WESTERDIJK FUNGALBIO DIVERSITY INSTITUTE

Section-level genome sequencing and comparative genomics of *Aspergillus* sections *Cavernicolus* and *Usti*

J.L. Nybo[1#], T.C. Vesth[1#], S. Theobald[1$], J.C. Frisvad[1], T.O. Larsen[1], I. Kjaerboelling[1#], K. Rothschild-Mancinelli[1#], E.K. Lyhne[1], K. Barry[2], A. Clum[2], Y. Yoshinaga[2], L. Ledsgaard[1], C. Daum[2], A. Lipzen[2], A. Kuo[2], R. Riley[2], S. Mondo[2], K. LaButti[2], S. Haridas[2], J. Pangalinan[2], A.A. Salamov[2], B.A. Simmons[3,4], J.K. Magnuson[3], J. Chen[5,6], E. Drula[7,8], B. Henrissat[1], A. Wiebenga[9], R.J.M. Lubbers[9], A. Müller[9], A.C. dos Santos Gomes[9], M.R. Mäkelä[10], J.E. Stajich[5,6], I.V. Grigoriev[2,11], U.H. Mortensen[1], R.P. de Vries[9*], S.E. Baker[3,12*], M.R. Andersen[1*#]

[1]*Department of Biotechnology and Biomedicine, Technical University of Denmark, Kgs. Lyngby, Denmark;* [2]*US Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA;* [3]*US Department of Energy Joint Bioenergy Institute, Berkeley, CA, USA;* [4]*Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA;* [5]*Department of Microbiology and Plant Pathology, University of California, Riverside, CA, USA;* [6]*Institute for Integrative Genome Biology, University of California, Riverside, CA, USA;* [7]*AFMB, UMR 7257 CNRS Aix-Marseille Univ., USC 1408 INRAE, Marseille, France;* [8]*Biodiversité et Biotechnologie Fongiques, UMR 1163, INRAE, Marseille, France;* [9]*Fungal Physiology, Westerdijk Fungal Biodiversity Institute, Utrecht, The Netherlands;* [10]*Department of Bioproducts and Biosystems, Aalto University, Aalto, Finland;* [11]*Department of Plant and Microbial Biology, University of California Berkeley, Berkeley, CA, USA;* [12]*Microbial Molecular Phenotyping Group, Environmental Molecular Sciences Division, Pacific Northwest National Laboratory, Richland, WA, USA*

[#]*Current address*: Novonesis A/S, Bagsværd, Denmark; [$]*Current address*: LifeMine Therapeutics, Cambridge MA, USA

[*]*Corresponding authors*: Mikael R. Andersen, mrra@novonesis.com; Ronald P. de Vries, r.devries@wi.knaw.nl; Scott E. Baker, scott.baker@pnnl.gov

Abstract: The genus *Aspergillus* is diverse, including species of industrial importance, human pathogens, plant pests, and model organisms. *Aspergillus* includes species from sections *Usti* and *Cavernicolus*, which until recently were joined in section *Usti*, but have now been proposed to be non-monophyletic and were split by section *Nidulantes*, *Aenei* and *Raperi*. To learn more about these sections, we have sequenced the genomes of 13 *Aspergillus* species from section *Cavernicolus* (*A. cavernicola, A. californicus*, and *A. egyptiacus*), section *Usti* (*A. carlsbadensis, A. germanicus, A. granulosus, A. heterothallicus, A. insuetus, A. keveii, A. lucknowensis, A. pseudodeflectus* and *A. pseudoustus*), and section *Nidulantes* (*A. quadrilineatus*, previously *A. tetrazonus*). We compared these genomes with 16 additional species from *Aspergillus* to explore their genetic diversity, based on their genome content, repeat-induced point mutations (RIPs), transposable elements, carbohydrate-active enzyme (CAZyme) profile, growth on plant polysaccharides, and secondary metabolite gene clusters (SMGCs). All analyses support the split of section *Usti* and provide additional insights: Analyses of genes found only in single species show that these constitute genes which appear to be involved in adaptation to new carbon sources, regulation to fit new niches, and bioactive compounds for competitive advantages, suggesting that these support species differentiation in *Aspergillus* species. Sections *Usti* and *Cavernicolus* have mainly unique SMGCs. Section *Usti* contains very large and information-rich genomes, an expansion partially driven by CAZymes, as section *Usti* contains the most CAZyme-rich species seen in genus *Aspergillus*. Section *Usti* is clearly an underutilized source of plant biomass degraders and shows great potential as industrial enzyme producers.

Key words: *Aspergillus*; CAZymes; comparative genomics; secondary metabolites; section *Cavernicolus*; section *Usti*

Citation: Nybo JL, Vesth TC, Theobald S, Frisvad JC, Larsen TO, Kjaerboelling I, Rothschild-Mancinelli K, Lyhne EK, Barry K, Clum A, Yoshinaga Y, Ledsgaard L, Daum C, Lipzen A, Kuo A, Riley R, Mondo S, LaButti K, Haridas S, Pangalinan J, Salamov AA, Simmons BA, Magnuson JK, Chen J, Drula E, Henrissat B, Wiebenga A, Lubbers RJM, Müller A, dos Santos Gomes AC, Mäkelä MR, Stajich JE, Grigoriev IV, Mortensen UH, de Vries RP, Baker SE, Andersen MR (2025). Section-level genome sequencing and comparative genomics of *Aspergillus* sections *Cavernicolus* and *Usti*. *Studies in Mycology* **111**: 101–114. doi: 10.3114/sim.2025.111.03

Received: 30 August 2024; **Accepted:** 4 February 2025; **Effectively published online:** 5 March 2025

Corresponding editor: J. Houbraken

INTRODUCTION

The filamentous ascomycete genus *Aspergillus* comprises more than 450 species (Visagie *et al.* 2024), several of which are medically, agriculturally, and biotechnologically important (Max *et al.* 2010, Meyer *et al.* 2011, Knuf & Nielsen 2012, Kocsubé *et al.* 2016, de Vries *et al.* 2017, Wakai *et al.* 2017). The species of this genus are highly adaptive with respect to biotope, resulting in *Aspergillus* being an extremely diverse fungal genus, producing a wide variety of morphological traits, enzymes, and secondary metabolites (Punt *et al.* 2002, Schuster *et al.* 2002, Ward *et al.* 2004, Jin *et al.* 2007, Dashtban *et al.* 2010, Meyer *et al.* 2011, Knuf & Nielsen 2012, Inglis *et al.* 2013, Richter *et al.* 2014, Frisvad & Larsen 2015, Yamada *et al.* 2015, Vries *et al.* 2017, Vesth *et al.* 2018, Sharma *et al.* 2019). Although industrially relevant species are well-studied, the total diversity within the genus remains largely unexplored. Studying this diversity will uncover novel industrially relevant enzymes and compounds, as well as the genetic basis underlying this diversity.

To study this diversity, the *Aspergillus* whole-genus sequencing project was initiated in 2013. Prior to this, only 10 *Aspergillus* species had been sequenced, whereas today, genomes of 253 species have been released (Grigoriev *et al.* 2012, Nordberg *et al.* 2014, Vesth *et al.* 2018, Kjærbølling *et al.* 2018), although some of these species have recently been merged (Bian *et al.* 2022). Most of the genomes are associated with the *Aspergillus* whole-genus sequencing project. The first large comparative genomics study published as part of the *Aspergillus* whole-genus sequencing project in 2018 focused on section *Nigri* (Vesth *et al.* 2018). Currently, *Aspergillus* is composed of 28 sections (Peterson *et*

Studies in Mycology

al. 2008, Varga *et al.* 2010, Samson *et al.* 2014, Hubka *et al.* 2015, Jurjevic *et al.* 2015, Kocsubé *et al.* 2016), one of which is section *Usti*.

The group of species defining section *Usti* was first described in 1965 by Raper and Fennell and was recognized as an individual section by Gams *et al.* (1985). In 2016, Hubka *et al.* published a phylogenetic tree in which *Usti* was split into two clades divided by sections *Nidulantes*, *Aenei* and *Raperi* (Hubka *et al.* 2016). Subsequently, Chen *et al.* proposed that the smaller monophyletic clade containing *A. californicus, A. egyptiacus, A. kassunensis, A. subsessilis*, and *A. cavernicola* represents a new section named *Cavernicolus* (Chen *et al.* 2016). The species comprising *Usti* and *Cavernicolus* inhabit food, soil, and indoor air environments, and have been isolated from a diversity of geographical regions from the Arctic permafrost to the African desert (Houbraken *et al.* 2007, Samson *et al.* 2011, Chen *et al.* 2016, Kozlovskii *et al.* 2016). This diverse adaptive ability can also be seen in the wide variety of secondary metabolites they produce (including toxins, pigments, acids, and antibiotics), many of which have only been found in species belonging to this section (Samson *et al.* 2011, Kozlovskii *et al.* 2016). This section also contains four opportunistic pathogens (*A. ustus, A. calidoustus, A. granulosus*, and *A. deflectus*) (Samson *et al.* 2011).

In this study, to investigate the species in more detail, and to confirm or disprove the division of section *Usti*, we have *de novo* sequenced the genomes of nine species from section *Usti*, three species from section *Cavernicolus*, and one from section *Nidulantes*, thus allowing for inter- and intra-section comparison of 13 species. For diversity reference, 16 additional *Aspergillus* genomes were included in this study, of which seven genomes have been sequenced using long read technology and five are established reference genomes. These were also selected to support analyses on transposable elements as a source of genome diversity and expansion. Moreover, we examined the intra-genus diversity to determine the pan and core genome of the genus *Aspergillus* as well as the proteins unique to each species in this study. In extension of this, we have examined the enzyme and secondary metabolite content in particular detail. These are essential traits of differentiation, as well as a being a highly attractive source of next generation bioproducts.

MATERIALS AND METHODS

Fungal strains

Unless stated otherwise, the species examined were taken from the IBT Culture Collection of Fungi at the Technical University of Denmark (DTU). Strains included in this study are listed in Table S1.

DNA and RNA preparation, sequencing and genome assembly

For all sequences generated for this study, spores were defrosted from storage at -80 °C and inoculated onto solid Czapek Yeast Agar (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 wk. Biomass for all fungal strains was obtained from shake flasks containing 200 mL of complex medium, either CYA, Malt Extract Agar (MEA), or CYA + 20 % sucrose (CY20), depending on which gave more mycelium, 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. The DNA isolation was performed using a modified version of the standard phenol extraction (Sambrook & Russell 2012) and checked for quality and concentration using a NanoDrop (BioNordika). The 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.* (1995) was used, with liquid nitrogen and mortar and pestle for breaking the mycelium, followed by lysis with proteinase K, and DNA extraction with phenol:chloroform:isoamylalcohol (25:24:1) [See the supplementary material in Vesth *et al.* (2018) for a full lab protocol].

One hundred nanogram of DNA was sheared to 300 bp using the Covaris LE220 and size selected using SPRI beads (Beckman Coulter). The fragments were treated with end-repair, A-tailing, and ligation of Illumina compatible adapters (IDT, Inc) using the KAPA-Illumina library creation kit (KAPA biosystems). Stranded cDNA libraries were generated using the Illumina Truseq Stranded RNA LT kit. Messenger RNA was purified from 1 ug of total RNA using magnetic beads containing poly-T oligos; it was fragmented and reversed transcribed using random hexamers and SSII (Invitrogen) followed by second strand synthesis. The fragmented cDNA was treated with end-pair, A-tailing, adapter ligation, and 8 cycles of PCR. The prepared libraries were quantified using KAPA Biosystem's next-generation sequencing library qPCR kit (Roche) and run on a Roche LightCycler 480 real-time PCR instrument. The quantified libraries were then multiplexed and the pool of libraries was prepared for sequencing on the Illumina HiSeq sequencing platform utilizing a TruSeq paired-end cluster kit, v. 4, and Illumina's cBot instrument to generate a clustered flow cell for sequencing. Sequencing of the flow cell was performed on the Illumina HiSeq 2500 sequencer using HiSeq TruSeq SBS sequencing kits, v. 4, following a 2× 150 indexed run recipe. Genomic DNA Illumina reads were QC filtered for artifact/process contamination and subsequently assembled together with Velvet v. 1.2.07 (Zerbino & Birney 2008). The resulting assembly was used to create a 25× of 2× 100 bp 3000 +/- 300 bp insert long mate-pair library with wgsim v. 0.3.1-r13 (https://github.com/lh3/wgsim) which was then assembled together with 125× of the original Illumina library with AllPathsLG release v. R49403 (Gnerre *et al*. 2010). Using BBDuk (https://sourceforge.net/projects/bbmap/), raw RNA-Seq reads were evaluated for artifact sequence by kmer matching (kmer = 25), allowing 1 mismatch and detected artifact was trimmed from the 3' end of the reads. RNA spike-in reads, PhiX reads and reads containing any Ns were removed. Quality trimming was performed using the phred trimming method set at Q6. Finally, following trimming, reads under the length threshold were removed (minimum length 25 bases or 1/3 of the original read length - whichever is longer). Filtered reads were assembled into consensus sequences using Trinity v. 2.1.1 (Grabherr *et al.* 2011).

Genome annotation

All genomes were annotated based on the JGI annotation pipeline (Grigoriev *et al.* 2014) as previously described (Kis-Papo *et al.* 2014).

Whole-genome phylogeny

Protein sequences of all organisms were compared using BLASTp (e-value cutoff $1e^{-05}$). Orthologous groups of sequences were constructed based on best bidirectional hits. Two-hundred groups with a single member from each species were selected and the sequences of each organism were concatenated into one long protein sequence. Concatenated sequences were aligned using MAFFT (thread 16) and well-aligned regions are extracted using Gblocks (-t = p -b4=5 -b5 = h). Trees were then constructed using multi- threaded RAxML, the PROTGAMMAWAG model and 100 bootstrap replicates.

Prediction of homologous protein families

All predicted sets of protein sequences for the 32 genomes analysed in this paper were aligned using the BLASTP function from the BLAST+ suite v. 2.2.27 with an e-value cut-off of $1e^{-10}$ (Altschul *et al.* 1990, Camacho *et al.* 2009). These 1 024 whole-genome BLAST tables were analysed to identify bidirectional hits in all pairwise comparisons. Using custom Python scripts, homologs were identified within and across the genomes and grouped into sequence-similar families using single linkage, if they met the following criterion: The sum of the alignment coverage between the pairwise sequences was >130 %, the alignment identity between the pairwise sequences was >50 %, and the hit must be found in both of the species' BLAST output (reciprocal hits). Singletons were assigned a family having only one gene member. This allowed for identification of species-unique genes as well as genes shared by sections, clades, and sub-clades of species (Vesth *et al.* 2018). All homologs were assigned functional and structural domains using InterPro v. 5 (Hunter *et al.* 2009) and checked for annotation and sequencing errors by investigating scaffold location and sequence identity. Core families were defined as families containing at least one protein from all species and species-unique families contained protein(s) from only one species.

Prediction of encoded CAZymes

CAZymes were predicted for all genomes using the Carbohydrate-Active Enzymes database [www.cazy.org (Drula *et al.* 2022)] and the method described in (Vesth *et al.* 2018). For this, each *Aspergillus* protein model was compared using BLASTp to proteins listed in the CAZy database. 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 manually inspected and conserved features such as the catalytic residues were searched whenever known. Sequence alignments with isolated functional domains were performed in the case of multimodular CAZymes.

Prediction of secondary metabolite gene clusters

For the prediction of secondary metabolite gene clusters (SMGCs), we developed a command-line Python script roughly following the SMURF algorithm (Khaldi *et al.* 2010) as also described in Vesth *et al.* (2018): We extract "backbone" genes and PFAM domains from the SMURF paper (Khaldi *et al.* 2010). 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.

For identification of SMGC families, we compared proteins of the resulting SMGCs with each other by alignment using BLASTp (BLAST+ suite v. 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 [See Vesth *et al.* (2018) for details on calculation of weighting scores]. Using these scores, we created a weighted network of SMGC clusters and used a random walk community detection algorithm (R v. 3.3.2, igraph_1.0.1 (Csárdi & Nepusz 2006) 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, adding up scores of the similarities to create an overall cluster similarity score.

Annotating SMGC families using MIBiG

Linking MIBiG to predicted clusters was performed as described in Vesth *et al.* (2018) and Theobald *et al.* (2018). Gene cluster annotations were downloaded from the MIBiG database (Medema *et al.* 2015) and 1461 sequences of backbone proteins extracted using biopython (Cock *et al.* 2009). Protein sequences were then compared against our dataset using BLAST. Hits reaching a percent identity, query coverage and hit coverage of over 95 % were retained to find best hits in our dataset. Corresponding SMGC families were annotated as related cluster of the hit.

Identification of shared SMGC families at nodes of the phylogenetic tree

The SMGC families were mapped on nodes on the phylogenetic tree if all species branching from that node contained the families.

Identification, classification and annotation of transposable elements (TEs)

The *Aspergillus*-specific transposable element library was built from seven PacBio genomes (see Table S1) and five well-annotated genomes (*A. nidulans*, *A. fumigatus* Af293, *A. oryzae, A. niger* ATTC 1015 and *A. terreus*) were selected. The *de novo* identification of repetitive elements in the unmasked genome assemblies was performed using the REPET TEdenovo pipeline (Quesneville *et al.* 2005, Flutre *et al.* 2011), which creates species-specific non-redundant consensus TE libraries. The libraries were then checked for potential host genes and were then combined into one. The combined library was then classified using RepeatClassifier from RepeatModeler (www.repeatmasker.org). RepeatClassifier will label sequences without similarity to any Repbase (Bao *et al.* 2015) entry as "unknown". The now classified library was combined with the RepeatMasker library of fungal repeats into one custom repeat library of consensus sequences. This custom library was used as library when running RepeatMasker (www.repeatmasker.org) on all genomes using the flags (-s, -no_is & -norna).

Data availability

All genomes can be accessed through JGI's Mycocosm portal (https://mycocosm.jgi.doe.gov/mycocosm/home). This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accessions JBFXLS000000000 (*Aspergillus amylovorus* CBS 600.67), JBDRAG000000000 (*Aspergillus quadrilineatus* CBS 591.65A), JBFXLU000000000 (*Aspergillus pseudoustus* CBS 123904), JBFXLR000000000 (*Aspergillus pseudodeflectus* CBS 756.74), JBFXLT000000000 (*Aspergillus granulosus* CBS 588.65), JBFTWV000000000 (*Aspergillus keveii* CBS 209.92), JBFXLQ000000000 (*Aspergillus lucknowensis* CBS 449.75), SWKD00000000 (*Aspergillus egyptiacus* CBS 656.73), SWKC00000000 (*Aspergillus carlsbadensis* CBS 123894), JBDRAI000000000 (*Aspergillus californicus* CBS 123895), JBDRAJ000000000 (*Aspergillus germanicus* CBS 123887), JBDRAH000000000 (*Aspergillus heterothallicus* CBS 489.65), JBJACX000000000 (*Aspergillus insuetus* CBS 107.25). The version described in this paper is v. JBFXLS010000000, JBDRAG010000000, JBFXLU010000000, JBFXLR010000000, JBFXLT010000000, JBFTWV010000000, JBFXLQ010000000, SWKD01000000, SWKC01000000, JBDRAI010000000, JBDRAJ010000000, JBDRAH010000000.

RESULTS AND DISCUSSION

Genome sequencing, quality control, and comparative analysis shows very large and gene-rich genomes in section *Usti*

We sequenced, assembled, and annotated 13 genomes using the JGI fungal sequencing pipeline (Grigoriev 2014): three from section *Cavernicolus* (*A. cavernicola* CBS 600.67, *A. californicus* CBS 123895, and *A. egyptiacus* CBS 656.73), nine from section *Usti* (*A. carlsbadensis* CBS 123894, *A. germanicus* CBS 123887, *A. granulosus* CBS 588.65, *A. heterothallicus* CBS 489.65, *A. insuetus* CBS 107.25, *A. keveii* CBS 209.92, *A. lucknowensis* CBS 449.75, *A. pseudodeflectus* CBS 756.74 and *A. pseudoustus* CBS 123904), and one from section *Nidulantes* (*A. quadrilineatus* CBS 591.65; previously *A. tetrazonus*). These genomes were compared to one previously published section *Usti* genome (*A. calidoustus*; Horn *et al.* 2016), a reference set of 15 published *Aspergillus* genomes from a diverse set of sections, and three non-*Aspergillus* genomes as a taxonomic outgroup (*Neurospora crassa, Penicillium chrysogenum*, and *Saccharomyces cerevisiae*) (Fig. 1A, Table S1).

The genomes from section *Cavernicolus* (28.3–35.7 Mbp) are generally similar in size to the reference species (av. 33.3 Mbp), whereas the genomes from *Usti* are larger (30.8–41.4 Mbp) with an average size of 38.6 Mbp (Fig. 1A, B). This makes the genomes of section *Usti* 20 % larger than both *Cavernicolus* and the 16 published Aspergilli. *Aspergillus keveii* has the largest (41.4 Mb) *Aspergillus* genome published apart from *A. ellipticus* which has 42.8 Mbp (Vesth *et al*. 2018).

Comparing the genome size to the predicted gene content shows that the large genomes of *Usti* are likely due to additional gene content rather than non-coding DNA. The gene content of *Cavernicolus* is 9890–13881 genes (av. 12209) whereas the gene content of *Usti* is 11541–15687 genes (av. 14324). The number of predicted genes correlates with the genome size being 1.2 times larger than *Cavernicolus* and 1.3 times larger than the 16 published Aspergilli (Fig. 1E). Further, each genome was annotated using both the Gene Ontology and InterPro databases (Ashburner *et al.* 2000, Finn *et al.* 2017, Jones *et al.* 2014, Gene Ontology Consortium 2017) to assign biological functions to the predicted proteins. The coverage of this functional annotation was homogeneous within the whole genus, with an average 55 % and 71 % of the proteins having at least one Gene Ontology or InterPro domain, respectively (Fig. 1F, G). This again supports that the larger genomes in *Usti* contain as much functional genetic information per Mb as the smaller genomes.

Patterns of transposable elements and RIPs support large information-rich genomes in section *Usti*

To assess the source of the expansion of the section *Usti* genomes, we first analysed the %GC content in the genomes as a symptom of repeat-induced point (RIP) mutations. This is a mechanism that mutates cytosine to thymine to silence transgenic repetitive DNA elements during meiosis (Galagan 2003). However, we saw no strong evidence of this; for most of the Aspergilli the %GC content is 50 % ± 2 %. *N. crassa* has a relatively low %GC content of 48 %, which is a consequence of RIPs. Interestingly, *A. ochraceoroseus*, *A. oryzae*, and *A. heteromorphus* all have a lower %GC content than *N. crassa*. Studies have shown that RIP occurs in *A. oryzae*, which correlates with the low transposable element content in the species (Fig. 1D, H) (Montiel *et al.* 2006). The exceptionally high content of transposable elements in both *A. ochraceoroseus* and *A. heteromorphus* suggests an alternative unknown reason for the low %GC content in these species.

Transgenic repeat elements do not exist in abundance among Aspergilli when compared to other ascomycetes (Montiel *et al.* 2006, Amselem *et al.* 2015), despite the genus showing large diversity in both classes and families. We thus used the 15 *Aspergillus* reference genomes, as mentioned above, to create an *Aspergillus*-specific repeat element library for transposable element analysis, and with that, analysed the genomes.

Genomes from both *Usti* and *Cavernicolus* have low repeat element content (Fig. 1H). This could be an artefact caused by the sole use of Illumina technology for sequencing (Mitra *et al.* 2015). Assemblies relying on short reads tend to underestimate the number of repeated sequences. We note that genomes that show high transposable element content have all been sequenced using long-read technologies (such as PacBio), where the longer reads allow detection of more genetic repeats (Table S1), suggesting that this is a prerequisite for getting an accurate picture of repeats. It does not detract from the finding of section *Usti* being very gene rich, as finding and adding more repeated elements would only increase the size of the genome. Thus, we trust that the sequences that we have produced are indeed very gene rich.

Genome analysis shows large variation in *Aspergillus* genomes

Overall, the genome diversity in genus *Aspergillus* is substantial (Fig. 1). The diversity is seen not only in the genome sizes, but also in the number of predicted proteins and transposable elements. The correlation between the genome sizes and the number of predicted proteins, especially for *Usti*, suggests that the large genomes are caused by additional protein encoding genes rather than non-coding DNA. Additionally, the large transposable element content found in *A. heteromorphus* and *A. ochraceoroseus* should be investigated further.

Whole genome phylogeny of 29 species validates the proposed split into sections *Cavernicolus* and *Usti*

In order to improve the section definition using whole-genome phylogeny, we used the whole-genome sequences of the 13 species in sections *Cavernicolus* and *Usti* as well as the 16 additional Aspergilli. A maximum likelihood (ML) tree was constructed from 200 conserved mono-core proteins, generating a genetically high-resolution phylogeny (Fig. 1A). This tree shows 100 % bootstrap support for all nodes except two (Fig. S1) and confirms previous phylogenetic proposals (Chen *et al.* 2016, Hubka *et al.* 2016, Houbraken *et al.* 2020, Visagie *et al.* 2024). One deviation is the relation between sections *Cavernicolus* and *Usti* to section *Nidulantes*. This tree strongly suggests that *Cavernicolus* is a sister clade that is closer to *Nidulantes*, whereas Chen proposed *Usti* to be the closest sister clade to *Nidulantes* of the two (Chen *et al.* 2016). This previous study was based on data from only four loci contrary to 200 in our study, the resolution is higher in this study, even though the previous study included more isolates. However, the increased sequence depth of our study suggests that section *Cavernicolus* is closer to section *Nidulantes* than *Usti*.

The core genome of genus *Aspergillus* is relatively stable between section-level comparisons

In order to examine the genetic variation across the three sections relative to genus *Aspergillus* in general, we first sorted the proteomes into families of homologous proteins using a previously published pipeline (Vesth *et al.* 2018). The families were further categorized into "pan", "core", and "accessory" proteomes (Soucy *et al.* 2015) for all 32 Aspergilli, including our 13 new genomes (Fig. 2, Fig. S2–S4).

The protein families that represent the core *in silico* proteome of genus *Aspergillus* cover 3 684 families (Fig. 2A), which correspond

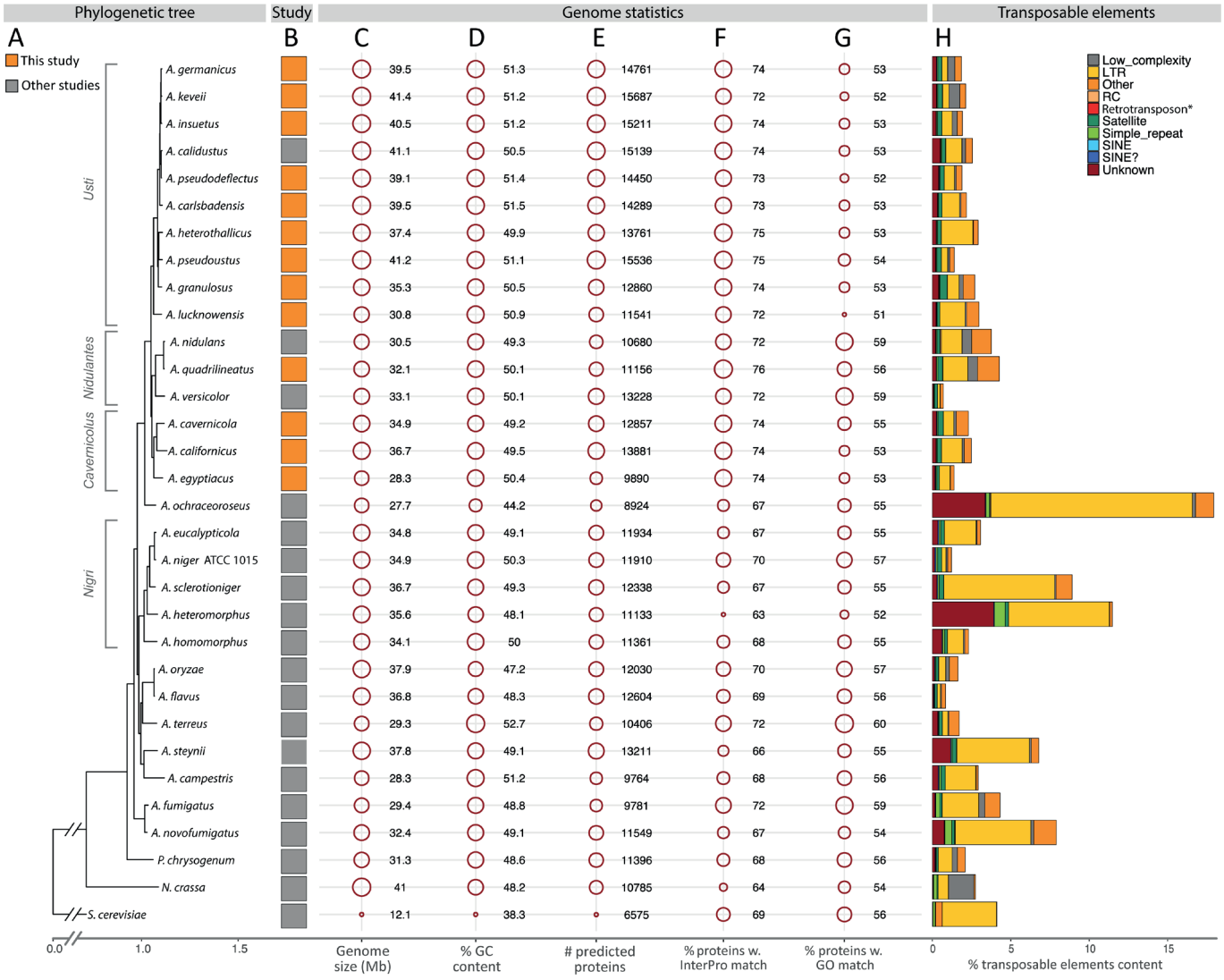

Fig. 1. Dendrogram, bar and bubble plots illustrating the phylogenetic distances and genomes statistics of 29 *Aspergillus* species, a *Penicillium*, a *Neurospora* and a *Saccharomyces* species (as taxonomic outgroup). **A.** Phylogenetic tree created using RAxML (Stamatakis 2014), MAFFT (Katoh and Standley 2013), and Gblocks (Talavera & Castresana 2007), based on 200 conserved proteins. **B.** Colours indicate whether the species is from this study (orange), or another sequencing project (grey). **C–G.** Bubble plots of descriptive numbers for each genome: **C.** Genome size (Mb). **D.** Percentage GC content. **E.** Number of predicted proteins. **F.** Number of proteins with at least one InterPro match. **G.** Number of proteins with at least one GO match (data was not available for *A. sclerotioniger*). The bubble sizes have been scaled to the panel and are not comparable across panels. **H.** Barplot illustrating the percentage transposable element content divided into families. (low_complexity: genomic regions of low complexity; LTR: long terminal repeat retrotransposons; RC: rolling circle repeats; Retrotransposons*: partial retrotransposons; Satellite: satellite DNA; Simple_repeat: regions with simple repeats; SINE: short interspersed nuclear elements; SINE?: partial SINEs; Unknown: repeat elements found by REPET (Quesneville *et al.* 2005, Flutre *et al.* 2011) but could not be classified by RepeatMasker (www.repeatmasker.org).

to 4206–5070 different core proteins per species (including paralogs, Fig. 2A). Of the core proteome, 89 % is covered by at least one functional annotation, most of which are found to be essential functions (Fig. S3, Table S2) (McInerney *et al.* 2017). It is both interesting and surprising to see that the core *Aspergillus* proteome is less than 1/3 of the genome in sections *Usti* and *Cavernicolus*. In a previous study of *Aspergillus* section *Nigri* including reference species (Vesth *et al.* 2018), we found the core genome to be larger (≈ 40 %). However, in that set, we did not have as many reference species with genomes of less than 10000 genes, as we do here. Comparing the two core proteomes, we find 88 % similarity (Table S2.1).

Genes unique to individual species contain functions required for diversification

Examining the species-unique proteins, the numbers range from 841–3019 proteins in the individual species (Fig. 2B), and contain functions involved in transport, regulation, and secondary metabolite production (Table S2.2). This is interesting, as these are functions expected to be involved in species differentiation: adaptation to new carbon sources, regulation to fit new niches, and bioactive compounds for competitive advantages. This indeed seems to be the case in these *Aspergillus* species.

Genetic variation at the section level shows more species diversity in section *Cavernicolus*

We investigated the inter- and intra-section genetic variation in *Usti* and *Cavernicolus* to further validate the division of *Usti*. *Usti* shares 56 % of its proteome with section *Nidulantes* and contains 13 % species-unique proteins, whereas *Cavernicolus* shares 62 % of its proteome with section *Usti* and contains 25 % species-unique proteins (Fig. 2D–E). While the large number of species-unique proteins in *Cavernicolus* are interesting from a diversity aspect, it could indicate that the genetic variation of this section is not properly reflected due to the small number of species representing this section compared to section *Usti*.

Examining the number of protein families specific to each section further supports the genetic variation within *Cavernicolus*. *Usti* has 45 section-specific protein families compared to eight in *Cavernicolus* (Fig. 2A). To see what biological functions might separate these sections, we examined the functions in the section-specific protein families, and as expected, the functions that are specific to each section are connected to environmental adaptation, such as regulation, metabolism, and catabolism including specific carbohydrate-active enzymes (Table S3).

Sections *Usti* and *Cavernicolus* are the most CAZyme-rich species observed in genus *Aspergillus*

Most species in genus *Aspergillus* are either saprobic or opportunistic pathogens, secreting a variety of carbohydrate-active enzymes (CAZymes) to degrade plant biomass that is used as a main carbon source. Thus, it is interesting to investigate the diversity of this trait among the species of this genus. We predicted the CAZyme content in the genomes using the CAZy database (Cantarel *et al.* 2009, Drula *et al.* 2022) (Fig. 3, Table S4).

Overall, the species of sections *Usti, Nidulantes*, and *Cavernicolus* have a higher number of CAZymes than nearly all other species in the comparison, 538–764 per species, with *A. egyptiacus* as an exception with only 427 predicted CAZymes (Fig.

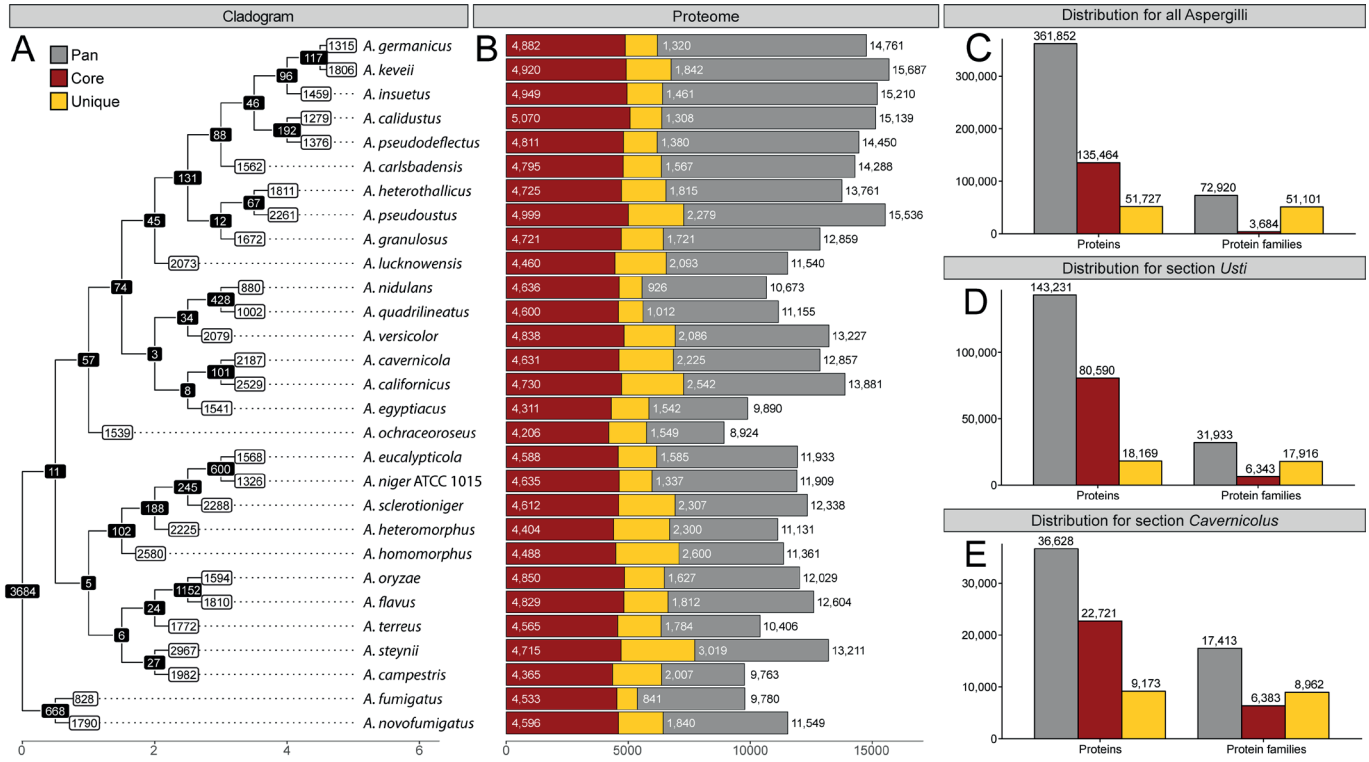

Fig. 2. Genetic diversity between fungal species with different taxonomic relations. **A.** A dendrogram representation of the phylogenetic relation between the 32 Aspergilli. The black boxes represent the homologous protein families shared among the species branching from the nodes. The white boxes represent the protein families unique to the individual species. **B.** A stacked histogram of the number of proteins included in the accessory (grey), core (red), and species-unique (yellow) protein families for each species. The numbers represent the total number of proteins (pan) for each species. **C–E.** Histogram representation of the number of proteins and families distributed over pan (grey), core (red), and species-unique (yellow) proteins for **C.** 29 Aspergilli (genus *Aspergillus*). **D.** Ten section *Usti* species. **E.** Three section *Cavernicolus* species.

3A). Section *Cavernicolus*, and in particular *Usti* contain the most CAZyme-rich species seen in genus *Aspergillus* (Fig. 3A). This is also the case when normalizing to the number of predicted genes, showing that particularly an expansion in glycoside hydrolases is the cause of the large CAZome in these species (Fig. 3D).

CAZyme potential and growth on plant polysaccharides shows that section *Usti* is an underutilized source of plant biomass degraders

In order to see how the CAZy potential was reflected in phenotypes and to assess inter-section variation, we further combined the CAZy analysis (Fig. 3A) with an analysis of the target polysaccharide for the CAZymes (Fig. 3B), and a growth of all species compared across 34 different carbon sources (Fig. S5).

Species from section *Usti* are particularly rich in the number of plant biomass degradation associated genes, with only *A. versicolor* from section *Nidulantes* having a similar number of genes. However, there is still clear variation visible within section *Usti*, with *A. lucknowensis* and *A. granulosus* exhibiting the lowest gene number and *A. insuetus* and *A. germanicus* the highest (Fig. 3B). Differences are particularly noticeable in the number of cellulolytic genes, while differences in the number of genes related to other polysaccharides are much smaller. This rich content related to plant biomass degradation is also reflected in the growth profile of these species, which all grow very well on a wide range of plant biomass related substrates (Fig. S5). In particular, the good growth on inulin of most species stands out, which is highly likely related to the presence of endoinulinases in the genomes of most species of this section (Table S5). The one species growing poorly on inulin, *A. heterothallicus*, also lacks genes encoding endoinulinase. This species also grew poorly on guar gum (mainly galactomannan), in contrast to all the other members of section *Usti*, which cannot be explained based on genome content. Considering their rich plant biomass degrading arsenal and their good growth on most plant biomass substrates, it is surprising that members of this section have not received attention as industrial enzyme producers. Possibly, the production of undesired secondary metabolites (see below) by some of the section members has discouraged exploration of the enzymatic potential of *Usti*.

In contrast, section *Cavernicolus* has relatively fewer plant biomass degradation-related CAZy genes, being more similar to *A. flavus* and *A. oryzae* especially for *A. californicus* and *A. cavernicola*, while *A. egyptiacus* has a number that is similar to *A. niger*, a well-described industrial enzyme producer (Vesth *et al.* 2018). Surprisingly, we were not able to compare the growth phenotype

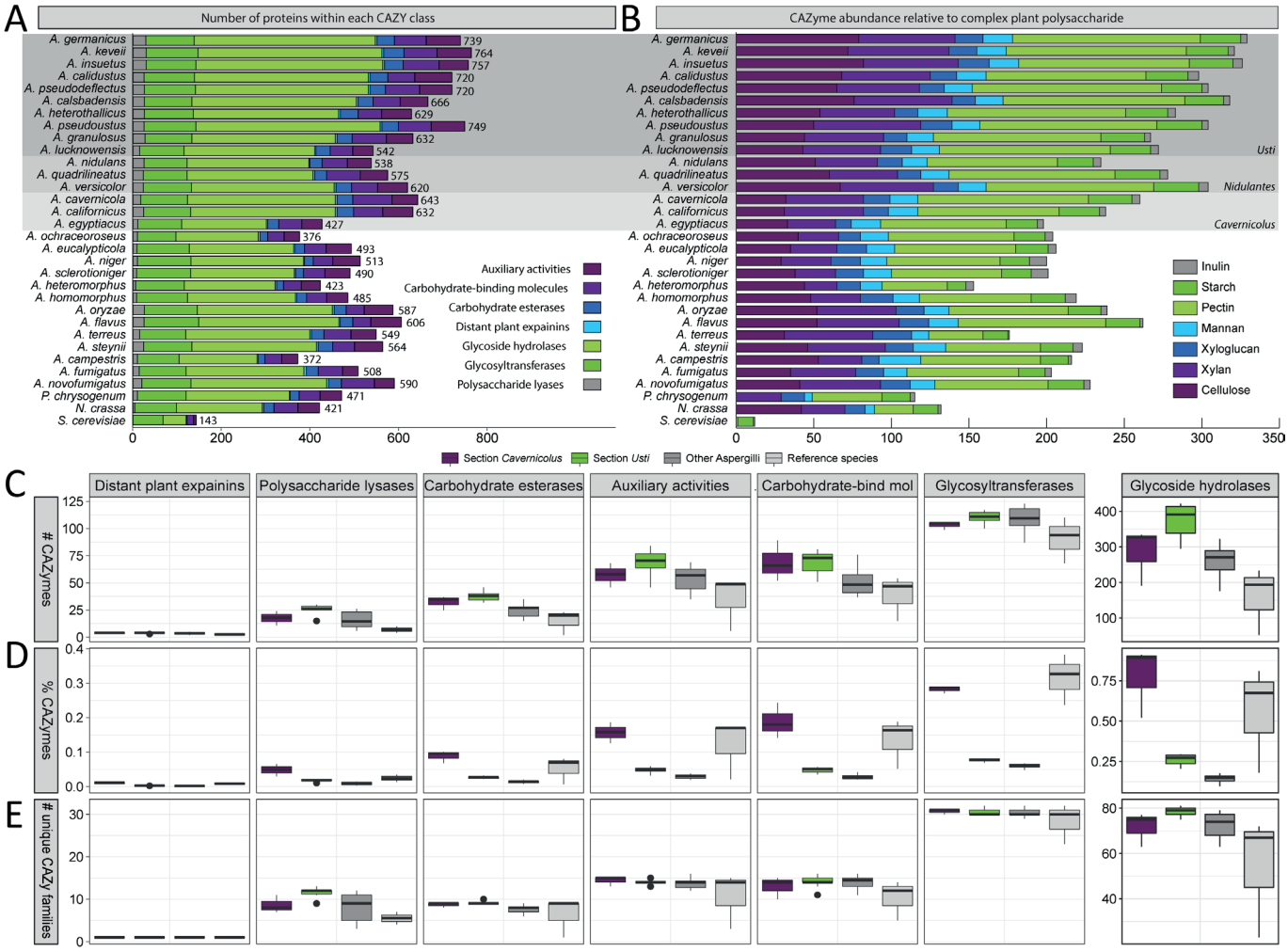

Fig. 3. Comparison of CAZy genome content by enzyme classes, associated modules, and target carbohydrate. The seven classes of enzyme activity are auxiliary activities, carbohydrate-binding molecules, carbohydrate esterases, distant plant expansins, glycoside hydrolases, glycosyltransferases, and polysaccharide lyases. **A.** Histogram representation of the seven overall enzyme classes per species. **B.** CAZyme abundance sorted according to target polysaccharide. **C–E.** Boxplot representing the diversity of CAZyme amount and content among section *Cavernicolus* (purple), section *Usti* (green), the 16 published Aspergilli (dark grey), and the non-*Aspergillus* species (light grey). For each CAZy: **C.** The total number of CAZymes. **D.** Percentage CAZymes in relation to proteome sizes. **E.** Number of unique CAZy families. Details on CAZy annotated proteins are available in Table S5.

and the CAZy content for *A. egyptiacus*, because this species appears to have an auxotrophic mutation that results in zero to very low growth on pure mono-, oligo- and polysaccharides. However, the other two species from section *Cavernicolus* grow in general well on polysaccharides, although somewhat less than species from section *Usti*. Species from section *Usti* also demonstrate better growth on a number of monomeric sugars, suggesting that they are adapted to a wider diversity of carbon sources than species from the section *Cavernicolus*.

Section *Usti* has a diverse CAZome as shown by species-unique CAZymes and CAZy-families

Due to the very high number of predicted CAZymes in sections *Usti* and *Cavernicolus*, we analyzed how many of the CAZymes are unique to the species and the sections, by examining the CAZy protein families and how they are shared between species. Although the numbers and percentages change between the different sets of *Aspergillus* species, they all show similar classes of CAZy protein families (Figs 3E, S6, S7).

In total, 1011 different CAZy protein families are found in sections *Cavernicola, Usti*, and *Nidulantes* (Table S4). The sections *Usti* and *Cavernicolus* share more than 89 % of their CAZy families. Interestingly, there are no CAZyme families which are found in all members of a section and nowhere else, but section *Usti* in particular contains species-unique CAZymes. Of the 202 CAZy families, 6 % are species-unique, whereas 83 % are shared by ten or more Aspergilli. This analysis confirms the potential of section *Usti* as a source of novel enzymes for biotechnology.

More than 10 % of secondary metabolite gene clusters in sections *Cavernicolus*, *Nidulantes*, and *Usti* are only found in a single species

A major reason for fungal adaptation is the production of secondary metabolites, which are involved in the interaction among species in their respective habitats as *e.g.*, communication molecules or antibiotics. We examined the secondary metabolism potential in these genomes to assess how related gene clusters have contributed to the genetic diversity in genus *Aspergillus*, and

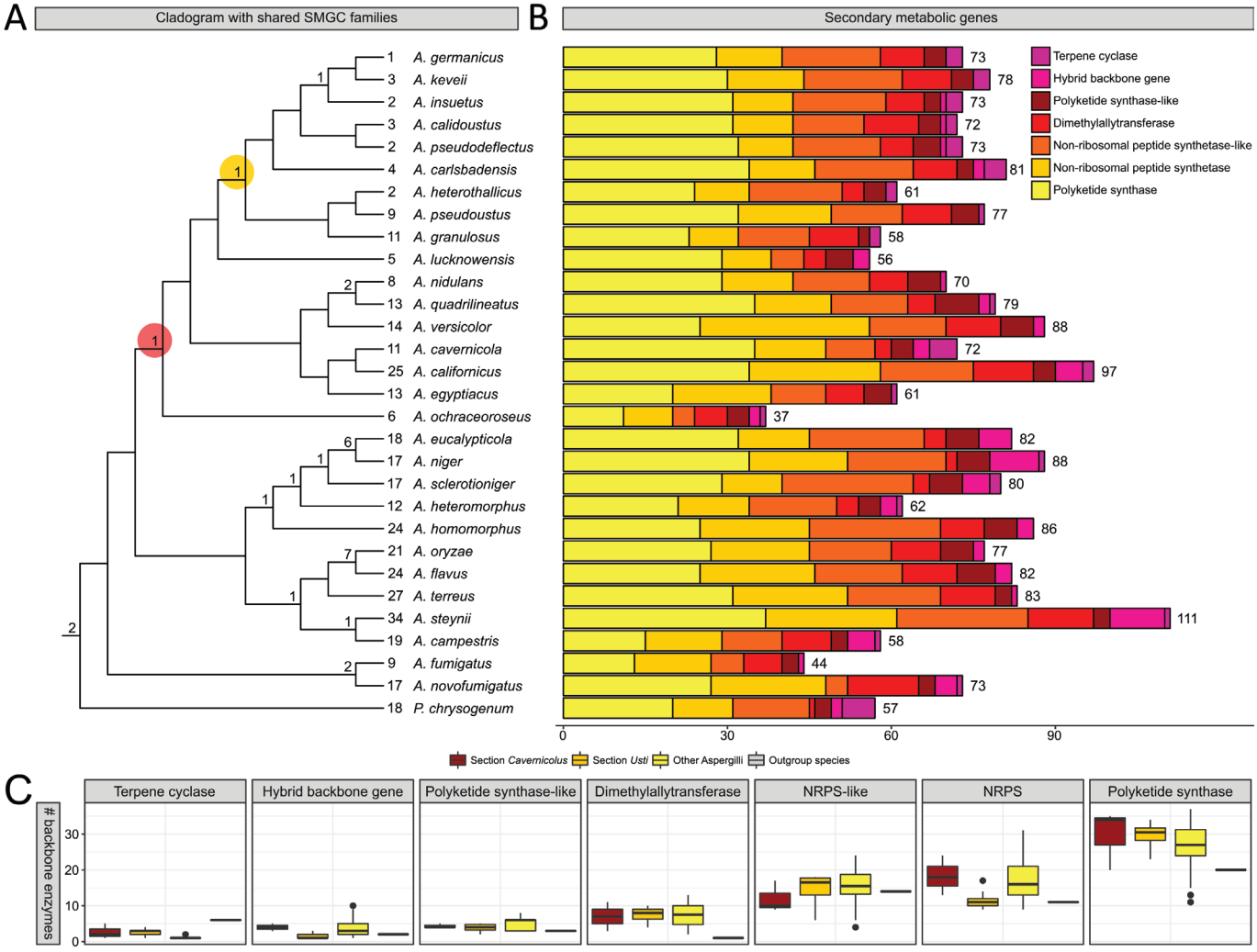

Fig. 4. Shared secondary metabolite gene clusters (SMGCs) and types of cluster backbone enzymes for 29 Aspergilli and *Penicillium chrysogenum*. **A.** Cladogram illustrating the SMGC families shared by species in the branch. Tip numbers display the number of families unique to each species. **B.** Histogram representation of the total SMGC profile of the species classified by backbone enzyme type. Boxplots representing the diversity of backbone enzymes among section *Cavernicolus* (red), section *Usti* (yellow), the 16 published Aspergilli (pale yellow), and the non-*Aspergillus* species (light grey). 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; Hybrid backbone gene: A backbone gene containing domains from NRPS and PKS backbones. Details on secondary metabolite gene clusters and backbone enzymes are available in Table S6.

potentially the divergence between sections *Cavernicolus* and *Usti*. We first identified secondary metabolite gene clusters (SMGCs) using a SMURF-like prediction tool (Khaldi *et al.* 2010, Vesth *et al.* 2018) (Fig. 4B), and grouped these into SMGC families believed to produce the same compound or variants thereof (Fig. 4A, Table S6). The analysis was performed on all species in the comparative set except *N. crassa* and *S. cerevisiae*.

We identified 1778 distinct SMGCs, which were grouped into 582 SMGC families in the *Aspergillus* species of Fig. 4A. To understand more about what type of SMGCs and families that are shared among the species, each of the 582 families were characterized by the enzyme that catalyses the formation of the backbone chemical structure of a given SM (the backbone enzyme) (Fig. 4B). The SMGC profile across the sections is highly dynamic, with a predominance of polyketide synthases (PKSs), non-ribosomal peptide synthetases (NRPSs), and NRPS-like enzymes (Fig. 4B, C). *Cavernicolus* contains a larger number of NRPSs and PKSs compared to *Usti*, which, comparatively, includes a larger number of NRPS-like enzymes (Fig. 4C).

Looking further into the diversification among the SMGC families in the *Aspergillus* species, we discovered large diversity: 92 % of the SMGC families cover in less than 10 species, and 64 % of the families are species-specific. However, a lot of this stems from the phylogenetic distance between the reference species. Even so, 126 out of the 1169 SMGCs (10.8 %) in sections *Cavernicolus*, *Nidulantes* and *Usti* are only found in one species (Fig. 4A, Table S6).

Sections *Cavernicolus* and *Usti* have obtained the majority of SM biosynthesis clusters after the divergence of the sections

We further wanted to investigate the SMGC diversity of sections *Cavernicolus* and *Usti*, and in particular, whether any have obtained section-specific SMGCs (Fig. 4A, Table S6). The species of section *Usti* has 604 SMGCs in 154 SMGC families, and *Cavernicolus* 195 SMGCs in 134 families (Table S6). Each section contains 36 % SMGC families that are only found in that section. In total 64 families are shared across the sections, but none of these were found only in the two sections (Table S6). The closest SMGC family shared by *Usti* and *Cavernicolus* includes additional species from sections *Nidulantes* and *Ochraceorosei* (Fig. 4A, Table S6). As they do not share unique sets of secondary metabolites, this supports that section *Nidulantes* indeed bifurcates sections *Usti* and *Cavernicolus*. Furthermore, the fact that both sections have as much as 36 % clusters unique to the section, suggests that the majority of SMGCs have been acquired after the divergence of the two sections.

Clade-specific SMGCs are found including a potential cluster for pigment biosynthesis

Searching for section- and clade- specific SMGCs that are found in all species members only resulted in one SMGC family present in all species in *Usti*, except *A. lucknowensis* (Fig. 4A, yellow dot). None were found exclusively in all members of section *Cavernicolus*,

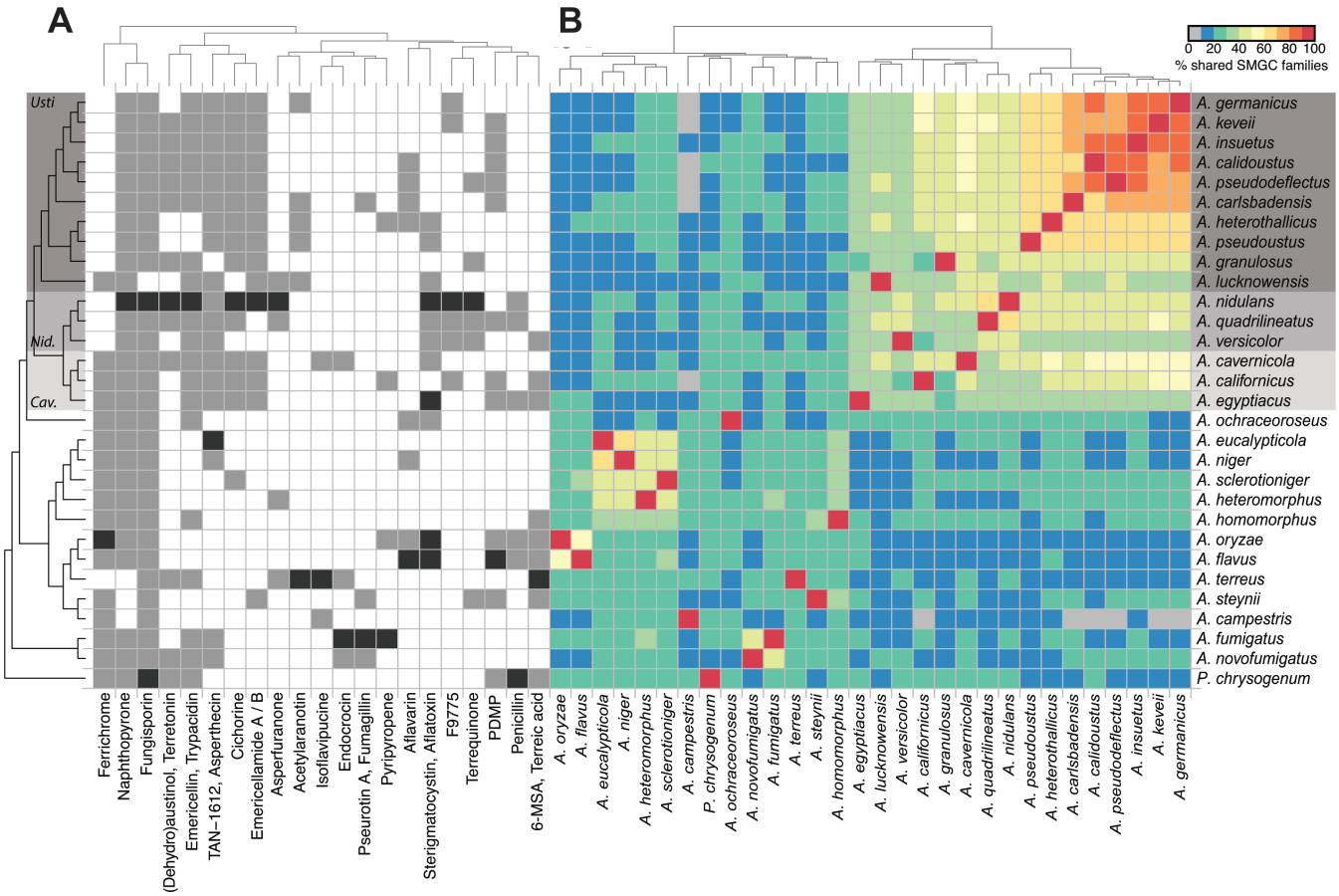

Fig. 5. Shared secondary metabolism gene clusters (SMGCs) and associated MIBiG related compounds. **A.** Association of known MIBiG compounds to SMGC families. Black boxes represent SMGCs with the highest score within a family to a gene cluster in the MIBiG database. Gray boxes illustrate the SMGC members in each family with an associated MIBiG compound. The three sections (*Usti, Nidulantes,* and *Cavernicolus*) are marked in the cladogram. PDMP: 4,4'-piperazine-2,5-diyldimethyl-bis-phenol. 6-MSA: 6-methylsalicylic acid. **B.** A heatmap representation of shared SMGC families among the species displayed in percentage by cell colour.

but one SMGC family was found in all included members of four sections (Fig. 4a, red dot).

These two SMGC families were further analysed to see if they may have divergence potential. The *Usti*-specific SMGC family was shown to include an NRPS-like gene (Fig. S8) and the four-section SMGC family was annotated to contain a PKS/NRPS-like hybrid gene (Fig. S9). A bioinformatic analysis of the respective clusters and comparison to compounds known to be produced by the species did not give direct leads as to which product(s) the clusters encode, although the second family bears close resemblance to a putative yellow pigment cluster found in *Talaromyces marneffii*. The conducted growth studies (Fig. S5) do not indicate a clear colour pattern for the four sections, suggesting multiple strategies for pigmentation. Further investigations will be required to accurately determine the activity of this cluster and its potential effect on species divergence.

De-replication of SMGCs similar to known *Aspergillus* gene clusters indicates dynamic loss and gain of clusters

The vast diversity of compounds produced in Aspergilli are known to only show low conservation across the genus, and therefore have been used to identify species in many cases (Frisvad *et al.* 2008, Frisvad & Larsen 2015). Hence, we were interested in identifying the conservation of known compounds within the genus and between sections *Usti* and *Cavernicolus*. The identified SMGC families (Theobald *et al.* 2018) were associated with characterized SMGCs from the Minimum Information on Biosynthetic Gene clusters (MIBiG) database (Li *et al.* 2016). Using a guilt-by-association approach we linked related SMGC families to known compounds from MIBiG, thus allowing us to find variants of compounds being produced by the identified SMGC families (Fig. 5A).

The analysis identified clusters largely conserved across the genus, such as YWA (naphtopyrone) (Mayorga & Timberlake 1992), nidulanin A (Andersen *et al.* 2013) and fungisporin (Ali *et al.* 2014) (Fig. 5A). It also identified clusters largely conserved across closely related sections such as emericellamide (Lukassen *et al.* 2015), asperthecin (Szewczyk *et al.* 2008), emericellin (Sanchez *et al.* 2011), terretonin (Guo *et al.* 2012), and cichorine (Ahuja *et al.* 2012), which have related SMGCs in sections *Usti, Nidulantes*, and *Cavernicolus*, and ferrichrome (Welzel *et al.* 2005) with related SMGCs in all sections besides *Usti* and *Nidulantes*. This pattern indicates that SMGCs have been acquired at different stages of differentiation with the potential of losses and horizontal transfers. Interestingly, the ferrichrome gene cluster seems to have been reacquired after the differentiation between sections *Cavernicolus* and *Nidulantes*, or it has been lost individually in sections *Usti* and *Nidulantes*.

The species in *Usti* are the most homogeneous sharing over 70 % of their SMGCs, whereas the species in *Cavernicolus* only share about 30–40 % (Fig. 5B, Fig. S10). Interestingly, *A. granulosus* and *A. lucknowensis* share fewer SMGCs (40–50 %) compared to the rest of the species in *Usti*, which makes them as diverse as species from section *Nidulantes*.

Sections *Cavernicolus* and *Usti* demonstrate a vast potential for production of varieties of known bioactive compounds

Apart from the conserved gene clusters, some SMGCs were only found in a few species with large phylogenetic distance, suggestive of recent horizontal gene transfers. Associating the MIBiG compounds to the SMGCs in *Cavernicolus* and *Usti*, we discovered that many of the related gene clusters have medical properties, such as compounds with anti-bacterial, -viral, -fungal, and -insectan activity (penicillin (Smith *et al.* 1990), acetylaranotin (Guo *et al.* 2013), and aflavarin (Cary *et al.* 2015), antiangiogenic, anticancer and immunosuppressant activity (fumagillin (Lin *et al.* 2013), pseurotin A (Zou *et al.* 2014), terrequinone (Bok *et al.* 2005), asperfuranone (Chiang *et al.* 2009), and endocrocin (Lim *et al.* 2012), as well as cholesterol-lowering abilities (pyripyropene (Itoh *et al.* 2010)). The sections further include SMGCs with industrial use and agricultural implications, such as acids (F9775 (orsellinic acid) (Sanchez *et al.* 2009), and terreic acid (Guo *et al.* 2014), and mycotoxins (aflatoxin (Ehrlich *et al.* 2004) and sterigmatocystin (Brown *et al.* 1996)). Especially the species in *Usti* include SMGCs with a large variety of antimicrobial and anticancer potential.

The guilt-by-association approach discovered both SMGCs that were conserved across genus- and section-level, but also SMGCs which could have been acquired through horizontal transfers. Of the 582 SMGC families, 36 were associated with a MIBiG gene cluster, which provides hints as to what type of compound the SMGC family might produce. Many of which have medically, agriculturally, and biotechnologically relevant applications, and should be investigated further.

CONCLUSIONS

The data presented in this study emphasizes the high genome diversity present across both section- and genus-level in *Aspergillus*. Even if Aspergilli only share 38 % of their proteomes, the functions are conserved over different taxonomic levels. Besides primary metabolism, genus *Aspergillus* share a large variety of CAZymes for substrate utilization, which could facilitate their ability to occupy a broad range of habitats. As opposed to the conservation seen in substrate utilization, genus *Aspergillus* express a large diversity in SMs, many of which are species-specific. This diversity in SMGCs is seen across both the amount and the type of compounds they might produce. Reinforced by the difference in genome sizes, core proteomes, species-unique proteins, whole-genome phylogeny, and SM profiles, we support the split of *Usti* into two sections as first proposed by Chen *et al.* 2016.

Both section *Cavernicolus* and *Usti* show high potential for new secondary metabolites as well as varieties of known bioactive compounds and CAZymes, two topics which based on this work can be investigated further.

AKNOWLEDGEMENTS

This work was supported by The Villum Foundation to JLN, TCV, ST, and MRA (grant number VKR023437), the Danish National Research Foundation to MRA, TCV and JCF (grant number DNRF137, CeMiSt), the Novo-Nordisk Foundation to BH and to MRM (grant number NNF21OC0067087), the Research Council of Finland to MRM (grant number 314102). The work (10.46936/10.25585/60001025) performed at the US Department of Energy (DOE) Joint BioEnergy Institute and the US DOE Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the US DOE.

DECLARATION ON CONFLICT OF INTEREST

BAS has a financial interest in Illium Technologies, Caribou Biofuels, and Erg Bio. All of the other authors declare no conflict of interest. Several of the authors are currently in the employ of the biotechnology company Novozymes A/S, which uses filamentous fungi for production of enzymes, but the company did not have a role in the design of the study, nor the publishing and editing of this paper.

REFERENCES

Ahuja M, Chiang YM, Chang SL. *et al.* (2012). Illuminating the diversity of aromatic polyketide synthases in *Aspergillus nidulans*. *Journal of the American Chemical Society* **134**: 8212–8221.

Ali H, Ries MI, Lankhorst PP. *et al.* (2014). A Non-Canonical NRPS is involved in the synthesis of fungisporin and related hydrophobic cyclic tetrapeptides in *Penicillium chrysogenum*. *PLoS ONE* **9**: e98212.

Altschul SF, Gish W, Miller W. *et al.* (1990). Basic local alignment search tool. *Journal of Molecular Biology* **215**: 403–410.

Amselem J, Lebrun MH, Quesneville H (2015). Whole genome comparative analysis of transposable elements provides new insight into mechanisms of their inactivation in fungal genomes. *BMC Genomics* **16**: 1–14.

Andersen MR, Salazar MP, Schaap PJ, *et al.* (2011). Comparative genomics of citric-acid-producing *Aspergillus niger* ATCC 1015 versus enzyme-producing CBS 513.88. *Genome Research* **21**: 885–897.

Andersen MR, Nielsen JB, Klitgaard AK, *et al.* (2013). Accurate prediction of secondary metabolite gene clusters in filamentous fungi. *Proceedings of the National Academy of Sciences USA* **110**: 99–107.

Arnaud MB, Cerqueira GC, Inglis DO, *et al.* (2012). The *Aspergillus* Genome Database (AspGD): recent developments in comprehensive multispecies curation, comparative genomics and community resources. *Nucleic Acids Research* **40**: D653–659.

Ashburner M, Ball CA, Blake JA, *et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* **25**: 25–29.

Bao W, Kojima KK, Kohany O (2015). Repbase update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **2**: 6–11.

Bian C, Kusuya Y, Sklenář F, *et al.* (2022). Reducing the number of accepted species in *Aspergillus* series *Nigri*. *Studies in Mycology* **102**: 95–132.

Bok JW, Hoffmeister D, Maggio-Hall LA, *et al.* (2005). Genomic mining for *Aspergillus* natural products. *Chemical Biology* **13**: 31–7.

Brown DW, Yu JH, Kelkar HS, *et al.* (1996). Twenty-five coregulated transcripts define a sterigmatocystin gene cluster in *Aspergillus nidulans*. *Proceedings of the National Academy of Sciences of the USA* **93**: 1418–1422.

Camacho C, Coulouris G, Avagyan V, *et al.* (2009). Blast+: architecture and applications. *BMC Bioinformatics* **10**: 1–9.

Cantarel BL, Coutinho PM, Rancurel C, *et al.* (2009). The carbohydrate-active enzymes database (CAZy): an expert resource for glycogenomics. *Nucleic Acids Research* **37**: 233–238.

Cary JW, Han Z, Yin Y, *et al.* (2015). Transcriptome analysis of *Aspergillus flavus* reveals *veA*-dependent regulation of secondary metabolite gene clusters, including the novel aflavarin cluster. *Eukaryotic Cell* **14**: 983–997.

Chen AJ, Frisvad JC, Sun BD, *et al.* (2016). *Aspergillus* section *nidulantes* (formerly *Emericella*): Polyphasic taxonomy, chemistry and biology. *Studies in Mycology* **84**: 1–118.

Chiang YM, Szewczyk E, Davidson AD, *et al.* (2009). A gene cluster containing two fungal polyketide synthases encodes the biosynthetic pathway for a polyketide, asperfuranone, in *Aspergillus nidulans*. *Journal of the American Chemical Society* **131**: 2965–2970.

Cock PJ, Antao T, Chang JT, *et al.* (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**: 1422–1423.

Csárdi G, Nepusz T (2006). The igraph software package for complex network research. *InterJournal Complex Systems* **1695**: 1–9.

Dashtban M, Schraft H, Syed TA, *et al.* (2010). Fungal biodegradation and enzymatic modification of lignin. *International Journal of Biochemistry and Molecular Biology* **1**: 36–50.

de Vries RP, Riley R, Wiebenga A, *et al.* (2017). Comparative genomics reveals high biological diversity and specific adaptations in the industrially and medically important fungal genus *Aspergillus*. *Genome Biology* **18**: 28.

Drula E, Garron ML, Dogan S, *et al.* (2022). The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Research* **50**: 571–577.

Ehrlich KC, Chang PK, Yu J, *et al.* (2004). Aflatoxin biosynthesis cluster gene *cypA* is required for g aflatoxin formation. *Applied and Environmental Microbiology* **70**: 6518–6524.

Finn RD, Attwood TK, Babbitt PC, *et al.* (2017). Interpro in 2017—beyond protein family and domain annotations. *Nucleic Acids Research* **45**: 190–199.

Flutre T, Duprat E, Feuillet C, *et al.* (2011). Considering transposable element diversification in *de novo* annotation approaches. *PLoS one* **6**: e16526.

Frisvad JC and Larsen TO (2015). Chemodiversity in the genus *Aspergillus*. *Applied Microbiology and Biotechnology* **99**: 7859–7877.

Frisvad JC, Andersen B, Thrane U (2008). The use of secondary metabolite profiling in chemotaxonomy of filamentous fungi. *Mycological Research* **112**: 231–240.

Fulton TM, Chunwongse J, Tanksley SD (1995). Microprep protocol for extraction of DNA from tomato and other herbaceous plants. *Plant Molecular Biology Reporter* **13**: 207–209.

Galagan JE, Calvo SE, Borkovich KA, *et al.* (2003). The genome sequence of the filamentous fungus Neurospora crassa. *Nature* **422**: 859–868.

Galagan JE, Calvo SE, Cuomo C, *et al.* (2005). Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*. *Nature* 438: 1105–1115.

Gams W, Christensen M, Onions AH, *et al.* (1986). Infrageneric taxa of *Aspergillus*. In: *Advances in Penicillium and Aspergillus systematics* (Samson RA, Pitt JI, eds). Springer, Boston, Massachusetts, USA: 55–56.

Gene Ontology Consortium T (2017). Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Research* **45**: 331–338.

Gnerre S, Maccallum I, Przybylski D, *et al.* (2010). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Science of the USA* **108**: 1513–1518.

Goffeau A, Barrell BG, Bussey H, *et al.* (1996). Life with 6000 genes. *Science* **274**: 563–567.

Grabherr MG, Haas BJ, Yassour M, *et al.* (2011). Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nature Biotechnology* **29**: 644–652.

Grigoriev IV, Nikitin R, Haridas S, *et al.* (2014). MycoCosm portal: Gearing up for 1000 fungal genomes. *Nucleic Acids Research* **42**: 699–704.

Grigoriev IV, Nordberg H, Shabalov I, *et al.* (2012). The genome portal of the department of energy joint genome institute. *Nucleic Acids Research* **40**: 26–32.

Guo CJ, Sun WW, Bruno KS, *et al.* (2014). Molecular genetic characterization of terreic acid pathway in *Aspergillus terreus*. *Organic Letters* **16**: 5250–5253.

Guo CJ, Knox BP, Chiang YM, *et al.* (2012). Molecular genetic characterization of a cluster in *A. terreus* for biosynthesis of the meroterpenoid terretonin. *Organic Letters* **14**: 5684–5687.

Guo CJ, Yeh HH, Chiang YM, *et al.* (2013). Biosynthetic pathway for the epipolythiodioxopiperazine acetylaranotin in *Aspergillus terreus* revealed by genome-based deletion analysis. *Journal of the American Chemical Society* **135**: 7205–7213.

Horn F, Linde J, Mattern DJ, *et al.* (2016). Draft genome sequences of fungus *Aspergillus calidoustus*. *Genome Announcements* **4**: e00102–16.

Houbraken J, Due M, Varga J, *et al.* (2007). Polyphasic taxonomy of *Aspergillus* section *Usti*. *Studies in Mycology* **59**: 107–128.

Houbraken J, Kocsube S, Visagie CM, *et al.* (2020). Classification of *Aspergillus*, *Penicillium*, *Talaromyces* and related genera (*Eurotiales*): An overview of families, genera, subgenera, sections, series and species. *Studies in Mycology* **95**: 5–169.

Hubka V, Nováková A, Kolařík M, *et al.* (2015). Revision of *Aspergillus* section *Flavipedes*: seven new species and proposal of section *Jani sect. nov*. *Mycologia* **107**: 169–208.

Hubka V, Nováková A, Peterson SW, *et al.* (2016). A reappraisal of *Aspergillus* section *nidulantes* with descriptions of two new sterigmatocystin-producing species. *Plant Systematics and Evolution* **302**: 1267–1299.

Hunter S, Apweiler R, Attwood TK, *et al.* (2009). InterPro: the integrative protein signature database. *Nucleic Acids Research* **37**: 211–215.

Inglis DO, Binkley J, Skrzypek MS, *et al.* (2013). Comprehensive annotation of secondary metabolite biosynthetic genes and gene clusters of *Aspergillus nidulans*, *A. fumigatus*, *A. niger* and *A. oryzae*. *BMC Microbiology* **13**: 91.

Itoh T, Tokunaga K, Matsuda Y, *et al.* (2010). Reconstitution of a fungal meroterpenoid biosynthesis reveals the involvement of a novel family of terpene cyclases. *Nature Chemistry* **2**: 858.

Jin FJ, Watanabe T, Juvvadi PR, *et al.* (2007). Double disruption of the proteinase genes, *tppA* and *pepE*, increases the production level of human lysozyme by *Aspergillus oryzae*. *Applied Microbiology and Biotechnology* **76**: 1059–1068.

Jurjevic Z, Kubátová A, Kolarík M, *et al.* (2015). Taxonomy of *Aspergillus* section *Petersonii sect. nov*. encompassing indoor and soil-borne species with predominant tropical distribution. *Plant Systematics and Evolution* **301**: 2441–2462.

Katoh K and Standley DM (2013). MAFFT: multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**: 772–780.

Khaldi N, Seifuddin FT, Turner G, *et al.* (2010). SMURF: Genomic mapping of fungal secondary metabolite clusters. *Fungal Genetics and Biology* **47**: 736–741.

Kildgaard S, Mansson M, Dosen I, *et al.* (2014). Accurate dereplication of bioactive secondary metabolites from marine-derived fungi by UHPLC-DAD-QTOFMS and a MS/HRMS library. *Marine Drugs* **12**: 3681–3705.

Kis-Papo T, Weig AR, Riley R, *et al.* (2014). Genomic adaptations of the halophilic dead sea filamentous fungus *Eurotium rubrum*. *Nature Communications* **5**: 4745.

Kjærbølling I, Vesth TC, Frisvad JC, *et al.* (2018). Linking secondary metabolites to gene clusters through genome sequencing of six diverse *Aspergillus* species. *Proceedings of the National Academy of Sciences of the USA* **115**: 753–761.

Kjærbølling I, Vesth T, Frisvad JC, *et al.* (2020). A comparative genomics study of 23 *Aspergillus* species from section *Flavi*. *Nature Communications* **11**: 1106.

Knuf C and Nielsen J (2012). Aspergilli: Systems biology and industrial applications. *Journal of Biotechnology* **7**: 1147–1155.

Kocsubé S, Perrone G, Magistà D, *et al.* (2016). *Aspergillus* is monophyletic: Evidence from multiple gene phylogenies and extrolites profiles. *Studies in Mycology* **85**: 199–213.

Kozlovskii AG, Antipova TV, Zhelifonova VP, *et al.* (2016). Secondary metabolites of fungi of the *Usti* section, genus *Aspergillus* and their application in chemosystematics. *Microbiology* **86**: 176–182.

Li YF, Tsai KJS, Harvey CJB, *et al.* (2016). Comprehensive curation and analysis of fungal biosynthetic gene clusters of published natural products. *Fungal Genetics and Biology* **89**: 18–28.

Lim FY, Hou Y, Chen Y, *et al.* (2012). Genome-based cluster deletion reveals an endocrocin biosynthetic pathway in *Aspergillus fumigatus*. *Applied and Environmental Microbiology* **78**: 4117-4125.

Lin HC, Chooi YH, Dhingra S, *et al.* (2013). The fumagillin biosynthetic gene cluster in *Aspergillus fumigatus* encodes a cryptic terpene cyclase involved in the formation of β-trans-bergamotene. *Journal of the American Chemical Society* **135**: 4616–4619.

Lukassen MB, Saei W, Sondergaard TE, *et al.* (2015). Identification of the scopulariade biosynthetic gene cluster in *Scopulariopsis brevicaulis*. *Marine Drugs* **13**: 4331–4343.

Machida M, Asai K, Sano M, *et al.* (2005). Genome sequencing and analysis of *Aspergillus oryzae*. *Nature* **438**: 1157–1161.

Max B, Salgado JM, Rodríguez N, *et al.* (2010). Biotechnological production of citric acid. *Brazilian Journal of Microbiology* **41**: 862–875.

Mayorga ME and Timberlake WE (1992). The developmentally regulated *Aspergillus nidulans wA* gene encodes a polypeptide homologous to polyketide and fatty acid synthases. *Molecular Genetics and Genomics* **235**: 205–212.

McInerney JO, McNally A, O'Connell MJ (2017). Why prokaryotes have pangenomes. *Nature Microbiology* **2**: e201740.

Medema MH, Kottmann R, Yilmaz P, *et al.* (2015). Minimum information about a biosynthetic gene cluster. *Nature Chemical Biology* **11**: 625–631.

Meyer V, Wu B, Ram AF (2011). *Aspergillus* as a multi-purpose cell factory: Current status and perspectives. *Biotechnology Letters* **33**: 469–476.

Mitra A, Skrzypczak M, Ginalski K, *et al.* (2015). Strategies for achieving high sequencing accuracy for low diversity samples and avoiding sample bleeding using Illumina platform. *PLoS ONE* **10**: e0120520.

Montiel MD, Lee HA, Archer DB (2006). Evidence of rip (repeat-induced point mutation) in transposase sequences of *Aspergillus oryzae*. *Fungal Genetics and Biology* **43**: 439–445.

Nielsen KF, Mogensen JM, Johansen M, *et al.* (2009). Review of secondary metabolites and mycotoxins from the *Aspergillus niger* group. *Analytical and Bioanalytical Chemistry* **395**: 1225–1242.

Nierman WC, Pain A, Anderson MJ, *et al.* (2005) Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*. *Nature* **438**: 1151–1156.

Nordberg H, Cantor M, Dusheyko S, *et al.* (2014). The genome portal of the department of energy joint genome institute: 2014 updates. *Nucleic Acids Research* **42**: 26–31.

Jones P, Binns D, Chang HY, *et al.* (2014). Interproscan 5: genome-scale protein function classification. *Bioinformatics* **30**: 1236–1240.

Peterson S, Varga J, Frisvad J, *et al.* (2008). Phylogeny and subgeneric taxonomy of *Aspergillus*. In: *Aspergillus in the Genomic Era* (Varga J, Samson RA, eds). Wageningen Academic Publisher, Wageningen, Netherlands: 33–56.

Punt PJ, van Biezen N, Conesa A, *et al.* (2002). Filamentous fungi as cell factories for heterologous protein production. *Trends in Biotechnology* **20**: 200–206.

Quesneville H, Bergman CM, Andrieu O, *et al.* (2005). Combined evidence annotation of transposable elements in genome sequences. *PLoS Computational Biology* **1**: 166–175.

Raffaele S and Kamoun S (2012). Genome evolution in filamentous plant pathogens: why bigger can be better. *Nature Reviews Microbiology* **10**: 417–430.

Raper KB and Fennell DI (1965). *The genus Aspergillus*. Williams & Wilkins, Baltimore, Maryland, USA.

Richter L, Wanka F, Boecker S, *et al.* (2014). Engineering of *Aspergillus niger* for the production of secondary metabolites. *Fungal Biology and Biotechnology* **1**: 1–13.

Sambrook J, Russell WD (2012). *Molecular cloning: A laboratory manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor.

Samson RA, Varga J, Meijer M, *et al.* (2011). New taxa in *Aspergillus* section *Usti*. *Studies in Mycology* **69**: 81–97.

Sanchez JF, Chiang YM, Szewczyk E, *et al.* (2009). Molecular genetic analysis of the orsellinic acid/F9775 gene cluster of *Aspergillus nidulans*. *Molecular BioSystems* **6**: 587–593.

Sanchez JF, Entwistle R, Hung JH, *et al.* (2011). Genome-Based deletion analysis reveals the prenyl xanthone biosynthesis pathway in *Aspergillus nidulans*. *Journal of the American Chemical Society* **133**: 4010–4017.

Schuster E, Dunn-Coleman N, Frisvad JC, *et al.* (2002). On the safety of *Aspergillus niger* - a review. *Applied Microbiology and Biotechnology* **59**: 426–435.

Sharma R, Katoch M, Srivastava P, *et al.* (2009). Approaches for refining heterologous protein production in filamentous fungi. *World Journal of Microbiology and Biotechnology* **25**: 2083–2094.

Smedsgaard J (1997). Micro-scale extraction procedure for standardized screening of fungal metabolite production in cultures. *Journal of Chromatography* **760**: 264–270.

Smith DJ, Burnham MK, Edwards J, *et al.* (1990). Cloning and heterologous expression of the penicillin biosynthetic gene cluster from *Penicillium chrysogenum*. *Biotechnology* **8**: 39–41.

Soucy SM, Huang J, Gogarten J (2015). Horizontal gene transfer: building the web of life. *Nature Reviews Genetics* **16**: 472–482.

Stamatakis A (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313.

Szewczyk E, Chiang YM, Oakley CE, *et al.* (2008). Identification and characterization of the asperthecin gene cluster of *Aspergillus nidulans*. *Applied and Environmental Microbiology* **74**: 7607–7612.

Talavera G, Castresana J (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology* **56**: 564–577.

Theobald S, Vesth TC, Rendsvig JK, *et al.* (2018). Uncovering secondary metabolite evolution and biosynthesis using gene cluster networks and genetic dereplication. *Scientific Reports* **8**: e17957.

Varga J, Frisvad JC, Samson RA (2010). *Aspergillus* sect. *Aeni sect. nov*., a new section of the genus for *A. karnatakaensis sp. nov*. and some allied fungi. *IMA Fungus* **1**: 197–205.

Vesth TC, Nybo JL, Theobald S, *et al.* (2018). Investigation of inter- and intraspecies variation through genome sequencing of *Aspergillus* section *Nigri*. *Nature Genetics* **50**: 1688–1695.

Visagie CM, Yilmaz Y, Kocsube S, *et al.* (2024). A review of recently introduced *Aspergillus*, *Penicillium*, *Talaromyces* and other *Eurotiales* species. *Studies in Mycology* **107**: 1–66.

Wakai S, Arazoe T, Ogino C, *et al.* (2017). Future insights in fungal metabolic engineering. *Bioresource Technology* **245**: 1314–1326.

Ward M, Lin C, Victoria DC, *et al.* (2004). Characterization of humanized antibodies secreted by *Aspergillus niger*. *Applied and Environmental Microbiology* **70**: 2567–2576.

Welzel K, Eisfeld K, Antelo L, *et al.* (2005). Characterization of the ferrichrome a biosynthetic gene cluster in the homobasidiomycete *Omphalotus olearius*. *FEMS Microbiology Letters* **249**: 157–163.

Yamada R, Yoshie T, Sakai S, *et al.* (2015). Effective saccharification of kraft pulp by using a cellulase cocktail prepared from genetically engineered *Aspergillus oryzae*. *Bioscience, Biotechnology, and Biochemistry* **79**:1034–1037.

Zerbino DR, Birney E (2008). Velvet: algorithms for *de novo* short read assembly using de Bruin graphs. *Genome Research* **18**: 821–829.

Zou Y, Xu W, Tsunematsu Y, *et al.* (2014). Methylation-dependent acyl transfer between polyketide synthase and nonribosomal peptide synthetase modules in fungal natural product biosynthesis. *Organic Letters* **16**: 6390–6393.

2017: The genome portal of the department of energy joint genome institute: Published *Aspergillus* genomes (https://genome.jgi.doe.gov/Aspergillus/Aspergillus.info.html). Accessed: 2017-12-30.

Supplementary Material: https://studiesinmycology.org

Fig. S1. A cladogram representation of the phylogenetic relations between the species in this paper. The red labels show bootstrap values of 100 % and the black labels show bootstrap values < 100 %.

Fig. S2. Phylogenetic relation and InterPro coverage of the fungal species included in this study. **A.** Dendrogram of the phylogenetic relation between the 32 species. The black boxes represent the homologous proteins among the species branching from the nodes. The white boxes represent the proteins unique to the specific species. **B.** Proteome sizes of each species (gray). **C.** Core proteome size of each species (red). **D.** Species unique proteins (yellow). The dark colored boxes represent proteins with InterPro annotation, the light-colored boxes represent the proteins with no annotation. The numbers at right side of the boxes indicates the total number of annotated and not annotated genes. The bar scales are unique to each graph.

Fig. S3. Phylogenetic relation and InterPro coverage of all Aspergilli. **A.** Dendrogram of the phylogenetic relation between the 29 *Aspergillus* species. The black boxes represent the homologous proteins among the species branching from the nodes. The white boxes represent the proteins unique to the specific species. **B.** Proteome sizes of each species (gray). **C.** Core proteome size of each species (red). **D.** Species unique proteins (yellow). The dark colored boxes represent proteins with InterPro annotation, the light-colored boxes represent the proteins with no annotation. The numbers at right side of the boxes indicates the total number of annotated and not annotated genes. The bar scales are unique to each graph.

Fig. S4. Phylogenetic relation and InterPro coverage of the species in section *Usti*. **A.** Dendrogram of the phylogenetic relation between the 13 species. The black boxes represent the homologous proteins among the species branching from the nodes. The white boxes represent the proteins unique to the specific species. **B.** Proteome sizes of each species (gray). **C.** Core proteome size of each species (red). **D.** Species unique proteins (yellow). The dark colored boxes represent proteins with InterPro annotation, the light-colored boxes represent the proteins with no annotation. The numbers at right side of the boxes indicates the total number of annotated and not annotated genes. The bar scales are unique to each graph.

Fig. S5. Growth profile of the different species using a wide range of plant biomass related substrates. *A. granulosus* did not grow on any media, so this one is not present in the picture.

Fig. S6. The collected number of CAZy families per species group. A histogram representation of the number of proteins within each CAZy class

for each CAZy family. The species groups are indicated by color. CAZy class; auxiliary activities (AA), carbohydrate-binding molecules (CBM), carbohydrate esterases (CE), polysaccharide lyases (PL), distant plant expansins (EXPN), glycosyltransferases (GT), and glycoside hydrolases (GH).

Fig. S7. The number of CAZy families in relation to species group proteome size (percentage). A histogram representation of the percentage of proteins within each CAZy class for each CAZy family. The species groups are indicated by color. CAZy class; auxiliary activities (AA), carbohydrate-binding molecules (CBM), carbohydrate esterases (CE), polysaccharide lyases (PL), distant plant expansins (EXPN), glycosyltransferases (GT), and glycoside hydrolases (GH).

Fig. S8. Synteny plot of a NRPS-like secondary metabolite gene cluster family shared by species from section *Usti*, section *Nidulantes*, section *Cavernicolus* and section *Ochraceorosei*. Annotated InterPro domains. **A.** NRPS-like. **B.** General substrate transporter. **C.** Carboxylesterase, type B. **D.** BTB/POZ. **E.** NAD-dependent epimerase/dehydratase. **H.** NmrA-like domain. **I.** Transcription factor domain. **J.** Alpha/beta hydrolase fold-1. **K.** Domain of unknown function. **L.** Glucose-methanol-choline oxidoreductase. **M.** Major facilitator superfamily. **N.** Male sterility. **O.** Short-chain dehydrogenase/reductase SDR. **P.** Transferase. **Q.** WD40 repeat. **R.** Zinc/iron permease. **S.** Zn(2)-C6 fungal-type DNA-binding domain. **T.** Transcription factor domain. NA: no annotation.

Fig. S9. Percentage shared secondary metabolite gene clusters between the 29 *Aspergillus* species and *Penicillium chrysogenum*. The heatmap has been clustered based on the number of shared secondary metabolites on both axis.

Fig. S10. Synteny plot of a NRPS-like secondary metabolite gene cluster family shared by all species from section *Usti* except for *A. lucknowensis.* Annotated InterPro domains. **A.** Cytochrome P450. **B.** NAD-dependent epimerase/dehydratase. **C.** PKS. **D.** Major facilitator superfamily. **E.** NRPS-Like. **H.** Oxoglutarate/iron-dependent dioxygenase. **I.** O-methyltransferase, family 2. **J.** FAD linked oxidase. **K.** FAD linked oxidase. **L.** GNAT domain. **M.** Rad1/Rec1/Rad17. **N.** SANT/Myb domain. **O.** General substrate transporter. **P.** Heavy metal-associated domain. **Q.** ABC transporter-like. **R.** Alpha/beta hydrolase fold-1. **S.** Alpha/beta hydrolase fold-5. **T.** Ankyrin repeat-containing domain. **U.** Cell wall galactomannoprotein. **V.** Cytochrome P450. **W.** Domain of unknown function. **X.** Translation elongation factor EF1B. **Y.** Fatty acid desaturase, type 1. **Z.** Fatty acid hydroxylase. **AA.** General substrate transporter. **AB.** Glutathione S-transferase. **AC.** Glycoside hydrolase. **AD.** Glycosyltransferase family 28. **AE.** Male sterility, NAD-binding. **AF.** Mitochondrial substrate/solute carrier. **AG.** MmgE/PrpD. **AH.** NmrA-like domain. **AI.** Protein kinase domain. **AJ.** Pyridine nucleotide-disulphide oxidoreductase. **AK.** Taurine catabolism dioxygenase TauD/TfdA. **AL.** Transcription factor domain. **AM.** Transcription factor domain, fungi. NA: no annotation.

Table S1. Sequencing and annotation statistics of the investigated species.

Table S2.1. The InterPro domains for the core proteome of genus *Aspergillus*. The InterPro domains found in Vesth *et al.* (2018) is marked by x.

Table S2.2. The InterPro domains for the species-unique proteins of genus *Aspergillus*. The InterPro domains found in Vesth *et al.* (2018) is marked by x.

Table S3. Identification of unique InterPro domains per section.

Table S3.1. The InterPro domains of the proteins unique to all species in section *Usti*.

Table S3.2. The InterPro domains of the proteins unique to all species in section *Cavernicolus*.

Table S4. Proteins with predicted CAZy families and definitions of all species in this study.

Table S5. Comparison of CAZy genome content.

Table S5.1. Total numbers of Glycoside Hydrolases (GH), Glycosyl Transferases (GT), Polysaccharide Lyases (PL), Carbohydrate Esterases (CE), Carbohydrate Binding Modules (CBM) and Auxiliary Activities (AA).

Table S5.2. Detailed comparison of Glycoside Hydrolases (GH).

Table S5.3. Detailed comparison of Glycosyl Transferases (GT).

Table S5.4. Detailed comparison of Polysaccharide Lyases (PL).

Table S5.5. Detailed comparison of Carbohydrate Esterases (CE).

Table S5.6. Detailed comparison of Carbohydrate Binding Modules (CBM).

Table S5.7. Detailed comparison of Auxiliary Activities (AA).

Table S5.8. Cellulose-related families.

Table S5.9. Xylan-related families.

Table S5.10. Mannan-related families.

Table S5.11. Xyloglucan-related families.

Table S5.12. Pectin-related families.

Table S5.13. Starch-related families.

Table S5.14. Inulin-related family GH32.

Table S5.15. Comparison of CAZymes related to plant biomass degradation.

Table S6. The predicted secondary metabolite gene clusters and their InterPro annotations and MIBiG associated compounds.