariaceae, Trapeliaceae, Stictidaceae, Trapeliaceae, Umbilicariaceae
	FAM03031	7	2	Parmeliaceae, Cladoniacea
	FAM03070	6	5	Parmeliaceae, Cladoniacea, Stereocaulaceae, Teloschistaceae, One uncertain
	FAM07143	8	1	Trapeliaceae
	FAM06129	5	1	Umbilicariaceae
Clan2	FAM03199	27	7	Parmeliaceae, Icmadophilaceae, Lobariaceae, Stereocaulaceae, Teloschistaceae, Umbilicariaceae
	FAM00121	26	6	Agyriaceae, Incertae sedis, Ramalinaceae, Teloschistaceae, Trapeliaceae, Umbilicariaceae
	FAM07263	15	6	Parmeliaceae, Peltigeraceae, Sporastatiaceae, Teloschistaceae, Trapeliaceae, Umbilicariaceae

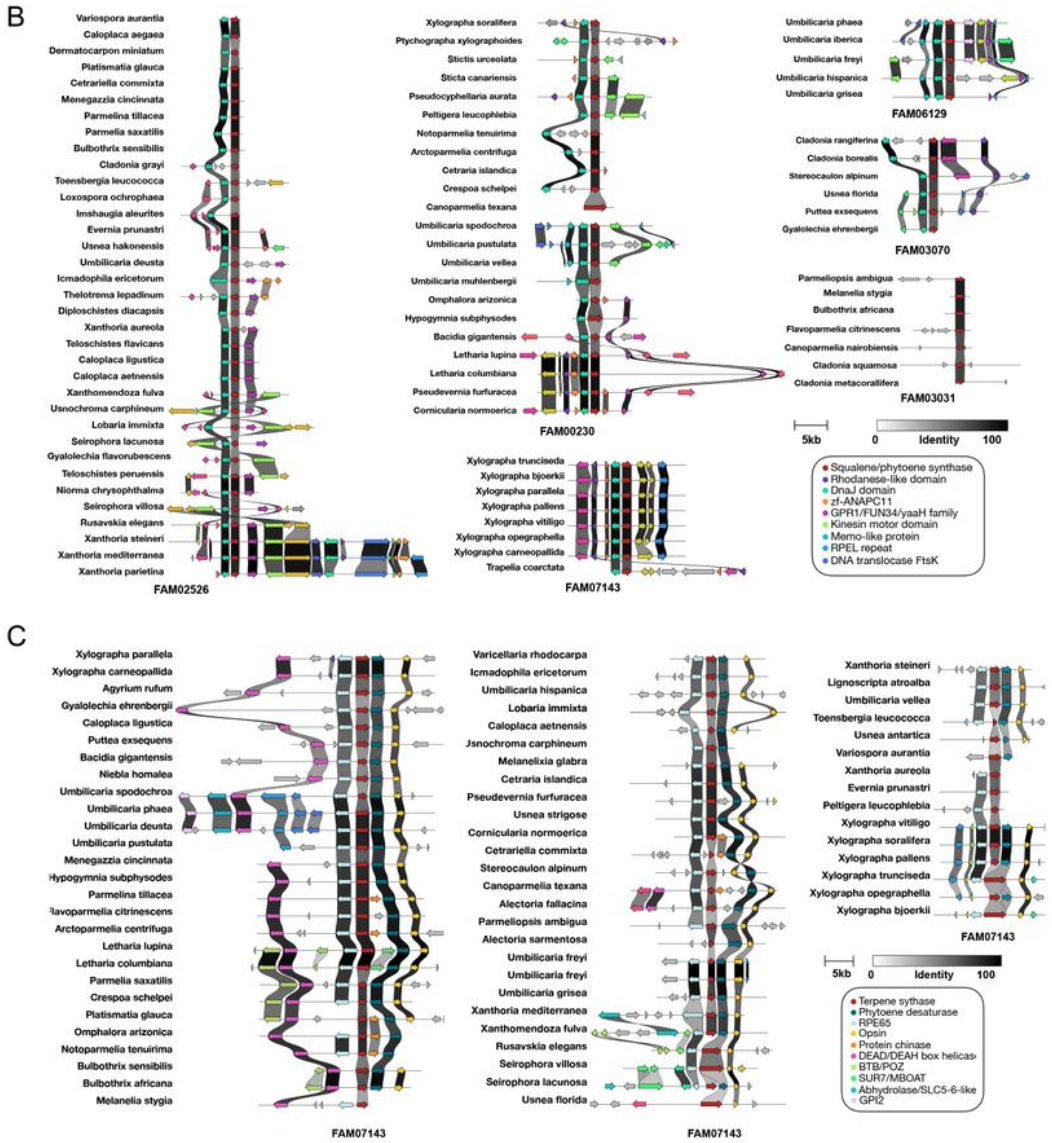

Figure 3

Clinker plot showing the cluster synteny among squalene/phytoene cluster **A**) Table giving the overview of the gene cluster families of the two terpene clans. A clan has a few gene cluster families, each containing several homologous terpene clusters. Gene cluster families represent homologous clusters grouped together based on their cluster homology and synteny as according to BiG-SCAPE (threshold 0.6). Clan1 has six cluster families whereas Clan 2 has three; **B)** Clinker plot showing the synteny among

squalene/phytoene clusters of the six families belonging to Clan1 among taxa. Notably while the position of accessory gene is not conserved among taxa and certain genes as DNAJ protein maybe present immediately upstream of the squalene synthase or after a few genes. Similarly, Keinesin motor domain (represented by florescent green arrow in FAM02526) can be present upstream or downstream of the squalene synthase. **C)** Clinker plot showing the synteny among squalene/phytoene clusters of the three families belonging to Clan2. As in clan1 gene organization and location is variable among taxa. Four genes – RPE65, squalene synthase (core gene of the cluster), phytoene desaturase and Opsin – are common among the three families of the Clan2. While gene arrangement in most cases follows the above-mentioned order, there are some interesting exceptions. For instance, in some cases protein Kinase is present between squalene synthase and phytoene desaturase. Similarly, although DEAD helicase is conserved and present in most taxa, its position of variable among taxa. In rare cases genes as protein kinase are sporadically present between the squalene synthase and phytoene desaturase. This indicates gene loss/gene gains are common in the evolution of this gene cluster family.

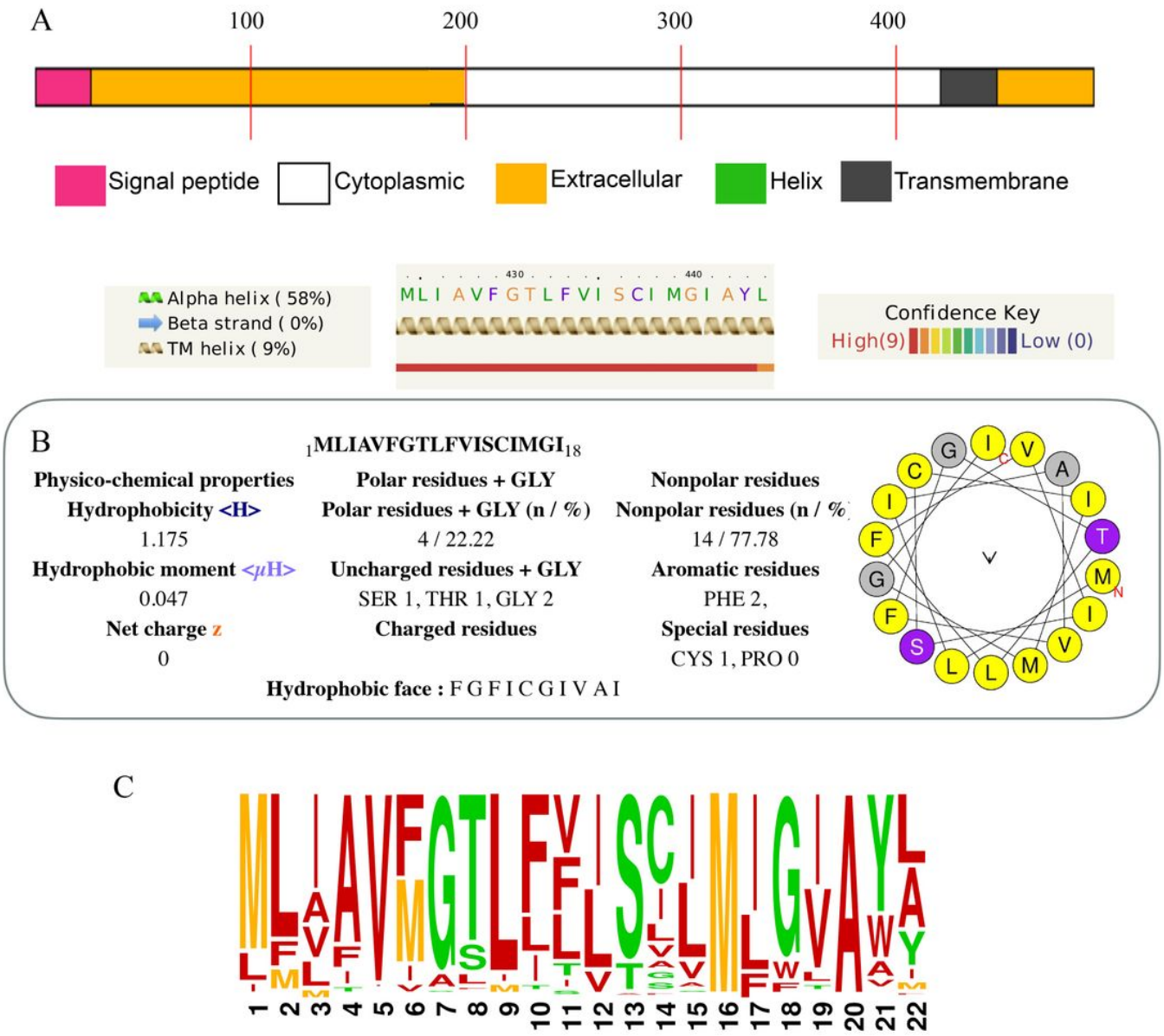

Figure 4

A) The domains and motifs of LFF squalene synthase derived from consensus sequence as predicted by DeepTMHMM. The program detected a signal peptide, extracellular and cytoplasmic regions at the N terminus and a transmembrane region at the C terminus. B) The high hydrophobicity of the predicted transmembrane region further supports the region to be transmembrane. C) Conservation of fungal-specific region of the squalene synthase across Lecanoromycetes.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarymaterial1.xlsx
- SupplementarymaterialMaterialsandMethods.docx
- Supplementarymaterial3.xlsx
- Supplementarymaterial4.pdf