Journal of *Fungi*

MDPI

Article

Chromosome-Level Comprehensive Genome of Mangrove Sediment-Derived Fungus *Penicillium variabile* HXQ-H-1

Ling Peng [1,†], Liangwei Li [1,†], Xiaochuan Liu [1], Jianwei Chen [1], Chengcheng Shi [1], Wenjie Guo [1], Qiwu Xu [1], Guangyi Fan [1,2], Xin Liu [1,2,*] and Dehai Li [3,*]

[1] BGI-Qingdao, BGI-Shenzhen, Qingdao 266555, China; pengling@genomics.cn (L.P.); liliangwei@genomics.cn (L.L.); Liuxiaochuan@genomics.cn (X.L.); chenjianwei@genomics.cn (J.C.); shichengcheng@genomics.cn (C.S.); guowenjie@genomics.cn (W.G.); xuqiwu@genomics.cn (Q.X.); fanguangyi@genomics.cn (G.F.)

[2] China National GeneBank, BGI-Shenzhen, Shenzhen 518120, China

[3] Key Laboratory of Marine Drugs, School of Medicine and Pharmacy, Ocean University of China, Qingdao 266100, China

[*] Correspondence: liuxin@genomics.cn (X.L.); dehaili@ouc.edu.cn (D.L.); Tel.: +86-532-5571-1134 (X.L.); +86-532-8203-1619 (D.L.)

[†] These authors contributed equally to this work.

Received: 18 November 2019; Accepted: 18 December 2019; Published: 23 December 2019

Abstract: *Penicillium* is an ascomycetous genus widely distributed in the natural environment and is one of the dominant fungi involved in the decomposition of mangroves, which can produce a variety of antitumor compounds and bioactive substances. However, in mangrove ecosystems there is no complete genome in this genus. In this study, we isolated a fungus strain named *Penicillium variabile* HXQ-H-1 from coast mangrove (Fujian Province, China). We generated a chromosome-level genome with total size of 33.32 Mb, scaffold N50 of 5.23 Mb and contig N50 of 96.74 kb. Additionally, we anchored about 95.91% assembly sequences into the longest seven scaffolds, and predicted 10,622 protein-coding genes, in which 99.66% could be annotated by eight protein databases. The secondary metabolites analysis reveals the strain has various gene clusters involving polyketide synthase (PKS), non-ribosomal peptide synthetase (NRPS) and terpene synthase that may have a largely capacity of biotechnological potential. Comparison genome analysis between *Penicillium variabile* and *Talaromyces islandicus* reveals a small difference in the total number of genes, whereas HXQ-H-1 has a higher gene number with COG functional annotation. Evolutionary relationship of *Penicillum* based on genome-wide data was carried out for the first time, showing the strain HXQ-H-1 is closely related to *Talaromyces islandicus*. This genomic resource may provide a new resource for development of novel bioactive antibiotics, drug candidates and precursors in *Penicillium variabile*.

Keywords: *Penicillium variabile*; chromosome-level genome; mangrove; secondary metabolites; phylogeny

1. Introduction

Penicillium is widely distributed in worldwide. It is known for its asexual forms and is found to inhabit soil, water, marine, mangrove swamps, decaying vegetation, air. Further *Penicillium* is a common contaminant of various substances. It plays an important role in human life for the function of decomposing various organic substances and produces antibiotics, toxic compounds (mycotoxins), extracellular enzyme, and antitumor active secondary metabolites [1–3]. Many species of *Penicillium* have been widely used in biotechnology, food, pharmaceutical, biological degradation, and many other applicable fields.

Marine-derived *Penicillium* has provided enormous excellent pharmaceutical leads over the past decades [3]. The manglicolous fungi represent the second largest marine fungi ecological group [4], and play an important role in mangrove ecosystems as the main decomposer and participant in the energy flow of mangroves, having significant effects on the survival, growth, and fitness of mangroves. In addition, secondary bioactive metabolites are derived from mangrove fungi, specially from mangrove endophytic fungi, which are used in pharmaceutical and nutraceutical industries to produce antimicrobial, anticancer, antioxidant, antidiabetic, and other therapeutic agents [5]. Here, we isolated a *Penicillium variabile* strain named HXQ-H-1 derived from mangrove. Multiple novel bioactive molecules were isolated from HXQ-H-1 by co-cultivation of *P. variabile* with *Talaromyces aculeatus* and DNA methyltransferase inhibitor respectively, of which nafuredin B and varitatin A were detected has the activity of anti-cancer cells [6,7]. The *P. variabile* strains derived from other environments have also been verified to have different metabolites. It's reported that *P. variabile* Sopp 1912 derived from permafrost deposits can synthesize clavine alkaloids, rugulovasines A and B which was toxic to animals and plants [8–10], *P. variabile* JN-A525 with urethanase activity which could directly catalyze the decomposition of urethane in wine [11]. *P. variabile* P16 produce glucose oxidase which was used for catalyzing the oxidation of glucose in industry [12]. Closely related species of *P. variabile* have been investigated at the genomic level, such as *Talaromyces islandicus* ("*Penicillium islandicum*") assembly genome with scaffold N50 of 2.35 Mb [13] and *Penicillium digitatum* assembly genome with scaffold N50 of 1.53 Mb [14]. *P. digitatum* has demonstrated an inability to utilize nitrogen due to the absence of the gene *crnA* or producing the metabolites patulin and penicillin based on the molecular level research [14]. Considering the diversity of active metabolites, interactions with the host, potential biomedical effects of *P. variabile*, and limited whole genome resources, we sequenced and assembled the first chromosome-level genome of *P. variabile* HXQ-H-1 in this study. The comprehensive, continuous genome could provide a new insight to the significant effects on the survival, growth, and fitness of mangroves and the biosynthesis mechanism of *P. variabile*.

2. Materials and Methods

2.1. Sampling and Isolation

The fungal strain HXQ-H-1 was isolated from the mangrove rhizosphere soil collected on the coast of Fujian Province (Fuzhou, China) and cultivated on modified martin solid plates with hygromycin 48 h, till forming the moderate clones. It was identified as *P. variabile* based on sequencing of the ITS region (GenBank no. KT429657) with 100% similarity to *P. variabile*. The strain was deposited at the Key Laboratory of Marine Drugs, the Ministry of Education of China, School of Medicine and Pharmacy, Ocean University of China, Qingdao, China.

2.2. DNA Extraction and WGS Library Construction

The fungi DNA was extracted using CTAB method and sheared into fragments between 100 and 500 bp in size by Covaris E220 ultrasonicator (Covaris, Brighton, UK). High-quality DNA were selected using AMPure XP beads (Agencourt, Beverly, MA, USA). After repairing using T4 DNA polymerase (Enzymatics, Beverly, MA, USA), the selected fragments were ligated at both ends to T-tailed adapters and amplified using KAPA HiFi HotStart ReadyMix (Kapa Biosystems, Wilmington, NC, USA). Then amplification products were subjected to a single-strand circularization process using T4 DNA Ligase (Enzymatics) and generated a single-stranded circular DNA library.

2.3. High-Through Chromosome Conformation Capture (Hi-C) Library Construction

Fresh fungi were cut into 10 mm fragments and resuspend in cold 1× phosphate buffered saline (PBS). Then 37% formaldehyde were added to fix the crosslinked to obtain 1% final concentration, and quenched the reaction by adding 2.5 M glycine solution to a final concentration of 0.2 M. The fungi were ground to fine powder in liquid nitrogen. Cross-linked nuclei isolated after centrifugation.

Then DNA was digested by the restriction enzyme (Mbo I) (NEB, Ipswich, MA, USA), followed by repairing the 5′ overhang using a biotinylated residue, and blunt-end fragments were ligated in situ using T4 DNA ligase. DNA was reverse-crosslinked and purified using a CTAB method. Finally, Hi-C library was created by shearing of DNA and capturing the biotin-containing fragments with streptavidin-coated beads using Dynabeads MyOne Streptavidin T1 (Invitrogen, Waltham, MA, USA). After DNA fragments end repairing, adenylation, and adaptor ligation, the fragments were amplified by PCR (8 cycles). A single-strand circularization process was carried out as WGS library described.

2.4. Genome Sequencing and Assembly

The WGS and Hi-C library were loaded and sequenced on the BGISEQ-500 platform [15]. Raw data are available in the CNGB Nucleotide Sequence Archive under the accession number CNP0000487. (CNSA: https://db.cngb.org/cnsa). The raw reads with high ratio of N (ambiguous base) and low-quality base were filtered out using SOAPnuke (v1.5.2) [16] with parameters "-l 20 -q 0.2 -n 0.05 -Q 2 -c 0 -d". Then the WGS clean data was assembled using SOAPdenovo (v2.0.4) [17] with parameters "-K 77 -d 1 -R". Subsequently, the Hi-C reads were mapped to the contigs of draft assembly using HiC-Pro (v2.8.0_devel) [18] with default parameters and proper mapped reads were extracted as the valid Hi-C reads. Then, the draft genome was anchored into chromosomes using 3D-DNA pipeline (v170123) [19] with parameters "-m haploid -s 4" according the interactive positions of Hi-C valid reads. BUSCO (v3.0.1) [20] was used to assess the confidence of the assembly with pezizomycotina_odb9. Simultaneously, we aligned the assembly sequences to the genome of *T. islandicus* WF-38-12 using lastz (v1.02.00) [21] with parameters "T = 2 C = 2 H = 2000 Y = 3400 L = 6000 K = 2200" and visualized using Circos (v0.69) [22].

2.5. Identification of Repetitive Elements and Non-Coding RNA Genes

Repetitive sequences were identified using multiple tools. TEs were identified by aligning against the Repbase [23] database using RepeatMasker (v4.0.5) [24] with parameters "-nolow -no_is -norna -engine wublast" and RepeatProteinMasker (v4.0.5) with parameters "-noLowSimple -pvalue 0.0001" at DNA and protein levels respectively. Meanwhile, de novo repeat library was detected using RepeatModeler (v1.0.8) and LTR-FINDER (v1.0.6) [25] with default parameters. Based on the de novo identified repeats, repeat elements were classified using RepeatMasker (v4.0.5) [24] with the same parameters. Furthermore, the tandem repeats were identified using Tandem Repeat Finder (v4.07) [26] with parameters "-Match 2 -Mismatch 7 -Delta 7 -PM 80 -PI 10 -Minscore 50 -MaxPeriod 2000".

For non-coding RNA (ncRNA), the tRNA genes were predicted using tRNAscan-SE (v1.3.1) [27] with default parameters. The rRNA fragments were identified using RNAmmer (v1.2). The snRNA and miRNA genes were predicted using CMsearch (v1.1.1) [28] with default parameters after aligning against the Rfam database [29] with blast (v2.2.30).

2.6. Gene Prediction and Function Annotation

For homolog gene prediction, 4 species protein sequences including *P. Expansum* (GCF_000769745.1), *P. rubens* (GCF_000226395.1), *T. marneffei* (GCF_000001985.1) and *T. stipitatus* (GCF_000003125.1) were download from NCBI for homology-based prediction. These protein sequences were aligned to *P. variabile* assembly sequences using BLAT (v36) [30] and for each protein 5 best matches were retained. Then, gene models were predicted using GeneWise [31] according the aligned loci with default parameters. For de novo prediction, Augustus (v2.4.1) [32] (v3.1) and Genemark (v4) [33] were used with parameters "–ES –fungus". Finally, all the predicted gene models were integrated using GLEAN [34] with parameters "–minlen 150 –minintron 11 –maxintron 10000" to obtain a consensus gene set. The gene set was assessed using BUSCO (v3.0.1) [20] with pezizomycotina_odb9 in protein mode.

The predicted genes were aligned to the KEGG [35], Swissprot [36], TrEMBL [37], COG [38], CAZy [39], NR and NT [40] database using Blast (v2.2.26) [41] with *E*-value $< 1 \times 10^{-5}$. Meanwhile, gene motifs and domains were identified against multiple protein databases including Pfam [42], SMART [43], PANTHER [44], PRINTS [45], ProSite [46] and ProDom [47] using InerProScan (5.16–55.0) [48], and Gene Ontology (GO) [49] terms were obtained based on the InterPro entries. The *P. variabile* assembly was uploaded to antiSMASH (v5.0) [50] website to identify the secondary metabolite gene cluster.

2.7. Core Genes, Homolog Genes, SNPs and Phylogeny Reconstruction

Gene sets of 33 *Penicillium* genomes published in NCBI, together with our strain were clustered using Cd-hit (v4.6.6) [51] with parameters "-c 0.5 -n 3 -p 1 -g 1 -d 0 -s 0.7 -aL 0.7 -aS 0.7" to generate a pan gene set, and the clusters present in all strains were extracted as a core gene set. To identify the homolog genes, all protein sequences were aligned to each other using blast+ (v2.2.31) and eliminated the redundancy by solar (v0.9.6). Clustering of the gene family was carried out by the alignment results with hcluster_sg (v0.5.1). Multiple sequences alignment of core genes and clustered gene families was performed by MUSCLE (v3.8.31) [52]. The SNPs among the 34 genomes were detected. The 33 downloaded genomes were aligned to the repeat marked reference genome HXQ-H-1 respectively using MUMmer (v3.23). The SNPs from each sample were aligned follow by reference locations. Phylogenetic trees were established by treebest (v1.9.2) with phyml [53] method and the number of bootstraps was 1000.

3. Results and Discussion

3.1. Genome Assembly

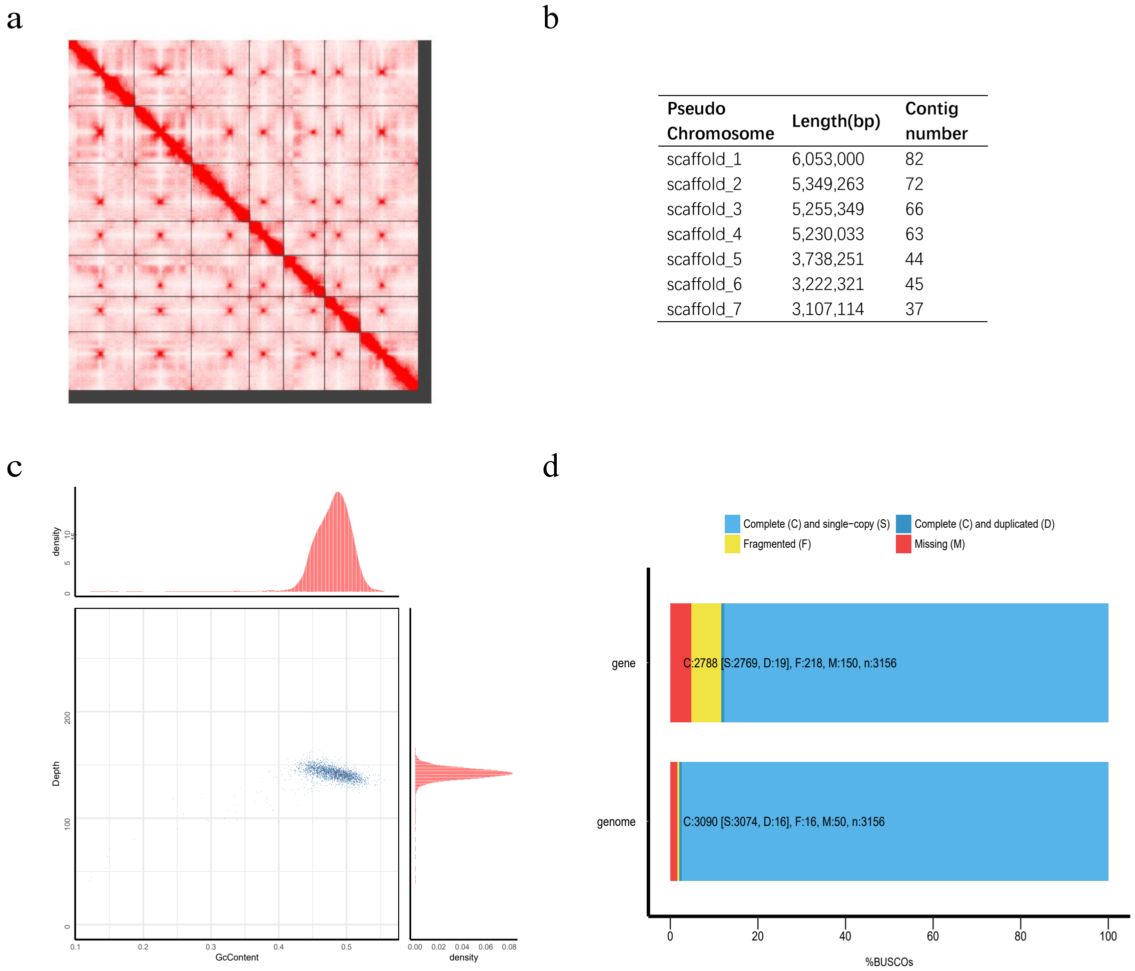
We obtained 20.24 Gb NGS clean data after quality control (Table S1). The genome size of *P. variabile* was estimated to be 35.56 Mb according to 17 bp k-mer frequency distribution (Figure S1). The draft genome was assembled with total length of 34.06 Mb, contig N50 of 94.65 kb (Table 1), closing to the estimated size. Hi-C sequencing produced about 112.76 Gb high quality data (Table S1, Figure S2a,b). We used 1.70 Gb valid data to anchor the draft assembly into seven chromosomes. The final chromosome-level assembly consists of 348 scaffolds with total length of 33.32 Mb, scaffold N50 of 5.23 Mb (Table 1). The longest seven chromosome-level scaffolds comprised 95.91% of the final assembly (Figure 1a,b).

Table 1. Statistic of genome assembly.

Values	Draft Genome		Chromosome Genome	
	Scaffold	Contig	Scaffold	Contig
Num	3860	3860	348	757
Length (bp)	34,059,084	34,059,084	33,318,899	33,114,399
N50 (bp)	93,482	93,482	5,232,202	96,738
N90 (bp)	24,393	24,393	2,951,000	30,747
GC%	46.58	47	48	48

To assess the quality of assembly, firstly WGS clean data were aligned back to the final assembly and a total of 97.93% reads were mapped. Based on the GC-depth distribution (Figure 1c), the GC content was concentrated in 40–55%, and the depth distribution was concentrated in 120–160, without GC or depth bias. Secondly, 3090 (97.9%) single-copy genes were found completely (Figure 1d) according to Benchmarking Universal Single-Copy Orthologs (BUSCO) assessment. The above analysis results indicating the reliability of our genome assembly and can support the downstream analysis.

Here, we assembled the high-quality genome of *P. variabile* into chromosomes using WGS sequences combined with Hi-C sequencing data. Though Hi-C technology has been frequently used in the research of various animal and plant genomes, and assisted assembly to obtain chromosome-level genomes and high-resolution chromatin three-dimensional structural information, it had not been reported in fungal genomes [54]. We constructed the fungal Hi-C library using the same experimental method as terrestrial plant Hi-C libraries [55]. It suggested that the Hi-C sequence method has a great boost to the auxiliary assembly of the fungi genome. With the decline in the cost of next generation sequencing, this strategy can be applied to other fungal species to improve the assembly qualities.



Pseudo Chromosome	Length(bp)	Contig number
scaffold_1	6,053,000	82
scaffold_2	5,349,263	72
scaffold_3	5,255,349	66
scaffold_4	5,230,033	63
scaffold_5	3,738,251	44
scaffold_6	3,222,321	45
scaffold_7	3,107,114	37

Figure 1. Summary of genome assembly and assessment. (**a**) The seven longest scaffolds interactive heatmap of *Penicillium variabile* assembly. The horizontal and vertical lines divide the genome into seven major pseudochromosomes. Gradient red indicates the intensity of contact between sequences. The intra-contact of sequences in pseudochromosome is stronger than the inter-contact. (**b**) The length of each chromosome and involved contig number. (**c**) GC-depth distribution of final assembly. The horizontal coordinate represents the average GC content per 10 kb windows, and the vertical coordinate represents the corresponding coverage depth. The bar charts on the top and right represent the distribution of GC and coverage depth. (**d**) BUSCO (benchmarking universal single-copy orthologs) assessment of genome and gene set. The bar chart below is a BUSCO assessment of genome, of which 3090 genes (97.91%) are predicted to be complete including 3074 single copies and 16 duplications. The upper bar chart is a BUSCO assessment of gene set, of which 2788 genes (88.34%) are predicted to be complete, including 2769 single copies and 19 duplications.

3.2. Genome Annotation

We combined the methods of de novo and Repbase to identify repetitive sequences in HXQ-H-1 genome. About 0.9819% (~327,160 bp) and 1.1729% (~390,774 bp) of the genome sequences were detected as repetitive sequences using de novo method and database respectively. In total, the repetitive sequences account for 1.88% (~627.53 kb), including 89.60 kb of DNA repeat elements (DNA), 52.67 kb of long interspersed nuclear elements (LINE), 2.88 kb of short interspersed nuclear elements (SINE), 93.08 kb of long terminal repeat elements (LTR) and 172.04 kb of tandem repeats (TRF) (Table S2).

Furthermore, we predicted 10,622 protein-coding genes in all, and there were 9927 *T. islandicus* WF-38-12 and 9000 *P. digitatum* [13,14]. For these 10,622 protein-coding genes, the average length of genes, coding sequences, exons, and introns were 1994.87 bp, 1568.29 bp, 473.94 bp, and 184.74 bp, respectively (Table 2). The gene set was assessed using BUSCO [20] with pezizomycotina_odb9 database, and 2788 (88.3%) out of 3156 single-copy orthologs can be found completely, indicating the high quality of gene set (Figure 1d).

Noncoding RNAs (ncRNAs) have complex regulatory functions at the transcriptional, translational and epigenetic level in fungi [56]. We predicted a total of 324 ncRNAs in HXQ-H-1, including 92 tRNAs, 3 rRNAs, 63 sRNAs, 50 snRNAs and 115 miRNAs (Table S3). Compared to the related species *T. islandicus*, the number of tRNAs was smaller (107 in WF-38-12) [13].

Table 2. Gene prediction results using homologous and de novo.

Values	Homolog				*denovo*	Glean
	P. expansum	*P. rubens*	*T. marneffei*	*T. stipitatus*	Augustus	
Gene_number	8947	9214	9510	9360	12,032	10,622
Average_gene_len (bp)	1456.62	1394.67	1459.34	1478.38	1773.88	1994.87
Average_cds_len (bp)	1317.95	1294.49	1349.26	1363.3	1581.44	1568.29
Average_exon_number	2.34	2.31	2.43	2.47	3.38	3.31
Average_exon_len (bp)	562.45	560.76	554.42	552.59	468.3	473.94
Average_intron_len (bp)	103.24	76.57	76.78	78.44	80.96	184.74

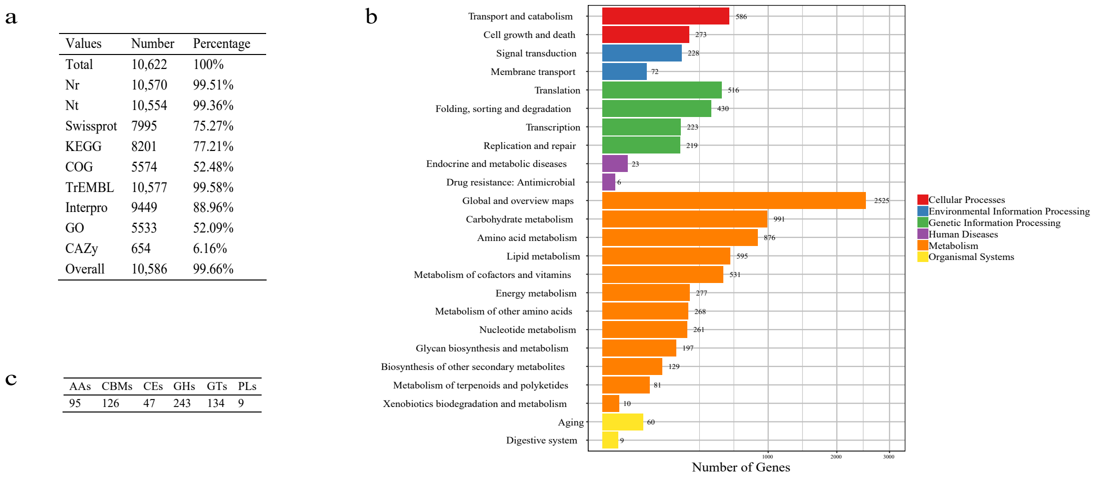
3.3. Functional Categorization

Of the 10,622 predicted genes, a total of 10,586 genes (99.66%) were functionally annotated by 9 databases (Figure 2a). In KEGG annotation result, 8201 genes (77.21%) were annotated and classified into 130 pathways belonging to 24 Level2 (Figure 2b). For COG functional annotation, 5574 genes (52.48%) were annotated to the 24 COG categories (Figure S3a). The most abundant cluster was related to the metabolism, including "amino acid transport and metabolism" and "carbohydrate transport and metabolism". Further, 5533 genes were classified according to the GO database, in which 8359, 8025 and 5679 genes were annotated to "cellular component", "biological process" and "molecular function", respectively. The "catalytic activity" was the largest annotated GO term (Figure S3b).

Furthermore, there were 654 (6.16%) genes were annotated by CAZymes. Among these genes, 243 genes encoding glycoside hydrolases (GHs) accounted for the largest proportion (2.29%). In addition, 134 genes were belonged to glycosyltransferases (GTs) (1.26%) (Figure 2c). Compared with the other 33 *Penicillium* genomes (Table S4), we found that HXQ-H-1 had the highest gene number in AA4, CBM35, CBM56, GH54, GH76, GT21, and PL7. Notably, CBM56 was annotated uniquely in HXQ-H-1 (Table S5), which may provide hints to understanding the energy flow and manglicolous fungus's lifestyle.

The secondary metabolism is an important characteristic of fungi and display a dichotomy within biological activities, some benefit the society by developing into medicines and agrochemicals, and some are pathogenic determinants of humans, animals and plants [57,58]. HXQ-H-1 secondary metabolites extractions such as nafuredin B showed cytotoxicity against a panel of human cancer cell lines, as well as varitatin A showed cytotoxicity against HCT-116 cell line (human colon cancer) and also inhibited the effects of protein tyrosine kinases [6,7]. To observe secondary metabolites biosynthetic gene clusters in HXQ-H-1, we further annotated the genome with antiSMASH pipeline [50]. A total of 77 gene clusters were detected (Table S6), including 29 polyketide synthase (PKS), 23 non-ribosomal peptide synthetase (NRPS), 10 mixed (NRPS-PKS, PKS-NRPS-NRPS like-terpene, T1PKS-NRPS like, T1PKS-terpene), eight terpene (Terpene), three indole cluster (indole), one β-lactone containing protease inhibitor (β-lactone), one Fungal RiPP with POP or UstH peptidase types and a modification (fungal-RiPP), one siderophore (siderophore), and one biosynthesis clusters (other). These results show more secondary metabolic clusters than those in *T.islandicus* [13] and the genes indole, betalactone, fungal-RiPP, and siderophore are absent in *T. islandicus*. It has been reported that most drug or drug-like metabolites are PKS, NRPS and terpene synthase and these enzymes have been extensively annotated in HXQ-H-1, reflecting that HXQ-H-1 has the large capacity of biotechnological potential as novel drugs [58].

Due to the special ecologic niche of mangroves, the ability of secondary metabolites of microorganisms is given to conquer harsh habitats. Most biosynthetic gene clusters of secondary metabolites are silent under standard laboratory culture conditions, new methods are needed to activate gene expression for metabolism exploration [59]. Our genomic analysis results here can provide guidance for gene regulation and facilitate the further studies of secondary metabolites in *P. variabile*.



a

Values	Number	Percentage
Total	10,622	100%
Nr	10,570	99.51%
Nt	10,554	99.36%
Swissprot	7995	75.27%
KEGG	8201	77.21%
COG	5574	52.48%
TrEMBL	10,577	99.58%
Interpro	9449	88.96%
GO	5533	52.09%
CAZy	654	6.16%
Overall	10,586	99.66%

c

AAs	CBMs	CEs	GHs	GTs	PLs
95	126	47	243	134	9

Figure 2. Function annotation of strain *P. variabile* HXQ-H-1 (**a**). Statistics of annotated genes by 9 databases. The third column indicates the ratio of the annotated genes number. (**b**). KEGG pathway classification. Histograms represent the gene number of pathways and grouped into 6 classifications which were tagged by colors. (**c**). Gene numbers of annotated by CAZyme.

3.4. Genome Comparison of *P. variabile* and *T.islandicus*

The genome collinearity between *P. variabile* HXQ-H-1 and *T. islandicus* WF-38-12 was highly consistent with 61.75% colinear ratio in HXQ-H-1. Our 7 chromosomes (scaffold 1–7) were primary matched to 13 chromosomes (CVMT01000001.1- CVMT01000013.1) of strain WF-38-12 (Figure 3). In detail, scaffold 1–6 corresponded to multiple scaffolds (scaffold1 vs. T01000003.1, T01000005.1, scaffold2 vs. T01000008.1, T01000001.1, scaffold3 vs. T01000013.1, T01000006.1, T01000004.1, T01000002.1, scaffold4 vs. T01000002.1, T01000012.1, scaffold5 vs. T01000002.1, T01000007.1, T01000011.1, scaffols6 vs. T01000010.1, T01000009.1). Scaffold 7 involved part of T01000001.1 (scaffold7 vs. T01000001.1). The GC content of HXQ-H-1 (47.59%) is similar to WF-38-12 (46.25%), and gene density are distributed without obvious bias along chromosomes (Figure 3). Through the COG annotation heatmap, we find that the percentage of HXQ-H-1 genes annotated by COG is relatively higher than WF-38-12 (Figure 3). Interestingly, there is a small difference in the total gene number (10,622 vs. 9927), whereas there is a relatively large difference in gene number with COG functional annotation (5574 vs. 5012). The main different functional categories were “signal transduction mechanisms”, “coenzyme transport and metabolism”, “energy production and conversion”, “lipid transport and metabolism”, and “secondary metabolites biosynthesis, transport, and catabolism”.

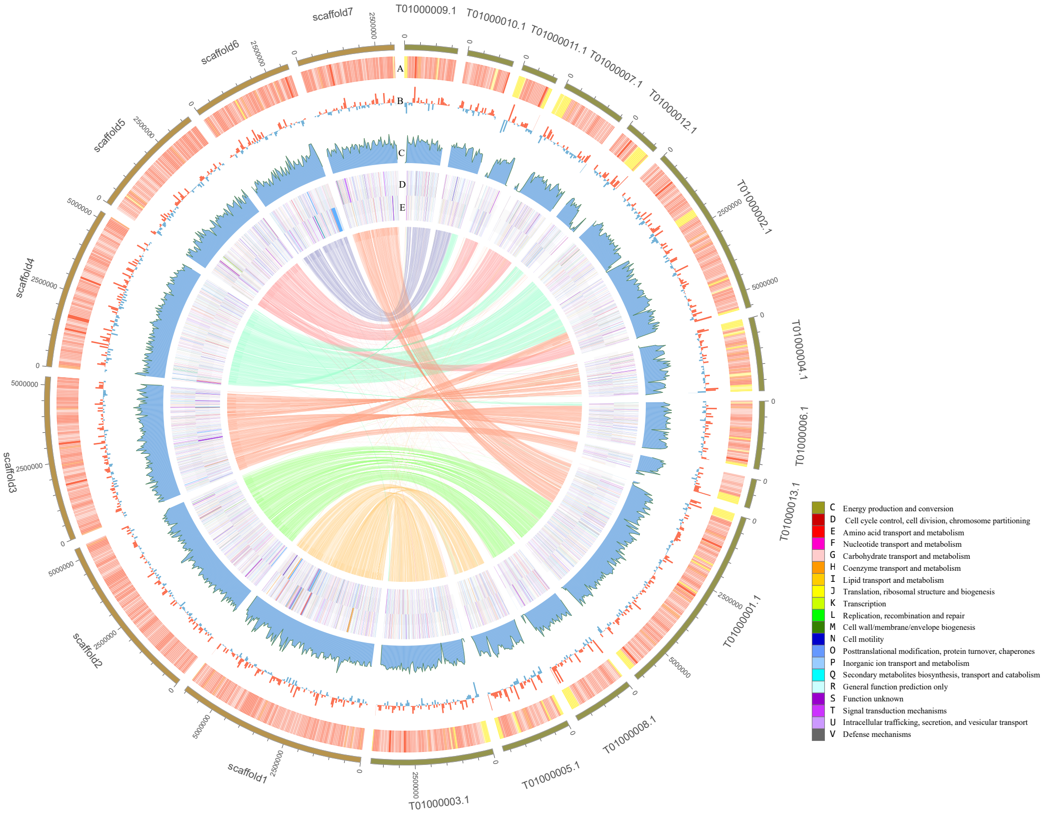


Figure 3. Genome comparison of *Penicillium variabile* HXQ-H-1 and *Talaromyces islandicus* WF-38-12. The lefts are the 7 scaffolds of HXQ-H-1, and the rights are the 13 scaffolds of WF-38-12. The collinear relationship between scaffolds is connected by center thin lines. The scale of each small square of the outermost circle represents 5 kb. Circle A is GC heatmap. Circle B is the histogram of GC (red: G > C; blue: G < C). Circle C is gene density in chromosomes. Circles D and E are the COG positive/negative annotation heatmaps, and the legend is shown in the bottom right.

3.5. Phylogenetic Analysis

To determine the HXQ-H-1 relationship among members of *Penicilliun*, we next reconstructed the species phylogenetic tree based on the whole genome of *P. variabile* HXQ-H-1 and other 33 *Penicillium* species (Table S4). We reconstructed trees using three kind of datasets using their core genes, homolog genes and SNPs, respectively. We obtained 582 core genes, 82 homolog genes and 3597 SNPs. All three phylogenetic trees are well-supported with boostrap values more than 95% (Figure 4). They are generally congruent and could be obviously divided into three clades. Previous studies reported the *Penicillium* was polyphyletic and clade2 were *Talaromyces* species described as a sexual state of *Penicillium* [60,61]. In our results, the *P. variabile* HXQ-H-1 is mostly close to the *T. islandicus* WF-38-12, consistent with previous phylogenetic tree construction using ribosomal DNA [62,63]. There are 48.18% (5118) genes in strain *P. variabile* HXQ-H-1 are orthologous with *T. islandicus* WF-38-12. It's the first time *Penicillium* phylogenetic tree reconstructed by genome-wide sequence data. We found that the traditional morphological classification of *Penicillium* species was confusing, leading to polyphyletic classification [60,61]. Biological classification needs to correctly reflect the phylogenetic relationship and evolutionary trends. With the development of molecular phylogenetics, the fungal classification has undergone tremendous transform. The reconstructed phylogeny three here may provide a reference for phylogenetic assessment or population genetics research.

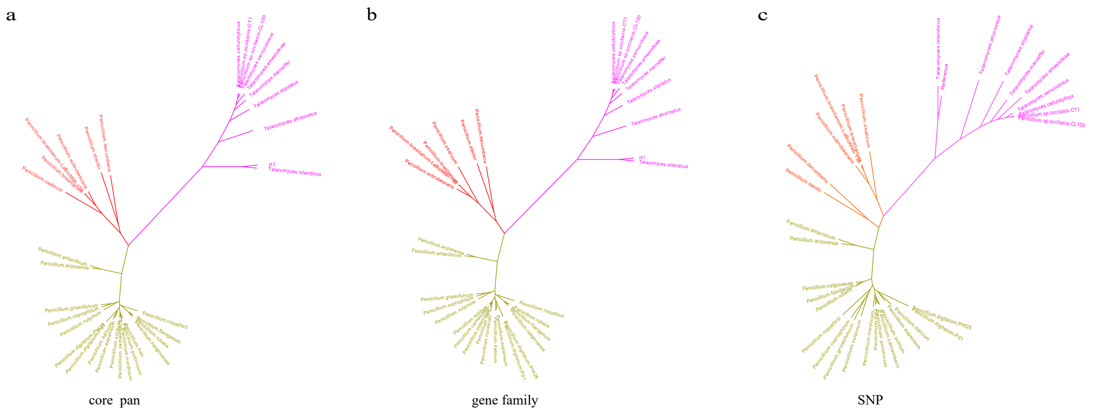


Figure 4. The species phylogenetic trees of *Penicillium variabile* HXQ-H-1 and other 33 *Penicillium* genomes. Trees were reconstructed based on core genes (**a**), homolog genes (**b**) and SNPs (**c**) respectively. The three clades are marked with different colors.

4. Conclusions

Mangrove is a special tropical coastal vegetation between the ocean and interior, an area home to a large number of marine fungi and terrestrial fungi. Manglicolous fungi plays an important role in mangrove ecosystems as a decomposer, and its metabolites not only help mangroves fight disease, but also help mankind to exploit new drugs. However, the research on mangrove fungi is not extensive. According a white paper, only 339 manglicolous fungal species have been identified and 46 marine (contain mangrove) fungi genome assembled [64]. As far as we know, this is the first *Penicillium* genome in mangrove ecosystems. Based on this assembly, we identified abundant repetitive sequences, protein-coding genes, and ncRNAs. According to different evaluation results, our high-quality genome assembly and accurate protein-coding genes prediction were also confirmed. We performed gene functional annotation including secondary metabolism, which reveals the large capacity, biotechnological potential of *P. variabile* HXQ-H-1 based on molecular biology. We created out genome-wide species phylogeny trees of *Penicillium* for the first time. All the resources provided here are convenient and helpful for comparative genomic analysis, phylogenetic analysis, and metabolic regulation analysis in the future.

Supplementary Materials: The following are available online at http://www.mdpi.com/2309-608X/6/1/7/s1, Figure S1: 17-mers frequency distribution of HXQ-H-1, Figure S2: Ratio of different types of reads qualified by HiC-Pro, Figure S3: Functional Categorization annotated by COG and GO database, Table S1: Statistic of sequencing data and quality control, Table S2: Types, length and proportion of repetitive sequences, Table S3: HXQ-H-1 ncRNA prediction result, Table S4: The 33 Penicillium genomes downloaded list, Table S5: CAZy annotation of 33 Penicillium genomes together with HXQ-H-1, Table S6: Secondary metabolites clusters annotated by antiSMASH pipeline in HXQ-H-1.

Author Contributions: Conceptualization, D.L., X.L. and G.F.; methodology, L.L. and C.S.; formal analysis, L.L.; Experimental work, Q.X. and W.G.; writing—original draft preparation, L.P. and X.C.L.; writing—review and editing, L.P., J.C. and C.S.; supervision, G.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China (No. 2016YFE0122000).

Acknowledgments: We thank Shandong Technology Innovation Center of Synthetic Biology for their advices with this work.

Conflicts of Interest: The authors declare no conflict of interest. BGI did not provide fund supporting for this study. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Li, C.S.; Li, X.M.; Gao, S.S.; Lu, Y.H.; Wang, B.G. Cytotoxic anthranilic acid derivatives from deep sea sediment-derived fungus penicillium paneum sd-44. *Mar. Drugs* **2013**, *11*, 3068–3076. [CrossRef] [PubMed]
2. Meng, L.H.; Zhang, P.; Li, X.M.; Wang, B.G. Penicibrocazines a–e, five new sulfide diketopiperazines from the marine-derived endophytic funguspenicillium brocae. *Mar. Drugs* **2015**, *13*, 276–287. [CrossRef] [PubMed]
3. Liu, S.; Su, M.; Song, S.J.; Jung, J.H. Marine-derived penicillium species as producers of cytotoxic metabolites. *Mar. Drugs* **2017**, *15*, 329. [CrossRef] [PubMed]
4. Latha, R.; Mitra, S.J.C.S. Mangrove fungi in india. *Curr. Sci.* **2004**, *86*, 1586.
5. Hrudayanath, T.; Bikash, C.B.; Rashmi, R.M. Ecological role and biotechnological potential of mangrove fungi: A review. *Mycobiology* **2013**, *4*, 54–71.
6. Zhang, Z.; He, X.; Zhang, G.; Che, Q.; Zhu, T.; Gu, Q.; Li, D. Inducing secondary metabolite production by combined culture of talaromyces aculeatus and penicillium variabile. *J. Nat. Prod.* **2017**, *80*, 3167–3171. [CrossRef]
7. He, X.; Zhang, Z.; Chen, Y.; Che, Q.; Zhu, T.; Gu, Q.; Li, D. Varitatin a, a highly modified fatty acid amide from penicillium variabile cultured with a DNA methyltransferase inhibitor. *J. Nat. Prod.* **2015**, *78*, 2841–2845. [CrossRef]
8. Zhelifonova, V.P.; Antipova, T.V.; Ozerskaya, S.M.; Ivanushkina, N.E.; Kozlovskii, A.G. The fungus penicillium variabile sopp 1912 isolated from permafrost deposits as a producer of rugulovasines. *Microbiology* **2006**, *75*, 644–648. [CrossRef]
9. Dorner, J.; Cole, R.; Hill, R.; Wicklow, D.; Cox, R.H. Penicillium rubrum and penicillium biforme, new sources of rugulovasines a and b. *Appl. Environ. Microbiol.* **1980**, *40*, 685–687.
10. Meurant, G. *Handbook of Toxic Fungal Metabolites*; Elsevier: Amsterdam, The Netherlands, 2012.
11. Gu, X.-L.; Tian, Y.P. Isolation and characterization of urethanase from penicillium variabile and its application to reduce ethyl carbamate contamination in chinese rice wine. *Appl. Biochem. Biotechnol.* **2013**, *170*, 718–728.
12. Petruccioli, M.; Federici, F.; Bucke, C.; Keshavarz, T. Enhancement of glucose oxidase production by penicillium variabile p16. *Enzym. Microb. Technol.* **1999**, *24*, 397–401. [CrossRef]
13. Schafhauser, T.; Wibberg, D.; Ruckert, C.; Winkler, A.; Flor, L.; van Pee, K.H.; Fewer, D.P.; Sivonen, K.; Jahn, L.; Ludwig-Muller, J.; et al. Draft genome sequence of talaromyces islandicus ("penicillium islandicum") wf-38-12, a neglected mold with significant biotechnological potential. *J. Biotechnol.* **2015**, *211*, 101–102. [CrossRef] [PubMed]
14. Marcet-Houben, M.; Ballester, A.R.; de la Fuente, B.; Harries, E.; Marcos, J.F.; Gonzalez-Candelas, L.; Gabaldon, T. Genome sequence of the necrotrophic fungus penicillium digitatum, the main postharvest pathogen of citrus. *BMC Genom.* **2012**, *13*, 646. [CrossRef] [PubMed]
15. Drmanac, R.; Sparks, A.B.; Callow, M.J.; Halpern, A.L.; Burns, N.L.; Kermani, B.G.; Carnevali, P.; Nazarenko, I.; Nilsen, G.B.; Yeung, G.; et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **2010**, *327*, 78–81. [CrossRef] [PubMed]
16. Chen, Y.; Chen, Y.; Shi, C.; Huang, Z.; Zhang, Y.; Li, S.; Li, Y.; Ye, J.; Yu, C.; Li, Z.; et al. Soapnuke: A mapreduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *Gigascience* **2018**, *7*, 1–6. [CrossRef]
17. Luo, R.; Liu, B.; Xie, Y.; Li, Z.; Huang, W.; Yuan, J.; He, G.; Chen, Y.; Pan, Q.; Liu, Y.; et al. Soapdenovo2: An empirically improved memory-efficient short-read de novo assembler. *Gigascience* **2012**, *1*, 18. [CrossRef]
18. Servant, N.; Varoquaux, N.; Lajoie, B.R.; Viara, E.; Chen, C.J.; Vert, J.P.; Heard, E.; Dekker, J.; Barillot, E. Hic-pro: An optimized and flexible pipeline for hi-c data processing. *Genome Biol.* **2015**, *16*, 259. [CrossRef]
19. Dudchenko, O.; Batra, S.S.; Omer, A.D.; Nyquist, S.K.; Hoeger, M.; Durand, N.C.; Shamim, M.S.; Machol, I.; Lander, E.S.; Aiden, A.P. , et al. De novo assembly of the aedes aegypti genome using hi-c yields chromosome-length scaffolds. *Science* **2017**, *356*, 92–95. [CrossRef]
20. Simao, 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* **2015**, *31*, 3210–3212. [CrossRef]
21. Harris, R. Improved Pairwise Alignment of Genomic DNA: The Pennsylvania State University. Ph.D. Thesis, The Pennsylvania State University, State College, PA, USA, 2007.

22. Krzywinski, M.; Schein, J.; Birol, I.; Connors, J.; Gascoyne, R.; Horsman, D.; Jones, S.J.; Marra, M.A. Circos: An information aesthetic for comparative genomics. *Genome Res.* **2009**, *19*, 1639–1645. [CrossRef]
23. Bao, W.; Kojima, K.K.; Kohany, O. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* **2015**, *6*, 11. [CrossRef] [PubMed]
24. Tarailo-Graovac, M.; Chen, N. Using repeatmasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* **2009**, *25*, 4.10.1–4.10.14. [CrossRef] [PubMed]
25. Xu, Z.; Wang, H. Ltr_finder: An efficient tool for the prediction of full-length ltr retrotransposons. *Nucleic Acids Res.* **2007**, *35*, W265–W268. [CrossRef] [PubMed]
26. Benson, G. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res* **1999**, *27*, 573–580. [CrossRef] [PubMed]
27. Lowe, T.M.; Eddy, S.R. Trnascan-se: A program for improved detection of transfer rna genes in genomic sequence. *Nucleic Acids Res.* **1997**, *25*, 955–964. [CrossRef]
28. Cui, X.; Lu, Z.; Wang, S.; Jing-Yan Wang, J.; Gao, X. Cmsearch: Simultaneous exploration of protein sequence space and structure space improves not only protein homology detection but also protein structure prediction. *Bioinformatics* **2016**, *32*, i332–i340. [CrossRef]
29. Kalvari, I.; Nawrocki, E.P.; Argasinska, J.; Quinones-Olvera, N.; Finn, R.D.; Bateman, A.; Petrov, A.I. Non-coding rna analysis using the rfam database. *Curr. Protoc. Bioinform.* **2018**, *62*, e51. [CrossRef]
30. Kent, W.J. Blat–the blast-like alignment tool. *Genome Res.* **2002**, *12*, 656–664. [CrossRef]
31. Birney, E.; Clamp, M.; Durbin, R. Genewise and genomewise. *Genome Res.* **2004**, *14*, 988–995. [CrossRef]
32. Stanke, M.; Keller, O.; Gunduz, I.; Hayes, A.; Waack, S.; Morgenstern, B. Augustus: Ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **2006**, *34*, W435–W439. [CrossRef]
33. Besemer, J.; Lomsadze, A.; Borodovsky, M. Genemarks: A self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.* **2001**, *29*, 2607–2618. [CrossRef] [PubMed]
34. Elsik, C.G.; Mackey, A.J.; Reese, J.T.; Milshina, N.V.; Roos, D.S.; Weinstock, G.M. Creating a honey bee consensus gene set. *Genome Biol.* **2007**, *8*, R13. [CrossRef] [PubMed]
35. Ogata, H.; Goto, S.; Sato, K.; Fujibuchi, W.; Bono, H.; Kanehisa, M. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **1999**, *27*, 29–34. [CrossRef] [PubMed]
36. UniProt, C. Reorganizing the protein space at the universal protein resource (uniprot). *Nucleic Acids Res.* **2012**, *40*, D71–D75.
37. Bairoch, A.; Apweiler, R. The swiss-prot protein sequence database and its supplement trembl in 2000. *Nucleic Acids Res.* **2000**, *28*, 45–48. [CrossRef] [PubMed]
38. Tatusov, R.L.; Fedorova, N.D.; Jackson, J.D.; Jacobs, A.R.; Kiryutin, B.; Koonin, E.V.; Krylov, D.M.; Mazumder, R.; Mekhedov, S.L.; Nikolskaya, A.N.; et al. The cog database: An updated version includes eukaryotes. *BMC Bioinform.* **2003**, *4*, 41. [CrossRef]
39. Cantarel, B.L.; Coutinho, P.M.; Rancurel, C.; Bernard, T.; Lombard, V.; Henrissat, B. The carbohydrate-active enzymes database (cazy): An expert resource for glycogenomics. *Nucleic Acids Res.* **2009**, *37*, D233–D238. [CrossRef]
40. Pruitt, K.D.; Tatusova, T.; Maglott, D.R. Ncbi reference sequence (refseq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **2005**, *33*, D501–D504. [CrossRef]
41. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]
42. Finn, R.D.; Bateman, A.; Clements, J.; Coggill, P.; Eberhardt, R.Y.; Eddy, S.R.; Heger, A.; Hetherington, K.; Holm, L.; Mistry, J.; et al. Pfam: The protein families database. *Nucleic Acids Res.* **2014**, *42*, D222–D230. [CrossRef]
43. Ponting, C.P.; Schultz, J.; Milpetz, F.; Bork, P. Smart: Identification and annotation of domains from signalling and extracellular protein sequences. *Nucleic Acids Res.* **1999**, *27*, 229–232. [CrossRef] [PubMed]
44. Mi, H.; Lazareva-Ulitsky, B.; Loo, R.; Kejariwal, A.; Vandergriff, J.; Rabkin, S.; Guo, N.; Muruganujan, A.; Doremieux, O.; Campbell, M.J.; et al. The panther database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res.* **2005**, *33*, D284–D288. [CrossRef] [PubMed]
45. Attwood, T.K. The prints database: A resource for identification of protein families. *Brief. Bioinform.* **2002**, *3*, 252–263. [CrossRef] [PubMed]

46. Hulo, N.; Bairoch, A.; Bulliard, V.; Cerutti, L.; De Castro, E.; Langendijk-Genevaux, P.S.; Pagni, M.; Sigrist, C.J. The prosite database. *Nucleic Acids Res.* **2006**, *34*, D227–D230. [CrossRef]
47. Bru, C.; Courcelle, E.; Carrere, S.; Beausse, Y.; Dalmar, S.; Kahn, D. The prodom database of protein domain families: More emphasis on 3d. *Nucleic Acids Res.* **2005**, *33*, D212–D215. [CrossRef]
48. Zdobnov, E.M.; Apweiler, R. Interproscan–an integration platform for the signature-recognition methods in interpro. *Bioinformatics* **2001**, *17*, 847–848. [CrossRef]
49. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; et al. Gene ontology: Tool for the unification of biology. The gene ontology consortium. *Nat. Genet.* **2000**, *25*, 25–29. [CrossRef]
50. Medema, M.H.; Blin, K.; Cimermancic, P.; de Jager, V.; Zakrzewski, P.; Fischbach, M.A.; Weber, T.; Takano, E.; Breitling, R. Antismash: Rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res.* **2011**, *39*, W339–W346. [CrossRef]
51. Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. Cd-hit: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **2012**, *28*, 3150–3152. [CrossRef]
52. Edgar, R.C. Muscle: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **2004**, *32*, 1792–1797. [CrossRef]
53. Guindon, S.; Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **2003**, *52*, 696–704. [CrossRef] [PubMed]
54. Schmitt, A.D.; Hu, M.; Ren, B. Genome-wide mapping and analysis of chromosome architecture. *Nat. Rev. Mol. Cell Biol.* **2016**, *17*, 743–755. [CrossRef] [PubMed]
55. Mascher, M.; Gundlach, H.; Himmelbach, A.; Beier, S.; Twardziok, S.O.; Wicker, T.; Radchuk, V.; Dockter, C.; Hedley, P.E.; Russell, J.; et al. A chromosome conformation capture ordered sequence of the barley genome. *Nature* **2017**, *544*, 427–433. [CrossRef] [PubMed]
56. Chacko, N.; Lin, X. Non-coding rnas in the development and pathogenesis of eukaryotic microbes. *Appl. Microbiol. Biotechnol.* **2013**, *97*, 7989–7997. [CrossRef]
57. Keller, N.P.; Bennett, J.; Turner, G. Secondary metabolism: Then, now and tomorrow. *Fungal Genet. Biol.* **2011**, *48*, 1–3. [CrossRef]
58. Bills, G.F.; Gloer, J.B. Biologically active secondary metabolites from the fungi. *Microbiol. Spectr.* **2016**, *4*. [CrossRef]
59. Ancheeva, E.; Daletos, G.; Proksch, P. Lead compounds from mangrove-associated microorganisms. *Mar. Drugs* **2018**, *16*, 319. [CrossRef]
60. Yilmaz, N.; Visagie, C.M.; Houbraken, J.; Frisvad, J.C.; Samson, R.A. Polyphasic taxonomy of the genus talaromyces. *Stud. Mycol.* **2014**, *78*, 175–341. [CrossRef]
61. Chen, A.J.; Sun, B.D.; Houbraken, J.; Frisvad, J.C.; Yilmaz, N.; Zhou, Y.G.; Samson, R.A. New talaromyces species from indoor environments in china. *Stud. Mycol.* **2016**, *84*, 119–144. [CrossRef]
62. Dupont, J.; Dennetière, B.; Jacquet, C.; Roquebert, M.-F. Pcr-rflp of its rdna for the rapid identification of penicillium subgenus biverticillium species. *Rev. Iberoam. Micol.* **2006**, *23*, 145–150. [CrossRef]
63. Prat, C.; Ruiz-Rueda, O.; Trias, R.; Antico, E.; Capone, D.; Sefton, M.; Baneras, L. Molecular fingerprinting by pcr-denaturing gradient gel electrophoresis reveals differences in the levels of microbial diversity for musty-earthy tainted corks. *Appl. Environ. Microbiol.* **2009**, *75*, 1922–1931. [CrossRef] [PubMed]
64. Fan, G.; Chen, J.; Jin, T.; Shi, C.; Du, X.; Zhang, H.; Zhang, Y.; Li, H.; Luo, T.; Yan, P.; et al. The report of marine life genomic research. *Preprints* **2018**. [CrossRef]