The birth, evolution and death of metabolic gene clusters in fungi

Antonis Rokas[1,2]*, *Jennifer H. Wisecaver*[1,3] *and Abigail L. Lind*[2,4]

Abstract | Fungi contain a remarkable diversity of both primary and secondary metabolic pathways involved in ecologically specialized or accessory functions. Genes in these pathways are frequently physically linked on fungal chromosomes, forming metabolic gene clusters (MGCs). In this Review, we describe the diversity in the structure and content of fungal MGCs, their population-level and species-level variation, the evolutionary mechanisms that underlie their formation, maintenance and decay, and their ecological and evolutionary impact on fungal populations. We also discuss MGCs from other eukaryotes and the reasons for their preponderance in fungi. Improved knowledge of the evolutionary life cycle of MGCs will advance our understanding of the ecology of specialized metabolism and of the interplay between the lifestyle of an organism and genome architecture.

Primary metabolism
The part of metabolism involving pathways associated with growth, such as those for macronutrients (for example, carbohydrates, fat and proteins).

Secondary metabolism
(Or specialized metabolism). The part of metabolism involving pathways associated with the production of small, bioactive molecules, such as mycotoxins, pigments and antibiotics.

[1]*Department of Biological Sciences, Vanderbilt University, Nashville, TN, USA.*

[2]*Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN, USA.*

[3]*Department of Biochemistry, Purdue University, West Lafayette, IN, USA.*

[4]*Gladstone Institutes, San Francisco, CA, USA.*

**e-mail: antonis.rokas@vanderbilt.edu*

https://doi.org/10.1038/s41579-018-0075-3

Fungi secrete enzymes to externally break down their surrounding food sources and absorb the products of this catabolic activity from the medium[1]. Extracellular digestion poses a considerable challenge; the food, which the fungal colony has spent substantial energy digesting, is accessible not only to the colony but also to all other nearby organisms[2]. To meet this challenge, fungi have evolved to consume a wide variety of foods, which is reflected in their diverse primary metabolism; they have also evolved a remarkable diversity of 'chemical weapons' to protect their food from competitors, which is reflected in their secondary metabolism (BOX 1). In addition, many fungi are pathogens of plants or animals, and secondary metabolites in particular have key roles as signals and toxins in these interactions[3–5].

Fungal primary and secondary metabolic pathways, especially those with ecologically specialized or accessory functions, are often encoded by metabolic gene clusters (MGCs)[5,6]. Similar to typical eukaryotic genes, the vast majority of fungal genes in MGCs are independently transcribed and not organized into operons, although exceptions exist[7]. MGCs participate in diverse functions related to primary metabolism, including the utilization of compounds (for example, nitrate[8,9] and allantoin[10]), sugars (for example, galactose[11]), vitamins (for example, biotin[12]) and amino acids (for example, tyrosine[13] and proline[14]); the degradation of xenobiotics (for example, arsenic[15] and cyanate[16]); and the production of various secondary metabolites, such as toxins (for example, aflatoxins[17], trichothecenes[18] and ergot alkaloids[19]), antibiotics (for example, penicillin[20]), drugs (for example, lovastatin[21]), pigments (for example, melanin[22]) and others (for example, kojic acid[23]) (FIG. 1).

Knowledge of primary and secondary MGCs far predates fungal genomics[6]; however, the close examination of the fungal DNA record enabled by genome sequencing has unearthed a much larger number of MGCs than anticipated (FIG. 2). This is especially true for MGCs involved in secondary metabolism; for example, the 498 genes in the 70 secondary MGCs[24] in *Aspergillus nidulans* (Ascomycota, Eurotiomycetes) account for 4.7% of the protein-coding genes of the organism. As only approximately a dozen secondary metabolites are known from this organism[25], there is optimism that functional characterization of the full complement of secondary MGCs in fungal genomes will lead to the discovery of a whole suite of novel metabolites[5,26].

The fact that natural selection has favoured and maintained the clustering of fungal metabolic pathways in certain fungi is surprising, not only because some of these pathways are often unclustered or partially clustered in other fungi[11,27,28] as well as in other eukaryotes[29,30] but also because comparative genomic analyses have revealed extensive structural rearrangements across fungal genomes[31]. In this Review, we explore the remarkable diversity in the organization of fungal MGCs; their variation within fungal populations and between fungal species; the mechanisms that underlie MGC formation, maintenance and decay in fungal populations; and how this organization of fungal metabolic pathways into MGCs facilitates adaptation to changing environments through the acquisition and loss of metabolic capacities. We also summarize what is known about MGCs involved in specialized metabolism in other eukaryotes and raise the question of why MGCs are less common or absent in other lineages.

Box 1 | **Metabolic gene clusters hold major clues to fungal ecology**

Examination of dentition can yield substantial insights into the feeding ecology of different mammals; for example, sharp, large incisors and pointed premolars and molars are characteristic of carnivorous mammals, whereas the absence of incisors and the presence of broad, ridged premolars and molars are characteristic of herbivores. In a similar manner, examination of the presence or absence patterns of metabolic pathways (including metabolic gene clusters (MGCs)) involved in primary metabolism in fungal genomes can shed light on the feeding ecology of different fungi. For example, the presence of an MGC comprising *gal1*, *gal7* and *gal10* is characteristic of species able to assimilate galactose, whereas the presence of a nitrate MGC denotes fungi that can assimilate nitrate.

The presence of claws, spines and other morphological characteristics can yield substantial insights into mammalian adaptations associated with interspecific interactions, such as predation, defence and territory establishment. For example, the sharp claws of tigers represent an adaptation to a predatory lifestyle, whereas the spines of hedgehogs are used for defence. In an analogous manner, examination of the presence or absence patterns of secondary metabolic pathways (including MGCs) can shed light on fungal community ecology. For example, the presence of the MGC for penicillin biosynthesis, a secondary metabolite with antibacterial activity, can be deduced to aid fungi in interspecific competition with bacteria, whereas the presence of the two MGCs for T-toxin production is associated with fungi that cause plant disease.

More generally, whereas all the mammalian traits discussed above represent morphological phenotypes (for example, phenotypes that are the product of development), their fungal trait counterparts represent biochemical phenotypes, arguing that, whereas much of animal evolution occurs in the context of development, fungal evolution occurs in the context of biochemistry.

Gene clusters

Cluster organization. The typical configuration of MGCs involved in primary metabolism includes genes encoding enzymes, transporters and transcription factors. For example, the five-gene MGC involved in nitrate assimilation in *Ogataea polymorpha* (Ascomycota, Saccharomycetes) contains two enzymes, two transcription factors and a transporter[32,33] (FIG. 1a). Similarly, the four-gene sugar utilization MGC[34] found in *Aspergillus parasiticus* and the four-gene proline catabolism MGC[14] in *A. nidulans* are both composed of a transporter, a regulatory protein and two structural genes encoding enzymes.

The standard configuration of MGCs involved in secondary metabolism includes a backbone gene whose protein product synthesizes the secondary metabolite precursor, such as a nonribosomal peptide synthetase (NRPS), a polyketide synthase (PKS), a dimethylallyl tryptophan synthetase (DMATS) or a terpene cyclase (TC), genes for one or more tailoring enzymes that chemically modify secondary metabolite precursors, transporter genes responsible for exporting the final product and transcription factors involved in cluster regulation. For example, the MGC for the synthesis of the mycotoxin gliotoxin in the opportunistic human pathogen *Aspergillus fumigatus* contains 13 genes, including genes encoding an NRPS (*gliP*), multiple tailoring structural genes (*gliI*, *gliJ*, *gliC*, *gliM*, *gliG*, *gliN* and *gliF*), a transporter (*gliA*), a transcription factor (*gliZ*) and a gliotoxin oxidase (*gliT*) that protects the fungus from its own gliotoxin[35] (FIG. 1a). Similarly, the MGC for the biosynthesis of the red pigment bikaverin in the rice pathogen *Fusarium fujikuroi* (Ascomycota, Sordariomycetes) includes a gene encoding a PKS (*bik1*) and two genes (*bik2* and *bik3*) encoding tailoring enzymes, as well as genes encoding a general transcriptional activator (*bik4*), a pathway-specific transcriptional activator (*bik5*) and a transporter (*bik6*)[36].

Diversity in gene content and structure. Although the paradigm is that primary and secondary MGCs contain all dedicated pathway genes, both gene content and structure can vary considerably across pathways (FIG. 1b).

Most observed diversity between primary MGCs concerns whether transcription factors and transporters are part of the MGC. For example, in contrast to the standard configurations of the nitrate assimilation MGC in *O. polymorpha* and the proline catabolism MGC in *A. nidulans* mentioned above, both of which encompass transporters, transcription factors and structural genes, the galactose MGC in *Saccharomyces cerevisiae* contains only the three structural genes responsible for the conversion of galactose to glucose-1-phosphate (*gal1*, *gal7* and *gal10*)[11,37,38]. The regulators of this pathway (*gal3*, *gal4* and *gal80*) as well as the gene encoding the galactose permease (*gal2*) all reside in distinct chromosomal locations, away from the MGC[37] (FIG. 1b). Other primary metabolic pathways consist of more than one MGC; this is the case for the six-gene biotin[12] pathway present in certain *S. cerevisiae* isolates, which consists of a two-gene MGC on chromosome 9, a three-gene MGC on chromosome 14 and one additional gene on chromosome 7 (FIG. 1).

Secondary MGCs also exhibit diversity in their gene content and structure. Some secondary MGCs deviate from the standard configuration and contain only enzymes, such as the endocrocin[39] MGC in *A. fumigatus*, the phenylalanine[13] and asperthecin[40] MGCs in *A. nidulans*, the penicillin[20] MGC in *Penicillium chrysogenum* (Ascomycota, Eurotiomycetes) and the melanin[22] MGC in *Alternaria alternata* (Ascomycota, Dothideomycetes). On the basis of the backbone gene in the MGC, secondary metabolites can be broadly classified into non-ribosomally synthesized peptides (encoded by NRPS genes), polyketides (by PKS), indole alkaloids (by DMATS) and terpenes (by TC)[5]. However, some MGCs contain multiple backbone genes; this is the case for the emericellamide MGC[41] in *A. nidulans* and for the cyclosporine MGC[42] in *Tolypocladium inflatum* (Ascomycota, Sordariomycetes), which contain both a PKS and an NRPS, the T-toxin MGC[43] in *Cochliobolus heterostrophus* (Ascomycota, Dothideomycetes), which contains two PKSs, and for the hexadehydroastechrome MGC[44] in *A. fumigatus*, which contains an NRPS and a DMATS. By contrast, other secondary MGCs lack backbone genes. For example, biosynthesis of the cyclic peptide ustiloxin B in *Aspergillus flavus* is not mediated via an NRPS but through a ribosomal peptide synthetic pathway encoded by a 16-gene MGC[45]. Similarly, the three-gene MGC involved in the biosynthesis of kojic acid in the domesticated fungus *Aspergillus oryzae* contains only a transcription factor, an oxireductase and a transporter[23] but none of the backbone genes typically found in secondary MGCs.

Secondary metabolic pathways, similar to primary metabolic pathways, can also consist of more than one MGC. For example, the cephalosporin biosynthetic pathway in *Acremonium chrysogenum* (Ascomycota,

Metabolic gene clusters (MGCs) A set of genes from the same metabolic pathway that is physically linked and occupies the same genetic locus in the chromosome; in other organisms, similarly organized co-adapted gene complexes are associated with non-metabolic traits and have come to be known as supergenes.

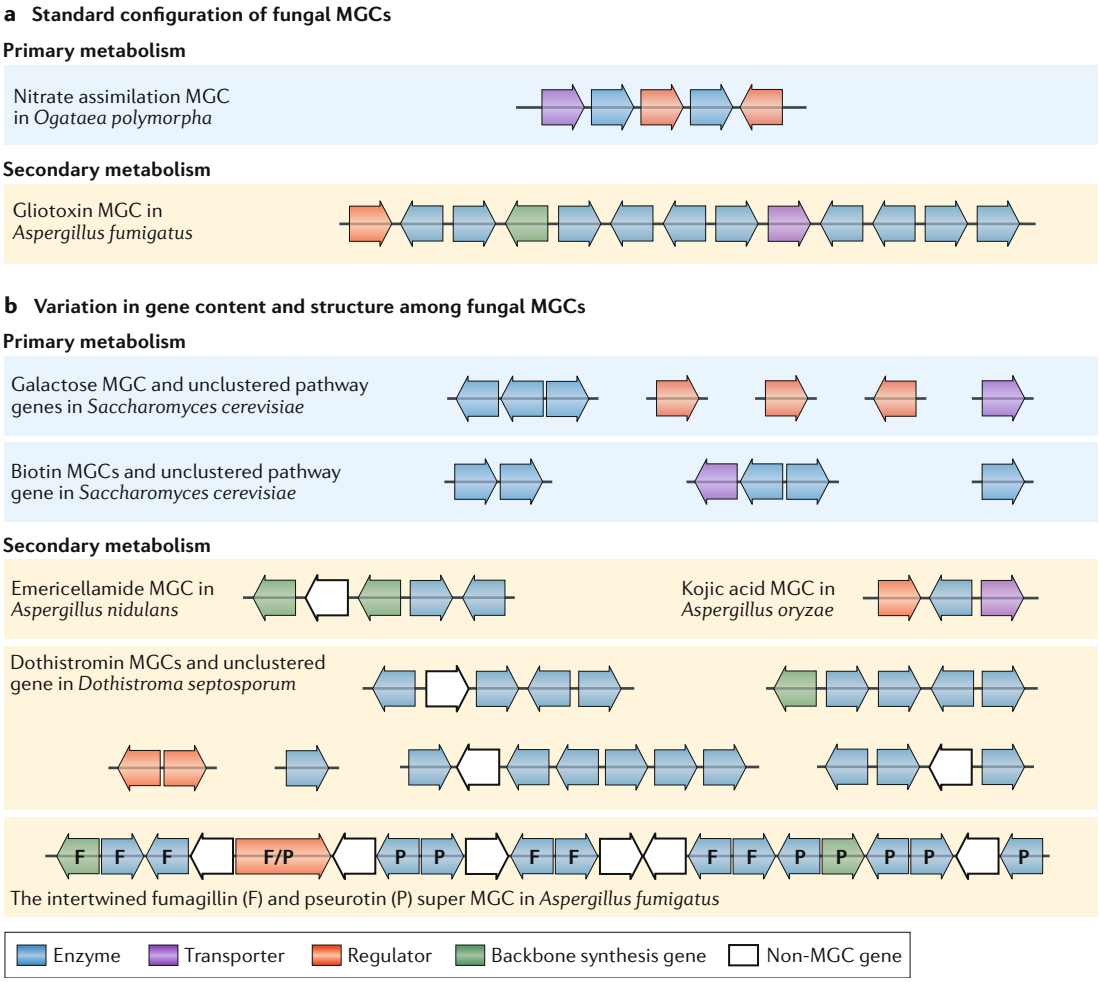
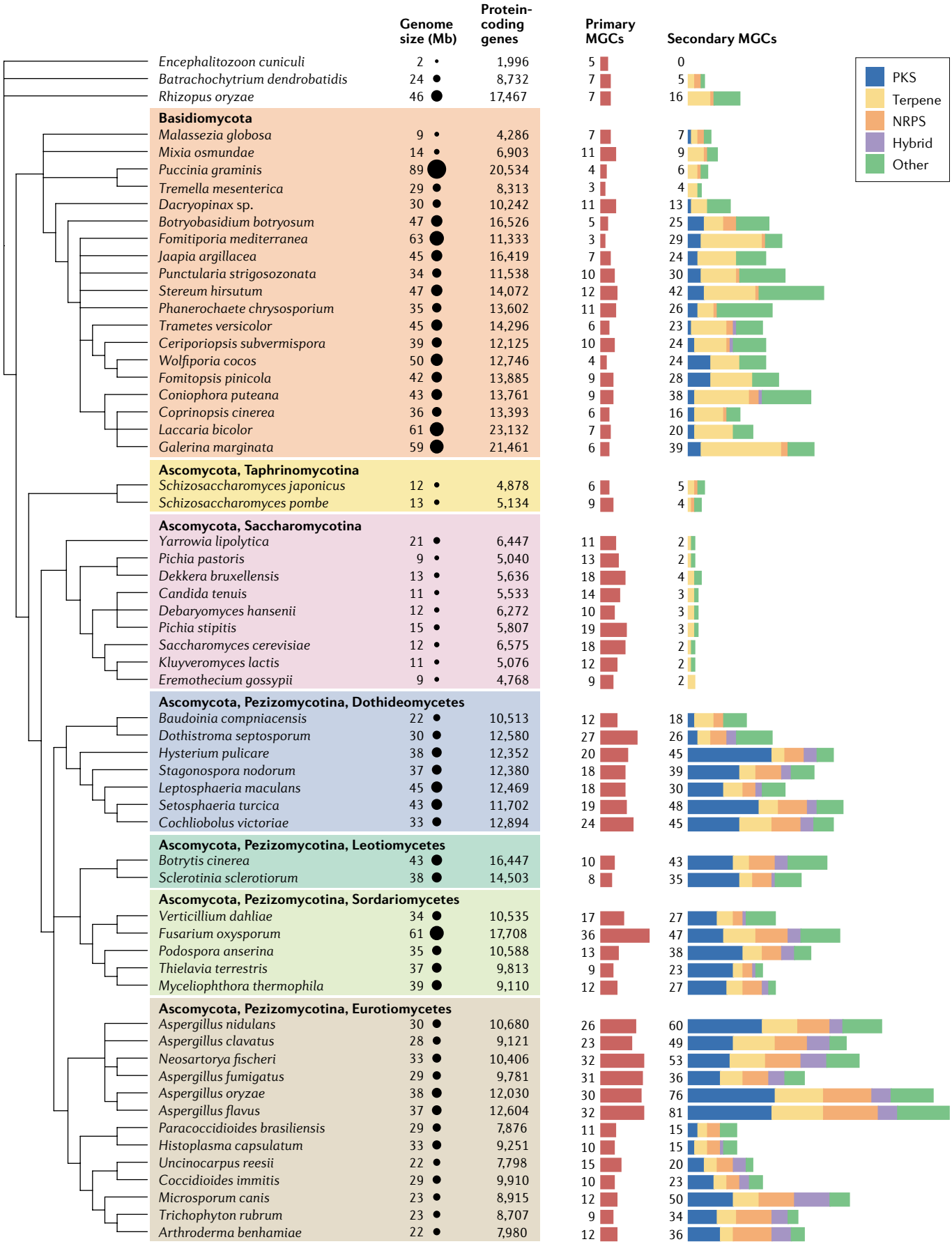


Fig. 1 | **Representative metabolic gene clusters from fungi. a** | The standard configuration of metabolic gene clusters (MGCs) in primary and secondary metabolism; such MGCs contain enzymes, transporters and regulatory genes, as well as (in the case of MGCs involved in secondary metabolism) backbone synthesis genes. The five-gene MGC involved in nitrate assimilation in *Ogataea polymorpha* contains two enzymes, two transcription factors and a transporter[33]. The MGC for the synthesis of the mycotoxin gliotoxin in the opportunistic human pathogen *Aspergillus fumigatus* contains 13 genes, including a backbone synthesis gene, multiple tailoring structural genes, a transporter, a transcription factor and a gliotoxin oxidase gene that provides self-protection to the fungus[35]. Gene size and spacing are not to scale. **b** | Both primary and secondary MGCs exhibit substantial variation in their gene content and structure. For example, in contrast to a single nitrate assimilation MGC in *O. polymorpha* (see part **a**), the galactose MGC in *Saccharomyces cerevisiae* contains only the three structural genes responsible for the conversion of galactose to glucose-1-phosphate (*gal1*, *gal7* and *gal10*), whereas the regulators of this pathway (*gal3*, *gal4* and *gal80*) as well as the galactose permease *gal2* all reside in distinct chromosomal locations[37]. Similarly, the biotin biosynthesis pathway is located in three different loci in the *S. cerevisiae* genome (a three-gene MGC, a two-gene MGC and an additional gene) and does not contain any regulatory genes[12]. Variation in gene content and structure is also evident for MGCs involved in secondary metabolism. For example, the emericellamide MGC in *Aspergillus nidulans* contains two different backbone synthesis genes[41], whereas the kojic acid MGC in *Aspergillus oryzae* does not contain any backbone synthesis genes[23]. Moreover, the genes involved in the biosynthesis of dothistromin in *Dothistroma septosporum* reside on six different loci (five of those are MGCs)[27], whereas the genes involved in the biosynthesis of fumagillin and pseurotin in *Aspergillus fumigatus* are intertwined and under the control of the same regulator[50].

Sordariomycetes) consists of at least two two-gene MGCs[46]; the dothistromin biosynthetic pathway[27] resides in six separate loci, five of which are MGCs (FIG. 1b), on a single chromosome in *Dothistroma septosporum* (Ascomycota, Dothideomycetes); and the biosynthetic pathway for the meroterpenoids austinol and dehydroaustinol[47] in *A. nidulans* consists of a four-gene MGC on chromosome 5 and a ten-gene MGC on chromosome 8. Multiple MGCs have been additionally described for *Fusarium* trichothecenes[48], *Aspergillus* aflatrem[49], *Cochliobolus* T-toxin[43] and *Aspergillus* prenyl xanthones[28].

Moreover, MGCs for different pathways can also reside next to each other on a chromosome or even be intertwined. For example, the MGC for the secondary metabolite aflatoxin in *A. parasiticus* is located right next to a sugar utilization MGC[34], which is interesting given that sugar availability is a known inducer of aflatoxin


Genome size (Mb)
Protein-coding genes
Primary MGCs
Secondary MGCs
PKS
Terpene
NRPS
Hybrid
Other
Encephalitozoon cuniculi 2 1,996 5 0
Batrachochytrium dendrobatidis 24 8,732 7 5
Rhizopus oryzae 46 17,467 7 16
Basidiomycota
Malassezia globosa 9 4,286 7 7
Mixia osmundae 14 6,903 11 9
Puccinia graminis 89 20,534 4 6
Tremella mesenterica 29 8,313 3 4
Dacryopinax sp. 30 10,242 11 13
Botryobasidium botryosum 47 16,526 5 25
Fomitiporia mediterranea 63 11,333 3 29
Jaapia argillacea 45 16,419 7 24
Punctularia strigosozonata 34 11,538 10 30
Stereum hirsutum 47 14,072 12 42
Phanerochaete chrysosporium 35 13,602 11 26
Trametes versicolor 45 14,296 6 23
Ceriporiopsis subvermispora 39 12,125 10 24
Wolfiporia cocos 50 12,746 4 24
Fomitopsis pinicola 42 13,885 9 28
Coniophora puteana 43 13,761 9 38
Coprinopsis cinerea 36 13,393 6 16
Laccaria bicolor 61 23,132 7 20
Galerina marginata 59 21,461 6 39
Ascomycota, Taphrinomycotina
Schizosaccharomyces japonicus 12 4,878 6 5
Schizosaccharomyces pombe 13 5,134 9 4
Ascomycota, Saccharomycotina
Yarrowia lipolytica 21 6,447 11 2
Pichia pastoris 9 5,040 13 2
Dekkera bruxellensis 13 5,636 18 4
Candida tenuis 11 5,533 14 3
Debaryomyces hansenii 12 6,272 10 3
Pichia stipitis 15 5,807 19 3
Saccharomyces cerevisiae 12 6,575 18 2
Kluyveromyces lactis 11 5,076 12 2
Eremothecium gossypii 9 4,768 9 2
Ascomycota, Pezizomycotina, Dothideomycetes
Baudoinia compniacensis 22 10,513 12 18
Dothistroma septosporum 30 12,580 27 26
Hysterium pulicare 38 12,352 20 45
Stagonospora nodorum 37 12,380 18 39
Leptosphaeria maculans 45 12,469 18 30
Setosphaeria turcica 43 11,702 19 48
Cochliobolus victoriae 33 12,894 24 45
Ascomycota, Pezizomycotina, Leotiomycetes
Botrytis cinerea 43 16,447 10 43
Sclerotinia sclerotiorum 38 14,503 8 35
Ascomycota, Pezizomycotina, Sordariomycetes
Verticillium dahliae 34 10,535 17 27
Fusarium oxysporum 61 17,708 36 47
Podospora anserina 35 10,588 13 38
Thielavia terrestris 37 9,813 9 23
Myceliophthora thermophila 39 9,110 12 27
Ascomycota, Pezizomycotina, Eurotiomycetes
Aspergillus nidulans 30 10,680 26 60
Aspergillus clavatus 28 9,121 23 49
Neosartorya fischeri 33 10,406 32 53
Aspergillus fumigatus 29 9,781 31 36
Aspergillus oryzae 38 12,030 30 76
Aspergillus flavus 37 12,604 32 81
Paracoccidioides brasiliensis 29 7,876 11 15
Histoplasma capsulatum 33 9,251 10 15
Uncinocarpus reesii 22 7,798 15 20
Coccidioides immitis 29 9,910 10 23
Microsporum canis 23 8,915 12 50
Trichophyton rubrum 23 8,707 9 34
Arthroderma benhamiae 22 7,980 12 36

◀ Fig. 2 | **Distribution of predicted primary and secondary metabolic gene clusters across the genomes of representative fungal species.** The genome size and number of protein-coding genes of representative fungal species are indicated. Primary metabolic gene clusters (MGCs) were derived using the previously published Kyoto Encyclopedia of Genes and Genomes (KEGG) network approach[111]; pathways participating in carbohydrate, energy, lipid, nucleotide, amino acid, glycan and cofactor or vitamin metabolism were classified as primary metabolic pathways. The categories and counts of secondary MGCs shown were obtained using AntiSMASH[167]. A list of predicted MGCs is provided in Supplementary Table 2. NRPS, nonribosomal peptide synthetase; PKS, polyketide synthase.

biosynthesis. Similarly, the genes for the secondary metabolites fumagillin and pseurotin are intertwined in a single MGC located at the subtelomeric region on chromosome 8 of *A. fumigatus*[50] (FIG. 1b). The fumitremorgin MGC is also nearby, residing ~35 kb away in the 5′ direction from the fumagillin MGC[50], whereas a single *A. fumigatus* isolate has been found to also contain a five-gene secondary MGC immediately adjacent to the 3′ end of the intertwined fumagillin and pseurotin MGC[51].

Polymorphism and divergence

Variation in fungal populations. MGCs harbour substantial amounts of genetic polymorphism[51]. Perhaps the most common type is associated with null alleles; a recent examination of variation in the 36 described MGCs in a sample of genomes from 66 *A. fumigatus* isolates revealed an abundance of non-functional gene polymorphisms, gene gain and loss polymorphisms, and whole cluster gain and loss polymorphisms[51]. Such alleles are common in populations of many fungal species and suggest that the full MGC diversity of the population is captured only by the pangenome[52] and not by the genomes of individual strains. For example, *A. flavus* field isolates often contain null alleles that render them unable to produce the potent mycotoxins aflatoxin and cyclopiazonic acid[53]. Similarly, null alleles have also been observed for the gibberellin and fumonisin MGCs in *Fusarium oxysporum* and *Fusarium fujikuroi*, respectively[54,55], for the ochratoxin MGC in *Aspergillus carbonarius*[56], and for diverse PKS-containing and NRPS-containing MGCs in the major pathogen of wheat *Zymoseptoria tritici* (Ascomycota, Dothideomycetes)[57]. In some cases, it is the null alleles that are present at high frequency in the population. For example, the major allele of the bikaverin MGC in the noble rot *Botrytis cinerea* (Ascomycota, Leotiomycetes) is a non-functional gene cluster composed of pseudogenes in three of the six pathway genes, including the backbone PKS gene *bik1* (REFS[58–60]). A recent screen showed that a functional MGC, which colours the mycelium of the fungus pink, was present in only 7 of the 33 isolates tested[60]. Finally, the genetic architecture of these null alleles can sometimes correlate with geography and be quite complex. For example, the entire galactose gene network, including the galactose MGC, is dispersed across several chromosomes in *Saccharomyces kudriavzevii*, a close relative of the baker's yeast; interestingly, the network exists as a functional gene network in Portuguese isolates but as a non-functional gene network of allelic pseudogenes in Japanese isolates[61].

Although the phenotypes associated with these null polymorphisms are well characterized, their adaptive importance, if any, is less well understood. For example, it has been hypothesized[62] that the atoxicity stemming from the null alleles in the aflatoxin MGC of *A. oryzae*, a fungus used in the making of the Japanese alcoholic drink sake, might have been driven by its impact on the survival of *S. cerevisiae* (aflatoxin is genotoxic[63]); the two fungi are co-cultured during the making of sake.

Much less frequently, MGCs can be found in various copy numbers in fungal populations. For example, the two-gene MGC involved in the synthesis of the osmoprotectant betaine shows different copy numbers across *A. fumigatus* isolates[64]. Another recently discovered and relatively rare type of polymorphism is associated with secondary MGCs that are located in different genomic locations in different isolates. For example, a secondary MGC found in six *A. fumigatus* isolates seems to be located on chromosome 1 in two isolates, on chromosome 4 in three isolates and on chromosome 8, immediately adjacent to the 3′ end of the intertwined fumagillin and pseurotin secondary metabolite gene supercluster[50], in another isolate[51]. In all six isolates, the secondary MGC is flanked by genomic regions containing transposable elements[51].

Perhaps the most unusual and least understood type of variation is observed at loci where non-homologous MGCs correspond to 'allelic' polymorphisms. Although these involve multiple genes, they are reminiscent of the idiomorph alleles in the fungal mating locus[65]. One example comes from *A. flavus*[62], where analysis of the genomes of eight different isolates identified a genomic region on chromosome 7 in which three isolates contain a nine-gene MGC highly similar to the MGC for the production of a volatile sesquiterpene in *Trichoderma virens* (Ascomycota, Sordariomycetes)[66], and the remaining five isolates contain a six-gene MGC[62]. Interestingly, only the gene encoding the terpene synthase, the *gapdh* gene and a few non-coding regions are homologous between the two 'alleles'[62]. Similarly, examination of 66 *A. fumigatus* isolates identified a locus on chromosome 3 that contains at least six different alleles that share no or partial homology to one another. Interestingly, the different alleles exhibit different metabolite profiles, which suggests that these idiomorph polymorphisms contribute to fungal chemodiversity[51].

Variation between species. Most primary MGCs show broad taxonomic distributions, whereas most secondary MGCs show narrow taxonomic distributions and are found in one or a few closely related species. For example, the galactose utilization and the nitrate assimilation MGCs are widespread across the more than 1,000 species comprising the Saccharomycotina subphylum[9,11], whereas a recent comparison of the 260 secondary MGCs collectively present in four distantly related *Aspergillus* species identified only a single two-gene MGC that was shared by two of the species examined[24]. However, exceptions exist[67], and the numbers of secondary MGCs shared between species increase in comparisons of close relatives[68,69].

Pangenome
The full complement of genes present in a population or species, which includes genes present in all individuals as well as genes that are present in only some individuals (or even in a single individual).

Idiomorph alleles
The non-homologous alleles that determine the fungal mating type of an isolate.

The types and trends of observed variation between species match well with those observed within species. A very common type is presence or absence variation across a lineage. In many cases, this variation stems from wholesale MGC loss or pseudogenization events. For example, both the galactose[11,70] and the L-rhamnose[70] utilization MGCs have been lost multiple times in the class Saccharomycetes. Similarly, multiple losses of secondary MGCs have been observed in various lineages (for example, loss of the bikaverin MGC in various *Botrytis* species[59] and of the *ACE1* MGC in several Pezizomycotina lineages[71]). However, in other instances, loss of MGC function has been accompanied by loss of some but not all its constituent genes (for example, several *Botrytis* species have lost the bikaverin MGC[59] but have retained one of its six genes, the non-pathway-specific regulator *bik4*; the same is true of the depudecin MGC across diverse fungal genera[72]). In addition, genomic comparisons between species have also uncovered loci whose synteny is conserved but whose MGC composition is variable[73], which are reminiscent of the idiomorph MGCs observed in fungal populations[51,62].

Conserved MGCs typically exhibit substantial variation in structure and gene content across species. For example, the nitrate assimilation MGC is clustered in *O. polymorpha*, partially clustered in *A. nidulans* and unclustered in *Neurospora crassa* (Ascomycota, Pezizomycotina)[33]. Furthermore, whereas the pathway is regulated by the orthologous transcription factors *nirA* and *nit4* in *A. nidulans* and *N. crassa*, respectively, the *O. polymorpha* MGC contains two unrelated transcription factors[33], which is suggestive of regulatory rewiring. Variation in MGC structure and gene content has been observed in several other MGCs, such as the allantoin utilization MGC[10] and the proline assimilation MGC[33]. Similarly, regulatory rewiring has been observed in the galactose MGC[38,74] as well as in many secondary MGCs in *Aspergillus* species[24]. The complex relationship between structural and phenotypic variation of MGCs across species is illustrated by examination of the galactose pathway across fungi (BOX 2).

The more narrow and discontinuous taxonomic distribution of secondary metabolic pathways means that there are fewer known examples of variation in structure and gene content between species. One well-studied case involves the MGCs for the biosynthesis of trichothecenes[48] in *Fusarium* species. Whereas the genes involved in this pathway reside in three distinct loci in *Fusarium graminearum* and *Fusarium sporotrichioides*, with the majority located in a 12-gene MGC, the pathway in *Fusarium equiseti* not only lacks one gene and includes another but also contains the pathway in just two loci, one of which is a 14-gene MGC[48].

Similar to primary MGCs, variation in gene content of secondary MGCs often translates to phenotypic variation. For example, the structural diversity of the potent trichothecene toxins produced by diverse plant pathogenic and entomopathogenic fungi stems from gene duplications, gene losses and functional divergence of genes in the trichothecene MGCs[75]. Perhaps the best known example is that of the MGCs involved in the production of the related compounds sterigmatocystin, *O*-methylsterigmatocystin (OMST) and aflatoxins found in several different species in the genus *Aspergillus*. Sterigmatocystin and aflatoxin MGCs are evolutionarily related but differ in gene content, order and orientation[76,77]. In these MGCs, which metabolite (or metabolites) is synthesized depends on three genes that are present in aflatoxin MGCs but absent from sterigmatocystin MGCs. Specifically, *aflP* is required for converting sterigmatocystin to OMST[17], whereas *aflU* and *aflQ* are required for converting OMST to the G and B classes of aflatoxins[17,78], respectively. In some cases, functional variation can be due to amino acid differences in proteins encoded by orthologous genes; for example, the biosynthesis of group B and C fumonisins is thought to be determined by sequence variation in a single gene (*fum8*) in the 17-gene MGC found in different *Fusarium* species[79].

MGC formation and maintenance mechanisms

MGC formation. How MGCs originate in fungal genomes and how they are maintained in natural populations are questions that have only recently begun to be examined. In all cases where the origins of a given MGC have been confidently inferred, they involve relocations of native genes as well as of duplicates of native genes. The best known example is the origin of the *DAL* MGC involved in allantoin metabolism in *S. cerevisiae* because the different degrees of clustering found in the genomes of close relatives of *S. cerevisiae* give us an evolutionary peek into the steps that occurred during the evolutionary assembly of this MGC[10]. In this case, four of the six genes in the MGC are native genes that relocated to the MGC locus, whereas the other two are duplicates of native genes residing elsewhere in the genome that relocated to the MGC locus after duplication. Other examples of MGCs that have arisen de novo by relocation of native genes (or of their duplicates) include the galactose[11], betaine[29] and cyanate detoxification[16] MGCs.

An alternative hypothesis is that certain fungal MGCs, particularly those associated with secondary metabolism, may have originated via horizontal gene transfer (HGT) from bacteria, as has been proposed for the penicillin and cephalosporin MGCs[46,80]. Under this hypothesis, secondary MGCs were originally formed as operons in bacteria and were subsequently horizontally transferred to fungi. However, given the paucity of examples of fungal MGCs that were horizontally acquired from bacteria, it is likely that most fungal MGCs originated via relocation (and sometimes prior duplication) of native genes.

MGC spread. Once an MGC is formed and established in a fungal species, it can occasionally spread beyond the species' boundaries via HGT. Examination of fungal genomes has revealed nearly a dozen instances of HGT between fungi involving both primary and secondary MGCs[81]. Notable examples include the HGT of a 23-gene MGC involved in the biosynthesis of the sterigmatocystin mycotoxin from an *Aspergillus* ancestor to a *Podospora* ancestor[77] (Ascomycota,

Regulatory rewiring
The same metabolic pathway (or other biological process) is regulated by different (non-homologous) transcriptional factors (and circuits) in different species.

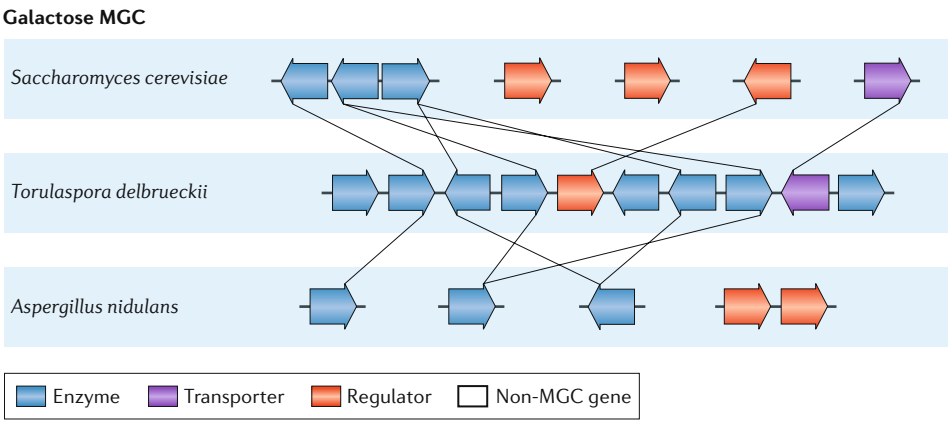
Horizontal gene transfer
(HGT; also known as lateral gene transfer). The transfer of genetic material from one organism to another through a process other than reproduction.

Box 2 | **Evolution of the galactose utilization metabolic gene cluster**

Most species in the class Saccharomycetes (Ascomycota), including *Saccharomyces cerevisiae*, *Kluyveromyces lactis* and *Candida albicans*, contain a galactose metabolic gene cluster (MGC) that includes *gal1*, *gal7* and *gal10*, the three structural genes responsible for the conversion of galactose to glucose-1-phosphate[11,37,38]. Even though this MGC is evolutionarily conserved, the regulation of the galactose MGC is markedly different between the three species[126]. Specifically, whereas the presence of galactose causes the expression of the MGC genes in *S. cerevisiae* to increase more than 1,000-fold[127], they increase by approximately 100-fold in *K. lactis*[128] and by less than 10-fold in *C. albicans*[38]. Furthermore, whereas the galactose MGC in *S. cerevisiae* and *K. lactis* is regulated by *gal4* (REF.[128]), the *C. albicans* MGC is regulated by the unrelated (to *gal4*) genes *rtg1* and *rtg3* (REF.[74]).

One major exception to the conserved genomic structure of the galactose MGC in Saccharomycetes is *Torulaspora delbrueckii* (see the figure). Instead of the typical three-gene galactose MGC, this yeast contains a ten-gene MGC comprising two *gal1*, two *gal10* and one *gal7* structural genes, one transcriptional activator (*gal4*), one permease (*gal2*) and an additional transporter (*hgt1*), as well as *mel1*, a secreted α-galactosidase that breaks down melibiose into glucose and galactose, and *pgm1*, which converts glucose-1-phosphate to glucose-6-phosphate. Thus, the three-gene galactose MGC in this organism seems to have been co-opted, along with other genes, into a larger ten-gene MGC, which is presumably involved in galactose assimilation as well as in additional catabolic activities[129].

Three-gene galactose MGCs are also present in the independently evolved yeasts in the genera *Schizosaccharomyces* (Ascomycota, Schizosaccharomycetes) and *Cryptococcus* (Basidiomycota, Tremellomycetes)[11].
The *Schizosaccharomyces pombe* MGC is functional and was acquired via horizontal gene transfer (see main text) from Saccharomycotina yeasts[11]. Remarkably, the organism does not grow on galactose and instead uses the MGC in glycosylation[85]. By contrast, the basidiomycete yeast *Cryptococcus neoformans* contains both a three-gene galactose MGC (which evolved independently of the MGC in Saccharomycetes) and unclustered *gal1*, *gal7* and *gal10* genes in its genome. Functional experiments have shown that, whereas the unclustered *gal10* is involved in galactose metabolism at 37 °C and is required for pathogenicity, involvement of the clustered *gal10* in galactose metabolism is temperature-independent, and the gene is not required for pathogenicity[130,131]. Finally, filamentous fungi, such as *Aspergillus nidulans*, contain the same three structural genes (*gal1*, *gal7* and *gal10*), but they are dispersed in the genome (see the figure). Interestingly, galactose utilization in filamentous fungi is governed by distinct regulatory mechanisms[132] and involves additional metabolic functions compared with yeasts[109].



Sordariomycetes), the HGT of a three-gene MGC involved in galactose utilization from a *Candida* ancestor to the *Schizosaccharomyces* ancestor[11], the HGT of a six-gene MGC involved in the biosynthesis of the red pigment bikaverin from a *Fusarium* ancestor to a *Botrytis* ancestor[58,59], the HGTs of a five-gene MGC associated with the biosynthesis of the hallucinogen psilocybin within fungi[82] and the multiple instances of HGT in filamentous fungi associated with the MGCs involved in the biosynthesis of gliotoxin and related mycotoxins[83]. Transferred MGCs may be subsequently lost (for example, the bikaverin MGC in several *Botrytis* species[59]), may retain their original function (for example, the sterigmatocystin MGC in *Podospora*[84]) or may diverge in function (for example, the galactose MGC in *S. pombe*, which does not use galactose as a source of energy but for glycosylation[85], as well as the MGCs involved in the biosynthesis of gliotoxin and related mycotoxins[83]).

MGC formation and maintenance are driven by selection. Unlike clusters of tandemly duplicated genes, where non-homologous recombination gives rise both to the genes and to their physical proximity in the genome[86], the observed clustering of the non-homologous genes that form MGCs must be the end result of an evolutionary process. The only known evolutionary force that can drive MGC formation, as well as their establishment and maintenance, in natural populations is natural selection[87].

The best evidence for the involvement of natural selection in MGC formation comes from the several known instances of convergent evolution of MGC formation. For example, the galactose MGCs present in the Saccharomycotina yeasts and in the unrelated basidiomycete yeast genus *Cryptococcus* have evolved independently; although the same three genes (*gal1*, *gal7* and *gal10*) are involved, the Saccharomycotina yeast *gal* genes are most closely related to the unclustered *gal*

Convergent evolution
The independent evolution of similar traits or features in organisms belonging to different, unrelated lineages.

genes of filamentous fungi (ASCOMYCETES), and the *Cryptococcus gal* genes are most closely related to the unclustered *gal* genes of other basidiomycete fungi. Independent origins of clustering have also been observed for the betaine MGC[29] as well as for a two-gene MGC putatively involved in cyanate detoxification[16].

However, the impact of natural selection on MGCs does not end when MGCs are formed. For example, several MGCs involved in primary metabolism are widely conserved both in terms of sequence and in terms of function (for example, the galactose MGC in the class Saccharomycetes[11]; BOX 2), which suggests that they have been under the long-term influence of purifying selection. Balancing selection has also been inferred to maintain multiple, functionally distinct MGC alleles involved in secondary metabolism for long time periods and across species boundaries, as in the cases of the trichothecene MGC in the *Fusarium graminearum* species complex[88] and of the aflatoxin MGC in *A. parasiticus* and closely related species[89]. More generally, genes participating in fungal MGCs are consistently among the top genes identified in genome-wide scans of selection[62,90,91].

Finally, selection may also be involved in the generation of the recurrent, and sometimes highly frequent, wholesale cluster loss or loss-of-function polymorphisms observed in the MGCs of many different fungi[51,57]. For example, many of the secondary metabolites that are biosynthesized by MGCs are secreted, which means that these metabolites can also be beneficial to conspecific organisms that are physically adjacent to the producing organism but that are not themselves producers (and do not carry the cost of synthesizing the metabolite)[2]. Thus, in a manner analogous to the black queen hypothesis[92], variants associated with MGC loss or inactivation may be adaptive.

Genetic mechanisms. Three non-mutually exclusive genetic models have been proposed to be involved in the formation and maintenance of fungal MGCs, namely, co-regulation, genetic linkage and selfishness. According to the co-regulation[93] model, fungal MGCs evolved via selection for tighter control of gene expression on pathway genes. That genes within both primary and secondary MGCs are tightly co-regulated is well established from transcriptomic comparisons of organisms grown in different conditions[94–98] and is considered a hallmark of specialized metabolism in general[99].

Multiple lines of evidence support the role of co-regulation in the evolutionary maintenance of fungal MGCs. Broad comparative studies examining the relationship between clustering and gene expression in *S. cerevisiae* have shown that genes with conserved synteny tend to be co-expressed[93]. Furthermore, the genes in both primary and secondary MGCs are more often divergently oriented in the genome, and their intergenic regions often contain shared regulatory binding sites, both of which are features that are strongly associated with transcriptional co-regulation[100]. For example, 18 of the 23 genes in the sterigmatocystin MGC of *A. nidulans* form pairs of divergently oriented genes whose shared intergenic regions contain regulatory binding sites for AflR, which is the transcriptional activator of the MGC[77,101]. Similarly, the intergenic region between the divergently oriented *gal1* and *gal10* genes in the *S. cerevisiae* galactose MGC contains shared regulatory binding sites for Gal4p and Mig1p, which are the transcriptional activator and the glucose-dependent repressor of the MGC[37], respectively. Finally, in recent years, several studies have furnished evidence in support of chromatin-mediated regulation of MGCs, which occurs via numerous chemical modifications of histones, the core constituents of chromatin. As this type of regulation establishes active and silent regions along a chromosome or chromosomal region, it is also highly likely to contribute to the maintenance of MGCs in fungal genomes[26,33,102].

The second genetic model argues that MGCs evolved through selection for tighter genetic linkage of their constituent genes[10]. Assuming that interactions between different alleles of genes in a pathway are non-additive, selection can favour the linkage of positively interacting alleles, leading to the formation of a co-adapted gene complex[87]. Support for this hypothesis comes from several studies showing that recombination rates are lower in regions containing MGCs or within MGCs themselves. For example, the *DAL* MGC involved in allantoin utilization in *S. cerevisiae* resides in a genomic region with one of the lowest recombination rates in the yeast genome[10]. Similarly, population analyses of secondary MGCs, such those responsible for the mycotoxins dothistromin[27] and aflatoxin[103,104], suggest that recombination rates within MGCs are low and identify blocks of sequence that seem to be in near complete linkage disequilibrium.

The final model, that of cluster selfishness[105], is a special case of the genetic linkage model described above. Under this model, the observed genetic linkage of MGCs was favoured by the repeated occurrence of HGT, enabling selection for the formations themselves, beyond any selective advantage that they confer to the organism[105]. A key part of this model is that MGCs will have a selective advantage over their non-clustered counterparts because they can propagate not only through vertical transmission but also via HGT, even when they reduce the fitness of the organism. Although numerous instances of HGT involving fungal MGCs have been described[81], the observed rate of HGT among MGCs seems to be too low to provide support for this model.

Very few studies have experimentally tested the effects of clustering and their implications for the different genetic models. One such study compared the growth rate of a *S. cerevisiae* isolate with the galactose MGC against the growth rate of an isolate that contained a non-clustered version of the pathway on a galactose medium[106]. Notably, the isolate with the galactose MGC exhibited equal fitness compared with the isolate that contained the unclustered pathway, which suggests that co-expression does not account for the origin or maintenance of the galactose MGC[106]. Furthermore, another study tested whether the order and orientation of genes within MGCs influence fitness[107]. By inverting the orientation of *dal2* in the *S. cerevisiae DAL* MGC — an

Black queen hypothesis
The idea that the loss of genes whose functions are associated with public goods (for example, the production of an antitoxin) may be individually advantageous (up to the point at which the cost associated with loss of public goods exceeds the benefit of loss).

orientation found in the distantly related budding yeast *Naumovia castellii* (Ascomycota, Saccharomycetes) — during growth in allantoin-containing medium, the authors showed that this inversion reduced not only the expression of the *dal2* gene but also those of the neighbouring *dal1* and *dal4* genes, arguing that the orientation of genes within an MGC is important for MGC regulation[107].

Phenotypic mechanisms. All three genetic models have been extensively discussed as explanations of the origin and maintenance of primary and secondary fungal MGCs. However, coordinated transcript abundance levels of genes in a pathway or co-adapted gene complexes are not actual phenotypes; rather, they represent genetic mechanisms by which the effects of selection on a phenotype could be mediated. If so, what is the phenotype favoured by selection?

One attractive hypothesis argues that the phenotype in question is toxicity avoidance, that is, the avoidance of accumulating toxic intermediates[11,29,106] (FIG. 3). The protective advantage offered by genes in a pathway that is organized as an MGC, as opposed to a pathway whose gene members are scattered across chromosomes, could be twofold. On the one hand, gene co-location may be a better safeguard against partial loss of a pathway whose intermediates are acutely toxic. On the other hand, gene co-location may enable more precise regulation of pathways with acutely toxic intermediates.

Several lines of evidence support the toxicity avoidance hypothesis. For example, several intermediates of both primary and secondary MGCs are known to be toxic[29], and the pathways associated with human metabolic disorders that stem from the accumulation of toxic intermediates, such as alkaptonuria and the different types of galactosaemia and tyrosinaemia, are often known to exist in fungi as MGCs[13,29] (FIG. 3). Disruption of any of the three *GAL* MGC genes[11], which are involved in the breakdown of galactose and are known to be conserved in fungi and humans, results in the human

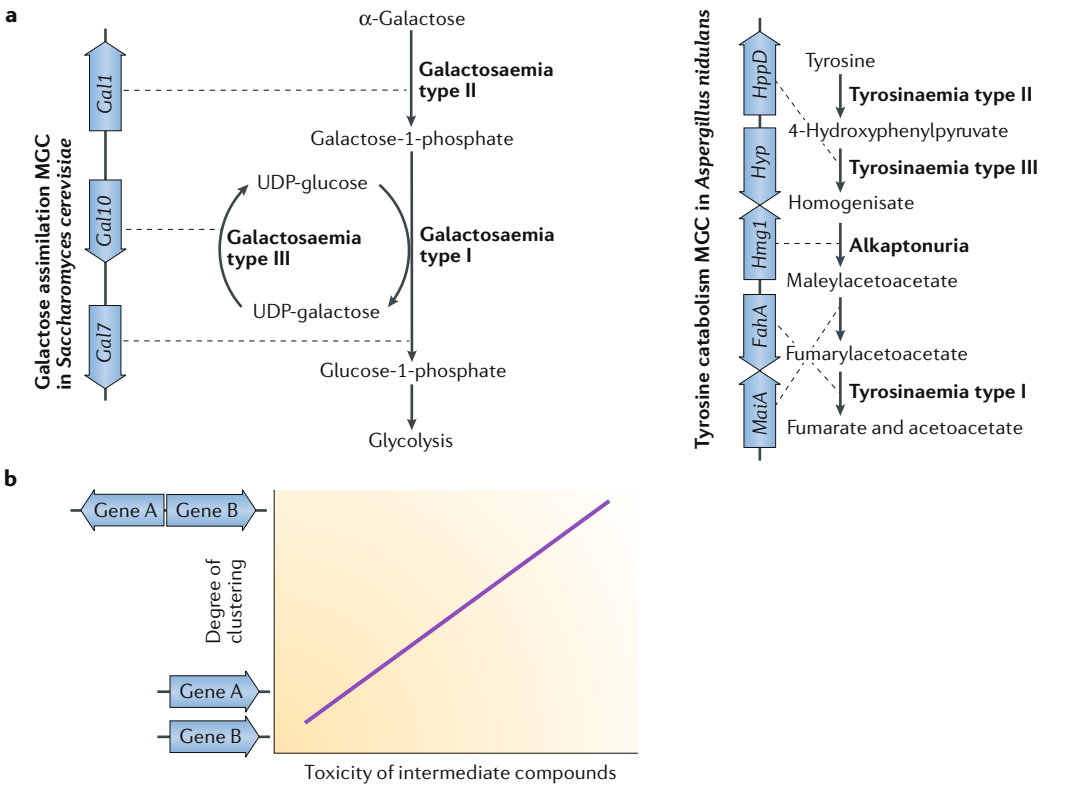


Fig. 3 | **The toxicity avoidance hypothesis.** The toxicity avoidance hypothesis argues that fungal metabolic gene clusters (MGCs) represent an adaptation against the accumulation of toxic intermediate compounds[29]. **a** | The intermediate compounds of several fungal MGCs, such as those involved in galactose assimilation and tyrosine catabolism, are toxic. Defects in the corresponding human genes lead to several metabolic disorders in humans. Disruption of any of the three human *GAL* genes, which are involved in the breakdown of galactose, results in galactosaemia. The most severe type (type I) results from the inactivation or loss of the *gal7* gene (*GALT* in human), which leads to the accumulation of galactose-1-phosphate, a highly toxic intermediate that disrupts glycolysis[115]. Similarly, disruption of one of the genes in the human tyrosine catabolism pathway results in alkaptonuria or black urine disease, whereas disruption of an additional three human genes in the pathway results in tyrosinaemia. The most severe type (type I) results from the inactivation or loss of the *fahA* gene, which leads to accumulation of fumarylacetoacetate, a reactive, genotoxic compound that can cause hepatic cancer[13]. **b** | According to the toxicity avoidance hypothesis, MGCs are expected to be more common in pathways that generate very toxic metabolic intermediates. UDP, uridine diphosphate. Figure adapted with permission from REF.[29], Proceedings of the National Academy of Sciences of the United States of America.

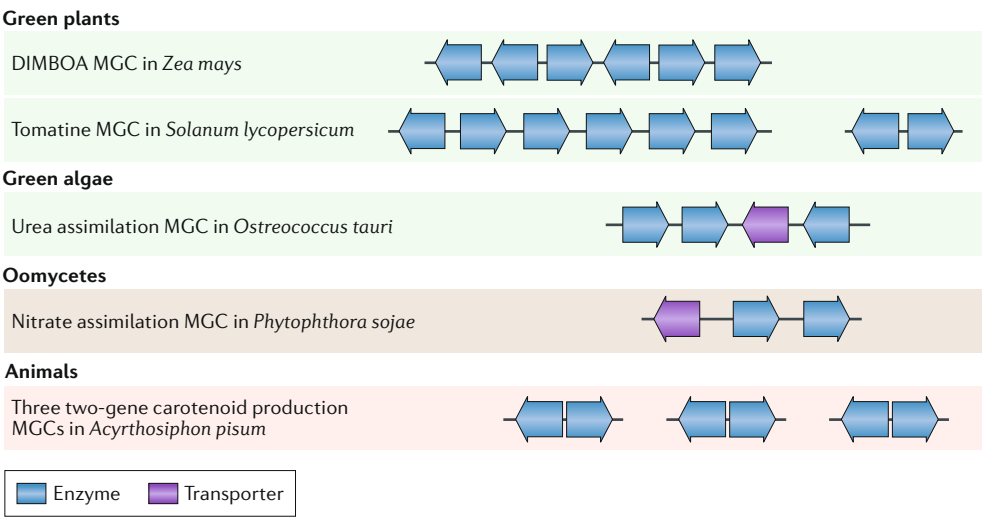
Box 3 | **Metabolic gene clusters in other eukaryotes**

Metabolic gene clusters (MGCs) have been identified in several other eukaryotic lineages (see the figure; Supplementary Table 1), but by far, most non-fungal MGCs are known from land plants[133,134]. The first plant MGC was identified in *Zea mays* (maize); this five-gene MGC is responsible for the biosynthesis of the benzoxazinoid 2,4-dihydroxy-7-methoxy-1,4-benzoxazin-3-one (DIMBOA), a precursor to several defensive chemicals that provide resistance towards insect herbivores and microbial pathogens[135,136]. Several other MGCs were subsequently described, including a polyketide MGC in *Hordeum vulgare* (barley)[137]; cyanogenic glucoside MGCs in *Manihot esculenta* (cassava), *Sorghum bicolor* (sorghum), *Trifolium repens* (white clover)[138] and *Lotus japonicus*[139]; alkaloid MGCs in *Solanum lycopersicum* (tomato), *Solanum tuberosum* (potato) and *Papaver somniferum* (opium poppy)[140,141]; and terpene MGCs in *Arabidopsis thaliana* (thale cress), *Lotus japonicus*, *Avena strigosa* (oat), *Oryza sativa* (rice), *S. lycopersicum*, *Cucumis sativus* (cucumber) and *Ricinus communis* (castor)[142–150]. Interestingly, all known plant MGCs seem to have originated via gene duplication, neofunctionalization and genome reorganization of native genes rather than through horizontal gene transfer (HGT) of genes or MGCs from microorganisms[133,151].

The prevalence of MGCs in plant genomes is currently under debate. For example, terpenes make up the largest class of specialized metabolites in plants, and there is evidence that more terpene MGCs may exist in plant genomes than are currently described. An examination of two key genes involved in terpene biosynthesis, terpenoid synthase and cytochrome P450 (*CYP*), found an average of seven terpenoid synthase–*CYP* gene pairs per plant genome (given a ≤50 kb intergenic distance), which is far more common than expected by chance[152]. This work, the growing list of functionally characterized plant MGCs and bioinformatic predictions of thousands of candidate MGCs have led some to hypothesize that MGCs in plants are widespread[153–155]. However, the fact remains that many characterized secondary metabolite pathways in plants do not form MGCs[156–158]. Moreover, a co-expression network analysis of bioinformatically predicted plant MGC candidates (for example, those that have not been functionally validated and linked to a metabolite product) suggests that the genes within these predicted MGCs are not co-regulated[99]. The lack of coordinated expression is in stark contrast to the pattern seen in fungi, where both characterized and predicted MGCs show strong co-expression[62,94,95,97,98], which suggests that many predicted MGCs in plants do not correspond to genuine secondary metabolite pathways and that the number of MGCs in plants may be overestimated.

In contrast to plant MGCs, the only two MGCs known in animals were acquired wholesale through HGT. Specifically, bdelloid rotifers have acquired a bacterial MGC comprising a racemase and a ligase thought to be involved in peptidoglycan cell wall biosynthesis[159]. Similarly, the three two-gene carotenoid MGCs encoding genes for carotenoid synthase–carotenoid cyclase and carotenoid desaturase found in aphids were obtained by HGT of a single ancestral MGC from fungi and its subsequent duplication in the insect lineage[160].

The few MGCs known in microbial protists and algae are all involved in the utilization of environmental nitrogen. For example, several oomycetes (water moulds) contain a three-gene MGC for nitrate assimilation, also found in various fungi, with phylogenetic analysis suggesting that the MGC arose once in oomycetes and was subsequently acquired via HGT in the ancestor of Dikarya fungi[9]. By contrast, green algae in the Chlorophyceae, Prasinophyceae and Trebouxiophyceae classes possess a 6–8-gene MGC for nitrate assimilation that evolved independently from the MGC in fungi and oomycetes[161–164]. Independent origins have also been observed in MGCs for urea assimilation in green algae, organisms that thrive in aquatic environments where nitrogen is often a limiting resource. For example, *Ostreococcus tauri* (Prasinophyceae) has a four-gene nickel-dependent urease MGC, which is evolutionarily distinct from the five-gene and three-gene urea assimilation MCGs present in *Chlamydomonas reinhardtii* and *Picochlorum* sp. (both Chlorophyceae), respectively[161,165,166].



metabolic disorder galactosaemia[108]. The most severe type results from the loss of the *gal7* gene (*GALT* in human), which leads to the accumulation of galactose-1-phosphate, which is a highly toxic intermediate that disrupts glycolysis[108] (FIG. 3a). Remarkably, deletion of the clustered *gal7* gene in *S. cerevisiae* is toxic in the presence of galactose, whereas deletion of its unclustered *A. nidulans* orthologue, *gald*, is not toxic[109].

One prediction of the toxicity avoidance hypothesis is that metabolic pathways encoded by MGCs should generate intermediate compounds that are more toxic than those of pathways whose genes are not known to form MGCs (FIG. 3b). Examination of the toxicity of intermediate compounds handled by pairs of enzymes that are neighbours in the metabolic network and whose genes are located immediately adjacent to each other in the genome shows that they are enriched for intermediate compounds that have lower lethal doses or contain reactive functional groups[29].

Facilitating fungal adaptation. The organization of several primary and secondary metabolic pathways into MGCs has important implications for fungal adaptation, largely by enabling the wholesale acquisition, duplication or loss of entire metabolic pathways[110]. For example, several large-scale genomic comparisons have provided evidence of duplication, loss and HGT of both primary and secondary MGCs[9,11,58,59,64,71,72,77,81,111]. Of the three processes, duplication (followed by divergence) and loss are far more frequent than HGT[111], providing a mechanistic explanation of why many MGCs, especially those involved in secondary metabolism, show very narrow taxonomic distributions and can vary extensively across closely related species or even between isolates of the same species[51,62,64,112]. Importantly, many such events can be linked to fungal ecology[81,110], occurring between species that seem to have overlapping ecological niches.

MGCs in other eukaryotes

MGCs are far more prevalent in fungi than in other eukaryotes (BOX 3). A comprehensive explanation of this higher prevalence continues to elude us, but several factors are likely to be involved. For example, population size, population structure and mating system are all known to influence the efficacy of natural selection to shape eukaryotic genome architecture[113,114], including the formation and maintenance of MGCs[61]. The large differences in metabolic capabilities between eukaryotic lineages (for example, not all have secondary metabolism) mean that their distributions of MGCs may also differ.

However, differences in the degree of gene clustering are also evident in metabolic pathways that are shared across many eukaryotes. For example, the galactose and tyrosine pathways are both genomically organized as MGCs in many fungi[11,13] but not in animals[30], even though disruption of these pathways has very strong fitness consequences in both lineages[13,115,116]. Such differences may be because different lineages have evolved different mechanisms for maintaining pathway efficiency[30]. For example, metabolic activity in animals is typically confined to specialized organs, such as the liver and kidneys, and thus selection for pathway efficiency may result in the tissue-specific co-expression of pathway genes. By contrast, because fungal metabolic activity occurs in the entire colony, selection for metabolic efficiency may result in their clustering. In support of this hypothesis, a recent study showed that primate orthologues of genes from fungal MGCs were significantly more correlated in their tissue-specific co-expression profiles than orthologues of genes from unclustered fungal metabolic pathways[30].

Many of the above explanations fail to fully account for the relative lack of MGCs in non-fungal microbial eukaryotes, particularly those whose population biology, mating system and ecology are similar to fungi. One possible explanation is the genome sampling bias towards eukaryotic microorganisms with pathogenic lifestyles[117], which tend to show reduced genomes, reduced metabolic diversity and increased ecological specialization[118–120]. Interestingly, PKSs, one of the major protein families involved in fungal secondary metabolism, have been identified in amoebae[121], apicomplexan parasites[122] and marine algae[123,124]. For example, the slime mould *Dictyostelium discoideum*, a soil-dwelling amoeba with a lifestyle similar to many filamentous fungi, contains 43 putative PKSs and numerous small-molecule transporters[121], although whether these genes are components of MGCs remains an open question. A recent bioinformatic screen identified 24 putative MGCs of 2 or more genes in 10 non-fungal, non-land plant, eukaryotic genomes[125], half of which were from *D. discoideum*.

Summary and outstanding questions

Fungal genomes contain a remarkable diversity of primary and secondary metabolic pathways, many of which are found physically linked on the chromosome, forming MGCs. These MGCs are fundamental to the fungal lifestyle, and they are involved in a wide variety of functions, from the utilization of compounds, such as sugars and amino acids, to the production of a wide array of antibiotics and toxins. The ever-increasing decoding of fungal genomes both within and between species has revealed a vast diversity of MGCs residing in fungal genomes and provided the raw materials for understanding how MGCs originate, how they evolve and how they are lost in fungal populations. Coupled with the advent of a wide variety of evolutionary, molecular and chemical approaches for linking MGCs to functions, we anticipate great progress on understanding the relationship between MGCs and their chemotypes across evolution in the near future.

However, several outstanding questions remain. What fraction of pathways forms MGCs and what fraction does not? What is the precise selective advantage conferred by the clustering of a metabolic pathway? Why do some pathways form MGCs in some species but not in others? Why do some genomes contain many MGCs and others very few? Is MGC occurrence associated with particular ecologies? How often does loss of MGC function take place without loss of the underlying genes? Why do fungi harbour so many MGCs while other eukaryotic lineages, especially microbial ones, contain so few? With the advent of genome sequencing technologies and the ever-growing development of molecular, chemical, evolutionary and bioinformatic tools for discovering MGCs and their functions, the time for understanding the making of fungal and eukaryotic chemodiversity is now.

Published online 7 September 2018

1. Whittaker, R. H. New concepts of kingdoms or organisms. *Science* **163**, 150–160 (1969).
2. Richards, T. A. & Talbot, N. J. Horizontal gene transfer in osmotrophs: playing with public goods. *Nat. Rev. Microbiol.* **11**, 720–727 (2013).
3. Pusztahelyi, T., Holb, I. J. & Pocsi, I. Secondary metabolites in fungus-plant interactions. *Frontiers Plant Sci.* **6**, 573 (2015).
4. Stergiopoulos, I., Collemare, J., Mehrabi, R. & De Wit, P. J. Phytotoxic secondary metabolites and peptides produced by plant pathogenic Dothideomycete fungi. *FEMS Microbiol. Rev.* **37**, 67–93 (2013).
5. Keller, N. P., Turner, G. & Bennett, J. W. Fungal secondary metabolism — from biochemistry to genomics. *Nat. Rev. Microbiol.* **3**, 937–947 (2005).
6. Keller, N. P. & Hohn, T. M. Metabolic pathway gene clusters in filamentous fungi. *Fungal Genet. Biol.* **21**, 17–29 (1997).
7. Yue, Q. et al. Functional operons in secondary metabolic gene clusters in *Glarea lozoyensis* (Fungi, Ascomycota, Leotiomycetes). *MBio* **6**, e00703 (2015).
8. Johnstone, I. L. et al. Isolation and characterisation of the crnA-niiA-niaD gene cluster for nitrate assimilation in *Aspergillus nidulans*. *Gene* **90**, 181–192 (1990).
9. Slot, J. C. & Hibbett, D. S. Horizontal transfer of a nitrate assimilation gene cluster and ecological transitions in fungi: a phylogenetic study. *PLOS One* **2**, e1097 (2007).
This is the first reported case in fungi of the acquisition of an entire metabolic pathway by horizontal transfer of an MGC.
10. Wong, S. & Wolfe, K. H. Birth of a metabolic gene cluster in yeast by adaptive gene relocation. *Nat. Genet.* **37**, 777–782 (2005).
This is the first study to show that fungal MGCs originate by relocation of native genes.
11. Slot, J. C. & Rokas, A. Multiple *GAL* pathway gene clusters evolved independently and by different mechanisms in fungi. *Proc. Natl Acad. Sci. USA* **107**, 10136–10141 (2010).
12. Hall, C. & Dietrich, F. S. The reacquisition of biotin prototrophy in *Saccharomyces cerevisiae* involved horizontal gene transfer, gene duplication and gene clustering. *Genetics* **177**, 2293–2307 (2007).
This study shows that the biotin pathway in the baker's yeast, whose five of the six genes reside in two subtelomeric gene clusters, was lost in an ancestor but was subsequently rebuilt through acquisition of bacterial genes via HGT and gene duplication followed by neofunctionalization.
13. Penalva, M. A. A fungal perspective on human inborn errors of metabolism: alkaptonuria and beyond. *Fungal Genet. Biol.* **34**, 1–10 (2001).
14. Hull, E. P., Green, P. M., Arst, H. N. Jr & Scazzocchio, C. Cloning and physical characterization of the L-proline catabolism gene cluster of *Aspergillus nidulans*. *Mol. Microbiol.* **3**, 553–559 (1989).
15. Bobrowicz, P., Wysocki, R., Owsianik, G., Goffeau, A. & Ulaszewski, S. Isolation of three contiguous genes. *ACR1, ACR2* and *ACR3*, involved in resistance to arsenic compounds in the yeast *Saccharomyces cerevisiae*. *Yeast* **13**, 819–828 (1997).
16. Elmore, M. H. et al. Clustering of two genes putatively involved in cyanate detoxification evolved recently and independently in multiple fungal lineages. *Genome Biol. Evol.* **7**, 789–800 (2015).
17. Yu, J. et al. Clustered pathway genes in aflatoxin biosynthesis. *Appl. Environ. Microbiol.* **70**, 1253–1262 (2004).
18. Brown, D. W., Dyer, R. B., McCormick, S. P., Kendra, D. F. & Plattner, R. D. Functional demarcation of the *Fusarium* core trichothecene gene cluster. *Fungal Genet. Biol.* **41**, 454–462 (2004).
19. Schardl, C. L. et al. Plant-symbiotic fungi as chemical engineers: multi-genome analysis of the clavicipitaceae reveals dynamics of alkaloid Loci. *PLOS Genet.* **9**, e1003323 (2013).
This is a comprehensive reconstruction of the evolutionary dynamics of MGCs responsible for four classes of alkaloid toxins, which are characterized by conserved cores of genes that specify the chemical skeleton structures of the toxins and various peripheral genes that are likely involved in chemical variations in toxin structures and their pharmacological effects.
20. Diez, B. et al. The cluster of penicillin biosynthetic genes. Identification and characterization of the pcbAB gene encoding the alpha-aminoadipyl-cysteinyl-valine synthetase and linkage to the pcbC and penDE genes. *J. Biol. Chem.* **265**, 16358–16365 (1990).
21. Kennedy, J. et al. Modulation of polyketide synthase activity by accessory proteins during lovastatin biosynthesis. *Science* **284**, 1368–1372 (1999).
22. Kimura, N. & Tsuge, T. Gene cluster involved in melanin biosynthesis of the filamentous fungus *Alternaria alternata*. *J. Bacteriol.* **175**, 4427–4435 (1993).
23. Terabayashi, Y. et al. Identification and characterization of genes responsible for biosynthesis of kojic acid, an industrially important compound from *Aspergillus oryzae*. *Fungal Genet. Biol.* **47**, 953–961 (2010).
24. Lind, A. L. et al. Examining the evolution of the regulatory circuit controlling secondary metabolism and development in the fungal genus Aspergillus. *PLOS Genet.* **11**, e1005096 (2015).
25. Sanchez, J. F., Somoza, A. D., Keller, N. P. & Wang, C. C. Advances in *Aspergillus* secondary metabolite research in the post-genomic era. *Nat. Prod. Rep.* **29**, 351–371 (2012).
26. Brakhage, A. A. Regulation of fungal secondary metabolism. *Nat. Rev. Microbiol.* **11**, 21–32 (2013).
27. Bradshaw, R. E. et al. Fragmentation of an aflatoxin-like gene cluster in a forest pathogen. *New Phytol.* **198**, 525–535 (2013).
28. Sanchez, J. F. et al. Genome-based deletion analysis reveals the prenyl xanthone biosynthesis pathway in *Aspergillus nidulans*. *J. Am. Chem. Soc.* **133**, 4010–4017 (2011).
29. McGary, K. L., Slot, J. C. & Rokas, A. Physical linkage of metabolic genes in fungi is an adaptation against the accumulation of toxic intermediate compounds. *Proc. Natl Acad. Sci. USA* **110**, 11481–11486 (2013).
This study shows that fungal metabolic genes that handle toxic compounds are more likely to be clustered, suggesting that the phenotype selection acts on to drive gene clustering is the mitigation of toxicity of pathway intermediates.
30. Eidem, H. R., McGary, K. L. & Rokas, A. Shared selective pressures on fungal and human metabolic pathways lead to divergent yet analogous genetic responses. *Mol. Biol. Evol.* **32**, 1449–1455 (2015).
31. Hane, J. K. et al. A novel mode of chromosomal evolution peculiar to filamentous Ascomycete fungi. *Genome Biol.* **12**, R45 (2011).
32. Siverio, J. M. Assimilation of nitrate by yeasts. *FEMS Microbiol. Rev.* **26**, 277–284 (2002).
33. Scazzocchio, C. Fungal biology in the post-genomic era. *Fungal Biol. Biotechnol.* **1**, 7 (2014).
34. Yu, J., Chang, P., Bhatnagar, D. & Cleveland, T. E. Cloning of a sugar utilization gene cluster in *Aspergillus parasiticus*. *Biochim. Biophys. Acta* **1493**, 211–214 (2000).
35. Schrettl, M. et al. Self-protection against gliotoxin — a component of the gliotoxin biosynthetic cluster, GliT, completely protects *Aspergillus fumigatus* against exogenous gliotoxin. *PLOS Pathog.* **6**, e1000952 (2010).
This study reports that a gene within the gliotoxin MGC is necessary and sufficient for conferring protection against gliotoxin poisoning, thus providing a competitive advantage for organisms carrying it when grown in the presence of competitors that produce gliotoxin.
36. Wiemann, P. et al. Biosynthesis of the red pigment bikaverin in *Fusarium fujikuroi*: genes, their function and regulation. *Mol. Microbiol.* **72**, 931–946 (2009).
37. Hittinger, C. T., Rokas, A. & Carroll, S. B. Parallel inactivation of multiple *GAL* pathway genes and ecological diversification in yeasts. *Proc. Natl Acad. Sci. USA* **101**, 14144–14149 (2004).
38. Martchenko, M., Levitin, A., Hogues, H., Nantel, A. & Whiteway, M. Transcriptional rewiring of fungal galactose metabolism circuitry. *Curr. Biol.* **17**, 1007–1013 (2007).
39. Lim, F. Y. et al. Genome-based cluster deletion reveals an endocrocin biosynthetic pathway in *Aspergillus fumigatus*. *Appl. Environ. Microbiol.* **78**, 4117–4125 (2012).
40. Szewczyk, E. et al. Identification and characterization of the asperthecin gene cluster of *Aspergillus nidulans*. *Appl. Environ. Microbiol.* **74**, 7607–7612 (2008).
41. Chiang, Y. M. et al. Molecular genetic mining of the *Aspergillus* secondary metabolome: discovery of the emericellamide biosynthetic pathway. *Chem. Biol.* **15**, 527–532 (2008).
42. Bushley, K. E. et al. The genome of *Tolypocladium inflatum*: evolution, organization, and expression of the cyclosporin biosynthetic gene cluster. *PLOS Genet.* **9**, e1003496 (2013).
43. Inderbitzin, P., Asvarak, T. & Turgeon, B. G. Six new genes required for production of T-toxin, a polyketide determinant of high virulence of *Cochliobolus heterostrophus* to maize. *Mol. Plant Microbe Interact.* **23**, 458–472 (2010).
44. Yin, W. B. et al. A nonribosomal peptide synthetase-derived iron(III) complex from the pathogenic fungus *Aspergillus fumigatus*. *J. Am. Chem. Soc.* **135**, 2064–2067 (2013).
45. Umemura, M. et al. Characterization of the biosynthetic gene cluster for the ribosomally synthesized cyclic peptide ustiloxin B in *Aspergillus flavus*. *Fungal Genet. Biol.* **68**, 23–30 (2014).
46. Gutierrez, S., Fierro, F., Casqueiro, J. & Martin, J. F. Gene organization and plasticity of the beta-lactam genes in different filamentous fungi. *Antonie Van Leeuwenhoek* **75**, 81–94 (1999).
47. Lo, H. C. et al. Two separate gene clusters encode the biosynthetic pathway for the meroterpenoids austinol and dehydroaustinol in *Aspergillus nidulans*. *J. Am. Chem. Soc.* **134**, 4709–4720 (2012).
48. Proctor, R. H., McCormick, S. P., Alexander, N. J. & Desjardins, A. E. Evidence that a secondary metabolic biosynthetic gene cluster has grown by gene relocation during evolution of the filamentous fungus *Fusarium*. *Mol. Microbiol.* **74**, 1128–1142 (2009).
49. Nicholson, M. J. et al. Identification of two aflatrem biosynthesis gene loci in *Aspergillus flavus* and metabolic engineering of *Penicillium paxilli* to elucidate their function. *Appl. Environ. Microbiol.* **75**, 7469–7481 (2009).
50. Wiemann, P. et al. Prototype of an intertwined secondary-metabolite supercluster. *Proc. Natl Acad. Sci. USA* **110**, 17065–17070 (2013).
This study is the first to report the existence of intertwined MGCs responsible for the production of at least two distinct secondary metabolites in fungal genomes.
51. Lind, A. L. et al. Drivers of genetic diversity in secondary metabolic gene clusters within a fungal species. *PLOS Biol.* **15**, e2003583 (2017).
This is the first study, together with that by Gibbons et al. (2012), to report that two non-homologous MGCs occupy the same locus (that is, an idiomorph polymorphism) and is the first to describe patterns of genome-wide variation of secondary MGCs in a fungal species.
52. Medini, D., Donati, C., Tettelin, H., Masignani, V. & Rappuoli, R. The microbial pan-genome. *Curr. Opin. Genet. Dev.* **15**, 589–594 (2005).
53. Chang, P. K., Horn, B. W. & Dorner, J. W. Sequence breakpoints in the aflatoxin biosynthesis gene cluster and flanking regions in nonaflatoxigenic *Aspergillus flavus* isolates. *Fungal Genet. Biol.* **42**, 914–923 (2005).
54. Wiemann, P. et al. Deciphering the cryptic genome: genome-wide analyses of the rice pathogen *Fusarium fujikuroi* reveal complex regulation of secondary metabolism and novel metabolites. *PLOS Pathog.* **9**, e1003475 (2013).
55. Chiara, M. et al. Genome sequencing of multiple isolates highlights subtelomeric genomic diversity within *Fusarium fujikuroi*. *Genome Biol. Evol.* **7**, 3062–3069 (2015).
56. Castella, G., Bragulat, M. R., Puig, L., Sanseverino, W. & Cabanes, F. J. Genomic diversity in ochratoxigenic and non ochratoxigenic strains of *Aspergillus carbonarius*. *Sci. Rep.* **8**, 5439 (2018).
57. Hartmann, F. E. & Croll, D. Distinct trajectories of massive recent gene gains and losses in populations of a microbial eukaryotic pathogen. *Mol. Biol. Evol.* **34**, 2808–2822 (2017).
58. Campbell, M. A., Rokas, A. & Slot, J. C. Horizontal transfer and death of a fungal secondary metabolic gene cluster. *Genome Biol. Evol.* **4**, 289–293 (2012).
59. Campbell, M. A., Staats, M., van Kan, J. L. A., Rokas, A. & Slot, J. C. Repeated loss of an anciently horizontally transferred gene cluster in *Botrytis*. *Mycologia* **105**, 1126–1134 (2013).
60. Schumacher, J. et al. A functional bikaverin biosynthesis gene cluster in rare strains of *Botrytis cinerea* is positively controlled by VELVET. *PLOS One* **8**, e53729 (2013).
61. Hittinger, C. T. et al. Remarkably ancient balanced polymorphisms in a multi-locus gene network. *Nature* **464**, 54–58 (2010).
62. Gibbons, J. G. et al. The evolutionary imprint of domestication on genome variation and function of the filamentous fungus *Aspergillus oryzae*. *Curr. Biol.* **22**, 1403–1409 (2012).
Together with the study by Lind et al., these are the first studies to report that two non-homologous MGCs occupy the same locus (that is, an idiomorph polymorphism) in the genome of a fungal species.

63. Keller-Seitz, M. U. et al. Transcriptional response of yeast to aflatoxin B1: recombinational repair involving RAD51 and RAD1. *Mol. Biol. Cell* **15**, 4321–4336 (2004).
64. Fedorova, N. D. et al. Genomic islands in the pathogenic filamentous fungus *Aspergillus fumigatus*. *PLOS Genet.* **4**, e1000046 (2008).
65. Metzenberg, R. L. & Glass, N. L. Mating type and mating strategies in *Neurospora*. *Bioessays* **12**, 53–59 (1990).
66. Mukherjee, P. K., Horwitz, B. A. & Kenerley, C. M. Secondary metabolism in *Trichoderma*—a genomic perspective. *Microbiology* **158**, 35–45 (2012).
67. Rank, C. et al. Distribution of sterigmatocystin in filamentous fungi. *Fungal Biol.* **115**, 406–420 (2011).
68. Khaldi, N. et al. SMURF: genomic mapping of fungal secondary metabolite clusters. *Fungal Genet. Biol.* **47**, 736–741 (2010).
69. de Vries, R. P. et al. Comparative genomics reveals high biological diversity and specific adaptations in the industrially and medically important fungal genus *Aspergillus*. *Genome Biol.* **18**, 28 (2017).
70. Riley, R. et al. Comparative genomics of biotechnologically important yeasts. *Proc. Natl Acad. Sci. USA* **113**, 9882–9887 (2016).
71. Khaldi, N., Collemare, J., Lebrun, M. H. & Wolfe, K. H. Evidence for horizontal transfer of a secondary metabolite gene cluster between fungi. *Genome Biol.* **9**, R18 (2008).
72. Reynolds, H. T. et al. Differential retention of gene functions in a secondary metabolite cluster. *Mol. Biol. Evol.* **34**, 2002–2015 (2017).
73. Zhang, H., Rokas, A. & Slot, J. C. Two different secondary metabolism gene clusters occupied the same ancestral locus in fungal dermatophytes of the Arthrodermataceae. *PLOS One* **7**, e41903 (2012).
74. Dalal, C. K. et al. Transcriptional rewiring over evolutionary timescales changes quantitative and qualitative properties of gene expression. *elife* **5**, e18981 (2016).
75. Proctor, R. H. et al. Evolution of structural diversity of trichothecenes, a family of toxins produced by plant pathogenic and entomopathogenic fungi. *PLOS Pathog.* **14**, e1006946 (2018).
76. Carbone, I., Ramirez-Prado, J. H., Jakobek, J. L. & Horn, B. W. Gene duplication, modularity and adaptation in the evolution of the aflatoxin gene cluster. *BMC Evol. Biol.* **7**, 111 (2007).
77. Slot, J. C. & Rokas, A. Horizontal transfer of a large and highly toxic secondary metabolic gene cluster between fungi. *Curr. Biol.* **21**, 134–139 (2011).
This study reports the horizontal transfer of a large MGC (at least 57 kb of sequence, 24 genes) involved in the making of the mycotoxin sterigmatocystin between fungi.
78. Ehrlich, K. C., Chang, P. K., Yu, J. & Cotty, P. J. Aflatoxin biosynthesis cluster gene cypA is required for G aflatoxin formation. *Appl. Environ. Microbiol.* **70**, 6518–6524 (2004).
79. Proctor, R. H., Busman, M., Seo, J. A., Lee, Y. W. & Plattner, R. D. A fumonisin biosynthetic gene cluster in *Fusarium oxysporum* strain O-1890 and the genetic basis for B versus C fumonisin production. *Fungal Genet. Biol.* **45**, 1016–1026 (2008).
80. Rosewich, U. L. & Kistler, H. C. Role of horizontal gene transfer in the evolution of fungi. *Annu. Rev. Phytopathol.* **38**, 325–363 (2000).
81. Wisecaver, J. H. & Rokas, A. Fungal metabolic gene clusters-caravans traveling across genomes and environments. *Front. Microbiol.* **6**, 161 (2015).
82. Reynolds, H. T. et al. Horizontal gene cluster transfer increased hallucinogenic mushroom diversity. *Evol. Lett.* **2**, 88–101 (2018).
83. Patron, N. J. et al. Origin and distribution of epipolythiodioxopiperazine (ETP) gene clusters in filamentous ascomycetes. *BMC Evol. Biol.* **7**, 174 (2007).
84. Matasyoh, J. C., Dittrich, B., Schueffler, A. & Laatsch, H. Larvicidal activity of metabolites from the endophytic *Podospora* sp. against the malaria vector *Anopheles gambiae*. *Parasitol. Res.* **108**, 561–566 (2011).
85. Matsuzawa, T. et al. New insights into galactose metabolism by *Schizosaccharomyces pombe*: isolation and characterization of a galactose-assimilating mutant. *J. Biosci. Bioeng.* **111**, 158–166 (2011).
86. Ortiz, J. F. & Rokas, A. CTDGFinder: a novel homology-based algorithm for identifying closely spaced clusters of tandemly duplicated genes. *Mol. Biol. Evol.* **34**, 215–229 (2017).
87. Nei, M. Modification of linkage intensity by natural selection. *Genetics* **57**, 625–641 (1967).
88. Ward, T. J., Bielawski, J. P., Kistler, H. C., Sullivan, E. & O'Donnell, K. Ancestral polymorphism and adaptive evolution in the trichothecene mycotoxin gene cluster of phytopathogenic *Fusarium*. *Proc. Natl Acad. Sci. USA* **99**, 9278–9283 (2002).
This is the first study to report evidence for trans-specific polymorphism in the gene cluster of the trichothecene mycotoxin, which is maintained by balancing selection acting on chemotype differences.
89. Carbone, I., Jakobek, J. L., Ramirez-Prado, J. H. & Horn, B. W. Recombination, balancing selection and adaptive evolution in the aflatoxin gene cluster of *Aspergillus parasiticus*. *Mol. Ecol.* **16**, 4401–4417 (2007).
90. Hartmann, F. E., McDonald, B. A. & Croll, D. Genome-wide evidence for divergent selection between populations of a major agricultural pathogen. *Mol. Ecol.* **27**, 2725–2741 (2018).
91. Grandaubert, J., Dutheil, J. Y. & Stukenbrock, E. H. The genomic determinants of adaptive evolution in a fungal pathogen. Preprint at *bioRxiv*. https://doi.org/10.1101/176727 (2018).
92. Morris, J. J., Lenski, R. E. & Zinser, E. R. The Black Queen hypothesis: evolution of dependencies through adaptive gene loss. *MBio* **3**, e00036–12 (2012).
93. Hurst, L. D., Pal, C. & Lercher, M. J. The evolutionary dynamics of eukaryotic gene order. *Nat. Rev. Genet.* **5**, 299–310 (2004).
94. Gibbons, J. G. et al. Global transcriptome changes underlying colony growth in the opportunistic human pathogen *Aspergillus fumigatus*. *Eukaryot. Cell* **11**, 68–78 (2012).
95. Lind, A. L., Smith, T. D., Saterlee, T., Calvo, A. M. & Rokas, A. Regulation of secondary metabolism by the Velvet complex is temperature-responsive in *Aspergillus*. *G3* **6**, 4023–4033 (2016).
96. Yu, J. J. et al. Tight control of mycotoxin biosynthesis gene expression in *Aspergillus flavus* by temperature as revealed by RNA-Seq. *FEMS Microbiol. Lett.* **322**, 145–149 (2011).
97. Andersen, M. R. et al. Accurate prediction of secondary metabolite gene clusters in filamentous fungi. *Proc. Natl Acad. Sci. USA* **110**, E99–E107 (2013).
98. Lawler, K., Hammond-Kosack, K., Brazma, A. & Coulson, R. M. Genomic clustering and co-regulation of transcriptional networks in the pathogenic fungus *Fusarium graminearum*. *BMC Syst. Biol.* **7**, 52 (2013).
99. Wisecaver, J. H. et al. A global coexpression network approach for connecting genes to specialized metabolic pathways in plants. *Plant Cell* **29**, 944–959 (2017).
100. Kensche, P. R., Oti, M., Dutilh, B. E. & Huynen, M. A. Conservation of divergent transcription in fungi. *Trends Genet.* **24**, 207–211 (2008).
101. Fernandes, M., Keller, N. P. & Adams, T. H. Sequence-specific binding by *Aspergillus nidulans* AflR, a C6 zinc cluster protein regulating mycotoxin biosynthesis. *Mol. Microbiol.* **28**, 1355–1365 (1998).
102. Gacek, A. & Strauss, J. The chromatin code of fungal secondary metabolite gene clusters. *Appl. Microbiol. Biotechnol.* **95**, 1389–1404 (2012).
103. Moore, G. G., Singh, R., Horn, B. W. & Carbone, I. Recombination and lineage-specific gene loss in the aflatoxin gene cluster of *Aspergillus flavus*. *Mol. Ecol.* **18**, 4870–4887 (2009).
Extensive analysis of patterns of linkage disequilibrium in the aflatoxin gene cluster of *A. flavus* suggests that recombination events are unevenly distributed across the cluster, including the absence of recombination for the subcluster of genes acting early in the aflatoxin synthesis pathway.
104. Olarte, R. A. et al. Effect of sexual recombination on population diversity in aflatoxin production by *Aspergillus flavus* and evidence for cryptic heterokaryosis. *Mol. Ecol.* **21**, 1453–1476 (2012).
105. Walton, J. D. Horizontal gene transfer and the evolution of secondary metabolite gene clusters in fungi: an hypothesis. *Fungal Genet. Biol.* **30**, 167–171 (2000).
106. Lang, G. I. & Botstein, D. A test of the coordinated expression hypothesis for the origin and maintenance of the *GAL* cluster in yeast. *PLOS One* **6**, e25290 (2011).
This study reports that disruption of the clustering in the galactose pathway does not reduce fitness, arguing against the hypothesis that gene clustering increases fitness because it enables greater coordination of gene expression.
107. Naseeb, S. & Delneri, D. Impact of chromosomal inversions on the yeast DAL cluster. *PLOS One* **7**, e42022 (2012).
108. Leslie, N. D. Insights into the pathogenesis of galactosemia. *Annu. Rev. Nutr.* **23**, 59–80 (2003).
109. Alam, M. K. & Kaminskyj, S. G. W. *Aspergillus* galactose metabolism is more complex than that of *Saccharomyces*: the story of GalD(GAL7) and GalE(GAL1). *Botany* **91**, 467–477 (2013).
110. Slot, J. C. Fungal gene cluster diversity and evolution. *Adv. Genet.* **100**, 141–178 (2017).
111. Wisecaver, J. H., Slot, J. C. & Rokas, A. The evolution of fungal metabolic pathways. *PLOS Genet.* **10**, e1004816 (2014).
112. Gibbons, J. G. & Rokas, A. The function and evolution of the *Aspergillus* genome. *Trends Microbiol.* **21**, 14–22 (2013).
113. Lynch, M. *The Origins of Genome Architecture*. (Sinauer, 2007).
114. Lewontin, R. C. On measures of gametic disequilibrium. *Genetics* **120**, 849–852 (1988).
115. Fridovich-Keil, J. L. Galactosemia: the good, the bad, and the unknown. *J. Cell. Physiol.* **209**, 701–705 (2006).
116. Mumma, J. O. et al. Distinct roles of galactose-1P in galactose-mediated growth arrest of yeast deficient in galactose-1P uridylyltransferase (GALT) and UDP-galactose 4′-epimerase (GALE). *Mol. Genet. Metab.* **93**, 160–171 (2008).
117. del Campo, J. et al. The others: our biased perspective of eukaryotic genomes. *Trends Ecol. Evol.* **29**, 252–259 (2014).
118. Loftus, B. et al. The genome of the protist parasite *Entamoeba histolytica*. *Nature* **433**, 865–868 (2005).
119. Brayton, K. A. et al. Genome sequence of *Babesia bovis* and comparative analysis of apicomplexan hemoprotozoa. *PLOS Pathog.* **3**, 1401–1413 (2007).
120. Corradi, N., Pombert, J. F., Farinelli, L., Didier, E. S. & Keeling, P. J. The complete sequence of the smallest known nuclear genome from the microsporidian *Encephalitozoon intestinalis*. *Nat. Commun.* **1**, 77 (2010).
121. Eichinger, L. et al. The genome of the social amoeba *Dictyostelium discoideum*. *Nature* **435**, 43–57 (2005).
122. Zhu, G. et al. *Cryptosporidium parvum*: the first protist known to encode a putative polyketide synthase. *Gene* **298**, 79–89 (2002).
123. John, U. et al. Novel insights into evolution of protistan polyketide synthases through phylogenomic analysis. *Protist* **159**, 21–30 (2008).
124. Monroe, E. A., Johnson, J. G., Wang, Z. H., Pierce, R. K. & Van Dolah, F. M. Characterization and expression of nuclear-encoded polyketide synthases in the brevetoxin-producing dinoflagellate. *Karenia brevis*. *J. Phycol.* **46**, 541–552 (2010).
125. Wang, H., Fewer, D. P., Holm, L., Rouhiainen, L. & Sivonen, K. Atlas of nonribosomal peptide and polyketide biosynthetic pathways reveals common occurrence of nonmodular enzymes. *Proc. Natl Acad. Sci. USA* **111**, 9259–9264 (2014).
126. Rokas, A. & Hittinger, C. T. Transcriptional rewiring: the proof is in the eating. *Curr. Biol.* **17**, R626–628 (2007).
127. Johnston, M. A model fungal gene regulatory mechanism: the *GAL* genes of *Saccharomyces cerevisiae*. *Microbiol. Rev.* **51**, 458–476 (1987).
128. Rubio-Texeira, M. A comparative analysis of the *GAL* genetic switch between not-so-distant cousins: *Saccharomyces cerevisiae* versus *Kluyveromyces lactis*. *FEMS Yeast Res.* **5**, 1115–1128 (2005).
129. Wolfe, K. H. et al. Clade- and species-specific features of genome evolution in the Saccharomycetaceae. *FEMS Yeast Res.* **15**, fov035 (2015).
130. Moyrand, F., Fontaine, T. & Janbon, G. Systematic capsule gene disruption reveals the central role of galactose metabolism on *Cryptococcus neoformans* virulence. *Mol. Microbiol.* **64**, 771–781 (2007).
131. Moyrand, F., Lafontaine, I., Fontaine, T. & Janbon, G. UGE1 and UGE2 regulate the UDP-glucose/UDP-galactose equilibrium in *Cryptococcus neoformans*. *Eukaryot. Cell* **7**, 2069–2077 (2008).
132. Christensen, U. et al. Unique regulatory mechanism for d-galactose utilization in *Aspergillus nidulans*. *Appl. Environ. Microbiol.* **77**, 7084–7087 (2011).
133. Boycheva, S., Daviet, L., Wolfender, J. L. & Fitzpatrick, T. B. The rise of operon-like gene clusters in plants. *Trends Plant Sci.* **19**, 447–459 (2014).
134. Nutzmann, H. W., Huang, A. & Osbourn, A. Plant metabolic clusters — from genetics to genomics. *New Phytol.* **211**, 771–789 (2016).

135. Frey, M. et al. Analysis of a chemical plant defense mechanism in grasses. *Science* **277**, 696–699 (1997).
136. Handrick, V. et al. Biosynthesis of 8-*O*-methylated benzoxazinoid defense compounds in maize. *Plant Cell* **28**, 1682–1700 (2016).
137. Schneider, L. M. et al. The Cer-cqu gene cluster determines three key players in a beta-diketone synthase polyketide pathway synthesizing aliphatics in epicuticular waxes. *J. Exp. Bot.* **67**, 2715–2730 (2016).
138. Olsen, K. M. & Small, L. L. Micro- and macroevolutionary adaptation through repeated loss of a complete metabolic pathway. *New Phytol.* **219**, 757–766 (2018).
139. Takos, A. M. et al. Genomic clustering of cyanogenic glucoside biosynthetic genes aids their identification in *Lotus japonicus* and suggests the repeated evolution of this chemical defence pathway. *Plant J.* **68**, 273–286 (2011).
140. Winzer, T. et al. A *Papaver somniferum* 10-gene cluster for synthesis of the anticancer alkaloid noscapine. *Science* **336**, 1704–1708 (2012).
141. Itkin, M. et al. Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes. *Science* **341**, 175–179 (2013).
142. Qi, X. et al. A gene cluster for secondary metabolism in oat: implications for the evolution of metabolic diversity in plants. *Proc. Natl Acad. Sci. USA* **101**, 8233–8238 (2004).
143. Shimura, K. et al. Identification of a biosynthetic gene cluster in rice for momilactones. *J. Biol. Chem.* **282**, 34013–34018 (2007).
144. Field, B. & Osbourn, A. E. Metabolic diversification — independent assembly of operon-like gene clusters in different plants. *Science* **320**, 543–547 (2008).
145. Swaminathan, S., Morrone, D., Wang, Q., Fulton, D. B. & Peters, R. J. CYP76M7 is an ent-cassadiene C11alpha-hydroxylase defining a second multifunctional diterpenoid biosynthetic gene cluster in rice. *Plant Cell* **21**, 3315–3325 (2009).
146. Field, B. et al. Formation of plant metabolic gene clusters within dynamic chromosomal regions. *Proc. Natl Acad. Sci. USA* **108**, 16116–16121 (2011).
147. Krokida, A. et al. A metabolic gene cluster in Lotus japonicus discloses novel enzyme functions and products in triterpene biosynthesis. *New Phytol.* **200**, 675–690 (2013).
148. Matsuba, Y. et al. Evolution of a complex locus for terpene biosynthesis in *Solanum*. *Plant Cell* **25**, 2022–2036 (2013).
149. King, A. J., Brown, G. D., Gilday, A. D., Larson, T. R. & Graham, I. A. Production of bioactive diterpenoids in the Euphorbiaceae depends on evolutionarily conserved gene clusters. *Plant Cell* **26**, 3286–3298 (2014).
150. Shang, Y. et al. Biosynthesis, regulation, and domestication of bitterness in cucumber. *Science* **346**, 1084–1088 (2014).
151. Chu, H. Y., Wegel, E. & Osbourn, A. From hormones to secondary metabolism: the emergence of metabolic gene clusters in plants. *Plant J.* **66**, 66–79 (2011).
152. Boutanaev, A. M. et al. Investigation of terpene diversification across multiple sequenced plant genomes. *Proc. Natl Acad. Sci. USA* **112**, E81–E88 (2015).
153. Schlapfer, P. et al. Genome-wide prediction of metabolic enzymes, pathways, and gene clusters in plants. *Plant Physiol.* **173**, 2041–2059 (2017).
154. Topfer, N., Fuchs, L. M. & Aharoni, A. The PhytoClust tool for metabolic gene clusters discovery in plant genomes. *Nucleic Acids Res.* **45**, 7049–7063 (2017).
155. Kautsar, S. A., Suarez Duran, H. G., Blin, K., Osbourn, A. & Medema, M. H. plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Res.* **45**, W55–W63 (2017).
156. Kliebenstein, D. J. & Osbourn, A. Making new molecules — evolution of pathways for novel metabolites in plants. *Curr. Opin. Plant Biol.* **15**, 415–423 (2012).
157. Itkin, M. et al. The biosynthetic pathway of the nonsugar, high-intensity sweetener mogroside V from *Siraitia grosvenorii*. *Proc. Natl Acad. Sci. USA* **113**, E7619–E7628 (2016).
158. Klein, A. P. & Sattely, E. S. Biosynthesis of cabbage phytoalexins from indole glucosinolate. *Proc. Natl Acad. Sci. USA* **114**, 1910–1915 (2017).
159. Gladyshev, E. A., Meselson, M. & Arkhipova, I. R. Massive horizontal gene transfer in bdelloid rotifers. *Science* **320**, 1210–1213 (2008).
160. Moran, N. A. & Jarvik, T. Lateral transfer of genes from fungi underlies carotenoid production in aphids. *Science* **328**, 624–627 (2010).
161. Derelle, E. et al. Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. *Proc. Natl Acad. Sci. USA* **103**, 11647–11652 (2006).
162. Fernandez, E. & Galvan, A. Inorganic nitrogen assimilation in *Chlamydomonas*. *J. Exp. Bot.* **58**, 2279–2287 (2007).
163. McDonald, S. M., Plant, J. N. & Worden, A. Z. The mixed lineage nature of nitrogen transport and assimilation in marine eukaryotic phytoplankton: a case study of *Micromonas*. *Mol. Biol. Evol.* **27**, 2268–2283 (2010).
164. Foflonker, F. et al. The unexpected extremophile: tolerance to fluctuating salinity in the green alga *Picochlorum*. *Algal Res.* **16**, 465–472 (2016).
165. Vallon, O. & Spalding, M. H. in *The Chlamydomonas Sourcebook Vol. 2* (ed. Stern, D.) 115–158 (Academic Press, 2009).
166. Foflonker, F. et al. Genome of the halotolerant green alga *Picochlorum* sp. reveals strategies for thriving under fluctuating environmental conditions. *Environ. Microbiol.* **17**, 412–426 (2015).
167. Medema, M. H. et al. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res.* **39**, W339–W346 (2011).

Acknowledgements
The authors thank past and present members of the Rokas laboratory, particularly J. Slot, J. Gibbons, M. Mead, K. McGary and J. Steenwyk, as well as long-time collaborator C. T. Hittinger for discussions over the years on the evolution of metabolic gene clusters in fungi. Research in the Rokas laboratory has been supported by the National Science Foundation, the Searle Scholars Program, the Guggenheim Foundation, the Burroughs Wellcome Trust, the National Institutes of Health, the Beckman Scholars Program and the March of Dimes.

Author contributions
A.R., J.H.W. and A.L.L. wrote the article, researched data for the article, made substantial contributions to discussions of the content and reviewed and/or edited the manuscript before submission.

Competing interests
The authors declare no competing interests.

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

Reviewer information
Nature Reviews Microbiology thanks B. McDonald and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Supplementary information
Supplementary information is available for this paper at https://doi.org/10.1038/s41579-018-0075-3.