Investigation of inter- and intraspecies variation through genome sequencing of Aspergillus section Nigri

Vesth, Tammi Camilla; Nybo, Jane L; Theobald, Sebastian; Frisvad, Jens Christian; Larsen, Thomas Ostenfeld; Nielsen, Kristian Fog; Hoof, Jakob Blæsbjerg; Brandl, Julian; Salamov, Asaf; Riley, Robert

Published in:
Nature Genetics

Link to article, DOI:
10.1038/s41588-018-0246-1

Publication date:
2018

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

Link back to DTU Orbit

Citation (APA):
Investigation of inter- and intraspecies variation through genome sequencing of *Aspergillus* section *Nigri*

Tammi C. Vesth¹, Jane L. Nybo¹, Sebastian Theobald¹, Jens C. Frisvad¹, Thomas O. Larsen¹, Kristian F. Nielsen¹, Jakob B. Hoof, Julian Brandl, Asaf Salamov², Robert Riley², John M. Gladden⁴, Pallavi Phatade³,⁵, Morten T. Nielsen¹, Ellen K. Lyhne¹, Martin E. Kogle¹, Kimchi Strasser⁶, Erin McDonnell⁶, Kerrie Barry², Alicia Clum², Cindy Chen², Kurt LaButti², Sajeet Haridas², Matt Nolan², Laura Sandor², Alan Kuo², Anna Lipzen², Matthieu Hainaut²,⁸, Elodie Drula²,⁸, Adrian Tsang⁶, Jon K. Magnuson³,⁵, Bernard Henriissat⁷,⁸,⁹, Ad Wiebenga¹⁰, Blake A. Simmons³,¹¹, Miia R. Mäkelä⁵,¹², Ronald P. de Vries¹⁰, Igor V. Grigoriev²,¹³, Uffe H. Mortensen¹, Scott E. Baker³,¹⁴ and Mikael R. Andersen¹⁵

*Aspergillus section Nigri* comprises filamentous fungi relevant to biomedicine, bioenergy, health, and biotechnology. To learn more about what genetically sets these species apart, as well as about potential applications in biotechnology and biomedicine, we sequenced 23 genomes de novo, forming a full genome compendium for the section (26 species), as well as 6 *Aspergillus niger* isolates. This allowed us to quantify both inter- and intraspecies genomic variation. We further predicted 17,903 carbohydrate-active enzymes and 2,717 secondary metabolite gene clusters, which we condensed into 455 distinct families corresponding to compound classes, 49% of which are only found in single species. We performed metabolomics and genetic engineering to correlate genotypes to phenotypes, as demonstrated for the metabolite auraspereone, and by heterologous transfer of citrate production to *Aspergillus nidulans*. Experimental and computational analyses showed that both secondary metabolism and regulation are key factors that are significant in the delineation of *Aspergillus* species.

Species in the genus *Aspergillus* are of broad interest to medical¹, applied², and basic research³. Members of *Aspergillus* section *Nigri* (‘black aspergilli’) are prolific producers of native and heterologous proteins⁴, organic acids (in particular citric acid⁷,⁸), and secondary metabolites (including biopharmaceuticals and mycotoxins like ochratoxin A). Furthermore, the section members are generally very efficient producers of extracellular enzymes⁹,¹⁰; they are the production organisms for 49 out of 260 industrial enzymes¹¹. Among the most important of these, in addition to *A. niger*, are *A. tubingensis*, *A. aculeatus*, and *A. luchuensis* (previously *A. acidus*, *A. kawachii*, and *A. awamori*¹², respectively).

Members of *Aspergillus section Nigri* are also known as destructive degraders of foods and feeds, and some isolates produce the potent mycotoxins ochratoxin A¹³ and fumonisins¹⁴-¹⁹. In addition, some species in this section have been proposed to be pathogenic to humans and other animals²⁰. It is thus of interest to further examine section *Nigri* for industrial exploitation, as well as prevention of food spoilage, toxin production, and pathogenicity caused by these fungi.

A combined phylogenetic and phenotypic approach has shown that section *Nigri* contains at least 27 species¹¹-²₃. Recent results have shown that the section contains species with high diversity and may consist of two separate clades: the biseriate species and the uniseriate species²⁴, which show differences in sexual states²⁵, sclerotium formation²⁶, and secondary metabolite production²⁷. In the section, only six species have had their genome sequenced: *A. niger*, *A. luchuensis*, *A. carbonarius*, *A. aculeatus*, *A. tubingensis*, and *A. brasiliensis*.

This section, with its combination of species richness and fungal species with a diverse impact on humanity, is thus particularly interesting for studying the diversification of fungi into species. In this study, we have de novo-sequenced the genomes of 20 species of section *Nigri*, thus completing a genome compendium of 26 described species in the section. Further, we have genome-sequenced three...
New genomes show high genetic diversity of section Nigri. We present 23 whole-genome draft sequences: 20 of section Nigri species previously unsequenced and 3 additional A. niger genomes for assessment of intraspecies diversity. All genomes were sequenced, assembled, and annotated using the Joint Genome Institute (JGI) fungal genome pipeline (Supplementary Table 1; genomes were sequenced by either Illumina or Pacific Biosciences sequencing). Figure 1 shows a phylogenetic tree as well as gene richness, number of scaffolds, and functional annotation (InterPro). The tree supports previous proposals that A. lactofermentus and A. phoenicis are synonyms of A. niger.

In comparing key statistics of the genomes, we found that some traits are quite similar and others surprisingly variable. Many of the investigated species have around the average number of genes (11,900), but there is considerable variation from the smallest number of predicted genes (10,066) to the largest (13,687). The smallest number of predicted genes in section Nigri is found in A. saccharolyticus, which supports the previous observation that this species is quite atypical in section Nigri.

We further evaluated the annotation of the 23 genome sequences we generated. The percentage of complete genes (including a start and stop codon) is in the range of 94–98%, and 67% of the proteins could be assigned one or more InterPro domains. The number of predicted genes in section Nigri shows genome flexibility.

The pan- and core-genome shows genome flexibility. Given the genetic diversity in section Nigri, we were interested in examining the extent of genome diversification. For this analysis, we focused on three conceptual groups of genes:

1. The pan-genome: all genes present in one or more species.
2. The core-genome: genes present in all included species,
including paralogs. This set is expected to encode cellular functions needed for all species.

(3) Species-unique genes: genes found in only one species in our analysis, with or without paralogs. Included in these, we would expect to find genes involved in environmental adaptation.

This group can also include annotation errors.

We first identified orthologs and paralogs with a BLASTp-based pipeline using reciprocal hits according to cut-offs specifically selected here for the Aspergillus genus (Methods). Groups of homologous proteins are referred to as families. Figure 2a–c shows the overall genetic diversity between 38 fungal strains (32 species) from closely related genera (Fig. 2a), within the Aspergillus genus (36 of the 38 strains; Fig. 2b), and from section Nigri (32 of the 38 strains; Fig. 2c).

The Aspergillus genus pan-genome comprises 433,116 genes across the 36 Aspergillus genomes, and from this, 62,996 gene families were constructed. Of those families, 6% are found in all genomes (3,769 core families), while 9% are genes without orthologs in the other genomes (40,424 unique genes; 39,929 unique genes) (Fig. 2b). We also found evidence of gene loss, duplication, and potential gene transfers between species of this section, as 23% of the pan-genome families are not present in groups of species fitting the phylogenetic relation between the 32 species in section Nigri (32 of the 38 strains; Fig. 2b). We also found evidence of gene loss, duplication, and potential gene transfers between species of this section, as 23% of the pan-genome families are not present in groups of species fitting the phylogenetic tree (Supplementary Table 2). This is consistent with previous work reporting extensive horizontal gene transfer in Aspergillus39.

We further performed an analysis defining the number of core-gene families in section Nigri and in all sub-clades thereof (Fig. 2d). The core-genome of section Nigri is 32% larger than that of the genus (4,983 families relative to 3,769; Fig. 2b,c). Conversely, 9% are unique to a specific species (32,378 unique genes in 32,036 families; Fig. 2c). The fraction of genes unique to a species is similar within the section and across the genus, meaning that adding a new section Nigri genome adds as many new genes as adding a more distantly related Aspergillus (within the analyzed group of species). This is rather interesting and shows a generally high genetic diversity of
genus Aspergillus. However, such a tendency could also be the result of overpredicting genes, considering the low rate of InterPro annotation in the unique genes (Fig. 2c).

The section Nigri core genome contains carbohydrate-active enzymes and secondary metabolism gene clusters. To associate biological functions to the pan-, core-, and unique genomes, and genes exclusive to only members of the black aspergilli, we employed the InterPro database13. An examination of the core-genome of 38 fungal genomes (Fig. 2a and Supplementary Fig. 1) revealed that only 4.5% of the genes lack InterPro domains (Supplementary Table 3a). What’s more, many genes across closely related fungal genera include generally known and conserved functions. For the pan-genomes of the 36 Aspergillus species compared with the section Nigri genomes, the percentages of unknown function are similar (32% compared with 33%; Supplementary Tables 3d and 4a and Supplementary Fig. 2), as are the corresponding percentages for the core-genes (14% compared with 17%, Fig. 2c; Supplementary Tables 3d and 4a). General functions like transporters, regulators, organelle-specific proteins, primary metabolism, and structural metabolism were found as core features across all 36 aspergilli (Supplementary Table 3f), which supports the general validity of the method.

We expected the section Nigri core-genome (gene families found in all the species of section Nigri but not in any other aspergilli examined) to contain Nigri signature genes, and we found this to be the case. These 1,214 gene families contain 580 InterPro domains conserved to a varying degree, including a many genes involved in the saprotrophic lifestyle and secondary metabolism (Supplementary Table 5). It is hypothesized that these genes are defining for the section compared with other aspergilli and will encode functions related to the phenotypes of species in this section.

Unique secondary metabolism genes in Aspergillus species. The genetic diversity seen in section Nigri led us to investigate whether the unique genes for each species show common trends in function. While these genes by definition do not have homologs in other species investigated in this work, we can predict general functions using InterPro domains. Unique genes of species in section Nigri matched 1,334 different InterPro domains (Supplementary Table 6a–c). Within the unique genes, we searched the list of InterPro domains in all sets of genes unique to individual section Nigri species (excluding the six A. niger isolates, to remove intraspecies redundancy). Surprisingly, we identified only ten domains that were found in nearly all Nigri species (25–26 species). Notably, nine of those are related to functions involved in secondary metabolism, gene regulation (transcription factors), or protein regulation (protein kinases) (Supplementary Table 7). Finding these functions in nearly all sets of species-specific genes suggests that secondary metabolite production and regulatory proteins are commonly identified as the species’-unique’ genes and are therefore critical differentiates for fungal species at the genetic level.

Intra- and interspecies genetic variations are similar. We were interested in comparing the diversity between isolates of the same species to the diversity among species in the same clade. We thus compared six A. niger isolates to the eight closely related species in the A. tubingensis clade (Fig. 2d). The A. niger isolates have a high
Aspergillus luchuensis IFO 4308
Aspergillus luchuensis CBS 106.47
Aspergillus piperis
Aspergillus eucalyptola
Aspergillus tubingenensis
Aspergillus costaricaensis
Aspergillus neoigiser
Aspergillus vadensis
Aspergillus niger ATCC 13496
Aspergillus niger CBS 513.88
Aspergillus niger (lactiçoferatus)
Aspergillus niger NRRL3
Aspergillus niger ATCC 1015
Aspergillus niger (phoecis)
Aspergillus weisethiae
Aspergillus brasiliensis
Aspergillus ibericus
Aspergillus sclerotialitabariunus
Aspergillus carbonarius
Aspergillus sclerotieniger
Aspergillus heteromorphus
Aspergillus ellipticus
Aspergillus fijiensis
Aspergillus brunneoiacialaeus
Aspergillus aculeatinus
Aspergillus aculeatus
Aspergillus violaceofuscus
Aspergillus japonicus
Aspergillus indologenius
Aspergillus uvarum
Aspergillus saccharolyticus
Aspergillus homomorphus
Aspergillus nidulans
Aspergillus flavus
Aspergillus oryzae
Aspergillus fumigatus AT293
Penicillium rubens
Neurospora crassa

### Fig. 4 | Comparison of CAZyme gene content divided by target polysaccharide.
Details on CAZy families are available in Supplementary Table 11. Growth profiles are available in Supplementary Fig. 6. All black aspergilli grew well on pectin and have a highly conserved and extensive set of genes encoding pectin-active enzymes. Growth on other plant polysaccharides such as xylan, starch, and guar gum was more variable, despite the presence of highly conserved genes related to xyloglucan and starch degradation. The growth and genetic variability on inulin are particularly high: nine species showed reduced growth. Moreover, endo-inulinase (GH32 INU) is only present in eight of the black aspergilli, while the remainder of inulin-related genes (GH32 INV and INX) are more commonly present (Supplementary Table 11). However, the growth phenotypes show no correlation with the gene content (Supplementary Fig. 6).

degree of genetic homogeneity, as 80% of the A. niger pan-genome is conserved across the six isolates and only 6% is unique to any of the isolates (Supplementary Fig. 3a). The same scale is seen in the A. tubingenensis clade (77% shared pan-genome, 7% unique; Supplementary Fig. 3b). Moreover, the percentage of genes with predicted functional domains within the two groups is similar to that within section Nigri as a whole (Supplementary Fig. 4 and Supplementary Tables 8a and 8a,d). The unique genes belonging to each of the two groups are largely of unknown function (A. tubingenensis clade 82%, A. niger complex 86%; Supplementary Tables 8a,d and 9a,b). The functions of the A. niger core-genome (3,798 domains) are, not surprisingly, very similar to those of section Nigri as a whole (Supplementary Tables 4c and 8c). In summary, the interspecies variation in the A. tubingenensis clade is of the same scale as the intraspecies variation in the A. niger isolates, showing that large genetic variation does not directly translate to the currently circumscribed species.

### Carbon utilization is not correlated with carbohydrate-active enzyme content.
Aspergilli have a particularly broad ability to degrade and convert plant biomass\(^{41}\). It is thus essential to examine the species diversity of this trait at the genotype and phenotype levels. We predicted the carbohydrate-active enzyme (CAZyme) gene content of the genomes across section Nigri (17,903 CAZyme domains; Fig. 4 and Supplementary Table 11) and performed growth profiling on plant biomass-related carbon sources (Supplementary Fig. 6). Growth on d-glucose was used to evaluate relative growth, showing variation between species.

In a previous study\(^{11}\), enzyme levels were measured in several black aspergilli, and significant differences were found. However, differences in enzyme levels do not reflect the copy number differences seen here (Supplementary Table 11). Considering the relative uniformity of the CAZyme content (Fig. 4), no correlation between genome content and growth on plant biomass-related carbon sources (Supplementary Fig. 6) was observed for the black aspergilli, suggesting that the differences in capability for plant biomass degradation reflect gene expression levels in the individual fungus. This confirms a proteome study of less-related aspergilli, in which the different response to plant biomass appeared to be mainly at the regulatory level\(^{41}\). The data suggest that this is the case for section Nigri: species-specific phenotypes are driven not generally by CAZyme content in closely related species, but by regulation.

### Secondary metabolism in section Nigri contains 455 families.
Secondary metabolism is thought to be a component of chemical defense, virulence, toxicity, mineral uptake, and communication in fungi\(^{14}\) and has a wide range of potential medical applications. As we had identified it to be commonly unique to individual species, we examined the exometabolite diversity of 37 Aspergillus and Penicillium species according to predictions of secondary
metabolism gene clusters (SMGCs) as well as chemical profiles of the species of section Nigri on multiple substrates.

We identified 2,717 SMGCs in the 37 genomes. This is an even higher number of SMGCs per species than a previous study found in 24 Penicillium genomes7. We were further interested in quantifying the actual diversity of the SMGCs in section Nigri and in analyzing presence patterns of SMGCs across species. We therefore defined SMGC 'families' as genetically similar SMGCs across genomes (Methods). Each SMGC family is expected to produce the same or similar compounds. This clustering resulted in the definition of 455 SMGC families across the 37 genomes (Supplementary Fig. 7), indicating the potential production of 455 different chemical families. Most families (82%) are found in fewer than 10 organisms, and 49% contain only one gene cluster (Supplementary Fig. 8 shows examples). On average there are 8.75 unique clusters per species, despite the close phylogenetic distance of the section.

Phylogenetic examination shows dynamic content of SMGCs. To reveal more about how SMGCs evolve and differentiate between species, each of the 455 SMGC families was characterized by the type of backbone enzyme and analyzed according to the phylogeny (Fig. 5a,b). Only five out of all SMGCs were present in all analyzed species, including clusters for the non-ribosomal peptide synthetase (NRPS)-derived siderophore ferrichrome, the circular NRP fungisporin⁵/nidulanin A⁶, and pigment (YW4) synthesis. Two shared SMGC families were false predictions, namely two fatty acid synthases.

Examining the dynamics of the families, only 32% and 19% of SMGCs found in two or three organisms, respectively, follow the whole-genome phylogeny and suggest intragenus horizontal gene transfer or SMGC loss to be relatively common. As an example, an SMGC is found in five distantly related species (Supplementary Fig. 8). Transferring or SMGC loss to be relatively common. As an example, an SMGC is found in five distantly related species (Supplementary Fig. 8). As an example, an SMGC is found in five distantly related species (Supplementary Fig. 8). Two shared SMGC families were false predictions, namely two fatty acid synthases.

Examining the dynamics of the families, only 32% and 19% of SMGCs found in two or three organisms, respectively, follow the whole-genome phylogeny and suggest intragenus horizontal gene transfer or SMGC loss to be relatively common. As an example, an SMGC is found in five distantly related species (Supplementary Fig. 8). Two shared SMGC families were false predictions, namely two fatty acid synthases.

Fig. 5 | SMGCs and types of cluster backbones for section Nigri. a, Cladogram of 36 Aspergillus (32 Nigri) genomes and one Penicillium genome. Numbers at nodes show SMGC families shared by species in the branch. End-points show families unique to an organism. b, Total SMGC family profile of the organism classified by backbone enzyme type. c, Presence/absence map of detected secondary metabolites in section Nigri species (Methods). Strains with no available metabolite data are marked in light gray. DMAT, dimethylallyltransferase (prenyl transferases); Hybrid, gene containing domains from NRPS and PKS backbones; NRPS, non-ribosomal peptide synthetase; NRPS-like, non-ribosomal peptide synthetase-like containing at least two NRPS-specific domains and another domain; TC, terpene cyclase.

Aspergillus a (32 SMGCs and types of cluster backbones for section Nigri).
species genomes (with the exception of *A. tubingenesis*, *A. niger*, *A. brasiliensis*, and *A. vadinensis*) encode one or more unique SMGCs. These patterns show the existence of high diversity of SMGCs between species and of a homogeneous set of SMGCs within isolates from the same species.

**Correlating secondary metabolisms with SMGC families links gene to function.** As a further application of the constructed SMGC families, we hypothesized that we can correlate SMGC families to classes of compounds. We performed extensive exometabolome analysis of 27 of the sequenced strains and identified 35 compound families (Fig. 5c and Supplementary Table 12).

The most abundant group was napthha-γ-pyrones, of which aurasperone B29 was identified in 14 of the isolates. We compared the presence patterns of SMGC families with the compound class (Fig. 5c) and combined it with a knowledge-based filtering of InterPro domains leaving one hit (Methods and Supplementary Fig. 8d). The candidate SMGC family is a nine-gene cluster found in 18 genomes—including the 14 where we detected the compound—and it contains all activities needed to synthesize aurasperone. In support of this identification, an SMGC for a closely related compound, aurofusarin, has been experimentally verified in *Fusarium graminearum*. The aurasperone cluster shares six genes (one of which is a duplication) with the aurofusarin cluster. This finding supports the assignment of this family of SMGCs to the production of aurasperone B and conceptually justifies this approach for efficient linking of clusters to compounds. We see this correlation approach as highly useful for future elucidation of fungal metabolites.

**Discussion**

We have sequenced the genomes of a whole section of filamentous fungi, and a diverse set of *A. niger* isolates, and found that the species are highly diverse in some traits, in particular secondary metabolism and to a lesser extent regulatory proteins, and homogeneous in others, such as glycolytic metabolism and CAZymes. The presented data furthermore provide an extensive compendium of 24 new genomes, which adds substantial information on fungal genetic diversity. We further combined genome analysis with metabolite profiling and heterologous gene expression to identify the genetic basis of several phenotypes within primary and secondary metabolism.

Of particular interest was the finding that the species-specific genes in all species share functions within gene/protein regulation and secondary metabolism, showing that unique sets of these functions exist for all species in the investigated set.

**References**


34. Grigoriev, I. V., Martínez, D. A. & Salamov, A. A. Fungal genomic annotation. 


37. Sørensen, A., Lübeck, P. S., Lübeck, M., Teller, P. J. & Ahring, B. K. 
   β-Glucosidases from a new Aspergillus species can substitute commercial 
   β-glucosidases for saccharification of lignocellulosic biomass. 

38. Sørensen, A. et al. Identifying and characterizing the most significant 
   β-glucosidase of the novel species Aspergillus saccharolyticus. 

   Genome-scale phylogenetic analysis finds extensive gene transfer among fungi. 

40. Karaffa, L. & Kubicek, C. P. 
   Secondary metabolism: regulation and role in fungal biology. 

41. Andersen, M. R., Nielsen, M. L. & Nielsen, J. Metabolic model integration of 

42. Frandsen, R. J. N. et al. The biosynthetic pathway for aurofusarin in 
   *Aspergillus niger* citric acid accumulation: do we understand this well working black box? 

43. Andersen, M. R., Nielsen, M. L. & Nielsen, J. Metabolic model integration of 

44. Bento, I. et al. Closely related fungi employ diverse enzymatic strategies to 

45. Fox, E. M. & Howlett, B. J. Secondary metabolism: regulation and role in 

46. Sørensen, A. et al. Global analysis of biosynthetic gene clusters reveals 
   vast potential of secondary metabolite production in *Penicillium* species. 

47. Nielsen, J. C. et al. Global analysis of biosynthetic gene clusters reveals 
   vast potential of secondary metabolite production in *Penicillium* species. 

   gene clusters in filamentous fungi. *Proc. Natl Acad. Sci.* USA **110**, 

49. Frandsen, R. J. N. et al. The biosynthetic pathway for aurofusarin in 
   *Paecilomyces griseofuscus* reveals a close link between the naphthoquinones 

50. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and 


52. Talavera, G. & Castresana, J. Improvement of phylogenies after removing 
   divergent and ambiguously aligned blocks from protein sequence alignments. 
Purification of DNA and RNA. For all sequences generated for this study (Supplementary Table 1), spores were defrosted from storage at −80°C and inoculated onto solid CYA medium. Fresh spores were harvested after 7–10 d and suspended in a 0.1% Tween solution. Spores were stored in solution at 5°C for up to 3 weeks. Biomass for all fungal strains was obtained from shake flasks containing 200 ml of complex medium, either CYA, MEAox, or CY20 depending on the strain (see Supplementary Table 1) cultivated for 5–10 d at 30°C. Biomass was isolated by filtering through Miracloth (Millipore, 475855-1R), freeze dried, and stored at 80°C. DNA isolation was performed using a modified version of the standard phenol extraction (ref. 1 and below) and checked for quality and concentration using a NanoDrop (ThermoFisher Scientific). RNA isolation was performed using the Qiagen RNeasy Plant Mini Kit according to the manufacturer’s instructions.

A sample of frozen biomass was subsequently used for RNA purification. First, hyphae were transferred to a 2 ml microtube together with a 5 mm steel bead (Qiagen), placed in liquid nitrogen, then lysed using the Qiagen TissueLyser LT at 45 Hz for 50 s. Then the Qiagen RNeasy Mini Plus Kit was used to isolate RNA. RLT Plus buffer (with 2-mercaptoethanol) was added to the samples, vortexed, and spun down. The lysate was then used in step 4 in the instructions provided by the manufacturer, and the protocol was followed from this step. For genomic DNA, a protocol based on Fulton et al.1 was used (See Supplementary Note).

DNA and RNA sequencing and assembly. All genomes in this study, except for those of A. heteromorphus, A. eucalypticola, and A. sclerotiorum, and all transcriptomes were sequenced with Illumina. The genomes of A. heteromorphus, A. eucalypticola, and A. sclerotiorum were sequenced with PacBio.

For all genomic Illumina libraries, 100 ng of DNA was sheared to 270 bp fragments using the Covaris LE220 (Covaris) and size selected using SPRI beads (Beckman Coulter). The fragments were treated with end-repair and A-tailing and ligated to Illumina-compatible adapters (IDT) using the KAPA-Illumina library creation kit (KAPA Biosystems). For transcriptomes, stranded complementary DNA libraries were generated using the Illumina TruSeq Stranded Total RNA LT Sample Prep Kit. Messenger RNA (mRNA) was purified from 1 μg of total RNA using magnetic beads containing poly(T) oligos. mRNA was fragmented using divalent cations and high temperature. The fragmented RNA was reverse transcribed using random hexamers and SSII (Invitrogen) followed by second-strand synthesis. The fragmented complementary DNA was treated with end-pair, A-tailing, adapter ligation, and 10 cycles of PCR. The prepared libraries were quantified using KAPA Biosystems’ next-generation sequencing library quantitative PCR kit and run on a Roche LightCycler 480 real-time PCR instrument. The quantified libraries were then multiplexed with other libraries, and library pools were prepared for sequencing on the Illumina HiSeq2000 sequencing platform using a TruSeq paired-end cluster kit, v3, and Illumina’s cbot instrument to generate clustered flow cells for sequencing. Sequencing of the flow cells was performed on the Illumina HiSeq2000 sequencer using a TruSeq SBS sequencing kit, v3, following a 2 x 150 indexed run recipe. After sequencing, the genomic FASTQ files were quality-control filtered to remove artifacts/process contamination and aligned using Velvet7. The resulting assemblies were used to create in silico long mate-pair libraries with resulting assemblies were used to create in silico long mate-pair libraries with inserts of 3,000–900 bp, which were then assembled with the target FASTQ using AllPaths-LG release version R47710. Illumina transcriptome reads were assembled into consensus sequences using Rnnotator v3.3.2.

For the genomes of A. heteromorphus, A. eucalypticola, and A. sclerotiorum, amplified libraries were generated using a modified shearing version of the Pacific Biosciences standard template preparation protocol. To generate each library, 5 μg of genomic DNA was used. The DNA was sheared using a Covaris LE220 focused-ultrasonicator with their Red miniTUBES to generate fragments 5 kb in length. The sheared DNA fragments were then treated according to the Pacific Biosciences protocol using the SMRTbell template preparation kit, where the fragments were treated with DNA damage repair (ends were repaired so that they were blunt ended and 5’ phosphorylated). Pacific Biosciences hairpin adapters were then ligated to the fragments to create the SMRTbell template for sequencing. The SMRTbell templates were then purified using exonuclease treatments and size-selected using AMPure PB beads. Sequencing primer was then annealed to the SMRTbell templates, and version P4 sequencing polymerase was bound to them. The prepared SMRTbell template libraries were sequenced on a Pacific Biosciences RS II sequencer using version C2 chemistry and 2 h sequencing movie run times. The three Pacific Biosciences genome datasets were assembled using HGAP3 (see URLs). Genome assembly and annotations are available at the JGI fungal genome portal MycoCosm11 (see URLs) and have been deposited in the DNA Data Bank of Japan (DDBJ)/European Molecular Biology Laboratory (EMBL)/GenBank under the following accession numbers: A. aculeata (PSTC00000000), A. brunneoviolaceus (PST00000000), A. costaricensis (PSTH00000000), A. ellipticus (PSTY00000000), A. eucalypticola (MSFU00000000), A. fijitensis (PSTG00000000), A. heteromorphus (MSFR00000000), A. homomorphus (PSTJ00000000), A. ibicus (PSTD00000000), A. japonicus (PSTD00000000), A. luteofoetus (MSFR00000000), A. noeinger (MSF00000000), A. niger ATCC 13157 (A. phoenicia) (QQUR00000000), A. niger ATCC 13496 (QQZP00000000), A. pipers (PSTD00000000), A. saccharolyticus (MSFQ00000000), A. sclerotiorum (PSS20000000), A. sclerotiorum (MSFK00000000), A. uvaram (MSFT00000000), A. vanudens (MSFS00000000), and A. velwitschiae (QIQZQ00000000).

See also the Nature Research Reporting Summary linked to this article.

Analysis of secondary metabolism. Cultivation for secondary metabolite analysis. Fungal strains were cultivated as three-point cultures on CYA, CYAS, and YES media for 7 d in the dark at 25°C. Three 6 mm inner diameter plugs taken across the cultures were then extracted using an (3:2:1) (ethylacetate—dichloromethane—methanol) mixture and dissolved in methanol22.

Extraction of fungal metabolites. Fungal metabolite extracts were prepared using one of the three following methods6,12 (1) chloroform–methanol–ethylacetate extraction, (2) micro-extraction using methanol–dichloromethane–ethylacetate, or (3) 75% methanol extraction.

Chemical analysis of secondary metabolites. All chemical analyses were done by reversed-phase ultrahigh-performance liquid chromatography (UHPLC) coupled to ultraviolet–visible diode array detection (DAD) combined with either fluorescence detection (FLD) or high-resolution mass spectrometry (HRMS). Three different methods were used:

Method 1. Pure UHPLC–DAD–FLD was performed using a rapid-separation liquid chromatography (UHPLC) UltiMate 3000 system (Dionex) linked to an 1100 Series FLD (Agilent). The system was equipped with an Agilent Poroshell phenyl-hexyl column (150×2.1 mm, 2.6 μm) and was run using a linear gradient of water–acetonitrile starting at 10% acetonitrile and increasing to 100% (both containing 50 μM trifluoroacetic acid) over 8 min, then using 100% acetonitrile for 2 min. The column temperature was 60°C, the flow rate 0.8 ml min−1, and the injection volume was 1 μl. The ultraviolet spectra 200–640 nm were matched against our internal database13.

Method 2. UHPLC–DAD–HRMS was conducted on a Dionex RSLC UltiMate system linked to a mXis high-definition quadrupole–time–of–flight mass spectrometer (Q–TOF MS) (Bruker Daltonics). Separation was done on a Kinetex C18 column (100×2.1 mm, 2.6 μm), with a linear gradient consisting of water and acetonitrile (both buffered with 20 mM formic acid), starting at 10% acetonitrile and increasing to 100% over 10 min, where it was held for 2 min and returned (0.4 ml min−1, 40°C). Injection volume, depending on sample concentration, typically varied between 0.1 and 1 μl. Some samples were analyzed in electrospray ionization (ESI)2 and some in ESI full-scan mode, scanning m/z 100–1,250. Data were analyzed by aggressive dereplication23 using lists of compounds considered to be from black aspergilli only (~350); a list with all Aspergillus compounds (~2,450); and a list of 1,600 reference standards, of which 500 are known to come from Aspergillus. Unknown peaks were matched against Antibase2012 and dereplicated using mass, isotope patterns, adduct patterns, log D, and ultraviolet visible data12.

Method 3. UHPLC–DAD–HRMS was conducted on an Agilent Infinity 1290 UHPLC system coupled to an Agilent 6550 Q–TOF MS. Separation was obtained on an Agilent Poroshell 120 phenyl-hexyl column (2.1×25 mm, 2.7 μm) using a linear gradient of water and acetonitrile (both buffered with 20 mM formic acid), progressing from 10% to 100% acetonitrile over 15 min, where it was held for 2 min. The flow was 0.35 ml min−1 and the temperature 60°C. Injection volume was between 0.1 and 1 μl, depending on the sample concentration. Some samples were analyzed in ESI+ and some in ESI full-scan mode, scanning m/z 100–1,700 and with automatic MS/MS enabled for ion counts above 100,000 and with a quarantine time of 0.06 min. MS/MS spectra were obtained at 10, 20, and 40 eV (ref. 19). Full-scan data were analyzed as above in MassHunter19. MS/MS data were matched to our internal MS library (~1,700 compounds) of reference standards and tentatively identified compounds.

Genome annotation and analysis. Genome annotation. All genomes were annotated based on the JGI annotation pipeline19 as previously described20.

Swiss-Prot comparison. Swiss-Prot comparisons were done using protein BLAST alignments with BLASTp (v2.3.0), c-value cut-off 1.0 × 10−10, target_seq 100, max_hsp 1, and locally optimal Smith–Waterman alignments (use_sw_back).

Whole-genome phylogeny. Protein sequences of all organisms were compared using BLASTp (c-value cut-off 1.0 × 10−10). Orthologous groups of sequences were
predicted as 'DMAT' genes.

- Genes that have ketoacyl-synth (PF00109) and Ketoacyl-synth_C (PF02801) but not Acyl_transf_1 (PF00698) were predicted as 'PKS-like' genes.
- Genes that have the three domains AMP-binding (PF00501), PP-binding (PF00550), and Condensation (PF00668) were predicted as 'NRPS' genes.
- Genes that have both 'PKS' and 'NRPS' domains were predicted as 'Hybrid' genes.
- Genes that have a Trp_DMAT domain were predicted as 'DMAT' genes.
- Genes that have Terpene_synth (PF01397) or Terpene_synth_C (PF03936) domains were predicted as 'Terpene cyclase/synthase' genes.

Secondary metabolite-specific PFAM domains were taken from Supplementary Table 2 of the SMURF paper.

As input, the program takes genomic coordinates and the annotated PFAM domains of the predicted genes. Based on the multidomain PFAM composition of identified 'backbone' genes, it can predict seven types of secondary metabolite clusters: (1) polyketide synthases (PKSs), (2) PKS-like, (3) non-ribosomal peptide-synthetases (NRPSs), (4) NRPS-like, (5) hybrid PKS-NRPS, (6) prenyltransferases (DMATS), and (7) terpene cyclases (TCs). Besides backbone genes, PFAM domains, which are enriched in experimentally identified secondary metabolite clusters (secondary metabolite-specific PFAMs), were used in determining the borders of gene clusters. The maximum allowed size of intergenic regions in a cluster was set to 3 kb, and each predicted cluster was allowed to have up to 6 genes without secondary metabolite-specific domains.

Prediction of secreted proteases. Secretome prediction was done using an in-house adaptation of SignalP6. 

According to SMURF the following genes were predicted as a 'backbone' genes:

- Genes that have at least three PFAM domains—ketocycl-synt (PF00109), Ketaacly-synt_C (PF02801), and Acyl_transf_1 (PF00698)—were predicted as 'PKS' genes.
- Genes that have ketocycl-synt (PF00109) and Ketaacly-synt-C (PF02801) but not Acyl_transf_1 (PF00698) were predicted as 'PKS-like' genes.
- Genes that have the three domains AMP-binding (PF00501), PP-binding (PF00550), and Condensation (PF00668) were predicted as 'NRPS' genes.
- Genes that have an AMP-binding (PF00501) domain and at least one of the domains PP-binding (PF00550), Condensation (PF00668), NAD_binding_4 (PF07995), and Epimerase (PF01370) were predicted as 'NRPS-like' genes.
- Genes that have both 'PKS' and 'NRPS' domains were predicted as 'Hybrid' genes.
- Genes that have a Trp_DMAT domain were predicted as 'DMAT' genes.
- Genes that have Terpene_synth (PF01397) or Terpene_synth_C (PF03936) domains were predicted as 'Terpene cyclase/synthase' genes.

Gene-compound assignment. Identification of conserved or highly similar fungal gene clusters was performed on the basis of the gene cluster predictions above. The genomes were compared using the BLASTp function from the BLAST+ suite (e-value cutoff of 1×10^{-10}). All genomes used in the study are available from Joint Genome Institute (JGI) website (see URLs). Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability. All genomes used in the study are available from Joint Genome Institute fungal genome portal MycoCosm (http://jgi.doe.gov/fung). All new genomes published in the study have been deposited in the DNA Data Bank of Japan (DDBJ)/European Molecular Biology Laboratory (EMBL)/GenBank under the following accession numbers: A. aculeatinus (PST200000000), A. brunneoviolaceus (PSTC00000000), A. casticariacus (PSTH00000000), A. ellipticus (PSY00000000), A. eucalyptolica (MSUF00000000), A. fijianensis (PSTG00000000), A. heteromorphus (MSLF00000000), A. homomorphus (PSTM00000000), A. ibericus (PSTP00000000), A. indolvensus (PSTR00000000), A. japonicus (PSTFO00000000), A. lactifluus (MSRF00000000), A. neomiger (MSFP00000000), A. niger ATCC 13157 (A. phoenicis (QQUR00000000), A. niger ATCC 13159 (QQZ00000000), A. piperis (PSTRD00000000), A. saccharolyticus (MSFR00000000), A. selerottiorubus (MSFQ00000000), A. sclerotiorubus (PSQ20000000), A. awarun (MSFT00000000), A. vaidentis (MSFS00000000), A. violaceofuscasus (PSTA00000000), and A. welwitschiae (QQQZ00000000).
References

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

<table>
<thead>
<tr>
<th>n/a</th>
<th>Confirmed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>✓ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement</td>
</tr>
<tr>
<td></td>
<td>✓ An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly</td>
</tr>
<tr>
<td></td>
<td>✓ The statistical test(s) used AND whether they are one- or two-sided</td>
</tr>
<tr>
<td></td>
<td>✓ Only common tests should be described solely by name; describe more complex techniques in the Methods section.</td>
</tr>
<tr>
<td></td>
<td>✓ A description of all covariates tested</td>
</tr>
<tr>
<td></td>
<td>✓ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons</td>
</tr>
<tr>
<td></td>
<td>✓ A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)</td>
</tr>
<tr>
<td></td>
<td>✓ For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted Give P values as exact values whenever suitable.</td>
</tr>
<tr>
<td></td>
<td>✓ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings</td>
</tr>
<tr>
<td></td>
<td>✓ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes</td>
</tr>
<tr>
<td></td>
<td>✓ Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated</td>
</tr>
<tr>
<td></td>
<td>✓ Clearly defined error bars State explicitly what error bars represent (e.g. SD, SE, CI)</td>
</tr>
</tbody>
</table>

Software and code

Policy information about availability of computer code

Data collection

No software was used for data collection

Data analysis

Custom code was used, this is available through GitHub: https://github.com/RoerdamAndersenLab/

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Genome assembly and annotations are available at the JGI fungal genome portal MycoCosm (6) (http://jgi.doe.gov/fungi) and have been deposited at DDBJ/EMBL/GenBank under the following accessions A. aculeatinus (PSTE00000000), A. brunneoviolaceus (PSTC00000000), A. costaricaensis (PSTH00000000), A. ellipticus (PSSY00000000), A. eucalyptcola (MSFU00000000), A. fijiensis PSTG00000000), A. heteromorphus (MSFL00000000), A. homomorphus (PSTJ00000000), A. ibericus
Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences       ☐ Behavioural & social sciences       ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/reportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>N/A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data exclusions</td>
<td>N/A</td>
</tr>
<tr>
<td>Replication</td>
<td>Genomes were only sequenced once each, but this is in accordance with current best practice.</td>
</tr>
<tr>
<td>Randomization</td>
<td>N/A</td>
</tr>
<tr>
<td>Blinding</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Reporting for specific materials, systems and methods

Materials & experimental systems

<table>
<thead>
<tr>
<th>n/a</th>
<th>Involved in the study</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unique biological materials</td>
</tr>
<tr>
<td>☑</td>
<td>Antibodies</td>
</tr>
<tr>
<td>☑</td>
<td>Eukaryotic cell lines</td>
</tr>
<tr>
<td>☑</td>
<td>Palaeontology</td>
</tr>
<tr>
<td>☑</td>
<td>Animals and other organisms</td>
</tr>
<tr>
<td>☑</td>
<td>Human research participants</td>
</tr>
</tbody>
</table>

Methods

<table>
<thead>
<tr>
<th>n/a</th>
<th>Involved in the study</th>
</tr>
</thead>
<tbody>
<tr>
<td>☑</td>
<td>ChIP-seq</td>
</tr>
<tr>
<td>☑</td>
<td>Flow cytometry</td>
</tr>
<tr>
<td>☑</td>
<td>MRI-based neuroimaging</td>
</tr>
</tbody>
</table>

Unique biological materials

Policy information about availability of materials

Obtaining unique materials

Strains are available from the authors, from strain collections and/or from the original isolators of the material.