RESEARCH ARTICLE

FunOrder: A robust and semi-automated method for the identification of essential biosynthetic genes through computational molecular co-evolution

Gabriel A. Vignolle, Denise Schaffer, Leopold Zehetner, Robert L. Mach, Astrid R. Mach-Aigner, Christian Derntl*

Institute of Chemical, Environmental and Bioscience Engineering, TU Wien, Vienna, Austria

* christian.derntl@tuwien.ac.at

Citation: Vignolle GA, Schaffer D, Zehetner L, Mach RL, Mach-Aigner AR, Derntl C (2021) FunOrder: A robust and semi-automated method for the identification of essential biosynthetic genes through computational molecular co-evolution. PLoS Comput Biol 17(9): e1009372. https://doi.org/10.1371/journal.pcbi.1009372

Editor: Samuel V. Scarpino, Northeastern University, UNITED STATES

Received: April 13, 2021

Accepted: August 23, 2021

Published: September 27, 2021

Data Availability Statement: The FunOrder tool, the relevant database, and the sequences and the FunOrder output of the negative control GCs and the positive control BGCs are available in the GitHub repository (https://github.com/gvignolle/FunOrder). We have also used Zenodo to assign a DOI to the repository: 10.5281/zenodo.5118984.

Funding: This study was supported by the Austrian Science Fund (FWF, https://www.fwf.ac.at/) [P 29556 to RM, P 34036 to CD] and TU Wien

Abstract

Secondary metabolites (SMs) are a vast group of compounds with different structures and properties that have been utilized as drugs, food additives, dyes, and as monomers for novel plastics. In many cases, the biosynthesis of SMs is catalysed by enzymes whose corresponding genes are co-localized in the genome in biosynthetic gene clusters (BGCs). Notably, BGCs may contain so-called gap genes, that are not involved in the biosynthesis of the SM. Current genome mining tools can identify BGCs, but they have problems with distinguishing essential genes from gap genes. This can and must be done by expensive, laborious, and time-consuming comparative genomic approaches or transcriptome analyses. In this study, we developed a method that allows semi-automated identification of essential genes in a BGC based on co-evolution analysis. To this end, the protein sequences of a BGC are blasted against a suitable proteome database. For each protein, a phylogenetic tree is created. The trees are compared by treeKO to detect co-evolution. The results of this comparison are visualized in different output formats, which are compared visually. Our results suggest that co-evolution is commonly occurring within BGCs, albeit not all, and that especially those genes that encode for enzymes of the biosynthetic pathway are co-evolutionary linked and can be identified with FunOrder. In light of the growing number of genomic data available, this will contribute to the studies of BGCs in native hosts and facilitate heterologous expression in other organisms with the aim of the discovery of novel SMs.

Author summary

The discovery and description of novel fungal secondary metabolites promises novel antibiotics, pharmaceuticals, and other useful compounds. A way to identify novel secondary metabolites is to express the corresponding genes in a suitable expression host. Consequently, a detailed knowledge or an accurate prediction of these genes is necessary. In fungi, the genes are co-localized in so-called biosynthetic gene clusters. Notably, the clusters may also contain genes that are not necessary for the biosynthesis of the secondary

(https://www.tuwien.at/) [PhD program TU Wien bioactive]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

metabolites, so-called gap genes. We developed a method to detect co-evolved genes within the clusters and demonstrated that essential genes are co-evolving and can thus be differentiated from the gap genes. This adds an additional layer of information, which can support researchers with their decisions on which genes to study and express for the discovery of novel secondary metabolites.

This is a *PLOS Computational Biology* Methods paper.

Introduction

Secondary metabolites (SMs) are a diverse group of compounds with a plethora of different chemical structures and properties which are found in all domains of life, but are predominantly studied in bacteria, fungi, and plants [1]. SMs are not necessary for the basic survival and growth of an organism but can be beneficial under certain conditions. For example, pigments help to sustain radiation, antibiotics help in competitive situations, and toxins can serve as defensive compounds or as virulence factors [2,3]. Notably, many SMs are used by humankind as drugs and pharmaceuticals, pigments and dyes, sweeteners and flavours, and most recently also as precursors for the synthesis of plastics [4]. The study of the secondary metabolism holds the promise for novel antibiotics, pharmaceuticals and other useful compounds [5].

A major hinderance in the discovery of yet undescribed SMs is the fact that most SMs are not produced under standard laboratory conditions, as they do not serve a purpose for the organisms then. Currently, different strategies are followed to circumvent this problem [6,7]. Untargeted approaches aim to induce the expression of any SM. To this end, biotic and abiotic stresses are applied, or global regulators and regulatory mechanisms are manipulated [8]. These strategies may lead to the discovery of novel compounds, whose corresponding genes have to be identified subsequently by time-consuming and expensive methods [7]. An extreme example are the aflatoxins, major food contaminants with serious toxicological effects [9]. It took over 40 years from the discovery of the aflatoxins as the causal agent of "turkey X" disease in the 1950s [10] until the corresponding genes were finally described in 1995 [11]. Targeted SM discovery approaches aim to induce the production of specific SMs by either overexpressing genes in the native host or by heterologous expression in another organism [12]. The targeted approaches, also called reverse strategy or bottom-up strategy allows a direct connection of SMs to the corresponding genes and does not rely on the inducibility of SM production in the native host. Inherently, the bottom-up approach is depending on modern genomics and accurate gene prediction tools [13].

In bacteria and fungi, the genes responsible for the biosynthesis of a certain SM are often co-localized in the genome, forming so called biosynthetic gene clusters (BGCs) [14,15]. The BGCs consists of one or more core genes, several tailoring enzymes, and genes involved in regulation and transport. As all these genes are essential for the production of a SM in the native host, we will refer to them as "essential genes" in this study. The core genes are responsible for assembling the basic chemical scaffold, which is further modified by the tailoring enzymes yielding the final SM [16]. We refer to the core genes and the tailoring genes as "biosynthetic genes" in this study. Depending on the class of the produced SM, the core genes differ. In fungi, the main SM classes are polyketides (e.g. the cholesterol-lowering drug lovastatin [17]

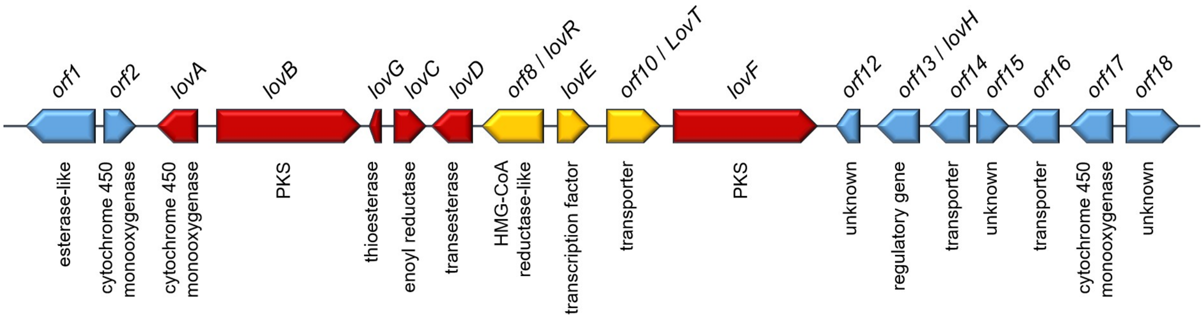

Fig 1. Schematic representation of the lovastatin BGC from *Aspergillus terreus* (lov). In red the biosynthetic genes for SM production, in gold the further essential genes, and in blue the genes not involved in the biosynthetic pathway.

https://doi.org/10.1371/journal.pcbi.1009372.g001

and the mycotoxin aflatoxin [9]) and non-ribosomal peptides (e.g. the immunosuppressant cyclosporine [18] and the antibiotic penicillin [19]), with polyketide synthases (PKS) or non-ribosomal peptide synthetases (NRPS) as core enzymes, respectively. Other SM classes are terpenoids, alkaloids, melanins [20,21], and ribosomally synthesized and posttranslationally modified peptides (RiPPs) [22,23], whose corresponding genes may also be organized in BGCs. As mentioned, BGCs may also contain genes encoding for transporters [24], transcription factors [25], or resistance genes [26]. While their gene products are not directly involved in the biosynthesis of a SM they are still essential for the biosynthesis; we will call them „further essential genes"in the following and differentiate them from the „biosynthetic genes". The biosynthetic genes and the further essential genes are both necessary for the biosynthesis of a SM in the native organisms. In contrast, only the biosynthetic genes and a selection of the further essential genes (e.g. transporters) are necessary for heterologous expression [reviewed in [27]]. Notably, fungal BGCs often also contain genes that are not necessary for the production of a SM, the so-called gap genes. The gap genes are not involved in the biosynthesis, regulation, or transport of the SM, but have an unrelated function (Fig 1). We would like to stress here, that this cannot be predicted based only on the class of the gene product. For instance, a gene encoding for a transporter in the aflatoxin BGC was reported to have no significant role in aflatoxin secretion [28].

As mentioned, the bottom-up approach for SM discovery is depending on modern genomics and the accurate prediction of genes and BGCs. Each important gene missing in the prediction is detrimental for obvious reasons, whereas each unnecessarily considered gap gene makes the study of a BGC more complicated and complex, and the construction and transformation processes for heterologous expression more challenging. Currently, several BGC prediction tools are available for fungi. Some tools for genome mining are antiSMASH [29], CASSIS and SMIPS [30], SMURF [31], TOUCAN, a supervised learning framework capable of predicting BGCs on amino acid sequences [32], and DeepBGC, an unrestricted machine learning approach using deep neural networks [33]. These tools are effective and successful in finding and predicting BGCs based solely on genomic data. AntiSMASH uses a rule-based approach to identify BGCs based on the identification of core or signature enzymes and applies a greedy approach to extend a cluster on either side. This may result in overlaps or combinations of closely situated clusters. However, the genes within the predicted BGCs are classified into core biosynthetic genes, additional biosynthetic genes, transport-related genes, regulatory genes, and other genes based on profile hidden Markov models by the antiSMASH tool. The BGC prediction method of CASSIS and SMIPS is based on the principle that the promoter regions of genes in a BGC contain one or more shared motif, as they are co-expressed and presumably regulated by the same regulatory factors and/or mechanisms [30].

As mentioned above, the class of an enzyme may be a good indication for a potential involvement in the biosynthesis of a SM but does not guarantee a correct prediction. This problem can be solved by the analysis of transcriptome data because the genes necessary for SM production within a BGC are normally co-expressed with each other but not with the gap genes [34]. Notably, this demands the knowledge of expression conditions and does not work for silent BGCs. However, it is an obvious advantage to have as much information as possible about a BGC before studying it in the native host or performing heterologous expression for a bottom-up approach for SM discovery.

We speculate that a comparative genomics analysis focusing on the evolutionary history of the genes in a BGC might be a feasible alternative to a transcriptomics analysis in fungi for the following reasons. In general, BGCs are suggested to undergo a distinct and faster evolution than the rest of the genome, based on different mechanisms and genetic drivers [16,35–40]. In bacteria, the evolution of BGCs is strongly influenced by the strong occurrence of horizontal gene transfer in these group of microorganism [39]. Medema et al. performed a large-scale computational analysis of bacterial BGCs and found that many BGCs consist of sub-clusters. These sub-clusters encode for enzymes that work together to form a distinct chemical structure. Notably, this sub-clusters were described as "independent evolutionary entities" and the contained genes are co-evolving. The authors suggested a "bricks and mortar" model. Therein, different sub-clusters, the "bricks" form different chemical building blocks for a secondary metabolite. Additional genes within the BGCs are encoding for enzymes that combine the building blocks, and fulfil other functions such as tailoring, regulation and transport. These individual genes are the "mortar" in the "brick and mortar" model [40]. The "bricks" correspond to what we term "biosynthetic genes" and the "mortar" to our "further essential genes". Through horizontal gene transfer, the "bricks" can be easily exchanged and recombined to form novel BGCs and secondary metabolites[40]. Notably, not all bacterial BGCs are composed of exchangeable sub-units but some BGCs keep a stable architecture over a long time [40].

In fungi, three molecular evolutionary processes were suggested to be responsible for shaping the BGCs in a recent study, i.e., functional divergence, horizontal gene transfer, and *de novo* assembly [41]. Rokas et al. define functional divergence as a "process by which homologous BGCs, through the accumulation of genetic changes, gradually diverge in their functions changes" [41] and horizontal gene transfer as a "process by which an entire BGC from the genome of one organism is transferred and stably integrated into the genome of another through non-reproduction related mechanisms" [41]. This implies in both cases, that fungal BGCs are staying intact. Further, the genes are suggested to undergo a co-evolution which is faster than the rest of the genome [41]. Medema's "brick and mortar" model would more or less correspond to what Rokas et al. describe as "*de novo* assembly". This is defined as a "process by which an entire BGC is evolutionarily assembled through the recruitment and relocation of native genes, duplicates of native genes, and horizontally acquired genes" [41]. Notably, Rokas et al. state that this is the"least well-documented evolutionary process involved in the generation of fungal chemodiversity" [41], suggesting that in known and described fungal BGCs functional divergence and horizontal gene transfer are the two main evolutionary process, during which BGCs are staying intact and genes undergo a similar evolution. Further, we hypothesize that especially the biosynthetic genes in a BGC are co-evolutionary linked by the selection pressure to keep the biosynthetic pathway intact. Notably, a co-evolution analysis is a laborious and time-consuming task because a phylogenetic tree has to be calculated for each gene and then the trees compared to each other manually [42]. Recently, a method for the detection of co-evolution in bacterial BGCs was developed with the aim to identify sub-clusters [43]. That method is based on the detection of orthologous genes that are present in close vicinity in many BGCs. This method is working unsupervised but requires a large set of BGCs as input [43].

In this study we describe a method (FunOrder) that allows a fast, semi-automated co-evolution analysis using individual BGCs as input. Based on this analysis and the assumption that the essential genes undergo a shared or similar evolution, FunOrder aims to identify essential genes in BGCs. To this end, we constructed a database of fungal proteomes as basis for the identification of co-evolutionary linked genes in ascomycetes. We determine the thresholds for the detection of co-evolution within different control gene sets. Then, we evaluated FunOrder and tested the underlying hypothesis, whether essential genes within a BGC could be identified based on the principle of co-evolution. We demonstrated the robustness and the applicability of the FunOrder method by analysing different control gene sets, including empirically validated BGCs and evaluated our method using stringent statistical tests.

Material and methods

Construction of a fungal proteome database

In this study we aim to identify co-evolutionary linked genes in ascomycetes. As the basis for the detection of co-evolution is a suitable database [42], we compiled an empirically optimized database consisting of 134 fungal proteomes from mainly ascomycetes and from two basidiomycetes for this method (Table 1). The two basidiomycete proteomes were included for the off chance of analysing gene clusters that do not originate from ascomycetes. The database covers the complete ascomycetes phylum and was iteratively tested and optimized for the detection of co-evolution in ascomycetes. The sequences were downloaded from the National Center for Biotechnology Information (NCBI) database and the Joint Genome Institute (JGI) [44]. A short identifier, unique in the database for each proteome, was introduced to enable multiple pairwise tree comparisons by the treeKO application [45]. A custom Perl script was used for removing duplicated entries in the database. The database is deposited in the GitHub repository https://github.com/gvignolle/FunOrder (doi:10.5281/zenodo.5118984).

Workflow

The workflow for the FunOrder method is depicted in Fig 2. First, the sequences of the BGC to be analysed are fed into the software bundle. FunOrder accepts a single file in either genbank file format or fasta format as input. The input files contain BGCs predicted by tools such as antiSMASH [29] or DeepBGC [33]. In case a genbank file is provided, a python script (Genbank to FASTA by Cedar McKay and Gabrielle Rocap, University of Washington) is called to extract the amino acid sequence of the genes in the BGC and create a fasta file. The multi-fasta file is then split into individual fasta files each containing a single protein sequence. These are placed in a subfolder created for the analysis of the BGC. Each file is named either after the position of the gene in the BGC or after the respective protein sequence description. This varies from the input file and the varying annotations used (If needed this can be changed in the script following the instructions of Genbank to FASTA by Cedar McKay and Gabrielle Rocap, University of Washington). Each header of the query sequences is tagged with the identifier "query" at the beginning of the header. The individual sequences are compared to the empirically optimized proteome database (Table 1) by a sequence similarity search using blastp 2.8.1+ (Protein-Protein BLAST) [133]. The output of this search is saved in a file with the ". tab" extension. Additionally, an optional remote search of the non-redundant National Center for Biotechnology Information (NCBI) protein database can be performed, yielding a file with the "ncbi.tab" extension. This allows a preliminary manual analysis of the input sequences and facilitates subsequent annotations of the BGCs.

Table 1. Fungal proteomes included in the empirically optimized database.

Organism	Source Database	Identifier	Reference
Acremonium chrysogenum	JGI	AcCh	[46]
Alternaria alternata	NCBI	AlAl	[47]
Alternaria arborescens	NCBI	AlAr	[48]
Alternaria gaisen	NCBI	AlGa	[49]
Alternaria sp. MG1	NCBI	AlSp	[50]
Alternaria tenuissima	NCBI	AlTe	[49]
Amanita muscaria	NCBI	AmMu	[51]
Amorphotheca resinae	JGI	AmRe	[52]
Arthrobotrys oligospora	JGI	ArOl	[53]
Arthroderma benhamiae	JGI	ArBe	[54]
Ascobolus immersus	JGI	AsIm	[55]
Aspergillus costaricaensis	NCBI	AsCo	[56]
Aspergillus fijiensis	NCBI	AsFi	[56]
Aspergillus flavus	NCBI	AsFl	[57]
Aspergillus fumigatus	NCBI	AsFu	[58]
Aspergillus homomorphus	NCBI	AsHo	[56]
Aspergillus ibericus	NCBI	AsIb	[56]
Aspergillus japonicus	NCBI	AsJa	[56]
Aspergillus niger	NCBI	AsNi	[59]
Aspergillus oryzae	NCBI	AsOr	[60]
Aspergillus phoenicis	NCBI	AsPh	[61]
Aspergillus terreus	NCBI	AsTe	[62]
Blumeria graminis	JGI	BlGr	[63]
Botryosphaeria dothidea	JGI	BoDo	[64]
Botrytis cinerea	NCBI	BoCi	[65]
Botrytis elliptica	NCBI	BoEl	[66]
Botrytis galanthina	NCBI	BoGa	[66]
Botrytis hyacinthi	NCBI	BoHy	[66]
Botrytis paeoniae	NCBI	BoPa	[66]
Botrytis porri	NCBI	BoPo	[66]
Botrytis tulipae	NCBI	BoTu	[66]
Cadophora sp.	JGI	CaSp	[67]
Capronia semiimmersa	JGI	CaSe	[68]
Chaetomium globosum	JGI	ChGl	[69]
Choiromyces venosus	JGI	ChVe	[55]
Cladonia grayi	JGI	ClGr	[70]
Cladophialophora bantiana	JGI	ClBa	[68]
Cladophialophora carrionii	JGI	ClCa	[68]
Cladophialophora immunda	JGI	ClIm	[68]
Cochliobolus heterostrophus	JGI	CoHe	[71]
Cochliobolus victoriae	JGI	CoVi	[72]
Colletotrichum nymphaeae	JGI	CoNy	[73]
Colletotrichum orchidophilum	JGI	CoOr	[74]
Colletotrichum salicis	JGI	CoSa	[73]
Colletotrichum simmondsii	JGI	CoSi	[73]
Colletotrichum tofieldiae	JGI	CoTo	[75]
Coniosporium apollinis	JGI	CoAp	[68]

(*Continued*)

Table 1. (Continued)

Organism	Source Database	Identifier	Reference
Coniosporium apollinis CBS 100218	JGI	Capo	[68]
Corynespora cassiicola	JGI	CoCa	[76]
Daldinia eschscholzii	JGI	DaEs	[77]
Diaporthe ampelina	JGI	DiAm	[78]
Diplodia seriata	JGI	DiSe	[78]
Erysiphe necator	JGI	ErNe	[79]
Eutypa lata	NCBI	EuLa	[80]
Exophiala aquamarina	JGI	ExAq	[68]
Exophiala dermatitidis	JGI	ExDe	[68]
Exophiala oligosperma	JGI	ExOl	[68]
Exophiala spinifera	JGI	ExSp	[68]
Exophiala xenobiotica	JGI	ExXe	[68]
Fonsecaea monophora	JGI	FoMo	[81]
Fusarium fujikuroi	NCBI	FuFu	[82]
Fusarium graminearum	NCBI	FuGr	[83]
Fusarium oxysporum	NCBI	FuOx	[84]
Fusarium proliferatum	NCBI	FuPr	[85]
Fusarium pseudograminearum	NCBI	FuPs	[86]
Fusarium verticillioides	NCBI	FuVe	[83]
Gaeumannomyces graminis	JGI	GaGr	[87]
Glonium stellatum	JGI	GlSt	[88]
Hypoxylon sp. EC38	JGI	HyEC	[77]
*Hypoxylon sp.*CO27	JGI	Hysp	[77]
Magnaporthe grisea	JGI	MaGr	[89]
Magnaporthiopsis poae	JGI	MaPo	[87]
Meliniomyces bicolor	JGI	MeBi	[52]
Meliniomyces variabilis	JGI	MeVa	[52]
Metarhizium acridum	NCBI	MeAc	[90]
Metarhizium album	NCBI	MeAl	[91]
Metarhizium anisopliae	NCBI	MeAn	[91]
Metarhizium brunneum	NCBI	MeBr	[91]
Metarhizium guizhouense	NCBI	MeGu	[91]
Metarhizium majus	NCBI	MeMa	[91]
Metarhizium rileyi	NCBI	MeRi	[92]
Metarhizium robertsii	NCBI	MeRo	[90]
Monacrosporium haptotylum	JGI	MoHa	[93]
Morchella importuna	JGI	MoIm	[94]
[Nectria] haematococca	NCBI	NeHa	[95]
Nectria haematococca	JGI	NeHa	[95]
Neurospora crassa	JGI	NeCr2	[96]
Neurospora crassa FGSC	JGI	NeCr	[97]
Neurospora tetrasperma	JGI	NeTe	[98]
Oidiodendron maius	JGI	OiMa	[51]
Ophiostoma piceae	JGI	OpPi	[99]
Paecilomyces variotii	JGI	PaVa	[100]
Panaeolus cyanescens	NCBI	PaCy	[101]
Paracoccidioides brasiliensis	JGI	PaBr	[102]

(*Continued*)

Table 1. (Continued)

Organism	Source Database	Identifier	Reference
Penicillium camemberti	NCBI	PeCa	[103]
Penicillium chrysogenum	NCBI	PeCh	[104]
Penicillium digitatum	NCBI	PeDi	[105]
Penicillium expansum	NCBI	PeEx	[106]
Penicillium nalgiovense	NCBI	PeNa	[107]
Penicillium oxalicum	NCBI	PeOx	[108]
Penicillium roqueforti	NCBI	PeRo	[103]
Penicillium rubens Wisconsin	NCBI	PeRu	[109]
Penicillium vulpinum	JGI	PeVu	[107]
Periconia macrospinosa	JGI	PeMa	[67]
Pestalotiopsis fici	NCBI	PeFi	[110]
Phaeoacremonium aleophilum	JGI	PhAl	[111]
Phaeomoniella chlamydospora	JGI	PhCh	[78]
Phialocephala scopiformis	JGI	PhSc	[112]
Pneumocystis jirovecii	JGI	PnJi	[113]
Pseudogymnoascus destructans	JGI	PsDe	[114]
Pseudomassariella vexata	JGI	PsVe	[115]
Rhizoctonia solani	NCBI	RhSo	[116]
Saccharomyces arboricola	NCBI	SaAr	[117]
Saccharomyces cerevisiae	NCBI	SaCe	[118]
Terfezia boudieri	JGI	TeBo	[55]
Tolypocladium ophioglossoides	NCBI	ToOp	[119]
Tolypocladium paradoxum	NCBI	ToPa	[120]
Trichoderma arundinaceum	NCBI	TrAr	[121]
Trichoderma asperellum	NCBI	TrAs	[122]
Trichoderma atroviride	NCBI	TrAt	[123]
Trichoderma citrinoviride	NCBI	TrCi	[122]
Trichoderma harzianum	NCBI	TrHa	[124]
Trichoderma longibrachiatum	NCBI	TrLo	[125]
Trichoderma reesei	NCBI	TrRe	[126]
Trichoderma virens	NCBI	TrVi	[123]
Trichophyton rubrum	JGI	TrRu	[127]
Tuber aestivum var. urcinatum	JGI	TuAe	[55]
Tuber magnatum	JGI	TuMa	[55]
Venturia inaequalis	JGI	VeIn	[128]
Verruconis gallopava	JGI	VeGa	[68]
Verticillium dahliae	JGI	VeDa	[129]
Xylona heveae	JGI	XyHe	[130]
Zymoseptoria brevis	JGI	ZyBr	[131]
Zymoseptoria pseudotritici	JGI	ZyPs	[132]

The sequences were downloaded from the National Center for Biotechnology Information (NCBI) database or the Joint Genome Institute (JGI). The identifiers were used in the FunOrder software package.

https://doi.org/10.1371/journal.pcbi.1009372.t001

Next, the top 20 results of the blastp analysis are extracted and combined with the query sequence for each gene. A custom Perl script removes potential duplicate entries based on sequence identity. Using emma, a multiple sequence alignment of these protein sequences is

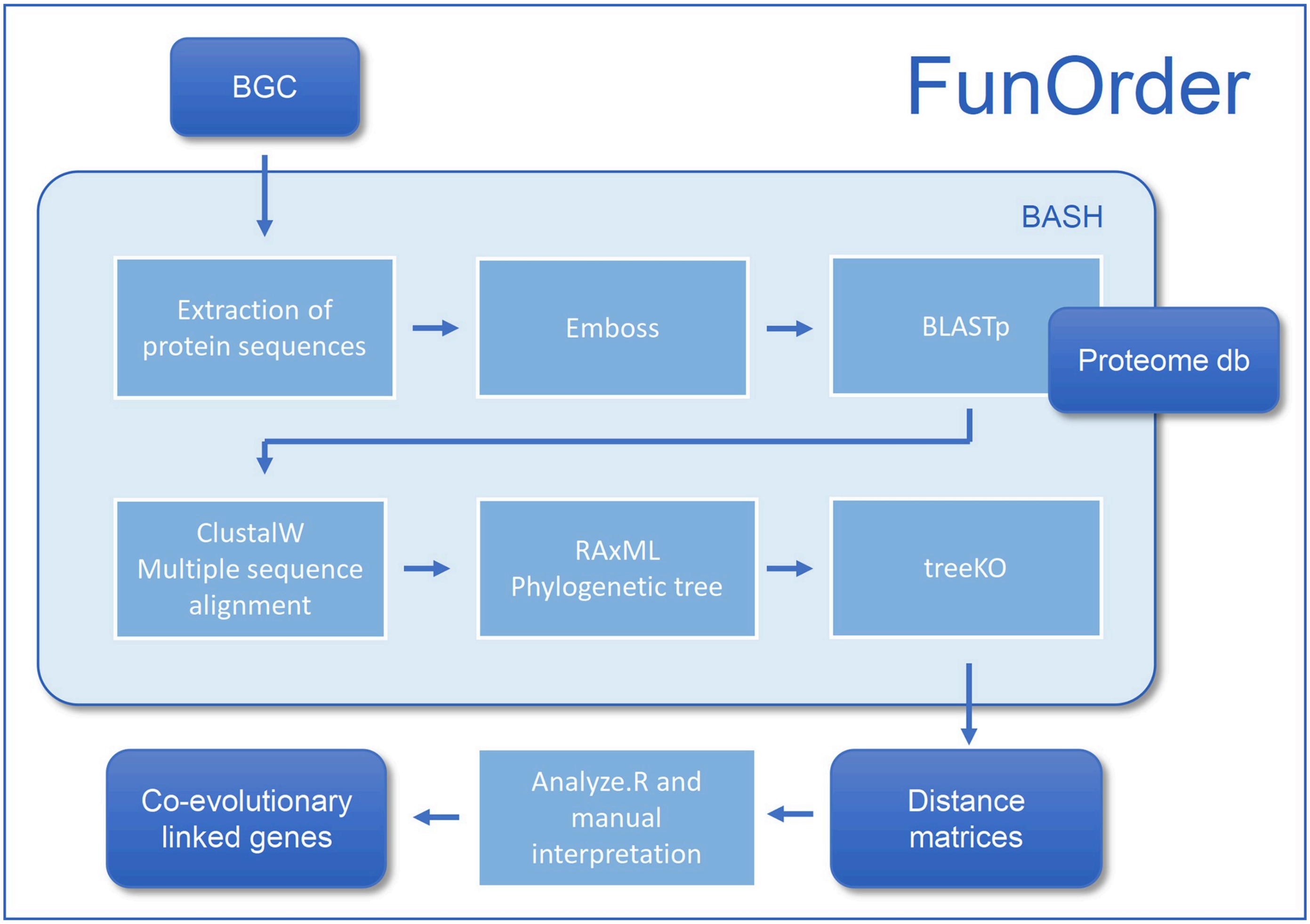
FunOrder
BGC
BASH
Extraction of protein sequences
Emboss
BLASTp
Proteome db
ClustalW Multiple sequence alignment
RAxML Phylogenetic tree
treeKO
Distance matrices
Analyze.R and manual interpretation
Co-evolutionary linked genes

Fig 2. Schematic representation of the workflow of FunOrder.

https://doi.org/10.1371/journal.pcbi.1009372.g002

calculated based on the ClustalW [134] algorithm, and a dendrogram computed. Based on the multiple sequence alignment, 100 rapid Bootstraps and a subsequent search for the best-scoring maximum likelihood (ML) tree are performed using RAxML (Randomized Axelerated Maximum Likelihood) [135]. The phylogenetic trees are computed using the LG amino acid substitution model. Furthermore, a standard ascertainment bias correction by Paul O. Lewis is performed. At this stage, we have obtained a phylogenetic tree (within the context of our empirically optimized database) for each protein of the input BGC.

To estimate if and to what extent the different genes within a BGC are co-evolved, the strict distance and speciation distance among the ML trees of the individual genes are calculated using the TreeKO algorithm [45]. This tool was designed for automated tree comparison and was already suggested to be used for the detection of co-evolution in protein families [45]. The tool compares the topology of different trees; a distance of 0 in both distance measures represents identical trees. In this context, a higher similarity between the different trees of the individual genes points towards a shared evolution. The strict distance is a weighted Robinson-Foulds (RF) distance measure that penalizes dissimilarities in evolutionarily important events such as gene losses and gene duplications; it has been suggested to be more significant in the detection of co-evolution than the evolutionary distance [45]. In contrast, the evolutionary or speciation distance is computed without taking evolutionary exceptions, such as duplication events or different species content of the two compared trees into account and infers shared "speciation history" based solely on topology without considering branch lengths and only considering shared species of the compared trees. Therefore, an evolutionary distance of 0 does not necessarily describe identical trees but shared "speciation history" of shared species. All pairwise strict and evolutionary distances are combined into matrices which are used as input for an R script [136–140].

In this R-script, first, the strict and evolutionary distances are summed up to a third combined distance matrix combining the information about co-evolution and shared speciation into a single measure. In our experience, this measure can be helpful to detect genes that share little co-evolution with the core-enzymes but are still essential for the biosynthesis, which is reflected in a shared speciation. The evolutionary distance is not directly part of the output of FunOrder as is not intended to be used for the detection of co-evolution. Second, the strict and the combined distance matrices are visualized as heatmaps with a dendrogram computed with the complete linkage method, to find similar clusters in these data sets. Next, the Euclidean distance within the matrices is computed and clustered using Ward's minimum variance method aiming at finding compact spherical clusters, with the implemented squaring of the dissimilarities before cluster updating, for the two distance matrices separately, with scaled input data [141]. Lastly, a principal component analysis (PCA) is performed on the two distance matrices and the score plot of the first two principal components visualized, respectively. These outputs enable the adoption of a larger view on the distance measures and thereby allow the analysis of co-evolution within the BGC from different perspectives. We describe in a following subchapter how to interpret these visualisations.

The software bundle is written in the BASH (Bourn Again Shell) environment and includes all necessary subprograms. As BASH is the default shell-language of all Linux distributions and MacOS, FunOrder can run on these two operation systems. The FunOrder software package is deposited in the GitHub repository https://github.com/gvignolle/FunOrder (doi:10.5281/zenodo.5118984). Notably, the software package includes scripts adapted to the use on servers and for the integration in various pipelines; details on these can be found in the ReadMe file on the GitHub repository. FunOrder requires some dependencies e.g., RAxML (Randomized Axelerated Maximum Likelihood) [135] and the EMBOSS (The European Molecular Biology Open Software Suite) package [142], for details and links to all dependencies please refer to the ReadMe file on the GitHub repository.

Compilation of benchmark gene clusters (GCs)

To test and evaluate the applicability of the FunOrder method, we used different control and test gene (or protein) sets. The sequences of all test and control sets are deposited in the GitHub repository https://github.com/gvignolle/FunOrder (doi: 10.5281/zenodo.5118984). The first set of negative control gene clusters (GCs) were 42 completely randomly generated synthetic GCs, which were created with a custom BASH script. Therein, ATGC strings of random composition and length were translated to amino acid strings using transeq from the EMBOSS package and the asterisks were removed. The second set of negative controls were 60 random GCs which were created by subsampling randomly the fungal proteome database with a Perl script from the MEME suit [143]. For each random GC a different seed number was given to guarantee non repetitive GCs, each random GC contained 3–10 randomly chosen protein sequences in a random order. These negative control GCs were subsampled from different genomes to maximize the randomness and use gene clusters that should not contain co-evolved genes.

Table 2. Empirically characterized biosynthetic gene clusters used as positive controls.

Product—BGC	Organism	MIBiG id	Reference(s)
2-Pyridon-Desmethylbassianin (dmb)	*Beauveria bassiana*	BGC0001136	[145]
Aflatoxin (afl)	*Aspergillus flavus*	BGC0000008	[146,147]
Botrydial (bot)	*Botrytis cinera*	BGC0000631	[148,149]
Cephalosporin (cef)	*Acremonium chrysogenum*	BGC0000317	[150]
Compactin (mlc)	*Penicillium citrinum*	BGC0000039	[151,152]
Cyclosporin (cyc2)	*Beauveria felina*	BGC0001565	[18,153–155]
Destruxin (dtxs)	*Metarhizium robertsii*	BGC0000337	[156]
Fumagillin (fma)	*Aspergillus fumigatus*	BGC0001067	[157]
Fumitremorgin (ftm)	*Aspergillus fumigatus*	-	[158–161]
Fumonisin (fum1)	*Fusarium oxysporum*	BGC0000063	[162]
Fumonisin (fum2)	*Fusarium verticilloides*	BGC0000062	[163–170]
Fusaric acid (FUB)	*Fusarium fujikuroi*	-	[171]
Ilicicolin H (ili)	*Neonectaria sp.* DH2	BGC0002035	[172]
Leporin (lep)	*Aspergillus flavus*	BGC0001445	[173]
Lovastatin (lov)	*Aspergillus terreus*	-	[17,62,174]
Mycophenolic acid (mpa1)	*Penicillium brevicompactum*	BGC0000104	[175–180]
Mycophenolic acid (mpa2)	*Penicillium roqueforti*	BGC0001360	[181]
Mycophenolic acid (mpa3)	*Penicillium roqueforti*	BGC0001677	[182]
Paxillin (pax)	*Penicillium paxilli*	BGC0001082	[183]
Penicillin (pen1)	*Penicillium chrysogenum*	BGC0000404	[184]
Penicillin (pen2)	*Penicillium chrysogenum*	BGC0000405	[19]
Pestheic acid (pta)	*Pestalotiopsis fici*	BGC0000121	[185]
Pneumocandin (GL)	*Glaera Iozoyensis*	BGC0001035	[186–188]
Sorbicillinol (sor1)	*Penicillium rubens*	BGC0001404	[189,190]
Sorbicillinol (sor2)	*Trichoderma reesei*	-	[191]
Tenellin (ten)	*Beauveria bassiana*	BGC0001049	[192,193]
Terrein (ter)	*Aspergillus terreus*	BGC0000161	[194]
Tetramic acid (tas)	*Hapsidospora irregularis*	-	[195]
Ustiloxin B (ust)	*Aspergillus flavus*	-	[196]
Xanthocillin (xan)	*Aspergillus fumigatus*	BGC0001990	[197]

https://doi.org/10.1371/journal.pcbi.1009372.t002

We used a set of 30 empirically well characterized BGCs from a broad range of different genera (Table 2) as positive controls. The BGC sequences were downloaded from NCBI or the MIBiG (Minimum information about a biosynthetic gene cluster) database [144]. The sequences are available at the GitHub repository https://github.com/gvignolle/FunOrder (doi:10.5281/zenodo.5118984). All BGCs were manually inspected for correctness and completeness based on the respective literature (S1 Table, references in Table 2). We further added 2 genes on each side of the BGC to mimic the greedy gain performed by antiSMASH, if possible (sequences available) and applicable (only few or no gap genes present). Next, we defined the class of each gene (biosynthetic gene, further essential gene, gap, or extra gene) according to the described function of the enzymes in the literature (S1 Table).

Further, we compiled 10 protein sets containing the sequences of enzymes of conserved metabolic pathways from organisms that were not included in the proteome database, termed „Biosynthetic_pathways“, or „BioPath"(S2 Table; sequences deposited at the GitHub repository https://github.com/gvignolle/FunOrder (doi:10.5281/zenodo.5118984)). As we anticipate a strong co-evolution among the corresponding genes, we used these sets as positive controls for co-evolution in general. Finally, we subsampled the genomes of organisms that were not included in the proteome database for 30 random loci containing 8 to 10 genes (S3 Table; sequences available at the GitHub repository https://github.com/gvignolle/FunOrder (doi:10.5281/zenodo.5118984)). We termed this control set „sequential GCs". This set should represent the random degree of co-evolution based only on genomic vicinity. Notably, due to the randomness of the sampling, the sequential GCs may also contain evolutionary linked genes.

Calculation of MEM and determination of thresholds for co-evolution

As the thresholds for the strict and/or evolutionary distance for the analysis of protein co-evolution are database dependent, we needed to define these thresholds manually. To this end, we performed a manual comparison of the phylogenetic trees of genes anticipated to be co-evolved and of not presumably co-evolved genes. As positive control datasets (anticipated co-evolution), we used the essential genes within the positive control BGCs. As negative control data set (anticipated to not have co-evolved), we used the genes in the random GCs. For the manual tree comparisons, we considered the topology (defined in S4 Table), branch lengths, number of nodes, and shared leaves of the trees and calculated the manual evaluation measure (MEM) according to the definitions in S5 Table. We calculated the MEM for each gene tree pair of the positive and the negative control data sets (S6 and S7 Tables, respectively). The measure ranges from 3 (same) to 0 (no shared leaves). The MEM values of each pair-wise tree comparison were then manually reconciled with the corresponding strict and the combined distance measures obtained from the treeKO analysis and the subsequent R script, respectively. The procedure is exemplary described for the 2-Pyridon-Desmethylbassianin (dmb) BGC from *Beauveria bassiana* in S1 File. Based on these manual comparisons, we defined the threshold values for strict and combined distances in the following: two genes are considered as co-evolved if the strict distance value is less than 0.7 or if the combined distance is equal to or less than 60 percent of the maximum value in the combined distance matrix of the analysed set.

Calculation of the Internal co-evolutionary quotient (ICQ)

The internal co-evolutionary quotient (ICQ) expresses how many genes in a GC or proteins in a protein set are co-evolved according to the previously defined threshold for strict and combined distances within the distance matrices of an analysed GC (or protein set). To calculate the ICQ, each protein is compared with every other protein. The total number of all possible pairwise comparisons is 2* [d*(d-1)] for d proteins. The ICQ was calculated using Eq 1, resulting in values between 0 and 1, with 1 representing no co-evolved genes, and 0 representing that most genes are co-evolved with each other in the insert GC.

$$ICQ = 1 - \left\{ \frac{g}{2 * [d * (d - 1)]} \right\} \qquad \text{Eq 1}$$

ICQ = internal co-evolutionary quotient; g = number of strict distances < 0.7 and combined distances < = (0.6 * max value of the combined distance matrix) in all matrices (visualized in the heatmaps); d = number of genes in the GC.

Manual interpretation of the FunOrder output

The FunOrder outputs three different visualizations (heatmap, dendrogram, PCA) each of the strict and combined distance matrices among the genes (or proteins) of an inserted GC (or protein set). These visualizations need to be interpreted manually. For the manual interpretation, we first searched for genes that clustered together with the core enzyme(s) in any of the three visualisations of the strict distance. The definition of the clusters needs to be performed carefully keeping the biological background (gene predictions) in mind. For instance, a cluster containing typical tailoring enzymes (e.g., hydrolases, P450 cytochrome oxidases, FAD-containing enzymes, etc.) and/or further essential genes (e.g., transcription factors or transporters) make sense, whereas clusters containing a lot of genes encoding for unknown genes and/or genes that are unlikely to be involved in the biosynthesis of a secondary metabolite) do not make sense. Next, clustering in the visualizations of the combined distances is considered. As the combined distance also contains information about the speciation history, it may be used to add further genes to the list of "detected genes". Notably, this needs to be critically evaluated and decided on a case-to-case basis, taking the gene predictions into account. Please also refer to S2 File for a detailed step-by-step description of the interpretation procedure, the exemplary analysis of the lovastatin BGC from *A*. *terreus* in the results, and S3 File and S4 File for the exemplary analysis of two unknown BGCs.

Performance evaluation

To test the robustness of FunOrder, we analysed 42 completely randomly generated synthetic GCs. To test whether the FunOrder method can be used to detect co-evolution within GCs (or protein sets), we calculated the ICQ for different control sets and compared the results in a kernel density plot. To evaluate the performance of the FunOrder method regarding its capability to identify presumably co-evolved essential genes (as defined in S1 Table) and to distinguish them from (presumably not co-evolved) gap genes and genes outside of the BGC via the detection of co-evolution, we performed a manual interpretation of 30 empirically characterized BGCs (Table 2) as described above. Genes that clustered together with the core enzyme(s) according to the procedure described above were considered as „detected". Then we counted the total number of (1a) detected essential genes or (1b) detected biosynthetic genes, (2a) not detected essential genes or (2b) not detected biosynthetic genes, (3) detected gap and extra genes, and (4) not detected gap or extra genes in all BGCs, and defined (1a or 1b) as true positives (TP), (2a or 2b) as false negatives (FN), (3) as false positives (FP), and (4) as true negatives (TN). The values were used for a final statistical evaluation of FunOrder as suggested by Chicco and Jurman [198].

Results and discussion

Applicability of FunOrder for the detection of co-evolution

First, we analyzed the 42 synthetic negative control GCs with the FunOrder software. We could not find any sequence similarities with the empirically optimized fungal proteome database, demonstrating the robustness of the FunOrder method towards non-biological random amino acid sequences. Consequently, the 42 synthetic negative control GCs were not considered in the following.

Next, we performed FunOrder analyses of different control GCs and protein sets and calculated the internal co-evolutionary quotients (ICQs) using Eq 1. The ICQ is a value for the relative amount of co-evolutionary relations among the genes (or proteins) in a given GC or protein set. An ICQ of 0 means that most genes (or proteins) are co-evolved with each other. An ICQ of 1 means, that no co-evolution can be detected using the defined thresholds. As negative control for co-evolution, we used 60 randomly assembled negative control GCs (random GCs, S8 Table). The random GCs were compiled by subsampling different proteomes, to minimize the chance of random, unwanted co-evolution in the clusters. As positive control for co-evolution we used 10 protein sets from conserved metabolic pathways of different ascomycetes (S2 Table), termed „Biosynthetic pathways", or „BioPath". Given, that the proteins are part of

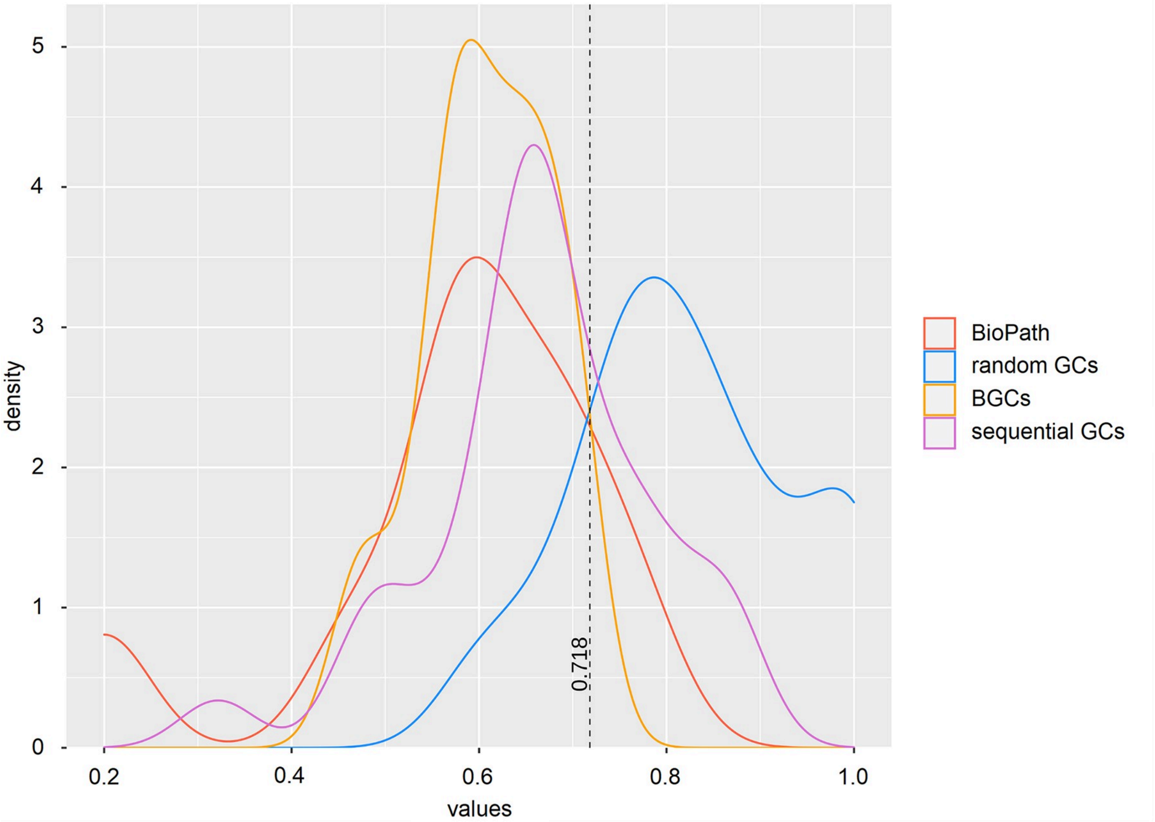

Fig 3. Kernel density plot of the ICQ values for co-evolutionary linked enzymes of different control sets. BioPath, protein sets of conserved biosynthetic pathways of the primary metabolism (S2 Table); random GCs, randomly assembled protein sets from 134 fungal proteomes (Table 1); BGCs, previously empirically characterized fungal BGCs (Table 2); sequential GCs, co-localized genes from random loci of different ascomycetes (S3 Table).

https://doi.org/10.1371/journal.pcbi.1009372.g003

the conserved primary metabolism and that their enzymatic functions are interrelated, we can assume a high level of internal co-evolution among the proteins within these protein sets. As control for the basic co-evolutionary value of co-localized (or sequential) genes, we used 30 random genetic loci containing 8 to 10 genes (S3 Table). We termed this control set „sequential GCs". As test set for BGCs of the secondary metabolism in ascomycetes we used 30 empirically characterized BGCs (Table 2, S1 Table), also termed positive control BGCs.

We compared the ICQs of the different sets in an ANOVA (S5 File) and in a kernel density plot (Fig 3). We found that the ICQs for the random GCs were significantly different from all the other sets, demonstrating that the workflow of the FunOrder method can be used to detect co-evolution, that the ICQ is a meaningful measure to represent the content of co-evolutionary relationships within a GC or protein set, and that the manually defined thresholds for strict and combined distances are applicable to define co-evolution within GC or proteins sets. Based on these results, we defined the threshold of the ICQ for biologically relevant co-evolution within a GC as the point of intersect between the random GCs and the BGCs (0.718). GCs with an ICQ above this threshold do not contain significantly more co-evolutionary connections among the contained genes than randomly assembled GCs.

To our surprise, we could not detect a statistically significant difference between the sequential GCs and the positive control GCs. However, the maxima for the BioPath proteins and the BGC are at the same value and the shape of the corresponding density plot is remarkably similar (Fig 3), whereas the maximum of the sequential GC is shifted towards the random GCs and the shape of the curve is different to the two positive control sets (Fig 3). These results indicate, that using only the absolute values of strict and combined distance may not be enough to distinguish co-evolutionary linked genes within the context of co-localized genes, but that the distances need to be assessed and interpreted in a case-by-case scenario considering the biological background and context of the analyzed GC.

Exemplary analysis of the lovastatin BGC (lov)

The FunOrder method allows the detection of co-evolved genes within a set of genes or proteins. As mentioned, we speculate that essential genes in BGCs are co-evolving and can therefore be differentiated from gap genes. In this context, the application of FunOrder might be used to detect the essential or at least the biosynthetic genes in BGCs. The software package of the FunOrder method calculates two distance matrices for the proteins within an input GC representing the evolutionary similarities (based on pair-wise comparisons of the phylogenetic trees using the treeKO tool [45]). First, we tried to use the previously defined thresholds for the strict and combined distances to automatically detect the co-evolutionary relations in BGCs. As insinuated above, this proofed not to be a successful strategy (not shown). We speculate, that the evolutionary similarities or distances among neighbouring genes are highly location specific and that the absolute values are therefore not meaningful as general thresholds. However, as the underlying strategy and method is clearly able to detect co-evolution (Fig 3), we speculated that the obtained data may need to be represented in different forms and/or reduced. Consequently, we added the following data visualizations to the FunOrder pipeline. The strict and combined distances are visualized in a heatmap and clustered by higher similarities (complete linkage method). Next, the Euclidean distances within the scaled distance matrices are calculated and clustered (hierarchical clustering) using the Wards minimum variance method aiming at finding compact spherical clusters, with the implemented squaring of dissimilarities before cluster updating. The clustering is visualized in dendrograms. Finally, the principal components of the data are represented in a score plot. Here, we exemplary describe the manual interpretation of these visualizations (S6 File and Fig 4) with the aim to detect co-evolution within the lovastatin BGC of *A*. *terreus* (lov, Fig 1). Please refer also to the step-by-step description on how to interpret the FunOrder output in S2 File.

For the analysis of the lovastatin BGC, we first had a look at the heatmap representing the strict distance matrix (S6 File). Therein all biosynthetic genes (*lovA-D*, *F*, *G*; Fig 1, red arrows) are clustering together with each other and with the gap gene *orf1*, although not all inter-gene distances were below the previously defined threshold (S6 File, heatmaps). This demonstrates again that, evaluating only the numerical values (regardless of the concrete thresholds) is not enough for a thorough analysis of a BGC. It is necessary to consider the distances within the genomic context by comparing all provided visualisations. The biosynthetic genes of lovastatin (*lovA-D*, *F*, *G*) also formed distinct clusters in the dendrograms and in the PCA of the strict distance (S6 File and Fig 4A) In our experience, it was often helpful to additionally take the combined distance values into consideration to get a more comprehensive picture of the BGC. As mentioned before, the combined distance also considers speciation history. In the case of the lovastatin BGC, *orf10* and *extra03* clustered together with *lovA*, *B*, *D*, *F*, *G* in the PCA of the combined distance (Fig 4B). The gene *orf10* encodes for an MFS (major facilitator superfamily) transporter, which warrants adding it to the „detected genes“; the transporter is actually necessary for the export of lovastatin [17] (Fig 1). The gene *extra03* is predicted to encode for an alpha-glucuronidase (AguA) which is involved in the hydrolysis of xylan. Therefore, the clustering only in combined distance matrix does not justify classifying the gene *extra03* as

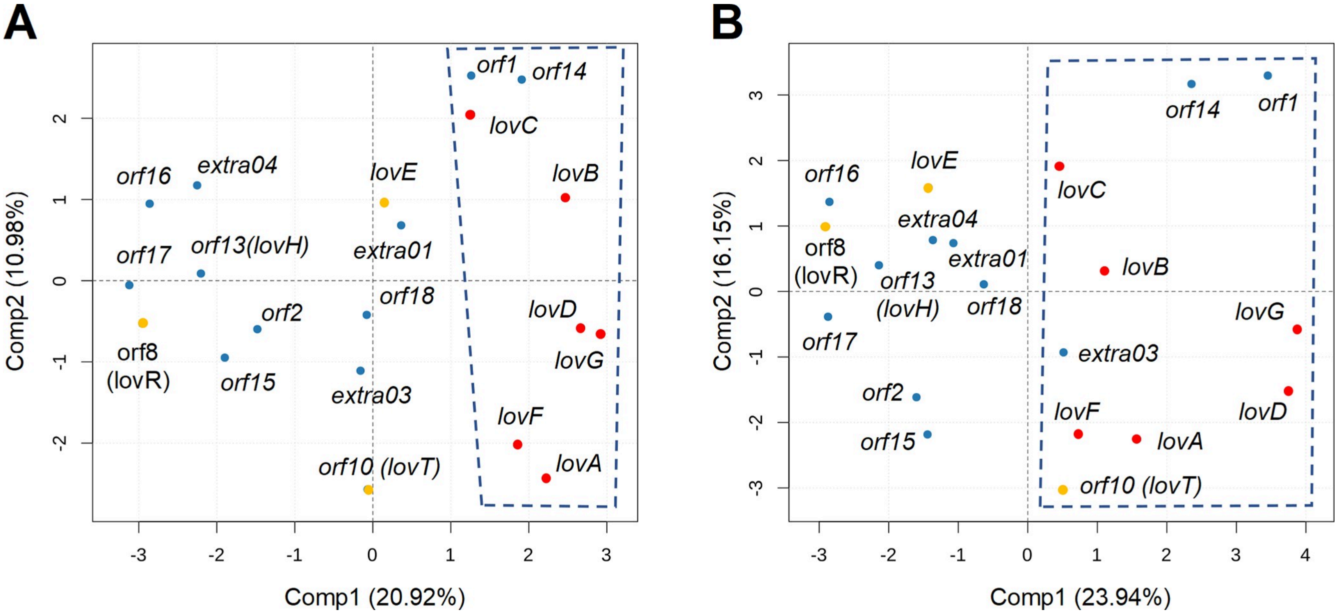

Fig 4. A selection of the standard output of the FunOrder analysis of the lovastatin BGC (lov). Score plots of the first two principal components from a PCA performed on the strict distance matrix (A) and on the combined distance matrix (B). The biosynthetic genes and the further essential genes are indicated in red and gold, respectively. Clusters in the PCA are indicated by the dashed boxes.

https://doi.org/10.1371/journal.pcbi.1009372.g004

„detected“. The other two „further essential genes“, *lovE and orf8* did not cluster together with the biosynthetic genes in any visualizations of the distance matrices (Fig 4 and S6 File)). LovE is a transcription factor and the main regulator of the lovastatin cluster [17] and essential for the lovastatin biosynthesis in the native organism, although it is not directly part of the biosynthetic pathway. The gene *orf8* encodes for a 3-hydroxy-3-methylglutaryl coenzyme-A (HMG-CoA) reductase, which is the target of statins [199] and in this case is conveying self-resistance to lovastatin [200]. These results suggest that these two genes did not undergo the same evolutionary process as the biosynthetic genes. This is in accordance with the „brick and mortar”model suggested by Medema et al. [40]. The biosynthetic genes represent a co-evolving „brick“, that is integrated into the biological context of *A*. *terreus* via the „mortar”that are the further essential genes.

This exemplary analysis demonstrates how the different data output formats of the software package need to be considered and compared manually, to decide on which genes are co-evolutionary linked and likely to be involved in the biosynthesis of a secondary metabolite. When considering only one output, one might get a distorted view of the analysed BGC. Notably, we did not intend to leave this step up to automation, because the human (expert or child) pattern recognition and mind still outperforms artificial intelligence (AI) algorithms and machine learning algorithms in this regard [201]. Please also refer to S3 File and S4 File in which we describe the analysis of two yet undescribed BGCs.

Speed and scalability of the software

As the empirically optimized proteome database contained only 134 fungal proteomes, we were able to use the blastp algorithm for sequence similarity search. The analysis of the lovastatin BGC of *A*. *terreus* (lov) with 17 genes, took 1 h 19 m 48 sec real time using 22 threads on an Ubuntu Linux system with 128 GB DDR4 RAM. The same analysis took 6 h 54 m 50 sec real time using 3 threads and 5 h 48 m 50 sec using 4 threads on a Linux Mint Laptop, demonstrating that the analysis of such a large cluster as the lovastatin cluster is fast and feasible. The number of threads can be defined, to increase the scalability and the overall performance.

Performance evaluation

Up to this point, we demonstrated that the FunOrder method can be used to detect the overall level of internal co-evolutionary relations within a GC or set of proteins. We demonstrated that similar levels of co-evolutionary relations occur among the genes in BGCs and among proteins of conserved metabolic pathways of the primary metabolism, and that these positive control sets can be distinguished from negative control GC, containing randomly stringed together proteins from different organisms with a threshold of 0.718 for the ICQ (Fig 3). Further, we showed that the values of strict and combined distances need to be visualized in different forms and then interpreted manually to detect co-evolution of individual genes within fungal BGCs. Next, we aimed to test, whether the detection of co-evolved genes is indeed a useful approach to identify the essential genes in fungal BGCs. To this end, we analysed the 30 empirically verified BGCs (Table 2) as described for the lovastatin cluster before. We looked for genes that are co-evolutionary linked with the core biosynthetic gene. These genes were considered as "detected". The "detected" genes sets were compared to the previously empirically obtained set of essential genes and classified the genes in true positives (TP), false negatives (FN), false positives (FP), or true negatives (TN) (S1 Table). To test and evaluate, how well FunOrder is performing in detecting either all essential or just the biosynthetic genes, we determined two different sets of TP and FN. TPs were either all detected essential genes, or all detected biosynthetic genes. Accordingly, FNs were either all not detected essential genes or all not detected biosynthetic genes (S1 Table). In both cases, FPs were all detected gap and extra genes, and TNs were all not detected gap and extra genes (S1 Table) because it makes biologically no sense to define a „detected"further essential gene as a FP, even when defining detected biosynthetic genes as TP. For an initial performance estimation, we calculated the percentages of detected essential and biosynthetic genes (S1 Table) and compiled them in a kernel density plot (Fig 5). More than 75% of all essential genes and biosynthetic genes were found to be co-evolving using the FunOrder method in 13 and 16 BGCs (out of 30 BGCs), respectively. The curves in the density plot also differ at high percentages; nearly all (above 90%) biosynthetic genes could be detect in more cases than nearly all essential genes. These two observations point in the direction, that especially the biosynthetic genes share a more coherent co-evolutionary history and can thus be identified by looking for co-evolved genes in BGCs. Obviously, not all essential genes in all BGCs are co-evolving and/or can be detected as co-evolved with this method. This is at least partly based on the biological background. Each BGC has a unique evolutionary background and needs to be interpreted individually. The FunOrder method offers additional information about co-evolution for already defined BGCs and may be useful in deciding which genes might be most relevant when studying a BGC.

For a stringent statistical evaluation, we calculated the normalized Matthews correlation coefficient (normMCC) and other classical metrics and global metrics (Table 3) as indicated by Chicco and Jurman [198] based on the previously defined TP, FN, FP, and TN (S1 Table). To determine the degree of balance between positive and negative controls we calculated the no-information error rate ni which is best for balanced test sets with the value 0.5. The obtained values of 0.5084 and 0.5444 allowed for the usage and confirmed the validity of the classical metrics such as F1 score and Accuracy. The FunOrder method displays overall high metrics in identifying essential and/or biosynthetic genes in a BGC. Despite the differences between biosynthetic and essential genes in Fig 5, we could not detect strong differences in the

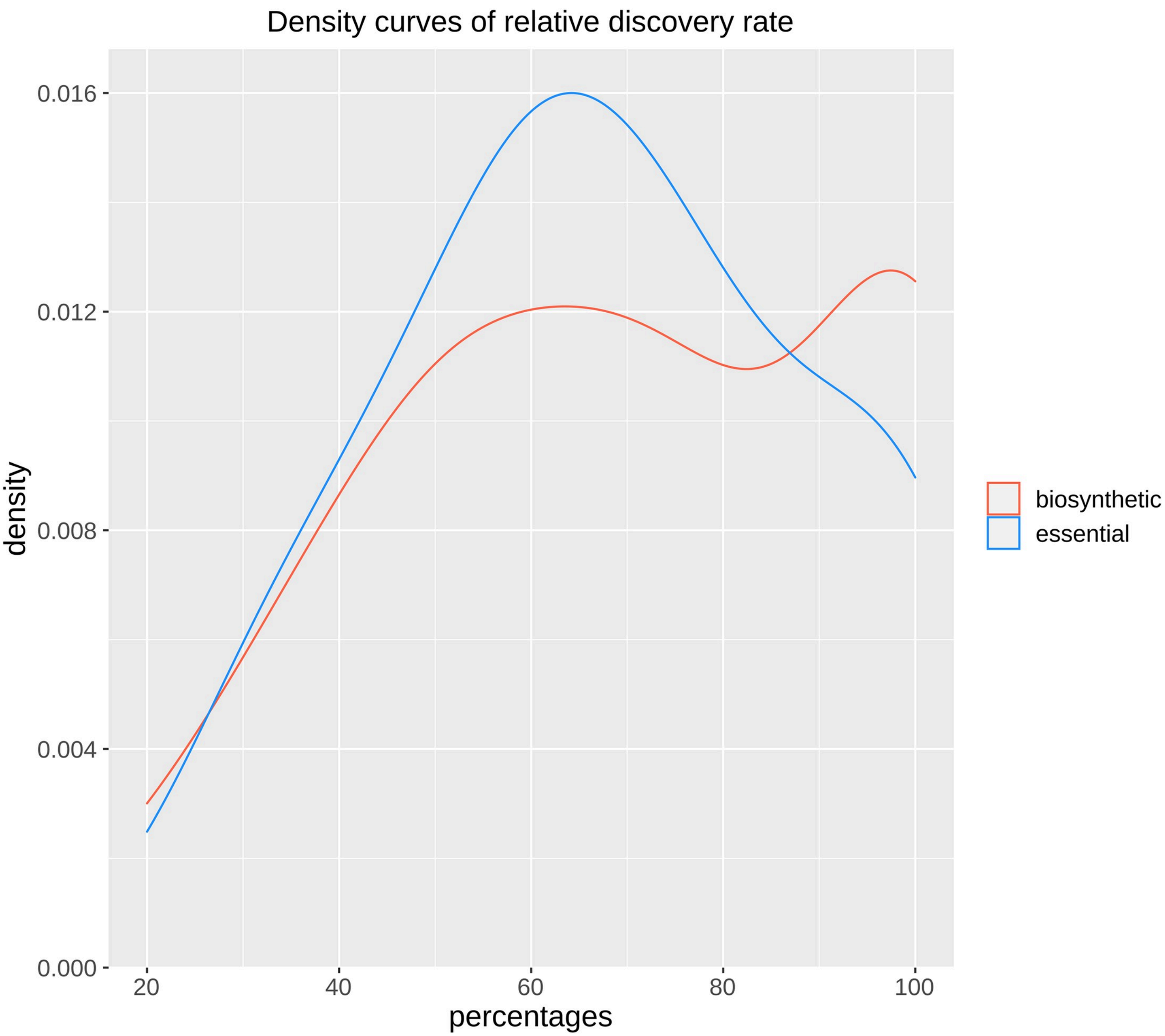

Fig 5. Kernel density plots of the relative discovery rate of essential or biosynthetic genes in 30 tested fungal BGCs.

https://doi.org/10.1371/journal.pcbi.1009372.g005

overall statistical assessment. FunOrder can be used to detect essential and biosynthetic genes in a BGC based on protein family co-evolution with a accuracy of 0.7215 and 0.743, respectively.

Concluding remarks

The FunOrder method was created to identify the essential genes in a BGC and distinguish them from gap genes based on the hypothesis that the essential genes are co-evolutionary linked. We evaluated this method and simultaneously tested the underlying hypothesis using different control sets of genes and proteins, respectively. We observed on the one hand that

Table 3. Statistical evaluation of the performance of FunOrder in detecting relevant genes in BGCs.

	essential genes	biosynthetic genes
Sensitivity	0.6349	0.6615
Specificity	0.8112	0.8112
Precision	0.7766	0.7457
Negative Predictive Value	0.6823	0.7412
False Positive Rate	0.1888	0.1888
False Discovery Rate	0.2234	0.2543
False Negative Rate	0.3651	0.3385
Accuracy	0.7215	0.743
F1 Score	0.6986	0.7011
Matthews Correlation Coefficient	0.4524	0.4797
Normalized Matthews Correlation Coefficient	0.7262	0.73985
No-information error rate ni	0.5084	0.5444

https://doi.org/10.1371/journal.pcbi.1009372.t003

co-evolutionary linkage in fungal BGCs is commonly occurring—especially within the biosynthetic genes, and on the other hand that the FunOrder method can be used to detect the biosynthetic genes within BGCs and to some extent also the further essential genes. We would like to stress that this method is delivering data on co-evolution, that needs to be critically evaluated and interpreted keeping the biological background in mind, and that FunOrder is not to be considered a stand-alone tool but meant to deliver supplementary data about co-evolution within predefined BGCs.

During the testing and evaluation, we encountered several cases of ambiguous results, where the different visualizations clustered different genes together. One way to handle such ambiguous results is to critically assess the results by considering the gene predictions. We further suggest adding and/or removing genes at the edges of the BGC and re-running the analysis. This might change the clustering behaviour and clarify the results. Alternatively, homologous BGCs from other fungi may be analysed by FunOrder and the clustering of the corresponding genes compared to the initial BGC.

The basis but also limitation for the method is the database [42]. Here we used a specific set of proteomes (Table 1) and were thus able to detect co-evolved genes in ascomycetes. Notably, the underlying strategy and workflow of FunOrder can be adapted to analysing genomic regions in other phyla, orders, or even kingdoms by using different databases. In case a larger database is integrated into the software package, alternative search algorithms, such as DIAMOND [202] or HMMER (similarity search using hidden Markov models) [203] might be used instead of blastp to enhance the performance. Nevertheless, each novel database, even if only one single proteome would be introduced in an existing database, will have to be verified and validated.

In this study, we looked for genes that share the same or a similar evolutionary background with the core genes of BGCs and could demonstrate that FunOrder is a fast and powerful method that can support scientists to decide which genes of a BGC are promising study objects. Notably, the application of this method is not limited to fungal BGC. It can be used for any applications where information of a shared co-evolution can contribute to a better understanding. FunOrder with the existing ascomycete database might already be used for a genome wide analysis of co-evolving transcription factors or detection of functionally connected protein-protein interactions [42]. As a future perspective, FunOrder might be even used for the analysis of total proteomes to detect evolutionary linked genes.

Supporting information

S1 Table. Empirically tested BGCs used as control set in this study.
(XLSX)

S2 Table. Protein sets of conserved metabolic pathways of the primary metabolism.
(XLSX)

S3 Table. Sequential GCs used in this study.
(XLS)

S4 Table. Definition of topology.
(PDF)

S5 Table. Parameters used to calculate the manual evaluation measure (MEM).
(PDF)

S6 Table. Calculation of MEM values for positive control BGCs.
(XLSX)

S7 Table. Calculation of MEM values for negative control GCs.
(XLSX)

S8 Table. Random GCs used in this study.
(XLSX)

S1 File. Exemplary MEM analysis of the dmb BGC.
(PDF)

S2 File. Step-by-step explanation for the manual interpretation of the FunOrder output.
(PDF)

S3 File. Exemplary interpretation of the FunOrder output of an unknown fungal BGC 1.
(PDF)

S4 File. Exemplary interpretation of the FunOrder output of an unknown fungal BGC 2.
(PDF)

S5 File. ANOVA for the ICQ values of the control and tests GCs and protein sets, respectively.
(PDF)

S6 File. FunOrder output of the Lovastatin BGC from *A. terreus* (lov).
(PDF)

Acknowledgments

We thank Dr. Ekaterina Shelest for reviewing a preprint of this study on https://www.biorxiv.org/ and providing us with helpful inputs and suggestions. We thank Matthias Schmal for testing the FunOrder installation and performance on a Linux Mint Laptop.

Author Contributions

Conceptualization: Gabriel A. Vignolle, Christian Derntl.

Data curation: Gabriel A. Vignolle, Denise Schaffer, Leopold Zehetner.

Formal analysis: Gabriel A. Vignolle, Leopold Zehetner.

Funding acquisition: Robert L. Mach, Christian Derntl.

Investigation: Gabriel A. Vignolle, Denise Schaffer.

Methodology: Gabriel A. Vignolle, Christian Derntl.

Project administration: Christian Derntl.

Resources: Robert L. Mach, Astrid R. Mach-Aigner.

Software: Gabriel A. Vignolle.

Supervision: Robert L. Mach, Astrid R. Mach-Aigner, Christian Derntl.

Validation: Gabriel A. Vignolle, Leopold Zehetner, Christian Derntl.

Visualization: Gabriel A. Vignolle, Christian Derntl.

Writing – original draft: Gabriel A. Vignolle, Christian Derntl.

Writing – review & editing: Denise Schaffer, Robert L. Mach, Astrid R. Mach-Aigner.

References

1. Thirumurugan D, Cholarajan A, Raja SSS, Vijayakumar R. An Introductory Chapter: Secondary Metabolites. In: Vijayakumar R, Raja SSS, editors. Secondary Metabolites—Sources and Applications. London, UK: IntechOpen Limited; 2018.
2. Malik VS. Microbial secondary metabolism. Trends in Biochemical Sciences. 1980; 5(3):68–72. https://doi.org/10.1016/0968-0004(80)90071-7.
3. Keller NP, Turner G, Bennett JW. Fungal secondary metabolism—from biochemistry to genomics. Nature Reviews Microbiology. 2005; 3(12):937–47. https://doi.org/10.1038/nrmicro1286 PMID: 16322742
4. Alberti F, Foster GD, Bailey AM. Natural products from filamentous fungi and production by heterologous expression. Applied microbiology and biotechnology. 2017; 101(2):493–500. Epub 2016/12/15. https://doi.org/10.1007/s00253-016-8034-2 PMID: 27966047; PubMed Central PMCID: PMC5219032.
5. Newman DJ, Cragg GM, Kingston DGI. Chapter 5—Natural Products as Pharmaceuticals and Sources for Lead Structures**Note: This chapter reflects the opinions of the authors, not necessarily those of the US Government. In: Wermuth CG, Aldous D, Raboisson P, Rognan D, editors. The Practice of Medicinal Chemistry (Fourth Edition). San Diego: Academic Press; 2015. p. 101–39.
6. Brakhage AA, Schroeckh V. Fungal secondary metabolites—strategies to activate silent gene clusters. Fungal genetics and biology: FG & B. 2011; 48(1):15–22. https://doi.org/10.1016/j.fgb.2010.04.004 PMID: 20433937.
7. Atanasov AG, Zotchev SB, Dirsch VM, Orhan IE, Banach M, Rollinger JM, et al. Natural products in drug discovery: advances and opportunities. Nature Reviews Drug Discovery. 2021. https://doi.org/10.1038/s41573-020-00114-z PMID: 33510482
8. Wiemann P, Keller NP. Strategies for mining fungal natural products. Journal of industrial microbiology & biotechnology. 2014; 41(2):301–13. https://doi.org/10.1007/s10295-013-1366-3 PMID: 24146366.
9. Kensler TW, Roebuck BD, Wogan GN, Groopman JD. Aflatoxin: a 50-year odyssey of mechanistic and translational toxicology. Toxicol Sci. 2011; 120 Suppl 1:S28–48. https://doi.org/10.1093/toxsci/kfq283 PMID: 20881231; PubMed Central PMCID: PMC3043084.
10. Blount W. Turkey “X” disease. Turkeys. 1961; 9(2):52–5.
11. Yu J, Chang PK, Cary JW, Wright M, Bhatnagar D, Cleveland TE, et al. Comparative mapping of aflatoxin pathway gene clusters in *Aspergillus parasiticus* and *Aspergillus flavus*. Applied and environmental microbiology. 1995; 61(6):2365–71. https://doi.org/10.1128/aem.61.6.2365-2371.1995 PMID: 7793957; PubMed Central PMCID: PMC167508.
12. Soldatou S, Eldjarn GH, Huerta-Uribe A, Rogers S, Duncan KR. Linking biosynthetic and chemical space to accelerate microbial secondary metabolite discovery. FEMS microbiology letters. 2019; 366 (13). https://doi.org/10.1093/femsle/fnz142 PMID: 31252431
13. Craney A, Ahmed S, Nodwell J. Towards a new science of secondary metabolism. The Journal of antibiotics. 2013; 66(7):387–400. https://doi.org/10.1038/ja.2013.25 PMID: 23612726

14. Osbourn A. Secondary metabolic gene clusters: evolutionary toolkits for chemical innovation. Trends in Genetics. 2010; 26(10):449–57. https://doi.org/10.1016/j.tig.2010.07.001 PMID: 20739089

15. Tran PN, Yen MR, Chiang CY, Lin HC, Chen PY. Detecting and prioritizing biosynthetic gene clusters for bioactive compounds in bacteria and fungi. Appl Microbiol Biotechnol. 2019; 103(8):3277–87. Epub 2019/03/13. https://doi.org/10.1007/s00253-019-09708-z PMID: 30859257; PubMed Central PMCID: PMC6449301.

16. Keller NP. Fungal secondary metabolism: regulation, function and drug discovery. Nat Rev Microbiol. 2019; 17(3):167–80. Epub 2018/12/12. https://doi.org/10.1038/s41579-018-0121-1 PMID: 30531948; PubMed Central PMCID: PMC6381595.

17. Mulder KC, Mulinari F, Franco OL, Soares MS, Magalhaes BS, Parachin NS. Lovastatin production: From molecular basis to industrial process optimization. Biotechnol Adv. 2015; 33(6 Pt 1):648–65. Epub 2015/04/15. https://doi.org/10.1016/j.biotechadv.2015.04.001 PMID: 25868803.

18. Weber G, Schörgendorfer K, Schneider-Scherzer E, Leitner E. The peptide synthetase catalyzing cyclosporine production in *Tolypocladium niveum* is encoded by a giant 45.8-kilobase open reading frame. Current Genetics. 1994; 26:120–5. https://doi.org/10.1007/BF00313798 PMID: 8001164

19. van den Berg MA, Westerlaken I, Leeflang C, Kerkman R, Bovenberg RA. Functional characterization of the penicillin biosynthetic gene cluster of *Penicillium chrysogenum* Wisconsin54-1255. Fungal genetics and biology: FG & B. 2007; 44(9):830–44. Epub 2007/06/06. https://doi.org/10.1016/j.fgb.2007.03.008 PMID: 17548217.

20. Nosanchuk JD, Stark RE, Casadevall A. Fungal Melanin: What do We Know About Structure? Front Microbiol. 2015; 6:1463. Epub 2016/01/07. https://doi.org/10.3389/fmicb.2015.01463 PMID: 26733993; PubMed Central PMCID: PMC4687393.

21. Wheeler MH, Bell AA. Melanins and their importance in pathogenic fungi. Curr Top Med Mycol. 1988; 2:338–87. https://doi.org/10.1007/978-1-4612-3730-3_10 PMID: 3288360.

22. Luo S, Dong SH. Recent Advances in the Discovery and Biosynthetic Study of Eukaryotic RiPP Natural Products. Molecules. 2019; 24(8). Epub 2019/04/21. https://doi.org/10.3390/molecules24081541 PMID: 31003555; PubMed Central PMCID: PMC6514808.

23. Montalban-Lopez M, Scott TA, Ramesh S, Rahman IR, van Heel AJ, Viel JH, et al. New developments in RiPP discovery, enzymology and engineering. Nat Prod Rep. 2020. Epub 2020/09/17. https://doi.org/10.1039/d0np00027b PMID: 32935693.

24. Wang DN, Toyotome T, Muraosa Y, Watanabe A, Wuren T, Bunsupa S, et al. GliA in *Aspergillus fumigatus* is required for its tolerance to gliotoxin and affects the amount of extracellular and intracellular gliotoxin. Medical mycology. 2014; 52(5):506–18. Epub 2014/05/23. https://doi.org/10.1093/mmy/myu007 PMID: 24847038.

25. Derntl C, Rassinger A, Srebotnik E, Mach RL, Mach-Aigner AR. Identification of the Main Regulator Responsible for Synthesis of the Typical Yellow Pigment Produced by *Trichoderma reesei*. Applied and environmental microbiology. 2016; 82(20):6247–57. https://doi.org/10.1128/AEM.01408-16 PMID: 27520818.

26. Schrettl M, Carberry S, Kavanagh K, Haas H, Jones GW, O'Brien J, et al. Self-protection against gliotoxin—a component of the gliotoxin biosynthetic cluster, GliT, completely protects Aspergillus fumigatus against exogenous gliotoxin. PLoS pathogens. 2010; 6(6):e1000952. https://doi.org/10.1371/journal.ppat.1000952 PMID: 20548963; PubMed Central PMCID: PMC2883607.

27. Anyaogu DC, Mortensen UH. Heterologous production of fungal secondary metabolites in *Aspergilli*. Frontiers in microbiology. 2015; 6(77). https://doi.org/10.3389/fmicb.2015.00077 PMID: 25713568

28. Chang PK, Yu J, Yu JH. aflT, a MFS transporter-encoding gene located in the aflatoxin gene cluster, does not have a significant role in aflatoxin secretion. Fungal genetics and biology: FG & B. 2004; 41 (10):911–20. Epub 2004/09/03. https://doi.org/10.1016/j.fgb.2004.06.007 PMID: 15341913.

29. Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee SY, et al. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. Nucleic Acids Res. 2019; 47(W1):W81–w7. Epub 2019/04/30. https://doi.org/10.1093/nar/gkz310 PMID: 31032519; PubMed Central PMCID: PMC6602434.

30. Wolf T, Shelest V, Nath N, Shelest E. CASSIS and SMIPS: promoter-based prediction of secondary metabolite gene clusters in eukaryotic genomes. Bioinformatics. 2016; 32(8):1138–43. Epub 2015/12/15. https://doi.org/10.1093/bioinformatics/btv713 PMID: 26656005; PubMed Central PMCID: PMC4824125.

31. Khaldi N, Seifuddin FT, Turner G, Haft D, Nierman WC, Wolfe KH, et al. SMURF: Genomic mapping of fungal secondary metabolite clusters. Fungal genetics and biology: FG & B. 2010; 47(9):736–41. https://doi.org/10.1016/j.fgb.2010.06.003 PMID: 20554054; PubMed Central PMCID: PMC2916752.

32. Almeida H, Palys S, Tsang A, Diallo AB. TOUCAN: a framework for fungal biosynthetic gene cluster discovery. NAR Genomics and Bioinformatics. 2020; 2(4). https://doi.org/10.1093/nargab/lqaa098 PMID: 33575642

33. Hannigan GD, Prihoda D, Palicka A, Soukup J, Klempir O, Rampula L, et al. A deep learning genome-mining strategy for biosynthetic gene cluster prediction. Nucleic acids research. 2019; 47(18):e110. https://doi.org/10.1093/nar/gkz654 PMID: 31400112; PubMed Central PMCID: PMC6765103.

34. Tai Y, Liu C, Yu S, Yang H, Sun J, Guo C, et al. Gene co-expression network analysis reveals coordinated regulation of three characteristic secondary biosynthetic pathways in tea plant (*Camellia sinensis*). BMC Genomics. 2018; 19(1):616. Epub 2018/08/17. https://doi.org/10.1186/s12864-018-4999-9 PMID: 30111282; PubMed Central PMCID: PMC6094456.

35. Lind AL, Wisecaver JH, Lameiras C, Wiemann P, Palmer JM, Keller NP, et al. Drivers of genetic diversity in secondary metabolic gene clusters within a fungal species. PLOS Biology. 2017; 15(11): e2003583. https://doi.org/10.1371/journal.pbio.2003583 PMID: 29149178

36. Rokas A, Wisecaver JH, Lind AL. The birth, evolution and death of metabolic gene clusters in fungi. Nature reviews Microbiology. 2018; 16(12):731–44. Epub 2018/09/09. https://doi.org/10.1038/s41579-018-0075-3 PMID: 30194403.

37. Palmer JM, Keller NP. Secondary metabolism in fungi: does chromosomal location matter? Current opinion in microbiology. 2010; 13(4):431–6. Epub 2010/07/16. https://doi.org/10.1016/j.mib.2010.04.008 PMID: 20627806; PubMed Central PMCID: PMC2922032.

38. Hoogendoorn K, Barra L, Waalwijk C, Dickschat JS, van der Lee TAJ, Medema MH. Evolution and Diversity of Biosynthetic Gene Clusters in *Fusarium*. Frontiers in microbiology. 2018; 9:1158. Epub 2018/06/21. https://doi.org/10.3389/fmicb.2018.01158 PMID: 29922257; PubMed Central PMCID: PMC5996196.

39. Fischbach MA, Walsh CT, Clardy J. The evolution of gene collectives: How natural selection drives chemical innovation. Proceedings of the National Academy of Sciences. 2008; 105(12):4601–8. https://doi.org/10.1073/pnas.0709132105 PMID: 18216259

40. Medema MH, Cimermancic P, Sali A, Takano E, Fischbach MA. A Systematic Computational Analysis of Biosynthetic Gene Cluster Evolution: Lessons for Engineering Biosynthesis. PLOS Computational Biology. 2014; 10(12):e1004016. https://doi.org/10.1371/journal.pcbi.1004016 PMID: 25474254

41. Rokas A, Mead ME, Steenwyk JL, Raja HA, Oberlies NH. Biosynthetic gene clusters and the evolution of fungal chemodiversity. Natural product reports. 2020; 37(7):868–78. https://doi.org/10.1039/c9np00045c PMID: 31898704

42. Ochoa D, Pazos F. Practical aspects of protein co-evolution. Frontiers in Cell and Developmental Biology. 2014; 2(14). https://doi.org/10.3389/fcell.2014.00014 PMID: 25364721

43. Del Carratore F, Zych K, Cummings M, Takano E, Medema MH, Breitling R. Computational identification of co-evolving multi-gene modules in microbial biosynthetic gene clusters. Communications Biology. 2019; 2(1):83. https://doi.org/10.1038/s42003-019-0333-6 PMID: 30854475

44. Nordberg H, Cantor M, Dusheyko S, Hua S, Poliakov A, Shabalov I, et al. The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. Nucleic Acids Res. 2014; 42(Database issue):D26–31. Epub 2013/11/15. https://doi.org/10.1093/nar/gkt1069 PMID: 24225321; PubMed Central PMCID: PMC3965075.

45. Marcet-Houben M, Gabaldon T. TreeKO: a duplication-aware algorithm for the comparison of phylogenetic trees. Nucleic Acids Res. 2011; 39(10):e66. Epub 2011/02/22. https://doi.org/10.1093/nar/gkr087 PMID: 21335609; PubMed Central PMCID: PMC3105381.

46. Terfehr D, Dahlmann TA, Specht T, Zadra I, Kurnsteiner H, Kuck U. Genome Sequence and Annotation of *Acremonium chrysogenum*, Producer of the beta-Lactam Antibiotic Cephalosporin C. Genome announcements. 2014; 2(5). https://doi.org/10.1128/genomeA.00948-14 PMID: 25291769; PubMed Central PMCID: PMC4175204.

47. Nguyen HD, Lewis CT, Lévesque CA, Gräfenhan T. Draft Genome Sequence of *Alternaria alternata* ATCC 34957. Genome announcements. 2016; 4(1). Epub 2016/01/16. https://doi.org/10.1128/genomeA.01554-15 PMID: 26769939; PubMed Central PMCID: PMC4714121.

48. Hu J, Chen C, Peever T, Dang H, Lawrence C, Mitchell T. Genomic characterization of the conditionally dispensable chromosome in *Alternaria arborescens* provides evidence for horizontal gene transfer. BMC Genomics. 2012; 13(1):171. https://doi.org/10.1186/1471-2164-13-171 PMID: 22559316

49. Armitage AD, Cockerton HM, Sreenivasaprasad S, Woodhall J, Lane CR, Harrison RJ, et al. Genomics Evolutionary History and Diagnostics of the *Alternaria alternata* Species Group Including Apple and Asian Pear Pathotypes. Frontiers in microbiology. 2020; 10(3124). https://doi.org/10.3389/fmicb.2019.03124 PMID: 32038562

50. Lu Y, Ye C, Che J, Xu X, Shao D, Jiang C, et al. Genomic sequencing, genome-scale metabolic network reconstruction, and in silico flux analysis of the grape endophytic fungus *Alternaria* sp. MG1. Microb Cell Fact. 2019; 18(1):13. Epub 2019/01/27. https://doi.org/10.1186/s12934-019-1063-7 PMID: 30678677; PubMed Central PMCID: PMC6345013.

51. Kohler A, Kuo A, Nagy LG, Morin E, Barry KW, Buscot F, et al. Convergent losses of decay mechanisms and rapid turnover of symbiosis genes in mycorrhizal mutualists. Nat Genet. 2015; 47(4):410–5. Epub 2015/02/24. https://doi.org/10.1038/ng.3223 PMID: 25706625.

52. Martino E, Morin E, Grelet GA, Kuo A, Kohler A, Daghino S, et al. Comparative genomics and transcriptomics depict ericoid mycorrhizal fungi as versatile saprotrophs and plant mutualists. New Phytol. 2018; 217(3):1213–29. Epub 2018/01/10. https://doi.org/10.1111/nph.14974 PMID: 29315638.

53. Yang J, Wang L, Ji X, Feng Y, Li X, Zou C, et al. Genomic and proteomic analyses of the fungus *Arthrobotrys oligospora* provide insights into nematode-trap formation. PLoS Pathog. 2011; 7(9): e1002179. Epub 2011/09/13. https://doi.org/10.1371/journal.ppat.1002179 PMID: 21909256; PubMed Central PMCID: PMC3164635.

54. Burmester A, Shelest E, Glöckner G, Heddergott C, Schindler S, Staib P, et al. Comparative and functional genomics provide insights into the pathogenicity of dermatophytic fungi. Genome Biol. 2011; 12 (1):R7. Epub 2011/01/21. https://doi.org/10.1186/gb-2011-12-1-r7 PMID: 21247460; PubMed Central PMCID: PMC3091305.

55. Murat C, Payen T, Noel B, Kuo A, Morin E, Chen J, et al. Pezizomycetes genomes reveal the molecular basis of ectomycorrhizal truffle lifestyle. Nat Ecol Evol. 2018; 2(12):1956–65. Epub 2018/11/14. https://doi.org/10.1038/s41559-018-0710-4 PMID: 30420746.

56. de Vries RP, Riley R, Wiebenga A, Aguilar-Osorio G, Amillis S, Uchima CA, et al. Comparative genomics reveals high biological diversity and specific adaptations in the industrially and medically important fungal genus *Aspergillus*. Genome Biology. 2017; 18(1):28. https://doi.org/10.1186/s13059-017-1151-0 PMID: 28196534

57. Nierman WC, Yu J, Fedorova-Abrams ND, Losada L, Cleveland TE, Bhatnagar D, et al. Genome Sequence of *Aspergillus flavus* NRRL 3357, a Strain That Causes Aflatoxin Contamination of Food and Feed. Genome announcements. 2015; 3(2). https://doi.org/10.1128/genomeA.00168-15 PMID: 25883274; PubMed Central PMCID: PMC4400417.

58. Nierman WC, Pain A, Anderson MJ, Wortman JR, Kim HS, Arroyo J, et al. Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*. Nature. 2005; 438(7071):1151–6. Epub 2005/12/24. https://doi.org/10.1038/nature04332 PMID: 16372009.

59. Pel HJ, de Winde JH, Archer DB, Dyer PS, Hofmann G, Schaap PJ, et al. Genome sequencing and analysis of the versatile cell factory *Aspergillus niger* CBS 513.88. Nature biotechnology. 2007; 25 (2):221–31. https://doi.org/10.1038/nbt1282 PMID: 17259976.

60. Zhao G, Yao Y, Qi W, Wang C, Hou L, Zeng B, et al. Draft genome sequence of *Aspergillus oryzae* strain 3.042. Eukaryotic cell. 2012; 11(9):1178–. https://doi.org/10.1128/EC.00160-12 PMID: 22933657.

61. Vesth TC, Nybo JL, Theobald S, Frisvad JC, Larsen TO, Nielsen KF, et al. Investigation of inter- and intraspecies variation through genome sequencing of *Aspergillus* section *Nigri*. Nature Genetics. 2018; 50(12):1688–95. https://doi.org/10.1038/s41588-018-0246-1 PMID: 30349117

62. Savitha J, Bhargavi SD, Praveen VK. Complete Genome Sequence of Soil Fungus *Aspergillus terreus* (KM017963), a Potent Lovastatin Producer. Genome Announc. 2016; 4(3). Epub 2016/06/11. https://doi.org/10.1128/genomeA.00491-16 PMID: 27284150; PubMed Central PMCID: PMC4901219.

63. Spanu PD, Abbott JC, Amselem J, Burgis TA, Soanes DM, Stüber K, et al. Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. Science. 2010; 330 (6010):1543–6. Epub 2010/12/15. https://doi.org/10.1126/science.1194573 PMID: 21148392.

64. Marsberg A, Kemler M, Jami F, Nagel JH, Postma-Smidt A, Naidoo S, et al. *Botryosphaeria dothidea*: a latent pathogen of global importance to woody plant health. Mol Plant Pathol. 2017; 18(4):477–88. Epub 2016/09/30. https://doi.org/10.1111/mpp.12495 PMID: 27682468; PubMed Central PMCID: PMC6638292.

65. Van Kan JA, Stassen JH, Mosbach A, Van Der Lee TA, Faino L, Farmer AD, et al. A gapless genome sequence of the fungus *Botrytis cinerea*. Mol Plant Pathol. 2017; 18(1):75–89. Epub 2016/02/26. https://doi.org/10.1111/mpp.12384 PMID: 26913498; PubMed Central PMCID: PMC6638203.

66. Valero-Jiménez CA, Veloso J, Staats M, van Kan JAL. Comparative genomics of plant pathogenic *Botrytis* species with distinct host specificity. BMC Genomics. 2019; 20(1):203. https://doi.org/10.1186/s12864-019-5580-x PMID: 30866801

67. Knapp DG, Németh JB, Barry K, Hainaut M, Henrissat B, Johnson J, et al. Comparative genomics provides insights into the lifestyle and reveals functional heterogeneity of dark septate endophytic fungi. Sci Rep. 2018; 8(1):6321. Epub 2018/04/22. https://doi.org/10.1038/s41598-018-24686-4 PMID: 29679020; PubMed Central PMCID: PMC5910433.

68. Teixeira MM, Moreno LF, Stielow BJ, Muszewska A, Hainaut M, Gonzaga L, et al. Exploring the genomic diversity of black yeasts and relatives (*Chaetothyriales, Ascomycota*). Stud Mycol. 2017; 86:1–28.

Epub 2017/03/30. https://doi.org/10.1016/j.simyco.2017.01.001 PMID: 28348446; PubMed Central PMCID: PMC5358931.

69. Cuomo CA, Untereiner WA, Ma LJ, Grabherr M, Birren BW. Draft Genome Sequence of the Cellulolytic Fungus *Chaetomium globosum*. Genome announcements. 2015; 3(1). https://doi.org/10.1128/genomeA.00021-15 PMID: 25720678; PubMed Central PMCID: PMC4342419.

70. Armaleo D, Müller O, Lutzoni F, Andrésson Ó S, Blanc G, Bode HB, et al. The lichen symbiosis re-viewed through the genomes of *Cladonia grayi* and its algal partner *Asterochloris glomerata*. BMC Genomics. 2019; 20(1):605. Epub 2019/07/25. https://doi.org/10.1186/s12864-019-5629-x PMID: 31337355; PubMed Central PMCID: PMC6652019.

71. Ohm RA, Feau N, Henrissat B, Schoch CL, Horwitz BA, Barry KW, et al. Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen *Dothideomycetes* fungi. PLoS pathogens. 2012; 8(12):e1003037. https://doi.org/10.1371/journal.ppat.1003037 PMID: 23236275; PubMed Central PMCID: PMC3516569.

72. Condon BJ, Leng Y, Wu D, Bushley KE, Ohm RA, Otillar R, et al. Comparative genome structure, secondary metabolite, and effector coding capacity across *Cochliobolus* pathogens. PLoS Genet. 2013; 9 (1):e1003233. Epub 2013/01/30. https://doi.org/10.1371/journal.pgen.1003233 PMID: 23357949; PubMed Central PMCID: PMC3554632.

73. Baroncelli R, Amby DB, Zapparata A, Sarrocco S, Vannacci G, Le Floch G, et al. Gene family expansions and contractions are associated with host range in plant pathogens of the genus *Colletotrichum*. BMC Genomics. 2016; 17:555. Epub 2016/08/09. https://doi.org/10.1186/s12864-016-2917-6 PMID: 27496087; PubMed Central PMCID: PMC4974774.

74. Baroncelli R, Sanz-Martin JM, Rech GE, Sukno SA, Thon MR. Draft Genome Sequence of *Colletotrichum sublineola*, a Destructive Pathogen of Cultivated Sorghum. Genome announcements. 2014; 2 (3). https://doi.org/10.1128/genomeA.00540-14 PMID: 24926053; PubMed Central PMCID: PMC4056296.

75. Hacquard S, Kracher B, Hiruma K, Münch PC, Garrido-Oter R, Thon MR, et al. Survival trade-offs in plant roots during colonization by closely related beneficial and pathogenic fungi. Nat Commun. 2016; 7:11362. Epub 2016/05/07. https://doi.org/10.1038/ncomms11362 PMID: 27150427; PubMed Central PMCID: PMC4859067.

76. Lopez D, Ribeiro S, Label P, Fumanal B, Venisse JS, Kohler A, et al. Genome-Wide Analysis of *Corynespora cassiicola* Leaf Fall Disease Putative Effectors. Front Microbiol. 2018; 9:276. Epub 2018/03/20. https://doi.org/10.3389/fmicb.2018.00276 PMID: 29551995; PubMed Central PMCID: PMC5840194.

77. Wu W, Davis RW, Tran-Gyamfi MB, Kuo A, LaButti K, Mihaltcheva S, et al. Characterization of four endophytic fungi as potential consolidated bioprocessing hosts for conversion of lignocellulose into advanced biofuels. Appl Microbiol Biotechnol. 2017; 101(6):2603–18. Epub 2017/01/13. https://doi.org/10.1007/s00253-017-8091-1 PMID: 28078400.

78. Morales-Cruz A, Amrine KC, Blanco-Ulate B, Lawrence DP, Travadon R, Rolshausen PE, et al. Distinctive expansion of gene families associated with plant cell wall degradation, secondary metabolism, and nutrient uptake in the genomes of grapevine trunk pathogens. BMC Genomics. 2015; 16(1):469. Epub 2015/06/19. https://doi.org/10.1186/s12864-015-1624-z PMID: 26084502; PubMed Central PMCID: PMC4472170.

79. Jones L, Riaz S, Morales-Cruz A, Amrine KC, McGuire B, Gubler WD, et al. Adaptive genomic structural variation in the grape powdery mildew pathogen, *Erysiphe necator*. BMC Genomics. 2014; 15 (1):1081. Epub 2014/12/10. https://doi.org/10.1186/1471-2164-15-1081 PMID: 25487071; PubMed Central PMCID: PMC4298948.

80. Blanco-Ulate B, Rolshausen PE, Cantu D. Draft Genome Sequence of the Grapevine Dieback Fungus *Eutypa lata* UCR-EL1. Genome announcements. 2013; 1(3). Epub 2013/06/01. https://doi.org/10.1128/genomeA.00228-13 PMID: 23723393; PubMed Central PMCID: PMC3668001.

81. Bombassaro A, de Hoog S, Weiss VA, Souza EM, Leão AC, Costa FF, et al. Draft Genome Sequence of *Fonsecaea monophora* Strain CBS 269.37, an Agent of Human Chromoblastomycosis. Genome Announc. 2016; 4(4). Epub 2016/07/30. https://doi.org/10.1128/genomeA.00731-16 PMID: 27469960; PubMed Central PMCID: PMC4966464.

82. Bashyal BM, Rawat K, Sharma S, Kulshreshtha D, Gopala Krishnan S, Singh AK, et al. Whole Genome Sequencing of *Fusarium fujikuroi* Provides Insight into the Role of Secretory Proteins and Cell Wall Degrading Enzymes in Causing Bakanae Disease of Rice. Front Plant Sci. 2017; 8:2013-. https://doi.org/10.3389/fpls.2017.02013 PMID: 29230233.

83. Cuomo CA, Guldener U, Xu JR, Trail F, Turgeon BG, Di Pietro A, et al. The *Fusarium graminearum* genome reveals a link between localized polymorphism and pathogen specialization. Science. 2007; 317(5843):1400–2. Epub 2007/09/08. 317/5843/1400 [pii] https://doi.org/10.1126/science.1143708 PMID: 17823352.

84. Ma LJ, van der Does HC, Borkovich KA, Coleman JJ, Daboussi MJ, Di Pietro A, et al. Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. Nature. 2010; 464(7287):367–73. Epub 2010/03/20. https://doi.org/10.1038/nature08850 PMID: 20237561; PubMed Central PMCID: PMC3048781.

85. Niehaus EM, Münsterkötter M, Proctor RH, Brown DW, Sharon A, Idan Y, et al. Comparative "Omics" of the *Fusarium fujikuroi* Species Complex Highlights Differences in Genetic Potential and Metabolite Synthesis. Genome Biol Evol. 2016; 8(11):3574–99. Epub 2017/01/04. https://doi.org/10.1093/gbe/evw259 PMID: 28040774; PubMed Central PMCID: PMC5203792.

86. Gardiner DM, Benfield AH, Stiller J, Stephen S, Aitken K, Liu C, et al. A high-resolution genetic map of the cereal crown rot pathogen *Fusarium pseudograminearum* provides a near-complete genome assembly. Mol Plant Pathol. 2018; 19(1):217–26. Epub 2016/11/27. https://doi.org/10.1111/mpp.12519 PMID: 27888554; PubMed Central PMCID: PMC6638115.

87. Okagaki LH, Nunes CC, Sailsbery J, Clay B, Brown D, John T, et al. Genome Sequences of Three Phytopathogenic Species of the *Magnaporthaceae* Family of Fungi. G3 (Bethesda). 2015; 5 (12):2539–45. Epub 2015/09/30. https://doi.org/10.1534/g3.115.020057 PMID: 26416668; PubMed Central PMCID: PMC4683626.

88. Peter M, Kohler A, Ohm RA, Kuo A, Krützmann J, Morin E, et al. Ectomycorrhizal ecology is imprinted in the genome of the dominant symbiotic fungus *Cenococcum geophilum*. Nat Commun. 2016; 7:12662. Epub 2016/09/08. https://doi.org/10.1038/ncomms12662 PMID: 27601008; PubMed Central PMCID: PMC5023957.

89. Dean RA, Talbot NJ, Ebbole DJ, Farman ML, Mitchell TK, Orbach MJ, et al. The genome sequence of the rice blast fungus *Magnaporthe grisea*. Nature. 2005; 434(7036):980–6. Epub 2005/04/23. https://doi.org/10.1038/nature03449 PMID: 15846337.

90. Gao Q, Jin K, Ying SH, Zhang Y, Xiao G, Shang Y, et al. Genome sequencing and comparative transcriptomics of the model entomopathogenic fungi *Metarhizium anisopliae* and *M. acridum*. PLoS Genet. 2011; 7(1):e1001264. Epub 2011/01/22. https://doi.org/10.1371/journal.pgen.1001264 PMID: 21253567; PubMed Central PMCID: PMC3017113.

91. Hu X, Xiao G, Zheng P, Shang Y, Su Y, Zhang X, et al. Trajectory and genomic determinants of fungal-pathogen speciation and host adaptation. Proc Natl Acad Sci U S A. 2014; 111(47):16796–801. Epub 2014/11/05. https://doi.org/10.1073/pnas.1412662111 PMID: 25368161; PubMed Central PMCID: PMC4250126.

92. Binneck E, Lastra CCL, Sosa-Gómez DR. Genome Sequence of *Metarhizium rileyi*, a Microbial Control Agent for Lepidoptera. Microbiology Resource Announcements. 2019; 8(36):e00897–19. https://doi.org/10.1128/MRA.00897-19 PMID: 31488537

93. Meerupati T, Andersson K-M, Friman E, Kumar D, Tunlid A, Ahrén D. Genomic Mechanisms Accounting for the Adaptation to Parasitism in Nematode-Trapping Fungi. PLOS Genetics. 2013; 9(11): e1003909. https://doi.org/10.1371/journal.pgen.1003909 PMID: 24244185

94. Tan H, Kohler A, Miao R, Liu T, Zhang Q, Zhang B, et al. Multi-omic analyses of exogenous nutrient bag decomposition by the black morel *Morchella importuna* reveal sustained carbon acquisition and transferring. Environ Microbiol. 2019; 21(10):3909–26. Epub 2019/07/18. https://doi.org/10.1111/1462-2920.14741 PMID: 31314937.

95. Coleman JJ, Rounsley SD, Rodriguez-Carres M, Kuo A, Wasmann CC, Grimwood J, et al. The genome of *Nectria haematococca*: contribution of supernumerary chromosomes to gene expansion. PLoS Genet. 2009; 5(8):e1000618. Epub 2009/08/29. https://doi.org/10.1371/journal.pgen.1000618 PMID: 19714214; PubMed Central PMCID: PMC2725324.

96. Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, et al. The genome sequence of the filamentous fungus *Neurospora crassa*. Nature. 2003; 422(6934):859–68. Epub 2003/04/25. https://doi.org/10.1038/nature01554 PMID: 12712197.

97. Baker SE, Schackwitz W, Lipzen A, Martin J, Haridas S, LaButti K, et al. Draft Genome Sequence of Neurospora crassa Strain FGSC 73. Genome announcements. 2015; 3(2):e00074–15. https://doi.org/10.1128/genomeA.00074-15 PMID: 25838471.

98. Ellison CE, Stajich JE, Jacobson DJ, Natvig DO, Lapidus A, Foster B, et al. Massive changes in genome architecture accompany the transition to self-fertility in the filamentous fungus *Neurospora tetrasperma*. Genetics. 2011; 189(1):55–69. Epub 2011/07/14. https://doi.org/10.1534/genetics.111.130690 PMID: 21750257; PubMed Central PMCID: PMC3176108.

99. Haridas S, Wang Y, Lim L, Massoumi Alamouti S, Jackman S, Docking R, et al. The genome and transcriptome of the pine saprophyte *Ophiostoma piceae*, and a comparison with the bark beetle-associated pine pathogen *Grosmannia clavigera*. BMC Genomics. 2013; 14:373. Epub 2013/06/04. https://doi.org/10.1186/1471-2164-14-373 PMID: 23725015; PubMed Central PMCID: PMC3680317.

100. Urquhart AS, Mondo SJ, Mäkelä MR, Hane JK, Wiebenga A, He G, et al. Genomic and Genetic Insights Into a Cosmopolitan Fungus, *Paecilomyces variotii* (*Eurotiales*). Front Microbiol. 2018; 9:3058. Epub 2019/01/09. https://doi.org/10.3389/fmicb.2018.03058 PMID: 30619145; PubMed Central PMCID: PMC6300479.

101. Reynolds HT, Vijayakumar V, Gluck-Thaler E, Korotkin HB, Matheny PB, Slot JC. Horizontal gene cluster transfer increased hallucinogenic mushroom diversity. Evolution Letters. 2018; 2(2):88–101. https://doi.org/10.1002/evl3.42 PMID: 30283667

102. Desjardins CA, Champion MD, Holder JW, Muszewska A, Goldberg J, Bailão AM, et al. Comparative genomic analysis of human fungal pathogens causing paracoccidioidomycosis. PLoS Genet. 2011; 7 (10):e1002345. Epub 2011/11/03. https://doi.org/10.1371/journal.pgen.1002345 PMID: 22046142; PubMed Central PMCID: PMC3203195.

103. Cheeseman K, Ropars J, Renault P, Dupont J, Gouzy J, Branca A, et al. Multiple recent horizontal transfers of a large genomic region in cheese making fungi. Nat Commun. 2014; 5:2876. Epub 2014/01/11. https://doi.org/10.1038/ncomms3876 PubMed Central PMCID: PMC3896755. PMID: 24407037

104. Specht T, Dahlmann TA, Zadra I, Kürnsteiner H, Kück U. Complete Sequencing and Chromosome-Scale Genome Assembly of the Industrial Progenitor Strain P2niaD18 from the *Penicillin Producer* Penicillium chrysogenum. Genome announcements. 2014; 2(4). Epub 2014/07/26. https://doi.org/10.1128/genomeA.00577-14 PMID: 25059858; PubMed Central PMCID: PMC4110216.

105. Marcet-Houben M, Ballester A-R, de la Fuente B, Harries E, Marcos JF, González-Candelas L, et al. Genome sequence of the necrotrophic fungus *Penicillium digitatum*, the main postharvest pathogen of citrus. BMC Genomics. 2012; 13(1):646. https://doi.org/10.1186/1471-2164-13-646 PMID: 23171342

106. Ballester AR, Marcet-Houben M, Levin E, Sela N, Selma-Lázaro C, Carmona L, et al. Genome, Transcriptome, and Functional Analyses of *Penicillium expansum* Provide New Insights Into Secondary Metabolism and Pathogenicity. Mol Plant Microbe Interact. 2015; 28(3):232–48. Epub 2014/10/23. https://doi.org/10.1094/MPMI-09-14-0261-FI PMID: 25338147.

107. Nielsen JC, Grijseels S, Prigent S, Ji B, Dainat J, Nielsen KF, et al. Global analysis of biosynthetic gene clusters reveals vast potential of secondary metabolite production in Penicillium species. Nat Microbiol. 2017; 2:17044. Epub 2017/04/04. https://doi.org/10.1038/nmicrobiol.2017.44 PMID: 28368369.

108. Liu G, Zhang L, Wei X, Zou G, Qin Y, Ma L, et al. Genomic and Secretomic Analyses Reveal Unique Features of the Lignocellulolytic Enzyme System of *Penicillium decumbens*. PloS one. 2013; 8(2): e55185. https://doi.org/10.1371/journal.pone.0055185 PMID: 23383313

109. van den Berg MA, Albang R, Albermann K, Badger JH, Daran JM, Driessen AJ, et al. Genome sequencing and analysis of the filamentous fungus *Penicillium chrysogenum*. Nature biotechnology. 2008; 26(10):1161–8. https://doi.org/10.1038/nbt.1498 PMID: 18820685.

110. Wang X, Zhang X, Liu L, Xiang M, Wang W, Sun X, et al. Genomic and transcriptomic analysis of the endophytic fungus *Pestalotiopsis fici* reveals its lifestyle and high potential for synthesis of natural products. BMC Genomics. 2015; 16(1):28. Epub 2015/01/28. https://doi.org/10.1186/s12864-014-1190-9 PMID: 25623211; PubMed Central PMCID: PMC4320822.

111. Blanco-Ulate B, Rolshausen P, Cantu D. Draft Genome Sequence of the Ascomycete *Phaeoacremonium aleophilum* Strain UCR-PA7, a Causal Agent of the Esca Disease Complex in Grapevines. Genome announcements. 2013; 1(3). Epub 2013/07/03. https://doi.org/10.1128/genomeA.00390-13 PMID: 23814032; PubMed Central PMCID: PMC3695428.

112. Walker AK, Frasz SL, Seifert KA, Miller JD, Mondo SJ, LaButti K, et al. Full Genome of *Phialocephala scopiformis* DAOMC 229536, a Fungal Endophyte of Spruce Producing the Potent Anti-Insectan Compound Rugulosin. Genome announcements. 2016; 4(1). Epub 2016/03/08. https://doi.org/10.1128/genomeA.01768-15 PMID: 26950333; PubMed Central PMCID: PMC4767923.

113. Cissé OH, Pagni M, Hauser PM. *De novo* assembly of the *Pneumocystis jirovecii* genome from a single bronchoalveolar lavage fluid specimen from a patient. mBio. 2012; 4(1):e00428–12. Epub 2012/12/28. https://doi.org/10.1128/mBio.00428-12 PMID: 23269827; PubMed Central PMCID: PMC3531804.

114. Chibucos MC, Crabtree J, Nagaraj S, Chaturvedi S, Chaturvedi V. Draft Genome Sequences of Human Pathogenic Fungus *Geomyces pannorum* Sensu Lato and Bat White Nose Syndrome Pathogen *Geomyces* (*Pseudogymnoascus*) *destructans*. Genome announcements. 2013; 1(6). Epub 2013/12/21. https://doi.org/10.1128/genomeA.01045-13 PMID: 24356829; PubMed Central PMCID: PMC3868853.

115. Mondo SJ, Dannebaum RO, Kuo RC, Louie KB, Bewick AJ, LaButti K, et al. Widespread adenine N6-methylation of active genes in fungi. Nat Genet. 2017; 49(6):964–8. Epub 2017/05/10. https://doi.org/10.1038/ng.3859 PMID: 28481340.

116. Cubeta MA, Thomas E, Dean RA, Jabaji S, Neate SM, Tavantzis S, et al. Draft Genome Sequence of the Plant-Pathogenic Soil Fungus *Rhizoctonia solani* Anastomosis Group 3 Strain Rhs1AP. Genome Announc. 2014; 2(5). Epub 2014/11/02. https://doi.org/10.1128/genomeA.01072-14 PMID: 25359908; PubMed Central PMCID: PMC4214984.

117. Liti G, Ba ANN, Blythe M, Müller CA, Bergström A, Cubillos FA, et al. High quality *de novo* sequencing and assembly of the *Saccharomyces arboricolus* genome. BMC Genomics. 2013; 14(1):69. https://doi.org/10.1186/1471-2164-14-69 PMID: 23368932

118. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, et al. Life with 6000 genes. Science. 1996; 274(5287):546, 63–7. Epub 1996/10/25. https://doi.org/10.1126/science.274.5287.546 PMID: 8849441.

119. Quandt CA, Bushley KE, Spatafora JW. The genome of the truffle-parasite *Tolypocladium ophioglossoides* and the evolution of antifungal peptaibiotics. BMC Genomics. 2015; 16(1):553. Epub 2015/07/29. https://doi.org/10.1186/s12864-015-1777-9 PMID: 26215153; PubMed Central PMCID: PMC4517408.

120. Quandt CA, Patterson W, Spatafora JW. Harnessing the power of phylogenomics to disentangle the directionality and signatures of interkingdom host jumping in the parasitic fungal genus *Tolypocladium*. Mycologia. 2018; 110(1):104–17. Epub 2018/06/05. https://doi.org/10.1080/00275514.2018.1442618 PMID: 29863984.

121. Proctor RH, McCormick SP, Kim HS, Cardoza RE, Stanley AM, Lindo L, et al. Evolution of structural diversity of trichothecenes, a family of toxins produced by plant pathogenic and entomopathogenic fungi. PLoS Pathog. 2018; 14(4):e1006946. Epub 2018/04/13. https://doi.org/10.1371/journal.ppat.1006946 PMID: 29649280; PubMed Central PMCID: PMC5897003.

122. Druzhinina IS, Chenthamara K, Zhang J, Atanasova L, Yang D, Miao Y, et al. Massive lateral transfer of genes encoding plant cell wall-degrading enzymes to the mycoparasitic fungus *Trichoderma* from its plant-associated hosts. PLoS Genet. 2018; 14(4):e1007322. Epub 2018/04/10. https://doi.org/10.1371/journal.pgen.1007322 PMID: 29630596; PubMed Central PMCID: PMC5908196.

123. Kubicek CP, Herrera-Estrella A, Seidl-Seiboth V, Martinez DA, Druzhinina IS, Thon M, et al. Comparative genome sequence analysis underscores mycoparasitism as the ancestral life style of *Trichoderma*. Genome Biol. 2011; 12(4):R40. Epub 2011/04/20. https://doi.org/10.1186/gb-2011-12-4-r40 PMID: 21501500; PubMed Central PMCID: PMC3218866.

124. Baroncelli R, Piaggeschi G, Fiorini L, Bertolini E, Zapparata A, Pè ME, et al. Draft Whole-Genome Sequence of the Biocontrol Agent *Trichoderma harzianum* T6776. Genome announcements. 2015; 3 (3):e00647–15. https://doi.org/10.1128/genomeA.00647-15 PMID: 26067977.

125. Xie BB, Qin QL, Shi M, Chen LL, Shu YL, Luo Y, et al. Comparative genomics provide insights into evolution of trichoderma nutrition style. Genome Biol Evol. 2014; 6(2):379–90. Epub 2014/02/01. https://doi.org/10.1093/gbe/evu018 PMID: 24482532; PubMed Central PMCID: PMC3942035.

126. Martinez D, Berka RM, Henrissat B, Saloheimo M, Arvas M, Baker SE, et al. Genome sequencing and analysis of the biomass-degrading fungus *Trichoderma reesei* (syn. *Hypocrea jecorina*). Nature biotechnology. 2008; 26(5):553–60. Epub 2008/05/06. https://doi.org/10.1038/nbt1403 PMID: 18454138.

127. Martinez DA, Oliver BG, Gräser Y, Goldberg JM, Li W, Martinez-Rossi NM, et al. Comparative genome analysis of *Trichophyton rubrum* and related dermatophytes reveals candidate genes involved in infection. mBio. 2012; 3(5):e00259–12. Epub 2012/09/07. https://doi.org/10.1128/mBio.00259-12 PMID: 22951933; PubMed Central PMCID: PMC3445971.

128. Deng CH, Plummer KM, Jones DAB, Mesarich CH, Shiller J, Taranto AP, et al. Comparative analysis of the predicted secretomes of Rosaceae scab pathogens *Venturia inaequalis* and *V. pirina* reveals expanded effector families and putative determinants of host range. BMC Genomics. 2017; 18(1):339. Epub 2017/05/04. https://doi.org/10.1186/s12864-017-3699-1 PMID: 28464870; PubMed Central PMCID: PMC5412055.

129. Klosterman SJ, Subbarao KV, Kang S, Veronese P, Gold SE, Thomma BP, et al. Comparative genomics yields insights into niche adaptation of plant vascular wilt pathogens. PLoS Pathog. 2011; 7(7): e1002137. Epub 2011/08/11. https://doi.org/10.1371/journal.ppat.1002137 PMID: 21829347; PubMed Central PMCID: PMC3145793.

130. Gazis R, Kuo A, Riley R, LaButti K, Lipzen A, Lin J, et al. The genome of *Xylona heveae* provides a window into fungal endophytism. Fungal Biol. 2016; 120(1):26–42. Epub 2015/12/24. https://doi.org/10.1016/j.funbio.2015.10.002 PMID: 26693682.

131. Grandaubert J, Bhattacharyya A, Stukenbrock EH. RNA-seq-Based Gene Annotation and Comparative Genomics of Four Fungal Grass Pathogens in the Genus Zymoseptoria Identify Novel Orphan Genes and Species-Specific Invasions of Transposable Elements. G3 (Bethesda). 2015; 5(7):1323–33. Epub 2015/04/29. https://doi.org/10.1534/g3.115.017731 PMID: 25917918; PubMed Central PMCID: PMC4502367.

132. Stukenbrock EH, Christiansen FB, Hansen TT, Dutheil JY, Schierup MH. Fusion of two divergent fungal individuals led to the recent emergence of a unique widespread pathogen species. Proceedings of the National Academy of Sciences. 2012; 109(27):10954. https://doi.org/10.1073/pnas.1201403109 PMID: 22711811

133. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25 (17):3389–402. Epub 1997/09/01. https://doi.org/10.1093/nar/25.17.3389 PMID: 9254694; PubMed Central PMCID: PMC146917.

134. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic acids research. 1994; 22(22):4673–80. Epub 1994/11/11. https://doi.org/10.1093/nar/22.22.4673 PMID: 7984417; PubMed Central PMCID: PMC308517.

135. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014; 30(9):1312–3. Epub 2014/01/24. https://doi.org/10.1093/bioinformatics/btu033 PMID: 24451623; PubMed Central PMCID: PMC3998144.

136. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. 2019.

137. Kucheryavskiy S. mdatools: Multivariate Data Analysis for Chemometrics. 2019.

138. Warnes GR, Bolker B, Bonebakker L, Gentleman R, Huber W, Liaw A, et al. gplots: Various R Programming Tools for Plotting Data. 2019.

139. Wickham H, Hester J, Francois R. readr: Read Rectangular Text Data. 2018.

140. Fox J, Weisberg S. An {R} Companion to Applied Regression. Sage. 2019.

141. Murtagh F, Legendre P. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? Journal of Classification. 2014; 31(3):274–95. https://doi.org/10.1007/s00357-014-9161-z

142. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 2000; 16(6):276–7. Epub 2000/05/29. https://doi.org/10.1016/s0168-9525(00)02024-2 PMID: 10827456.

143. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE: tools for motif discovery and searching. Nucleic acids research. 2009; 37(Web Server issue):W202–8. https://doi.org/10.1093/nar/gkp335 PMID: 19458158; PubMed Central PMCID: PMC2703892.

144. Kautsar SA, Blin K, Shaw S, Navarro-Muñoz JC, Terlouw BR, van der Hooft JJJ, et al. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. Nucleic Acids Research. 2019; 48(D1): D454–D8. https://doi.org/10.1093/nar/gkz882 PMID: 31612915

145. Heneghan MN, Yakasai AA, Williams K, Kadir KA, Wasil Z, Bakeer W, et al. The programming role of trans-acting enoyl reductases during the biosynthesis of highly reduced fungal polyketides. Chemical Science. 2011; 2(5):972–9. https://doi.org/10.1039/C1SC00023C

146. Ehrlich KC, Chang PK, Yu J, Cotty PJ. Aflatoxin biosynthesis cluster gene *cypA* is required for G aflatoxin formation. Appl Environ Microbiol. 2004; 70(11):6518–24. Epub 2004/11/06. https://doi.org/10.1128/AEM.70.11.6518-6524.2004 PMID: 15528514; PubMed Central PMCID: PMC525170.

147. Bhatnagar D, Cary JW, Ehrlich K, Yu J, Cleveland TE. Understanding the genetics of regulation of aflatoxin production and *Aspergillus flavus* development. Mycopathologia. 2006; 162(3):155–66. Epub 2006/09/01. https://doi.org/10.1007/s11046-006-0050-9 PMID: 16944283.

148. Porquier A, Morgant G, Moraga J, Dalmais B, Luyten I, Simon A, et al. The botrydial biosynthetic gene cluster of Botrytis cinerea displays a bipartite genomic structure and is positively regulated by the putative Zn(II)2Cys6 transcription factor BcBot6. Fungal Genet Biol. 2016; 96:33–46. Epub 2016/10/22. https://doi.org/10.1016/j.fgb.2016.10.003 PMID: 27721016.

149. Pinedo C, Wang CM, Pradier JM, Dalmais B, Choquer M, Le Pecheur P, et al. Sesquiterpene synthase from the botrydial biosynthetic gene cluster of the phytopathogen *Botrytis cinerea*. ACS Chem Biol. 2008; 3(12):791–801. Epub 2008/11/28. https://doi.org/10.1021/cb800225v PMID: 19035644; PubMed Central PMCID: PMC2707148.

150. Hamed RB, Gomez-Castellanos JR, Henry L, Ducho C, McDonough MA, Schofield CJ. The enzymes of β-lactam biosynthesis. Natural product reports. 2013; 30(1):21–107. Epub 2012/11/09. https://doi.org/10.1039/c2np20065a PMID: 23135477.

151. Abe Y, Suzuki T, Ono C, Iwamoto K, Hosobuchi M, Yoshikawa H. Molecular cloning and characterization of an ML-236B (compactin) biosynthetic gene cluster in *Penicillium citrinum*. Mol Genet Genomics. 2002; 267(5):636–46. Epub 2002/08/13. https://doi.org/10.1007/s00438-002-0697-y PMID: 12172803.

152. Abe Y, Ono C, Hosobuchi M, Yoshikawa H. Functional analysis of mlcR, a regulatory gene for ML-236B (compactin) biosynthesis in *Penicillium citrinum*. Mol Genet Genomics. 2002; 268(3):352–61. Epub 2002/11/19. https://doi.org/10.1007/s00438-002-0755-5 PMID: 12436257.

153. Hoffmann K, Schneider-Scherzer E, Kleinkauf H, Zocher R. Purification and Characterization of Eucaryotic Alanine Racemase Acting as Key Enzyme in Cyclosporin Biosynthesis. Biological Chemistry. 1994; 269(17):12710–4. PMID: 8175682

154. Yang X, Feng P, Yin Y, Bushley K, Spatafora JW, Wang C. Cyclosporine Biosynthesis in *Tolypocladium inflatum* Benefits Fungal Adaptation to the Environment. Molecular Biology and Physiology. 2018; 9(5). https://doi.org/10.1128/mBio.01211-18 PMID: 30279281

155. Weber G, Leitner E. Disruption of the cyclosporin synthetase gene of *Tolypocladium niveum*. Current Genetics. 1994; 26:461–7. https://doi.org/10.1007/BF00309935 PMID: 7874740

156. Wang B, Kang Q, Lu Y, Bai L, Wang C. Unveiling the biosynthetic puzzle of destruxins in *Metarhizium* species. Proc Natl Acad Sci U S A. 2012; 109(4):1287–92. Epub 2012/01/11. https://doi.org/10.1073/pnas.1115983109 PMID: 22232661; PubMed Central PMCID: PMC3268274.

157. Lin H-C, Chooi Y-H, Dhingra S, Xu W, Calvo AM, Tang Y. The Fumagillin Biosynthetic Gene Cluster in *Aspergillus fumigatus* Encodes a Cryptic Terpene Cyclase Involved in the Formation of β-trans-Bergamotene. Journal of the American Chemical Society. 2013; 135(12):4616–9. https://doi.org/10.1021/ja312503y PMID: 23488861

158. Kato N, Suzuki H, Takagi H, Uramoto M, Takahashi S, Osada H. Gene disruption and biochemical characterization of verruculogen synthase of *Aspergillus fumigatus*. Chembiochem. 2011; 12(5):711–4. Epub 2011/03/16. https://doi.org/10.1002/cbic.201000562 PMID: 21404415.

159. Kato N, Suzuki H, Takagi H, Asami Y, Kakeya H, Uramoto M, et al. Identification of cytochrome P450s required for fumitremorgin biosynthesis in *Aspergillus fumigatus*. Chembiochem. 2009; 10(5):920–8. Epub 2009/02/20. https://doi.org/10.1002/cbic.200800787 PMID: 19226505.

160. Maiya S, Grundmann A, Li SM, Turner G. Improved tryprostatin B production by heterologous gene expression in *Aspergillus nidulans*. Fungal Genet Biol. 2009; 46(5):436–40. Epub 2009/04/18. https://doi.org/10.1016/j.fgb.2009.01.003 PMID: 19373974.

161. Kato N, Suzuki H, Okumura H, Takahashi S, Osada H. A point mutation in *ftmD* blocks the fumitremorgin biosynthetic pathway in *Aspergillus fumigatus* strain Af293. Biosci Biotechnol Biochem. 2013; 77 (5):1061–7. Epub 2013/05/08. https://doi.org/10.1271/bbb.130026 PMID: 23649274.

162. Proctor RH, Busman M, Seo JA, Lee YW, Plattner RD. A fumonisin biosynthetic gene cluster in *Fusarium oxysporum* strain O-1890 and the genetic basis for B versus C fumonisin production. Fungal genetics and biology: FG & B. 2008; 45(6):1016–26. Epub 2008/04/01. https://doi.org/10.1016/j.fgb.2008.02.004 PMID: 18375156.

163. Zaleta-Rivera K, Xu C, Yu F, Butchko RA, Proctor RH, Hidalgo-Lara ME, et al. A Bidomain Nonribosomal Peptide Synthetase Encoded by FUM14 Catalyzes the Formation of Tricarballylic Esters in the Biosynthesis of Fumonisins. Biochemistry 2006; 45:2561–9. https://doi.org/10.1021/bi052085s PMID: 16489749

164. Butchko RA, Plattner RD, Proctor RH. Deletion Analysis of FUM Genes Involved in Tricarballylic Ester Formation during Fumonisin Biosynthesis. J Agric Food Chem. 2006; 54:9398−404. https://doi.org/10.1021/jf0617869 PMID: 17147424

165. Butchko RA, Plattner RD, Proctor RH. FUM13 Encodes a Short Chain Dehydrogenase/Reductase Required for C-3 Carbonyl Reduction during Fumonisin Biosynthesis in *Gibberella moniliformis*. J Agric Food Chem 2003; 51:3000−6. https://doi.org/10.1021/jf0262007 PMID: 12720383

166. Butchko RA, Plattner RD, Proctor RH. FUM9 is required for C-5 hydroxylation of fumonisins and complements the meitotically defined Fum3 locus in *Gibberella moniliformis*. Appl Environ Microbiol. 2003; 69(11):6935–7. Epub 2003/11/07. https://doi.org/10.1128/AEM.69.11.6935-6937.2003 PMID: 14602658; PubMed Central PMCID: PMC262316.

167. Brown DW, Butchko RA, Busman M, Proctor RH. The *Fusarium verticillioides* FUM gene cluster encodes a Zn(II)2Cys6 protein that affects FUM gene expression and fumonisin production. Eukaryot Cell. 2007; 6(7):1210–8. Epub 2007/05/08. https://doi.org/10.1128/EC.00400-06 PMID: 17483290; PubMed Central PMCID: PMC1951116.

168. Du L, Zhu X, Gerber R, Huffman J, Lou L, Jorgenson J, et al. Biosynthesis of sphinganine-analog mycotoxins. J Ind Microbiol Biotechnol. 2008; 35(6):455–64. Epub 2008/01/25. https://doi.org/10.1007/s10295-008-0316-y PMID: 18214562.

169. Proctor RH, Desjardins AE, Plattner RD, Hohn TM. A Polyketide Synthase Gene Required for Biosynthesis of Fumonisin Mycotoxins in *Gibberella fujikuroi* Mating Population A. Fungal Genetics and Biology 1999; 27:100–12. https://doi.org/10.1006/fgbi.1999.1141 PMID: 10413619

170. Lia Y, Lou L, Cerny RL, Butchko RA, Proctor RH, Shen Y, et al. Tricarballylic ester formation during biosynthesis of fumonisin mycotoxins in *Fusarium verticillioides*. Mycology. 2013; 4(4):179–86. Epub

2014/03/04. https://doi.org/10.1080/21501203.2013.874540 PMID: 24587959; PubMed Central PMCID: PMC3933019.

171. Studt L, Janevska S, Niehaus EM, Burkhardt I, Arndt B, Sieber CM, et al. Two separate key enzymes and two pathway-specific transcription factors are involved in fusaric acid biosynthesis in *Fusarium fujikuroi*. Environ Microbiol. 2016; 18(3):936–56. Epub 2015/12/15. https://doi.org/10.1111/1462-2920.13150 PMID: 26662839.

172. Lin X, Yuan S, Chen S, Chen B, Xu H, Liu L, et al. Heterologous Expression of Ilicicolin H Biosynthetic Gene Cluster and Production of a New Potent Antifungal Reagent, Ilicicolin J. Molecules. 2019; 24 (12). Epub 2019/06/21. https://doi.org/10.3390/molecules24122267 PMID: 31216742; PubMed Central PMCID: PMC6631495.

173. Cary JW, Uka V, Han Z, Buyst D, Harris-Coward PY, Ehrlich KC, et al. An *Aspergillus flavus* secondary metabolic gene cluster containing a hybrid PKS-NRPS is necessary for synthesis of the 2-pyridones, leporins. Fungal Genet Biol. 2015; 81:88–97. Epub 2015/06/09. https://doi.org/10.1016/j.fgb.2015.05.010 PMID: 26051490.

174. Manzoni M, Rollini M. Biosynthesis and biotechnological production of statins by filamentous fungi and application of these cholesterol-lowering drugs. Appl Microbiol Biotechnol. 2002; 58(5):555–64. Epub 2002/04/17. https://doi.org/10.1007/s00253-002-0932-9 PMID: 11956737.

175. Zhang W, Cao S, Qiu L, Qi F, Li Z, Yang Y, et al. Functional characterization of MpaG', the O-methyltransferase involved in the biosynthesis of mycophenolic acid. Chembiochem. 2015; 16(4):565–9. Epub 2015/01/30. https://doi.org/10.1002/cbic.201402600 PMID: 25630520.

176. Zhang W, Du L, Qu Z, Zhang X, Li F, Li Z, et al. Compartmentalized biosynthesis of mycophenolic acid. Proc Natl Acad Sci U S A. 2019; 116(27):13305–10. Epub 2019/06/19. https://doi.org/10.1073/pnas.1821932116 PMID: 31209052; PubMed Central PMCID: PMC6613074.

177. Hansen BG, Genee HJ, Kaas CS, Nielsen JB, Regueira TB, Mortensen UH, et al. A new class of IMP dehydrogenase with a role in self-resistance of mycophenolic acid producing fungi. BMC Microbiol. 2011; 11:202. Epub 2011/09/20. https://doi.org/10.1186/1471-2180-11-202 PMID: 21923907; PubMed Central PMCID: PMC3184278.

178. Hansen BG, Salomonsen B, Nielsen MT, Nielsen JB, Hansen NB, Nielsen KF, et al. Versatile enzyme expression and characterization system for *Aspergillus nidulans*, with the *Penicillium brevicompactum* polyketide synthase gene from the mycophenolic acid gene cluster as a test case. Appl Environ Microbiol. 2011; 77(9):3044–51. Epub 2011/03/15. https://doi.org/10.1128/AEM.01768-10 PMID: 21398493; PubMed Central PMCID: PMC3126399.

179. Regueira TB, Kildegaard KR, Hansen BG, Mortensen UH, Hertweck C, Nielsen J. Molecular basis for mycophenolic acid biosynthesis in *Penicillium brevicompactum*. Appl Environ Microbiol. 2011; 77 (9):3035–43. Epub 2011/03/15. https://doi.org/10.1128/AEM.03015-10 PMID: 21398490; PubMed Central PMCID: PMC3126426.

180. Hansen BG, Mnich E, Nielsen KF, Nielsen JB, Nielsen MT, Mortensen UH, et al. Involvement of a natural fusion of a cytochrome P450 and a hydrolase in mycophenolic acid biosynthesis. Appl Environ Microbiol. 2012; 78(14):4908–13. Epub 2012/05/01. https://doi.org/10.1128/AEM.07955-11 PMID: 22544261; PubMed Central PMCID: PMC3416377.

181. Del-Cid A, Gil-Duran C, Vaca I, Rojas-Aedo JF, Garcia-Rico RO, Levican G, et al. Identification and Functional Analysis of the Mycophenolic Acid Gene Cluster of *Penicillium roqueforti*. PLoS One. 2016; 11(1):e0147047. Epub 2016/01/12. https://doi.org/10.1371/journal.pone.0147047 PMID: 26751579; PubMed Central PMCID: PMC4708987.

182. Gillot G, Jany JL, Dominguez-Santos R, Poirier E, Debaets S, Hidalgo PI, et al. Genetic basis for mycophenolic acid production and strain-dependent production variability in *Penicillium roqueforti*. Food Microbiol. 2017; 62:239–50. Epub 2016/11/28. https://doi.org/10.1016/j.fm.2016.10.013 PMID: 27889155.

183. Scott B, Young CA, Saikia S, McMillan LK, Monahan BJ, Koulman A, et al. Deletion and gene expression analyses define the paxilline biosynthetic gene cluster in *Penicillium paxilli*. Toxins (Basel). 2013; 5(8):1422–46. Epub 2013/08/21. https://doi.org/10.3390/toxins5081422 PMID: 23949005; PubMed Central PMCID: PMC3760044.

184. Fierro F, Garcia-Estrada C, Castillo NI, Rodriguez R, Velasco-Conde T, Martin JF. Transcriptional and bioinformatic analysis of the 56.8 kb DNA region amplified in tandem repeats containing the penicillin gene cluster in *Penicillium chrysogenum*. Fungal Genet Biol. 2006; 43(9):618–29. Epub 2006/05/23. https://doi.org/10.1016/j.fgb.2006.03.001 PMID: 16713314.

185. Xu X, Liu L, Zhang F, Wang W, Li J, Guo L, et al. Identification of the first diphenyl ether gene cluster for pestheic acid biosynthesis in plant endophyte *Pestalotiopsis fici*. Chembiochem. 2014; 15(2):284–92. Epub 2013/12/05. https://doi.org/10.1002/cbic.201300626 PMID: 24302702.

186. Chen L, Yue Q, Zhang X, Xiang M, Wang C, Li S, et al. Genomics-driven discovery of the pneumocandin biosynthetic gene cluster in the fungus *Glarea lozoyensis*. BMC Genomics. 2013; 14(339). https://doi.org/10.1186/1471-2164-14-339 PMID: 23688303

187. Chen L, Li Y, Yue Q, Loksztejn A, Yokoyama K, Felix EA, et al. Engineering of New Pneumocandin Side-Chain Analogues from *Glarea lozoyensis* by Mutasynthesis and Evaluation of Their Antifungal Activity. ACS Chem Biol. 2016; 11(10):2724–33. Epub 2016/10/22. https://doi.org/10.1021/acschembio.6b00604 PMID: 27494047; PubMed Central PMCID: PMC5502478.

188. Chen L, Yue Q, Li Y, Niu X, Xiang M, Wang W, et al. Engineering of *Glarea lozoyensis* for exclusive production of the pneumocandin B0 precursor of the antifungal drug caspofungin acetate. Appl Environ Microbiol. 2015; 81(5):1550–8. Epub 2014/12/21. https://doi.org/10.1128/AEM.03256-14 PMID: 25527531; PubMed Central PMCID: PMC4325176.

189. Salo O, Guzman-Chavez F, Ries MI, Lankhorst PP, Bovenberg RAL, Vreeken RJ, et al. Identification of a Polyketide Synthase Involved in Sorbicillin Biosynthesis by *Penicillium chrysogenum*. Appl Environ Microbiol. 2016; 82(13):3971–8. Epub 2016/04/24. https://doi.org/10.1128/AEM.00350-16 PMID: 27107123; PubMed Central PMCID: PMC4907180.

190. Guzman-Chavez F, Salo O, Nygard Y, Lankhorst PP, Bovenberg RAL, Driessen AJM. Mechanism and regulation of sorbicillin biosynthesis by *Penicillium chrysogenum*. Microb Biotechnol. 2017; 10 (4):958–68. https://doi.org/10.1111/1751-7915.12736 PMID: 28618182; PubMed Central PMCID: PMC5481523.

191. Derntl C, Guzman-Chavez F, Mello-de-Sousa TM, Busse HJ, Driessen AJM, Mach RL, et al. In Vivo Study of the Sorbicillinoid Gene Cluster in *Trichoderma reesei*. Frontiers in microbiology. 2017; 8:2037. https://doi.org/10.3389/fmicb.2017.02037 PMID: 29104566; PubMed Central PMCID: PMC5654950.

192. Heneghan MN, Yakasai AA, Halo LM, Song Z, Bailey AM, Simpson TJ, et al. First heterologous reconstruction of a complete functional fungal biosynthetic multigene cluster. Chembiochem. 2010; 11 (11):1508–12. Epub 2010/06/25. https://doi.org/10.1002/cbic.201000259 PMID: 20575135.

193. Halo LM, Heneghan MN, Yakasai AA, Song Z, Williams K, Bailey AM, et al. Late Stage Oxidations during the Biosynthesis of the 2-Pyridone Tenellin in the Entomopathogenic Fungus *Beauveria bassiana*. J Am Chem Soc. 2008; 130:17988–96. https://doi.org/10.1021/ja807052c PMID: 19067514

194. Zaehle C, Gressler M, Shelest E, Geib E, Hertweck C, Brock M. Terrein biosynthesis in *Aspergillus terreus* and its impact on phytotoxicity. Chem Biol. 2014; 21(6):719–31. Epub 2014/05/13. https://doi.org/10.1016/j.chembiol.2014.03.010 PMID: 24816227.

195. Kakule TB, Zhang S, Zhan J, Schmidt EW. Biosynthesis of the Tetramic Acids Sch210971 and Sch210972. Organic Letters. 2015; 17(10):2295–7. https://doi.org/10.1021/acs.orglett.5b00715 PMID: 25885659

196. Umemura M, Nagano N, Koike H, Kawano J, Ishii T, Miyamura Y, et al. Characterization of the biosynthetic gene cluster for the ribosomally synthesized cyclic peptide ustiloxin B in *Aspergillus flavus*. Fungal Genet Biol. 2014; 68:23–30. Epub 2014/05/21. https://doi.org/10.1016/j.fgb.2014.04.011 PMID: 24841822.

197. Lim FY, Won TH, Raffa N, Baccile JA, Wisecaver J, Rokas A, et al. Fungal Isocyanide Synthases and Xanthocillin Biosynthesis in Aspergillus fumigatus. mBio. 2018; 9(3). Epub 2018/05/31. https://doi.org/10.1128/mBio.00785-18 PMID: 29844112; PubMed Central PMCID: PMC5974471.

198. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020; 21(1):6. Epub 2020/01/04. https://doi.org/10.1186/s12864-019-6413-7 PMID: 31898477; PubMed Central PMCID: PMC6941312.

199. Frishman WH, Rapier RC. Lovastatin: an HMG-CoA reductase inhibitor for lowering cholesterol. The Medical clinics of North America. 1989; 73(2):437–48. Epub 1989/03/01. https://doi.org/10.1016/s0025-7125(16)30681-2 PMID: 2645482.

200. Hutchinson CR, Kennedy J, Park C, Kendrew S, Auclair K, Vederas J. Aspects of the biosynthesis of non-aromatic fungal polyketides by iterative polyketide synthases. Antonie van Leeuwenhoek. 2000; 78(3):287–95. https://doi.org/10.1023/a:1010294330190 PMID: 11386351

201. Gauglitz G. Artificial vs. human intelligence in analytics. Analytical and bioanalytical chemistry. 2019; 411(22):5631–2. https://doi.org/10.1007/s00216-019-01972-2 PMID: 31240356

202. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nature methods. 2015; 12(1):59–60. https://doi.org/10.1038/nmeth.3176 PMID: 25402007

203. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic acids research. 2011; 39(Web Server issue):W29–W37. Epub 05/18. https://doi.org/10.1093/nar/gkr367 PMID: 21593126.