nature microbiology

ARTICLES

PUBLISHED: 3 APRIL 2017 | VOLUME: 2 | ARTICLE NUMBER: 17044

Global analysis of biosynthetic gene clusters reveals vast potential of secondary metabolite production in *Penicillium* species

Jens Christian Nielsen[1], Sietske Grijseels[2], Sylvain Prigent[1], Boyang Ji[1], Jacques Dainat[3,4], Kristian Fog Nielsen[2], Jens Christian Frisvad[2], Mhairi Workman[2] and Jens Nielsen[1,5]*

Filamentous fungi produce a wide range of bioactive compounds with important pharmaceutical applications, such as antibiotic penicillins and cholesterol-lowering statins. However, less attention has been paid to fungal secondary metabolites compared to those from bacteria. In this study, we sequenced the genomes of 9 *Penicillium* species and, together with 15 published genomes, we investigated the secondary metabolism of *Penicillium* and identified an immense, unexploited potential for producing secondary metabolites by this genus. A total of 1,317 putative biosynthetic gene clusters (BGCs) were identified, and polyketide synthase and non-ribosomal peptide synthetase based BGCs were grouped into gene cluster families and mapped to known pathways. The grouping of BGCs allowed us to study the evolutionary trajectory of pathways based on 6-methylsalicylic acid (6-MSA) synthases. Finally, we cross-referenced the predicted pathways with published data on the production of secondary metabolites and experimentally validated the production of antibiotic yanuthones in Penicillia and identified a previously undescribed compound from the yanuthone pathway. This study is the first genus-wide analysis of the genomic diversity of Penicillia and highlights the potential of these species as a source of new antibiotics and other pharmaceuticals.

In the past decades, the discovery of new antibiotics has not kept pace with the development of antibiotic-resistant infectious microorganisms[1]. Historically, actinomycetes have been the main source of antibiotics, and filamentous fungi have not been studied to the same extent, despite their ability to produce diverse arrays of secondary metabolites (SMs)[2]. Filamentous fungi are less studied and characterized at the genomic level, and genomic insight is a key element in efficiently unlocking their potential for new antibiotics and other pharmaceuticals.

In recent years, secondary metabolite research has benefited extensively from the development of bioinformatics analysis tools, which enable researchers to mine genomes for the identification of genes involved in the synthesis of secondary metabolites[3–5]. These tools exploit the fact that genes involved in the biosynthetic pathway of secondary metabolites tend to cluster together in the genome in biosynthetic gene clusters (BGCs). BGCs are often composed of a backbone gene, predominantly polyketide synthases (PKSs) and non-ribosomal peptide synthetases (NRPSs), which gene products, polymerize monomers to generate a chemical backbone structure that is subject to modifications by tailoring enzymes. PKSs and NRPSs are multidomain megaenzymes that use short-chain carboxylic acids and amino acids, respectively, as the substrate for condensation reactions carried out by ketoacyl synthase (KS) domains in PKSs and condensation (C) domains in NRPSs (ref. 2).

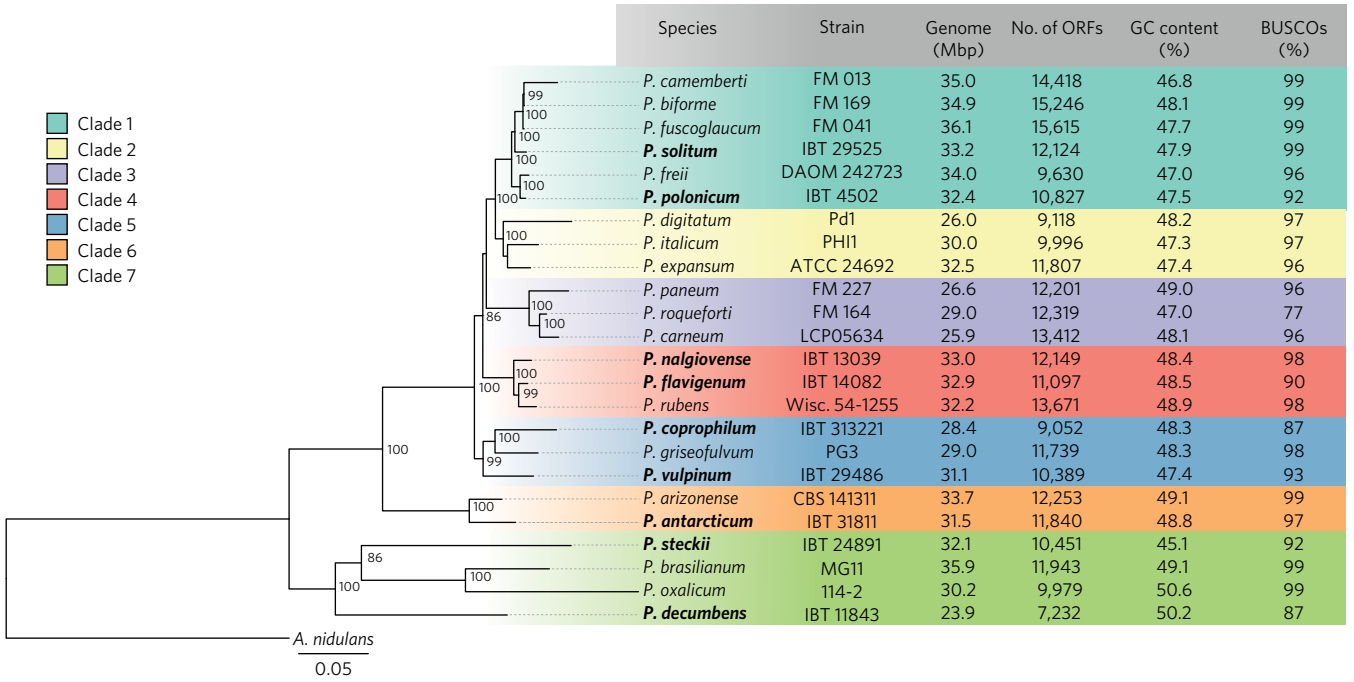
The *Penicillium* genus consists of more than 354 species, of which several are of industrial importance[6]. Most prominent is the industrial penicillin producer *Penicillium rubens* (often identified as *P. chrysogenum*), but other Penicillia are used for the production of other pharmaceutical secondary metabolites[7], enzymes such as cellulases[8] and in the manufacture of food products[9]. They play an important role in mediating phosphate to plants[10], which has great impact on agriculture[11], and their production of mycotoxins upon the contamination of food and feed is a serious health concern in humans and animals[12]. The implications *Penicillium* species have on humans are widely associated with their ability to produce bioactive secondary metabolites, and to better understand how we can harness the production of these for biotechnological exploitation, or to limit them where they are unwanted, it is important to assess the entire arsenal of secondary metabolite biosynthetic pathways in these species.

Here, we have sequenced the genomes of nine *Penicillium* species and gathered available *Penicillium* genomes from the literature. In total, we mined 24 genomes for BGCs and associated PKS and NRPS BGCs to known pathways. This enabled the prediction of yanuthone production in *Penicillium*, which was experimentally validated, together with the identification of a previously undescribed compound from the yanuthone pathway, in *Penicillium flavigenum*.

Results

Genome features of the *Penicillium* genus. The genomes of nine *Penicillium* species, *P. antarcticum*, *P. coprophilum*, *P. decumbens*, *P. flavigenum*, *P. nalgiovense*, *P. polonicum*, *P. solitum*, *P. steckii* and *P. vulpinum*, were sequenced, assembled and annotated as described in the Methods, and analysed together with 15 published *Penicillium* genomes (Supplementary Table 1). All genomes were 23.9–36.1 Mbp in size and contained 7,232–15,615 genes (Fig. 1). The genomes sequenced in this study generally showed characteristics similar to those of previously published

[1]Department of Biology and Biological Engineering, Chalmers University of Technology, SE412 96 Gothenburg, Sweden. [2]Department of Biotechnology and Biomedicine, Technical University of Denmark, DK2800 Kgs. Lyngby, Denmark. [3]Department of Medical Biochemistry and Microbiology, Uppsala University, 751 23 Uppsala, Sweden. [4]National Bioinformatics Infrastructure Sweden (NBIS), SciLifeLab, Uppsala University, 752 37 Uppsala, Sweden. [5]Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, DK2800 Kgs. Lyngby, Denmark. *e-mail: nielsenj@chalmers.se



Species	Strain	Genome (Mbp)	No. of ORFs	GC content (%)	BUSCOs (%)
P. camemberti	FM 013	35.0	14,418	46.8	99
P. biforme	FM 169	34.9	15,246	48.1	99
P. fuscoglaucum	FM 041	36.1	15,615	47.7	99
P. solitum	IBT 29525	33.2	12,124	47.9	99
P. freii	DAOM 242723	34.0	9,630	47.0	96
P. polonicum	IBT 4502	32.4	10,827	47.5	92
P. digitatum	Pd1	26.0	9,118	48.2	97
P. italicum	PHI1	30.0	9,996	47.3	97
P. expansum	ATCC 24692	32.5	11,807	47.4	96
P. paneum	FM 227	26.6	12,201	49.0	96
P. roqueforti	FM 164	29.0	12,319	47.0	77
P. carneum	LCP05634	25.9	13,412	48.1	96
P. nalgiovense	IBT 13039	33.0	12,149	48.4	98
P. flavigenum	IBT 14082	32.9	11,097	48.5	90
P. rubens	Wisc. 54-1255	32.2	13,671	48.9	98
P. coprophilum	IBT 313221	28.4	9,052	48.3	87
P. griseofulvum	PG3	29.0	11,739	48.3	98
P. vulpinum	IBT 29486	31.1	10,389	47.4	93
P. arizonense	CBS 141311	33.7	12,253	49.1	99
P. antarcticum	IBT 31811	31.5	11,840	48.8	97
P. steckii	IBT 24891	32.1	10,451	45.1	92
P. brasilianum	MG11	35.9	11,943	49.1	99
P. oxalicum	114-2	30.2	9,979	50.6	99
P. decumbens	IBT 11843	23.9	7,232	50.2	87

Figure 1 | Maximum likelihood phylogram and genome statistics of *Penicillium* species analysed in this study. The phylogram was determined from the protein sequence of 1,389 single-copy genes present in all genomes. *Aspergillus nidulans* was used as outgroup, bootstrap support is given from 100 bootstrap replicates, and the scale bar indicates the mean expected substitutions per site. Species marked in bold were sequenced in this study. BUSCO[57] describes the completeness of the genomes by assessing the presence of ubiquitous fungal single-copy genes. ORFs, open reading frames.

Penicillium species, albeit *P. steckii* showed a slightly lower GC content (45.1%) than the rest, and *P. decumbens* had the smallest genome (Fig. 1).

Phylogenetic analysis showed that the 24 species clustered into seven clades, with *P. decumbens*, *P. steckii*, *P. brasilianum* and *P. oxalicum* being the most distinct species, representing *Penicillium* subgenus *Aspergilloides* (clade 7) (Fig. 1). The remaining species represented members of *Penicillium* subgenus *Penicillium*, with the recently discovered *P. arizonense*[13] and the marine fungus *P. antarcticum*[14] diverging earliest and representing section *Canescentia* (clade 6), while 18 species formed a closely related branch, and were divided into their recently redefined sections[15], which was in agreement with the topology of the phylogram, that is, *Fasciculata* (clade 1), *Robsamsonia* (clade 2), *Roquefortorum* (clade 3), *Chrysogena* (clade 4) and *Penicillium* (clade 5) (Fig. 1).

Functional versatility. All 278,862 genes from the 24 species were clustered into orthologous groups based on protein sequence, and 3,249 gene families were shared by all species, hence representing the core genome, while 8,784 genes families were seen in a subset of the species (at least two) and represented the dispensable genome (Supplementary Fig. 1). A functional annotation using euKaryotic Orthologous Groups (KOGs) (Fig. 2a) proved that the core genome contained 39% non-metabolic and 37% metabolic genes, while the dispensable genome was dominated by 47% non-metabolic genes and 27% metabolic genes. The remaining genes were annotated within poorly characterized KOG categories. In the pangenome, secondary metabolism (Q) showed the greatest variation in gene content, followed by the remaining major parts of metabolism: carbohydrate metabolism (G), amino acid metabolism (E) and lipid metabolism (I) (Fig. 2a and Supplementary Fig. 2). Interestingly, secondary metabolism (Q) and carbohydrate metabolism (G) were mainly encoded in the core genome, while amino acid (E) and lipid metabolism (I) were more equally distributed among core and dispensable genes (Fig. 2a). We further analysed gene copy number relative to genome size, and showed that secondary ($P = 0.003$) and carbohydrate ($P = 0.047$) metabolic genes were enriched among gene families correlating with genome size (Supplementary Fig. 3). This highlights that these subsystems are under positive selection and drive the genome expansion in *Penicillium*, and also explains the high variation within these KOG categories. More specifically, the variation within secondary metabolism was mainly explained by variations in the content of cytochrome P450 genes (Supplementary Fig. 4), while carbohydrate active enzymes (CAZys) showed the greatest variation within glycoside hydrolases (GHs), ranging from 759 in *P. arizonense*, which was recently shown to encode a large number of CAZys[13], to 321 in *P. decumbens*.

Overview of BGCs. Secondary metabolism was further investigated by mining the genomes for BGCs using antiSMASH[16] (Fig. 2b). A total of 1,317 putative BGCs were detected in the 24 genomes, corresponding to an average of 54.9 BGCs per species (median 55), with the highest number of 78 BGCs being observed in *P. polonicum*. The most abundant class was PKS BGCs, with 463 in total and an average of 17.2 per species (median 18), and the second most abundant class was NRPS BGCs, with 275 in total and an average of 11.3 per species (median 12). Among the hybrid BGCs, the majority (140 of 195) were PKS–NRPS hybrids. Interestingly, the species from clade 1 contained more terpene BGCs than the rest, and generally showed a high number of BGCs encoded in their genomes (Fig. 2b).

BGC diversity. The two major BGC classes in *Penicillium* were based on PKSs and NRPSs and these were, including hybrids thereof, grouped into gene cluster families using core domains of the backbone genes (in total 467 PKSs, 260 NRPSs and 71 PKS–NRPS hybrids). We used the most conserved regions, the KS domain of PKSs and the C domains of NRPSs, to serve as a proxy for the similarity of the entire BGC, because these

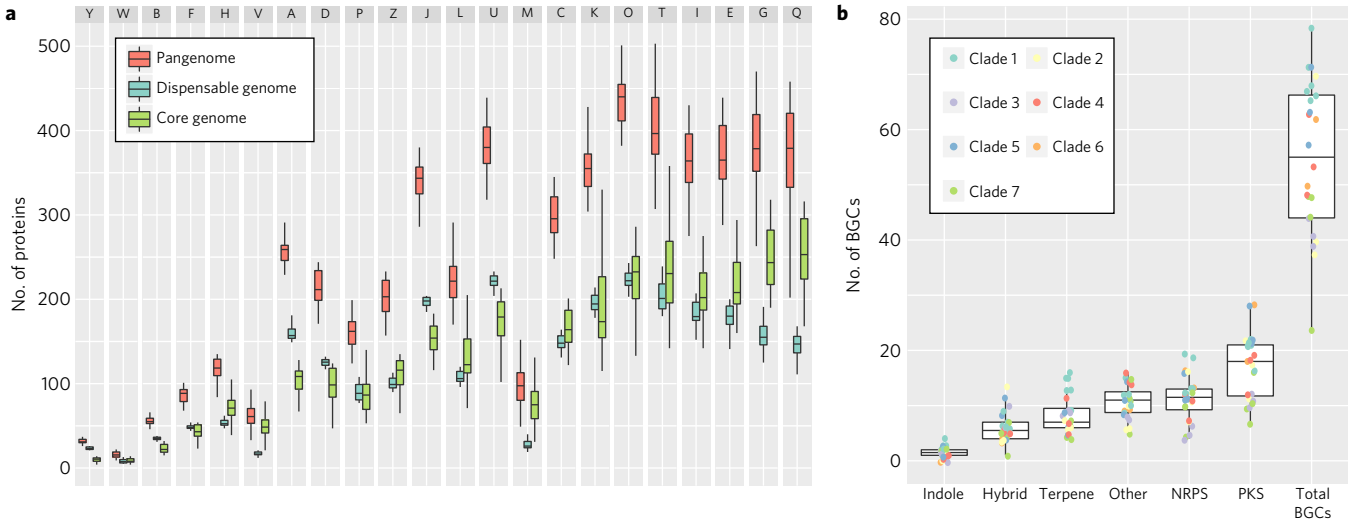


Figure 2 | Functional analysis of *Penicillium* species. a, Distribution of proteins allocated to different subsystems as defined by euKaryotic Orthologous Groups (KOGs). KOGs are sorted according to standard deviation in the pangenome. KOG categories are as follows. For cellular processes and signalling, M is cell wall/membrane/envelope biogenesis, O is post-translational modification, protein turnover and chaperones, T is signal transduction mechanisms, U is intracellular trafficking, secretion and vesicular transport, V is defence mechanisms, W is extracellular structures, Y is nuclear structure, and Z is cytoskeleton. For information storage and processing, A is RNA processing and modification, B is chromatin structure and dynamics, J is translation, ribosomal structure and biogenesis, K is transcription, and L is replication, recombination and repair. For metabolism, C is energy production and conversion, D is cell cycle control, cell division and chromosome partitioning, E is amino-acid transport and metabolism, F is nucleotide transport and metabolism, G is carbohydrate transport and metabolism, H is coenzyme transport and metabolism, I is lipid transport and metabolism, P is inorganic ion transport and metabolism, and Q is secondary metabolites biosynthesis, transport and catabolism. **b**, Distribution of BGC classes. Each boxplot represents the distribution of a class of BGCs as defined by antiSMASH[16]. The class 'Other' contains BGCs that do not fit into any of the predefined categories of antiSMASH and rare BGC classes that were present in a few species only. In the boxplots, the box indicates the first, second and third quartiles and the whiskers extend to the lowest/highest value that is within 1.5 IQR (inter-quartile range).

domains have previously been shown to be good evolutionary determinants[17–19].

To identify the diversity between the domains and to evaluate whether their similarity could be used to infer the similarity of BGCs, an all versus all alignment was performed (Fig. 3a). The majority of the KS domains showed a similarity between 20 and 60% identity, while the C domains were less similar and distributed mainly between 10 and 40% identity (Supplementary Fig. 5). However, a number of highly similar domains clustered together, and several of these generated super clusters, as indicated by the dark colours in Fig. 3a. These super clusters mainly represented PKSs with the same degree of reduction, or C domains with the same catalytic activity. The tendency to form distinct super clusters was more pronounced for the KS domains and might encompass PKSs generating scaffolds with similar chemical structure, as was observed (Supplementary Fig. 6).

Based on a detailed inspection of the similarity of the domains (Supplementary Fig. 5), an empirical cutoff of 74 and 64% identity for KS and C domains, respectively, was chosen to select domains (and by extension BGCs) that were shared between multiple species (Fig. 3b). Among the KS and C domains, 81 and 77%, respectively, grouped into a gene cluster family, meaning that 104 PKS and 148 NRPS BGCs in the *Penicillium* species studied here were unique to one species. In all species analysed, we found examples of both unique and shared BGCs.

The putative products of the BGC pathways were predicted by mapping the KS and C domains to the MIBiG database[20] of characterized BGCs, from which all fungal KS and C domains were extracted and aligned against the corresponding domains predicted in the *Penicillium* genomes. Among the predicted PKS-containing BGCs, 80 could be annotated to one of 21 different pathways (Fig. 3b and Supplementary Data 1). In the KS network, nine highly abundant gene cluster families were present in more than half the species, but only one of these could be assigned to a pathway, that is, biosynthesis of the green conidial pigment YWA1, which was found in all species. Second most abundant was a gene cluster family showing the same level of similarity to monodictyphenone and pestheic acid, which are both derived from emodin. Based on the low connectivity of this cluster in the network (Fig. 3b), we expect these pathways to have a diverse set of end products based on a conserved chemical core, similar to the emodin pathway in *Aspergillus nidulans*, which has been shown to encode more than ten different stable products[21]. Besides this gene cluster family, all other clusters in the network showed a high connectivity, indicating a consensus in the pathway predictions (Fig. 3b).

For the NRPS-containing BGCs, 51 could be annotated to one of nine different pathways (Fig. 3b and Supplementary Data 2). The most abundant NRPS BGC that could be annotated was fungisporin, which was observed in all species, and because the fungisporin NRPS contains two C domains, the same number of fungisporin clusters was seen in the network (Fig. 3b). The second most abundant annotated BGC was the roquefortin C/meleagrin pathway identified in ten species. Three hybrid PKS–NRPSs were among the predicted pathways (pseurotin A, chaetoglobosin and aspyridone), and because these contain both KS and C domains, they were predicted in both the PKS and NRPS analyses.

A total of three BGCs were predicted to be present in all 24 *Penicillium* genomes and represented one PKS BGC and two NRPS BGCs (containing two C domains each). In the previous section, we showed that the genes categorized as secondary metabolism are, in the majority, part of the core genome in *Penicillium* (Fig. 2). However, when grouping BGCs based on the encoded biosynthetic pathways, only few proved to be conserved in all species (Fig. 3b).

Evolution of 6-MSA pathways. The BGCs of the mycotoxin patulin and the antibiotic yanuthone D were further investigated because

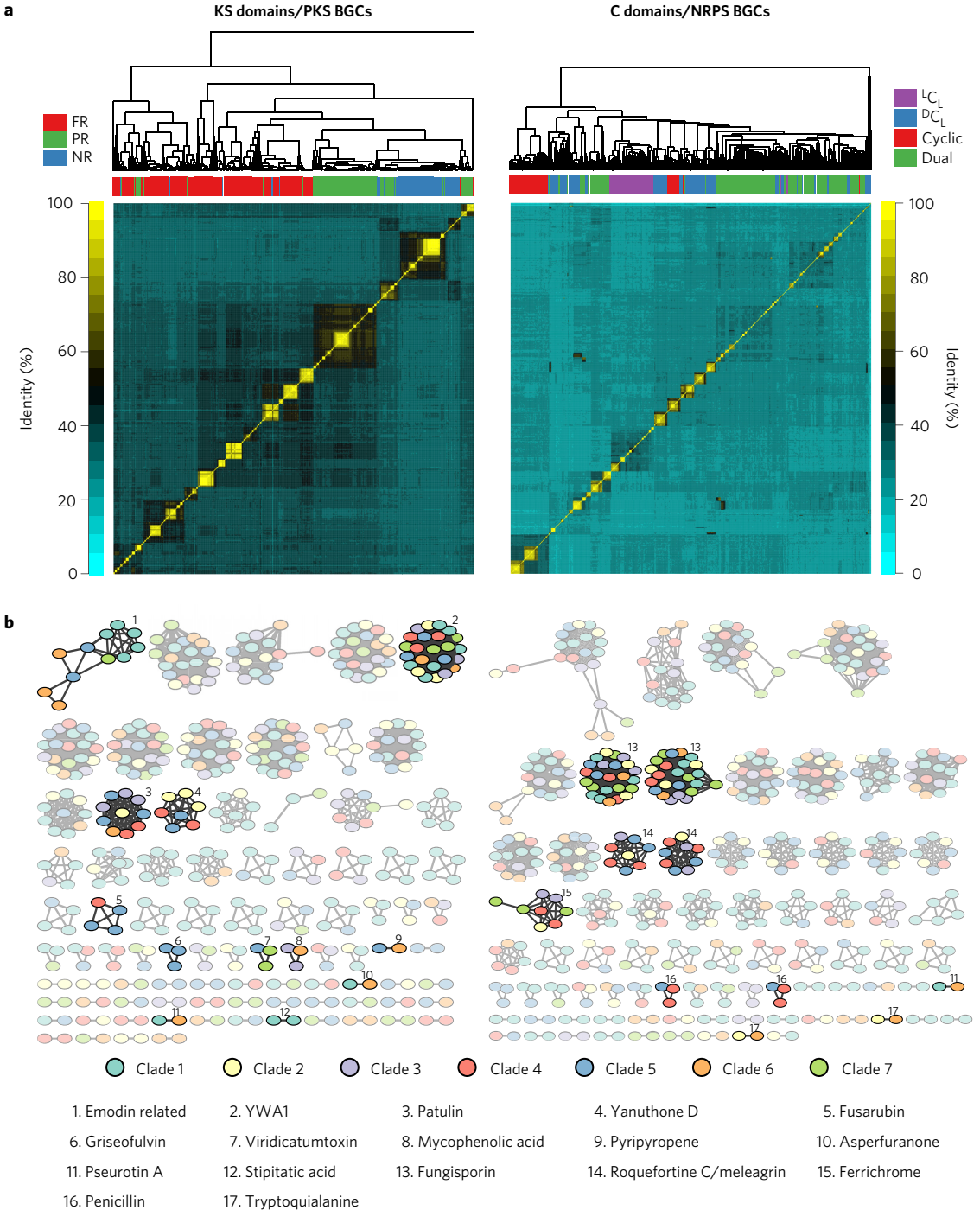


Figure 3 | Overview of the similarity of PKS and NRPS BGCs in *Penicillium* species. a, All versus all comparison of conserved protein domains (ketoacyl synthase (KS) and condensation (C) domains), based on pairwise sequence alignments. The KS domains were annotated based on the number of reducing domains in the corresponding PKS into either fully reducing (FR), partially reducing (PR) or non-reducing (NR). Clustering of the KS domains was based on a maximum likelihood phylogram. The C domains were annotated based on catalytic specificity into the following classes: 'cyclic', catalysing both peptide bond formation and subsequent cyclization; '$^{D}C_{L}$', linking an L-amino acid to a D-amino acid; 'dual', catalysing condensation and epimerization; '$^{L}C_{L}$', linking two L-amino acids. Clustering of C domains was based on Bray–Curtis dissimilarity. **b**, Network representation of PKS (left) and NRPS (right) BGCs shared by multiple species. Each node represents a domain and by extension a BGC, and the edges connect domains that are similar. Nodes are coloured according to the clade of the species and highlighted clusters represent gene cluster families mapped to the MIBiG database (Supplementary Data 1 and 2).

these two pathways synthesize similar secondary metabolites that are both based on a 6-methylsalicylic acid (6-MSA) core (Fig. 4a). Despite both PKSs (encoding 6-MSA synthases) catalyse the same reaction, they could be clearly distinguished based on KS domain similarity (Supplementary Fig. 7). The first three steps in the pathways are catalysed by orthologous enzymes (Supplementary Fig. 8) and convert one acetyl-CoA and three malonyl-CoA to toluquinol. Further conversion to gentisyl alcohol in the patulin pathway is catalysed by PatH, a paralogue of PatI (refs 22, 23) (Fig. 4a). The next step in both pathways is unknown, but could be catalysed by another set of orthologues, PatJ and YanE, as previously proposed[23], because these enzymes have co-evolved

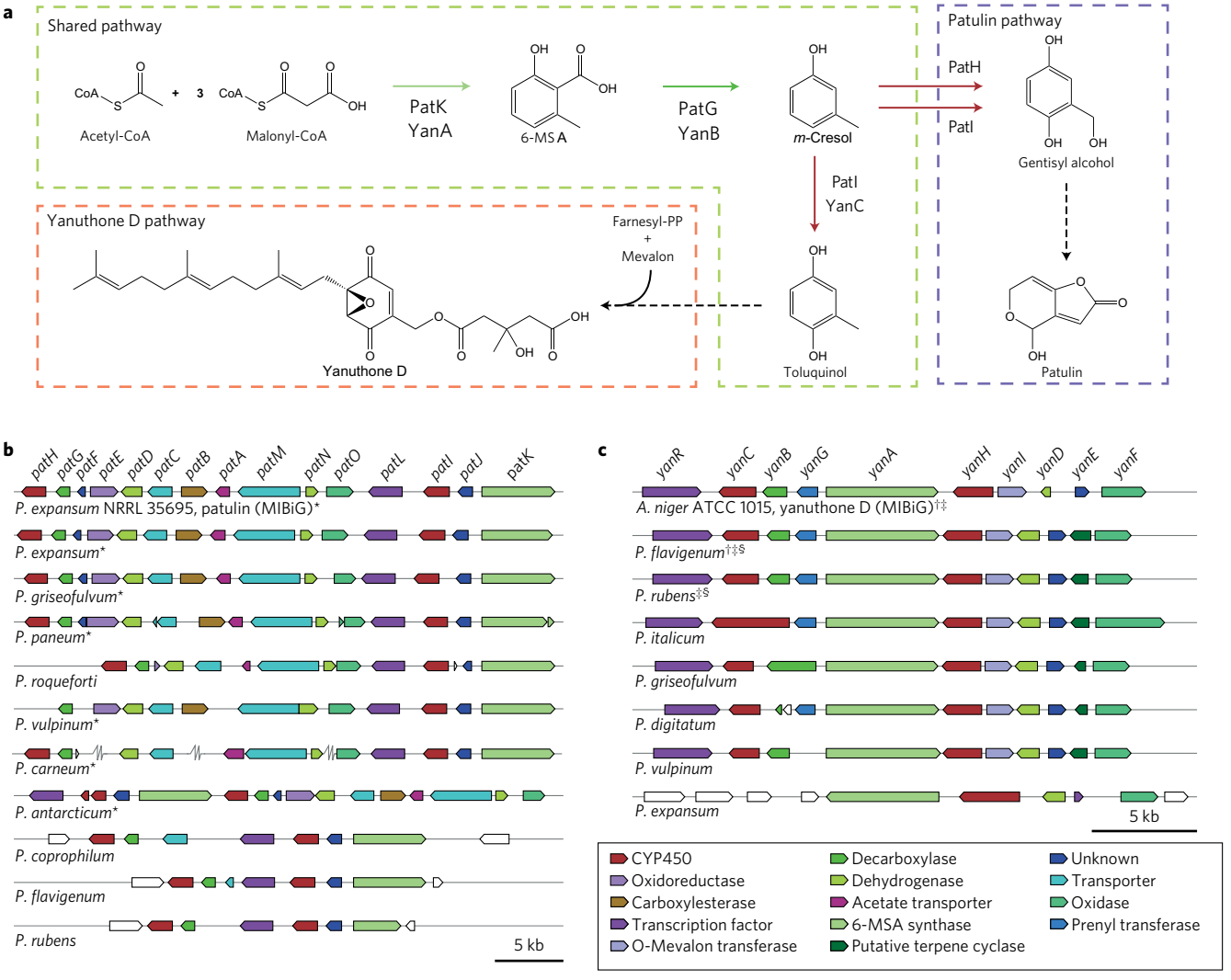


Figure 4 | Patulin and yanuthone D biosynthesis and BGCs. a, Biosynthetic pathways of patulin and yanuthone D. Dashed arrows indicate multiple steps. **b**, Predicted patulin BGCs relative to the version in *P. expansum* (MIBiG, BGC0000120) (ref. 76), **c**, Predicted yanuthone D BGCs relative to the version in *A. niger* (MIBiG, BGC0000170) (ref. 23). Gene functions are given according to this study and previous work[22,23]. Verified producers of *patulin, †yanuthone D and ‡yanuthone E (Supplementary Data 1). §Verified in this study.

with the remaining shared genes (Supplementary Fig. 9) and were conserved in all 6-MSA BGCs identified in this study, except in the partial yanuthone D BGC in *P. expansum* (Fig. 4b,c). To investigate the variation within the patulin and yanuthone D BGCs, gene synteny was assessed relative to the corresponding experimentally verified BGCs from the MIBiG database (Fig. 4b,c).

Patulin BGCs were predicted in ten species, of which seven contained orthologues of at least 12 of 15 genes in the reference cluster from *P. expansum* (Fig. 4b). Previous studies have confirmed these species as patulin producers[24,25], except *P. roqueforti*, which has not been seen to produce patulin, but is closely related to the patulin producers *P. paneum* and *P. carneum*[26]. For the remaining species (*P. coprophilum*, *P. flavigenum* and *P. rubens*) partial patulin BGCs were observed, which is in agreement with observations that *P. flavigenum* and *P. coprophilum* can synthesize intermediates of the patulin pathway such as the bioactive compound isoepoxydon (Supplementary Data 1). We also observed that the patulin BGCs in *P. antarcticum* had a unique gene architecture, which was rearranged compared to the other species, and is a possible evolutionary bridge between two previously described patulin BGC conformations in *A. clavatus*[22] and *P. expansum*[27] (Supplementary Fig. 10). Our findings also showed that the *P. expansum* patulin BGC architecture is dominant in *Penicillium*, correlating with the fact that *P. antarcticum* branches off as the first species containing the patulin BGC in the phylogenetic tree (Fig. 1 and Supplementary Fig. 10). These frequent rearrangements, even within the same genus, confirm the high level of plasticity of BGCs, which enables rapid development of alternative pathways and end products[28].

Yanuthone D BGCs were predicted in six species, but one additional species (*P. expansum*) clustered with these in the network (Fig. 3b) and hence was included in the syntenic evaluation (Fig. 4c). The synteny plot revealed, however, that the *P. expansum* BGC was fragmented compared to the reference cluster, and shared only five of ten genes with the yanuthone D BGC in *A. niger*. The other predicted yanuthone D BGCs showed homology with at least nine of ten genes. For *P. italicum* and *P. griseofulvum*, one missing gene could be due to a fusion of the *yanC-yanB* and *yanB-yanG* genes, respectively. Furthermore, an additional highly conserved gene in the yanuthone D BGCs of the *Penicillium* species was located between *yanE* and *yanF*, as previously observed in *P. griseofulvum*[29]. To annotate the gene, a protein BLAST against the NCBI non-redundant database was conducted, and the best hit was a putative integral membrane protein from *Botrytis cinerea* (EMR85854.1), showing 35% identity. Another gene predicted as

an integral membrane protein was previously annotated in the pyripyropene BGC in *A. fumigatus* (Pyr4), but proved to encode a membrane-bound terpene cyclase[30], and this enzyme showed 32% identity to the protein of the additional gene in the *Penicillium* yanuthone BGCs, indicating that it could be modifying the prenyl group of yanuthones and generate alternative yanuthone derivatives not seen in *A. niger*[23].

Examining the prevalence of 6-MSA pathways in *Penicillium* species, we found that 6-MSA synthase was present in either patulin or yanuthone D BGCs in 12 of 14 species in clades 2–6, but not in any species from clade 1 or 7, indicating that 6-MSA synthase could have been lost by the divergence of these clades (Supplementary Fig. 10). For the 6-MSA synthesizing species, five contained both patulin and yanuthone BGCs (or partial versions), while seven species contained only one of the two (Supplementary Fig. 10). The fact that the majority of the species have lost one of the two pathways suggests that a weak evolutionary pressure exists to lose either of the two pathways and could be a way to avoid cross-reactions of the pathways. Other 6-MSA synthases have been linked to pathways in *Aspergillus* species, for example, aculinic acid in *A. aculeatus*[31], terreic acid in *A. terreus*[32] and isoasperlactone and asperlactone in *A. westerdijkiae*[33], with the former two showing a closer evolutionary relationship to patulin 6-MSA synthases than those from yanuthone BGCs (Supplementary Fig. 11).

Production of secondary metabolites. In the above sections, sequence similarity was used to group PKS- and NRPS-containing BGCs into gene cluster families and annotate them to known pathways. To evaluate these predictions, we cross-referenced the available literature as well as secondary metabolites detected in this study (Supplementary Data 1 and 2) and could validate the production of the predicted or related compounds for 89 of the 127 pathway predictions. For the remaining unconfirmed cases, explanations could be that the BGCs are either silent under most conditions[34] or that we identified an inactive evolutionary reminiscence of a BGC (ref. 35).

Among the predicted pathways, which were initially unconfirmed, we decided to further investigate the production of yanuthones, for which a corresponding BGC was seen in seven species (Fig. 4c). We tested all seven for the production of yanuthones and related compounds, using liquid chromatography combined with high-resolution tandem mass spectrometry (LC-MS/HRMS) to search for the known compounds, as well as isomers generating fragment ions, as the available yanuthone standards (Supplementary Fig. 12).

In the crude extracts of *P. rubens* and *P. flavigenum*, we identified yanuthone E, and in the latter, also yanuthone D. In the extract of *P. flavigenum*, we also identified a compound that had a very similar fragmentation pattern as the yanuthones, but a different molecular mass (Supplementary Fig. 12). The MS/HRMS showed that it had the same polyketide core and prenyl unit as yanuthone E, but with the carboxylate unit having an elemental composition of $C_3H_4O_4$, which is most probably malonic acid (Supplementary Fig. 12). We also searched for compounds with a C_{10} prenyl chain, but did not detect any. The absence of detectable yanuthones in the remaining five species could be due to a lack of induction under the applied growth conditions, as production was found to be very specific for *A. niger*, where only solid YES medium could induce production and, even under these conditions, the yanuthones were only represented by minor peaks compared to other secondary metabolites[23]. In *P. rubens* and *P. flavigenum*, only PDA medium induced production, and here still as minor peaks.

Discussion

In this study, we have sequenced and annotated nine *Penicillium* species and hence significantly contributed to the number of publicly available *Penicillium* genomes. Together with the published genomes of additional *Penicillium* species, a comparative genomic analysis with a focus on secondary metabolism was conducted.

We found that secondary metabolism is the most variable cellular subsystem in terms of number of genes within Penicillia, and that genes involved in secondary metabolite biosynthesis also show the greatest expansion rate, meaning that these genes are enriched among the genes correlating with genome size. Expansion rate has previously been assessed in plant genomes (*Arabidopsis*, Soybean and Sorghum), and an enrichment of secondary metabolite biosynthetic genes was also observed[36]. Thus, it might be general for secondary-metabolite-producing organisms that adaptation by the acquisition of additional secondary metabolite biosynthetic genes is a fast and cost-efficient way to improve fitness within an environment.

In the *Penicillium* genomes, BGCs were identified, and we hypothesized that the conserved domains KS (from PKS) and C (from NRPS) could serve as a proxy for evaluating the similarity between BGCs containing these genes, but also for mapping them to other BGCs from more distant fungal genera. The idea for doing this was sparked by a previous bacterial study that used KS and C domains for comparing BGCs in 75 isolates of three different *Salinispora* species[19]. Our approach was similar, although we used a more promiscuous cutoff for grouping BGCs to account for the greater phylogenetic diversity in our genomes.

In total, we found 1,317 BGCs in 24 *Penicillium* genomes, of which 61% (798 in total) contained a PKS, NRPS or hybrid PKS–NRPS. We grouped these into gene cluster families and mapped them to characterized fungal pathways in the MIBiG database, which was recently updated with an additional 197 fungal BGCs (ref. 37). Despite this, only 16% (127 in total) of the PKS- and NRPS-derived *Penicillium* BGCs could be connected to a pathway, and the unannotated BGCs included 13 gene cluster families that were seen in more than half the species. This highlights the knowledge gap that exists between compounds and biosynthetic genes, even among widely abundant BGCs, and demonstrates the potential in further studying fungal BGCs with the objective of identifying novel antibiotics and other pharmaceutical compounds. There might be several of these BGCs where the product is known but has not been connected to the biosynthetic genes, and our work can aid in establishing these connections. One way of doing this is to assess the overlap in BGCs by two or more species and correlate this with their overlap in produced secondary metabolites, as previously demonstrated[38,39]. The work conducted in this study can greatly assist in this kind of analysis and provides a framework for looking up BGC overlaps in Penicillia and selecting BGCs of interest for further investigation.

Moreover, our work can be used as guidance for the discovery of previously unknown producers of known compounds that might be suitable as secondary-metabolite-producing cell factories[3] or for the identification of BGCs containing auxiliary genes, which might produce derivatives of known compounds that potentially could be of pharmaceutical value. We demonstrated this by identifying yanuthone BGCs in *Penicillium* species containing an extra gene with homology to a terpene cyclase, which is absent in the corresponding characterized BGC in *A. niger*[23]. We showed that *P. flavigenum* and *P. rubens* can produce yanuthone E, while *P. flavigenum* also produces the end product yanuthone D as well as a previously undescribed yanuthone, identified in this study. This yanuthone has the same polyketide core and prenyl group as yanuthone E, but probably with a malonic acid as carboxylate unit. Yanuthones have been shown to exhibit antifungal activity against *Candida albicans*[23,40], demonstrating their pharmaceutical potential. An intermediate in the *A. niger* yanuthone D pathway, 7-deacetoxyyanuthone, has previously been reported from a marine *Penicillium* species[41] as well as from *P. chrysogenum*[42].

Methods

Genome sequencing and assembly. Nine *Penicillium* species (*P. coprophilum*, *P. nalgiovense*, *P. polonicum*, *P. antarcticum*, *P. vulpinum*, *P. solitum*, *P. decumbens*, *P. flavigenum*, *P. steckii*) obtained from the culture collection (IBT) at the Department of Biotechnology and Biomedicine at the Technical University of Denmark were propagated, and genomic DNA was extracted as previously described[13]. DNA was prepared for sequencing using the TruSeq DNA PCRfree library kit (Illumina) according to the manufacturers protocol (#1503618, rev B) and sequenced on an Illumina HiSeq 2500 instrument. The resulting 125 bp paired-end reads, with an average insert size of 600 bp, were quality checked and assembled using various assembly tools: (ABySS (v.1.5.2)[43], SOAPdenovo (v.2.04)[44], SPAdes (v.3.5.0)[45] and MIRA (v.4.0.2)[46]). De Bruijn graph-based assemblers were tested with k-mers between 55 and 117, and the quality of the assemblies was assessed using QUAST[47] and FRCalign[48]. Genome assemblies from either SOAPdenovo (*P. nalgiovense*, *P. polonicum* and *P. flavigenum*) or ABySS (*P. coprophilum*, *P. antarcticum*, *P. vulpinum*, *P. solitum*, *P. decumbens* and *P. steckii*) showed the best overall quality and were used for further analysis.

Gene prediction. To increase the accuracy of gene prediction, repeat regions were annotated using the existing reference repeat library included in RepeatMasker (v.4.0.3)[49] supplemented with novel repeats detected with the RepeatModeler package (v.1.0.8)[50]. Candidate repeat sequences that were identified were checked against protein sequences from the manually curated database Swiss-prot[51] (retrieved 8th January 2015) to exclude any nucleotide motif stemming from low-complexity coding sequences. From this repeat library we used RepeatMasker and RepeatRunner to assign repeat sequences to genomic loci.

Genes were predicted in the masked genome by first training an *ab initio* gene predictor using GeneMark-ET (v.4.3)[52], which integrates RNA-seq evidence. For this we used an in-house data set of species-specific RNA-seq samples or 39 RNA-seq samples from related *Penicillium* species in the case of *P. antarcticum* and *P. solitum*, where species-specific RNA-seq data were not available. The RNA-seq reads were mapped to the individual genome assemblies using TopHat2 (v.2.0.9)[53] and cufflinks (v.2.2.1)[54] with settings –F 0.15 and –j 0.1. Further evidence was collected at the protein level from sequences belonging to Swiss-prot (547,357 proteins). Finally, the gene build was computed using Maker (v.2.31.6)[55] by combining the evidence sequences (proteins and transcripts) with the *ab initio* gene prediction generated by GeneMark.

Functional annotation. Predicted genes were functionally annotated based on the protein sequence queried against Swiss-prot to retrieve gene name and protein functions following the best-hit principle when the BLAST e-value was inferior to 1e-6. In addition, InterProScan (v.5.7–48)[56] was used to retrieve pathway information from several databases.

Comparative functional analyses. Genome sequences of *Penicillium* species not generated in this study were downloaded from the Joint Genome Institute (JGI) or the National Center for Biotechnology Information (NCBI) (Supplementary Table 1). The qualities of all genomes were assessed with BUSCO (v.1.1) using the data set for fungi, which detects the presence of ubiquitous fungal genes to evaluate the completeness of the gene prediction[57].

Proteins in all genomes were classified into KOGs (retrieved 21st August 2015 from http://ftp.ncbi.nih.gov/pub/mmdb/cdd/little_endian) using rpsblast, with an e-value cutoff set to 0.01, and core and dispensable genome fractions were defined based on protein groups generated with orthoMCL[58] (*evalueExponentCutoff*=1e-5 and *percentMatchCutoff*=50). Enzyme expansion rate was calculated as described previously[36] with EC numbers defined based on orthoMCL groups and the EC numbers given in the *P. rubens* genome-scale metabolic model[59] and hypergeometric test for determination of significance. For the annotation of CAZys, all proteins were mapped to the CAZy database using family-specific profile HMMs, downloaded from dbCAN[60] and scanned with the hmmscan program from the HMMer (v.3.1b2) software package. The resulting output was parsed with the hmmscan parser script supplied by dbCAN.

Phylogenetic analysis

Species phylogeny. Proteins in the 24 *Penicillium* genomes and *Aspergillus nidulans* (AACD00000000.1) were divided into orthologous groups using orthoMCL[58], and 1,389 single copy genes were shared in all genomes. For each of these single-copy genes, all sequences were aligned using MUSCLE[61], and the resulting multiple sequence alignment was trimmed for poorly aligned regions using Gblocks[62] with parameters –b2 = 50% –b5 = a. The aligned and trimmed sequences were concatenated to a single sequence of 479,008 amino acids. The data set was resampled using the seqboot bootstrap implementation from the phylip package[63], into one data set with 1e5 amino-acid positions for tree topology determination and 100 data sets with 1e4 amino-acid positions for evaluation of the bootstrap support values. A gamma distributed JTT substitution model (PROTGAMMAJTT) was selected as amino-acid substitution model, based on having the lowest AIC score among the models tested in ProtTest3 (ref. 64). RAxML[65] was used to calculate 100 maximum likelihood trees and the one with the highest likelihood score was selected for tree topology. Bootstrap values were generated based on the 100 resampled data sets.

KS domain clustering. Extracted amino-acid sequences of KS domains (see below) were aligned using MAFFT (v.7.273)[66] with the globalpair setting turned on and trimmed using Gblocks with parameter –b5 = h. RAxML was used as described in the previous section.

Protein phylogeny. Multiple sequence alignments were generated using MUSCLE[61] and subsequently trimmed such that each position contained less than 60% gaps. Based on ProtTest3 (ref. 64), gamma distributed JTT or LG were chosen as amino-acid substitution models to infer maximum likelihood phylograms using phyml[67] with 100 boostrap replicates. The Robinson–Foulds distance between phylogenetic trees was calculated using the implementation in ETE 3 (ref. 68). Protein sequences not generated in this study were obtained from the MIBiG database[20] for patulin (BGC0000120) and yanuthone D (BGC0000170) or from NCBI for the patulin BGC in *A. clavatus* (A1CFL4.1, A1CFL5.1, A1CFL6.1, A1CFL7.1, A1CFL8.1), the terreic acid BGC in *A. terreus* (Q0CJ59.1) and the aculinic acid BGC in *A. aculeatus* (OJJ97578.1), with the latter identified through a protein BLAST.

Secondary metabolite gene cluster detection and clustering. Annotated genomes in genbank format were submitted to antiSMASH (v.3.0.4)[16] for the detection of BGCs. The output was parsed with custom Python scripts to extract backbone genes, that is, PKSs, NRPSs, as well as additional metadata for each BGC. KS and C domains of all extracted backbone genes were detected using the online tool NaPDoS[69]. The detected KS and C domains were compared by an all versus all Smith–Waterman alignment[70] as implemented in biopython[71]. The resulting distribution of pairwise percent identities allowed us to infer a cutoff of 74% identity for KS domains and 64% identity for C domains, to distinguish whether BGCs should be grouped into gene cluster families. Synteny plots were generated using MultiGeneBlast[72].

To annotate the predicted BGCs, the MIBiG database (v.1.2)[20] was downloaded and fungal PKS, NRPS and hybrid PKS–NRPS protein sequences were extracted using custom Python scripts. The corresponding KS and C domains were identified and mapped to the predicted *Penicillium* BGCs using the same approach as described above.

Experimental secondary metabolite detection. Seven *Penicillium* species (*P. flavigenum*, *P. rubens*, *P. italicum*, *P. griseofulvum*, *P. digitatum*, *P. vulpinum* and *P. expansum*) were grown on YES (yeast extract sucrose agar: 20 g l^{-1} yeast extract, 150 g l^{-1} sucrose, 0.5 g l^{-1} $MgSO_4{\cdot}7H_2O$, 1 ml l^{-1} trace metal solution, 15 g l^{-1} agar), CYA (Czapek yeast autolysate agar: 5 g l^{-1} yeast extract, 35 g l^{-1} Czapek dox broth, 1 ml l^{-1} trace metal solution, 15 g l^{-1} agar) and PDA (potato dextrose agar: 5 g l^{-1} potato dextrose agar and 1 ml l^{-1} trace metal solution) medium for the detection of yanuthones and fungisporin. Three agar plugs were sampled from one colony on each medium, and 1.0 ml of extraction solvent, isopropanol:ethylacetate (1/3) containing 1% formic acid, was added. After ultrasonification for 1 h, the extract was transferred to a clean vial, evaporated to dryness and redissolved in 100 µl methanol. After centrifugation for 5 min, the supernatant was directly used for chemical analysis. Identification of secondary metabolites was done with UHPLC-DAD-TOFMS using an Agilent 1290 ultra-high-performance liquid chromatography (UHPLC) system with UV/Vis diode array detection (DAD) and coupled to an Agilent 6545 quadrupole time-of-flight mass spectrometer (TOFMS) (Agilent Technologies, Torrance, CA) equipped with an electrospray ionization (ESI) source. Separation was done on a Poroshell column (120 Å, 2.6 µm particle, 2 mm ID × 250 mm, maintained at 60 °C with a flow rate of 0.35 ml min^{-1}). A linear gradient system composed of 20 mmol l^{-1} formic acid in water and 20 mmol l^{-1} formic acid in acetonitrile was used, starting from 10% (vol/vol) acetonitrile and increasing to 100% in 15 min, maintaining this rate for 3 min before returning to the starting conditions within 0.1 min and maintaining these for 2.4 min before the following run. Samples were analysed in both positive and negative electrospray scanning, *m/z* 50–1,600. Tandem MS/HRMS spectra were obtained at fixed collision-induced dissociation (CID) energies of 10, 20 and 40 eV and matched to the internal library of ~1,500 reference standards and previously tentatively identified compounds[73].

To identify potential alternative yanuthones, all ions generating fragment ions identical to fragment ions observed from the yanuthone reference standards were investigated (pseudo-parent ion scanning). Reference standards of yanuthones D, E, F, H, I, J, X2, L and M, as well as 22-deacetylyanuthone A and 7-deacetoxyyanuthone A, were available. Pseudo-parent ion scanning was performed for a *m/z* 191.1794 fragment (prenyl unit) and the yanuthone E losing the carboxylate moiety at *m/z* 343.2267 and with an additional water loss at *m/z* 325.216207, as well as *m/z* 323.2006 and 341.2111 for yanuthone D. We also searched for compounds with a C_{10} prenyl chain using the fragment ion *m/z* 123.1168.

Fungisporins were characterized as yanuthones with fragment ions already identified and validated by isotope labelling[74] and all tested species were found to produce fungisporins, although with different elemental compositions ($C_{29}H_{38}N_4O_4$, $C_{28}H_{38}N_4O_5$, $C_{29}H_{38}N_4O_5$, $C_{28}H_{38}N_4O_6$, $C_{30}H_{39}N_5O_5$, $C_{30}H_{37}N_5O_6$ and $C_{30}H_{39}N_5O_6$).

For the identification of puberulonic acid (*P. polonicum* and *P. freii*), pyripyropenes (*P. coprophilum*) and isoepoxydon (*P. flavigenum* and *P. coprophilum*), extracts were prepared from cultures grown on solid YES medium, as described above, and compounds were identified using LC-DAD-TOFMS, as previously described[75].

Code availability. Python scripts referred to in the Methods are available at https://github.com/JensChrNielsen/Penicillium.

Data availability. All BGCs linked to entries in the MIBiG databases are provided in the Supplementary Data. *Pencillium* whole-genome shotgun projects for this study have been deposited at DDBJ/ENA/GenBank under accession numbers MDDG00000000 (*P. coprophilum*), MOOB00000000 (*P. nalgiovense*), MDYM00000000 (*P. polonicum*), MDYN00000000 (*P. antarcticum*), MDYP00000000 (*P. vulpinum*), MDYO00000000 (*P. solitum*), MDYL00000000 (*P. decumbens*), MLQL00000000 (*P. flavigenum*) and MLKD00000000 (*P. steckii*). Other data that support the findings of this study are available from the corresponding author upon reasonable request.

Received 3 December 2016; accepted 2 March 2017; published 3 April 2017

References

1. Aminov, R. I. A brief history of the antibiotic era: lessons learned and challenges for the future. *Front. Microbiol.* **1,** 134 (2010).
2. Keller, N. P., Turner, G. & Bennett, J. W. Fungal secondary metabolism—from biochemistry to genomics. *Nat. Rev. Microbiol.* **3,** 937–947 (2005).
3. Nielsen, J. C. & Nielsen, J. Development of fungal cell factories for the production of secondary metabolites: linking genomics and metabolism. *Synth. Syst. Biotechnol.* http://dx.doi.org/10.1016/j.synbio.2017.02.002 (2017).
4. Ziemert, N., Alanjary, M. & Weber, T. The evolution of genome mining in microbes—a review. *Nat. Prod. Rep.* **33,** 988–1005 (2016).
5. Medema, M. H. & Fischbach, M. A. Computational approaches to natural product discovery. *Nat. Chem. Biol.* **11,** 639–648 (2015).
6. Visagie, C. M. *et al.* Identification and nomenclature of the genus *Penicillium*. *Stud. Mycol.* **78,** 343–371 (2014).
7. Barrios-González, J. & Miranda, R. U. Biotechnological production and applications of statins. *Appl. Microbiol. Biotechnol.* **85,** 869–883 (2010).
8. Fang, X., Shen, Y., Zhao, J., Bao, X. & Qu, Y. Status and prospect of lignocellulosic bioethanol production in China. *Bioresour. Technol.* **101,** 4814–4819 (2010).
9. García-Estrada, C. & Martín, J.-F. Biosynthetic gene clusters for relevant secondary metabolites produced by *Penicillium roqueforti* in blue cheeses. *Appl. Microbiol. Biotechnol.* **100,** 8303–8313 (2016).
10. Chai, B., Wu, Y., Liu, P., Liu, B. & Gao, M. Isolation and phosphate-solubilizing ability of a fungus, *Penicillium* sp. from soil of an alum mine. *J. Basic Microbiol.* **51,** 5–14 (2011).
11. Richardson, A. E. & Simpson, R. J. Soil microorganisms mediating phosphorus availability update on microbial phosphorus. *Plant Physiol.* **156,** 989–996 (2011).
12. Puel, O., Galtier, P. & Oswald, I. P. Biosynthesis and toxicological effects of patulin. *Toxins* **2,** 613–631 (2010).
13. Grijseels, S. *et al. Penicillium arizonense*, a new, genome sequenced fungal species, reveals a high chemical diversity in secreted metabolites. *Sci. Rep.* **6,** 35112 (2016).
14. Park, M. S., Lee, E. J., Fong, J. J., Sohn, J. H. & Lim, Y. W. A new record of *Penicillium antarcticum* from marine environments in Korea. *Mycobiology* **42,** 109–113 (2014).
15. Houbraken, J., Wang, L., Lee, H. B. & Frisvad, J. C. New sections in *Penicillium* containing novel species producing patulin, pyripyropens or other bioactive compounds. *Persoonia* **36,** 299–314 (2015).
16. Weber, T. *et al.* antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res.* **43,** W237–W243 (2015).
17. Kroken, S., Glass, N. L., Taylor, J. W., Yoder, O. C. & Turgeon, B. G. Phylogenomic analysis of type I polyketide synthase genes in pathogenic and saprobic ascomycetes. *Proc. Natl Acad. Sci. USA* **100,** 15670–15675 (2003).
18. Rausch, C., Hoof, I., Weber, T., Wohlleben, W. & Huson, D. H. Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. *BMC Evol. Biol.* **7,** 78 (2007).
19. Ziemert, N. *et al.* Diversity and evolution of secondary metabolism in the marine Actinomycete genus *Salinispora*. *Proc. Natl Acad. Sci. USA* **111,** E1130–E1139 (2014).
20. Medema, M. H. *et al.* Minimum information about a biosynthetic gene cluster. *Nat. Chem. Biol.* **11,** 625–631 (2015).
21. Klejnstrup, M. L. *et al.* Genetics of polyketide metabolism in *Aspergillus nidulans*. *Metabolites* **2,** 100–133 (2012).
22. Artigot, M. P. *et al.* Molecular cloning and functional characterization of two CYP619 cytochrome P450s involved in biosynthesis of patulin in *Aspergillus clavatus*. *Microbiology* **155,** 1738–1747 (2009).
23. Holm, D. K. *et al.* Molecular and chemical characterization of the biosynthesis of the 6-MSA-derived meroterpenoid yanuthone D in *Aspergillus niger*. *Chem. Biol.* **21,** 519–529 (2014).
24. Frisvad, J. C., Smedsgaard, J., Larsen, T. O. & Samson, R. A. Mycotoxins, drugs and other extrolites produced by species in *Penicillium* subgenus *penicillium*. *Stud. Mycol.* **49,** 201–241 (2004).
25. Vansteelandt, M. *et al.* Patulin and secondary metabolite production by marine-derived *Penicillium* strains. *Fungal Biol.* **116,** 954–961 (2012).
26. Boysen, M., Skouboe, P., Frisvad, J. & Rossen, L. Reclassification of the *Penicillium roqueforti* group into three species on the basis of molecular genetic and biochemical profiles. *Microbiology* **142,** 541–549 (1996).
27. Ballester, A. *et al.* Genome, transcriptome, and functional analyses of *Penicillium expansum* provide new insights into secondary metabolism and pathogenicity. *Mol. Plant–Microbe Interact.* **28,** 232–248 (2015).
28. Medema, M. H., Cimermancic, P., Sali, A., Takano, E. & Fischbach, M. A. A systematic computational analysis of biosynthetic gene cluster evolution: lessons for engineering biosynthesis. *PLoS Comput. Biol.* **10,** e1004016 (2014).
29. Banani, H. *et al.* Genome sequencing and secondary metabolism of the postharvest pathogen *Penicillium griseofulvum*. *BMC Genomics* **17,** 19 (2016).
30. Itoh, T. *et al.* Reconstitution of a fungal meroterpenoid biosynthesis reveals the involvement of a novel family of terpene cyclases. *Nat. Chem.* **2,** 858–864 (2010).
31. Petersen, L. M., Holm, D. K., Gotfredsen, C. H., Mortensen, U. H. & Larsen, T. O. Investigation of a 6-MSA synthase gene cluster in *Aspergillus aculeatus* reveals 6-MSA-derived aculinic acid, aculins A-B and Epi-Aculin A. *ChemBioChem* **16,** 2200–2204 (2015).
32. Guo, C.-J., Sun, W.-W., Bruno, K. S. & Wang, C. C. C. Molecular genetic characterization of terreic acid pathway in *Aspergillus terreus*. *Org. Lett.* **16,** 5250–5253 (2014).
33. Bacha, N. *et al.* Cloning and characterization of novel methylsalicylic acid synthase gene involved in the biosynthesis of isoasperlactone and asperlactone in *Aspergillus westerdijkiae*. *Fungal Genet. Biol.* **46,** 742–749 (2009).
34. Brakhage, A. A. Regulation of fungal secondary metabolism. *Nat. Rev. Microbiol.* **11,** 21–32 (2013).
35. Wisecaver, J. H. & Rokas, A. Fungal metabolic gene clusters—caravans traveling across genomes and environments. *Front. Microbiol.* **6,** 161 (2015).
36. Chae, L., Kim, T., Nilo-Poyanco, R. & Rhee, S. Y. Genomic signatures of specialized metabolism in plants. *Science* **344,** 510–513 (2014).
37. Li, Y. F. *et al.* Comprehensive curation and analysis of fungal biosynthetic gene clusters of published natural products. *Fungal Genet. Biol.* **89,** 18–28 (2016).
38. Gao, X. *et al.* Fungal indole alkaloid biosynthesis: genetic and biochemical investigation of the tryptoquialanine pathway in *Penicillium aethiopicum*. *J. Am. Chem. Soc.* **133,** 2729–2741 (2011).
39. Chooi, Y.-H., Cacho, R. & Tang, Y. Identification of the viridicatumtoxin and griseofulvin gene clusters from *Penicillium aethiopicum*. *Chem. Biol.* **17,** 483–494 (2010).
40. Petersen, L. M. *et al.* Characterization of four new antifungal yanuthones from *Aspergillus niger*. *J. Antibiot.* **68,** 201–205 (2015).
41. Li, X., Choi, H. D., Kang, J. S., Lee, C.-O. & Son, B. W. New polyoxygenated farnesylcyclohexenones, deacetoxyyanuthone A and its hydro derivative from the marine-derived fungus *Penicillium* sp. *J. Natural Prod.* **66,** 1499–1500 (2003).
42. Maskey, R. P., Grün-Wollny, I. & Laatsch, H. Sorbicillin analogues and related dimeric compounds from *Penicillium notatum*. *J. Natural Prod.* **68,** 865–870 (2005).
43. Simpson, J. *et al.* ABySS: a parallel assembler for short read sequence data. *Genome Res.* **19,** 1117–1123 (2009).
44. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience* **1,** 18 (2012).
45. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19,** 455–477 (2012).
46. Chevreux, B., Thomas, W. & Suhai, S. Genome sequence assembly using trace signals and additional sequence information. *Comp. Sci. Biol.* **99,** 45–56 (1999).
47. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29,** 1072–1075 (2013).
48. Vezzi, F., Narzisi, G. & Mishra, B. Reevaluating assembly evaluations with feature response curves: gAGE and assemblathons. *PLoS ONE* **7,** e52210 (2012).
49. Smit, A., Hubley, R. & Green, P. *RepeatMasker Open-4.0* (2015); http://www.repeatmasker.org.
50. Smit, A. & Hubley, R. *RepeatModeler Open-1.0* (2015); http://www.repeatmasker.org.
51. The UniProt consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* **43,** D204–D212 (2014).
52. Lomsadze, A., Burns, P. D. & Borodovsky, M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res.* **42,** e119 (2014).
53. Kim, D. *et al.* Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14,** R36 (2013).
54. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28,** 511–515 (2010).
55. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12,** 491 (2011).
56. Zdobnov, E. M. & Apweiler, R. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17,** 847–848 (2001).

57. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31,** 3210–3212 (2015).
58. Li, L., Stoeckert, C. J. J. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13,** 2178–2189 (2003).
59. Agren, R. *et al.* The RAVEN toolbox and its use for generating a genome-scale metabolic model for *Penicillium chrysogenum*. *PLoS Comput. Biol.* **9,** e1002980 (2013).
60. Yin, Y. *et al.* dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* **40,** W445–W451 (2012).
61. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).
62. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17,** 540–552 (2000).
63. Felsenstein, J. PHYLIP—phylogeny inference package (version 3.2). *Cladistics* **5,** 164–166 (1989).
64. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. Prottest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27,** 1164–1165 (2011).
65. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22,** 2688–2690 (2006).
66. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30,** 772–780 (2013).
67. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59,** 307–321 (2010).
68. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **33,** 1635–1638 (2016).
69. Ziemert, N. *et al.* The natural product domain seeker NaPDoS: a phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. *PLoS ONE* **7,** e34064 (2012).
70. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **147,** 195–197 (1981).
71. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25,** 1422–1423 (2009).
72. Medema, M. H., Takano, E. & Breitling, R. Detecting sequence homology at the gene cluster level with multigeneblast. *Mol. Biol. Evol.* **30,** 1218–1223 (2013).
73. Kildgaard, S. *et al.* Accurate dereplication of bioactive secondary metabolites from marine-derived fungi by UHPLC-DAD-QTOFMS and a MS/HRMS library. *Mar. Drugs* **12,** 3681–3705 (2014).
74. Klitgaard, A., Nielsen, J. B., Frandsen, R. J. N., Andersen, M. R. & Nielsen, K. F. Combining stable isotope labeling and molecular networking for biosynthetic pathway characterization. *Anal. Chem.* **87,** 6520–6526 (2015).
75. Nielsen, K. F., Månsson, M., Rank, C., Frisvad, J. C. & Larsen, T. O. Dereplication of microbial natural products by LC-DAD-TOFMS. *J. Natural Prod.* **74,** 2338–2348 (2011).
76. Tannous, J. *et al.* Sequencing, physical organization and kinetic expression of the patulin biosynthetic gene cluster from *Penicillium expansum*. *Int. J. Food Microbiol.* **189,** 51–60 (2014).

Acknowledgements

This work was supported by the European Commission Marie Curie Initial Training Network Quantfung (FP7-People-2013-ITN, grant no. 607332), the Novo Nordisk Foundation and the Knut and Alice Wallenberg Foundation. The computations were performed using resources at the Chalmers Centre for Computational Science and Engineering (C3SE) provided by the Swedish National Infrastructure for Computing (SNIC). Sequencing support was provided by the Science for Life Laboratory (SciLifeLab), National Genomics Infrastructure (NGI) and UPPMAX (UPPNEX project ID no. b2014081). Support on genome annotation by the National Bioinformatics Infrastructure Sweden (NBIS) is acknowledged. Agilent Technologies is acknowledged for the Thought Leader Donation of the 6545 UHPLC-QTOF. The authors thank H. Wang for comments on the manuscript.

Author contributions

J.C.N., J.C.F., M.W. and J.N. conceived the study. J.C.N. designed and performed the bioinformatics computations, and analysed and interpreted the data. S.P. and B.J. assisted with bioinformatics design and interpretation. J.D. carried out the annotation of the genomes. S.G. and K.F.N. generated culture extracts and performed LC–MS analysis. J.C.N., S.G. and J.N. wrote the manuscript. All authors read and approved the final version of the manuscript.

Additional information

Supplementary information is available for this paper.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to J.N.

How to cite this article: Nielsen, J. C. *et al.* Global analysis of biosynthetic gene clusters reveals vast potential of secondary metabolite production in *Penicillium* species. *Nat. Microbiol.* **2,** 17044 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Competing interests

The authors declare no competing financial interests.