bioRxiv preprint doi: https://doi.org/10.1101/149856; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Drivers of genetic diversity in secondary metabolic gene clusters in a fungal population

Abigail L. Lind[1], Jennifer H. Wisecaver[2], Catarina Lameiras[3], Fernando Rodrigues[4,5], Gustavo H. Goldman[6], Antonis Rokas[1,2]

1. Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, Tennessee, USA.

2. Department of Biology, Vanderbilt University, Nashville, Tennessee, USA.

3. Department of Microbiology, Portuguese Oncology Institute of Porto, Porto, Portugal

4. Life and Health Sciences Research Institute (ICVS), School of Medicine, University of Minho, Braga, Portugal

5. ICVS/3B's - PT Government Associate Laboratory, Braga/Guimarães, Portugal.

6. Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Universidade de São Paulo, São Paulo, Brazil

†Corresponding author and lead contact: antonis.rokas@vanderbilt.edu

bioRxiv preprint doi: https://doi.org/10.1101/149856; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Summary

Filamentous fungi produce a diverse array of secondary metabolites (SMs) critical for defense, virulence, and communication. The metabolic pathways that produce SMs are found in contiguous gene clusters in fungal genomes, an atypical arrangement for metabolic pathways in other eukaryotes. Comparative studies of filamentous fungal species have shown that SM gene clusters are often either highly divergent or uniquely present in one or a handful of species, hampering efforts to determine the genetic basis and evolutionary drivers of SM gene cluster divergence. Here we examined SM variation in 66 cosmopolitan strains of a single species, the opportunistic human pathogen *Aspergillus fumigatus.* Investigation of genome-wide population-level variation showed that *A. fumigatus* strains contained five general types of variation in SM gene clusters: non-functional gene polymorphisms, gene gain and loss polymorphisms, whole cluster gain and loss polymorphisms, allelic polymorphisms where different alleles corresponded to distinct, non-homologous clusters, and location polymorphisms in which a cluster was found to differ in its genomic location across strains. These polymorphisms affect the function of representative *A. fumigatus* SM gene clusters, such as those involved in the production of gliotoxin, fumigaclavine, and helvolic acid, as well as the function of clusters with undefined products. In addition to enabling the identification of polymorphisms whose detection requires extensive genome-wide synteny conservation (e.g., mobile gene clusters and non-homologous cluster alleles), our population genomics approach also implicated multiple underlying genetic drivers, including point mutations, recombination, genomic deletion and insertion events, as well as horizontal gene transfer from distant fungi. Finally, most of the population variants that we uncover have been previously hypothesized to contribute to SM gene cluster diversity across entire fungal classes and phyla. We suggest that the drivers of genetic diversity operating within a fungal population shown here are sufficient to explain SM cluster macroevolutionary patterns.

Keywords: chemodiversity, specialized metabolism, genome evolution, genome architecture, gene loss, genomic rearrangement

bioRxiv preprint doi: https://doi.org/10.1101/149856; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

Filamentous fungi produce a diverse array of small molecules that function as toxins, antibiotics, and pigments [1]. Though by definition secondary metabolites (SMs) are not strictly necessary for growth and development, they are critical to the lifestyle of filamentous fungi [2]. For example, antibiotic SMs gain give their fungal producers a competitive edge in environments crowded with other microbes [3]. SMs can additionally mediate communication between and within species, as well as contribute to virulence on animal and plant hosts in pathogenic fungi [4,5].

A genomic hallmark of SMs in filamentous fungi is that the biosynthetic pathways that produce them are typically organized into contiguous gene clusters in the genome [6]. These gene clusters contain the chemical backbone synthesis genes whose enzymatic products produce a core metabolite, such as non-ribosomal peptide synthases (NRPS) and polyketide synthases (PKS), tailoring enzymes that chemically modify the metabolite, transporters involved in product export, and transcription factors that control the expression of the clustered genes [6]. Filamentous fungal genomes, particularly those in the phylum Ascomycota [6], typically contain dozens of SM gene clusters. However, most individual SM gene clusters appear to be either species-specific or narrowly taxonomically distributed in only a handful of species [6,7]. SM gene clusters that are more broadly distributed show discontinuous taxonomic distributions and are often highly divergent between species. Consequently, the identity and total number of SM gene clusters can vary widely even between very closely related species whose genomes exhibit very high sequence and synteny conservation [8,9].

In the last decade, several comparative studies have described macroevolutionary patterns of SM gene cluster diversity. For example, studies centered on genomic comparisons of closely related species, such as members of the same genus, have identified several different types of inter-species divergence, from single nucleotide substitutions (e.g., differences in fumonisins produced by *Fusarium* species are caused by variants in one gene [10]), to gene gain / loss events (e.g., the trichothecene gene clusters in *Fusarium* species and the aflatoxin family

bioRxiv preprint doi: https://doi.org/10.1101/149856; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

SM gene clusters in *Aspergillus* species) [11–16], and genomic rearrangements (e.g., the trichothecene gene clusters in *Fusarium*) [11]. Additionally, genetic and genomic comparisons across fungal orders and classes have identified several instances of gene gain or loss [17–19] and horizontal gene transfer [13,20–23] acting on individual genes or on entire gene clusters, providing explanations for the diversity and discontinuity of the taxonomic distribution of certain SM gene clusters across fungal species.

Although inter-species comparative studies have substantially contributed to our understanding of SM diversity, the high levels of evolutionary divergence of SM clusters make inference of the genetic drivers of SM gene cluster evolution challenging; simply put, it has been difficult to "catch" the mechanisms that generate SM gene cluster variation "in the act". Several previous studies have examined intra-species or population-level differences in individual SM gene clusters, typically focusing on the presence and frequency of non-functional alleles of clusters involved in production of mycotoxins. Examples of clusters exhibiting such polymorphisms include the gibberellin gene cluster in *Fusarium oxysporum* [24], the fumonisin gene cluster in *Fusarium fujikuroi* [25], the aflatoxin and cyclopiazonic gene clusters in *Aspergillus flavus* [26], and the bikaverin gene cluster in *Botrytis cinerea* [27]. While these studies have greatly advanced our understanding of SM gene cluster genetic variation and highlighted the importance of population-level analyses, studies examining the entirety of SM gene cluster polymorphisms in fungal populations are so far lacking. We currently do not understand the types and frequency of SM gene cluster polymorphisms in populations, whether these polymorphisms affect all types of SM gene clusters, as well as the genetic drivers of SM gene cluster evolution.

To address these questions, we investigated the genetic diversity of all 36 known and predicted SM gene clusters in whole genome sequence data from 66 strains of the opportunistic human pathogen *Aspergillus fumigatus*, 8 of which were sequenced in this study. We found that 13 SM gene clusters were generally conserved and harbored low amounts of variation. In contrast, the remaining 23 SM gene clusters were highly variable and contained

bioRxiv preprint doi: https://doi.org/10.1101/149856; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

one or more of five different types of genetic variation: single-nucleotide polymorphisms including nonsense and frameshift variants, individual gene gain and loss polymorphisms, entire cluster gain and loss polymorphisms, polymorphisms associated with changes in cluster genomic location, and clusters with non-homologous alleles resembling the idiomorphs of fungal mating loci. Many clusters contained interesting combinations of these different polymorphisms, such as pseudogenization in some strains and entire cluster loss in others. The types of variants we find are likely generated by a combination of DNA replication and repair errors, recombination, genomic insertions and deletions, and horizontal transfer. We additionally find an enrichment for transposable elements (TEs) around horizontally transferred clusters, clusters that change in genomic locations, and idiomorphic clusters. Taken together, our results provide a guide to both the types of polymorphisms and the genetic drivers of SM gene cluster diversification in filamentous fungi. As most of the genetic variants that we observe have been previously associated with SM gene cluster diversity across much larger evolutionary distances and timescales, we argue that population-level processes influencing SM gene cluster diversity are sufficient to explain SM cluster macroevolutionary patterns.

bioRxiv preprint doi: https://doi.org/10.1101/149856; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Results

We analyzed the genomes of 66 globally distributed strains of *Aspergillus fumigatus* for polymorphisms in SM gene clusters. We performed whole-genome sequencing on 8 strains, and collected the remaining 58 strains from publicly available databases including NCBI Genome and the NCBI Short Read Archive (Figure 1, Table S1) [28–32]. We analyzed all strains for polymorphisms in 33 curated SM gene clusters present in the reference Af293 genome and additionally searched for novel SM gene clusters (see Methods). These examinations revealed five distinct types of polymorphisms which influence SM gene cluster variation (Table 1):

a) Single nucleotide and short indel polymorphisms. 33 / 33 SM gene clusters (present in the reference Af293 strain) contained multiple genes with missense SNPs and short indel variants in at least one strain in the population. 23 / 33 SM gene clusters contained one or more genes with frameshift or nonsense variants in the population.
b) Gene content polymorphisms involving loss or gain of one or more genes. 6 / 33 SM gene clusters contained a gene content polymorphism in the population.
c) Whole SM gene cluster gain and loss polymorphisms. 3 / 33 SM gene clusters present in the genome of the reference Af293 strain were absent in at least one strain in the population and an additional 3 previously unknown SM gene clusters were present in the population.
d) Idiomorphic polymorphisms. One locus contained multiple non-homologous SM gene cluster alleles in different strains of the population.
e) Genomic location polymorphisms. 2 / 33 SM gene clusters were found in different genomic locations (e.g., different chromosomes) between strains.

Single-nucleotide and indel polymorphisms

It is well established that single nucleotide polymorphisms (SNPs) and short indel polymorphisms are caused by errors in DNA replication and repair, and are a major source of genomic variation [33]. Non-synonymous SNPs and indels with missense, frameshift, and nonsense effects were widespread across the 33 SM reference gene clusters (Table S2). Every strain contained numerous missense mutations and at least one nonsense or frameshift

bioRxiv preprint doi: https://doi.org/10.1101/149856; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

mutation in its SM gene clusters. Although missense mutations are likely to influence SM production, the functional effects of nonsense and frameshift mutations are comparatively easier to infer from genomic sequence data because they often lead to truncated proteins lacking a significant portion of their amino acid sequence. For example, a frameshift mutation in the polyketide synthase (PKS) of the trypacidin gene cluster in the A1163 strain results in loss of trypacidin production [34]. Interestingly, we identified a premature stop codon (Gln273*) in a transcription factor required for trypacidin production, *tpcD,* in a strain sequenced in this study (MO79587EXP). These data suggest that function of this SM gene cluster has been lost through at least two independent genetic events in *A. fumigatus*.

Individual nonsense or frameshift variants ranged from very common in the population to rarer variants present in one or a handful of strains. For example, the non-ribosomal peptide synthase (NRPS) *pes3* gene (Afu5g12730) in SM gene cluster 21 harbors 16 nonsense or frameshift polymorphisms in 55 strains. Seven of these polymorphisms are common and present in 10 or more strains, while seven are rarer and found in 5 or fewer strains. Strains with lab-mutated null alleles of the *pes3* gene are more virulent than strains with functional copies [35], which may explain the widespread occurrence of null *pes3* alleles in the *Aspergillus* population.

Gene content polymorphisms

We additionally identified several SM gene clusters that gained or lost genes in some strains. These gene content polymorphisms were most likely generated through genomic deletion or insertion events and were often present in high frequencies in the population (Table 1). In three cases, these polymorphisms impact backbone synthesis genes, rendering the SM gene cluster non-functional. One example involves SM gene cluster 14, whose standard composition includes a pyoverdine synthase gene, an NRPS-like gene, an NRPS backbone gene, and several additional modification genes (Figure 2). We discovered that 4 / 66 strains lack an 11-kb region on the 3’ end of the cluster which normally contains an NRPS gene and two additional cluster genes, and the first non-SM genes on the 3’ end flanking the cluster. All *A.*

bioRxiv preprint doi: https://doi.org/10.1101/149856; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

fumigatus strains contain a *copia* family transposable element [36] at the 3' end of the cluster, suggesting that transposable elements may have been involved in the generation of this polymorphism. While this polymorphism could have arisen through a deletion event, a homologous cluster lacking the 11-kb region is also present in the reference genomes of *Aspergillus lentulus* and *Aspergillus fischerianus*, close relatives of *A. fumigatus* (Figure 2). The most parsimonious explanation is that the genome of the *A. fumigatus* ancestor contained an SM gene cluster that lacked the 11-kb region, and that this genomic region was subsequently gained and increased in frequency in the *A. fumigatus* population.

Two additional gene content polymorphisms affecting SM backbone genes were restricted to one strain each and appear to have arisen through genomic deletion events. Specifically, strain IF1SWF4 lacks an 8-Kb region near the helvolic acid SM gene cluster, resulting in the loss of the backbone oxidosqualene cyclase gene as well an upstream region containing two non-SM genes (Figure S1A). Strain LMB35Aa lacks a 54-kb region on the end of chromosome 2, which includes five genes from the telomere-proximal fumigaclavine C cluster (Figure S1B).

In three other cases, gene content polymorphisms involved gene loss or truncation events of non-backbone structural genes. We found that the second half of the ORF of the *gliM O*-methyltransferase gene in the gliotoxin gene cluster has been lost in 2 / 66 strains (Figure S1C), that the first half of the permease *fmqE* in the fumiquinazoline gene cluster has been lost in 4 / 66 strains (Figure S1D), and that an ABC transporter gene in SM cluster 21 has been almost entirely lost in 21 / 66 strains (Figure S1E-F).

Whole gene cluster gain and loss polymorphisms

Several SM gene clusters were gained or lost entirely in 13 / 66 strains. We observed instances where a cluster present in the genome of the reference Af293 strain was absent or pseudogenized in other strains as well as cases in which SM clusters present in other strains were absent from the reference Af293 strain.

bioRxiv preprint doi: https://doi.org/10.1101/149856; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The most notable example of an SM gene cluster that was present in the Af293 reference genome but absent or pseudogenized in others was SM cluster 4. This cluster contains 5 genes on the tip of the Af293 chromosome 1 and contains orthologs to five of the six genes in the fusarielin gene cluster in *Fusarium graminearum* [37]. This cluster is also present in several other *Aspergillus* species, including *A. clavatus* and *A. niger* [37]. Phylogenetic analysis of the genes in this SM gene cluster is consistent with horizontal gene transfer between fungi in the class Sordariomycetes and fungi in the class Eurotiomycetes, or alternatively with extensive gene loss in both Sordariomycetes and Eurotiomycetes (Figure S2). This gene cluster is entirely absent in 4 / 66 strains, and its genes are undergoing pseudogenization in an additional 44 strains via multiple independent mutational events (Figure 3A). Specifically, 19 strains shared a single frameshift variant in the polyketide synthase gene (4380_4381insAATGGGCT; frameshift at Glu1461 in Afu1g17740) and an additional 13 strains shared a single frameshift variant (242delG; frameshift at Gly81) in an aldose 1-epimerase gene (Afu1g17723). Twelve other strains each contained one to several frameshift or nonsense polymorphisms involving nine unique mutational sites, suggesting that this pathway is undergoing multiple independent pseudogenization events. Five of these strains contained multiple distinct frameshifts and premature stop codons in more than one gene in the cluster, indicating that the entire pathway is pseudogenized in these strains.

By searching for novel SM gene clusters in the genomes of the other 65 *A. fumigatus* strains, we found three SM gene clusters that were absent from the genome of the Af293 reference strain. As SM gene clusters are often present in repeat-rich and subtelomeric regions that are challenging to assemble [38,39], these strains might harbor additional novel SM gene clusters.

One of the novel SM gene clusters that we identified, cluster 34, was present in all but two of the strains (Af293 and F7763). This cluster contains a PKS backbone gene, one PKS-like gene with a single PKS associated domain, nine genes with putative biosynthetic functions involved in secondary metabolism, and six hypothetical proteins (Figure 3B). The two strains

bioRxiv preprint doi: https://doi.org/10.1101/149856; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

246 that lack this cluster contain a likely non-functional cluster fragment that includes the PKS-like
247 gene, two biosynthetic genes, and three hypothetical proteins. Interestingly, the 3' region
248 flanking this cluster is syntenic across all 66 strains but the 5' region is not, suggesting that a
249 recombination or deletion event may have resulted in the loss of this cluster in the Af293 and
250 F7763 strains.
251
252 The other SM gene clusters that were absent from the Af293 genome are present at
253 lower frequencies in the population; cluster 35 is present in 2 / 66 strains and cluster 36 in 4 /
254 66 strains. Cluster 35 is located in a region syntenic with an Af293 chromosome 4 region and is
255 flanked on both sides by transposable elements (Figure S3). Eight of the 14 genes in this SM
256 gene cluster are homologous to genes in an SM gene cluster in the genome of the insect
257 pathogenic fungus *Metarhizium anisopliae* (Figure S3). Phylogenetic analysis of these 8 genes is
258 consistent with a horizontal transfer event (Figure S4). Cluster 36 is an NRPS containing cluster
259 located on genomic scaffolds that lack homology to either the Af293 or A1163 genomes,
260 making it impossible to determine on which chromosome this cluster is located (Figure S3). The
261 evolutionary histories of the genes in the cluster are consistent with vertical inheritance and are
262 present in multiple *Aspergillus* species.
263
264 **Idiomorph polymorphisms**
265 One of the most peculiar types of polymorphisms that we identified is a locus containing
266 different unrelated alleles of SM gene clusters, reminiscent of the idiomorph alleles at the
267 fungal mating loci [40]. This locus, which resides on chromosome 3 and corresponds to cluster
268 10 in the Af293 genome (Figure 4), was previously described as being strain-specific in a
269 comparison between Af293 and A1163 [29] and is thought to reside in a recombination hot
270 spot [30]. Our analysis showed that this locus contained at least 6 different alleles present in
271 two or more of the 66 strains as well as 2 additional alleles that were each only present in one
272 strain (Figure S5).
273

bioRxiv preprint doi: https://doi.org/10.1101/149856; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In the Af293 reference genome, the cluster present at this locus contains one full-length PKS gene along with multiple genes that contain NRPS- or PKS-associated domains (Allele C). In the A1163 reference genome and 17 other strains, there is a full-length NRPS and a full-length PKS (Allele B). These alleles show an almost complete lack of sequence similarity except for a conserved hypothetical protein and a fragment of the full-length A1163 PKS in the Af293 allele; in contrast, the upstream and downstream flanking regions of the two alleles, which do not contain any backbone genes, are syntenic. Remarkably, another allele, present in 12 strains, contains all of the genes from both the Af293 and A1163 clusters (Allele D). The remaining three alleles contain various combinations of these genes. One allele found in 22 strains contains some A1163-specific genes and no Af293-specific genes (Allele A), while another allele found in 3 strains contains some Af293-specific genes but no A1163 genes (Allele F). The final allele, present in 8 strains, contains the entire Af293 allele as well as part of the A1163 allele (Allele E). Every allele is littered with long terminal repeat sequence fragments from *gypsy* and *copia* TE families as well as with sequence fragments from DNA transposons from the *mariner* family [36]. In some cases, these TEs correspond with breakpoints in synteny between alleles, suggesting that the diverse alleles of this SM gene cluster may arise via TE-driven recombination. Further, both of the alleles that are restricted to a single strain had an insertion event of several genes near a TE, though the rest of the locus is highly similar to one of the more common alleles (Figure S5). The evolutionary history of this highly diverse locus is unclear. While it is tempting to speculate that the largest allele containing all observed genes represents the ancestral state, it does not explain the presence of a shared hypothetical protein and PKS gene fragment between the Af293 locus (Allele C) and the A1163 locus (Allele B).

Genomic location polymorphisms

The final type of polymorphism that we observed is associated with SM gene clusters that are located in different genomic locations in different strains, suggesting that these SM gene clusters are behaving like mobile elements. This type of polymorphism was observed in SM gene clusters 1 and 33, both of which produce as yet identified products, and are present at low frequencies in the population.

bioRxiv preprint doi: https://doi.org/10.1101/149856; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

SM gene cluster 1, which is present in six strains at three different genomic locations (Figure 5A), consists of a PKS and four other modification genes that are always flanked by a 15 Kb region (upstream) and a 43 Kb region (downstream) containing TEs. In the reference Af293 strain and in strain F7763, this SM gene cluster and its flanking regions are located on chromosome 1, while in strains dutch3, F13619, and Z5 they are located between Afu4g07320 and Afu4g07340 on chromosome 4. In contrast, in strain JCM_10253, the cluster and flanking regions are located on chromosome 8 immediately adjacent to the 3' end of the intertwined fumagillin and pseurotin SM gene supercluster [41].

In 5 / 6 strains, the cluster appears to be functional and does not contain nonsense SNPs or indels. However, the cluster found on chromosome 1 in strain F7763 contains two stop codons in the oxidoreductase gene (Gln121* and Gln220*) and two premature stop codons in the polyketide synthase (Gln1156* and Gln1542*), suggesting this strain contains a null allele.

This "jumping" gene cluster is not present in any other sequenced *Aspergillus* genus, and phylogenetic analysis of its constituent genes is consistent with horizontal gene transfer between fungi (Figure S6). Specifically, this gene cluster is also present in *Phaeosphaeria nodorum,* a plant pathogen from the class Dothideomycetes, *Pseudogymnoascus pannorum*, a fungus isolated from permafrost from the Leotiomycetes, *Escovopsis weberi*, and a fungal parasite of fungus-growing ants from the Sordariomycetes (Figure 5B). One additional species, the endophyte *Hypoxylon* sp. CI4A from the class Sordariomycetes, contains four of the five cluster genes but is missing Afu1g00970, an MFS drug transporter. However, this species contains an unrelated gene annotated as an MFS drug transporter immediately adjacent to this cluster, so this species may be using a different transporter (Figure 5B). None of these fungi contain the upstream or downstream TE-rich flanking regions present in *A. fumigatus*, and each fungus contains additional unique genes with putative biosynthetic functions adjacent to the transferred cluster. The most likely explanation for this change in flanking regions is that this

bioRxiv preprint doi: https://doi.org/10.1101/149856; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

SM gene cluster was transferred into *A. fumigatus* once and has subsequently moved its location in the genome.

The second SM gene cluster that shows variation in its genomic location across strains, cluster 33, contains a terpene synthase. This cluster is present in only 5 strains at 3 distinct locations (Figure S7). Similar to cluster 1, this cluster is also flanked by TEs and in one strain the clusters is located 58 Kb from another SM gene cluster. In contrast to cluster 1, this cluster does not appear to have been horizontally transferred between fungi and its genes are present in other sequenced *Aspergillus* species. As this mobile cluster was not horizontally transferred, it is possible that the horizontal transfer of cluster 1 and its mobile nature throughout the *A. fumigatus* genome are driven by different mechanisms.

bioRxiv preprint doi: https://doi.org/10.1101/149856; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Discussion

Our examination of 66 genomes from strains of *Aspergillus fumigatus* revealed that five general types of polymorphisms describe variation in SM gene clusters in a fungal population. These polymorphisms include variation in single nucleotides, gene and gene cluster gains and losses, non-homologous clusters at the same genomic position, and changes in genomic locations of clusters (Figure 6). In several cases, the genetic mechanisms that gave rise to these polymorphisms and their functional consequences are clear. Polymorphisms in single-nucleotides most likely arose from errors in DNA replication or repair, while gene loss polymorphisms likely arose from genomic deletion or recombination events. Others, such as SM gene clusters that exist as non-homologous alleles or that entered the population through horizontal transfer, could arise through multiple genetic mechanisms. Transposable elements are common in polymorphic SM gene clusters and are found flanking mobile and horizontally transferred clusters, as well as in regions adjacent to non-homologous alleles and where gene gain has occurred, suggesting they contribute to SM gene cluster diversity. Using a population genomics approach to identify SM gene cluster variants allowed us to also capture and describe novel polymorphisms, including mobile gene clusters and idiomorphic clusters, which are difficult to identify in comparative genomic studies between species whose conservation of genome-wide synteny is low. Strikingly, the variants and genetic drivers we observe at the population level are also implicated as driving SM gene cluster variation between fungal species, suggesting that the observed microevolutionary processes are sufficient to explain macroevolutionary patterns of SM gene cluster evolution. Below, we discuss our key results and place them in the broader context of SM gene cluster evolution.

The first novel type of polymorphism was observed in SM gene clusters 1 and 33, which by occupying different genomic locations in different strains, appear to behave in a manner similar to mobile genetic elements. Interestingly, both clusters are located near or immediately adjacent to other SM gene clusters in some strains. For example, cluster 1 is located immediately adjacent to the intertwined fumagillin and pseurotin supercluster [41] in one strain. This supercluster is regulated by the transcriptional factor *fapR* and is located in a

bioRxiv preprint doi: https://doi.org/10.1101/149856; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

chromosomal region controlled by the master SM regulators *laeA* and *veA* [41,42], raising the hypothesis that these mobile gene clusters might be co-opting the regulatory machinery already in place. Previous work has hypothesized that the fumagillin and pseurotin supercluster formed through genomic rearrangement events placing the once-independent gene clusters in close proximity to each other [41]. Our observation that this mobile gene cluster is located in this same region not only supports this hypothesis but also implicates transposable elements as one of the mechanisms by which such genomic rearrangements are formed. These superclusters may also represent an intermediate stage in the formation of new SM gene clusters. Supercluster formation, potentially mediated by mobile gene clusters, and followed by gene loss, could explain macroevolutionary patterns of SM gene clusters have shown that clustered genes in one species can be dispersed over multiple gene clusters in other species [9,11].

The second novel type of polymorphism was the presence of multiple non-homologous alleles at the cluster 10 locus, echoing the structure of the idiomorphic mating type locus [40]. This region has previously been reported to be a recombination hotspot in *A. fumigatus* [30] and all alleles contain numerous transposable elements. Thus, it is possible that polymorphism at this locus originated via SM gene cluster fusion or splitting events driven by transposable elements. Interestingly, two other previously described instances of SM gene cluster variation bear close resemblance to the *A. fumigatus* idiomorphic SM gene cluster 10 locus. The first is the presence of two non-homologous *Aspergillus flavus* alleles, where some strains contain a 9-gene sesquiterpene-like SM gene cluster and others contain a non-homologous 6-gene SM gene cluster at the same genomic location [43]. The second is the presence of two non-homologous SM gene clusters at the same, well-conserved, locus in a comparison of six species of dermatophyte fungi [44]. Based on these results, we hypothesize that idiomorphic clusters may be common in fungal populations and contribute to the broad diversity of SM gene clusters across filamentous fungi.

bioRxiv preprint doi: https://doi.org/10.1101/149856; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The remaining types of SM polymorphism in the *A. fumigatus* population have previously been described at the species level. For example, null alleles have been reported for many individual SM gene clusters across diverse fungal species [10,17,26,27] and our results show that it is a widespread phenomenon, affecting two-thirds of the *A. fumigatus* SM gene clusters. Consistent with previous literature reporting the presence of additional SM gene clusters not present in the reference strain [24,25], we also identify four low-frequency SM gene clusters in the *A. fumigatus* population, two of which are horizontally transferred from distantly related fungi and two that appear vertically inherited. We find numerous cases of gene loss and relatively fewer cases of gene gain; previous studies have implicated these processes across numerous types of SM gene clusters and reported their effects on metabolite production [45]. Previous work has indicated that gene and gene cluster gain and loss tends to occur near telomeres, suggesting that higher rates of genetic events like recombination lead to loss [46]. While we do find several cases of gene and gene cluster loss in subtelomeric clusters, including the fumigaclavine gene cluster and the helvolic acid gene cluster, we also find gene loss in clusters that are not located near telomeres, such as the gliotoxin gene cluster and the fumiquinazoline gene cluster. These findings suggest that while gene and gene cluster loss can occur in telomeric regions, clusters in other genomic regions also experience high rates of loss.

Previous work has demonstrated that SM gene clusters, like the metabolites that they produce, are highly divergent between fungal species. Our examination of genome-wide variation shows that these SM gene clusters are correspondingly diverse within individual strains of a single fungal species. Furthermore, the observed genetic changes in SM gene clusters are widespread across different types of gene clusters and are caused by many underlying genetic drivers, including gene and gene cluster gain, loss, non-homologous cluster alleles, and mobile gene clusters. The net effect of these substitutions, gains, losses, and rearrangements raises the hypothesis that fungal SM gene clusters are likely in a state of evolutionary flux, constantly altering their SM gene cluster repertoire, and consequently modifying and diversifying their chemodiversity.

bioRxiv preprint doi: https://doi.org/10.1101/149856; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Methods

Strains analyzed

Eight strains of *A. fumigatus* were isolated from four patients with recurrent cases of aspergillosis in the Portuguese Oncology Institute in Porto, Portugal. Each strain was determined to be *A. fumigatus* using macroscopic features of the culture and microscopic morphology observed in the slide preparation from the colonies with lactophenol solution [47]. Based on the morphological characterization, all clinical strains were classified as *A. fumigatus complex*-Fumigati. The genomes of all eight strains were sequenced using 150bp Illumina paired-end sequence reads at the Genomic Services Lab of Hudson Alpha (Huntsville, Alabama, USA). Genomic libraries were constructed with the Illumina TruSeq library kit and sequenced on an Illumina HiSeq 2500 sequencer. Samples of all eight strains were sequenced at greater than 180X coverage or depth (Table S1). Short read sequences for these 8 strains are available in the Short Read Archive under accession SRP109032.

In addition to the 8 strains sequenced in this study, we retrieved 58 *A. fumigatus* strains with publicly available whole genome sequencing data, resulting in a population genomics dataset of 66 strains (Table S1). The strains used included both environmental and clinical strains and were isolated from multiple continents. Genome assemblies for 10 of these strains, including the Af293 and A1163 reference strains, were available for download from GenBank [28–32,48]. For 6 of these strains, short read sequences were also available from the NCBI Short Read Archive (SRA), which were used for variant discovery only (see Single nucleotide variant (SNV) and indel discovery) and not for genome assembly. Short read sequences were not available for the remaining 4 strains. Short read sequences were downloaded for an additional 48 strains from the Short Read Archive if they were sequenced with paired-end reads and at greater than 30x coverage.

Single nucleotide variant (SNV) and indel discovery

bioRxiv preprint doi: https://doi.org/10.1101/149856; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

All strains with available short read data (62 of 66 strains) were aligned to both the Af293 and A1163 reference genomes using BWA mem version 0.7.12-r1044 [49]. Coverage of genes present in the reference genome was calculated using bedtools v2.25.0 [50]. SNV and indel discovery and genotyping was performed relative to the Af293 reference genome and was conducted across all samples simultaneously using the Genome Analysis Toolkit version 3.5-0-g36282e4 with recommended hard filtering parameters [51–53] and annotated using snpEff version 4.2 [54].

De novo genome assembly and gene annotation

All 56 strains without publicly available genome assemblies were *de novo* assembled using the iWGS pipeline [55]. Specifically, all strains were assembled using SPAdes v3.6.2 and MaSuRCA v3.1.3 and resulting assemblies were evaluated using QUAST v3.2 [56–58]. The average N50 of assemblies constructed with this strategy was 463 KB (Table S1). Genes were annotated in these assemblies as well as in five GenBank assemblies with no predicted genes using augustus v3.2.2 trained on *A. fumigatus* gene models [59]. Repetitive elements were annotated in all assemblies using RepeatMasker version open-4.0.6 [60].

Secondary metabolic gene cluster annotation and discovery

Secondary metabolic gene clusters in the Af293 reference genome were taken from two recent reviews, both of which considered computational and experimental data to delineate cluster boundaries [61,62] (Table S3). The genomes of the other 65 strains were scanned for novel SM gene clusters using identified using antiSMASH v3.0.5.1 [63]. To prevent potential assembly errors from confounding the analysis, any inference about changes in genomic locations of genes or gene clusters was additionally verified by manually inspecting alignments and ensuring that paired end reads supported an alternative genomic location (see SNV and indel discovery). Cases where paired end reads did not support the change in genomic location or where mapping was ambiguous or low quality were discarded.

Phylogenetic analysis

bioRxiv preprint doi: https://doi.org/10.1101/149856; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To construct a SNP-based strain phylogeny, biallelic SNPs with no missing data were pruned using SNPRelate v1.8.0 with a linkage disequilibrium threshold of 0.8 [64]. A phylogeny was constructed using RAxML v8.0.25 using the ASC_BINGAMMA substitution model [65]. The tree was midpoint rooted and all branches with bootstrap support less than 80% were collapsed.

To understand the evolutionary histories of specific SM gene clusters showing unusual taxonomic distributions, we reconstructed the phylogenetic trees of their SM genes. Specifically, SM cluster protein sequences were queried against a local copy of the NCBI non-redundant protein database (downloaded May 30, 2017) using phmmer, a member of the HMMER3 software suite [66] using acceleration parameters --F1 1e-5 --F2 1e-7 --F3 1e-10. A custom perl script sorted the phmmer results based on the normalized bitscore (nbs), where nbs was calculated as the bitscore of the single best-scoring domain in the hit sequence divided by the best bitscore possible for the query sequence (i.e., the bitscore of the query aligned to itself). No more than five hits were retained for each unique NCBI Taxonomy ID. Full-length proteins corresponding to the top 100 hits (E-value < 1 × 10-10) to each query sequence were extracted from the local database using esl-sfetch [66]. Sequences were aligned with MAFFT v7.310 using the E-INS-i strategy and the BLOSUM30 amino acid scoring matrix [67] and trimmed with trimAL v1.4.rev15 using its gappyout strategy [68]. The topologies were inferred using maximum likelihood as implemented in RAxML v8.2.9 [65] using empirically determined substitution models and rapid bootstrapping (1000 replications). The phylogenies were midpoint rooted and branches with less than 80% bootstrap support were collapsed using the ape and phangorn R packages [69,70]. Phylogenies were visualized using ITOL version 3.0 [71].

bioRxiv preprint doi: https://doi.org/10.1101/149856; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

509

510 **Acknowledgements**

511 This work was supported in part by the National Science Foundation (DEB-1442113 to A.R.) and
512 by the Northern Portugal Regional Operational Programme (NORTE 2020), under the Portugal
513 2020 Partnership Agreement, through the European Regional Development Fund (FEDER)
514 (NORTE-01-0145-FEDER-000013 to F.R.). G.H.G. thanks the Fundação de Amparo à Pesquisa do
515 Estado de São Paulo (FAPESP) and Conselho Nacional de Desenvolvimento Científico e
516 Tecnológico (CNPq), both from Brazil, for funding and support. This work was conducted in part
517 using the resources of the Advanced Computing Center for Research and Education at
518 Vanderbilt University (http://www.accre.vanderbilt.edu/).

519

520 **Author contributions**

521 Conceptualization, A.R., G.H.G., A.L.L.; Methodology, A.L.L., J.H.W, C.L.; Investigation, A.L.L.;
522 Visualization, A.L.L., J.H.W; Resources, G.H.G., F.R., C.L.; Writing, A.L.L, A.R.

523

bioRxiv preprint doi: https://doi.org/10.1101/149856; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

524 **Table 1. Types and rates of SM gene cluster variants in *A. fumigatus* strains.** See also Figure S3
525 and S4.

526

Description	Phenotype	Drivers	Frequency at cluster level	Frequency at strain level	Previous reports
Single-nucleotide polymorphisms and indels	Potential for protein function change (missense); abrogation of protein function (nonsense and frameshift)	DNA replication errors; relaxation of purifying selection	100% (33/33 clusters; missense); 70% (23/33 clusters; nonsense and frameshift)	Every strain affected	Bikaverin in *Botrytis* [17,27], aflatoxin in *Aspergillus oryzae* and *Aspergillus flavus* [26], fumonisins in *Fusarium* [10], many others
Gene content polymorphisms	Loss of gene cluster function; structural changes in the metabolite; change in cluster expression or metabolite transport	Deletion and insertion events; recombination; transposable elements	6 clusters	27 / 66 strains	Trichothecene in *Fusarium*, aflatoxin and sterigmatocystin in *Aspergillus* [11–15]
Whole gene cluster polymorphisms	Loss or gain of novel metabolites	Deletion and insertion events; horizontal gene transfer; transposable elements	6 clusters	13 / 66 strains	Gibberellin and fumonisin in *Fusarium* [24,25]
fCluster idiomorphs	Changes in metabolites produced or structure of metabolites	Transposable elements; recombination; other mechanisms?	1 gene cluster	8 unique identified alleles	Putative SM gene clusters in dermatophytes; putative SM gene cluster in *Aspergillus flavus* and *Aspergillus oryzae* [43,44]
Mobile gene clusters	Potential for change in gene regulation	Transposable elements; horizontal gene transfer; other mechanisms?	2 gene clusters	8 / 66 strains	None

527
528

bioRxiv preprint doi: https://doi.org/10.1101/149856; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table S1. Summary of strains, sequence data, and assemblies used.

Table S2. All nonsynonymous variants in SM gene cluster genes.

Table S3. Description of reference Af293 SM gene clusters.

bioRxiv preprint doi: https://doi.org/10.1101/149856; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Vining LC. Functions of secondary metabolites. Annu Rev Microbiol. Annual Reviews 4139 El Camino Way, P.O. Box 10139, Palo Alto, CA 94303-0139, USA; 1990;44: 395–427. doi:10.1146/annurev.mi.44.100190.002143

2. Schimek C. Evolution of Special Metabolism in Fungi : Concepts , Mechanisms , and Pathways. In: Pöggler S, Wöstmeyer J, editors. Evolution of Fungi and Fungal-Like Organisms, The Mycota. XIV. Berlin, Heidelberg: Springer-Verlag; 2011. pp. 293–328.

3. Fox EM, Howlett BJ. Secondary metabolism: regulation and role in fungal biology. Curr Opin Microbiol. 2008;11: 481–7. doi:10.1016/j.mib.2008.10.007

4. Scharf DH, Heinekamp T, Brakhage AA. Human and Plant Fungal Pathogens: The Role of Secondary Metabolites. Heitman J, editor. PLoS Pathog. Public Library of Science; 2014;10: e1003859. doi:10.1371/journal.ppat.1003859

5. Yim G, Wang HH, Davies J. Antibiotics as signalling molecules. Philos Trans R Soc Lond B Biol Sci. 2007;362: 1195–200. doi:10.1098/rstb.2007.2044

6. Keller NP, Turner G, Bennett JW. Fungal secondary metabolism - from biochemistry to genomics. Nat Rev Microbiol. 2005;3: 937–47. doi:10.1038/nrmicro1286

7. Bennett J, Bentley R. What's in a name?—Microbial secondary metabolism. Adv Appl Microbiol. 1989;34.

8. Khaldi N, Seifuddin FT, Turner G, Haft D, Nierman WC, Wolfe KH, et al. SMURF: Genomic mapping of fungal secondary metabolite clusters. Fungal Genet Biol. 2010;47: 736–41. doi:10.1016/j.fgb.2010.06.003

9. Lind AL, Wisecaver JH, Smith TD, Feng X, Calvo AM, Rokas A. Examining the evolution of the regulatory circuit controlling secondary metabolism and development in the fungal genus Aspergillus. PLoS Genet. Public Library of Science; 2015;11: e1005096. doi:10.1371/journal.pgen.1005096

10. Proctor RH, Busman M, Seo J-A, Lee YW, Plattner RD. A fumonisin biosynthetic gene cluster in Fusarium oxysporum strain O-1890 and the genetic basis for B versus C fumonisin production. Fungal Genet Biol. 2008;45: 1016–1026. doi:10.1016/j.fgb.2008.02.004

11. Proctor RH, McCormick SP, Alexander NJ, Desjardins AE. Evidence that a secondary metabolic biosynthetic gene cluster has grown by gene relocation during evolution of the filamentous fungus Fusarium. Mol Microbiol. 2009;74: 1128–42. doi:10.1111/j.1365-2958.2009.06927.x

12. Berry DB, Guan Q, Hose J, Haroon S, Gebbia M, Heisler LE, et al. Multiple Means to the Same End: The Genetic Basis of Acquired Stress Resistance in Yeast. Copenhaver GP,

bioRxiv preprint doi: https://doi.org/10.1101/149856; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

editor. PLoS Genet. Public Library of Science; 2011;7: e1002353. doi:10.1371/journal.pgen.1002353

13. Slot JC, Rokas A. Horizontal transfer of a large and highly toxic secondary metabolic gene cluster between fungi. Curr Biol. Elsevier Ltd; 2011;21: 134–9. doi:10.1016/j.cub.2010.12.020

14. Ehrlich KC, Chang P-K, Yu J, Cotty PJ. Aflatoxin Biosynthesis Cluster Gene cypA Is Required for G Aflatoxin Formation. Appl Environ Microbiol. 2004;70: 6518–6524. doi:10.1128/AEM.70.11.6518-6524.2004

15. Carbone I, Ramirez-Prado JH, Jakobek JL, Horn BW. Gene duplication, modularity and adaptation in the evolution of the aflatoxin gene cluster. BMC Evol Biol. 2007;7: 111. doi:10.1186/1471-2148-7-111

16. Yu J, Chang P-K, Ehrlich KC, Cary JW, Bhatnagar D, Cleveland TE, et al. Clustered pathway genes in aflatoxin biosynthesis. Appl Environ Microbiol. American Society for Microbiology; 2004;70: 1253–62. doi:10.1128/AEM.70.3.1253-1262.2004

17. Campbell M a, Rokas A, Slot JC. Horizontal transfer and death of a fungal secondary metabolic gene cluster. Genome Biol Evol. 2012;4: 289–93. doi:10.1093/gbe/evs011

18. Kroken S, Glass NL, Taylor JW, Yoder OC, Turgeon BG. Phylogenomic analysis of type I polyketide synthase genes in pathogenic and saprobic ascomycetes. Proc Natl Acad Sci U S A. 2003;100: 15670–5. doi:10.1073/pnas.2532165100

19. Bushley KE, Turgeon BG. Phylogenomics reveals subfamilies of fungal nonribosomal peptide synthetases and their evolutionary relationships. BMC Evol Biol. 2010;10: 26. doi:10.1186/1471-2148-10-26

20. Patron NJ, Waller RF, Cozijnsen AJ, Straney DC, Gardiner DM, Nierman WC, et al. Origin and distribution of epipolythiodioxopiperazine (ETP) gene clusters in filamentous ascomycetes. BMC Evol Biol. 2007;7: 174. doi:10.1186/1471-2148-7-174

21. Khaldi N, Collemare J, Lebrun M-H, Wolfe KH. Evidence for horizontal transfer of a secondary metabolite gene cluster between fungi. Genome Biol. BioMed Central; 2008;9: R18. doi:10.1186/gb-2008-9-1-r18

22. Khaldi N, Wolfe KH. Evolutionary Origins of the Fumonisin Secondary Metabolite Gene Cluster in Fusarium verticillioides and Aspergillus niger. Int J Evol Biol. Hindawi Publishing Corporation; 2011;2011: 423821. doi:10.4061/2011/423821

23. Reynolds HT, Slot JC, Divon HH, Lysøe E, Proctor RH, Brown DW. Differential Retention of Gene Functions in a Secondary Metabolite Cluster. Mol Biol Evol. Oxford University Press (OUP); 10: e1004816. doi:10.1093/molbev/msx145

24. Wiemann P, Sieber CMK, von Bargen KW, Studt L, Niehaus E-M, Espino JJ, et al. Deciphering the Cryptic Genome: Genome-wide Analyses of the Rice Pathogen Fusarium

bioRxiv preprint doi: https://doi.org/10.1101/149856; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

fujikuroi Reveal Complex Regulation of Secondary Metabolism and Novel Metabolites. Mitchell AP, editor. PLoS Pathog. Public Library of Science; 2013;9: e1003475. doi:10.1371/journal.ppat.1003475

25. Chiara M, Fanelli F, Mulè G, Logrieco AF, Pesole G, Leslie JF, et al. Genome Sequencing of Multiple Isolates Highlights Subtelomeric Genomic Diversity within *Fusarium fujikuroi*. Genome Biol Evol. 2015;7: 3062–3069. doi:10.1093/gbe/evv198

26. Chang P-K, Horn BW, Dorner JW. Sequence breakpoints in the aflatoxin biosynthesis gene cluster and flanking regions in nonaflatoxigenic Aspergillus flavus isolates. Fungal Genet Biol. 2005;42: 914–923. doi:10.1016/j.fgb.2005.07.004

27. Schumacher J, Gautier A, Morgant G, Studt L, Ducrot P-H, Le Pêcheur P, et al. A Functional Bikaverin Biosynthesis Gene Cluster in Rare Strains of Botrytis cinerea Is Positively Controlled by VELVET. Yun S-H, editor. PLoS One. Public Library of Science; 2013;8: e53729. doi:10.1371/journal.pone.0053729

28. Nierman WC, Pain A, Anderson MJ, Wortman JR, Kim HS, Arroyo J, et al. Genomic sequence of the pathogenic and allergenic filamentous fungus Aspergillus fumigatus. Nature. 2005;438: 1151–6. doi:10.1038/nature04332

29. Fedorova ND, Khaldi N, Joardar VS, Maiti R, Amedeo P, Anderson MJ, et al. Genomic islands in the pathogenic filamentous fungus Aspergillus fumigatus. Richardson PM, editor. PLoS Genet. Public Library of Science; 2008;4: e1000046. doi:10.1371/journal.pgen.1000046

30. Abdolrasouli A, Rhodes J, Beale MA, Hagen F, Rogers TR, Chowdhary A, et al. Genomic Context of Azole Resistance Mutations in Aspergillus fumigatus Determined Using Whole-Genome Sequencing. MBio. 2015;6. doi:10.1128/mBio.00536-15

31. Knox BP, Blachowicz A, Palmer JM, Romsdahl J, Huttenlocher A, Wang CCC, et al. Characterization of Aspergillus fumigatus Isolates from Air and Surfaces of the International Space Station. mSphere. 2016;1.

32. Paul S, Zhang A, Ludeña Y, Villena GK, Yu F, Sherman DH, et al. Insights from the genome of a high alkaline cellulase producing Aspergillus fumigatus strain obtained from Peruvian Amazon rainforest. J Biotechnol. 2017;251: 53–58. doi:10.1016/j.jbiotec.2017.04.010

33. Roberts JD, Kunkel TA. Fidelity of DNA replication. In: DePamphilis ML, editor. DNA replication in eukaryotic cells. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press; 1996. pp. 217–247.

34. Throckmorton K, Lim FY, Kontoyiannis DP, Zheng W, Keller NP. Redundant synthesis of a conidial polyketide by two distinct secondary metabolite clusters in Aspergillus fumigatus. Environ Microbiol. 2015; doi:10.1111/1462-2920.13007

35. O'Hanlon KA, Cairns T, Stack D, Schrettl M, Bignell EM, Kavanagh K, et al. Targeted

bioRxiv preprint doi: https://doi.org/10.1101/149856; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

643 Disruption of Nonribosomal Peptide Synthetase pes3 Augments the Virulence of
644 Aspergillus fumigatus. Infect Immun. 2011;79: 3978–3992. doi:10.1128/IAI.00192-11

645 36. Kapitonov V V., Jurka J. A universal classification of eukaryotic transposable elements
646 implemented in Repbase. Nat Rev Genet. Nature Publishing Group; 2008;9: 411–412.
647 doi:10.1038/nrg2165-c1

648 37. Sørensen JL, Hansen FT, Sondergaard TE, Staerk D, Lee TV, Wimmer R, et al. Production
649 of novel fusarielins by ectopic activation of the polyketide synthase 9 cluster in Fusarium
650 graminearum. Environ Microbiol. Blackwell Publishing Ltd; 2012;14: 1159–1170.
651 doi:10.1111/j.1462-2920.2011.02696.x

652 38. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational
653 challenges and solutions. Nat Rev Genet. NIH Public Access; 2011;13: 36–46.
654 doi:10.1038/nrg3117

655 39. Palmer JM, Keller NP. Secondary metabolism in fungi: does chromosomal location
656 matter? Curr Opin Microbiol. 2010;13: 431–6. doi:10.1016/j.mib.2010.04.008

657 40. Metzenberg RL, Glass NL. Mating type and mating strategies inNeurospora. BioEssays.
658 Wiley Subscription Services, Inc., A Wiley Company; 1990;12: 53–59.
659 doi:10.1002/bies.950120202

660 41. Wiemann P, Guo C-J, Palmer JM, Sekonyela R, Wang CCC, Keller NP. Prototype of an
661 intertwined secondary-metabolite supercluster. Proc Natl Acad Sci U S A. 2013;110:
662 17065–70. doi:10.1073/pnas.1313258110

663 42. Lin H-C, Chooi Y-H, Dhingra S, Xu W, Calvo AM, Tang Y. The Fumagillin Biosynthetic Gene
664 Cluster in Aspergillus fumigatus Encodes a Cryptic Terpene Cyclase Involved in the
665 Formation of β-trans-Bergamotene. J Am Chem Soc. 2013;135: 4616–9.
666 doi:10.1021/ja312503y

667 43. Gibbons JG, Salichos L, Slot JC, Rinker DC, McGary KL, King JG, et al. The evolutionary
668 imprint of domestication on genome variation and function of the filamentous fungus
669 Aspergillus oryzae. Curr Biol. 2012;22: 1403–9. doi:10.1016/j.cub.2012.05.033

670 44. Zhang H, Rokas A, Slot JCJ. Two different secondary metabolism gene clusters occupied
671 the same ancestral locus in fungal dermatophytes of the Arthrodermataceae. PLoS One.
672 2012;7: e41903. doi:10.1371/journal.pone.0041903

673 45. Schardl CL, Young CA, Hesse U, Amyotte SG, Andreeva K, Calie PJ, et al. Plant-Symbiotic
674 Fungi as Chemical Engineers: Multi-Genome Analysis of the Clavicipitaceae Reveals
675 Dynamics of Alkaloid Loci. Heitman J, editor. PLoS Genet. Public Library of Science;
676 2013;9: e1003323. doi:10.1371/journal.pgen.1003323

677 46. Young CA, Schardl CL, Panaccione DG, Florea S, Takach JE, Charlton ND, et al. Genetics,
678 genomics and evolution of ergot alkaloid diversity. Toxins. 2015. pp. 1273–1302.

bioRxiv preprint doi: https://doi.org/10.1101/149856; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

679 doi:10.3390/toxins7041273

680 47. de Hoog G, Guarro J, Gené J, Figueras M. Atlas of Clinical Fungi. Washington, DC: ASM
681 Press; 2001.

682 48. Liu D, Zhang R, Yang X, Wu H, Xu D, Tang Z, et al. Thermostable cellulase production of
683 Aspergillus fumigatus Z5 under solid-state fermentation and its application in
684 degradation of agricultural wastes. Int Biodeterior Biodegradation. 2011;65: 717–725.
685 doi:10.1016/j.ibiod.2011.04.005

686 49. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform.
687 Bioinformatics. 2009;25: 1754–1760. doi:10.1093/bioinformatics/btp324

688 50. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic
689 features. Bioinformatics. Oxford University Press; 2010;26: 841–2.
690 doi:10.1093/bioinformatics/btq033

691 51. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome
692 Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing
693 data. Genome Res. Cold Spring Harbor Laboratory Press; 2010;20: 1297–303.
694 doi:10.1101/gr.107524.110

695 52. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et
696 al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best
697 Practices Pipeline. Current Protocols in Bioinformatics. Hoboken, NJ, USA: John Wiley &
698 Sons, Inc.; 2013. p. 11.10.1-11.10.33. doi:10.1002/0471250953.bi1110s43

699 53. DePristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, et al. A framework for
700 variation discovery and genotyping using next-generation DNA sequencing data. Nat
701 Genet. 2011;43: 491–498. doi:10.1038/ng.806

702 54. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for
703 annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in
704 the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). Taylor &
705 Francis; 2012;6: 80–92. doi:10.4161/fly.19695

706 55. Zhou X, Peris D, Kominek J, Kurtzman CP, Hittinger CT, Rokas A. in silico Whole Genome
707 Sequencer & Analyzer (iWGS): A Computational Pipeline to Guide the Design and
708 Analysis of de novo Genome Sequencing Studies. G3:
709 Genes|Genomes|Genetics. 2016; doi:10.1534/g3.116.034249

710 56. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new
711 genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol.
712 Mary Ann Liebert, Inc.; 2012;19: 455–77. doi:10.1089/cmb.2012.0021

713 57. Zimin A V., Marcais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. The MaSuRCA genome
714 assembler. Bioinformatics. 2013;29: 2669–2677. doi:10.1093/bioinformatics/btt476

bioRxiv preprint doi: https://doi.org/10.1101/149856; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

58. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29: 1072–1075. doi:10.1093/bioinformatics/btt086

59. Stanke M, Morgenstern B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. Oxford University Press; 2005;33: W465-7. doi:10.1093/nar/gki458

60. Smit A, Hubley R, Green P. Repeatmasker Open-4.0 [Internet]. Available: http://www.repeatmasker.org

61. Inglis DO, Binkley J, Skrzypek MS, Arnaud MB, Cerqueira GC, Shah P, et al. Comprehensive annotation of secondary metabolite biosynthetic genes and gene clusters of Aspergillus nidulans, A. fumigatus, A. niger and A. oryzae. BMC Microbiol. 2013;13: 91. doi:10.1186/1471-2180-13-91

62. Bignell E, Cairns TC, Throckmorton K, Nierman WC, Keller NP. Secondary metabolite arsenal of an opportunistic pathogenic fungus. Philos Trans R Soc B Biol Sci. 2016;371.

63. Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach M a, et al. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res. 2011;39: W339-46. doi:10.1093/nar/gkr466

64. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics. Oxford University Press; 2012;28: 3326–3328. doi:10.1093/bioinformatics/bts606

65. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30: 1312–1313. doi:10.1093/bioinformatics/btu033

66. Eddy SR. A new generation of homology search tools based on probabilistic inference. Genome Inform. 2009;23: 205–11.

67. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol. 2013;30: 772–780. doi:10.1093/molbev/mst010

68. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25: 1972–1973. doi:10.1093/bioinformatics/btp348

69. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics. Oxford University Press; 2004;20: 289–290. doi:10.1093/bioinformatics/btg412

70. Schliep KP. phangorn: phylogenetic analysis in R. Bioinformatics. Oxford University Press; 2011;27: 592–3. doi:10.1093/bioinformatics/btq706

bioRxiv preprint doi: https://doi.org/10.1101/149856; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

751 71. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and
752 annotation of phylogenetic and other trees. Nucleic Acids Res. 2016;44: W242–W245.
753 doi:10.1093/nar/gkw290

754

bioRxiv preprint doi: https://doi.org/10.1101/149856; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

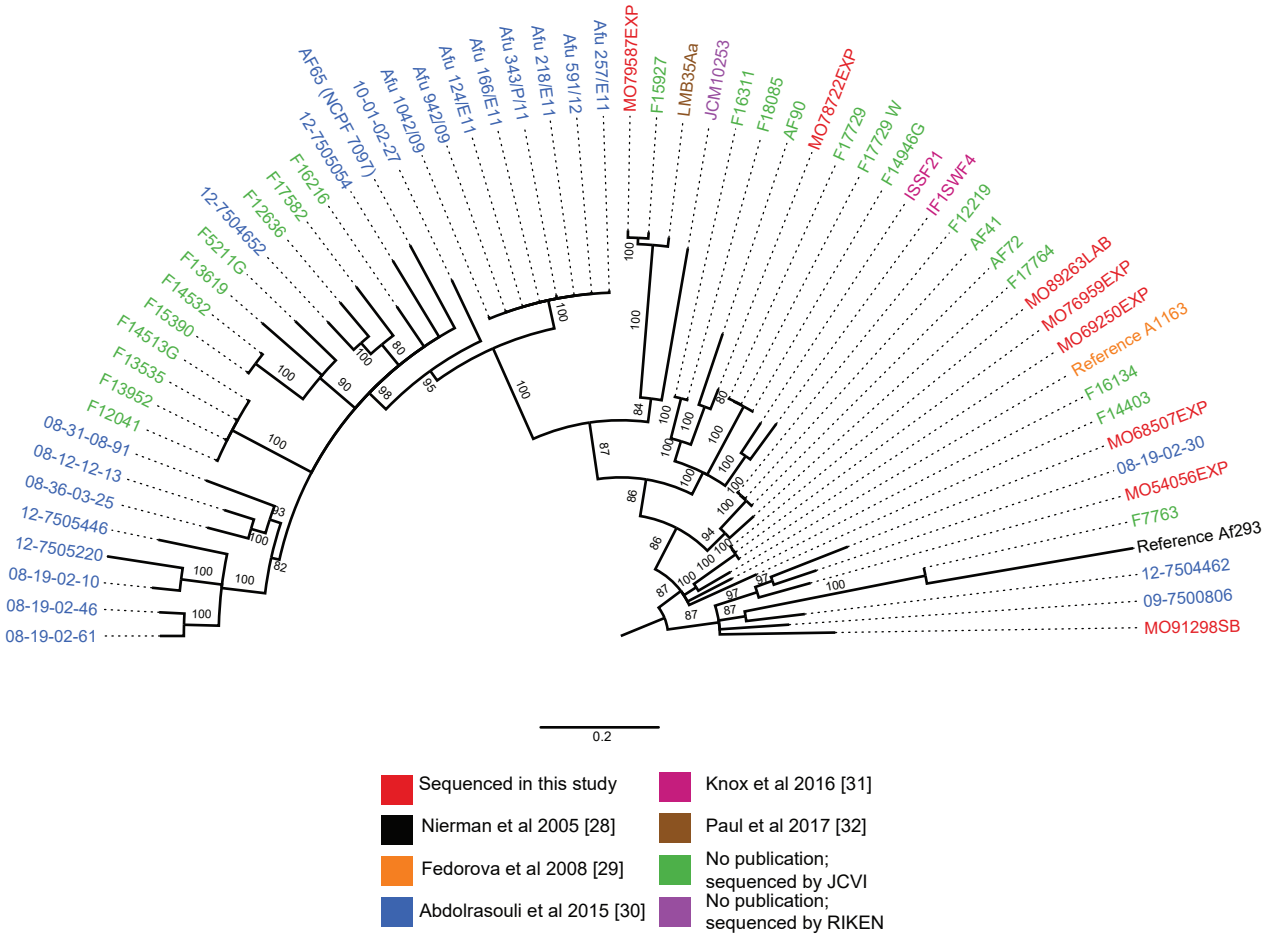


Figure 1. SNP-based phylogeny of *A. fumigatus* strains. The phylogeny was constructed using biallelic SNPs with no missing data. The tree is midpoint rooted and all branches with bootstrap support less than 80% are collapsed.

bioRxiv preprint doi: https://doi.org/10.1101/149856; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

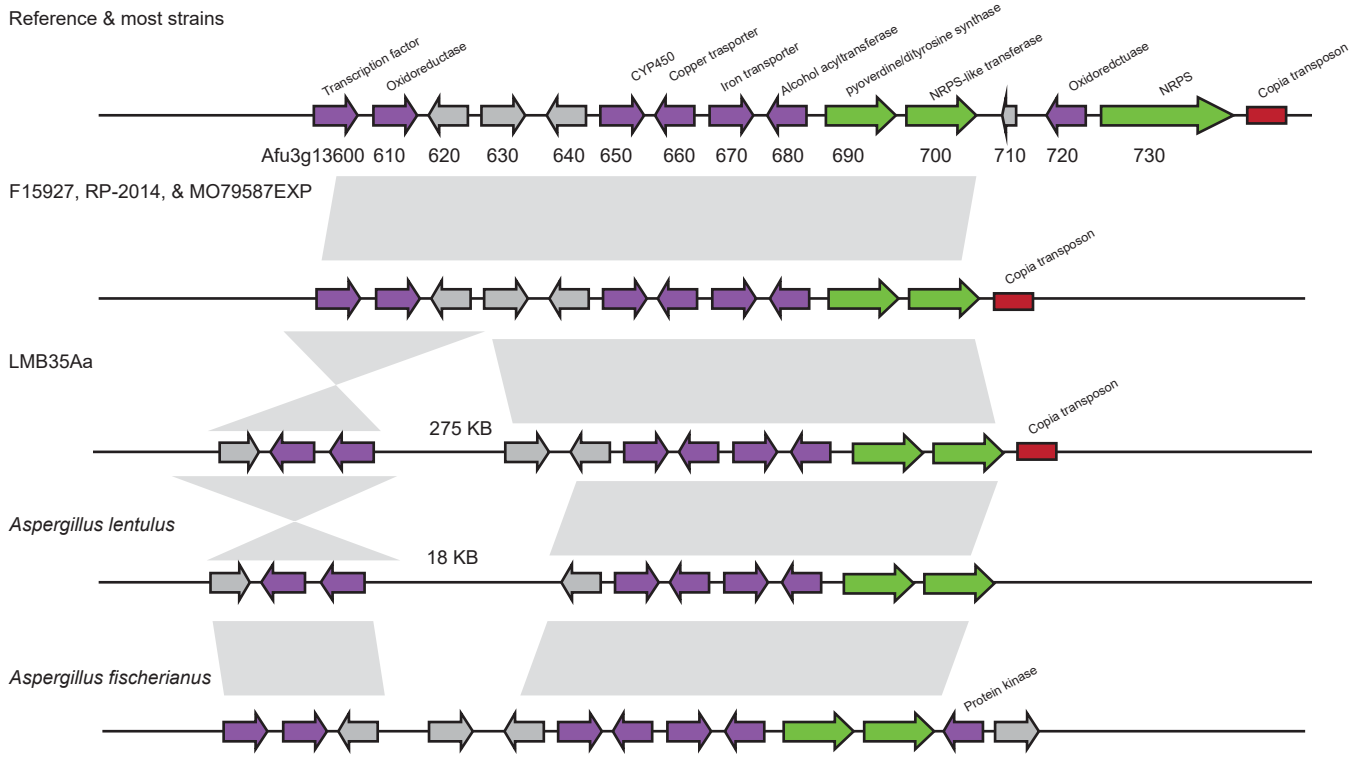
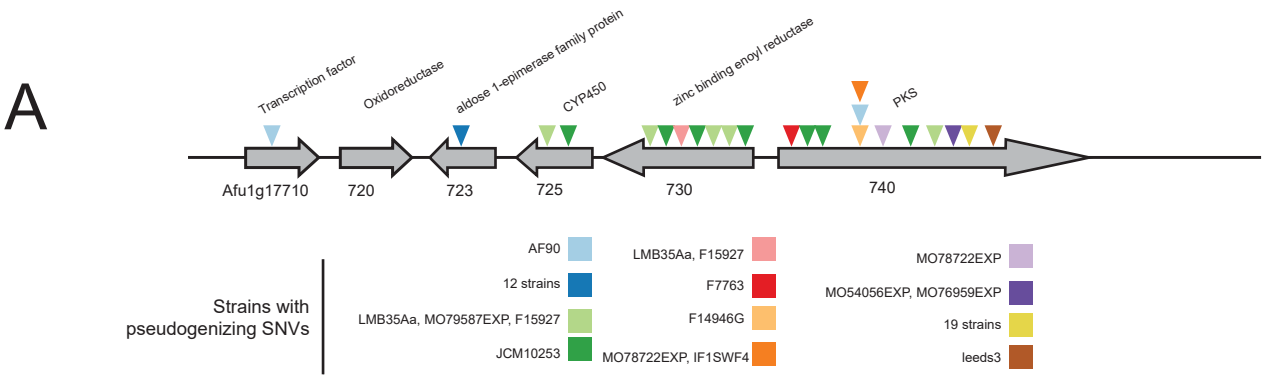


Figure 2. Differences in gene content in SM gene cluster 14 in A. fumigatus strains and closely related species. Four *A. fumigatus* strains lack an 11-Kb region in this cluster, including an NRPS backbone gene. Regions upstream and downstream of this cluster are syntenic. LMB35Aa also contains a large inversion that moves a transcription factor, oxidoreductase, and hypothetical protein 275 kb away from the cluster. *Aspergillus fischerianus* and *Aspergillus lentulus*, close relatives of *A. fumigatus*, contain a cluster lacking the 11-kb region.

bioRxiv preprint doi: https://doi.org/10.1101/149856; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



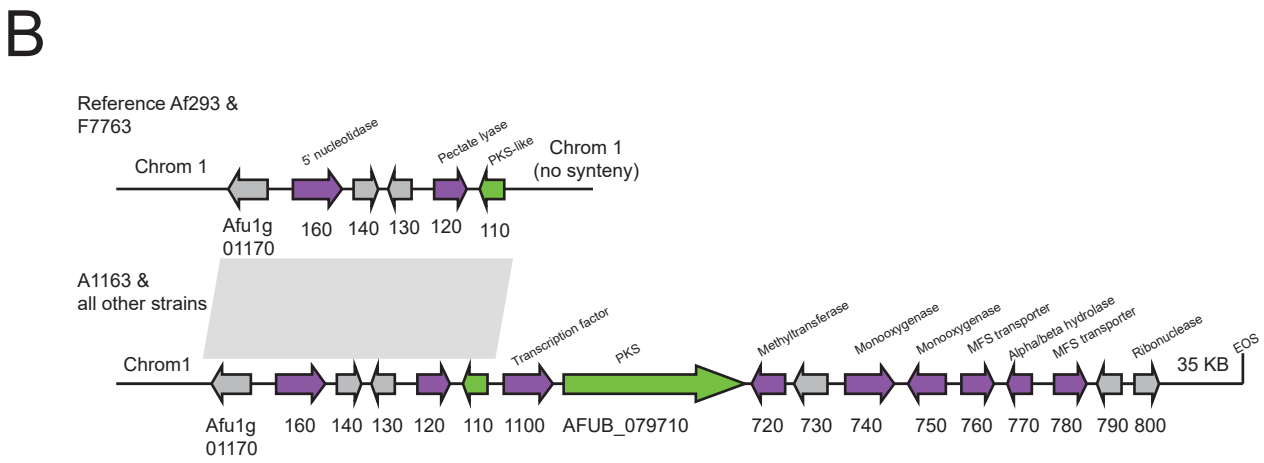


Figure 3. Pseudogenization and gene loss in SM gene clusters. A) SM gene cluster found in most *A. fumigatus* strains but absent from the Af293 reference and from the F7763 strain. EOS denotes end of scaffold. B) Positions of frameshift variants and nonsense variants in the fusarielin-like SM gene cluster 4.

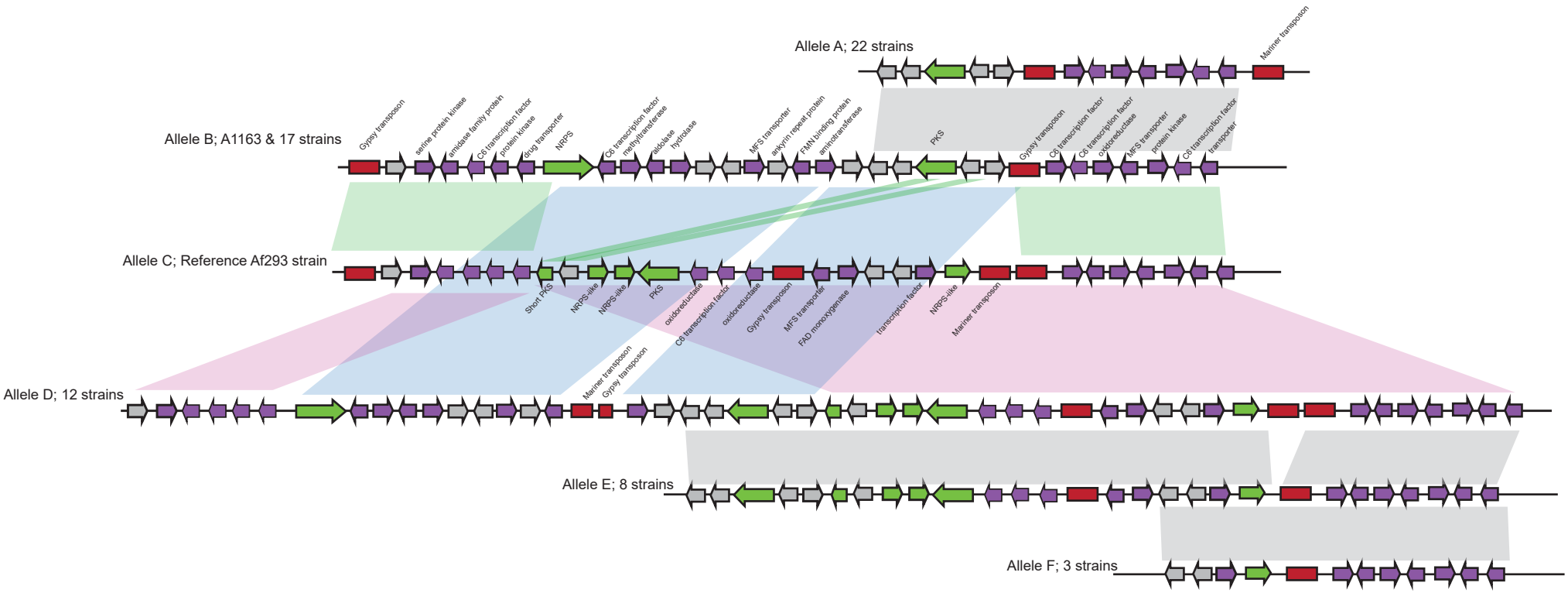


Figure 4. Six alleles of an idiomorphic SM gene cluster. Alleles of SM gene cluster 10 on chromosome 3. Red boxes denote transposable elements. Green arrows denote backbone genes (PKS or NRPS) or genes containing domains associated with PKS or NRPS genes. Purple arrows denote genes likely involved in SM biosynthesis, transport, or cluster regulation. Gray arrows denote genes without a putative function or with functions unrelated to SM.

bioRxiv preprint doi: https://doi.org/10.1101/149856; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

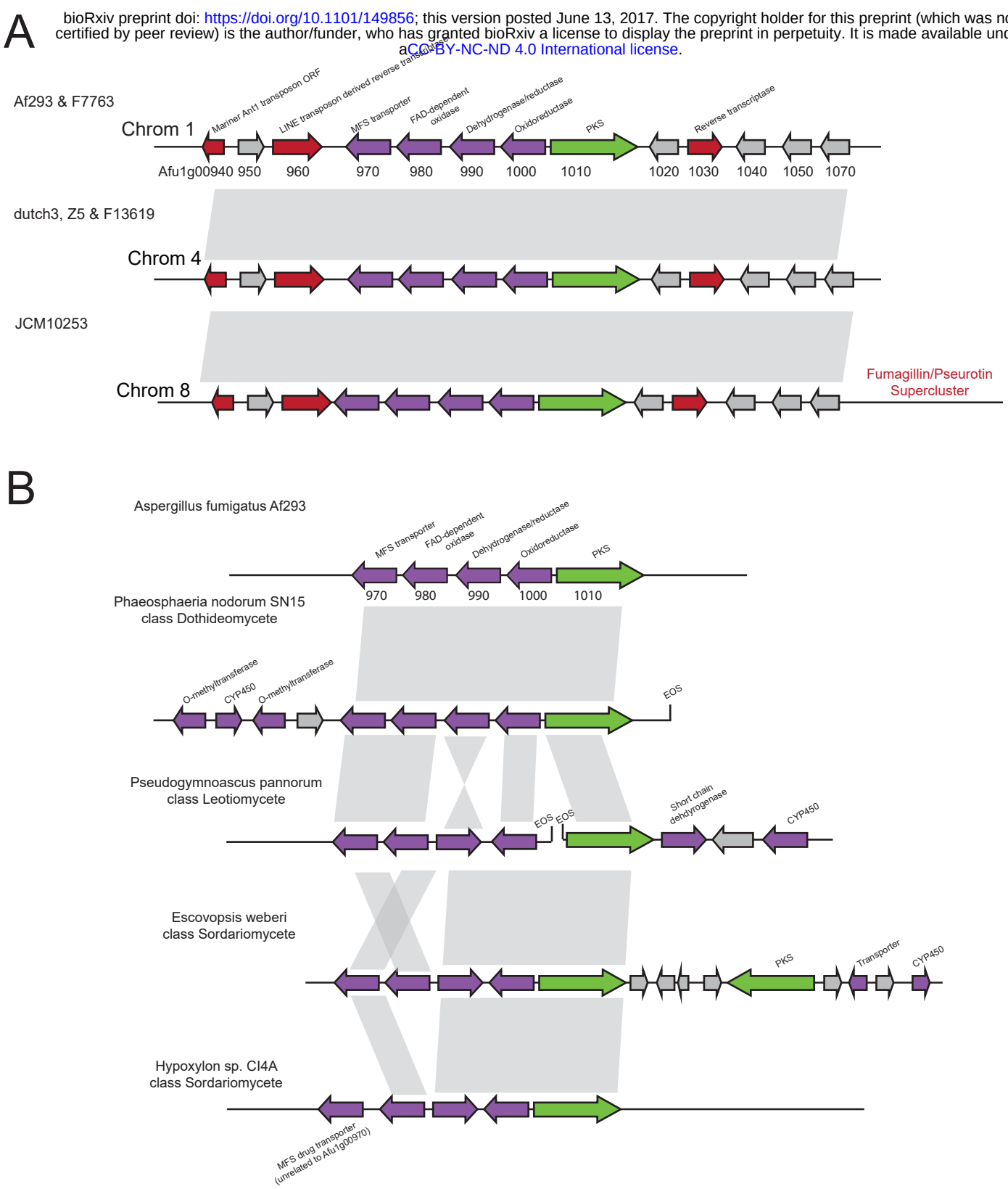


Figure 5. Multiple genomic locations of a horizontally transferred SM gene cluster.
A) Genomic location of SM gene cluster 1 (Afu1g00970-01010) and flanking region in all strains. This cluster is on chromosome 1 in two strains, chromosome 4 in three strains, and adjacent to the intertwined fumagillin and pseurotin supercluster on chromosome 8 in one strain. The flanking regions contain transposon-derived open reading frames including two putative reverse transcriptases. B) Synteny of A. fumigatus SM gene cluster 1 with clusters in *Phaeosphaeria nodorum*, *Pseudogymnoascus pannorum*, *Escovopsis weberi*, and *Hypoxylon sp. CI4A*. EOS denotes end of scaffold. All species contain non-syntenic genes predicted by antiSMASH to be part of a biosynthetic gene cluster.

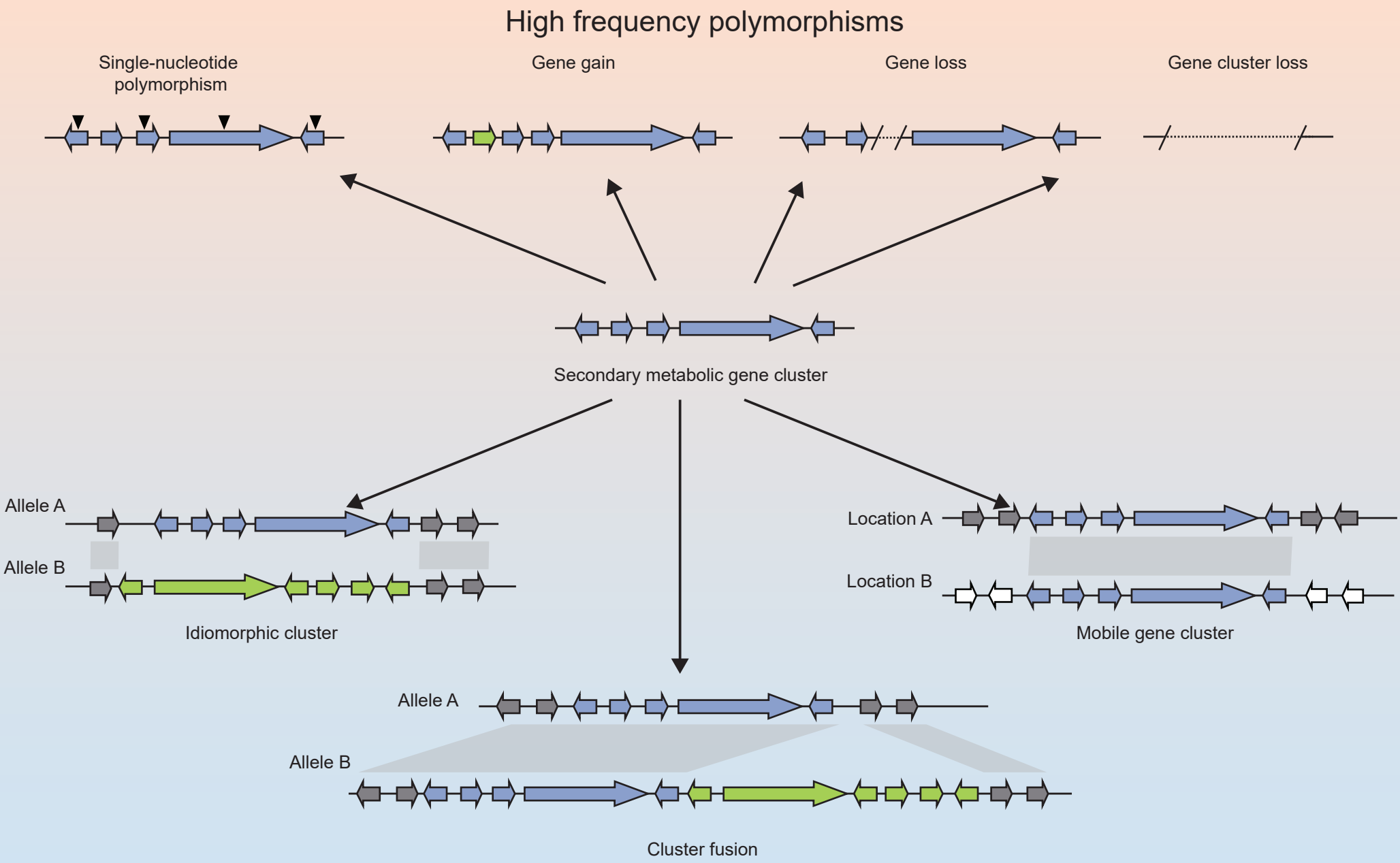


Figure 6. Types and frequencies of all SM gene cluster variants in the A. fumigatus population.