Title

A comparative genomics study of 23 Aspergillus species from section Flavi

Permalink

https://escholarship.org/uc/item/3mr4c31c

Journal

Nature Communications, 11(1)

ISSN

2041-1723

Authors

Kjærbølling, Inge
Vesth, Tammi
Frisvad, Jens C
et al.

Publication Date

2020

DOI

10.1038/s41467-019-14051-y

Peer reviewed

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-14051-y OPEN

A comparative genomics study of 23 *Aspergillus* species from section *Flavi*

Inge Kjærbølling[1], Tammi Vesth[1], Jens C. Frisvad[1], Jane L. Nybo[1], Sebastian Theobald[1], Sara Kildgaard[1], Thomas Isbrandt Petersen[1], Alan Kuo[2], Atsushi Sato[3], Ellen K. Lyhne[1], Martin E. Kogle[1], Ad Wiebenga[4], Roland S. Kun[4], Ronnie J.M. Lubbers[4], Miia R. Mäkelä[5], Kerrie Barry[2], Mansi Chovatia[2], Alicia Clum[2], Chris Daum[2], Sajeet Haridas[2], Guifen He[2], Kurt LaButti[2], Anna Lipzen[2], Stephen Mondo[2], Jasmyn Pangilinan[2], Robert Riley[2], Asaf Salamov[2], Blake A. Simmons[6], Jon K. Magnuson[6], Bernard Henrissat[7], Uffe H. Mortensen[1], Thomas O. Larsen[1], Ronald P. de Vries[4], Igor V. Grigoriev[2,8], Masayuki Machida[9], Scott E. Baker[6,10] & Mikael R. Andersen[1*]

Section *Flavi* encompasses both harmful and beneficial *Aspergillus* species, such as *Aspergillus oryzae*, used in food fermentation and enzyme production, and *Aspergillus flavus*, food spoiler and mycotoxin producer. Here, we sequence 19 genomes spanning section *Flavi* and compare 31 fungal genomes including 23 *Flavi* species. We reassess their phylogenetic relationships and show that the closest relative of *A. oryzae* is not *A. flavus*, but *A. minisclerotigenes* or *A. aflatoxiformans* and identify high genome diversity, especially in sub-telomeric regions. We predict abundant CAZymes (598 per species) and prolific secondary metabolite gene clusters (73 per species) in section *Flavi*. However, the observed phenotypes (growth characteristics, polysaccharide degradation) do not necessarily correlate with inferences made from the predicted CAZyme content. Our work, including genomic analyses, phenotypic assays, and identification of secondary metabolites, highlights the genetic and metabolic diversity within section *Flavi*.

[1] Department of Biotechnology and Bioengineering, Technical University of Denmark, Søltoft Plads 223, 2800 Kongens Lyngby, Denmark. [2] US Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA. [3] Kikkoman Corporation, 250 Noda, 278-0037 Noda, Japan. [4] Fungal Physiology, Westerdijk Fungal Biodiversity Institute & Fungal Molecular Physiology, Utrecht University, Uppsalalaan 8, 3584 CT Utrecht, The Netherlands. [5] Department of Microbiology, Faculty of Agriculture and Forestry, University of Helsinki, Viikinkaari 9, Helsinki, Finland. [6] US Department of Energy Joint BioEnergy Institute, 5885 Hollis St., Emeryville, CA 94608, USA. [7] Architecture et Fonction des Macromolécules Biologiques, (CNRS UMR 7257, Aix-Marseille University, 163 Avenue de Luminy, Parc Scientifique et Technologique de Luminy, 13288 Marseille, France. [8] Department of Plant and Microbial Biology, University of California, 111 Koshland Hall, Berkeley, CA 94720, USA. [9] Kanazawa Institute of Technology, 3 Chome-1, 924-0838 Yatsukaho, Hakusan-shi, Ishikawa-ken, Japan. [10] Environmental Molecular Sciences Division, Earth and Biological Sciences Directorate, Pacific Northwest National Laboratory, 902 Battelle Blvd, Richland, WA 99354, USA. *email: MRRA@novozymes.com

Aspergillus section *Flavi* encompasses a large number of species, many of which have a significant impact on human life: some species (e.g., *A. oryzae* and *A. sojae*) are routinely used in production of sake, miso, soy sauce, and other fermented foods. Moreover, *A. oryzae* is used industrially for production of enzymes and secondary metabolite production[1–4]. In contrast, other *Flavi* species (e.g., *A. flavus* and *A. parasiticus*) are notorious for producing highly toxic fungal compounds (e.g., aflatoxins), in addition to infecting and damaging crops[5–7]. Furthermore, *A. flavus* has been shown to infect immunocompromised humans, and is currently the second most common cause of human aspergillosis[8,9].

In addition, the section includes less known species that, similar to their (in)famous relatives, display both beneficial and harmful properties. The benefits are found in producers of bioactive compounds (such as the anti-insectant N-alkoxypyridone metabolite, leporin A, from *A. leporis*; an antibiotic with antifungal activity, avenaciolide, from *A. avenaceus*) and enzyme producers (including amylases, proteases, and xylanolytic enzymes in *A. tamarii* and pectin-degrading enzymes in *A. alliaceus*). On the harmful side, plant pathogens (*A. alliaceus* on onion bulb, *A. nomius* on nuts, seeds, and grains) and toxin producers (ochratoxin from *A. alliaceus*, aflatoxin from *A. nomius*) are also found among these less studied *Flavi* species[10,11], for which no genome sequences have previously been available.

Given the importance of section *Flavi*, it is highly valuable to examine the full genetic potential of the section in order to assess alternative species for industrial use, combat pathogenicity, find novel bioactives, and to identify useful enzymes. Prior to this project, whole-genome sequences were only available for five species from section *Flavi* (*A. oryzae, A. flavus, A. sojae, A. luteovirescens* (formerly *A. bombycis*), and *A. parasiticus*[3,12–15]). They all belong to a closely related clade within the section and thus cover only a small part of the diversity.

In this study, as part of the *Aspergillus* genus-sequencing project[16,17], we have generated genome sequences for 18 additional species plus an additional *A. parasiticus* isolate, permitting genomic comparisons across 23 members of section *Flavi* containing at least 29 species[10]. We apply these sequences in tandem with experimental and phenotypic data on secondary metabolite production, growth characteristics, and plant polysaccharide degradation to link phenotypes to genotypes and quantify the genetic potential of the section. Our analysis is useful for (1) exploring novel enzymes and secondary metabolites, (2) optimizing food fermentation and industrial use, and (3) improving food and feed protection and control.

Results and discussion

Assessment of 19 newly sequenced section *Flavi* genomes. In this study, we present the whole-genome sequences of 19 species from *Aspergillus* section *Flavi* (Fig. 1b). Two of these (*A. nomius* and *A. arachidicola*[18,19]) were also published by other groups in parallel to this work. We compare these 19 to previously sequenced section *Flavi* species (*A. oryzae, A. flavus, A. sojae, and*

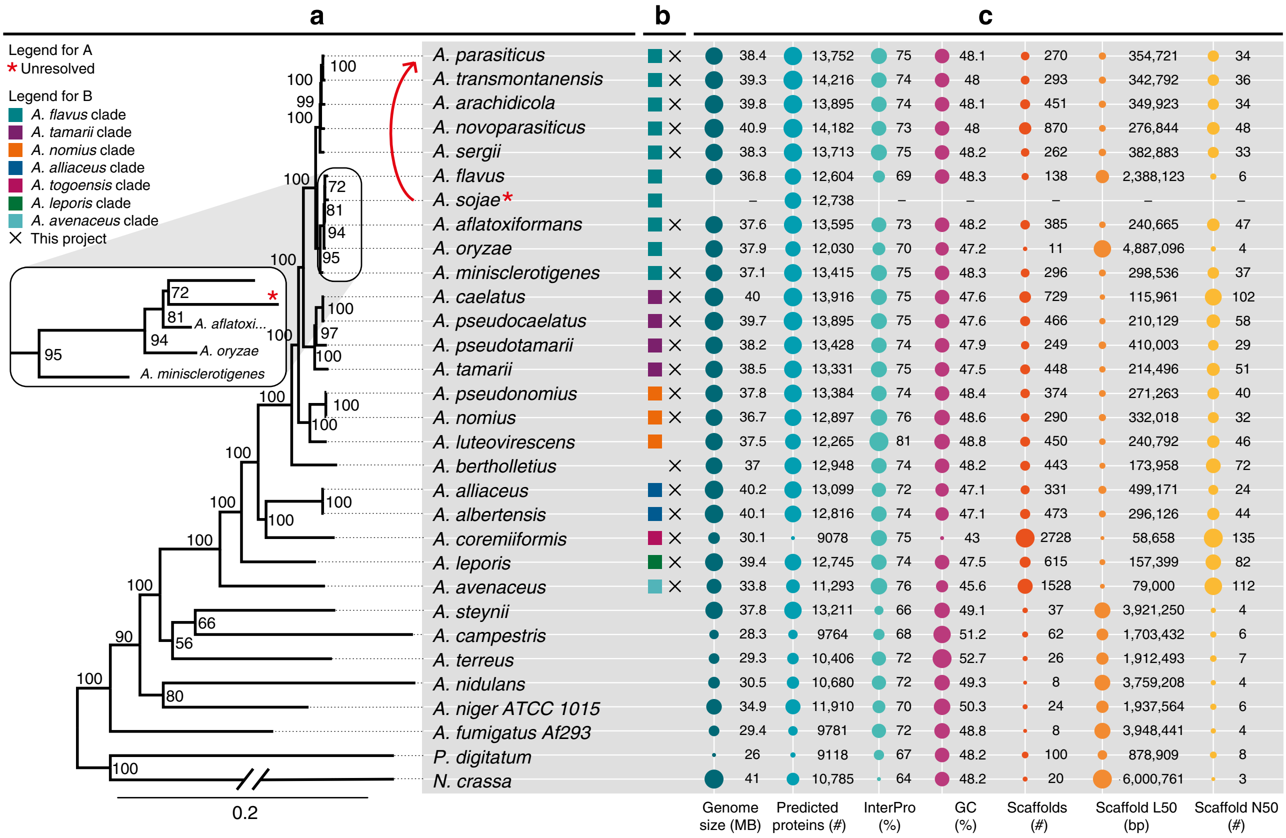

Fig. 1 Phylogeny and genome statistics of section *Flavi* plus eight other *Aspergillus, Penicillium,* and *Neurospora* species. a Phylogenetic tree constructed using RAxML, MUSCLE, and Gblocks based on 200 monocore genes (a single homolog in each of the species). The red star indicates an uncertain leaf most likely caused by a different gene calling method[98–100], and the arrow shows where *A. sojae* should be placed in the phylogenetic tree. The zoom shows the branching in a clade around *A. oryzae*. **b** The colors illustrate the clades found within section *Flavi* and X indicates species sequenced in this study. Earlier sequenced genomes such as *A. oryzae* and *A. fumigatus* were assembled using optical mapping and genetic maps. **c** Seven bubble plots illustrating key genome numbers and sequencing quality parameter. The bubble sizes have been scaled to each panel and are not comparable across panels.

A. luteovirescens[3,12–14]) as well as eight reference species: six from the rest of genus *Aspergillus* plus *Neurospora crassa* and *Penicillium digitatum* as outgroups (Fig. 1a, b).

As a first basis test, the quality of the genome assemblies was compared based on genome size, GC content, and number of predicted proteins (Fig. 1c). This showed a reasonable draft genome quality with 13 out of the 18 genomes assembled into fewer than 500 scaffolds (Fig. 1c, column 5). One cause of alarm was *A. coremiiformis* with 2728 scaffolds, which made us concerned with the quality of the gene content. However, the genome covers 99.78% of the Benchmarking Universal Single-Copy Orthologs (BUSCO[20]), and 96% of the expressed sequence tag (EST) clusters can be mapped to the genome. We thus conclude that the genome annotation is of a high enough quality for comparisons of the gene content despite the large number of scaffolds.

Section *Flavi* species generally have expanded genomes. The genome sizes of *Aspergillus* section *Flavi* are generally large compared with other representative Aspergilli (average of 37.96 Mbp vs. 31.7 Mbp (Fig. 1c)), as was previously reported for *A. oryzae*[21]. One major exception is *A. coremiiformis*, which has both fewer genes and a notably smaller genome, making it unique in the section.

Multigene phylogeny shows complex heritage of *A. oryzae*. Next we examined the evolutionary relationships in section *Flavi* based on a phylogeny derived from 200 genes (Fig. 1a). The support of the branching within the tree is high (100 out of 100 bootstraps in most branches). The tree confirms that section *Flavi* is a monophyletic group. The clades in Fig. 1a correspond to a previously reported phylogenetic tree based on the beta-tubulin gene[10,11,22] and the distances between sections correspond to previous work[23].

One potential error in the tree is that *A. sojae* is found closest to *A. flavus*, since *A. sojae* is perceived as a domesticated version of *A. parasiticus*. This branching indeed also has the lowest bootstrap value in the tree. The most likely explanation is that since the *A. sojae* gene predictions are based on the *A. flavus* and *A. oryzae* genome annotations[24,25], a bias is created in the predicted genes and this bias is likely reflected in the tree. As a test, we have generated phylogenetic trees using alternative methods not dependent on gene annotation (CVTree[26,27]). These clearly show that *A. sojae* is closest to *A. parasiticus*, both when using whole-genome and proteome sequences (Supplementary Fig. 1 and Supplementary Fig. 2). We hence think that *A. sojae* should be placed next to *A. parasiticus* in the phyogenetic tree as the arrow indicated in Fig. 1a.

Furthermore, *A. oryzae*, perceived as a domesticated version of *A. flavus*[10,28–30], is not directly next to it in the tree. However, it has previously been suggested that *A. oryzae* descends from an ancestor that was the ancestor of *A. minisclerotigenes* or *A. aflatoxiformans*[31]. The phylogeny (Fig. 1a, zoom) supports this suggestion, showing that *A. minisclerotigenes* and *A. aflatoxiformans* are closer relatives of *A. oryzae* than *A. flavus*.

Analysis of shared proteins confirms high genetic diversity. In order to examine core features shared by all section *Flavi* species, clades, as well as features of individual species, we made an analysis of shared homologous genes within and across species[16], and sorted these into homologous protein families (Fig. 2). This allowed the identification of (1) The core genome—protein families with at least one member in all compared species. This is expected to cover essential proteins. (2) Section-specific and clade-specific genes—genes that have homologs in all members of a clade/section, but not with any other species. (3) Species-specific genes—genes without homologs in any other species in the comparison.

The *core genome* of all 31 species in this dataset is 2082 protein families. For the 29 *Aspergillus* species this number is 3853, and for section *Flavi* species alone constitutes 4903 protein families. Thus, more than half the genome of the section *Flavi* species varies across the species.

Examining the clade-specific protein families, only very few (27–54) are found (Fig. 2a), which is low compared with section *Nigri* examined previously[16]. As sections *Nigri* and *Flavi* are roughly equally species-rich, this could indicate that the species in section *Flavi* are more distinct. This is supported by the fact that

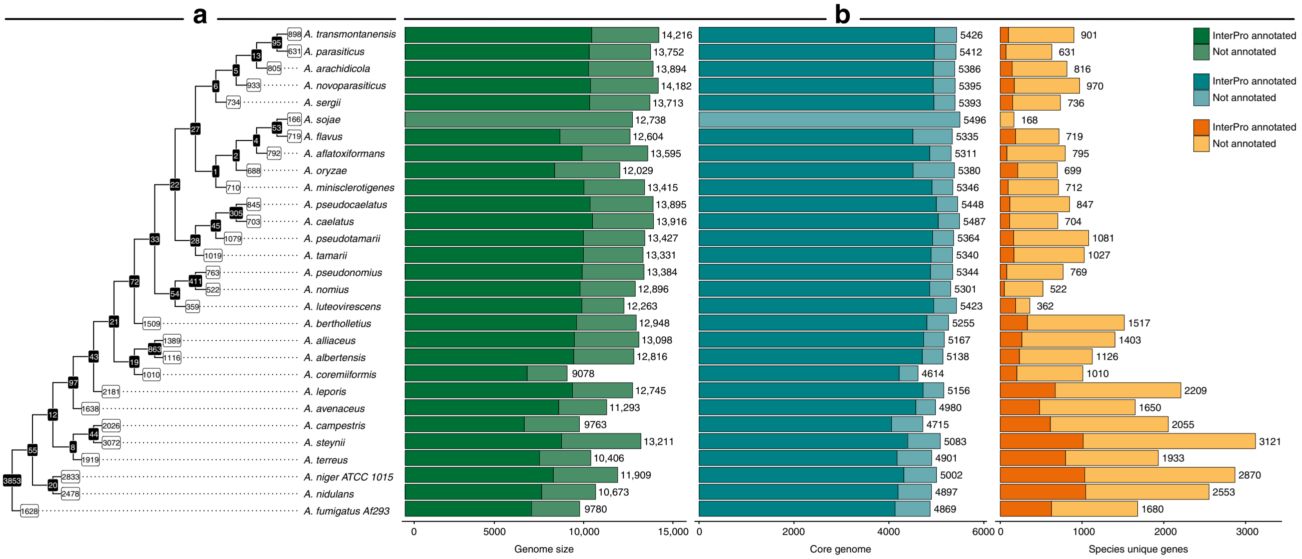

Fig. 2 Core-specific, section-specific, and clade-specific and species-unique genes. a A dendrogram representing the phylogenetic relationship between the 29 Aspergilli. The black boxes in the nodes represent the homologous protein families shared among the species branching from that node. The white boxes at the tips represent the protein families unique to that individual species. **b** A barplot showing the number of total (green), core (turquoise), and species-specific (orange) proteins for each species. The dark shading illustrates the number of proteins with a least one functional annotation based on InterPro[32].

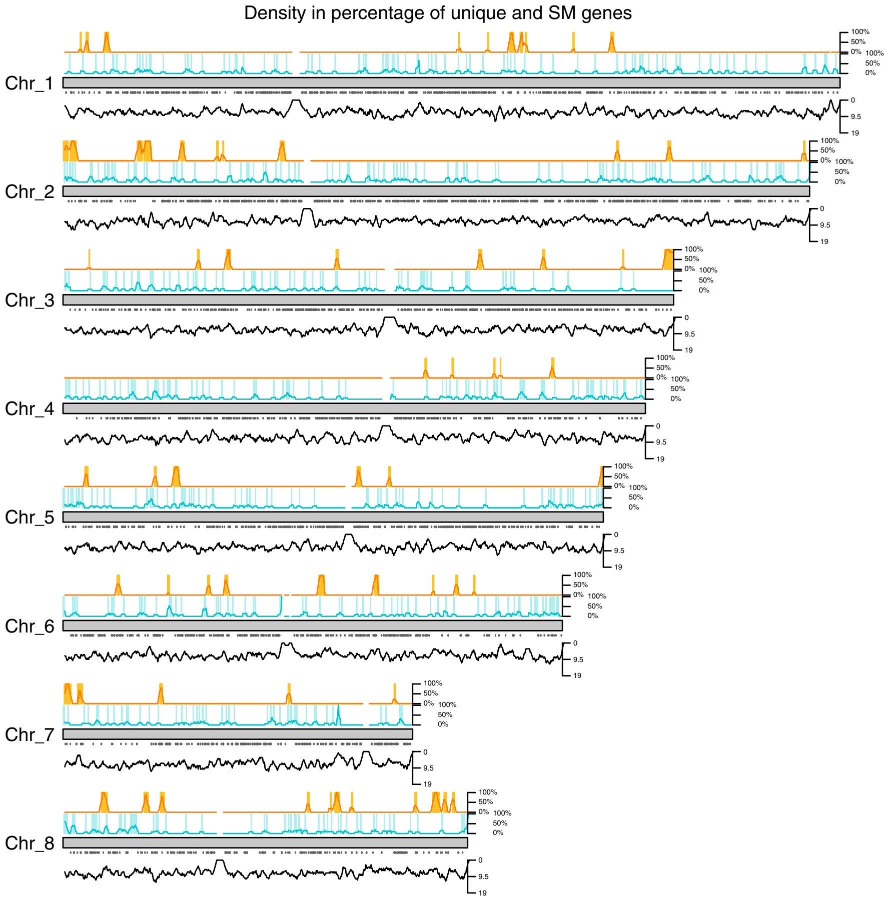

Fig. 3 Location of species unique and secondary metabolite genes in the *A. oryzae* genome. The gray bars represent the *A. oryzae* genome. Above the chromosome the species-specific (turquoise) and secondary metabolite genes (orange) are mapped to the genome, each line represents a gene. The curve shows the percentage of the density calculated from the total number of genes within 30 kbp in steps of 5 kb. Below the genome, the core genes are mapped by the gray dots and the density of the total number of genes is shown by the black graph (with a window of 30 kbp).

the number of species-specific genes are very high (166–2181), where we see 166 (*A. sojae*) to be an artificially low number, due to the gene calling in this genome being based on *A. flavus* and *A. oryzae* genomes.

Species-specific genes often encode regulation and P450s. We wanted to see whether the species-specific genes could be linked to known *Flavi* functions such as food fermentation and plant and human pathogenicity. In order to do this, we examined predicted functions of the species-specific genes using InterPro, GO and KOG annotations[32–35]. The portion with a functional annotation was low; 20, 12, and 9% for InterPro, GO, and KOG, respectively; in total 21% had an annotation (Supplementary Figs. 3–5). This is a very high—but not unusual—percentage of unidentifiable functions.

We will focus on InterPro since it covers more genes: the most common InterPro functions include transcription factors, protein kinases, transporters, and P450s (Supplementary Fig. 3), which are also significantly overrepresented. While these traits cannot directly be linked to food fermentation and pathogenicity, regulation is involved in adaptation and P450s play roles in both substrate degradation and production of bioactive compounds, both of which are relevant for fungal pathogenicity.

Species genes are over-represented in sub-telomeric regions. It has been shown that the sub-telomeric sequences are extensively rearranged regions in *A. nidulans, A. oryzae*, and *A. fumigatus*[21]. This is also seen in mammals, nematodes, and yeasts[36]. Previous studies[37,38] showed that sub-telomeric regions have a bias for unique, diverged, or missing genes. Another study has shown secondary metabolite gene clusters (SMGCs) to be enriched in sub-telomeric regions in *A. nidulans* and *A. fumigatus*[21].

We therefore examined the gene density and location of species-specific genes, secondary metabolite clusters, and core genome, by using the telomere-to-telomere *A. oryzae* genome as a reference in order to assess the potential overrepresentation of these genes in the sub-telomeric regions (Fig. 3).

Both visual inspection and Fisher's exact test confirmed that both species-specific (p-value = 7.266e–07) and SMGCs (p-value < 2.2e–16) are enriched toward the sub-telomeric regions (100 kbp from the chromosomal ends), where core genes are found less often at the sub-telomeric regions. The fact that the species-specific genes are not randomly distributed argues against that they are simply annotation or gene modeling errors, therefore indicating that they are, indeed, legitimate genes. The distribution of the species-specific genes suggests that new genes are more frequently successfully incorporated into the sub-telomeric regions than other locations. Whether this is the result of a selection for the sub-telomeric region, or a counterselection against other regions, or both, the data do not reveal.

Synteny analysis reveales islands of highly variable gene content. Syntenic and non-syntenic regions are another factor to consider when analyzing genome location. It has been shown that the *A. oryzae* genome has a mosaic pattern of syntenic and

Table 1 Percentage of genome with conserved synteny relative to *A. oryzae*.

Species	# Syntenic genes	% of *A. oryzae*
A. parasiticus	8199	68.15
A. transmontanensis	8238	68.48
A. arachidicola	8817	73.29
A. novoparasiticus	8102	67.35
A. sergii	8091	67.26
A. flavus	8686	72.20
A. aflatoxiformans	9094	75.59
A. oryzae	-	-
A. minisclerotigenes	8498	70.64
A. caelatus	7411	61.60
A. pseudocaelatus	7503	62.37
A. pseudotamarii	7494	62.29
A. tamarii	7471	62.10
A. pseudonomius	7179	59.68
A. nomius	7269	60.42
A. luteovirescens	7863	65.36
A. bertholletius	6801	56.53
A. alliaceus	6021	50.05
A. albertensis	5998	49.86
A. coremiiformis	5425	45.10
A. leporis	5800	48.21
A. avenaceus	5351	44.48
A. nidulans	4272	35.51
A. fumigatus	4876	40.53

non-syntenic regions relative to distantly related Aspergilli[1,2]. We examined the synteny across section *Flavi* and into *A. nidulans* and *A. fumigatus* using *A. oryzae* RIB40 as reference (Table 1). This analysis supports our earlier finding that *A. oryzae* is closely related to *A. aflatoxiformans* than *A. flavus*.

An overview of shared syntenic genes are illustrated in Supplementary Fig. 6. In general, there are fewer regions of synteny toward the telomeric ends as previously seen[1,2] in a comparison of *A. nidulans*, *A. fumigatus*, and *A. oryzae*. We further observed that chromosomes 1 and 2 have a very high degree of conserved synteny, while chromosomes 6 and 8 have a much lower conservation of synteny.

We find dense islands of non-syntenic genes in non-subtelomeric regions on chromosomes 4, 6, and 8. These could be caused by horizontal gene transfer (HGT), gene shuffling, or de novo gene formation. We investigated for HGTs using BLASTp to examine the best hits in the NCBI nonredundant database. Recent HGTs are expected to have high sequence identity with another group of species where it would have been transferred from, and not be found in the closely related species[39]. None of these islands showed signs of recent HGTs. Furthermore, only 23 of the 80 genes in the non-syntenic blocks were *A. oryzae*-specific. It thus seems likely that these non-syntenic islands are caused by a mix of significant rearrangements, duplication events, and the emergence of *A. oryzae*-specific genes.

Taken together, the fact that we observe some very conserved chromosomes and some highly rearranged non-syntenic blocks could indicate an evolutionary pressure for stability in some regions while other regions are frequently subject to gene shuffling and rearrangements, i.e., rearrangement hot spots.

Section *Flavi* is a rich source of carbohydrate-active enzymes. Carbohydrate-Active enZymes (CAZymes) are essential for what carbon sources a species can degrade and utilize. Within section *Flavi* the CAZymes/carbon utilization is mainly described for *A. oryzae*[1,2,40] and to a lesser extent for *A. flavus*[41–45] and *A. sojae*[46,47], while only incidental studies have been performed with other species of this group[48–54], often describing production or characterization of a certain CAZyme activity or protein, respectively.

We used the CAZy database to predict the CAZyme content in the genomes of the section (Fig. 4). A total of 13,759 CAZymes were predicted for the 23 *Flavi* species (average 598/species). This is quite rich compared with included reference Aspergilli (508/species).

It is clear from this analysis that there is a distinct difference between the clades of section *Flavi* (Fig. 4b), showing again a variation in gene content in the section.

Variable CAZyme content does not reflect the ability to degrade plant biomass. To evaluate the actual carbon utilization ability across section *Flavi*, we performed growth profiling of 31 species (29 Aspergilli, including 23 species from section *Flavi*) on 35 plant biomass-related substrates (Fig. 5, Supplementary Data 1) and compared this with the CAZyme gene content prediction related to plant biomass degradation (Supplementary Data 2). In a previous study, the variation in growth between distantly related Aspergilli could be linked to differences in CAZyme gene content[55], but this was not the case for closer related species from *Aspergillus* section *Nigri*[16].

Glucose resulted in the best growth of all monosaccharides for all species and was therefore used as an internal reference for growth (Supplementary Fig. 7). Growth on other carbon sources was compared with growth on D-glucose and this relative difference was compared between the species. Growth on monosaccharides was largely similar between the species of section *Flavi* (Fig. 5, Supplementary Fig. 7, and Supplementary Data 1).

The CAZyme sets related to plant biomass degradation are overall highly similar for section *Flavi* (Fig. 5), with the exception of *A. coremiiformis*, which has a strongly reduced gene set. This is mainly due to reduction in glycoside hydrolase families, but also a number of families related to pectin, xylan, and xyloglucan degradation. Surprisingly, this species showed better relative growth on xylan than most other species, while the growth on other polysaccharides was mainly similar to that of section *Flavi*. Thus, the reduced gene set has not reduced its ability to degrade plant biomass. This could be similar to the case of *T. reesei*, which also has a reduced CAZyme gene set, but produces the corresponding enzymes at very high levels[56]. However, the origin of this approach is likely very different as its CAZyme content was shaped by loss and then massive HGT gain of plant cell wall degrading enzymes[57], while no indications for this are present for *A. coremiiformis*.

Hydrolytic differences are clade-specific within section *Flavi* (Supplementary Data 2). The *A. togoensis* clade has a reduced set of xylanolytic and xyloglucanolytic genes, but this is not reflected in the growth. In contrast, GH115 (alpha-glucuronidase) genes are expanded in clades *A. flavus, A. tamarii*, and *A. nomius* (xylanolytic enzymes or activity have been reported from several species from these clades[49–51,53,58–62]), GH62 (arabinoxylan arabinofuranohydrolase) was expanded in clade *A. leporis*, and clades *A. leporis* and *A. avenaceus* were the only clades with CE15 (glucuronoyl esterases), which were also found in *Aspergillus* species outside section *Flavi*.

The galactomannan degrading ability was nearly fully conserved in section *Flavi*, but interestingly growth on guar gum that consists mainly of galactomannan was variable between the species. Similarly, the reduced amylolytic ability of clades *A. togoensis* and *A. avenaceus* did not result in reduced growth on starch or maltose.

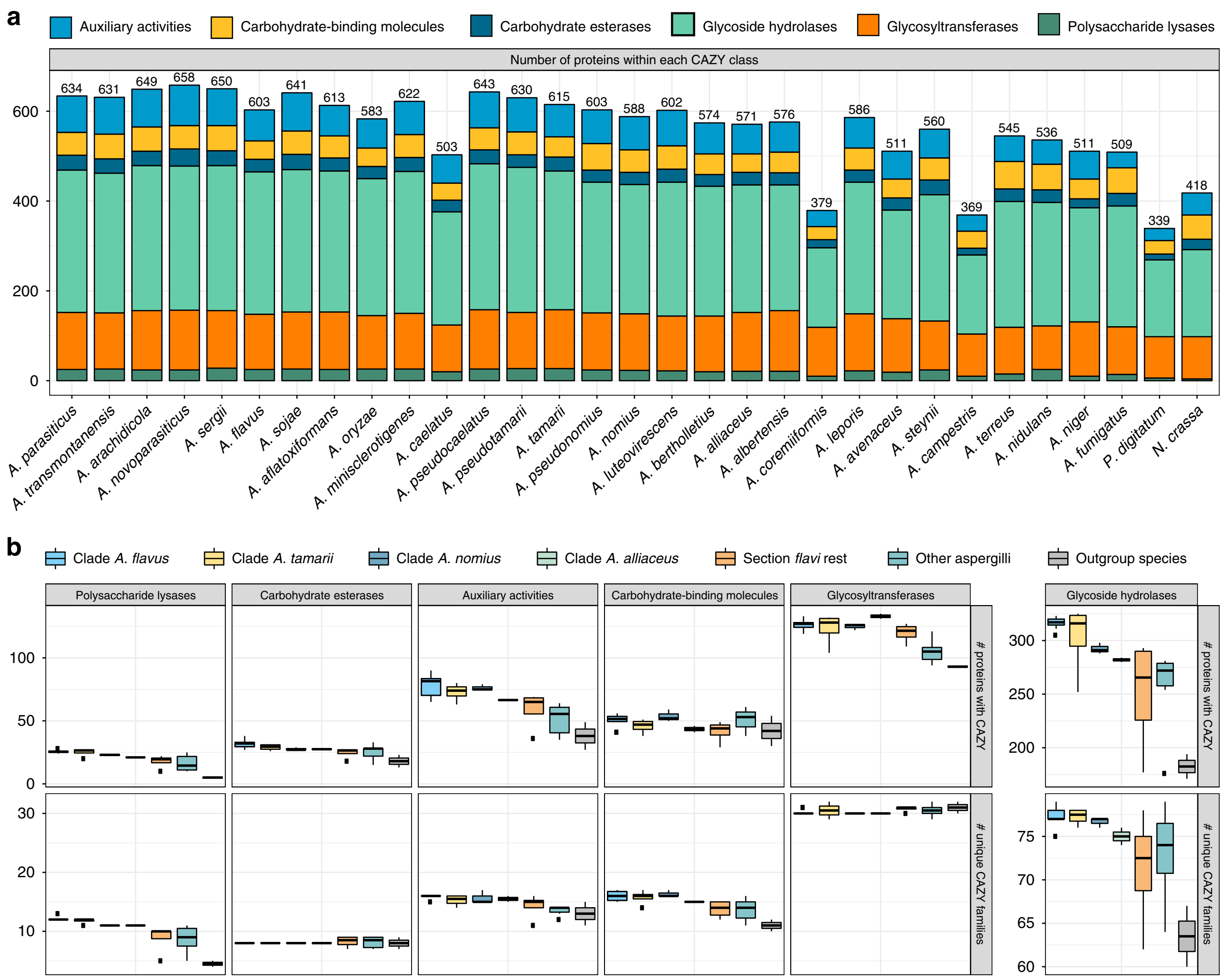

Fig. 4 Carbohydrate-active enzymes (CAZymes) in section *Flavi*. a The total number of CAZymes in each species distributed on six categories of enzyme activity: auxiliary activities, carbohydrate-binding molecules, carbohydrate esterases, glycoside hydrolases, glycosyltransferases, and polysaccharide lyases. **b** Boxplot representing the diversity of CAZyme family content and abundance among clade *A. flavus* (light blue), *A. tamarii* (yellow), *A. nomius* (dark blue), *A. alliaceus* (light turquoise), the rest of the *Flavi* section (orange), other Aspergilli (dark turquoise), and non-Aspergillus species (gray). For each CAZyme class the total number of CAZymes (top row) and the number of unique CAZyme families (bottom row) are displayed. In the boxplot the midline represents the median, the upper and lower limit of the box represents the third and first quartile, and the whiskers extend up to 1.5 times the interquartile.

Variation was observed in the number of pectinolytic genes. The most pronounced differences were the absence of PL11 (rhamnogalacturonan lyase) genes from most species of section *Flavi*, and the expansion of GH78 (alpha-rhamnosidase) in clades *A. flavus* and *A. tamarii*. However, these differences and the smaller ones in other families did not result in large variation in growth on pectin.

More obvious differences were present during growth on cellobiose, lactose, and lignin. Most species grew poorly on cellobiose despite similar numbers of beta-glucosidase-encoding genes in most species (Supplementary Data 2). Similarly, only *A. arachidicola*, and to a lesser extent *A. albertensis* grew well on lactose, while the number of beta-galactosidases in these species is similar to that of the other species. Most interesting was the finding that *A. albertensis* grew as well on lignin as on D-glucose, suggesting potential applications in biofuel production.

In summary, CAZyme potential in section *Flavi* is largely conserved (with the exception of *A. coremiiformis)* with some variations in copy numbers, but the genomic potential and variations are not necessarily reflected in the growth. It is therefore likely that as suggested previously[55], the observed differences are largely at the regulatory level.

CAZyme family GH28 is inflated in clade *A. flavus*. We were particularly interested in GH28 CAZymes, as they are important for food fermentation and the quality of the final fermented product[63]. A phylogenetic tree was created of all members of GH28 from section *Flavi* (Supplementary Fig. 8). The tree consists of 429 proteins, on average 18.7 per species.

Within the tree there are different groupings. Five groups have members from all 23 species, nine groups are missing one to four species (usually *A. coremiiformis* and *A. caelatus*), and two groups are specific to the *A. flavus, A. tamarii*, and *A. nomius* clades. Last there are eight groups containing 2–13 species, which do not follow the phylogeny—suggesting these to be sources of GH28 variation.

In general, species from clade *A. flavus* have a high number of GH28 members. *A. sojae* is known to have a high number of GH28, which is also seen here with 24 members; however, *A. sergii* has an even higher number with 25 members. It could be interesting to investigate if this could be exploited either by using *A. sergii* as a new species in food fermentation and/or as a source of novel enzymes.

Analysis of secondary metabolism. The genus *Aspergillus* is known to produce a large number of SMs and the number of

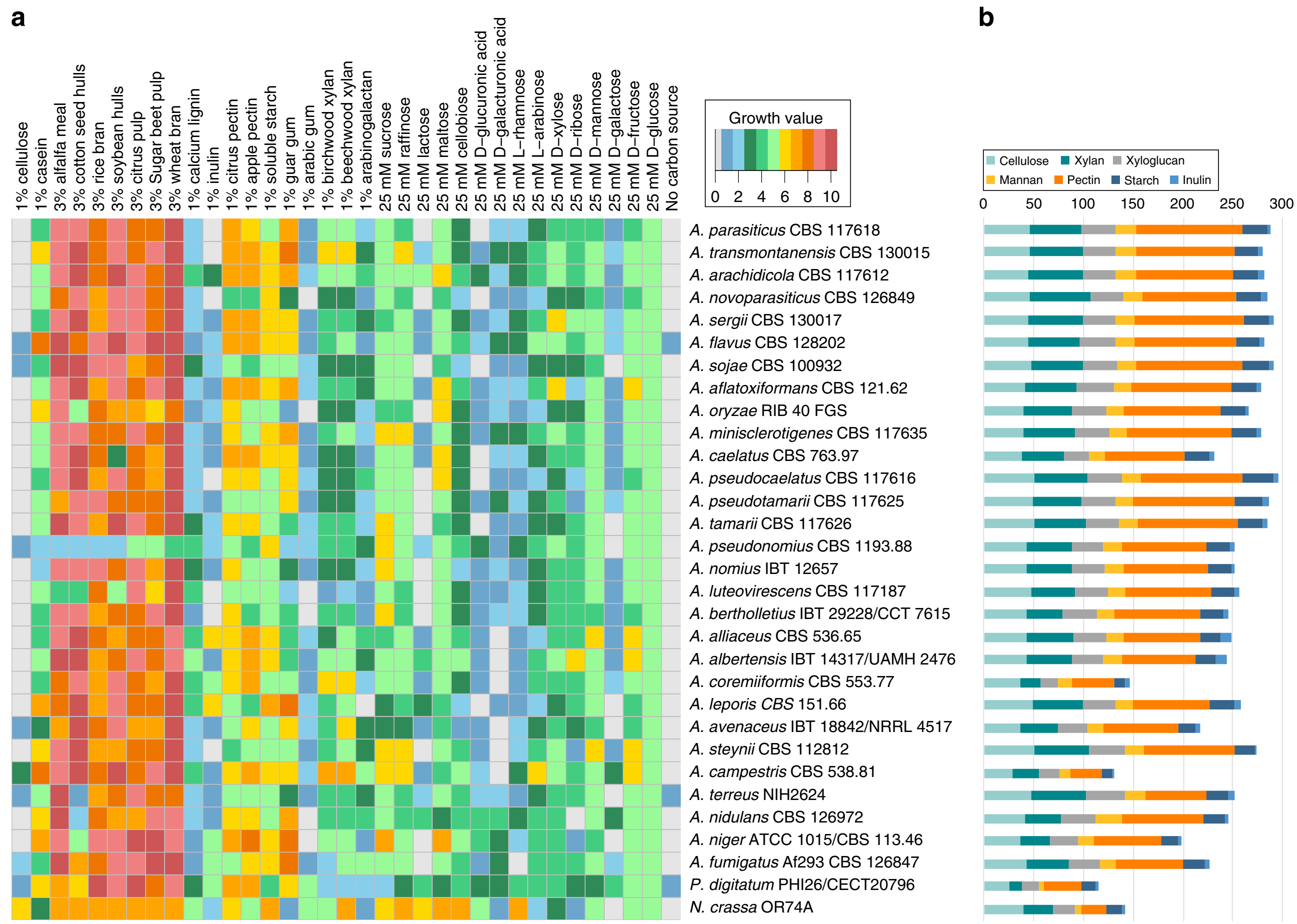

Fig. 5 Carbohydrate-active enzymes in section *Flavi* sorted according to the phylogram of Fig. 1. a Heatmap representing the growth profiles of 23 *Flavi* species and 8 additional species on 35 different media. **b** Comparison of the CAZyme sets related to plant biomass degradation in the genomes of species from *Aspergillus* section *Flavi*, and some other fungi. The colors reflect the polysaccharides the enzymes are active toward.

predicted SMGCs is even higher. The majority of predicted SMGCs are uncharacterized and therefore have the potential to produce a diversity of novel, bioactive compounds. We examined the diversity and potential for SM production in section *Flavi*, both quantitatively in terms of numbers of clusters, and qualitatively in terms of the compounds these clusters could potentially produce.

Secondary metabolism in section *Flavi is* diverse and prolific. To quantitatively assess the potential for SM production, SMGCs were predicted using a SMURF-like prediction tool[64] for all species except *N. crassa* and *A. sojae*, since these were sequenced by other methods and with dissimilar gene calling methods (Fig. 6c). Within the 28 *Aspergillus* species, there is a total of 1972 predicted SMGCs and for section *Flavi* genomes, the total is 1606 SMGCs (73/species). This is more than 15 extra per species compared with the very prolific *Penicillium* genus[65].

We wanted to examine how unique the SMGCs are, and thus constructed families of SMGCs (Supplementary Data 3). For the entire dataset, we could collapse it into 477 SMGC families, and for section *Flavi* 308 SMGC families. Out of these, 150 SMGC clusters are only found in one section *Flavi* species (Fig. 6a), showing a large number of unique clusters in each species (6.8 unique SMGCs/species). Compared with *Aspergillus* section *Nigri*, the number of clusters per species in this study is slightly lower, but the number of members in each SMGC family is also lower, demonstrating greater diversity in secondary metabolism in section *Flavi* compared with section *Nigri*.

Dereplicating secondary metabolism predicts toxin producers. To assess the potential for SM production qualitatively, we used a pipeline of "genetic dereplication" where predicted clusters are associated with verified characterized clusters (from the MIBiG database[66]) in a guilt-by-association method[67]. Based on this, 20 cluster families were coupled to a compound family (Fig. 6b). Some cluster families were found in all or nearly all *Flavi* genomes, e.g., those similar to the naphthopyrone[68], nidulanin A[69], azanigerone[70], 4,4′-piperazine-2,5-diyldimethyl-bis-phenol, and aflavarin[71]/endocrocin[72,73] clusters. Most families generally follow the phylogenetic groups, suggesting a loss-based distribution pattern, but some, like the SMGC families similar to the asperfuranone[74], pseurotin A[75], or fumagillin[76] clusters did not follow the phylogeny. Moreover, potential producers of known toxins such as aflatoxin and aspirochlorine were identified (Fig. 6b).

Combination of data and analysis links a compound to a cluster. Extending from the known SMGC clusters, we were interested in linking compounds and clusters based on the presence/absence pattern of produced compounds and predicted clusters. We therefore created a heatmap of all the cluster families found in at least five species, added the predicted compound families from the MIBiG dereplication, in addition to manually curated compound families from a literature survey (Supplementary Fig. 9). In addition to this, we measured the SM production of the *Flavi* species (Supplementary Data 4).

Of particular interest was miyakamides. They are originally isolated from an *A. flavus* isolate and shown to have antibiotic

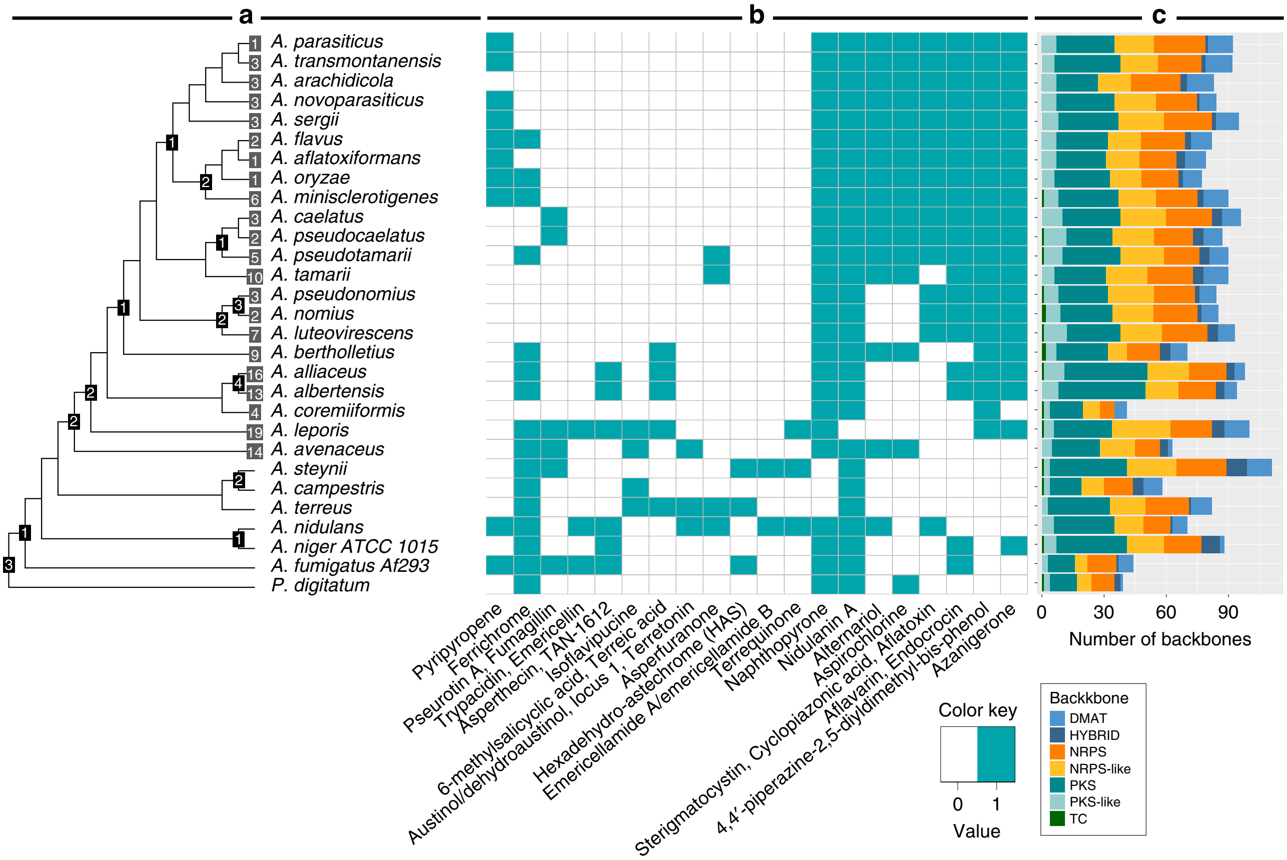

Fig. 6 Dereplication of known compounds and predicted secondary metabolite backbone genes per species. a A dendrogram representing the phylogenetic relationship between the species. The black boxes in the nodes represent the secondary metabolite gene cluster (SMGC) families shared among the species branching from that node. If there is no black box there are zero clusters shared. The gray boxes at the tips show the number of unique SMGC families only found in one species for the *Flavi* section. **b** Matrix indicating the presence and absence of SMGC families coupled to known clusters from the MIBiG database[66] for each species. Overview of the cluster family for aflatoxin can be found in Supplementary Figure 11. **c** Predicted secondary metabolite genes for each species divided by the backbone enzyme. DMAT: dimethylallyltransferase (prenyl transferases), HYBRID: a backbone 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 or one NRPS A domain in combination with NAD binding 4 domain or short-chain dehydrogenase, PKS: polyketide synthase, PKS-like: polyketide synthase like, containing at least two PKS-specific domains and another domain, TC: terpene cyclase.

properties[77], but the biosynthetic gene cluster is not known. Our chemical analysis showed production in *A. sojae, A. nomius, A. parasiticus, A. novoparasiticus*, and *A. transmontanensis*.

We performed retro-biosynthesis from the chemical structure and predicted that the biosynthetic gene cluster should contain a nonribosomal peptide synthetase (NRPS) with 2–3 adenylation domains (since two of the three amino acids are similar), an N-methyltransferase, an acetyltransferase, and potentially a decarboxylase/dehydrogenase (Supplementary Fig. 10A). Searching for cluster families with members in all the miyakamide-producing species having NRPS backbones with 2–3 adenylation domains and a methyltransferase domain, only one cluster family met the requirements. The cluster family has a NRPS backbone with a methyltransferase domain, three A domains in most species, and two in *A. novoparasiticus*. The prediction of only two A domains is most likely caused by annotation error since the sequence similarity is conserved before the start of the gene (Supplementary Fig. 10B). The size of the predicted cluster is 1–9 genes, the difference is likely caused by SMGC prediction errors (Synteny plot in Supplementary Fig. 10B). The synteny plot shows that the NRPS and two small genes with unknown function are widely conserved. We thus propose that the identified NRPS along with the two conserved genes of unknown function are likely candidates for miyakamide biosynthesis.

The aflatoxin biosynthetic gene cluster is highly conserved. Perhaps the best known secondary metabolite in section *Flavi* is the highly carcinogenic aflatoxin. Aflatoxins are known to be produced by many section *Flavi* species (*A. arachidicola, A. luteovirescens, A. flavus, A. minisclerotigenes, A. nomius, A. aflatoxiformans, A. pseudocaelatus, A. pseudonomius, A. pseudotamarii*, and some *A. oryzae* isolates)[4,10].

The dereplication analysis (Fig. 6b) identified a SMGC family predicted to be involved in sterigmatocystin and aflatoxin production, which is all the species in the *A. flavus, A. nomius*, and *A. tamarii* clades except *A. tamarii*. A synteny plot of the SMGC family (Supplementary Fig. 11) shows that the cluster is extremely well conserved with no rearrangements and a high alignment identity for the aflatoxin genes. Only *A. caelatus* has a truncated form with only the *aflB, aflC*, and *aflD* genes and *A. tamarii* seems to have a complete loss of the cluster. Interestingly, most of the predicted clusters did not include the *aflP* and *aflQ* genes that are responsible for the last step of aflatoxin biosynthesis. We searched the genomes for *aflP* (Supplementary Fig. 12), and found it in all genomes, but with different start sites and extra sequence in the middle of the proteins. RNA-seq data support these models (Supplementary Fig. 13) and suggest errors in the *A. flavus* gene models. Similarly, the *aflQ* gene is found in all the other species, but 5–10 genes away from the predicted clusters. Thus, detailed analysis shows that all these species have the genes required for aflatoxin biosynthesis.

Conclusion

We de novo sequenced the genomes of species representing various parts of the *Flavi* section, which allowed a section-wide comparison illustrating the similarities and diversity within the section. We show that *A. oryzae* is closely related to *A. minisclerotigenes* or *A. aflatoxiformans* based on a 200-gene phylogeny.

Members of the *Flavi* section have a large genome size compared with other Aspergilli. The large genome is reflected in the high number of SMGCs and CAZymes that could be a source of novel compounds and enzymes in the future.

We have shown that the aflatoxin cluster is highly conserved both concerning identity and synteny in the *A. flavus, A. nomius* clade, and partly in the *A. tamarii* clade where the cluster is partly lost in *A. caelatus* and completely lost in *A. tamarii*.

The number of species unique proteins is varying, but even with the very closely related *A. flavus* clade, most species have above 700 unique proteins illustrating the high diversity. Localization analysis of *A. oryzae* has shown the distribution of species unique genes and SMGC across the genome but with a higher density in the sub-telomeric ends. Synteny analyses have highlighted some tendencies of some highly conserved chromosomes and a few dense non-syntenic blocks that could represent rearrangement hot spots.

Overall the data and analysis presented here provide the fungal research community with a substantial resource, and set the stage for future research in the field.

Methods

Fungal strains. The species examined in this study (Supplementary Table 1) were from the IBT Culture Collection of Fungi at the Technical University of Denmark (DTU) or from the Westerdijk Fungal Biodiversity Institute (CBS), unless otherwise noted. Strains can be obtained from these sources.

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 days 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, CYA, MEAox, or CY20 depending on the strain (see Supplementary Table 1) cultivated for 5–10 days 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 (see ref. [78] and below) and checked for quality and concentration using a NanoDrop (BioNordika). 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.[79] was used. The same procedure was used previously[16,80].

DNA and RNA sequencing and assembly. All genomes and transcriptomes in this study were sequenced with Illumina. 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 repair, A tailing, adapter ligation, and ten 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 HiSeq 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 × 150 indexed run recipe.

After sequencing, the genomic FASTQ files were quality control-filtered to remove artifacts/process contamination and assembled using Velvet54. The resulting assemblies were used to create in silico long mate-pair libraries with inserts of 3000 ± 90 bp, which were then assembled with the target FASTQ using AllPathsLG release version R4771055. Illumina transcriptome reads were assembled into consensus sequences using Rnnotator v3.3.256.

Genome annotation. All genomes were annotated using the JGI annotation pipeline[81,82] as previously described[16,80]. Genome assembly and annotations are available at the JGI fungal genome portal MycoCosm[81] (see URLs) and have been deposited in the DNA Data Bank of Japan (DDBJ)/European Molecular Biology Laboratory (EMBL)/GenBank under the accession numbers provided in the Data Availability Statement.

Homologous protein families. All predicted proteins from the 31 genomes used in this study were aligned using the BLASTp function from the BLAST + suite version 2.2.27 with an (e-value < 10^{10}). The resulting 961 whole-genome BLAST tables were analyzed to identify homologous proteins and group them into families as described previously[16].

Protein families containing at least one protein from all species were defined as core families, while species-unique families were defined as families containing one or more protein(s) from only one species.

Functional annotation. Functional domains were identified in all the proteins using InterPro[32], GO[34], and KOG[35].

Phylogeny. Monocore genes were identified as protein families having exactly one member in each species. Each protein family was aligned using MUSCLE version 3.8.31 (default settings) and then trimmed using gblocks version 0.91b ($-t = p\ -b4 = 5\ -b5 = \mathrm{h}$). Following 200 of these monocore sequences (with length between 150 and 1000 AA) were selected randomly and concatenated and used to construct a phylogentic tree using RaxML version 8.2.8 using the PROTGAMMAWAG substitution model and 1000 bootstraps.

Prediction of CAZymes. CAZymes were predicted using the CAZymes database (CAZy, www.cazy.org[83]) and the method described in our previous work[16]. Each *Aspergillus* protein model was compared using BLASTp with proteins listed in the CAZYmes database (CAZy)[83,84]. Models with over 50% identity over the entire length of an entry in CAZy were directly assigned to the same family (or subfamily when relevant). Proteins with less than 50% identity to a protein in CAZy were all manually inspected, and conserved features, such as the catalytic residues, were searched whenever known. Because 30% sequence identity results in widely different *e*-values (from nonsignificant to highly significant), for CAZy family assignments, we examined sequence conservation (percentage identity over CAZy domain length). Sequence alignments with isolated functional domains were performed in the case of multimodular CAZymes. The same methods were used for *Penicillium digitatum* and *Neurospora crassa*.

Prediction of secondary metabolite gene clusters. Secondary metabolite gene clusters (SMGCs) and SMGC families were predicted based on the SMURF algorithm[64] and the method described in our previous work[16]. For the prediction of SMGCs, we developed a command-line Python script roughly following the SMURF algorithm:

According to SMURF the following genes were predicted as "backbone" genes:

Genes that have at least three PFAM domains—ketoacyl-synt (PF00109), Ketoacyl-synt_C (PF02801), and Acyl_transf_1 (PF00698)—were predicted as 'PKS' genes.

Genes that have ketoacyl-synt (PF00109) and Ketoacyl-synt_C (PF02801) but not Acyl_transf_1 (PF00698) were predicted as "PKS-like" genes.

Genes that have at least 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 (PF07993), 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.

Secondary metabolite-specific PFAM domains were taken from Supplementary Data 1 of the SMURF paper[64]. 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 six genes without secondary metabolite-specific domains.

SMGC families were created based on the SMURF prediction comparisons using BLASTp (BLAST + suite version 2.2.27, *e*-value $\leq 1 \times 10^{-10}$). Subsequently, a score based on BLASTp identity and shared proteins was created to determine the similarity between gene clusters as depicted in the formula below. Using these scores, we created a weighted network of SMGC clusters and used a random walk community detection algorithm (R version 3.3.2, igraph_1.0.166) to determine families of SMGC clusters. Finally, we ran another round of random walk clustering on the communities that contained more members than species in the analysis (ptailoring/pbackbone = sum of percentage BLAST alignment of tailoring/backbone enzymes, respectively; ntailoring/nbackbone = number of tailoring/backbone enzymes with significant hits, respectively; ttailoring/tbackbone = total number of tailoring/backbone enzymes):

$$\text{ptailoring} \times \frac{\text{ntailoring}}{\text{ttailoring}} \times 0.35 + \text{pbackbone} \times \frac{\text{nbackbone}}{\text{tbackbone}} \times 0.65 \tag{1}$$

To create a cluster similarity score, a combined score of tailoring and backbone enzymes was created. The sum of the BLASTp percent identity (ptailoring/pbackbone) of all hits for tailoring enzymes between two clusters was divided by the maximum amount of tailoring enzyme (ttailoring/tbackbone) and multiplied by 0.35. Then the score for the backbone enzymes was calculated in the same manner but multiplied by 0.65 to give more weight to the backbone enzymes. The scores were added to create an overall cluster similarity score:

$$\text{avg}\left(\text{pident}_{\text{tailoring}}\right) \times 0.35 + \text{avg}(\text{pident}_{\text{backbones}}) \times 0.65 \tag{2}$$

Annotation of SMGC families using MIBiG (genetic dereplication). SMGC families were annotated based on the MIBiG database[67]. Known gene clusters were coupled to SMGC families, making it possible to predict the compounds or derivatives thereof a species can potentially produce. Cluster families containing one cluster highly similar to a known compound cluster are labeled after the known compound.

Genome synteny analysis. Orthologs were defined as a pair of genes found between two genomes from different species by bidirectional best hits using BLASTP with *e*-value < 10^{10}. When two genes within 10 kbp on the first genome have corresponding orthologs within 10 kbp on the second genome, the region between the two genes was defined as a syntenic block. The distance between the two genes was calculated by the formula, $|PC1 - PC2| - 1/2\ (LN1 + LN2)$, where PCn and LNn are nucleotide position of the center and nucleotide length of gene "*n*" ($n = 1$ or 2), respectively.

Analysis of chromosomal localization. For visualization of chromosome, chromosomal location, and gene density the R package karyoploteR was used[85].

Secondary metabolite gene cluster analysis and visualization. For visualization of cluster synteny and similarity EasyFig was used[86]. The parameters minimum length and minimum identity were set to 50 bp and 50%, respectively.

Profiling of growth on different carbon sources. The species were grown on 35 different media plates using the same method as first described by deVries et al.[55] All strains were grown on MM[87] containing monosaccharides/oligosaccharides, polysaccharides, and crude substrates at 25 mM, 1%, and 3% final concentration, respectively.

Chemical analysis of secondary metabolism. The section *Flavi* strains were cultivated as three-point cultures on CYA and YES media for 7 days in the dark at 30 °C; subsequently three plugs (6 mm inner diameter) were taken across the colony, 800 μL of isopropanol ethyl acetate (1:3 v/v) with 1% formic acid was added and ultrasonicated for 1 h. The liquid sample was transferred to another tube and evaporated; after this 300 μL of methanol was added to dissolve the pellets and the samples were ultrasonicated for 20 min[88–90]. Samples were then centrifuged at max g-power for 2–3 min, and afterward 150 μL of the supernatant was transferred to HPLC vials[91–97].

Ultra-high-performance liquid chromatography–diode array detection–quadrupole time-of-flight mass spectrometry (UHPLC–DAD–QTOFMS) was performed on an Agilent Infinity 1290 UHPLC system equipped with a diode array detector. Separation was obtained on a 250 × 2.1 mm i.d., 2.7 μm, Poroshell 120 Phenyl Hexyl column (Agilent Technologies, Santa Clara, CA) held at 60 °C. The sample, 1 μL, was eluted with a flow rate of 0.35 mL min^{-1} using A: a linear gradient 10% acetonitrile in Milli-Q water buffered with 20 mM formic acid increased to 100% in 15 min, staying there for 2 min before returning to 10% in 0.1 min, held for 3 min before the following run.

Mass spectrometry (MS) detection was performed on an Agilent 6545 QTOF MS equipped with an Agilent dual-jet stream electro spray ion (ESI) source with a drying gas temperature of 160 °C, gas flow of 13 L min^{-1}, sheath gas temperature of 300 °C, and flow of 16 L min^{-1}. Capillary voltage was set to 4000 V, and nozzle voltage, to 500 V in positive mode. MS spectra were recorded as centroid data, at an *m/z* of 100–1700, and auto MS/HRMS fragmentation was performed at three collision energies (10, 20, and 40 eV), on the three most intense precursor peaks per cycle. The acquisition was 10 spectra s^{-1}. Data were handled in the software Agilent MassHunter Qualitative Analysis (Agilent Technologies, Santa Clara, CA).

Reporting summary. 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/fungi). All new genomes published in the study have been deposited in GenBank under the following NCBI accession numbers: Aspergillus nomius IBT 12657—BioProject: PRJNA333904, Accession No: SWDY00000000; Aspergillus minisclerotigenes CBS 117635—BioProject: PRJNA333902, Accession No: SWDZ00000000; Aspergillus parasiticus CBS 117618—BioProject: PRJNA333906, Accession No: SWCZ00000000; Aspergillus novoparasiticus CBS 126849—BioProject: PRJNA333905, Accession No: SWDA00000000; Aspergillus leporis CBS 151.66—BioProject: PRJNA333901, Accession No: SWBU00000000; Aspergillus alliaceus CBS 536.65—BioProject: PRJNA334014, Accession No: SWAS00000000; Aspergillus parvisclerotigenus CBS 121.62—BioProject: PRJNA333907, Accession No: SWAT00000000; Aspergillus pseudotamarii CBS 117625— BioProject: PRJNA333910, Accession No: STFH00000000; Aspergillus avenaceus IBT 18842— BioProject: PRJNA333897, Accession No: STFI00000000; Aspergillus tamarii CBS 117626—BioProject: PRJNA333912, Accession No: STFJ00000000; Aspergillus transmontanensis CBS 130015—BioProject: PRJNA333913, Accession No: STFK00000000; Aspergillus sergii CBS 130017—BioProject: PRJNA333911, Accession No: STFL00000000; Aspergillus arachidicola CBS 117612—BioProject: PRJNA333896, Accession No: STFM00000000; Aspergillus coremiiformis CBS 553.77—BioProject: PRJNA333900, Accession No: STFN00000000; Aspergillus caelatus CBS 763.97—BioProject: PRJNA333899, Accession No: STFO00000000; Aspergillus bertholletius IBT 29228—BioProject: PRJNA333898, Accession No: STFP00000000; Aspergillus albertensis IBT 14317—BioProject: PRJNA333895, Accession No: STFQ00000000; Aspergillus pseudonomius CBS 119388—BioProject: PRJNA333909, Accession No: STFR00000000; Aspergillus pseudocaelatus CBS 117616—BioProject: PRJNA333908, Accession No: STFS00000000.

Received: 16 April 2019; Accepted: 2 December 2019;
Published online: 27 February 2020

References

1. Machida, M., Yamada, O. & Gomi, K. Genomics of *Aspergillus oryzae*: learning from the history of Koji Mold and exploration of its future. *DNA Res.* **15**, 173–183 (2008).
2. Kobayashi, T. et al. Genomics of *Aspergillus oryzae*. *Biosci. Biotechnol. Biochem.* **71**, 646–670 (2007).
3. Sato, A. et al. Draft genome sequencing and comparative analysis of *Aspergillus sojae* NBRC4239. *DNA Res.* **18**, 165–176 (2011).
4. Frisvad, J. C., Møller, L. L. H., Larsen, T. O., Kumar, R. & Arnau, J. Safety of the fungal workhorses of industrial biotechnology: update on the mycotoxin and secondary metabolite potential of *Aspergillus niger*, *Aspergillus oryzae*, and *Trichoderma reesei*. *Appl. Microbiol. Biotechnol.* **102**, 9481–9515 (2018).
5. Yu, J. et al. What can the *Aspergillus flavus* genome offer to mycotoxin research? *Mycology* **2**, 218–236 (2011).
6. Klich, M. A. *Aspergillus flavus*: the major producer of aflatoxin. *Mol. Plant Pathol.* **8**, 713–722 (2007).
7. Gourama, H. *Aspergillus flavus* and *Aspergillus parasiticus*: aflatoxigenic fungi of concern in foods and feeds: a review. *J. Food Prot.* **58**, 1395–1404 (1995).
8. Hedayati, M. T., Pasqualotto, A. C., Warn, P. A., Bowyer, P. & Denning, D. W. *Aspergillus flavus*: human pathogen, allergen and mycotoxin producer. *Microbiology* **153**, 1677–1692 (2007).
9. Krishnan, S., Manavathu, E. K. & Chandrasekar, P. H. *Aspergillus flavus*: an emerging *non-fumigatus Aspergillus* species of significance. *Mycoses* **52**, 206–222 (2009).

10. Varga, J., Frisvad, J. C. & Samson, R. A. Two new aflatoxin producing species, and an overview of *Aspergillus* section Flavi. *Stud. Mycol.* **69**, 57–80 (2011).
11. Frisvad, J. C. et al. Taxonomy of *Aspergillus* section Flavi and their production of aflatoxins, ochratoxins and other mycotoxins. *Stud. Mycol.* **93**, 1–63 (2019).
12. Machida, M. et al. Genome sequencing and analysis of *Aspergillus oryzae. Nature* **438**, 1157–1161 (2005).
13. Nierman, W. C. et al. Genome sequence of *Aspergillus flavus* NRRL 3357, a strain that causes aflatoxin contamination of food and feed. *Genome Announc.* **3**, e00168–15 (2015).
14. Moore, G. G., Mack, B. M., Beltz, S. B. & Gilbert, M. K. Draft genome sequence of an aflatoxigenic *Aspergillus* species, *A. bombycis. Genome Biol. Evol.* **8**, 3297–3300 (2016).
15. Linz, J. E., Wee, J. & Roze, L. V. *Aspergillus parasiticus* SU-1 genome sequence, predicted chromosome structure, and comparative gene expression under aflatoxin-inducing conditions: evidence that differential expression contributes to species phenotype. *Eukaryot. Cell* **13**, 1113–1123 (2014).
16. Vesth, T. C. et al. Investigation of inter- and intraspecies variation through genome sequencing of *Aspergillus* section Nigri. *Nat. Genet.* https://doi.org/10.1038/s41588-018-0246-1 (2018).
17. Kjærbølling, I. et al. Linking secondary metabolites to gene clusters through genome sequencing of six diverse *Aspergillus* species. *Proc. Natl Acad. Sci.* **115**, E753–E761 (2018).
18. Moore, G. G., Mack, B. M., Beltz, S. B. & Puel, O. Genome sequence of an aflatoxigenic pathogen of Argentinian peanut, *Aspergillus arachidicola. BMC Genom.* **19**, 189 (2018).
19. Moore, G. G., Mack, B. M. & Beltz, S. B. Genomic sequence of the aflatoxigenic filamentous fungus *Aspergillus nomius. BMC Genom.* **16**, 551 (2015).
20. Simã, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
21. Galagan, J. E. et al. Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae. Nature* **438**, 1105–1115 (2005).
22. Varga, J. et al. New and revisited species in *Aspergillus* section Nigri. *Stud. Mycol.* **69**, 1–17 (2011).
23. Kocsubé, S. et al. *Aspergillus* is monophyletic: evidence from multiple gene phylogenies and extrolites profiles. *Stud. Mycol.* **85**, 199–213 (2016).
24. Sato, A., Matsushima, K., Ito, K. & Mituyama, T. Comparative genomics of the *Aspergillus* section Flavi. in *29th Fungal Genetics Conference Abstract Book* 175 (2017).
25. Sato, A., Matsushima, K., Ito, K. & Mituyama, T. Genome sequence data analysis portal. (2017). https://genome.cbrc.jp/sojae/.
26. Qi, J., Luo, H. & Hao, B. CVTree: a phylogenetic tree reconstruction tool based on whole genomes. *Nucleic Acids Res.* **32**, 45–47 (2004).
27. Zuo, G. & Hao, B. CVTree3 web server for whole-genome-based and alignment-free prokaryotic phylogeny and taxonomy. *Genomics Proteom. Bioinform.* **13**, 321–331 (2015).
28. Kurtzman, C. P., Smiley, M. J., Robnett, C. J., Wicklow, D. T. & Wickl, D. T. DNA relatedness among wild and domesticated species in the *Aspergillus flavus* group. *Mycologia* **78**, 955–959 (1986).
29. Yuan, G.-F., Liu, C.-S. & Chen, C.-C. Differentiation of *Aspergillus parasiticus* from *Aspergillus sojae* by random amplification of polymorphic DNA. *Appl. Environ. Microbiol.* **61**, 2384–2387 (1995).
30. Gibbons, J. G. et al. The evolutionary imprint of domestication on genome variation and function of the filamentous fungus *Aspergillus oryzae. Curr. Biol.* **22**, 1403–1409 (2012).
31. Chang, P.-K., Ehrlich, K. & Fujii, I. Cyclopiazonic acid biosynthesis of *Aspergillus flavus* and *Aspergillus oryzae. Toxins* **1**, 74–99 (2009).
32. Finn, R. D. et al. InterPro in 2017−−beyond protein family and domain annotations. *Nucleic Acids Res.* **45**, D190–D199 (2017).
33. Gene Ontology Consortium, T. et al. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
34. Gene Ontology Consortium, T. Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Res.* **45**, D331–D338 (2017).
35. Tatusov, R. L. et al. The COG database: an updated version includes eukaryotes. *BMC Bioinform.* **4**, 41 (2003).
36. Eichler, E. E. & Sankoff, D. Structural dynamics of eukaryotic chromosome evolution. *Science* **301**, 793–797 (2003).
37. Nierman, W. C. et al. Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus. Nature* **438**, 1151–1156 (2005).
38. Fedorova, N. D. et al. Genomic islands in the pathogenic filamentous fungus *Aspergillus fumigatus. PLoS Genet.* **4**, e1000046 (2008).
39. Fitzpatrick, D. A. Horizontal gene transfer in fungi. *FEMS Microbiol. Lett.* **329**, 1–8 (2012).
40. Barbesgaard, P., Heldt-Hansen, H. P. & Diderichsen, B. On the safety of *Aspergillus oryzae*: a review. *Appl. Microbiol. Biotechnol.* **36**, 569–572 (1992).
41. Benoit, I. et al. Closely related fungi employ diverse enzymatic strategies to degrade plant biomass. *Biotechnol. Biofuels* **8**, 107 (2015).
42. de Siqueira, F. G. et al. Evaluation of holocellulase production by plant-degrading fungi grown on agro-industrial residues. *Biodegradation* **21**, 815–824 (2010).
43. Kim, J. Production of xylanolytic enzyme complex from *Aspergillus flavus* using agricultural wastes. *Mycobiology* **33**, 84–89 (2005).
44. Mahmoud, M. A., Al-Othman, M. R., Abd-El-Aziz, A. R. M., Metwaly, H. A. & Mohamed, H. A. Expression of genes encoding cellulolytic enzymes in some *Aspergillus* species. *Genet. Mol. Res.* **15**, 15048913 (2016).
45. Mäkelä, M. R. et al. Genomic and exoproteomic diversity in plant biomass degradation approaches among Aspergilli. *Stud. Mycol.* in press (2018).
46. Ichishima, E. Development of enzyme technology for *Aspergillus oryzae, A. sojae*, and *A. luchuensis*, the national fungi of Japan. *Biosci. Biotechnol. Biochem.* **80**, 1681–1692 (2016).
47. Mata-Gómez, M. A. et al. A novel pectin-degrading enzyme complex from *Aspergillus sojae* ATCC 20235 mutants. *J. Sci. Food Agric.* **95**, 1554–1561 (2015).
48. Civas, A., Eberhard, R., Le Dizet, P. & Petek, F. Glycosidases induced in *Aspergillus tamarii*. Secreted alpha-D-galactosidase and beta-D-mannanase. *Biochem. J.* **219**, 849–855 (1984).
49. da Silva, A. C. et al. Production and characterization of xylanase from *Aspergillus parasiticus* URM 5963 isolated from soil of Caatinga. *J. Microbiol. Biotechnol.* **5**, 71–75 (2016).
50. Heinen, P. R. et al. GH11 xylanase from *Aspergillus tamarii* Kita: Purification by one-step chromatography and xylooligosaccharides hydrolysis monitored in real-time by mass spectrometry. *Int. J. Biol. Macromol.* **108**, 291–299 (2018).
51. Makhuvele, R., Ncube, I., van Rensburg, E. L. J. & La Grange, D. C. Isolation of fungi from dung of wild herbivores for application in bioethanol production. *Braz. J. Microbiol.* **48**, 648–655 (2017).
52. Moreira, F. G., Lenartovicz, V. & Peralta, R. M. A thermostable maltose-tolerant alpha-amylase from *Aspergillus tamarii. J. Basic Microbiol.* **44**, 29–35 (2004).
53. Saroj, P., P, M. & Narasimhulu, K. Characterization of thermophilic fungi producing extracellular lignocellulolytic enzymes for lignocellulosic hydrolysis under solid-state fermentation. *Bioresour. Bioprocess.* **5**, 31 (2018).
54. Sen, S., Ray, L. & Chattopadhyay, P. Production, purification, immobilization, and characterization of a thermostable beta-galactosidase from *Aspergillus alliaceus. Appl. Biochem. Biotechnol.* **167**, 1938–1953 (2012).
55. de Vries, R. P. et al. Comparative genomics reveals high biological diversity and specific adaptations in the industrially and medically important fungal genus *Aspergillus. Genome Biol.* **18**, 28 (2017).
56. Martinez, D. et al. Genome sequencing and analysis of the biomass-degrading fungus *Trichoderma reesei* (syn. *Hypocrea jecorina*). *Nat. Biotechnol.* **26**, 553–560 (2008).
57. Druzhinina, I. S. 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.* **14**, e1007322 (2018).
58. de Souza, W. R. et al. Transcriptome analysis of *Aspergillus niger* grown on sugarcane bagasse. *Biotechnol. Biofuels* **4**, 40 (2011).
59. Ferreira, G., Boer, C. G. & Peralta, R. M. Production of xylanolytic enzymes by *Aspergillus tamarii* in solid state fermentation. *FEMS Microbiol. Lett.* **173**, 335–339 (1999).
60. Kimura, I., Sasahara, H. & Tajima, S. Purification and characterization of two xylanases and an arabinofuranosidase from *Aspergillus sojae. J. Ferment. Bioeng.* **80**, 334–339 (1995).
61. Mellon, J. E., Cotty, P. J., Callicott, K. A. & Abbas, H. Identification of a major xylanase from *Aspergillus flavus* as a 14-kD protein. *Mycopathologia* **172**, 299–305 (2011).
62. de Souza, D. F., de Souza, C. G. M. & Peralta, R. M. Effect of easily metabolizable sugars in the production of xylanase by *Aspergillus tamarii* in solid-state fermentation. *Process Biochem.* **36**, 835–838 (2001).
63. Terada, M., Hayashi, K. & Mizunuma, T. Distinction between *Aspergillus oryzae* and *Aspergillus sojae* by the productivity of some hydrolytic enzymes. *Nihon shoyu kenkyusho zasshi* **6**, 75–81 (1980).
64. Khaldi, N. et al. SMURF: genomic mapping of fungal secondary metabolite clusters. *Fungal Genet. Biol.* **47**, 736–741 (2010).
65. Nielsen, J. C. et al. Global analysis of biosynthetic gene clusters reveals vast potential of secondary metabolite production in *Penicillium* species. *Nat. Microbiol.* **2**, 17044 (2017).
66. Medema, M. H. et al. The minimum information about a biosynthetic gene cluster. *Nat. Chem. Biol.* **11**, 625–631 (2015).
67. Theobald, S. et al. Uncovering secondary metabolite evolution and biosynthesis using gene cluster networks and genetic dereplication. *Sci. Rep.* https://doi.org/10.1038/s41598-018-36561-3 (2018).
68. Mayorga, M. E. & Timberlake, W. E. The developmentally regulated *Aspergillus nidulans wA* gene encodes a polypeptide homologous to polyketide and fatty acid synthases. *Mol. Gen. Genet.* **235**, 205–212 (1992).

69. Andersen, M. R. et al. Accurate prediction of secondary metabolite gene clusters in filamentous fungi. *Proc. Natl Acad. Sci. USA* **110**, E99–E107 (2013).
70. Zabala, A. O., Xu, W., Chooi, Y. H. & Tang, Y. Characterization of a silent azaphilone gene cluster from *Aspergillus niger* ATCC 1015 reveals a hydroxylation-mediated pyran-ring formation. *Chem. Biol.* **19**, 1049–1059 (2012).
71. Cary, J. W. et al. Transcriptome analysis of *Aspergillus flavus* reveals veA-dependent regulation of secondary metabolite gene clusters, including the novel aflavarin cluster. *Eukaryot. Cell* **14**, 983–997 (2015).
72. Lim, F. Y. et al. Genome-based cluster deletion reveals an endocrocin biosynthetic pathway in *Aspergillus fumigatus*. *Appl. Environ. Microbiol.* **78**, 4117–4125 (2012).
73. Berthier, E. et al. Low-volume toolbox for the discovery of immunosuppressive fungal secondary metabolites. *PLoS Pathog.* **9**, e1003289 (2013).
74. Chiang, Y.-M. M. et al. A gene cluster containing two fungal polyketide synthases encodes the biosynthetic pathway for a polyketide, asperfuranone, in *Aspergillus nidulans*. *J. Am. Chem. Soc.* **131**, 2965–2970 (2009).
75. Maiya, S., Grundmann, A., Li, X., Li, S. M. & Turner, G. Identification of a hybrid PKS/NRPS required for pseurotin A biosynthesis in the human pathogen *Aspergillus fumigatus*. *ChemBioChem* **8**, 1736–1743 (2007).
76. Lin, H. C. et al. The fumagillin biosynthetic gene cluster in *Aspergillus fumigatus* encodes a cryptic terpene cyclase involved in the formation of beta-trans-bergamotene. *J. Am. Chem. Soc.* **135**, 4616–4619 (2013).
77. Shiomi, K. et al. New antibiotics miyakamides produced by a fungus. *J. Antibiot.* **55**, 952–961 (2002).
78. Sambrook, J., Fritsch, E. F. & Maniatis, T. *Molecular Cloning: A Laboratory Manual.* (Cold Spring Harbor Laboratory Press, 2012).
79. Fulton, T. M., Chunwongse, J. & Tanksley, S. D. Microprep protocol for extraction of DNA from tomato and other herbaceous plants. *Plant Mol. Biol. Rep.* **13**, 207–209 (1995).
80. Kis-Papo, T. et al. Genomic adaptations of the halophilic dead sea filamentous fungus *Eurotium rubrum*. *Nat. Commun.* **5**, 3745 (2014).
81. Grigoriev, I. V. et al. MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Res.* **42**, 699–704 (2014).
82. Grigoriev, I. V., Martinez, D. A. & Salamov, A. A. Fungal genomic annotation. *Appl. Mycol. Biotechnol.* **6**, 123–142 (2006).
83. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* **42**, 490–495 (2014).
84. Cantarel, B. L. et al. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res.* **37**, 233–238 (2009).
85. Gel, B. & Serra, E. KaryoploteR: An R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **33**, 3088–3090 (2017).
86. Sullivan, M. J., Petty, N. K. & Beatson, S. A. Easyfig: a genome comparison visualizer. *Bioinformatics* **27**, 1009–1010 (2011).
87. De Vries, R. P. et al. A new black *Aspergillus* species, *A. vadensis*, is a promising host for homologous and heterologous protein production. *Appl. Environ. Microbiol.* **70**, 3954–3959 (2004).
88. Arnaud, M. B. et al. The Aspergillus Genome Database (AspGD): recent developments in comprehensive multispecies curation, comparative genomics and community resources. *Nucleic Acids Res.* **40**, 653–659 (2012).
89. Andersen, M. R. et al. Comparative genomics of citric-acid-producing *Aspergillus niger* ATCC 1015 versus enzyme-producing CBS 513.88. *Genome Res.* **21**, 885–897 (2011).
90. Marcet-Houben, M. et al. Genome sequence of the necrotrophic fungus *Penicillium digitatum*, the main postharvest pathogen of citrus. *BMC Genom.* **13**, 646 (2012).
91. Galagan, J. E. et al. The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature* **422**, 859–868 (2003).
92. Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290 (2004).
93. Lebar, M. D. et al. Identification and functional analysis of the aspergillic acid gene cluster in *Aspergillus flavus*. *Fungal Genet. Biol.* **116**, 14–23 (2018).
94. Nicholson, M. J. et al. Identification of two aflatrem biosynthesis gene loci in *Aspergillus flavus* and metabolic engineering of *Penicillium paxilli* to elucidate their function. *Appl. Environ. Microbiol.* **75**, 7469–7481 (2009).
95. Wollenberg, R. D. et al. Chrysogine biosynthesis is mediated by a two-module nonribosomal peptide synthetase. *J. Nat. Prod.* **80**, 2131–2135 (2017).
96. Yun, C. S., Motoyama, T. & Osada, H. Biosynthesis of the mycotoxin tenuazonic acid by a fungal NRPS-PKS hybrid enzyme. *Nat. Commun.* **6**, 8758 (2015).
97. Watanabe, K. Effective use of heterologous hosts for characterization of biosynthetic enzymes allows production of natural products and promotes new natural product discovery. *Chem. Pharm. Bull.* **62**, 1153–1165 (2014).
98. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
99. Edgar, R. C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
100. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540–552 (2000).

Acknowledgements

Genome sequencing was kindly supported by Joint BioEnergy Institute and Joint Genome Institute. M.R.A., J.L.N., S.T. and T.C.V. gratefully acknowledge funding from the Villum Foundation, Grant VKR023437. M.R.A. and T.C.V. further acknowledge funding from the Danish National Research Foundation, grant number DNRF137. M.R.A. further acknowledges support from the Jorck Foundation. The work conducted by the US Department of Energy Joint Genome Institute, a US Department of Energy Office of Science User Facility, is supported by the Office of Science of the US Department of Energy under Contract No. DE-AC02-05CH11231. The US Department of Energy Joint BioEnergy Institute (www.jbei.org) is supported by the US Department of Energy, Office of Science, and Office of Biological and Environmental Research, through Contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the US Department of Energy.

Author contributions

I.K. analyzed data, contrived data analysis methods, contributed to design of research, and wrote most parts of the paper. T.V. conceived the overall project, analyzed data, contributed to design of research, wrote parts of the paper, and coordinated the project. J.C.F. contributed to design of research, contributed analytical tools and data for species selection and verification, wrote parts of the paper, and analyzed data. J.L.N. and S.T. analyzed data and contributed to design of research. S.K. and T.I. generated data on secondary metabolism and analyzed chemical data. E.K.L. and M.E.K. contributed to design of research, developed methods, conducted experiments, and analyzed data. A. Sat analyzed data and contributed to design of research. A.W., R.S.K. and R.J.M.L. performed part of the experiments. M.R.M. analyzed data and wrote parts of the paper. A.K., A. Sal, S.H., R.R. and S.M. annotated genomes and analyzed data. A.C., A.L., K.L., and J.P. assembled the genomes. C.D., G.H. and M.C. sequenced RNA and DNA. J.K.M. and B.A.S. contributed to design of research. B.H. and E.D. contributed analytical tools and analyzed CAZyme data. U.H.M. contributed to design of research and developed methods. T.O.L. generated data on secondary metabolism, analyzed data, and wrote parts of the paper. R.P.dV. analyzed data and wrote parts of the paper. K.B. and I.G. coordinated the DNA and RNA sequencing and annotation. M.M. analyzed data and wrote parts of the paper. S.E.B. conceived the overall project, analyzed data, contributed to design of research, and contributed to writing and editing the paper. M.R.A. conceived the overall project, analyzed data, contributed to design of research, wrote parts of the paper, and coordinated the project. All authors commented on the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-019-14051-y.

Correspondence and requests for materials should be addressed to M.R.A.

Peer review information *Nature Communications* thanks Yi-Ming Chiang, William Nierman and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

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