Comparative genomic study of the Penicillium genus elucidates a diverse pangenome and 15 lateral gene transfer events

Petersen, Celine; Sørensen, Trine; Nielsen, Mikkel Rank; Sondergaard, Teis Esben; Sørensen, Jens Laurids; Fitzpatrick, David; Frisvad, Jens Christian; Nielsen, Kåre Lehmann

Published in:
I M A Fungus

DOI (link to publication from Publisher):
10.1186/s43008-023-00108-7

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Petersen, C., Sørensen, T., Nielsen, M. R., Sondergaard, T. E., Sørensen, J. L., Fitzpatrick, D., Frisvad, J. C., & Nielsen, K. L. (2023). Comparative genomic study of the Penicillium genus elucidates a diverse pangenome and 15 lateral gene transfer events. *I M A Fungus*, *14*(1), Article 3. https://doi.org/10.1186/s43008-023-00108-7

RESEARCH

Open Access

Comparative genomic study of the *Penicillium* genus elucidates a diverse pangenome and 15 lateral gene transfer events

Celine Petersen[1], Trine Sørensen[1], Mikkel R. Nielsen[2], Teis E. Sondergaard[1], Jens L. Sørensen[2], David A. Fitzpatrick[3], Jens C. Frisvad[4] and Kåre L. Nielsen[1*]

Abstract

The Penicillia are known to produce a wide range natural products—some with devastating outcome for the agricultural industry and others with unexploited potential in different applications. However, a large-scale overview of the biosynthetic potential of different species has been lacking. In this study, we sequenced 93 *Penicillium* isolates and, together with eleven published genomes that hold similar assembly characteristics, we established a species phylogeny as well as defining a *Penicillium* pangenome. A total of 5612 genes were shared between ≥ 98 isolates corresponding to approximately half of the average number of genes a *Penicillium* genome holds. We further identified 15 lateral gene transfer events that have occurred in this collection of *Penicillium* isolates, which might have played an important role, such as niche adaption, in the evolution of these fungi. The comprehensive characterization of the genomic diversity in the *Penicillium* genus supersedes single-reference genomes, which do not necessarily capture the entire genetic variation.

Keywords Long read sequencing, MinION, Filamentous fungi, *Penicillium*, Phylogenetics, Pangenome, Lateral gene transfer, Mycotoxins

INTRODUCTION

The genus *Penicillium* is broad, ecologically diverse, and comprised of more than 483 known species. They are widespread and occupy various habitats including air, soil, indoor environments, vegetation, and food products (Houbraken et al. 2020; Visagie et al. 2014a). In nature, members of the genus are known decomposers of organic matter, which includes pre- and post-harvesting rotting of food crops (Frisvad and Samson 2004; Samson et al. 2019). On the other hand, the food industry utilizes the same decomposition, but also flavor and texture enhancing capability of some species in e.g., cheese production (Thom 1906; Nelson 1970; Karahadian et al. 1985; Giraud et al. 2010; Bodinaku et al. 2019) and fermented sausages (López-Díaz et al. 2001; Ludemann et al. 2010). A wide range of mycotoxins are produced by *Penicillium* (Frisvad et al. 2004; Perrone and Susca 2017*)*, for example patulin, whose presence in fruits is tightly monitored and controlled (Perrone and Susca 2017). Naturally, the most famous extrolite produced by *Penicillium* spp. is penicillin, which revolutionized the medical treatment of bacterial diseases in the twentieth century (Fleming 1929). Generally, the *Penicillium* genus has received much

*Correspondence:
Kåre L. Nielsen
kln@bio.aau.dk
[1] Department of Chemistry and Bioscience, Aalborg University, Fredrik Bajers Vej 7H, 9220 Ålborg, Denmark
[2] Department of Chemistry and Bioscience, Aalborg University, Niels-Bohrs Vej 8, 6700 Esbjerg, Denmark
[3] Department of Biology, Maynooth University, Maynooth W23 F2K8, Ireland
[4] Department of Biotechnology and Biomedicine, Technical University of Denmark, Søltofts Plads B221, 2800 Kgs, Lyngby, Denmark

attention and numerous studies screening for degradative abilities, exoenzyme production, and bioremediation but also for production of commercial antibiotics, cholesterol lowering agents, antioxidants, immunosuppressants etc. (Perrone and Susca 2017; Barrett et al. 2020; Schneider et al. 2016; Terrasan et al. 2010; Adsul et al. 2007; Li et al. 2007). A major challenge of such studies is that interesting compounds are often only produced under very specific conditions, which are very difficult to reproduce in laboratory settings. In contrast, whole genome sequencing opens the door to a more comprehensive view of the biosynthetic potential of the Penicillia. Advancements in long read sequencing technologies have provided opportunities to create affordable high-quality genome drafts (Petersen et al. 2022; Sørensen et al. 2014). Such high-quality genomes create higher comparative power when assessing the divergence of genes, pathways, and evolutionary development, which in turn allows more precise identification of species, orthologous genes, and gene clusters, and thus a better foundation for genome mining for novel compounds and enzymes, as well as a deeper understanding of evolutionary relationship. The use of a limited number of marker genes (e.g., ITS, *BenA*, *CaM,* and *RPB2*) sometimes leads to ambiguous results for closely related species due to little observed variation, and this can be overcome by the analysis of more comprehensive sets of genes, comprising hundreds or even thousands of conserved genes (Rokas et al. 2003). Incongruence may also arise in such datasets, but by increasing the number of genes and taxa, it can be minimized (Steenwyk et al. 2019). Furthermore, the structure of gene clusters, which is important for identifying orthologous gene clusters, synthesizing the same or similar compounds across species, can be analyzed in contiguous high-quality genome drafts (Blin et al. 2019).

In this study, we have sequenced and de novo assembled the genome drafts of 93 *Penicillium* isolates, of which 68 were considered individual species. The purpose is to reevaluate the established phylogeny with a more comprehensive gene set, as well as establish a *Penicillium* pangenome to gather insight into the entire gene pool and establish groups of orthologs. Furthermore, we analyzed the genome sequences for signs of lateral gene transfer (LGT) from bacteria to identify transfer events that might have an important role in the evolution of *Penicillium* genus.

MATERIALS AND METHODS

Penicillium isolates

91 *Penicillium* isolates were collected from the IBT Culture Collection of Fungi at the Technical University of Denmark (DTU, Denmark). *P. riverlandense* (CBS 135883) and *P. lagena* (CBS 129212) were purchased from CBS strain collection from Westerdijk Fungal Biodiversity Institute (Utrecht, The Netherlands). Further information regarding species and accession number of these isolates can be found in Additional file 1: Table S1. The isolates were grown and treated, sequenced, and assembled as described in Petersen et al. (2022).

Assemblies from an additional ten *Penicillium* isolates produced by others from long read sequencing data were downloaded from NCBI. These are: *P. camemberti* (FM013 = LCP06093 (NCBI ref. nr. PcamFM013r2_polished)), *P. capsulatum* (ATCC48735 (NCBI ref. nr. ASM94376v1)), *P. capsulatum* (LiaoWQ-2011 (NCBI ref. nr. ASM94377v1)), *P. digitatum* (DSM62840 (NCBI ref. nr. ASM1229554v2)), *P. digitatum* (PdW03 (NCBI ref. nr. ASM1676781v1)), *P. expansum* (R19 (NCBI ref. nr. Pexp_R19)), *P. oxalicum* (SGAir0226 (NCBI ref. nr. SGAir0226)), *P. polonicum* (F7 (NCBI ref. nr. ASM1346617v1)), *P. solitum* (RS1 (NCBI ref. nr. ASM95277v2)), and *P. solitum* (#12 (NCBI ref. nr. ASM1313803v1)). Furthermore, we included *P. aurantiogriseum* (IBT 35659 (NCBI ref. nr. ASM1997785v1)) that has recently been sequenced by our group (Petersen et al. 2022).

High molecular weight DNA extraction and sequencing

High molecular weight DNA was extracted from the 93 *Penicillium* isolates using either Genomic Buffer Set (Qiagen, Germany) in combination with QIAGEN Genomic-Tips 20/G or using phenol–chloroform extraction in combination with QIAGEN Genomic-Tips 20/G as described in Petersen et al. (2022). Specific extraction methods are listed in Additional file 1: Table S1. Quality control of DNA was performed, followed by a removal of small DNA fragments to increase efficiency of sequencing, and finally another quality control was performed as described in Petersen et al. (2022). DNA Library preparations of two to four fungi were performed following the Native barcoding genomic DNA (EXP-NBD104, EXPNBD114, and SQK-LSK109) protocol from Oxford Nanopore Technologies (Oxford, UK) and the isolates were sequenced either on a R9.4.1 or R10.3 flow cell (Additional file 1: Table S1).

Genome assembly and annotation

The raw reads were processed to generate 93 genome draft assemblies as described in Petersen et al. (Petersen et al. 2022). In short, Guppy (Technologies and pyguppyclient. Available from: 702https:, , github.com, nanoporetech, pyguppyclient(accessed 2022 September 23). 2022) was used to basecall, remove adaptors, and demultiplex the reads. Filtlong (Wick 2018) was used for trimming the reads for low confidence basecalls as well as short reads. Evaluation of the reads was performed with

NanoPlot (Coster et al. 2018). Minimap2 (Li 2018) with default setting was used to map overlaps of the reads before the reads were assembled by Miniasm (Li 2016) using default setting. Subsequently, the assembly was polished using Racon (Vaser et al. 2017) and Medaka (Oxford Nanopore Technologies 2018) both with default setting. The different software versions used are listed in Additional file 1: Table S2 and trimming criteria can be found in Additional file 1: Table S1. Evaluation of assembly completeness was performed with Benchmarking Universal Single-Copy Orthologs (BUSCO) v5.0.0 using Eurotiales BUSCO dataset v10 (Manni et al. 2021).

FunGAP v1.1.0 (Min et al. 2017) was used to annotate the 93 genome draft assemblies and the 11 assemblies from NCBI. FunGAP was guided by an RNA-seq dataset that was generated by compiling RNA-seq data from several recent experiments of various *Penicillium* species: *P. bilaiae* (SRX2770370), *P. capsulatum* (SRX2022597), *P. digitatum* (SRX5587704), *P. glabrum* (SRX2770363), *P. novae-zelandiae* (SRX3064154), *P. solitum* (SRX5209310, SRX5209311), *P. steckii* (SRX3398614, SRX3398615, SRX3398616, SRX3398617, SRX3398618, SRX3398619), *P. swiecickii* (SRX3959645), *P. citreonigrum* (SRX7101841), and *P. subrubescens* (SRX7664426). One gigabase (gb) of sequences was randomly subsampled from each set and pooled. Protein sequences from *P. digitatum PHI26* (SAMN02399970), *P. oxalicum 114-2* (SAMN02981374), *P. freii* DAOM 242723 (SAMN03784687), and *P. flavigenum* (SAMN05200882) were downloaded from NCBI with the download_sister_orgs.py script from FunGAP to obtain the proteome database.

Secondary metabolite genes were predicted from the 104 proteomes using AntiSMASH v5.1 (Blin et al. 2019) and visualized using the R package ggplot (Wickham 2009). BiG-SCAPE (Navarro-Muñoz et al. 2020) was used to investigate gene cluster duplication of the biosynthetic gene cluster predicted by AntiSMASH in *P. soppii*. The Pfam database (Mistry et al. 2021) and InterProScan v5.38-76.0 (Jones et al. 2014) was used to functionally annotate the genes in each genome draft.

Alignment and phylogeny.

BUSCO v5 analysis revealed 1,142 BUSCO families present in all 104 *Penicillium* isolates and two relevant outgroups (*Aspergillus fumigatus* (AF293) and *Aspergillus flavus* (NRRL3357)). Each BUSCO family was individually aligned with MUSCLE v3.8.1551 (Edgar 2004) and trimmed using trimAl v1.2 (Capella-Gutiérrez et al. 2009) with the parameter "-automated1" to remove poorly aligned regions. Trimmed alignments were concatenated together resulting in a supermatrix alignment of 885,360 amino acids. To speed up computation, phylogenetically uninformative sites were removed from the alignment, generating a final alignment of 364,162 amino acids. Maximum-likelihood phylogenetic reconstruction was performed using IQ-TREE v2.1.2 (Minh et al. 2020) with the JTT + F + R5 model, which was the best-fit model according to ModelFinder (Kalyaanamoorthy et al. 2017), and 100 bootstrap replicates were undertaken to infer branch support values.

Due to computational memory limitations, it was not possible to conduct whole genome alignment of all 104 genome sequences in a single analysis. Considering that additional information gained by whole genome alignment is most likely to be relevant for closely related species, four subtrees were pruned from the aforementioned phylogenetic tree (Additional file 2: Figure S1-S4). Whole genome alignment of the isolates in each subtree was then performed using CLC Genomics Workbench version 20.0 (Qiagen, Århus) using default setting. The resulting average nucleotide identity matrices—one for each subtree—was then used to generate four neighbour joining trees. The R package ggtree (Yu et al. 2017) was used to illustrate the figures.

Mating type identification

The mating loci were identified from annotated genomes in GBK file format using the BLASTx algorithm with default settings in CLC Main Workbench version 7 (Qiagen, Århus). Coding sequences of the following genes were used as query: *Fusarium graminearum* PH-1 *MAT1-1-1* (FGSG_08892) and *MAT1-2-1* (FGSG_08893), *P. Roquefort* FM 317 *MAT1-1-1* (JX627318), *P. Roquefort* FM 164 *MAT1-2-1*, *SLA2* and *APN2* (KC469511), *P. chrysogenum* ATCC 28089 *MAT1-1-1* (AM904544), *P. chrysogenum NRRL* 1249B21 *MAT1-2-1* (AM904545). After identification of putative homologs, *P. alfredii* 34,128 *MAT1-1–1* (gene_00424), *APN2* (gene_00425), *SLA2* (gene_00423), and *P. antarticum* 31339 *MAT1-2-1* (gene_05584) were added to the list of query sequences and the BLASTx was repeated for every *Penicillium* genome. Synteny alignment plots were generated with the tBLASTx feature of EasyFig v2.1 (Sullivan et al. 2011) with BLAST options set to Min. length = 50 bp, Max. e-value = 0.001, and Min. Identity value = 50% for *P. malachiteum* or 25% for *P. macrosclerotiorum*.

Pangenome construction

GET_HOMOLOGUES v3.3.2 with OMCL (Contreras-Moreira and Vinuesa 2013) was used to construct the *Penicillium* pangenome by clustering the 104 predicted proteomes into orthogroups with default setting. The pangenome was divided into core (genes present in all isolates), softcore (genes present in ≥ 95% of all isolates), shell (genes present in < 95% but > 2% of all isolates), and cloud (genes present in $\leq 2\%$ of all isolates) genes. The pan and core gene size were estimated by random sampling of the genomes during the construction of the pangenome and fitted using Willenbrock model and Tettelin model for core and pan-genome, respectively. The estimate of the pan gene size was performed again with a subset of data (proteins with a length ≤ 150 aa, proteins with a length between 151 and 750 aa, proteins with a length between 751 and 1500 aa, and proteins with a length > 1500 aa) and all data to evaluate the difference in the saturation curve. However, DIAMOND (Buchfink et al. 2021) was used instead of BLASTp (Altschul et al. 1997) during similarity searches this time. A representative protein sequence of each orthogroup was chosen by taking the longest protein sequence within the average plus two standard deviations. Enrichment analysis of gene ontology (GO) was performed on these representatives, using GOATOOLS (Klopfenstein et al. 2018) and a false discovery rate of 0.05 was applied.

Estimation of the ratio of substitution rates at non-synonymous and synonymous sites (d_N/d_S) was calculated for each orthogroup in the softcore genome and each orthogroup with at least 10 genes in the shell genome. MUSCLE v5.0.1428 (Edgar 2021) was used to align the protein-coding sequences of each orthogroup. The alignments were converted to a nucleotide alignment that determined whether a substitution was caused by a synonymous or non-synonymous change using PAL2NAL (Suyama et al. 2006). CODEML from PAML (Yang 2007) was used to calculate pairwise d_N/d_S values in each orthogroup with the site model M0. Lastly, median value for each orthogroup (9530 orthogroups for the shell genome and 5605 orthogroups for the softcore genome) was selected for further comparison. The values were log-normal transformed to find d_N/d_S values that were significantly lower or higher than mean with a false discovery rate cutoff of 0.05. Mann Whitney U statistical testing was performed between core, softcore, shell, and cloud orthogroups for protein length, as well as softcore and shell orthogroups for d_N/d_S values in R.

The pangenome was classified into euKaryotic Orthologous Groups (KOGs) by taking the representative protein sequence of each orthogroup used in the GO enrichment analysis and searching these against the KOG database (retrieved July 1st, 2022 from (http://ftp.ncbi.nih.gov/pub/mmdb/cdd/little_endian) using rpsblast (E value 0.01). Only the best hit from each orthogroup was used. If the KOG annotation had multiple KOG categories, all of them were included.

The pangenome matrix generated from GET_HOMOLOGUES was used to create a maximum likelihood tree using GET_PHYLOMAKERS (Vinuesa et al. 2018). Five independent IQ-TREE runs with ultrafast bootstrapping were performed. The best model was GTR2 + FO + R6. This tree is based on whether genes are present or absent. The tree was visualized and annotated using ITOL v5 (Letunic and Bork 2021).

Figures were illustrated with the R package ggplot (Wickham 2009) unless otherwise indicated.

Identification of lateral gene transfer

The extent of LGT in each genome was assessed by combining all 104 proteomes together to give a query database of 1,244,528 proteins. Individual proteins were searched against a protein database representative of fully sequenced prokaryotic and eukaryotic species using BLASTp v2.10.1 + with an E value cut-off of $1e^{-10}$. This database consists of 1,698 genomes sampled from all three domains of life that had been used in a previous interdomain LGT analysis (McCarthy and Fitzpatrick 2016). *Penicillium* proteins that had a top hit to a bacterial species (3200 in total) were retained for a second round of database searches.

The 3200 *Penicillium* proteins were subsequently searched against the 20,671 Uniprot reference proteomes (Release 2022_01) using BLASTp with an E value cut-off set to $1e^{-10}$. Proteins with top hits to bacteria were again retained. A BLAST filter was applied where the query percentage identity to the top bacterial hits must be greater than 50% and the length of the query must be at least 80% of the subject (or vice versa). Furthermore, at least 80% of the top 200 hits for the query protein must come from bacteria. To maximise taxonomical coverage, a third round of database searches was then undertaken against NCBI's non-redundant protein database (NR, last access 4th February 2022) to confirm a top hit to a bacterial source.

To avoid redundancy, candidate *Penicillium* LGT sequences were then grouped into 56 orthogroups using OrthoFinder (Emms and Kelly 2019). Representative and singleton sequences from these orthogroups were queried using BLASTp with an E value cutoff of $1e^{-10}$ against the Uniprot reference proteomes to locate homologs for phylogenetic analysis. For computational reasons, an arbitrary limit of 500 maximum hits per query sequence was imposed. We reconstructed 56 maximum likelihood (ML) phylogenetic trees for the candidate LGT families. Briefly, candidate LGT gene families were aligned using MUSCLE v3.8.1551. ML trees were inferred using IQ-TREE v2.1.2 with the TEST model and 100 bootstrap replicates. The resultant phylogenetic trees were inspected manually and visualized and annotated using ITOL v5. SignalP v6.0 (Teufel et al. 2022) was used to investigate whether the products of the LGT event genes were secreted.

RESULTS

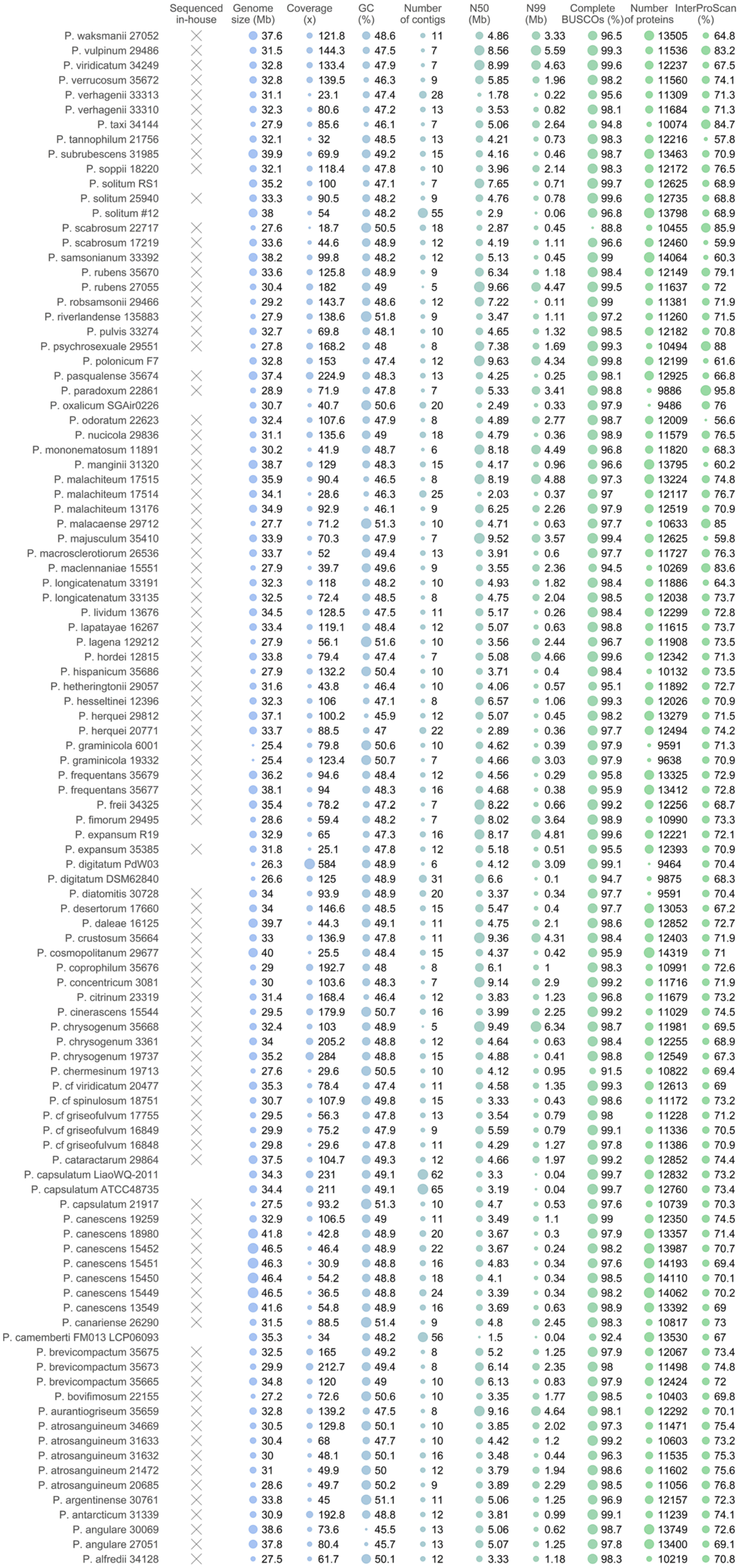
Long read sequencing yielded high-quality genome drafts

In this study, high molecular weight DNA from 93 *Penicillium* isolates was sequenced using the MinION long read sequencing platform, and de novo assembled into highly contiguous assemblies. The isolates were carefully chosen to ensure adequate phylogenetic/taxonomical coverage of the *Penicillium* genus. On average, 5.12 gb were sequenced per genome, of which, on average, 62.61% remained after filtering for minimal length and subsequently used in the assembly process (Additional file 1: Table S1). This corresponds to an average of 742,415 reads per genome. An average, 18.20% of these had the required minimal length to be included. The average assembled genome size was 33,27 Mb (25.4—46.5 Mb) and the number of contigs ranged from five to 28 (Fig. 1). To measure genome contiguity, we computed the N99. The N99 represents the minimum contig length needed to cover 99% of the genome. The average N99 across all genomes was 1.51 Mb. To evaluate genome completeness, BUSCO analysis was performed. An average of 97.87% complete BUSCOs were observed. Furthermore, 11 *Penicillium* genome assemblies made from long read data and thus with similar assembly characteristics were downloaded from NCBI, including *P. aurantiogriseum* that we recently sequenced (Petersen et al. 2022) (Fig. 1). All 104 *Penicillium* genome assemblies were annotated with FunGAP. The number of predicted proteins ranged from 9591 to 14,319 with an average of 11,976. The proportion of proteins assigned to protein superfamilies (containing a pfam domain) ranged from 56.6 to 95.8% per genome. Typically, secondary metabolite gene cluster contain one or a few core genes (e.g. polyketide synthases, non-ribosomal peptide synthases), as well a number of cluster specific genes (transporter, transcription factors etc.). The number of predicted secondary metabolite core genes range from 24 (in 24 biosynthesis gene clusters in *P. diatomitis*) to 112 (in 89 biosynthesis gene clusters in *P. soppii*) (Additional file 2: Figure S5). To ensure that we do not overestimate the number of unique biosynthetic gene clusters by including duplications, the gene clusters identified by antiSMASH in *P. soppii* were compared using BiG-SCAPE. We did not observe any clustering which would indicate duplications of gene clusters.

Phylogenetics based on 1142 BUSCO genes reveals misidentification or contamination of accessions but generally supports previous suggested phylogentic trees

To evaluate the phylogenetic relationship of the different *Penicillium* spp., we performed an alignment of 1,142 BUSCO gene shared by all 104 *Penicillium* isolates as well as two *Aspergillus* spp. outgroups. Annotations of subgenera and sections in accordance with Houbraken et al. (2020) were added (* MERGEFORMAT Fig. 2). In general, the BUSCO genes yield a more robust tree with all nodes supported with bootstrap values of 100, expect two nodes with values 97 and 86. The lowest bootstrap value of 86 is found in the node separating the three highly similar *P. brevicompactum.* In comparison, bootstrap values in Houbraken et al. study vary from below 70 to 100, with many values below 85. A clear division of the two subgenera *Penicillium* and *Aspergilloides* was observed in the BUSCO tree. In this analysis, the phylogenetic placement of *P. riverlandense* and *P. lagena* (sect. *Torulomyces*) together with *P. alfredii* (sect. *Alfrediorum*) were distinct from the two subgenera. The placement of *P. alfredii* was consistent with that obtained by analysis of presence/absence of genes in the entire gene pool (Additional file 2: Figure S6). This, however, is not the case for *P. riverlandense* and *P. lagena*, which are placed in the middle of the subgenus *Aspergilloides*. Regardless, their placement outside of the two subgenera could indicate that *P. alfredii*, *P. riverlandense, and P. lagena* perhaps are closely related to ancestors of the rest of the *Penicillium* spp..

The BUSCO phylogeny assigned *P. capsulatum* (ATCC48735) and *P. capsulatum* (LiaoWQ-2011), together with *P. scabrosum* (IBT 22717) and *P. atrosanguineum* (IBT 31633) in unexpected subgenera. However, we believe these incongruences are likely to represent misidentifications or simple tissue collection inventory errors. Houbraken et al. (2021) reidentified *P. capsulatum* (ATCC48735) as a *P. canescens* (ATCC48735). This agrees with our observations (* MERGEFORMAT Fig. 2, Additional file 2: Figure S3 and Additional file 2: Figure S6), and we further provide evidence that the same applies for *P. capsulatum* (LiaoWQ-2011). Thus, *P. capsulatum* (LiaoWQ-2011) should be named *P. canescens* (LiaoWQ-2011). Additionally, *P. solitum* (RS1) have been reidentified as *P. polonicum* (RS1) (Houbraken et al. 2021), and in agreement with this, our conserved BUSCO phylogeny shows close relationship to *P. polonicum*. *P. scabrosum* (IBT 22717) clustered close to *P. chermesinum* and is therefore most likely not a *P. scabrosum*, but instead one species of the sect. *Charlesia*. In addition, the other *P. scabrosum* analyzed (IBT 17219) is in sister-group relationship with *P. chrysogenum*. *P. atrosanguineum* (IBT 31633) does not cluster with the other *P. atrosanguineum* strains but close to *P. paradoxum* and should likely belong to sect. *Paradoxa*. In all cases, consistent phylogeny was observed based on presence/absence of genes (Additional file 2: Figure S6) as well as based



	Sequenced in-house	Genome size (Mb)	Coverage (x)	GC (%)	Number of contigs	N50 (Mb)	N99 (Mb)	Complete BUSCOs (%)	Number of proteins	InterProScan (%)
P. waksmanii 27052	X	37.6	121.8	48.6	11	4.86	3.33	96.5	13505	64.8
P. vulpinum 29486	X	31.5	144.3	47.5	7	8.56	5.59	99.3	11536	83.2
P. viridicatum 34249	X	32.8	133.4	47.9	7	8.99	4.63	99.6	12237	67.5
P. verrucosum 35672	X	32.8	139.5	46.3	9	5.85	1.96	98.2	11560	74.1
P. verhagenii 33313	X	31.1	23.1	47.4	28	1.78	0.22	95.6	11309	71.3
P. verhagenii 33310	X	32.3	80.6	47.2	13	3.53	0.82	98.1	11684	71.3
P. taxi 34144	X	27.9	85.6	46.1	7	5.06	2.64	94.8	10074	84.7
P. tannophilum 21756	X	32.1	32	48.5	13	4.21	0.73	98.3	12216	57.8
P. subrubescens 31985	X	39.9	69.9	49.2	15	4.16	0.46	98.7	13463	70.9
P. soppii 18220	X	32.1	118.4	47.8	10	3.96	2.14	98.3	12172	76.5
P. solitum RS1		35.2	100	47.1	7	7.65	0.71	99.7	12625	68.9
P. solitum 25940	X	33.3	90.5	48.2	9	4.76	0.78	99.6	12735	68.8
P. solitum #12		38	54	48.2	55	2.9	0.06	96.8	13798	68.9
P. scabrosum 22717	X	27.6	18.7	50.5	18	2.87	0.45	88.8	10455	85.9
P. scabrosum 17219	X	33.6	44.6	48.9	12	4.19	1.11	96.6	12460	59.9
P. samsonianum 33392	X	38.2	99.8	48.2	12	5.13	0.45	99	14064	60.3
P. rubens 35670	X	33.6	125.8	48.9	9	6.34	1.18	98.4	12149	79.1
P. rubens 27055	X	30.4	182	49	5	9.66	4.47	99.5	11637	72
P. robsamsonii 29466	X	29.2	143.7	48.6	12	7.22	0.11	99	11381	71.9
P. riverlandense 135883	X	27.9	138.6	51.8	9	3.47	1.11	97.2	11260	71.5
P. pulvis 33274	X	32.7	69.8	48.1	10	4.65	1.32	98.5	12182	70.8
P. psychrosexuale 29551	X	27.8	168.2	48	8	7.38	1.69	99.3	10494	88
P. polonicum F7		32.8	153	47.4	12	9.63	4.34	99.8	12199	61.6
P. pasqualense 35674	X	37.4	224.9	48.3	13	4.25	0.25	98.1	12925	66.8
P. paradoxum 22861	X	28.9	71.9	47.8	7	5.33	3.41	98.8	9886	95.8
P. oxalicum SGAir0226		30.7	40.7	50.6	20	2.49	0.33	97.9	9486	76
P. odoratum 22623	X	32.4	107.6	47.9	8	4.89	2.77	98.7	12009	56.6
P. nucicola 29836	X	31.1	135.6	49	18	4.79	0.36	98.9	11579	76.5
P. mononematosum 11891	X	30.2	41.9	48.7	6	8.18	4.49	96.8	11820	68.3
P. manginii 31320	X	38.7	129	48.3	15	4.17	0.96	96.6	13795	60.2
P. malachiteum 17515	X	35.9	90.4	46.5	8	8.19	4.88	97.3	13224	74.8
P. malachiteum 17514	X	34.1	28.6	46.3	25	2.03	0.37	97	12117	76.7
P. malachiteum 13176	X	34.9	92.9	46.1	9	6.25	2.26	97.9	12519	70.9
P. malacaense 29712	X	27.7	71.2	51.3	10	4.71	0.63	97.7	10633	85
P. majusculum 35410	X	33.9	70.3	47.9	7	9.52	3.57	99.4	12625	59.8
P. macrosclerotiorum 26536	X	33.7	52	49.4	13	3.91	0.6	97.7	11727	76.3
P. maclennaniae 15551	X	27.9	39.7	49.6	9	3.55	2.36	94.5	10269	83.6
P. longicatenatum 33191	X	32.3	118	48.2	10	4.93	1.82	98.4	11886	64.3
P. longicatenatum 33135	X	32.5	72.4	48.5	8	4.75	2.04	98.5	12038	73.7
P. lividum 13676	X	34.5	128.5	47.5	11	5.17	0.26	98.4	12299	72.8
P. lapatayae 16267	X	33.4	119.1	48.4	12	5.07	0.63	98.8	11615	73.7
P. lagena 129212	X	27.9	56.1	51.6	10	3.56	2.44	96.7	11908	73.5
P. hordei 12815	X	33.8	79.4	47.4	7	5.08	4.66	99.6	12342	71.3
P. hispanicum 35686	X	27.9	132.2	50.4	10	3.71	0.4	98.4	10132	73.5
P. hetheringtonii 29057	X	31.6	43.8	46.4	10	4.06	0.57	95.1	11892	72.7
P. hesseltinei 12396	X	32.3	106	47.1	8	6.57	1.06	99.3	12026	70.9
P. herquei 29812	X	37.1	100.2	45.9	12	5.07	0.45	98.2	13279	71.5
P. herquei 20771	X	33.7	88.5	47	22	2.89	0.36	97.7	12494	74.2
P. graminicola 6001	X	25.4	79.8	50.6	10	4.62	0.39	97.9	9591	71.3
P. graminicola 19332	X	25.4	123.4	50.7	7	4.66	3.03	97.9	9638	70.9
P. frequentans 35679	X	36.2	94.6	48.4	12	4.56	0.29	95.8	13325	72.9
P. frequentans 35677	X	38.1	94	48.3	16	4.68	0.38	95.9	13412	72.8
P. freii 34325	X	35.4	78.2	47.2	7	8.22	0.66	99.2	12256	68.7
P. fimorum 29495	X	28.6	59.4	48.2	7	8.02	3.64	98.9	10990	73.3
P. expansum R19		32.9	65	47.3	16	8.17	4.81	99.6	12221	72.1
P. expansum 35385	X	31.8	25.1	47.8	12	5.18	0.51	95.5	12393	70.9
P. digitatum PdW03		26.3	584	48.9	6	4.12	3.09	99.1	9464	70.4
P. digitatum DSM62840		26.6	125	48.9	31	6.6	0.1	94.7	9875	68.3
P. diatomitis 30728	X	34	93.9	48.9	20	3.37	0.34	97.7	9591	70.4
P. desertorum 17660	X	34	146.6	48.5	15	5.47	0.4	97.7	13053	67.2
P. daleae 16125	X	39.7	44.3	49.1	11	4.75	2.1	98.6	12852	72.7
P. crustosum 35664	X	33	136.9	47.8	11	9.36	4.31	98.4	12403	71.9
P. cosmopolitanum 29677	X	40	25.5	48.4	15	4.37	0.42	95.9	14319	71
P. coprophilum 35676	X	29	192.7	48	8	6.1	1	98.3	10991	72.6
P. concentricum 3081	X	30	103.6	48.3	7	9.14	2.9	99.2	11716	71.9
P. citrinum 23319	X	31.4	168.4	46.4	12	3.83	1.23	96.8	11679	73.2
P. cinerascens 15544	X	29.5	179.9	50.7	16	3.99	2.25	99.2	11029	74.5
P. chrysogenum 35668	X	32.4	103	48.9	5	9.49	6.34	98.7	11981	69.5
P. chrysogenum 3361	X	34	205.2	48.8	12	4.64	0.63	98.4	12255	68.9
P. chrysogenum 19737	X	35.2	284	48.8	15	4.88	0.41	98.8	12549	67.3
P. chermesinum 19713	X	27.6	29.6	50.5	10	4.12	0.95	91.5	10822	69.4
P. cf viridicatum 20477	X	35.3	78.4	47.4	11	4.58	1.35	99.3	12613	69
P. cf spinulosum 18751	X	30.7	107.9	49.8	15	3.33	0.43	98.6	11172	73.2
P. cf griseofulvum 17755	X	29.5	56.3	47.8	13	3.54	0.79	98	11228	71.2
P. cf griseofulvum 16849	X	29.9	75.2	47.9	9	5.59	0.79	99.1	11336	70.5
P. cf griseofulvum 16848	X	29.8	29.6	47.8	11	4.29	1.27	97.8	11386	70.9
P. cataractarum 29864	X	37.5	104.7	49.3	12	4.66	1.97	99.2	12852	74.4
P. capsulatum LiaoWQ-2011		34.3	231	49.1	62	3.3	0.04	99.7	12832	73.2
P. capsulatum ATCC48735		34.4	211	49.1	65	3.19	0.04	99.7	12760	73.4
P. capsulatum 21917	X	27.5	93.2	51.3	10	4.7	0.53	97.6	10739	70.3
P. canescens 19259	X	32.9	106.5	49	11	3.49	1.1	99	12350	74.5
P. canescens 18980	X	41.8	42.8	48.9	20	3.67	0.3	97.9	13357	71.4
P. canescens 15452	X	46.5	46.4	48.9	22	3.67	0.24	98.2	13987	70.7
P. canescens 15451	X	46.3	30.9	48.8	16	4.83	0.34	97.6	14193	69.4
P. canescens 15450	X	46.4	54.2	48.8	18	4.1	0.34	98.5	14110	70.1
P. canescens 15449	X	46.5	36.5	48.8	24	3.39	0.34	98.2	14062	70.2
P. canescens 13549	X	41.6	54.8	48.9	16	3.69	0.63	98.9	13392	69
P. canariense 26290	X	31.5	88.5	51.4	9	4.8	2.45	98.3	10817	73
P. camemberti FM013 LCP06093		35.3	34	48.2	56	1.5	0.04	92.4	13530	67
P. brevicompactum 35675	X	32.5	165	49.2	8	5.2	1.25	97.9	12067	73.4
P. brevicompactum 35673	X	29.9	212.7	49.4	8	6.14	2.35	98	11498	74.8
P. brevicompactum 35665	X	34.8	120	49	10	6.13	0.83	97.9	12424	72
P. bovifimosum 22155	X	27.2	72.6	50.6	10	3.35	1.77	98.5	10403	69.8
P. aurantiogriseum 35659	X	32.8	139.2	47.5	8	9.16	4.64	98.1	12292	70.1
P. atrosanguineum 34669	X	30.5	129.8	50.1	10	3.85	2.02	97.3	11471	75.4
P. atrosanguineum 31633	X	30.4	68	47.7	10	4.42	1.2	99.2	10603	73.2
P. atrosanguineum 31632	X	30	48.1	50.1	16	3.48	0.44	96.3	11535	75.3
P. atrosanguineum 21472	X	31	49.9	50	12	3.79	1.94	98.6	11602	75.6
P. atrosanguineum 20685	X	28.6	49.7	50.2	9	3.89	2.29	98.5	11056	76.8
P. argentinense 30761	X	33.8	45	51.1	11	5.06	1.25	96.9	12157	72.3
P. antarcticum 31339	X	30.9	192.8	48.8	12	3.81	0.99	99.1	11239	74.1
P. angulare 30069	X	38.6	73.6	45.5	13	5.06	0.62	98.7	13749	72.6
P. angulare 27051	X	37.8	80.4	45.7	13	5.07	1.25	97.8	13400	69.1
P. alfredii 34128	X	27.5	61.7	50.1	12	3.33	1.18	98.3	10219	70.8

Fig. 1 Genome characteristics. X indicates genomes sequenced in this study. Sequencing quality parameters and key genome numbers are illustrated by nine bubble plots. The bubble sizes are normalized and thus comparable within each column only, and not across columns

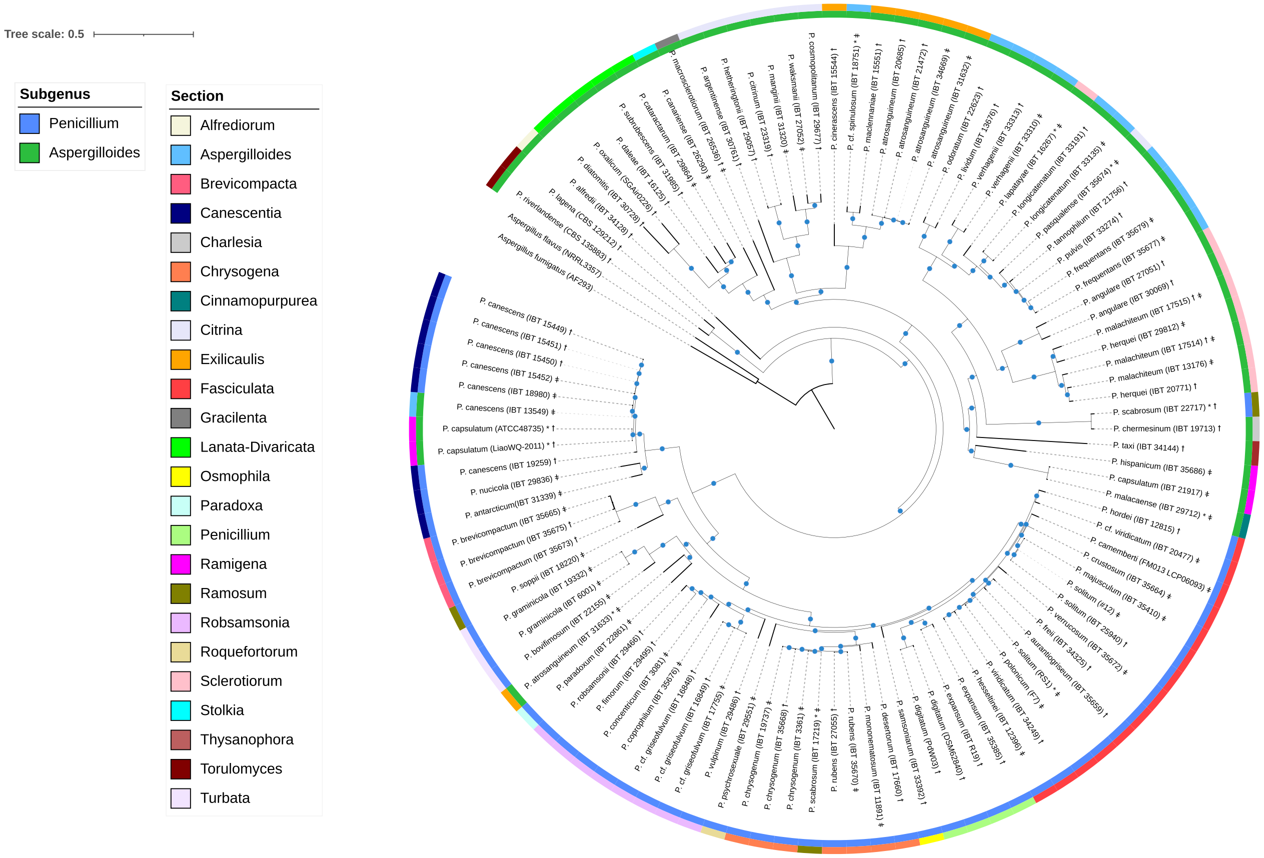

Tree scale: 0.5
Subgenus
Penicillium
Aspergilloides
Section
Alfrediorum
Aspergilloides
Brevicompacta
Canescentia
Charlesia
Chrysogena
Cinnamopurpurea
Citrina
Exilicaulis
Fasciculata
Gracilenta
Lanata-Divaricata
Osmophila
Paradoxa
Penicillium
Ramigena
Ramosum
Robsamsonia
Roquefortorum
Sclerotiorum
Stolkia
Thysanophora
Torulomyces
Turbata
Aspergillus fumigatus (AF293)
Aspergillus flavus (NRRL3357)
P. riverlandense (CBS 135883) †
P. lagena (CBS 129212) †
P. alfredii (IBT 34128) †
P. diatomitis (IBT 30728) †
P. oxalicum (SGAir0226) †
P. daleae (IBT 16125) †
P. subrubescens (IBT 31985) †
P. cataractarum (IBT 29864) ‡
P. canariense (IBT 26290) ‡
P. macrosclerotiorum (IBT 26536) † ‡
P. argentinense (IBT 30761) †
P. hetheringtonii (IBT 29057) †
P. citrinum (IBT 23319) †
P. manginii (IBT 31320) ‡
P. waksmanii (IBT 27052) ‡
P. cosmopolitanum (IBT 29677) †
P. cinerascens (IBT 15544) †
P. cf. spinulosum (IBT 18751) * ‡
P. maclennaniae (IBT 15551) †
P. atrosanguineum (IBT 20685) †
P. atrosanguineum (IBT 21472) †
P. atrosanguineum (IBT 34669) ‡
P. atrosanguineum (IBT 31632) ‡
P. odoratum (IBT 22623) †
P. lividum (IBT 13676) †
P. verhagenii (IBT 33313) †
P. verhagenii (IBT 33310) ‡
P. lapatayae (IBT 16267) * ‡
P. longicatenatum (IBT 33191) †
P. longicatenatum (IBT 33135) ‡
P. pasqualense (IBT 35674) * ‡
P. tannophilum (IBT 21756) †
P. pulvis (IBT 33274) †
P. frequentans (IBT 35679) ‡
P. frequentans (IBT 35677) ‡
P. angulare (IBT 27051) †
P. angulare (IBT 30069) †
P. malachiteum (IBT 17515) † ‡
P. herquei (IBT 29812) ‡
P. malachiteum (IBT 17514) † ‡
P. malachiteum (IBT 13176) ‡
P. herquei (IBT 20771) †
P. scabrosum (IBT 22717) * †
P. chermesinum (IBT 19713) †
P. taxi (IBT 34144) †
P. hispanicum (IBT 35686) ‡
P. capsulatum (IBT 21917) ‡
P. malacaense (IBT 29712) * ‡
P. hordei (IBT 12815) †
P. cf. viridicatum (IBT 20477) ‡
P. camemberti (FM013 LCP06093) ‡
P. crustosum (IBT 35664) ‡
P. majusculum (IBT 35410) ‡
P. solitum (#12) ‡
P. solitum (IBT 25940) †
P. verrucosum (IBT 35672) ‡
P. freii (IBT 34325) †
P. aurantiogriseum (IBT 35659) †
P. solitum (RS1) * ‡
P. polonicum (F7) ‡
P. viridicatum (IBT 34249) †
P. hesseltinei (IBT 12396) ‡
P. expansum (IBT 35385) †
P. expansum (IBT R19) †
P. digitatum (DSM62840) †
P. digitatum (PdW03) †
P. samsoniarum (IBT 33392) †
P. desertorum (IBT 17660) †
P. mononematosum (IBT 11891) ‡
P. rubens (IBT 35670) ‡
P. rubens (IBT 27055) †
P. scabrosum (IBT 17219) * ‡
P. chrysogenum (IBT 3361) ‡
P. chrysogenum (IBT 35668) †
P. chrysogenum (IBT 19737) ‡
P. psychrosexuale (IBT 29551) ‡
P. vulpinum (IBT 29486) †
P. cf. griseofulvum (IBT 17755) ‡
P. cf. griseofulvum (IBT 16849) †
P. cf. griseofulvum (IBT 16848) †
P. coprophilum (IBT 35676) ‡
P. concentricum (IBT 3081) ‡
P. fimorum (IBT 29495) †
P. robsamsonii (IBT 29466) †
P. paradoxum (IBT 22861) ‡
P. atrosanguineum (IBT 31633) * ‡
P. bovifimosum (IBT 22155) ‡
P. graminicola (IBT 6001) ‡
P. graminicola (IBT 19332) ‡
P. soppii (IBT 18220) ‡
P. brevicompactum (IBT 35673) †
P. brevicompactum (IBT 35675) †
P. brevicompactum (IBT 35665) ‡
P. antarcticum(IBT 31339) ‡
P. nucicola (IBT 29836) ‡
P. canescens (IBT 19259) †
P. capsulatum (LiaoWQ-2011) * †
P. capsulatum (ATCC48735) * †
P. canescens (IBT 13549) ‡
P. canescens (IBT 18980) ‡
P. canescens (IBT 15452) ‡
P. canescens (IBT 15450) †
P. canescens (IBT 15451) †
P. canescens (IBT 15449) †

Fig. 2 Representation of the phylogenetic relationship between the 104 *Penicillium* isolates. The phylogeny is based on 1142 BUSCO families. *Aspergillus fumigatus* AF293 and *Aspergillus flavus* NRRL 3357 were chosen as outgroups. The inner circle is colored according to the subgenera whereas the outer circle is colored according to sections as described in Houbraken et al. (2020). The blue circles state nodes with a bootstrap value of ≥ 95%. The tree scale is amino acid substitutions per site. *P. capsulatum* (ATCC48735)* and *P. solitum* (RS1)* have been reidentified as *P. canescens* (ATCC48735) and *P. polonicum* (RS1), respectively by Houbraken et al. (2021). Asterisk indicates isolates that have been misclassified. † and ‡ list mating type MAT1-1 and MAT1-2, respectively

on whole genome alignment (Additional file 2: Figure S1-S4).

The BUSCO phylogeny and the placement of subgenera displays a high degree of congruence with a previous phylogenetic analysis of the *Penicillium* genus (Houbraken et al. 2020); that analysis was based on a set of three gene markers (*BenA*, *CaM* and *RPB2*). The *Ramosum* sect. consisting of *P. soppii*, *P. scabrosum* (IBT 17219), and *P. scabrosum* (IBT 22717) were distinct from each other in the conserved BUSCO gene phylogeny. As mentioned before, we believe a misidentification of the two *P. scabrosum* accessions (IBT 22717, IBT17219) has occurred (* MERGEFORMAT Fig. 2, Additional file 2: Figure S1-4, and Additional file 2: Figure S6). Conversely, the phylogenetic position of *P. soppii* was in congruence with Houbraken et al. (2020), and it is presumed to show the correct phylogenetic position for the sect. *Ramosum*. According to Houbraken et al. (2020), *P. samsoniarum* (sect. *Osmophila*) and *P. psychrosexuale* (sect. *Roquefortorum*) should cluster together, but this is not the case in the phylogenetic based on the conserved BUSCO gene set (* MERGEFORMAT Fig. 2). However, the phylogeny based on whole genome alignment showed that *P. samsoniarum* is closely related to *P. psychrosexuale*. Given the expected increased power of whole genome alignment to elucidate close relationships, we believe that *P. samsoniarum* and *P. psychrosexuale* are more closely related phylogenetically than suggested in conserved BUSCO phylogeny. Furthermore, the clustering pattern of sect. *Charlesia*, *Ramigena*, and *Thysanophora* was slightly different compared to Houbraken et al. (2020). We further observed that the *Cinnamopurpurea* sect. consisting of only *P. malacaense* was included in the *Ramigena* sect. due to the close clustering with *P. capsulatum* (IBT 21917). We propose that *P. malacaense* is a *P. capsulatum* spp. resulting in the *Cinnamopurpurea* sect. not being represented in the phylogeny at all. A few other examples of inconsistencies can likewise be observed in the sect. *Exilicaulis* and *Aspergilloides*: *P. pasqualense*, *P. lapatayae*, and *P. cf. Spinulosum* (* MERGEFORMAT Fig. 2).

Given the aforementioned observations *P. atrosanguineum* (IBT 31633), *P. scabrosum* (IBT 17219), *P. scabrosum* (IBT 22717), *P. malacaense* (IBT 29712), *P. pasqualense* (IBT 35674), *P. lapatayae* (IBT 16267), and *P. cf. Spinulosum* (IBT 18751) should for now be reidentified as *Penicillium* sp. (IBT 31633x), *P. chrysogenum* (IBT 17219x), *P. chermesinum* (IBT 22717x), *P. capsulatum* (IBT 29712x), *Penicillium* sp. (IBT 35674x), *Penicillium* sp. (IBT 16267x), and *Penicillium* sp. (IBT 18751x), respectively.

Penicillia can be heterothallic or homothallic. This is determined by the two loci MAT1-1 and MAT1-2. Heterothallic strains harbor either MAT1-1 or MAT1-2 locus whereas homothallic strains harbor both loci next to each other (Dyer and Kück 2017; Dyer et al. 2016). MAT1-1 and MAT1-2 can be found across the phylogeny (* MERGEFORMAT Fig. 2). The majority of the *Penicillium* isolates share the same gene structure with the mating loci located between the *SLA2* gene and the *APN2* gene. A total of 99 isolates are heterothallic as expected, with 53 isolates carrying the *MAT1-1-1* gene and 46 isolates carrying the *MAT1-2-1* gene. Surprisingly, however, both *MAT1-1-1* and *MAT1-2-1* gene can be found in *P. macrosclerotiorum* (IBT 26536), *P. malachiteum* (IBT 17514), *and P. malachiteum* (IBT 17515) making them homothallic (Figure S7-S8). By manual inspection of read data we were able to identify single reads spanning both loci, confirming that both MAT1-1-1 and MAT1-2-1 genes are present in the same genome molecule.

The Penicillium pangenome contains a large accessory gene pool

The predicted proteomes from the 104 genome assemblies were used to define and characterize the pangenome of *Penicillium* to uncover the genetic diversity of the genus. A pangenome is defined as the entire set of genes of a strain, species, or genus. It can be divided into core and accessory genome, which are orthogroups present in all isolates and orthogroups that are not shared by all isolates, respectively. As part of the core genome, 2249 orthogroups were defined since these were present in all isolates, whereas 5612 orthogroups were identified as part of the softcore genome, a subdivision of core genome where an orthogroup can be found in at least 98 isolates (95% of 104 isolates) (Fig. 3a). The softcore genome also includes the orthogroups from core genome. The accessory genome can be further divided into a shell genome corresponding to orthogroups found in 3–97 isolates and a cloud genome corresponding to orthogroups found in two isolates or singleton genes. Respectively, 24,607 and 106,563 orthogroups belongs to shell and cloud genome. Due to their wide distribution within the genus, orthogroups belonging to the core/softcore and shell genomes likely represent true protein coding genes, conversely, a proportion of the cloud genome is likely to not represent functional genes, but instead gene fragments, pseudogenes, and errors during gene calling. This is supported by the observation that the median lengths of every protein sequence in the core (486 aa), softcore (471 aa), and shell (356 aa) are rather similar, whereas the median length of the cloud (145 aa) orthogroups is much lower. (Fig. 3b). For this reason, the shell genome likely best describes the accessory genome and further work with the accessory genome was performed using the shell

(See figure on next page.)

Fig. 3 Pangenome characteristics. **a** The distribution of orthogroups in the core, softcore, shell, and cloud genome of *Penicillium*. The x-axis lists the different orthogroups whereas the y-axis lists the different assemblies. Purple and grey list whether the orthogroup is present or absent, respectively. The number under each panel lists the number of orthogroups in the different part of the pangenome. The panels were scaled after the number of orthogroups. Thus, softcore genome is approximately two and half times larger than core genome, whereas shell genome is approximately four times larger than softcore genome. The exception is the cloud genome that should be four times larger than shell genome, but instead is four times smaller. **b** The distribution of lengths of all protein sequences in the pangenome. The y-axis is discontinuous. All distributions were significantly different from each other (two-sided Mann–Whitney U-test) as indicated by stars. **c** The cumulative size of the *Penicillium* gene pool. The analysis was performed ten times with random addition of genomes. **d** The distribution of the median d_N/d_S value for each orthogroup. The y-axis is discontinuous. The two distributions are significantly different from each other as indicated by stars corresponding to a p value of $< 2 \times 10^{-16}$ using two-sided Mann–Whitney U-test. **e** KOG functional classification of the pangenome. Only the best hit from each orthogroup was included and if this hit had multiple categories, all of them are plotted. The R and S category were discarded in the plot. A, RNA processing and modification; B, chromatin structure and dynamics; C, energy production and conversion; D, cell cycle control, cell division and chromosome partitioning; E, amino-acid transport and metabolism; F, nucleotide transport and metabolism; G, carbohydrate transport and metabolism; H, coenzyme transport and metabolism; I, lipid transport and metabolism; J, translation, ribosomal structure, and biogenesis; K, transcription; L, replication, recombination, and repair; M, cell wall/membrane/envelope biogenesis; N, cell motility; O, post-translational modification, protein turnover, and chaperones; P, inorganic ion transport and metabolism; Q, secondary metabolites biosynthesis, transport, and catabolism; T, signal transduction mechanisms; U, intracellular trafficking, secretion, and vesicular transport; V, defense mechanisms; W, extracellular structures; Y, nuclear structure; and Z, cytoskeleton

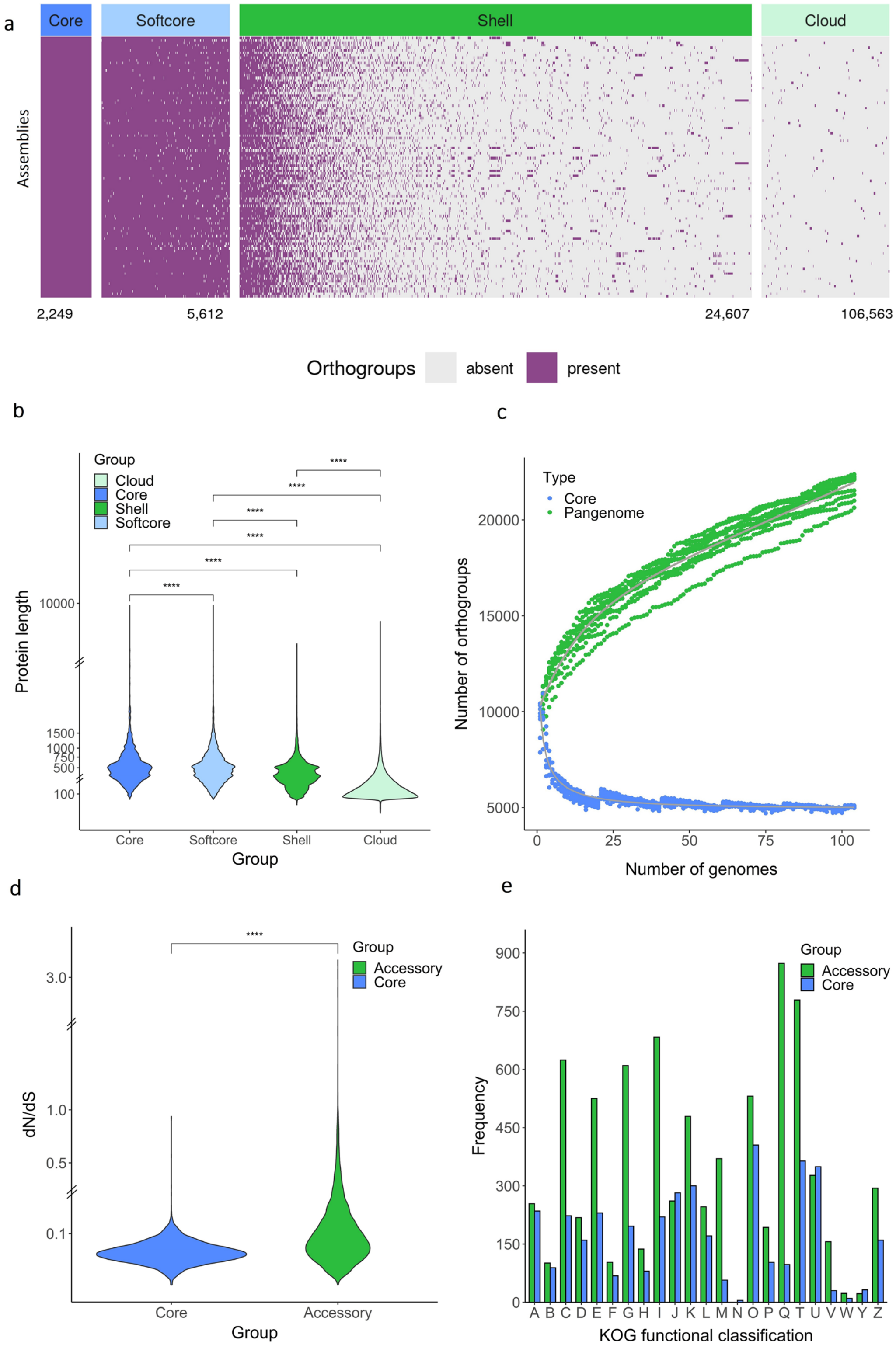


Fig. 3 (See legend on previous page.)

genome only. Henceforth, the term accessory genome will be used instead of shell. Furthermore, from the observed BUSCO completeness of 97.87% we can expect that the chance of observing a true core genome orthogroup in all genomes is only $0.9787^{104} = 10.66\%$, rendering the strict definition of core genome meaningless. As a consequence, the softcore genome is taken as representing the true core genome and henceforth the term core genome will be used.

Predicted proteins with a length > 150 aa were used to estimate the size of the pan- and core genome size by random addition of genome sequences. This process was replicated 10 times. The saturation of the core genome indicates a relatively closed pangenome, meaning that the addition of new genome assemblies did not substantially decrease the number of orthogroups (Fig. 3c) in the core genome. In contrast, addition of new genome sequences to the pangenome adds new orthogroups to the gene pool, indicating an open pangenome. The same analysis was also performed on subsets of data that were extracted based on protein length from Fig. 3b. From here, it is clear that protein sequences ≤ 150 aa did not show any saturation whereas three subsets, 151–750 aa, 751–1500 aa, and > 1500 aa showed saturation as protein length increased (Additional file 2: Figure S9). It is likely that predicted proteins with a length < 150 aa contains a larger proportion of gene fragments, pseudogenes, and other errors than larger proteins. Taken together, these results show that while the core genome of *Penicillium* is well characterized in this pangenome model, the accessory genome is less comprehensively described and while showing signs of saturation, considerable gene, and therefore potential for metabolic diversity, is not included in the present pangenome model. Regardless, the large accessory genome observed agrees with the observed diversity of *Penicillium* habitats, enabling isolates to survive in diverse environmental niches.

The distribution of gene ontology (GO) terms between the core genes and the entire gene pool as well as the accessory genes and the entire gene pool was also analyzed (Additional file 3). As expected, GO terms involved in general housekeeping such as development and basic cellular function, were significantly enriched in the core genome, and included "cellular metabolic process", "protein binding", and "ATP binding"; in total, 140 GO terms were enriched. Conversely, "transmembrane transport", "monooxygenase activity", "heme binding", and "FAD binding" etc. were significantly unrepresented in the core genome; in total, 12 GO terms were underrepresented. The opposite was observed in the accessory genome where GO terms which originate from general housekeeping functions were significantly underrepresented; in total, 603 GO terms were underrepresented. Furthermore, the accessory genome was significantly enriched for "secondary metabolite biosynthetic process" together with "mycotoxin biosynthetic process" and "heme binding"; in total, 25 GO were enriched. This validates the assumption that the core genome controls general housekeeping systems that are needed in all members of the genus, whereas the accessory genome varies within the genus and enables differential response to external stimuli, including the capacity to synthesize commercially interesting secondary metabolites and enzymes. In agreement with this analysis, heme binding was also observed to be significantly enriched in the accessory genome of *Aspergillus fumigatus* (Barber et al. 2021). Heme is found in decaying organic matter and can either be utilized as cofactors in many enzymes directly, or converted to iron (Kornitzer and Roy 2020). Additionally, "transition metal ion binding" and "zinc ion binding" were likewise significantly enriched in the accessory genome together with "monooxygenase activity" and "oxidoreductase activity", which are tailoring enzymes and their cofactors involved in the biosynthesis of several secondary metabolites. However, significant underrepresentation of these tailoring enzymes and their cofactors was observed in the core genome, indicating the versatility of secondary metabolites across the genus of *Penicillium*.

The rate of non-synonymous-to-synonymous substitutions (d_N/d_S) for each orthogroup were calculated to reveal evolutionary selection in the core and accessory genome (Fig. 3d). Overall, both core and accessory genes may be under a general trend of purifying selection as indicated by d_N/d_S values lower than one. This observation is in line with another analysis on a pangenome from a filamentous fungus (Wyka et al. 2022). Not surprisingly, lower average d_n/d_s ratio was observed in core genes compared to accessory genes. The core genome consists of conserved genes that are generally under high purifying selection whereas the purifying selective pressure is expected to be less on the accessory genome. Since only a few genes were found to be under positive selection (d_N/d_S values > 1), d_N/d_S values significantly less or higher than mean in both the core and accessory genome were found (Additional file 4—Table S11-14) and GO enrichment analyzes were performed on these genes (Additional file 4: Table S15). Within the core genome, 124 orthogroups were found to have d_N/d_S values significantly less than the mean, and seven orthogroups were found to have significantly higher values than mean. Within the accessory genomes, 45 and 27 orthogroups were found to have significantly less or higher d_N/d_S values than the mean values, respectively. Considering the core genome, no GO enrichment was observed among the seven orthogroups that had d_N/d_S values significantly higher than the mean, simply because of the low number of orthogroups. However, one notable orthogroup encodes an ABC transporter. Fungal ABC transporter are involved in the efflux of, amongst others, natural toxic compounds such as xenobiotics (Víglaš and Olejníková 2021). Positive selection of such genes has been observed within competing organisms in the same ecological niches or in symbiotic relationships (Derbyshire 2020; Vallender and Lahn 2004). Twenty-five GO terms were enriched between the 124 orthogroups that had d_N/d_S values significantly less than the mean, with the majority related to "GTP binding" and "GTPases" such as genes encoding RAS family, ADP-ribosylation factor family, and septins. Furthermore, enrichment involving "transferase activity, transferring phosphorus-containing groups" was observed within genes encoding protein kinase domain, 4′-phosphopantetheinyl transferase superfamily, FGGY family of carbohydrate kinases, and RNA polymerase. None of the 27 orthogroups that had significantly larger d_N/d_S values than the mean in the accessory genome had any GO functional annotation. Further, NCBI database searches of these 27 orthogroups returned five genes with results (zinc finger, protein kinase-like domain, and rRNA intron-encoded homing endonuclease), four genes with no significant results, and the rest were results that matched hypothetical proteins. None of these hits were immediately noteworthy, nor consistent, and we therefore conclude that no functional categories under positive selection can be identified. Further, we find it likely that the majority of these genes are in fact not truly functional genes, but pseudogenes and gene fragments that do not give rise to functional gene products. Such nonfunctional genes are not expected to be under the same selective pressure as functional genes. In contrast, enrichment towards "mitotic spindle" (two orthogroups) was observed in the 45 orthogroups that had significantly lower d_N/d_S values than mean in the accessory genome. The orthogroups encode DASH complex subunit Dad4. However, it may not be surprising that a component of the DASH complex, a central part of the mitotic spindle, apparently diverse enough to fall in the accessory genome category in our pangenome analysis, is among the genes under most purifying selection in this category, considering the importance of the mitotic spindle to one of the most essential biological processes—mitosis.

A KOG functional classification of the pangenome showed that 77.5% and 37.3% of the orthogroups belonging to core and accessory genome, respectively, could be classified (Fig. 3e). The core genome contained 25% metabolic genes and 54.4% non-metabolic genes whereas the accessory genome was more equally distributed with 36.2% metabolic genes and 39.2% non-metabolic genes. "Posttranslational modification, protein turnover, chaperones" (O) was mostly represented in the core genome along with "signal transduction mechanisms" (T) and "intracellular trafficking, secretion, and vesicular transport" (U). As expected, the accessory genome was dominated by "secondary metabolites biosynthesis, transport and catabolism" (Q) followed by "signal transduction mechanisms" (T) and "lipid transport and metabolism" (I).

Fiftteen unique LGT events have happen in Penicillia

In order to locate putative LGT events, we concatenated all 104 *Penicillium* proteomes, yielding a query database of 1,244,528 proteins. Each query protein was searched against a taxonomically diverse sequence database and queries with top hits to bacterial species were retained for further searches against the UniProt reference proteomes and NCBI's non-redundant protein database. In total, 625 *Penicillium* proteins were found to have a top hit to a bacterial species. These 625 proteins were grouped into 56 orthogroups, and phylogenetic trees were reconstructed for each orthogroup and their homologs (see Methods). Individual phylogenetic trees were manually inspected to confirm incongruent relationships relative to the species phylogeny. A study by Wang and Ruan et al. (2020) have identified ancestral LGT events into other fungal genera. Therefore, we discarded phylogenies that displayed evidence of ancestral LGT into other fungal genera and found support for 15 unique LGT events in the *Penicillium* genus only (Additional file 2: Figure S10-24, * MERGEFORMAT Table 1, Additional file 5). These 15 events represent 139 genes in total (* MERGEFORMAT Table 1 and Additional file 5) and are distributed across 82 of the 104 *Penicillium* genomes under consideration (Figure S25). Furthermore, when multiple strains per species are excluded, these 15 events are located in 57 unique *Penicillium* species. Overall, 95 of the 139 LGT genes lack introns (68%) compared to 287,428 of all 1,244,528 (23.1%) *Penicillium* proteins in our dataset, indicating the LGT genes have a different intron profile relative to the genomes they are located in. Based on SignalP v6.0 predictions, none ofthe products of the LGT event genes were secreted.

Ten of the putative LGT genes (LGT1-10, * MERGEFORMAT Table 1) are located in > 2 genomes and are considered high confidence inferences. Three of these LGT events (LGT1,7 & 8) have been reported previously (Wang and Ruan 2020). The number of LGT events per genome ranges between 0 and 3 (Additional file 2: Figure S25). LGT1 and LGT8 are the most ancestral transfers and occurred before the speciation of the majority of *Penicillium* species (Additional file 2: Figure S25); however, their patchy phylogenetic distribution is indicative of extensive gene loss in individual species and

Table 1 Functional annotation of LGT events

	# LGT proteins in Tree	# Genomes represented	Annotation	Average length of LGT proteins	Average % ID to bacterial species	Average # introns
LGT1[1]	14	14	Glycoside hydrolase 105 family protein	389	59	1.1
LGT2	21	21	FAD-binding protein	524	62	0.1
LGT3	6	6	10 kDa chaperonin	96	73	3.0
LGT4	18	10	Amidohydro_3 domain-containing protein	562	54	0.4
LGT5	2	2	Zinc-binding alcohol dehydrogenase	331	61	0.5
LGT6	2	2	Elongation factor Tu	481	69	0.0
LGT7[2]	52	51	Short-chain dehydrogenase	227	79	0.0
LGT8[3]	13	12	SMI1_KNR4 domain-containing protein	187	43	0.8
LGT9	4	4	Acyl carrier protein	78	78	3.0
LGT10	2	2	Peptidyl-prolyl cis–trans isomerase	162	61	0.0
LGT11	1	1	NAD-dependent epimerase/dehydratase	304	61	0.0
LGT12	1	1	NADP octopine/nopaline dehydrogenase	362	51	0.0
LGT13	1	1	DUF4976 domain-containing protein	483	62	1.0
LGT14	1	1	Cold-shock protein	71	66	0.0
LGT15	1	1	Extracellular solute-binding protein	278	60	0.0

[1,2,3] correspond to HGT56, HGT1, and HGT18 previously located by Wang and Ruan et al. (Wang and Ruan 2020)

lineages. Alternatively, other LGT events, such as LGT2 and LGT7, occurred along specific lineages and have been maintained. The majority of LGT genes are single copy, but LGT4 shows evidence of a tandem duplication event, and eight of the 10 genomes contain a paralog. Synteny (not shown) and phylogenetic analysis (Additional file 2: Figure S13) indicate that this duplication occurred in the ancestor of the species that contain these genes (Additional file 2: Figure S25).

DISCUSSION

Vast improvements in sequencing technologies have occurred in recent time, including long read DNA sequencing technologies such as Oxford Nanopore Technologies. In turn, these developments facilitate increased insight into fungal genetics by permitting mid or high throughput sequencing projects of groups of organisms. In this study, we have sequenced, de novo assembled, and annotated 93 *Penicillium* isolates and provided a significant contribution of high contiguity and completeness to the number of publicly available *Penicillium* genome drafts (358 genome drafts were available at NCBI in October 2022 of which 15 originate from long read sequencing data). While the genomes are of high contiguity, completeness, and elucidate the potential *Penicillium* spp. have to produce a vast number of secondary metabolites with pharmaceutical, agricultural, or industrial value, a note of caution is warranted. Sequencing errors are inherent to DNA sequencing and the error profile of Oxford Nanopore Technologies is not completely random but biased towards indel errors in homopolymer regions. This problem cannot be easily overcome by increasing read coverage but requires different sequencing technology, such as Illumina short-reads (Stoler and Nekrutenko 2021) or newer versions of Oxford Nanopore Technologies (Sereika et al. 2022).

A benefit of the use of whole genome data is that it facilitates the use of genome-scale phylogeny instead of a few marker genes only. The inclusion of more data could potentially add more resolution and power to the analysis. However, it would be expected that, long range phylogenetic relationships are better represented by hyperconserved regions of the genome, e.g. a limited set of hyperconserved marker genes, since more variable parts would undergo multiple substitutions acting as "noise" in long range analysis. Conversely, the variable parts of the genome are better at resolving evolutionary relationships of closely related organisms, and it is to the analysis of these variable parts that whole genome alignment can contribute. Further, a relevant set of conserved genes, which can serve as "intermediate" variable genes, is the BUSCO gene set. In this paper, we have conducted phylogenetic analysis with the whole genome, BUSCO gene set, as well as presence/absence of orthologous genes (see Methods for details) and compared it with already published phylogeny based on three marker genes (Houbraken et al. 2020). In general, the four different approaches are highly congruent with one other, indicating convergence of the phylogenetic analyses to the true evolutionary relationship of the *Penicillium* genomes. There are, however, a few noteworthy exceptions. The unique placement of *P. alfredii*, *P. lagena*, and

P. riverlandense outside the two subgenera is unexcepted and stands in contrary to what was observed by Houbraken et al. (2020). *P. alfredii* was isolated from house dust on an island in the Federated States of Micronesia (Visagie et al. 2014b). *P. lagena and P. riverlandense* were isolated from soil in the USA, and bract from *Protea repens* infructescens in South Africa, respectively, according to the CBS strains database. Visagie et al. (2016) also observed inconsistencies between the individual phylogenetic trees forITS, BenA, CaM, and RPB2, respectively, for the *Torulomyces* sect. (*P. lagena* and *P. riverlandense*). Assuming that the unique placement of *P. alfredii*, *P. lagena*, and *P. riverlandense* in our analysis is true, it opens the possibility that they diverged from the *Penicillium* spp. prior to the divergence of the two subgenera *Penicillium* and *Aspergilloides* and thereby form the basis for a new subgenus. However, further studies regarding the matter should be investigated to support this suggestion.

With the increasing accessibility of sequencing data, pangenomic analyses are becoming more widespread in comparative genomics. A pangenome is a visualization of all available genomic information of a group of organisms to get a better understanding of the genomic makeup of a genus or species. Pangenome analyses are more common for bacteria and plants but recently some studies of pangenomes in fungi have been published as well (Barber et al. 2021; Wyka et al. 2022; Nielsen et al. 2017; Wang et al. 2021). To our knowledge, this is the first study that investigates a fungal pangenome of an entire genus with such a large number of high-quality de novo assemblies. By revealing the genomic diversity of a genus, variations and common characteristics of numerous species can be assessed in a relatively simple way. The pangenome can therefore function as an atlas to provide an easy overview of a genus.

Inherent to the lack of precise methods to identify true gene models, and relying on probabilistic predictions, artefact gene predictions are very likely to be observed and result in an inflated cloud genome. To assess the extent of this, we assessed the proportion of BUSCO genes observed in the cloud genome (20.6% completeness equaling to 742 genes) but 669 of these were fragmented, supporting that the major proportion of the cloud genome consists of gene fragments. In general, the choice of prediction algorithm and their tendency for overprediction may also contribute to artefactual inflation of the cloud genome. However, we observed that FunGAP (based on Braker, Maker, and Augustus (Min et al. 2017)) gene prediction on *P. digitatum* PDW03 leads to 9464 gene models, which is quite similar to the 9003 gene models of the original publication, using Braker alone (Wang et al. 2021).

Nielsen et al. (2017) presented a pangenome from 24 *Penicillium* species and found 3249 core genes and 8784 accessory genes. From their analysis, the pangenome appeared closed, meaning that including additional members to the group would provide few new additional genes and, hence, gene functions. However, members from the subgenus *Penicillium* were overrepresented (20 species) compared to subgenus *Aspergilloides* (four species). In contrast, we do not observe a closed pangenome, but indeed expected signs of saturation of the pangenome. In our dataset, the distribution of *Penicillium* (53 isolates) and *Aspergilloides* (51 isolates) are comparable, and species were selected to represent a wide diversity. We believe that the increased diversity causes the number of observed core and accessory genes to increase to 5612 and 24,607, respectively. However, we cannot exclude that our pangenome was based on a stricter alignment parameter (minimum percent match cutoff of 75% vs. 50%). The relatively open pangenome in this study indicates that, even though we have sampled a broader diversity of the genus compared to Nielsen et al. it is not sufficient to capture all the genetic variation of *Penicillium*. Of course, given the extremely high microbial diversity on Earth, with diversity estimates ranging from millions to trillions, it might not be practically feasible to sample and thereby capture all, or even nearly all, of the genetic information of a genus. However, the concept will provide an indication of the end of the scale. Thus, we find that conducting the analysis on groups of genes sorted according to size reveals that larger genes are approximating saturation faster than smaller genes. Assuming that shorter gene models are more likely to include non-functional pseudo-genes and gene fragments than rather genes, we hypothesize that the larger gene analysis is more likely to represent the true diversity. However, we cannot rule out that larger genes in general are under higher purifying selection than shorter, which may also impact this result. Further, comparison to the pangenome obtained by Nielsen et al. (2017) found that the KOG category responsible for secondary metabolism was mainly encoded in the core genome, whereas our analysis showed that it was predominantly in the accessory genome. This again is likely caused by some orthogroups involved in secondary metabolism being categorized as core genes due to the non-representative and limited diversity sample in the study by Nielsen et al. (2017). In this study, increasing the sampled diversity resulted in the placement of secondary metabolite genes as categories in the accessory genome, which is to be expected, considering the considerable literature available on diversity of observed secondary metabolites themselves across *Penicillium* (e.g., Houbraken et al. 2020).

Overall, the *Penicillium* pangenome reconstructed in this study seems reasonable with 5,612 core genes. Given the average gene count is 11,976 per genome, this infers that approximately half of all genes are shared between the *Penicillium* isolates. The result can form a solid basis for further comparisons between *Penicillium* isolates.

It is well known that LGT plays an important role in the evolution of fungi (Fitzpatrick 2012). For example, an early LGT analysis of *S. cerevisiae* S288C showed it has acquired 13 bacterial genes and these have contributed to important functional innovations, including the ability to synthesize biotin and the ability to grow under anaerobic conditions (Hall et al. 2005). One of the most infamous incidences of LGT relates to the acquisition of a toxin gene (ToxA) by *Pyrenophora tritici-repentis* from *Stagonospora nodorum* resulting in *Pyrenophora* infestations of wheat (Friesen et al. 2006). There are also multiple studies showing that several metabolic gene clusters, whose functions are often associated with fungal virulence, have experienced LGT (Campbell et al. 2012; Greene et al. 2014). In this study, we analyzed the genomes of 104 *Penicillium* genomes (71 unique species) to determine the frequency of recent bacterial lateral gene transfers into the *Penicillium* genus. Overall, we located 15 gene transfer events, encompassing 139 genes into the genus. Overall, this accounts for ~ 0.01% of the total protein dataset utilized here and is significantly lower than the 2.9% and 3.5% reported in other fungal species (Murphy et al. 2019; Wisecaver et al. 2014), but similar to comparisons previously undertaken in the CTG clade (Fitzpatrick et al. 2008). However, our analysis ignores more ancient bacterial to fungal transfers that may have occurred before the differentiation from other closely related genera in the Eurotiales order, or indeed even more ancient gene transfer events. Furthermore, we have not accounted for recent/ancient fungal-to-fungal gene transfer events, so the overall component of the dataset that has arisen through LGT will be larger. A previous LGT study utilizing 23 unique species of *Penicillium* located 60 LGT events into the genus (Wang and Ruan 2020). However, only seven of these (termed HGT1,3,18,26,45,53,56) LGTs were unique to the genus as the remainder were located in closely related fungi and most likely acquired before speciation between phyla and genera. Our analysis successfully recovered three of these gene *Penicillium* specific transfers (HGT1,18,56,* MERGEFORMAT Table 1). We did not recover HGT1,3,18 because the proteins were either below our sequence length or similarity cutoff (see Methods). HGT26 was also discarded as there was an ortholog in *Aspergillus sydowii* (tree not shown) and it is therefore not deemed to be specific to the *Penicillium* genus. The frequency (32%) of genes with introns in the 139 putative LGT genes is lower than that observed in the *Penicillium* dataset (76.9%) and is consistent with their bacterial origin, as previously reported (Wang and Ruan 2020). Overall, the retention or loss of transferred genes varies. For example, the LGT2 event occurred at the last common ancestor to 15 species (21 genomes) and has been maintained by all of these (Figure S25). A similar pattern is observed for LGT7 where the gene transfer has occurred before the speciation of 28 unique species and has been maintained in all 36 representative genomes. Conversely, LGT1 and LGT8 are LGTs that occurred before the speciation of most of the *Penicillium* species considered here; however, their retention is very patchy and most likely associated with niche adaptation (Wang and Ruan 2020).

CONCLUSION

In this study, we have sequenced, assembled, and annotated 93 *Penicillium* isolates. A combined *Penicillium* pangenome was generated with these genome assemblies together with eleven previously published *Penicillium* genome models. The pangenome consist of a core genome containing 5612 genes and a larger accessory genome of 24,607 genes, elucidating a diverse pangenome, likely with a high potential of producing a wide range of secondary metabolites. Analysis of the phylogenetic relationship based on shared BUSCO genes of the 104 *Penicillium* isolates largely confirmed previous results obtained with three maker genes. Finally, we identified 15 lateral gene transfer events from bacteria that have occurred during evolution of *Penicillium*.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s43008-023-00108-7.

Additional file 1. Overview of isolates and software versions used in the assembly process.

Additional file 2. Additional figures S1-S25.

Additional file 3. GO enrichment.

Additional file 4. dN/dS.

Additional file 5. LGT.

Additional file 6. Orthogroups.

Acknowledgements
Not applicable.

Adherence to national and international regulations
Not applicable.

Author contributions
Conceptualization: C.P., T.E.S., J.C.F., J.L.S. and K.L.N. Investigation: C.P. Formal analysis C.P., T.S, M.R.N, D.A.F and K.L.N. Writing—original draft preparation: C.P. Writing—review and editing: C.P., T.S., M.R.N, T.E.S., J.L.S., D.A.F., J.C.F, and K.L.N. Visualization: C.P. Supervision: K.L.N. Funding acquisition, K.L.N., T.E.S., J.C.F., and J.L.S. All authors read and approved the final manuscript.

Funding
This research was funded by Novo Nordisk Foundation, Grant NNF18OC0034952.

Availability of data and materials
Data for this study have been deposited at NCBI under the BioProject number PRJNA867151. Individual references for genome assemblies and their reads can be found in Additional file 1: Table S1. The *Penicillium* pangenome structed in orthogroups can be found in Additional file 6.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 28 November 2022 Accepted: 26 January 2023
Published online: 01 February 2023

References

Adsul MG, Bastawde KB, Varma AJ, Gokhale DV (2007) Strain improvement of *Penicillium janthinellum* NCIM 1171 for increased cellulase production. Bioresour Technol 98:1467–1473. https://doi.org/10.1016/J.BIORTECH.2006.02.036

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402. https://doi.org/10.1093/NAR/25.17.3389

Barber AE, Sae-Ong T, Kang K, Seelbinder B, Li J et al (2021) Aspergillus fumigatus pan-genome analysis identifies genetic variants associated with human infection. Nat Microbiol 6(12):1526–1536. https://doi.org/10.1038/s41564-021-00993-x

Barrett K, Jensen K, Meyer AS, Frisvad JC, Lange L (2020) Fungal secretome profile categorization of CAZymes by function and family corresponds to fungal phylogeny and taxonomy: example *Aspergillus* and *Penicillium*. Sci Rep 10:5158. https://doi.org/10.1038/s41598-020-61907-1

Blin K, Shaw S, Steinke K, Villebro R, Ziemert N et al (2019) antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. Nucleic Acids Res 47:W81–W87. https://doi.org/10.1093/nar/gkz310

Bodinaku I, Shaffer J, Connors AB, Steenwyk JL, Biango-Daniels MN et al (2019) Rapid Phenotypic and Metabolomic Domestication of Wild *Penicillium* Molds on Cheese. Mbio 10:e02445-e2519. https://doi.org/10.1128/mBio.02445-19

Buchfink B, Reuter K, Drost HG (2021) Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods 18(4):366–368. https://doi.org/10.1038/s41592-021-01101-x

Campbell MA, Rokas A, Slot JC (2012) Horizontal transfer and death of a fungal secondary metabolic gene cluster. Genome Biol Evol 4:289–293. https://doi.org/10.1093/GBE/EVS011

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972. https://doi.org/10.1093/BIOINFORMATICS/BTP348

Contreras-Moreira B, Vinuesa P (2013) GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. Appl Environ Microbiol 79:7696. https://doi.org/10.1128/AEM.02411-13

De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C (2018) NanoPack: visualizing and processing long-read sequencing data. Bioinformatics 34:2666–2669. https://doi.org/10.1093/bioinformatics/bty149

Derbyshire MC (2020) Bioinformatic detection of positive selection pressure in plant pathogens: the neutral theory of molecular sequence evolution in action. Front Microbiol. https://doi.org/10.3389/fmicb.2020.00644

Dyer PS, Inderbitzin P, Debuchy R (2016) 14 Mating-type structure, function, regulation and evolution in the Pezizomycotina BT - growth, differentiation and sexuality. In: Wendland J (ed) Springer International Publishing, Cham. pp. 351–385. https://doi.org/10.1007/978-3-319-25844-7_14.

Dyer PS, Kück U (2017) Sex and the imperfect fungi. Microbiol Spectr 5(3):5–3. https://doi.org/10.1128/microbiolspec.FUNK-0043-2017

Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinform 5:1–19. https://doi.org/10.1186/1471-2105-5-113/FIGURES/16

Edgar R (2021) MUSCLE v5 enables improved estimates of phylogenetic tree confidence by ensemble bootstrapping. https://doi.org/10.1101/2021.06.20.449169.

Emms DM, Kelly S (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol 20:1–14. https://doi.org/10.1186/S13059-019-1832-Y/FIGURES/5

Fitzpatrick DA (2012) Horizontal gene transfer in fungi. FEMS Microbiol Lett 329:1–8. https://doi.org/10.1111/J.1574-6968.2011.02465.X

Fitzpatrick DA, Logue ME, Butler G (2008) Evidence of recent interkingdom horizontal gene transfer between bacteria and *Candida parapsilosis*. BMC Evol Biol. https://doi.org/10.1186/1471-2148-8-181

Fleming A (1929) On the antibacterial action of cultures of a Penicillium, with special reference to their use in the isolation of B. influenzae. Br J Exp Pathol 10:226–236

Friesen TL, Stukenbrock EH, Liu Z, Meinhardt S, Ling H et al (2006) Emergence of a new disease as a result of interspecific virulence gene transfer. Nat Genet 38:953–956. https://doi.org/10.1038/NG1839

Frisvad JC, Samson RA (2004) Polyphasic taxonomy of *Penicillium* subgenus *Penicillium* a guide to identification of food and air-borne terverticillate Penicillia and their mycotoxins. Stud Mycol 49:1–174

Frisvad JC, Smedsgaard J, Larsen TO, Samson RA (2004) Mycotoxins, drugs and other extrolites produced by species in *Penicillium* subgenus *Penicillium*. Mycol 49:201–241

Giraud F, Giraud T, Aguileta G, Fournier E, Samson R et al (2010) Microsatellite loci to recognize species for the cheese starter and contaminating strains associated with cheese manufacturing. Int J Food Microbiol 137:204–213. https://doi.org/10.1016/J.IJFOODMICRO.2009.11.014

Greene GH, McGary KL, Rokas A, Slot JC (2014) Ecology drives the distribution of specialized tyrosine metabolism modules in fungi. Genome Biol Evol 6:121–132. https://doi.org/10.1093/GBE/EVT208

Hall C, Brachat S, Dietrich FS (2005) Contribution of horizontal gene transfer to the evolution of *Saccharomyces cerevisiae*. Eukaryot Cell 4:1102–1115. https://doi.org/10.1128/EC.4.6.1102-1115.2005

Houbraken J, Kocsubé S, Visagie CM, Yilmaz N, Wang XC et al (2020) Classification of *Aspergillus*, *Penicillium*, *Talaromyces* and related genera (*Eurotiales*): an overview of families, genera, subgenera, sections, series and species. Stud Mycol 95:5–169. https://doi.org/10.1016/J.SIMYCO.2020.05.002

Houbraken J, Visagie CM, Frisvad JC (2021) Recommendations to prevent taxonomic misidentification of genome-sequenced fungal strains. Microbiol Resour Announc 10:e01074-e1120. https://doi.org/10.1128/MRA.01074-20

Jones P, Binns D, Chang H-Y, Fraser M, Li W et al (2014) InterProScan 5: genome-scale protein function classification. Bioinformatics 30:1236–1240. https://doi.org/10.1093/bioinformatics/btu031

Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods 14:587. https://doi.org/10.1038/NMETH.4285

Karahadian C, Josephson DB, Lindsay RC (1985) Volatile compounds from *Penicillium* sp. contributing musty-earthy notes to Brie and Camembert cheese flavors. J Food Sci 33:435

Klopfenstein DV, Zhang L, Pedersen BS, Ramírez F, Vesztrocy AW et al (2018) GOATOOLS: a python library for gene ontology analyses. Sci Rep 8(1):1–17. https://doi.org/10.1038/s41598-018-28948-z

Kornitzer D, Roy U (2020) Pathways of heme utilization in fungi. Biochim Biophys Acta Mol Cell Res 1867:118817. https://doi.org/10.1016/J.BBAMCR.2020.118817

Letunic I, Bork P (2021) Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res 49:W293–W296. https://doi.org/10.1093/NAR/GKAB301

Li H (2016) Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics 32:2103–2110. https://doi.org/10.1093/bioinformatics/btw152

Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li Y, Cui F, Liu Z, Xu Y, Zhao H (2007) Improvement of xylanase production by *Penicillium oxalicum* ZH-30 using response surface methodology. Enzyme Microb Technol 40:1381–1388. https://doi.org/10.1016/J.ENZMICTEC.2006.10.015

López-Díaz TM, Santos JA, García-López ML, Otero A (2001) Surface mycoflora of a Spanish fermented meat sausage and toxigenicity of *Penicillium* isolates. Int J Food Microbiol 68:69–74. https://doi.org/10.1016/S0168-1605(01)00472-X

Ludemann V, Greco M, Rodríguez MP, Basílico JC, Pardo AG (2010) Conidial production by *Penicillium nalgiovense* for use as starter cultures in dry fermented sausages by solid state fermentation. LWT Food Sci Technol 43:315–318. https://doi.org/10.1016/J.LWT.2009.07.011

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM (2021) BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol 38:4647–4654. https://doi.org/10.1093/MOLBEV/MSAB199

McCarthy CGP, Fitzpatrick DA (2016) Systematic search for evidence of interdomain horizontal gene transfer from prokaryotes to oomycete lineages. mSphere. https://doi.org/10.1128/MSPHERE.00195-16

Min B, Grigoriev IV, Choi IG (2017) FunGAP: Fungal Genome Annotation Pipeline using evidence-based gene model evaluation. Bioinformatics 33:2936–2937. https://doi.org/10.1093/BIOINFORMATICS/BTX353

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD et al (2020) IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol 37:1530–1534. https://doi.org/10.1093/MOLBEV/MSAA015

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA et al (2021) Pfam: the protein families database in 2021. Nucleic Acids Res 49:D412–D419. https://doi.org/10.1093/NAR/GKAA913

Murphy CL, Youssef NH, Hanafy RA, Couger MB, Stajich JE et al (2019) Horizontal gene transfer as an indispensable driver for evolution of Neocallimastigomycota into a distinct gut-dwelling fungal lineage. Appl Environ Microbiol. https://doi.org/10.1128/AEM.00988-19

Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW, Kautsar SA, Tryon JH et al (2020) A computational framework to explore large-scale biosynthetic diversity. Nat Chem Biol 16:60–68. https://doi.org/10.1038/s41589-019-0400-9

Nelson JH (1970) Production of blue cheese flavor via submerged fermentation by *Penicillium roqueforti*. J Agric Food Chem 18:567–569

Nielsen JC, Grijseels S, Prigent S, Ji B, Dainat J et al (2017) Global analysis of biosynthetic gene clusters reveals vast potential of secondary metabolite production in *Penicillium* species. Nat Microbiol 2:17044. https://doi.org/10.1038/nmicrobiol.2017.44

Oxford Nanopore Technologies (2018) Medaka. Available from: 715 https://github.com/nanoporetech/medaka. Accessed 23 Sept 2021

Oxford Nanopore Technologies. pyguppyclient. Available from: 702 https://github.com/nanoporetech/pyguppyclient. Accessed 23 Sept 2022

Perrone G, Susca A (2017) Penicillium species and their associated mycotoxins. Methods Mol Biol 1542:107–119. https://doi.org/10.1007/978-1-4939-6707-0_5/TABLES/1

Petersen C, Sørensen T, Westphal KR, Fechete LI, Sondergaard TE et al (2022) High molecular weight DNA extraction methods lead to high quality filamentous ascomycete fungal genome assemblies using Oxford Nanopore sequencing. Microb Genom 8:000816. https://doi.org/10.1099/MGEN.0.000816

Rokas A, Williams BL, King N, Carroll SB (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804. https://doi.org/10.1038/nature02053

Samson RA, Houbraken J, Thrane U, Frisvad JC, Andersen B (2019) Food and indoor fungi, 2nd edn. CBS Fungal Biodiversity Centre, Utrecht

Schneider WDH, Gonçalves TA, Uchima CA, Couger MB, Prade R et al (2016) *Penicillium echinulatum* secretome analysis reveals the fungi potential for degradation of lignocellulosic biomass. Biotechnol Biofuels. https://doi.org/10.1186/S13068-016-0476-3

Sereika M, Kirkegaard RH, Karst SM, Michaelsen TY, Sørensen EA et al (2022) Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. Nat Methods. https://doi.org/10.1038/s41592-022-01539-7

Sørensen T, Petersen C, Fechete LI, Nielsen KL, Sondergaard TE (2022) A highly contiguous genome assembly of *Arthrinium puccinoides*. Genome Biol Evol 14:evac010. https://doi.org/10.1093/gbe/evac010

Steenwyk JL, Shen X-X, Lind AL, Goldman GH, Rokas A (2019) A robust Phylogenomic time tree for biotechnologically and medically important fungi in the genera *Aspergillus* and *Penicillium*. Mbio 10:e00925-e1019. https://doi.org/10.1128/mBio.00925-19

Stoler N, Nekrutenko A (2021) Sequencing error profiles of Illumina sequencing instruments. NAR Genomics Bioinforma 3:019. https://doi.org/10.1093/nargab/lqab019

Sullivan MJ, Petty NK, Beatson SA (2011) Easyfig: a genome comparison visualizer. Bioinformatics 27:1009–1010. https://doi.org/10.1093/bioinformatics/btr039

Suyama M, Torrents D, Bork P (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res 34:W609–W612. https://doi.org/10.1093/NAR/GKL315

Terrasan CRF, Temer B, Duarte MCT, Carmona EC (2010) Production of xylanolytic enzymes by *Penicillium janczewskii*. Bioresour Technol 101:4139–4143. https://doi.org/10.1016/J.BIORTECH.2010.01.011

Teufel F, Almagro Armenteros JJ, Johansen AR, Gíslason MH, Pihl SI et al (2022) SignalP 6.0 predicts all five types of signal peptides using protein language models. Nat Biotechnol. https://doi.org/10.1038/S41587-021-01156-3

Thom C (1906) Fungi in cheese ripening: Camembert and Roquefort. US Dep Agric Bur Anim Ind Bull 82:1–39

Vallender EJ, Lahn BT (2004) Positive selection on the human genome. Hum Mol Genet 13:R245–R254. https://doi.org/10.1093/hmg/ddh253

Vaser R, Sović I, Nagarajan N, Šikić M (2017) Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res 27:737–746. https://doi.org/10.1101/gr.214270.116

Víglaš J, Olejníková P (2021) An update on ABC transporters of filamentous fungi–from physiological substrates to xenobiotics. Microbiol Res 246:126684. https://doi.org/10.1016/J.MICRES.2020.126684

Vinuesa P, Ochoa-Sánchez LE, Contreras-Moreira B (2018) GET_PHYLOMARKERS, a software package to select optimal orthologous clusters for phylogenomics and inferring pan-genome phylogenies, used for a critical geno-taxonomic revision of the genus Stenotrophomonas. Front Microbiol 9:771. https://doi.org/10.3389/FMICB.2018.00771/BIBTEX

Visagie CM, Houbraken J, Frisvad JC, Hong SB, Klaassen CHW et al (2014a) Identification and nomenclature of the genus *Penicillium*. Stud Mycol 78:343–371. https://doi.org/10.1016/J.SIMYCO.2014.09.001

Visagie CM, Hirooka Y, Tanney JB, Whitfield E, Mwange K et al (2014b) *Aspergillus*, *Penicillium* and *Talaromyces* isolated from house dust samples collected around the world. Stud Mycol 78:63–139. https://doi.org/10.1016/J.SIMYCO.2014.07.002

Visagie CM, Houbraken J, Dijksterhuis J, Seifert KA, Jacobs K et al (2016) A taxonomic review of *Penicillium* species producing conidiophores with solitary phialides, classified in section *Torulomyces*. Pers Mol Phylogeny Evol Fungi 36:134. https://doi.org/10.3767/003158516X690952

Wang M, Ruan R (2020) Genome-wide identification and functional analysis of the horizontally transferred genes in *Penicillium*. Genomics 112:5037–5043. https://doi.org/10.1016/J.YGENO.2020.09.025

Wang M, Ruan R, Li H (2021) The completed genome sequence of the pathogenic ascomycete fungus *Penicillium digitatum*. Genomics 113:439–446. https://doi.org/10.1016/J.YGENO.2021.01.001

Wick R (2018) Filtlong. Available from: https://github.com/rrwick/Filtlong. Accessed 704 23 Sept 2022

Wickham H (2009) ggplot2: elegant graphics for data analysis. Springer-Verlag New York

Wisecaver JH, Slot JC, Rokas A (2014) The evolution of fungal metabolic pathways. PLoS Genet. https://doi.org/10.1371/JOURNAL.PGEN.1004816

Wyka S, Mondo S, Liu M, Nalam V, Broders K (2022) A large accessory genome and high recombination rates may influence global distribution and broad host range of the fungal plant pathogen *Claviceps* *purpurea*. PLoS ONE 17:e0263496. https://doi.org/10.1371/JOURNAL.PONE.0263496

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586–1591. https://doi.org/10.1093/MOLBEV/MSM088

Yu G, Smith DK, Zhu H, Guan Y, Lam TTY (2017) ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol 8:28–36. https://doi.org/10.1111/2041-210X.12628

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.