Cellular and Molecular Life Sciences

REVIEW

Operons

Anne E. Osbourn · Ben Field

Abstract Operons (clusters of co-regulated genes with related functions) are common features of bacterial genomes. More recently, functional gene clustering has been reported in eukaryotes, from yeasts to filamentous fungi, plants, and animals. Gene clusters can consist of paralogous genes that have most likely arisen by gene duplication. However, there are now many examples of eukaryotic gene clusters that contain functionally related but non-homologous genes and that represent functional gene organizations with operon-like features (physical clustering and co-regulation). These include gene clusters for use of different carbon and nitrogen sources in yeasts, for production of antibiotics, toxins, and virulence determinants in filamentous fungi, for production of defense compounds in plants, and for innate and adaptive immunity in animals (the major histocompatibility locus). The aim of this article is to review features of functional gene clusters in prokaryotes and eukaryotes and the significance of clustering for effective function.

A. E. Osbourn (✉) · B. Field
Department of Metabolic Biology, John Innes Centre,
Colney Lane, Norwich NR4 7UH, UK
e-mail: anne.osbourn@bbsrc.ac.uk

Present Address:
B. Field
UMR 6191, Laboratoire de Génétique et Biophysique des
Plantes, iBEB, UMR 6191 CNRS/CEA/Université
Aix-Marseille, Faculté des Sciences de Luminy,
Case 901, 163 av de Luminy, 13009 Marseille, France

Introduction

Operons (clusters of co-regulated genes with related functions) are a well-known feature of prokaryotic genomes. Archeal and bacterial genomes generally contain a small number of highly conserved operons and a much larger number of unique or rare ones [1]. Functional gene clustering also occurs in eukaryotes, from yeasts to filamentous fungi, mammals, nematodes, and plants [2]. The members of these eukaryotic gene clusters contribute to a common function but do not usually share sequence similarity. These gene clusters therefore represent functional gene organizations with operon-like features (physical clustering and co-regulation), although the genes are not usually transcribed as a single mRNA as is the case in prokaryotes. This article reviews facets of genome organization in prokaryotes and eukaryotes that are of relevance for understanding the significance of the establishment, maintenance, and dissipation of functional gene clusters and the evolutionary forces that shape genome architecture.

Classical operons

The term "operon" was coined by Jacob and Monod [3–5], who characterized the first defined classical operon, the *lac* operon, in *Escherichia coli*. The *lac* operon consists of three structural genes that are required for lactose utilisation, *lacZ*, *lacY*, and *lacA* (Fig. 1). These genes encode a β-galactosidase, a *lac* permease, and a transacetylase, respectively. The first step in lactose metabolism is catalyzed by β-galactosidase, which cleaves lactose into galactose and glucose. Lactose is then taken up by the transmembrane protein, *lac* permease. The transacetylase transfers an acetyl group from coenzyme A (CoA) to the

lacZYA operon

Fig. 1 The *Escherichia coli lac* regulon. *lacI* encodes the *lac* repressor, which in its active form inhibits the transcription of the structural genes by binding an operator. The genes within the *lac* operon (*lacZYA*) are transcribed as a single mRNA. Expression is induced by lactose under conditions of carbon starvation. Expression of *lacZYA* is positively regulated by an activator whose production is controlled by the concentration of extracellular glucose

hydroxyl group of the galactosides. The regulatory gene *lacI* encodes the *lac* repressor, which in its active form inhibits the transcription of the structural genes by binding an operator. The *lacZYA* genes are co-ordinately regulated, and expression is induced by lactose under conditions of carbon starvation. Regulation is mediated by co-transcription, the *lacZYA* genes being transcribed in a single mRNA. Numerous examples of bacterial operons have since been described [6–8]. Operons usually code for genes in the same functional pathway. However, there are also examples of operons that contain genes with no obvious functional relationship. Genes within operons of this kind may be required under the same environmental conditions despite being involved in different pathways [9].

Although the *lacI* repressor gene is able to regulate expression of the *lacZYA* genes when placed anywhere in the chromosome, this gene is normally positioned immediately upstream of *lacZYA*. Co-localization of the regulatory gene with the genes that it regulates has been suggested to be a "selfish" property of the *lac* gene cluster, since such organization may increase the fitness of this group of genes by allowing horizontal as well as vertical inheritance [10, 11]. Operons are certainly known to undergo horizontal gene transfer (HGT) [10, 12, 13]. However, essential genes are commonly found in operons, as are other genes that are not known to be transmitted by HGT, providing evidence against the selfish operon theory [14, 15]. The selfish cluster theory also does not explain the many operons that contain functionally unrelated genes [9]. The emergence of new operons will require not only the coming together of genes but also the establishment and maintenance of co-ordinate regulation of these genes. A more likely explanation for the existence of operons is concerned with regulation.

The regulatory model provides several potential explanations for the existence of operons [15]. It has been argued that co-regulation could be evolved more easily by modifying two independent promoters than by placing two genes in proximity [10]. A counter-argument to this is that, for complex regulation, an operon with one complex promoter might be expected to arise more readily than would two independent complex promoters [12]. The dependence

of several genes on a single regulatory sequence is expected to put this sequence under stronger selection, so allowing for the emergence of more complex regulatory strategies [15]. The observation that operons do tend to have more complex conserved regulatory sequences than individually transcribed genes is consistent with this hypothesis [9, 12, 16]. Genes within new operons are significantly less likely to be optimally spaced when compared to old operons, which is consistent with the idea that canonical spacings form by deletion after the operon has already formed [9]. In laboratory experiments, the expression level of the *lac* operon evolves to optimality in a few hundred generations [17]. Thus, changes in operon spacing could reflect fine tuning of expression levels.

Transcription of genes in a single transcript is expected to diminish gene expression noise and ensure more precise stoichiometry [15]. The most highly conserved operons do tend to code for components of protein complexes [9, 14, 18, 19]. However, there are many examples of operons that do not code for protein complexes [9]. In addition, only a few percent of known protein–protein interactions involve genes encoded by the same operon [20]. Furthermore, the optimal expression level of each gene is not the same for all genes in an operon [15].

Rapid and reliable gene regulation may require the transcription factor gene to be in close proximity to the sites within the genome to which it binds (the rapid search hypothesis [2–24]). In prokaryotes, transcription and translation are coupled spatially and temporally. Such organization would therefore provide for synthesis of transcription factors near the genes that they regulate, enabling rapid binding of co-localized sites and tight co-ordinate regulation. Computer simulations have been used to address the relationship between genome organization and biophysical constraints and have provided evidence to support the rapid search hypothesis [21]. Transcription–translational coupling may also facilitate co-ordinate downstream processes such as assembly of protein complexes through co-translational folding [25], and cell compartmentalization [26].

The birth and death of operons is a dynamic process. The complete genome sequences of over 1,000 prokaryotes have now been determined. This wealth of sequence information has enabled genome-wide studies of bacterial genomes to be carried out, and the evolutionary life-cycle of operons to be inferred by comparison of related bacterial species [9]. HGT represents one source of operons, and is the primary origin of these functional gene clusters as predicted by the selfish operon model [15]. In this case, the genes within the operon have homology with genes from other sequenced organisms and can readily be identified as the products of a HGT event. In addition, new operons are highly enriched for "ORFan" genes—genes that lack

identifiable homologues outside a group of related bacteria. Most ORFans are functional protein-encoding genes that contribute to the fitness of the organism (i.e., are under purifying selection) and were probably acquired from bacteriophage [27]. Such ORFan genes tend to be located 3′ of native genes in operons (Fig. 2). This arrangement may be selected for because the ORFan gene is transcribed from a native promoter without perturbing the expression of the native gene. The grouping of more than one ORFan gene in an operon may indicate that the ORFan genes have been introduced by HGT from a source that has not yet been sequenced. Finally, operons consisting of native genes can form without HGT. These new operons can arise as a consequence of rearrangements that bring more distant genes into proximity, or by deletion of intervening genes (Fig. 2).

Because few operons are conserved across all or even most bacteria, it is clear that after operons form many of them "die". Operons could be lost by the deletion of one/more genes or alternatively by splitting the operon apart. Since operon formation often brings functionally related genes together, it seems unlikely to be a neutral process. If operon formation is driven by gene expression, then it should be associated with changes in the expression patterns of the constituent genes. Studies of the expression patterns of genes in *E. coli* operons and of orthologous genes in "not-yet" operons in the related bacterium *Shewanella oneidensis* MR-1 have provided compelling evidence that operon formation has a major effect on gene expression patterns [9]. Similarly, "dead" operons are significantly less co-expressed than live operons but significantly more co-expressed than random pairs of genes. Thus, operon destruction also has a major effect on gene expression patterns, but it does not entirely eliminate the similarity of expression. The turnover of operon structure may accompany switching between constitutive and inducible expression. Although constitutive expression may seem wasteful and hence deleterious, favorable gene combinations cannot be selected for unless the genes are

expressed. Constitutive gene expression, even at very low levels, would logically be expected to be necessary to enable the selection-mediated garnering of beneficial combination of genes into operons with subsequent fine tuning of expression. While genes within the same operon are under much stronger selection to remain together than genes that are in different operons [28, 29], there is also evidence of weaker selection for high level interactions between operons, some of which are so widely conserved that they are known as super- or uberoperons [21, 30].

Secondary metabolism in filamentous fungi

Filamentous fungi produce a huge array of secondary metabolites (sometimes also referred to as natural products). These compounds are commonly synthesized by groups of genes that form metabolic gene clusters [32–34]. The genes within these clusters are physically linked and co-regulated, but unlike bacterial operons they are not transcribed as a single mRNA. Nevertheless, these fungal gene clusters represent functional gene organizations with operon-like features (physical clustering and co-regulation). Examples include gene clusters for important pharmaceuticals (such as the *β*-lactam antibiotics, penicillin, and cephalosporin), the anti-hypercholesterolaemic agent lovastatin, ergopeptines, and carcinogenic toxins (aflatoxin and sterigmatocystin). Secondary metabolites are also important in mediating interactions between fungal pathogens and their hosts. For example, an unknown metabolite synthesized by the ACE1 gene cluster in the rice blast fungus *Magnaporthe grisea* is involved in recognition of particular rice cultivars [35, 36], and the host selective HC-toxin produced by the corn pathogen *Cochliobolus carbonum* is critical for ability to cause disease on cultivars that have the *Hm* resistance gene [37–39]. Such gene clusters generally involve "signature" secondary metabolic genes such as non-ribosomal peptide synthases (NRPS), type I polyketide synthases (PKS), terpene cyclases (TS),

Fig. 2 Mechanisms of operon formation in bacteria (adapted from Ref. [9]). ORFans are genes that lack identifiable homologues outside a group of closely related bacteria. Most ORFans are functional protein-encoding genes that are under purifying selection and so are likely to contribute to the fitness of the organism. They were probably acquired from bacteriophage

(a) Insertion of foreign (ORFan) gene

(b) Deletion of intervening genes

(c) Genome rearrangement
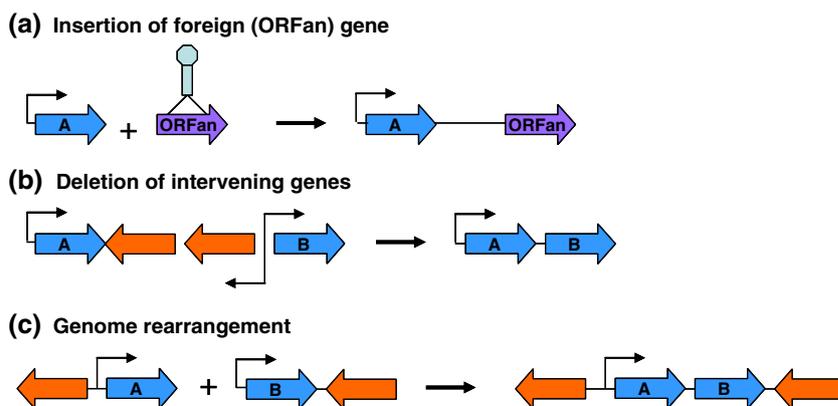
and dimethylallyl tryptophan synthetases (DMATS) genes, for the synthesis of non-ribosomal peptides, polyketides, terpenes, and indole alkaloids, respectively [32]. These signature genes are clustered along with various combinations of genes for further metabolite elaboration (e.g., oxidoreductases, methylases, acetylases, esterases), transporters, and sometimes (but not always) regulators [32, 33, 40]. While the first fungal gene clusters were identified using a combination of genetic, molecular genetic and biochemical approaches, the advent of fungal genome sequencing has enabled facile discovery of new candidate secondary metabolic gene clusters based on genome browsing and co-expression analysis [32–34, 40].

Many fungal secondary metabolites have important biological properties such as antibacterial, antifungal, or antitumor activity [41]. The exploitation of fungal metabolites for pharmaceutical purposes was pioneered by the discovery of penicillin in 1929 and its subsequent development for large-scale use as an antibiotic during the Second World War. The fungal natural product repertoire has since been the subject of extensive screening programmes for drug discovery [32]. Although some secondary metabolites have been shown to be important in mediating interactions between fungal pathogens and their hosts [33, 36, 37, 42, 43], in many cases the biological significance of these compounds for the producing fungus is unknown. The production of different types of compound is often restricted to particular fungal lineages, and the most likely function of these compounds is in niche adaptation and survival. Many filamentous fungi live saprophytically in the soil where they are exposed to a diverse range of organisms. Secondary metabolites may therefore act as mediators of interactions within soil communities. Secondary metabolite production has recently been shown to protect *Aspergillus nidulans* from predation by an arthropod [44]. In a more complex three-way interaction, colonization of perennial ryegrass by secondary metabolite-producing fungal endophytes has been shown to confer protection against insect herbivory [45].

Secondary metabolite gene clusters in filamentous fungi commonly contain a gene for a pathway-specific transcription factor that positively regulates expression of the associated biosynthetic genes, each of which has its own promoter. These transcription factors are often $Zn(II)_2Cys_6$ zinc binuclear cluster proteins, a class of protein so far found only in fungi [32]. One such example is AflR, which is required for aflatoxin and sterigmatocystin biosynthesis in Aspergilli [46, 47]. Other transcription factors that are encoded in biosynthetic gene clusters include $Cys_2 His_2$ zinc-finger proteins (Tri6 and MRT16 for trichothecene production) [48] and an ankyrin repeat protein (ToxE for HC-toxin production) [38]. Transcription factor genes for synthesis of ergovaline and lolitrem in the endophytic fungi

Neotyphodium lolii and *Epichloe festuca* do not reside within the cognate biosynthetic gene clusters and are as yet unidentified [49–52]. In these latter cases, identification of pathway regulators will necessarily rely on forward screens for mutants with altered regulation of expression of the pathway, on reverse approaches that involve searching for regulatory proteins that bind to promoters of characterized biosynthetic genes, and/or on genome-wide co-expression analysis.

The *Aspergillus laeA* gene encodes a nuclear protein with homology to arginine and histone methyltransferases [32, 53–57]. LaeA was identified in a screen for mutants of *A. nidulans* that were affected in their ability to synthesize sterigmatocystin [53]. Deletion of *laeA* results in loss of *aflR* gene expression and sterigmatocystin and aflatoxin synthesis in *A. nidulans* and *A. flavus*. It transpires that *laeE* deletion and overexpression strains are also affected in the synthesis of several other characterized and novel *Aspergillus* secondary metabolites, indicating a role for LaeA as a global regulator of secondary metabolism [53, 57, 58]. Comparison of global gene expression in *A. flavus* wild-type, laeA mutant, and complemented strains has revealed that LaeA controls transcription of around 10% of the genes in the *A. fumigatus* genome. Many of these genes are in 13 of the 22 secondary metabolite clusters. Seven of the LaeA-regulated clusters are in subtelomeric regions of chromosomes with a high degree of heterochromatin [58]. Collectively, these data suggest that LaeA acts as a master regulator of secondary metabolism in *Aspergillus* through chromatin remodeling [53, 54]. This hypothesis is supported by chromatin immunoprecipitation (ChIP) studies. Heterochromatin mutants with enhanced sterigmatocystin biosynthesis show decreased trimethylation of lysine residue K9 of histone H3, whilst *laeA* mutants show increased trimethylation of this lysine residue and a concomitant decrease in sterigmatocystin production [33]. Deletion of the histone deacetylase (HdaA) gene results in early and increased expression of biosynthetic genes not only for sterigmatocystin but also for another secondary metabolic gene cluster for penicillin synthesis [59]. Treatment of *Aspergillus* species and other fungi with selective drugs that inhibit histone deacetylase or DNA methyltransferase gives altered secondary metabolite profiles compared to control treatments [59–61]. It thus appears that epigenetic processes have important functions in the regulation of secondary metabolic gene clusters. Chemical genetic techniques are now being exploited along with genetic and genomics-based approaches to identify new cryptic pathways and metabolites in diverse filamentous fungi [55, 58, 60, 61]. Overexpression of transcription factors implicated in regulation of secondary metabolic pathways offers a further route to identification of new pathways and their end-products [55, 62]. Synthesis of secondary metabolites

by filamentous fungi is also under environmental and developmental control. The signaling cascades associated with these tiers of regulation are integrated into the hierarchical regulation of expression of the gene clusters [32, 33, 63] (Fig. 3). LaeA has also recently been shown to control penicillin biosynthesis, pigmentation, and sporulation in *Penicillium chrysogenum* [64].

There is evidence to suggest HGT of the penicillin gene cluster from bacteria to fungi [65–67]. HGT of clusters between fungi may in part explain the discontinuous distribution of gene clusters for synthesis of secondary metabolites within the Ascomycetes [68, 69]. The selfish cluster hypothesis has been put forward to explain clustering of functionally related genes in fungi [66]. However, fungal genomes are very plastic, and it is likely that the formation and maintenance of metabolic gene clusters in fungal genomes is driven by selection for optimized production of metabolites that fulfil an adaptive function. Many fungal metabolic gene clusters are located close to telomeres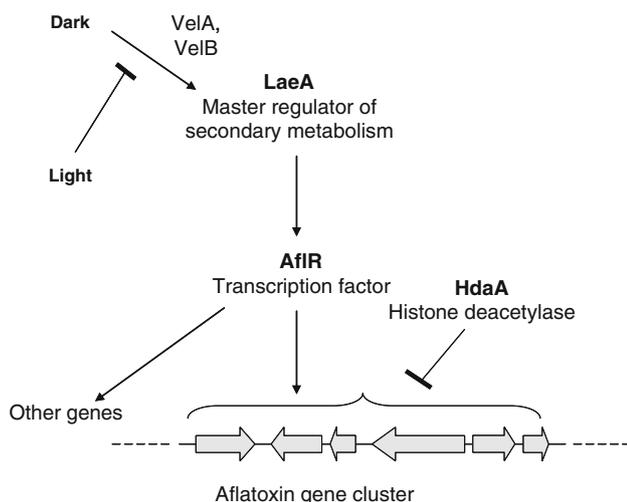, a chromosomal location that would be expected to facilitate recombination, DNA inversions, partial deletions, translocations, and other genomic rearrangements [70–73]. Intragenic reorganization followed by vertical descent is therefore a more satisfactory explanation. Clustering may facilitate co-regulation of gene expression, although it is clearly not a prerequisite for this since expression of unlinked genes for other metabolic pathways can be readily co-regulated. As the number of complete genome sequences of filamentous fungi increases, it should become possible to elucidate and perhaps model the mechanisms that drive cluster formation and maintenance, following approaches similar to those used to study the life and death of bacterial operons. While this section has focused on gene clusters for the synthesis of secondary metabolites in filamentous fungi, it is noteworthy that clusters of diverse virulence genes with no obvious function in metabolism have recently been identified in the corn smut fungus *Ustilago maydis* following completion of the full genome sequence of this organism [74].

Metabolic gene clusters in yeast

Baker's yeast *Saccharomyces cerevisiae*, unlike filamentous fungi, does not produce arrays of diverse secondary metabolites. Clusters of genes of related function are relatively unusual in *S. cerevisiae* by comparison with filamentous fungi, presumably for this reason. However, the *S. cerevisiae* genome does contain several functional gene clusters that are required for growth under certain conditions. These include gene clusters for utilization of specific carbon sources [e.g., the galactose (*GAL*) gene cluster], the DAL gene cluster for use of allantoin as a nitrogen source, and gene clusters for biotin synthesis, vitamin B1/B6 metabolism, and for arsenic resistance [75–77]. Studies of the distribution, origin, and fate of these gene clusters have provided important insights into the mechanisms underpinning adaptation of yeasts to new ecological niches.

The *DAL* gene cluster of *S. cerevisiae* consists of six contiguous genes located close to a telomere [76, 78]. These genes enable yeast to use allantoin (a purine degradation product) as a nitrogen source (Fig. 4). The *DAL* gene cluster is completely conserved in the four closest relatives of *S. cerevisiae*. In less closely related species, the six genes are also clustered, but there are differences in the internal arrangement of gene order and the cluster is located in a different part of the genome (although probably still subtelomeric). The *DAL* cluster is not present in the genomes of more distantly related hemiascomycetes. However, homologues of the six *DAL* genes are found scattered around the genomes of these species. The species that possess a *DAL* cluster form a monophyletic group.
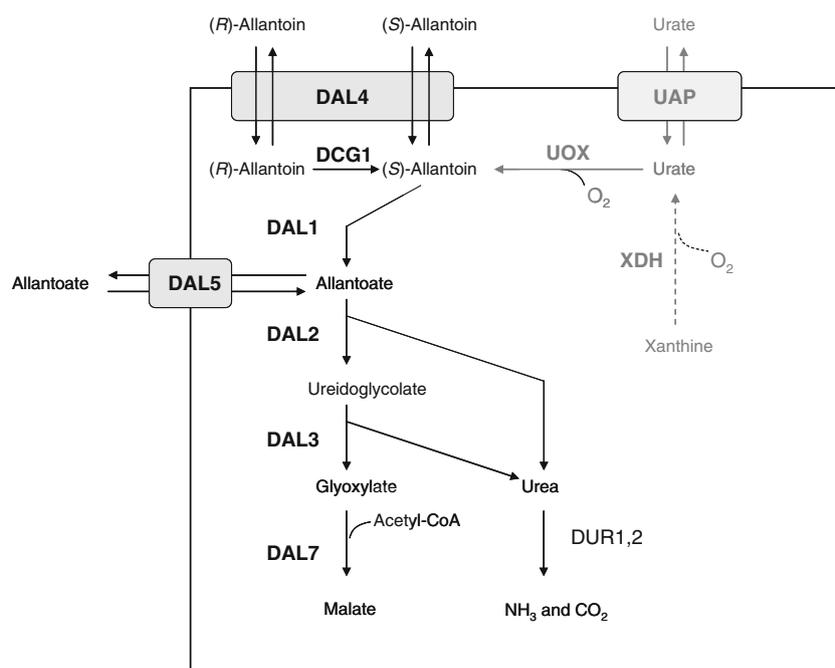

Fig. 3 Different levels of regulation of the aflatoxin gene cluster in the fungus, *Aspergillus nidulans*. The genes for aflatoxin biosynthesis are clustered in a 70-kb region and encode at least 23 co-regulated transcripts (not all cluster genes are shown in this diagram). The positive regulatory gene *aflR* lies within the gene cluster and is required for the transcriptional activation of most, if not all, pathway genes. *aflR* also regulates three genes outside the aflatoxin gene cluster. LaeA is a master regulator of secondary metabolism in the Aspergilli and is required for AlfR-mediated expression of the aflatoxin biosynthetic cluster. LaeA also controls transcription of gene clusters for other secondary metabolites and is believed to act via chromatin remodeling. Histone deacetylase (HdaA) performs an opposing role to LaeA and functions to repress gene expression in this region. Synthesis of secondary metabolites in the Aspergilli is linked to environmental stimuli and development. In the light asexual development (sporulation) is induced and genes involved in aflatoxin are not expressed. Darkness induces sexual development with associated induction of metabolite production, mediated by the proteins VelB and VeA. Adapted from [33]

Fig. 4 The allantoin degradation pathway—an adaptation to growth under oxygen-limiting conditions in *Saccharomyces cerevisiae*. In the classical purine degradation pathway, xanthine is converted to urate and then to allantoin, in two successive oxidation steps catalyzed by the peroxisomal enzymes xanthine dehydrogenase (XDH) and urate oxidase (UOX). *XDH* genes are present in filamentous fungi but not in yeasts. To use purine derivatives as a nitrogen source yeasts must therefore import urate, allantoin or allantoate from outside the cell. Yeasts that lack the *DAL* cluster (e.g., *Kluyveromyces lactis*, *Saccharomyces kluyveri*, and *Zygosaccharomyces rouxii*) are all able to grow on urate as a sole nitrogen source, presumably using urate permease (UAP) to import it, UOX to oxidize it and the *DAL* pathway enzymes to break it down to urea. In contrast, *S. cerevisiae* and *Saccharomyces castelli* have lost the ability to use urate, have no *UOX* or *UAP* genes, and instead import allantoin. The allantoin permease gene *DAL4* is a duplicate of the uracil permease gene *FUR4*. The Dal4 and Fur4 proteins are members of a purine-related transporter family. The subsequent degradation steps involve the same *DAL* pathway genes in all yeasts, but in *S. cerevisiae* and *S. castelli* the genes have been organized into a cluster and the malate synthase gene *MLS1* has been duplicated to produce *DAL7*. *MLS1* and *DAL7* both encode malate synthase but they are regulated differently. The glyoxylate cycle gene *MLS1* is glucose repressed, whereas *DAL7* is nitrogen-repressed. The reaction catalyzed by UOX requires molecular oxygen as a substrate and takes place in the peroxisome. Biochemical reorganization of the purine degradation pathway to enable import of allantoin instead of urate eliminates the oxygen-requiring step mediated by UOX and coincided with the formation of the *DAL* gene cluster. This biochemical reorganization may have been driven by selection for ability to grow under conditions of oxygen limitation. Reproduced from [76]

Comparative analysis of the *DAL* genes and clusters in the genomes of different yeast species in combination with phylogenetic information has provided compelling evidence to suggest that the *DAL* cluster was assembled quite recently in evolutionary terms, after the split of the sensu stricto group of yeasts from other yeasts and hemiascomycetes [76]. Six of the eight genes involved in allantoin degradation, which were previously scattered around the genome, became relocated to a single subtelomeric site in an ancestor of *S. cerevisiae* and *S. castelli*. Four of these genes (*DAL1*, *DAL2*, *DAL3*, and *DCG1*) are single-copy genes in *S. cerevisiae* and seem to have transposed to their new locations in the cluster. This could have occurred by gene duplication followed by loss of the gene at the original locus. The other two genes, *DAL4* and *DAL7*, were generated by duplication of progenitor genes that remain at their original sites in the *S. cerevisiae* genome (Fig. 5).

These genomic rearrangements coincided with a biochemical reorganization of the purine degradation pathway, which switched to importing allantoin instead of urate. This change circumvented the need for urate oxidase, one of several oxygen-consuming enzymes lost by yeasts that can grow vigorously in anaerobic conditions (Fig. 4). It has therefore been proposed that selection for reduced dependence on oxygen led to a switch from urate to allantoin utilization in an ancestor of the sensu stricto group of yeasts [76]. Natural sources of allantoin for yeasts are plants [79] and insect excretion [80].

The selection for formation of new metabolic gene clusters such as the *DAL* gene cluster is likely to be intense, driven by the need to adapt to growth under different environmental conditions. Gene clusters that have been formed by epistatic selection are expected to be recombination cold spots and so to be in linkage disequilibrium
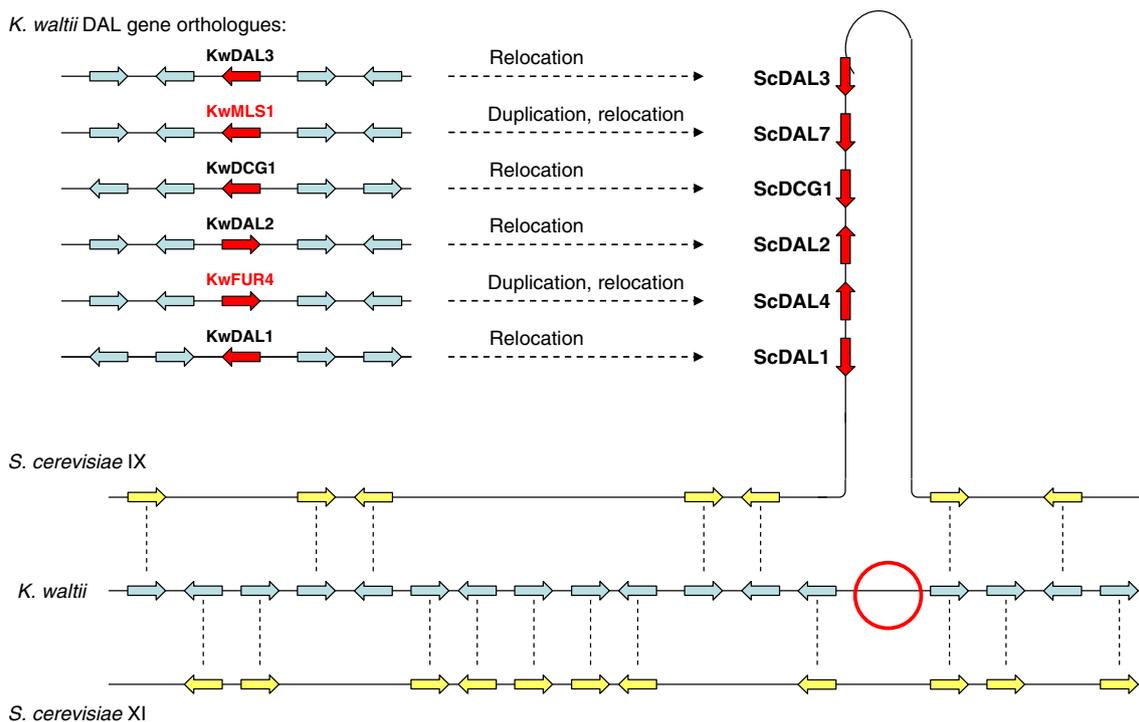
Fig. 5 The birth of the *DAL* gene cluster. The *DAL* gene cluster on *S. cerevisiae* chromosome IX (*red*) lies within a sister region between chromosomes IX and XI (genes shown in *pale yellow*). The corresponding region in *K. waltii*, a yeast species that does not contain the *DAL* gene cluster, contains a merge from these two chromosomes (genes shown at the *bottom* in *pale blue*). The *K. waltii* orthologues of the six *DAL* genes are scattered around the genome (*upper left*). *DAL1*, *DAL2*, *DAL3*, and *DCG1* apparently transposed to the cluster site, while *DAL4* and *DAL7* were formed by duplication of *FUR4* and *MLS1*. The region of the *K. waltii* corresponding to the site into which the *DAL* cluster has inserted is indicated with a *red circle*. Adapted from [76]

[81], and this is indeed the case for the *DAL* gene cluster [76]. Epistatic selection for linkage may in addition be driven by the need to select for combinations of alleles that interact well in order to avoid the accumulation of toxic pathway intermediates within cells. For example, glyoxylate, which is an intermediate in the DAL pathway (Fig. 4), is toxic to yeast [78]. Glyoxylate is produced by the Dal3 reaction and removed by the Dal7 reaction. There may therefore be selection for alleles of *DAL3* and *DAL7* that interact well and facilitate metabolic channeling. The finding that Dal3 enzyme activity is reduced in a *dal7* mutant is consistent with this channeling hypothesis [76, 82].

The *DAL* gene cluster is subtelomeric and lies in an Htz1-activated domain (HZAD) of the *S. cerevisiae* genome. HZADs are regions of the genome that are decorated with the histone H2A variant H2A.Z (Htz1) rather than with the normal histone H2A [83]. Htz1 preferentially associates with narrow regions within the promoters of genes that are normally maintained in repressed form but are strongly induced under specific growth conditions or during particular gene expression programmes. A model has been proposed in which Hzt1 associates to specific nucleosomes in the promoters of inactive genes in order to

poise, and perhaps organize, chromatin structure in a manner that is permissive to transcription initiation [84]. The genes in the *DAL* cluster are negatively regulated by the Hst-Sum1 system, which represses expression of mid-sporulation genes during mitotic growth (Fig. 6). Assuming that there is a selective advantage to repressing *DAL* gene expression when nitrogen is not limiting, there would have been an incremental selective advantage to relocating each gene into the chromatin modification (HZAD) domain. Thus, one way in which alleles could interact well is by being amenable to the same type of chromatin modification [76].

Under new selection regimes, adaptations may evolve while established functions may become less important. The *GAL* genes, which are required for galactose utilization, are clustered in the genomes of every yeast species in which they are present [75]. This pathway converts galactose into glucose-6-phosphate, a substrate for glycolysis. Galactose utilization is widespread amongst yeasts and is likely to be ancestral. However, several yeast species have lost the ability to use this carbon source. Comparisons of the genomes of galactose-utilizing and non-utilizing yeast species have revealed that three out of the four non-utilizing species examined lack any trace of the pathway
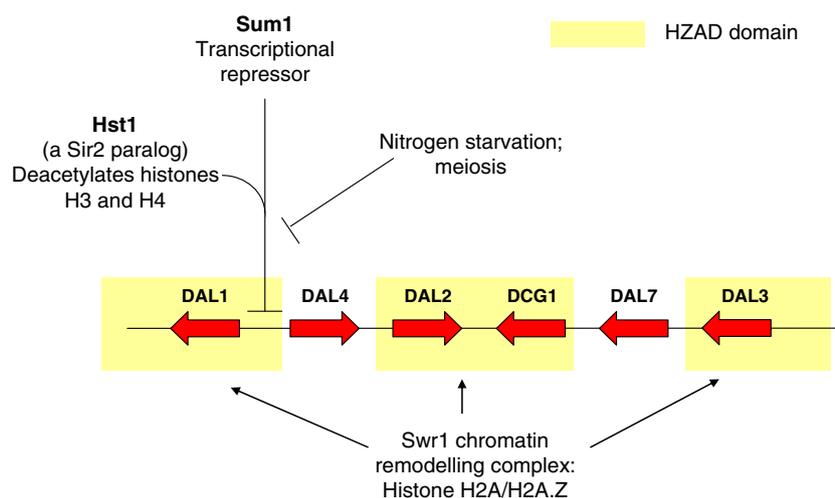
Fig. 6 Activation of expression of the *DAL* genes. The *DAL* gene cluster lies within an HZAD domain, in which histone H2A is replaced by the histone variant H2A.Z (Htz1) [83]. H2A.Z protects euchromatin from the ectopic spread of silent heterochromatin (the region of Sir2-effected silencing closer to the telomere), and is present in (and enables expression of) the *DAL1*, *DAL2*, *DCG1* and *DAL3* genes but not *DAL4* or *DAL7*. Interestingly, while *DAL1*,

DAL2, *DCG1* and *DAL3* are all single copy genes, *DAL4* and *DAL7* were generated by duplication of progenitor genes that remain at their original sites in the *S. cerevisiae* genome. The exchange of histone H2A for H2A.Z is mediated by the Swr1 chromatin remodeling complex [182]. The region between the *DAL1* and *DAL4* genes binds the transcription repressor Sum1, which in turn recruits Hst1 (a paralog of Sir2), which deacetylates histones H3 and H4

except for a single gene. However, *S. kudriavzevii*, a close relative of *S. cerevisiae*, retains remnants of all seven dedicated *GAL* genes as syntenic pseudogenes, providing a rare glimpse of an entire pathway in the process of degradation [75]. Thus, whilst a newly formed functional gene cluster confers a selective advantage in a new ecological niche, rapid and irreversible gene inactivation and pathway degeneration can occur under non-selective conditions. It has been suggested for *S. kudriavzevii* that this change may be associated with adaptation to growth on decaying leaves and soil rather than on sugar-rich substrates [75]. The loss of genes and pathways through reductive evolution has been inferred for many organisms that have adapted to pathogenic or endosymbiotic lifestyles [85–92]. Adaptation to a new niche has been shown to result in a "cost" in terms of lost ancestral capabilities. These capabilities may be lost either because they are no longer under selection (neutral) or because of a deleterious effect on fitness in a new niche [75, 93–95].

Operon-like gene clusters in plants

Genes for metabolic pathways in plants are generally not clustered, at least for the majority of the pathways that have been characterized in detail to date. However, several examples of functional gene clusters for plant metabolic pathways have recently emerged. These are the cyclic hydroxamic acid (DIBOA) pathway in maize [96–98], triterpene biosynthetic gene clusters in oat [99, 100] and

Arabidopsis [101] (the avenacin and thalianol gene clusters, respectively), and the diterpenoid momilactone cluster in rice [102, 103]. These gene clusters all appear to have been assembled from plant genes by gene duplication, acquisition of new function, and genome reorganization and are not likely to be a consequence of horizontal gene transfer from microbes. The existence of these clusters, of which at least three are implicated in plant defense [98, 99, 102–104], implies that plant genomes are able to assemble functional gene clusters that confer an adaptive advantage. The selection for rapid and recent formation of such metabolic gene clusters is likely to be intense, driven by the need to adapt to growth under different environmental conditions, and implies remarkable genome plasticity.

Benzoxazinoids

The benzoxazinoids are defense-related compounds that occur constitutively as glucosides in certain members of the Gramineae and in some dicots. 2,4-Dihydroxy-1,4-benzoxazin-3-one (DIBOA) is the primary hydroxamic acid in rye while its methoxy derivative 2,4-dihydroxy-7-methoxy-1,4-benzoxazin-3-one (DIMBOA) is predominant in maize and wheat [105–108]. In the Poaceae, the production of benzoxazinoids is developmentally regulated with highest levels being found in the roots and shoots of young seedlings. The glucosides are hydrolyzed in response to infection or physical damage to produce DIBOA and DIMBOA, which are antimicrobial and also have pesticidal and allelopathic activity. Induction of benzoxazinoid

accumulation has also been reported in response to *cis*-jasmone treatment [109].

The complete molecular pathway for benzoxazinoid biosynthesis has been elucidated in maize (reviewed in [98]). The first committed step towards DIBOA and DIMBOA biosynthesis is the conversion of indole-3-glycerol phosphate to indole, which is catalyzed by the tryptophan synthase α (TSA) homologue BX1. *Bx1* is likely to have been recruited from primary metabolism either directly or indirectly by duplication of the maize gene encoding TSA. BX1 and TSA are both chloroplast-localized indole-3-glycerolphosphate lyases (IGLs). BX1 functions as a monomer and produces free indole, while TSA forms a complex with the β-subunit of tryptophan synthase TSB to convert indole-3-glycerol phosphate to tryptophan [97, 110]. The subsequent conversion of indole into DIBOA is catalyzed by four related but highly substrate-specific cytochrome P450s (BX2-5) [96]. The glucosyltransferases BX8 and 9 catalyse glucosylation of benzoxazinoids. The glucosides of DIBOA and DIMBOA have reduced chemical reactivity when compared to the aglycones, suggesting that glucosylation may reduce phytotoxicity and so be important for storage [111]. Glucosylation takes place prior to hydroxylation by the 2-oxoglutarate dioxygenase (2-ODD) BX6 [112] and O-methylation by *O*-methyltransferase (OMT) BX7 [113]. All the *Bx* genes with the exception of *Bx9* are linked within 6 cM of *Bx1* on maize chromosome 4 [97, 111].

The distribution of benzoxazinoids across the Gramineae is sporadic. Maize, wheat, rye, and certain wild barley species are capable of the synthesis of these compounds while oats, rice, and cultivated barley varieties are not [106, 108]. The pathway to DIBOA is conserved in maize, wheat, and wild barley [97, 114–117]. The *Bx* gene cluster is believed to be of ancient origin. Wheat and rye have undergone a shared genomic event that has led to the splitting of the *Bx* gene cluster into two parts that are located on different chromosomes. This can be explained by a reciprocal translocation in the ancestor of wheat and rye [114]. Bx-deficient variants of a diploid accession of wild wheat *Triticum boeoticum* have recently been identified. Molecular characterization suggests that Bx deficiency in these accessions arose by disintegration of the *Bx1* coding sequence, followed by degeneration and loss of all five *Bx* biosynthetic genes examined [116]. Barley species that do not produce benzoxazinoids have also lost all *Bx* genes [114, 117]. The precise physical distances between all of the genes within the *Bx* cluster are not known. However, in maize, *Bx1* and *Bx2* genes are 2.5 kb apart [97] while *Bx8* is 44 kb from *Bx1* [118]. In hexaploid wheat, the *Bx3* and *Bx4* genes are 7–11 kb apart within the three genomes [119]. Although several of the *Bx* genes are in close physical proximity this gene cluster appears to be less tightly linked than the other examples that have been considered so far in this review. Interestingly, barley lines that produce benzoxazinoids do not synthesize gramine, a defense compound that is also derived from the tryptophan pathway. Conversely, gramine-accumulating barley species are deficient in benzoxazinoids. This has led to the suggestion that the biosynthetic pathways for these two different classes of defense compound are mutually exclusive, possibly due to competition for common substrates [117].

Outside the Poaceae, benzoxazinoids (in particular DIBOA and its glucoside) are found in certain isolated eudicot species belonging to the orders Ranunculales (e.g., larkspur, *Consolida orientalis*; yellow archangel, *Lamium galeobdolon*) and Lamiales (e.g., zebra plant, *Acanthus squarrosa*) [120]. Comparison of the BX1 enzymes of grasses and benzoxazinone-producing eudicots indicates that these enzymes do not share a common monophyletic origin. Furthermore, the CYP71C family of CYP450s to which BX2-5 belong is not represented in the model eudicot, thale cress *Arabidopsis thaliana*, and all members of this family described to date originate from the Poaceae. It therefore seems likely that the ability to synthesize benzoxazinones has evolved independently in grasses and eudicots.

Terpenes

Investigation of triterpene biosynthesis in plants has led to the discovery of two other examples of operon-like metabolic gene clusters, namely the avenacin gene cluster in oat (*Avena* species) and the thalianol gene cluster in *A. thaliana* [99, 101, 104].

Avenacins are antimicrobial triterpene glycosides that confer broad spectrum disease resistance to soil-borne pathogens [104, 121]. Analysis of the genes and enzymes for avenacin synthesis has revealed that the pathway has evolved recently, since the divergence of oats from other cereals and grasses [99, 100, 122, 123, 186]. Transferal of genes for the synthesis of antimicrobial triterpenes into cereals such as wheat holds potential for crop improvement but first requires the necessary genes and enzymes to be characterized. Synthesis of avenacins is developmentally regulated and occurs in the epidermal cells of the root meristem. The major avenacin, A-1, has strong fluorescence under ultra-violet light and can be readily visualized in these cells. This fluorescence, which is an extremely unusual property amongst triterpenes, has enabled isolation of over 90 avenacin-deficient mutants using a simple screen for reduced root fluorescence [100, 104]. This mutant collection has facilitated gene cloning and pathway elucidation.

Sad1 encodes an oxidosqualene cyclase enzyme that catalyses the first committed step in the avenacin pathway [99, 122], while *Sad2* encodes a second early pathway enzyme—a novel cytochrome P450 enzyme belonging to the newly described monocot-specific CYP51H subfamily [100]. *Sad1* and *Sad2* are likely to have been recruited from the sterol pathway (from cycloartenol synthase and obtusifoliol 14α-demethylase, respectively) by gene duplication and acquisition of new functions [99, 100, 122]. Sterols and avenacins are both synthesized from the mevalonate pathway [124]. While the genes for sterol synthesis are generally regarded as being constitutively expressed throughout the plant, the expression of *Sad1*, *Sad2*, and other cloned genes for avenacin biosynthesis is tightly regulated and is restricted to the epidermal cells of the root meristem [99, 100, 122]. Recruitment of *Sad1* and *Sad2* from the sterol pathway by gene duplication has therefore involved a change in expression pattern as well as neofunctionalisation.

The *Sad1* and *Sad2* genes are adjacent, and lie ~70 kb apart in the *A. strigosa* genome [100]. A third gene has recently been cloned and shown to encode a serine carboxypeptidase-like acyltransferase that is required for avenacin acylation. This gene (*Sad7*) [104, 186] is ~60 kb from *Sad1* on the opposite side to *Sad2*. Four other loci that are required for avenacin synthesis also co-segregate with these cloned genes, indicating that most of the genes for the pathway are likely to be clustered [99]. Since avenacins confer broad spectrum disease resistance, the gene cluster is likely to have arisen through strong epistatic selection for maintenance and co-inheritance of this gene collective. In addition, interference with the integrity of the gene cluster can in some cases lead to the accumulation of toxic intermediates, with detrimental consequences for plant growth, so providing further selection for cluster maintenance [123].

Gene clustering may also facilitate co-ordinate regulation of gene expression at the level of chromatin [2].

The thalianol gene cluster in *A. thaliana* consists of four contiguous co-expressed genes encoding an oxidosqualene cyclase (THAS), two different types of cytochrome P450 (THAH and THAD), and a BAHD acyltransferase [101] (Fig. 7). THAS, THAH, and THAD catalyze the synthesis and subsequent modification of thalianol. The BAHD acyltransferase gene is predicted to be part of the cluster based on its location and expression pattern, but an acylated downstream product has not as yet been identified. In the related crucifer, *A. lyrata*, the structure of the BAHD acyltransferase gene is different, and the two BAHD genes are more divergent than the other pairs of genes within these regions (Fig. 7). This may be indicative of paralogy rather than orthology [125]. Alternatively it may indicate that the BAHD acyltransferase genes are not under strong selection and so are divergent. It is noteworthy that there is also a large insertion of >100 kb (consisting mainly of transposons) in the cluster in *A. lyrata* that is not present in *A. thaliana*. *A. thaliana* thalianol pathway mutants do not show any obvious effects on plant defense and the function of the pathway is as yet unknown. However, the *THAS*, *THAH*, and *THAD* sequences are highly conserved across different *A. thaliana* accessions, implying that the gene cluster confers an important and as yet unidentified selective advantage in this species [101].

The genes within the *A. thaliana* thalianol gene cluster are expressed predominantly in the root epidermis, as is the case for the oat avenacin gene cluster [99–101]. The THAS, THAH, THAD, and BAHD acyltransferase genes all have marked histone H3 lysine 27 trimethylation, whereas the immediate flanking genes do not, suggesting co-ordinate regulation of expression of the gene cluster at the level of chromatin [101]. As is the case for the avenacin
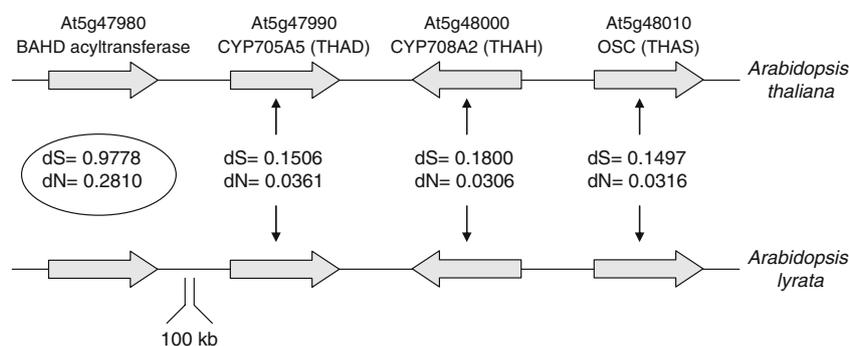

Fig. 7 The thalianol gene cluster in *Arabidopsis*. The *A. thaliana* gene cluster consists of four contiguous co-expressed genes encoding an oxidosqualene cyclase (THAS), two different types of cytochrome P450 (THAH and THAD) and a BAHD acyltransferase [101]. The organization of the equivalent region from the related crucifer, *A. lyrata*, is shown. The structure of the BAHD acyltransferase gene in *A. lyrata* is different (indicated by a split gene) and there is a 100-kb insertion in this region. *dS* number of synonymous substitutions per synonymous site; *dN* number of nonsynonymous substitutions per nonsynonymous site. Adapted from Ref. [125]

pathway, tight regulation of the pathway appears to be critical since accumulation of thalianol pathway intermediates can impact on plant growth and development.

There are superficial similarities between the avenacin and thalianol gene clusters in that they are both required for triterpene synthesis and contain genes for oxidosqualene cyclases, CYP450s, and acyltransferases. However, phylogenetic analysis indicates that the genes within these clusters are monocot and eudicot specific, respectively, and that the assembly of these clusters has occurred recently and independently in the two plant lineages [101]. This suggests that selection pressure may act during the formation of certain plant metabolic pathways to drive gene clustering, and that triterpene pathways are predisposed to such clustering.

A third example of a gene cluster for synthesis of terpenes in plants has been reported from rice, in this case for synthesis of diterpene defense compounds known as momilactones [102, 103]. Momilactones were originally identified as dormancy factors from rice seed husks and are also constitutively secreted from the roots of rice seedlings. In rice cell suspension cultures and in leaves, expression of the rice momilactone genes can be co-ordinately induced in response to challenge with pathogens, elicitor treatment, or exposure to UV irradiation [102, 103]. Synthesis of momilactones is initiated by terpene synthases that are distinct from the oxidosqualene cyclases that catalyze the first committed step in triterpene synthesis. The 168-kb momilactone gene cluster lies on rice chromosome 4 and consists of two diterpene synthase genes, a dehydrogenase gene and two functionally uncharacterized P450 genes, all of which are involved in/implicated in momilactone

synthesis [103]. These genes are all co-ordinately induced in response to treatment with a chitin oligosaccharide elicitor. Analysis of the promoters of the genes within this cluster has revealed the presence of potential recognition sites for WRKY and basic leucine zipper (bZIP) transcription factors, proteins that are associated with activation of defense responses. Gene clustering has been suggested to facilitate efficient coordinated expression of the momilactone gene cluster in response to elicitation [103].

Functions of gene clusters in animal defense and development

Global gene expression analysis has revealed extensive clustering of non-homologous genes that are co-ordinately expressed in eukaryotes, including in animals (for reviews, see [2, 126, 127]). These groups of genes may be expressed during development, or in certain tissues and diseased states, and have been reported in studies of *Drosophila*, nematode, mouse, and humans. Such co-expression domains may therefore be an important source for the discovery of new functional gene clusters in animals and other eukaryotes. However, more research is needed before we can fully understand the functional significance of co-expression domains [127]. Of the known functional gene clusters in animals, the best characterized is the major histocompatability complex (MHC), which encodes proteins involved in innate and adaptive immunity. Other classes of mammalian gene clusters include the *Hox* and β-globin loci, which are required for development and for
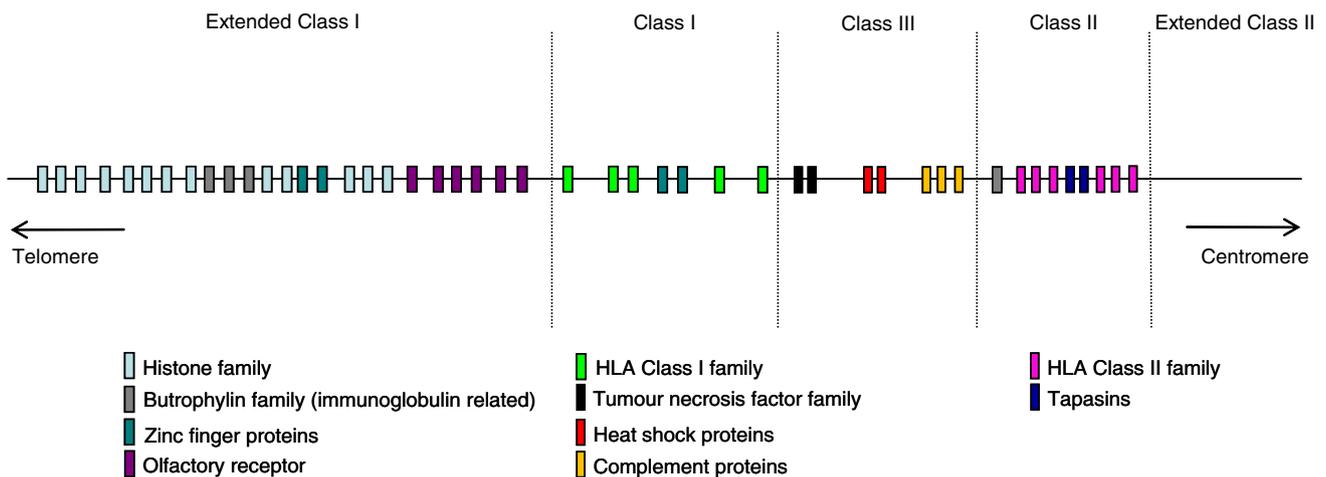

Fig. 8 The human major histocompatibility complex (MHC). A representation of the distribution of genes from large or immunologically important gene families across the ~7.5 Mb extended MHC on the short arm of human chromosome 6. Each *coloured block* represents multiple genes and the number of blocks indicates their approximate abundance. Single genes, which are interspersed throughout, are not shown. Also not shown are the 157 tRNA loci that lie in the extended Class I region. The drawing is not to scale

the synthesis of haemoglobin, respectively. The latter two examples consist of genes that share sequence similarity and so are distinct from classical operons and from the functional gene clusters discussed above. However, investigation of these loci has revealed important insights into the mechanisms of regulation of arrayed gene clusters in eukaryotes and so these gene clusters will also be considered here.

The major histocompatability complex (MHC)

The extended human major histocompatability complex (MHC) spans ∼7.5 Mb of chromosome 6, contains over 282 protein coding-genes and is the most important immunological region of the human genome [128] (Fig. 8). The MHC was discovered in mouse in 1936 and was investigated for its genetic role in tissue graft and organ transplant compatibility (histocompatibility). It later transpired that the primary function of the classical MHC gene-products was to provide protection against pathogens in innate and adaptive immunity. The MHC has been subdivided into a number of regions: the Class I region which contains the classical HLA Class I family genes that are involved in antigen presentation to CD8 T lymphocytes; the Class III region, which contains diverse immune and non-immune related genes and is the most gene dense region of the human genome; and the Class II region, which is dominated by the HLA Class II family genes that are involved in antigen presentation to CD4 T lymphocytes. The boundaries of the MHC have been pushed out since its discovery to delineate an extended MHC region [128]. Together, immune-related genes make up about 30% of the expressed transcripts in the extended MHC and their gene products function in antigen processing and presentation, inflammation, leukocyte maturation, complement signaling, immune regulation, and responses to stress [128]. However, three of the largest gene families in the extended MHC are not obviously immune-related. These are the tRNA, histone, and olfactory receptor gene families. Histones and tRNAs are required for DNA replication and protein synthesis, and enormous quantities of transcript must be produced to meet the needs of a single cell. Therefore, in the majority of eukaryotes, histone and tRNA genes tend to be found in tandem duplicated arrays that are also regions of high transcription [129]. In the human genome, the largest tRNA and histone gene arrays reside within the extended MHC (comprising 157 and 66 loci, respectively). For this reason, it has been suggested that the immune-related genes of the MHC may be hitchhiking with the histone and tRNA gene arrays, and in doing so benefiting from the inherent high transcriptional activity of the region [128].

The majority of genes from the MHC class I and III regions are constitutively expressed in all somatic cell types, although expression levels can vary over two orders of magnitude depending on the cell-type or extracellular stimuli [130]. By contrast, genes in the MHC class II region are expressed only in antigen-presenting cells or in other cell-types after induction by cytokines such as interferon-gamma. The Class II gene transactivator (CIITA) acts as a master regulator for the expression of genes in the MHC Class II region. Mutations in CIITA are one of the causes of bare lymphocyte syndrome in humans, a hereditary immunodeficiency disease characterized by a complete lack of MHC class II gene expression. CIITA acts by stabilizing a multi-subunit complex on the promoters of target genes to activate their transcription [131]. Formation of the CIITA-complex results in a large wave of histone acetylation and intergenic transcription that spreads out bi-directionally from the target promoter [132]. CIITA-mediated transcription is highly specific and appears to target only around 25 genes in the human genome, including the HLA Class II family genes within the MHC Class II region, two unrelated genes in the MHC Class I region, and other immune-related genes on different chromosomes [133]. Despite the presence of such specific transcription factors, the physical clustering of genes in the MHC does appear to be important for their regulation. The chromatin fibre of the extended MHC forms extra-chromosomal loops that are periodically anchored to the nuclear matrix at matrix attachment regions (MARs) by MAR binding proteins to form a chromatin "loopscape" [134, 135]. Treatment with interferon-gamma results in remodeling of the loopscape and differential regulation of genes within the affected regions [136]. Remodeling also leads to alterations in the composition of MAR binding proteins associated with the chromatin fibre and the subsequent recruitment of promyelocytic leukemia nuclear bodies, which may act as transcriptional factories for specific chromosomal loci [135, 137].

The origin of the major histocompatibility complex in metazoans is likely to predate the emergence of the jawed fish ∼500 million years ago [138]. The MHC region varies greatly in structure and size between species. This diversity is likely to have been driven by adaptive changes in response to selection pressure from pathogens and parasites [138]. The linkage between MHC Class I and Class II genes has been preserved from the cartilaginous fish to humans, except in the bony fish where linkage has disintegrated [138]. The MHC is also highly variable within species; in humans, as many as 300 alleles can be found at a single locus. Some alleles are so divergent that their common ancestor is likely to predate the formation of the species. The MHC is also a site of strong linkage disequilibrium, and large conserved blocks of specific alleles,

or haplotypes, of up to 3.2 Mb can be detected [139]. Therefore, particular patterns of MHC alleles may be fine-tuned to work together, and this may be one of the mechanisms by which clustering of the MHC is maintained [126].

A number of regions in the human genome are paralogous to the MHC, and these are thought to have arisen by the duplication and fragmentation of a single proto-MHC after two rounds of whole genome duplication [140]. One of these paralogous regions is the natural killer complex (NKC) [140]—a functional gene cluster that spans >1.5 Mb and contains C-type lectin receptor genes important for natural killer cell (NK) function in addition to other immune related genes. Remarkably, the MHC of chicken contains two C-type lectin receptor genes that are highly homologous to the receptor genes in the human NKC. The presence of these NK receptors in the chicken MHC suggests that the proto-MHC may have contained at least one ancestral C-type lectin receptor gene that was differentially retained at different loci after whole genome duplication in the lineages that gave rise to chicken and man. The leukocyte receptor complex (LRC), which contains a large array of immunoglobulin family NK receptors, is similarly thought to be derived from the proto-MHC [140]. Together, these results suggest that the clustering of immune-related genes at the MHC, NKC, and LRC was not independent, but instead was derived from the more ancient clustering of immune-related genes at the proto-MHC. How the ancestral genes became clustered remains a matter of debate.
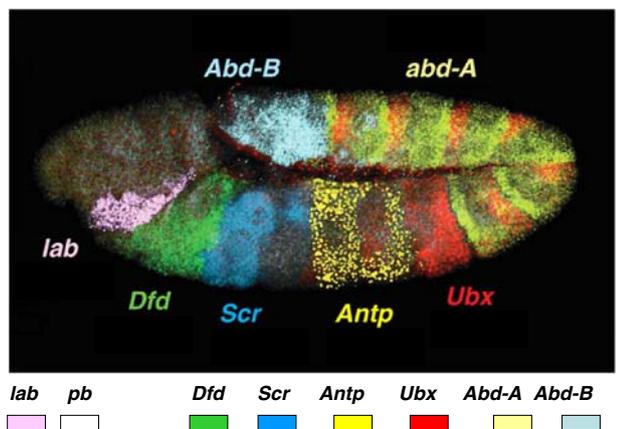

Fig. 9 Confocal image of septuple in situ hybridization exhibiting the spatial expression of *Hox* gene transcripts in a developing *Drosophila* embryo, with the chromosomal organization of the *Hox* gene cluster shown below: *lab* labial, *pb* proboscipedia (not shown in the in situ hybridization), *Dfd* deformed, *Scr* sex combs reduced, *Antp* antennapedia, *Ubx* ultrabithorax, *Abd-A* abdominal-A, *Abd-B* abdominal-B. The in situ hybridization image has been reproduced with permission from [183]
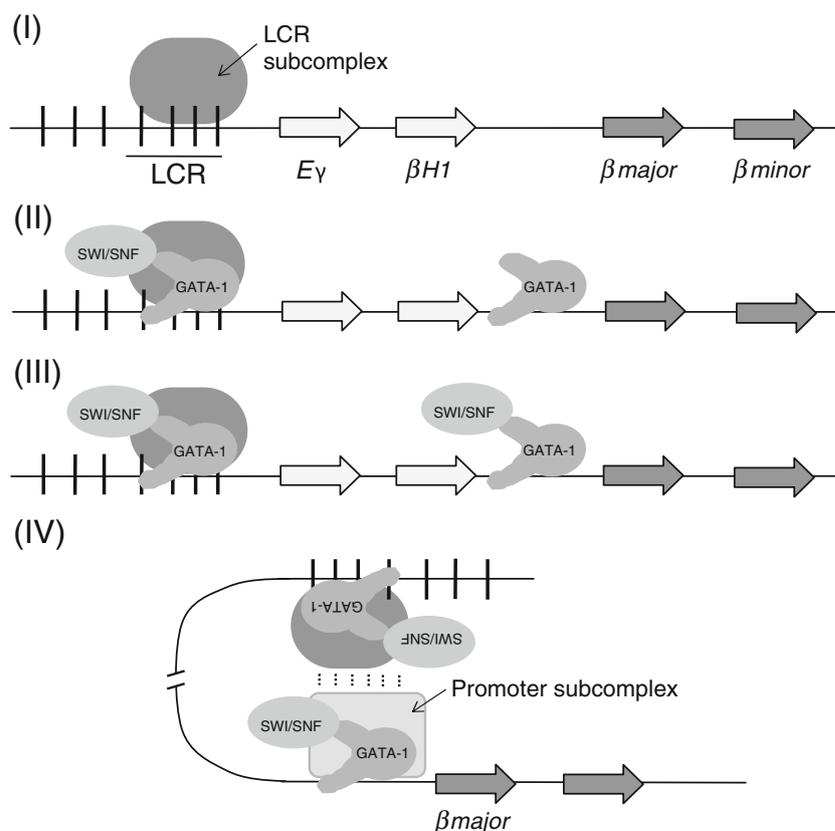
Hox *genes*

Homeotic mutations in animals result in the transformation of one body segment into another and have played a crucial role in shaping our understanding of animal development. In the late 1970s, it was discovered that many homeotic mutations in Drosophila mapped to the Bithorax and Antennapedia complexes [141, 142], complexes that we now know to consist of tandem duplicated arrays of genes encoding homeodomain (Hox) transcription factors [143, 144]. The first ancestral *Hox* cluster is thought to have appeared before the separation of the Cnidarians and Bilaterians, some 600 million years ago [145]. Diversification of the ancestral cluster facilitated the development of diverse and complex body plans across the metazoa. For example, in mammals, there are four *Hox* clusters that have arisen through whole genome duplications with up to 14 *Hox* genes within each cluster.

Expression of *Hox* genes is regulated in a remarkable fashion. The order of the genes along the chromosome corresponds to the region of expression along the length of the developing animal (spatial collinearity of expression) (Fig. 9). Clustered *Hox* genes are also activated in a wave that starts from the 3′ most end of the cluster and moves progressively to the 5′ end (temporal collinearity of expression). In some species, the *Hox* cluster has become disrupted. This is at its most extreme in the tunicate *Oikopleura dioica* where the *Hox* cluster has completely disintegrated and the remaining *Hox* genes are dispersed throughout the genome [146]. Cluster disintegration in *O. dioica* is accompanied by the loss of temporal collinearity of *Hox* gene expression. However, the spatial expression pattern of the *Hox* genes appears to be largely maintained. Through comparisons with other species lacking complete *Hox* clusters, it appears likely that cluster disintegration marks a switch to a more rapid life-cycle and a determinative mode of development where cell fate is fixed and flexible *Hox* regulation is no longer required [146]. Physical linkage of the *Hox* genes therefore appears to be important for choreographing *Hox* expression.

Retinoic acid (RA), a vitamin A-derived morphogen, can stimulate the sequential induction of *Hox* genes [147]. RA stimulation is mediated through a conserved retinoic acid responsive element (RARE) upstream of the 3′ most *Hox* gene, *Hox1*. Through the course of embryonic development, conserved RAREs upstream of progressively 5′ *Hox* genes propagate a wave of induction across the *Hox* cluster. Components of the RA signaling pathway either predate the *Hox* cluster or appeared shortly after its emergence, suggesting that temporal collinearity of expression may be an ancestral feature of the *Hox* cluster [148]. The sequential activation of *Hox* genes in different metazoans is accompanied by directional changes in

Fig. 10 The mouse β-globin locus. The four mouse β-like globin genes are located downstream of an array of DNAse I hypersensitive sites (*vertical bars*). The first four DNAse I hypersensitive sites upstream of *Eγ* form the locus control region (*LCR*). Activation of the *βmajor* locus occurs in four steps [151, 184]: (I) an LCR subcomplex is assembled over the LCR; (II), (III) GATA-1 occupies the LCR and *βmajor* promoter and recruits the SWI/SNF chromatin remodeling complex; and (IV) SWI/SNF-dependent chromatin looping and final assembly of the promoter complex occur simultaneously, and are followed by recruitment of Pol II and activation of *βmajor* transcription. Adapted from [184]

histone modifications, the opening up of the chromatin, and in some cases looping out of the chromatin fibre [126]. The orchestration of this complex series of events is still not fully understood. Transcriptional repression of *Hox* genes through histone modifications and chromatin condensation is equally important for the establishment of appropriate *Hox* expression domains. *Hox* expression can in addition be controlled post-transcriptionally and perhaps also epigenetically by non-coding RNAs and micro RNAs (miRNAs), the genes for which are embedded in the *Hox* cluster [149].

The β-globin gene cluster

Cis-regulated clusters of structurally related genes such as the *Hox* cluster are a recurring feature of animal genomes. The β-globin cluster is a particularly well-characterized example which has provided many insights into *cis*-regulation of gene expression in eukaryotes (for other examples, see [150]). The mouse β-globin cluster contains four related functional genes, *Ey*, *βH1*, *βmajor*, and *βminor*, encoding β-globin polypeptides that combine with α-globin polypeptides to form the haemoglobin tetramer. The β-globin genes are differentially expressed through development; *Eγ* and *βH1* are expressed in embryonic/fetal erythroid cells and *βmajor* and *βminor*

are expressed in the adult [151] (Fig. 10). Upstream of the β-globin genes are a number of DNase hypersensitive sites (HSs), indicating the presence of regions of low nucleosome density and enhanced chromatin accessibility [152]. A cluster of four upstream HSs comprise a type of enhancer known as a locus control region (LCR). An LCR is defined as a sequence that is able to confer copy number-dependent and position-independent expression of a transgene. The β-globin LCR was the first to be described [151]. Deletion studies in humans and mice have shown that the LCR is required for high level expression levels, but that when used to drive expression of a transgene it cannot recapitulate the tissue-specific developmental expression pattern of the β-globin genes. Expression from the β-globin cluster is accompanied by dynamic changes in both permissive and repressive histone modifications which are at least in part mediated by *trans*-acting factors such as the zinc-finger DNA binding transcription factor GATA-1 and the SWI/SNF chromatin remodeling complex [151] (Fig. 10). The intergenic transcription that occurs across the β-globin locus has also been proposed to play a role in the establishment of the histone pattern via polymerase II or RNAi dependent mechanisms [151]. Interestingly a major proportion of this intergenic transcription is actually initiated at the β-globin LCR [132].
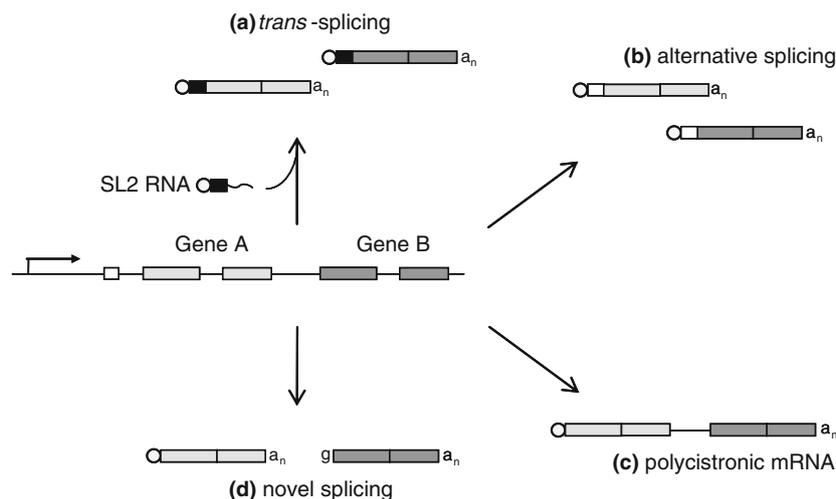
Fig. 11 Different strategies for the processing of eukaryotic operons. **a** The operons of *Caenorhabditis elegans* produce a polycistronic pre-mRNA that is processed into two or more mature monocistronic mRNAs by the *trans*-splicing machinery. The spliced leader sequence (*black box*) is derived from a spliced leader (SL) RNA which carries the 5′ RNA cap (*circle*) [156]. **b** Alternative splicing of a polycistronic pre-mRNA can produce two mature mRNAs with a common first exon (*white box*) as observed at the conserved cholinergic locus [185]. **c** In some cases, such as for the *Drosophila stoned* locus, mature polycistronic mRNAs can be translated directly [167]. The downstream CDS is likely to be translated through ribosome reinitiation, leaky scanning or internal ribosome entry. **d** Also in *Drosophila*, at least one polycistronic pre-mRNA is spliced via a novel mechanism that results in a conventional capped monocistronic mRNA for the upstream gene and an uncapped monocistronic mRNA for the downstream gene [170]. The uncapped transcript is stable and appears to be translated

Classical operons in eukaryotes?

In the early 1990s, structures with remarkable similarities to classical prokaryotic operons were discovered in the genome of the nematode *Caenorhabditis elegans* [153, 154]. We now know that 15% of genes in *C. elegans* are organized into such operons. These operons consist of linked genes that are under the control of a single promoter and are transcribed as a single polycistronic mRNA [155, 156] (Fig. 11). However, unlike the situation in prokaryotes, this polycistronic mRNA is then *trans*-spliced into monocistronic mRNAs that are translated individually. In *trans*-splicing, a donor splice-leader RNA donates a short leader sequence to a splice acceptor site within the polycistronic mRNA that becomes the 5′ end of a mature monocistronic mRNA. Genes within *C. elegans* operons are rarely related by sequence or function, and housekeeping and constitutively expressed genes are overrepresented [155–157]. This suggests that *C. elegans* operons arose in situ without gene duplication or genome rearrangement and therefore that these operons are of a fundamentally different nature to the majority of prokaryotic operons and also to functional gene clusters in eukaryotes. The adoption of *trans*-splicing is likely to be the key event that led to the flourishing of operons in nematodes. The benefits offered by *trans*-splicing are not well understood, but once established, this mRNA processing mechanism appears to favor operon formation and maintenance [158–160]. Operon breakup may be disfavored because downstream genes within an operon usually lack promoters and so are unlikely to be expressed following segmental duplication or disruption of the operon. Beyond the nematodes, *trans*-spliced operons have been detected in the genomes of several other animals including the tunicates *Oikopleura dioica* and *Ciona intestinalis* [161–163] and the chaetognath arrow worm, *Spadella cephaloptera* [164]. It has been predicted that operons may be present in all species able to carry out *trans*-splicing [165]. The origins of *trans*-splicing are unclear, and the scattered distribution of this mechanism across the animal kingdom could be due to convergent evolution or to extensive loss of an ancient feature [166]. A number of operon-like structures have also been identified in *Drosophila*, which lacks the *trans*-splicing machinery. Some *Drosophila* operons produce dicistronic mRNAs that are translated directly [167–169] while others, such as the *lounge lizard* (*llz*)/*CheB42a* locus produce a bicistronic pre-mRNA that is processed to produce two monocistronic mRNAs [170]. Subsequent bioinformatic analysis identified 409 *Drosophila* gene pairs that may also produce bicistronic mRNA [170]. Unlike *C. elegans*, the operons identified in *Drosophila* commonly contain genes with related functions and/or sequence similarity, and mRNA processing does not involve *trans*-splicing. Further research is required to determine whether *Drosophila* has particular mechanisms in place that promote operon

formation. Operon-like loci have also been identified in mammals [171–174], plants [175–177], and the filamentous fungi [178]. The small number of examples identified so far would suggest that such structures are very rare in the genomes of these intensively studied organisms.

Conclusions and perspectives

Here we have reviewed the literature on gene clusters in prokaryotes and eukaryotes, with particular emphasis on functional gene clusters. This theme is very broad and inevitably we have not been able to cover the entire swathe of literature in this field. We apologise to those whose work we have not cited. Nevertheless, we hope that by bringing together different and disparate facets of this area we have been able to highlight some of the similarities and differences between the ways in which gene clusters are organized and regulated in different organisms. The reasons for gene clustering can be considered at both the functional level and at the population level. These two considerations are not mutually exclusive [179]. Considering the population level first, where the fitness of an allele at one locus depends on the genotype at another locus then a selective advantage may arise for genomic rearrangements that reduce the distance between the two loci [180]. Significantly, this ratcheting effect may be enhanced when the fitness of recombinant haplotypes is low, for example where the combination of a functional and non-functional allele at two loci results in the premature disruption of a biochemical pathway and accumulation of toxic intermediates [76, 123, 181]. Operons or gene complexes can thus be regarded as units of strong gene interaction, with tightening of linkage between structural genes [179]. The selection for new gene clusters is likely to be intense, driven by the need to adapt to growth under different environmental conditions. At the functional level, physical clustering may be advantageous because it allows groups of genes to be co-ordinately regulated at the levels of nuclear organization and/or chromatin. Thus, one way in which alleles could interact well is by being co-localized in regions of chromosomes that facilitate co-ordinate regulation at this level and by being amenable to the same type of chromatin modification. In the future, it is likely that in-depth analysis of the levels at which functional gene clusters are post-transcriptionally regulated will reveal new facets of co-ordinate regulation that will shed further light on the mechanistic benefits of physical clustering.

References

1. Koonin E (2009) Evolution of genome architecture. Int J Biochem Cell Biol 41:298–306
2. Hurst LD, Pal C, Lercher MJ (2004) The evolutionary dynamics of eukaryotic gene order. Nat Rev Genet 5:299–310
3. Jacob F, Perrin D, Sanchez C, Monod J (1960) L'operon: Groupe de genes a l'expression coordonne par un operateur. C R Acad Sci 245: 1727–729
4. Jacob F, Monod J (1961) On the regulation of gene activity. In: Cold Spring Harbor Symposium Quantitative Biology 26, pp 193–211
5. Jacob F, Monod J (1961) Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol 3:318–356
6. Warren PB, ten Wolde PR (2004) Statistical analysis of the spatial distribution of operons in the transcriptional regulation network of *Escherichia coli*. J Mol Biol 342:1379–1390
7. Korbel JO, Jensen LJ, von Mering C, Bork P (2004) Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. Nat Biotechnol 22:911–917
8. Kepes F (2004) Periodic transcriptional organization of the *E. coli* genome. J Mol Biol 340:957–964
9. Price MN, Arkin AP, Alm EJ (2006) The life-cycle of operons. PLoS Genet 2:0859–0873
10. Lawrence JG, Roth JR (1996) Selfish operons: horizontal transfer may drive the evolution of gene clusters. Genetics 143:1843–1860
11. Lawrence JG (2002) Shared strategies in gene organization among prokaryotes and eukaryotes. Cell 110:407–413
12. Price MN, Huang KH, Alm EJ, Arkin AP (2005) Operon formation is driven by co-regulation and not by horizontal gene transfer. Genome Res 15:809–819
13. Omelchenko MV, Makarova KS, Wolf YI, Rogozin IB, Koonin EV (2003) Evolution of mosaic operons by horizontal gene transfer and gene displacement in situ. Genome Biol 4:R55
14. Pal C, Hurst JD (2004) Evidence against the selfish operon theory. Trends Genet 20:232–234
15. Rocha EPC (2008) The organization of the bacterial genome. Annu Rev Genet 42:211–233
16. Hazkani-Covo E, Graur D (2005) Evolutionary conservation of bacterial operons: does transcriptional connectivity matter? Genetica 124:145–166
17. Dekel E, Alon U (2005) Optimality and evolutionary tuning of the expression level of a protein. Nature 436:588–592
18. Dandekar T, Snel B, Huynen M, Bork P (1998) Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci 23:324–328
19. Swain PS (2004) Efficient attenuation of stochasticity in gene expression through post-transcriptional control. J Mol Biol 344:965–976
20. Butland G, Peregrine-Alvarez JM, Li J, Yang W, Yang X, Canadien V, Starostine A, Richards D, Beattie B, Krogan N, Davey M, Parkinson J, Greenblatt J, Emili A (2005) Interaction network containing conserved and essential protein complexes in *Escherichia coli*. Nature 433:531–537
21. Kolesov G, Wunderlich Z, Laikova ON, Gelfand MS, Mirny LA (2007) How gene order is influenced by the biophysics of transcription regulation. Proc Natl Acad Sci USA 104:13948–13953
22. Hershberg R, Yeger-Lotem E, Margalit H (2005) Chromosomal organization is shaped by the transcription regulatory network. Trends Genet 21:138–142
23. McFall E (1986) *cis*-Acting proteins. J Bacteriol 167:429–432
24. Golding I, Cox EC (2006) Physical nature of bacterial cytoplasm. Phys Rev Lett 96:98102–98104

25. Thanaraj TA, Argos P (1996) Ribosome-mediated translational pause and protein domain organization. Protein Sci 5:1594–1612

26. Danchin A, Guerdoux-Jamet P, Moszer I, Nitschke P (2000) Mapping the bacterial cell architecture into the chromosome. Philos Trans R Soc Lond B 355:179–190

27. Daubin V, Ochman H (2004) Bacterial genomes as new gene homes: the genealogy of ORFans in *E. coli*. Genome Res 14:1036–1042

28. de Daruvar A, Collado-Vides J, Valencia A (2009) Analysis of the cellular functions of *Escherichia coli* operons and their conservation in *Bacillus subtilis*. J Mol Evol 55:211–221

29. Ermolaeva MD, White O, Salzberg SL (2001) Prediction of operons in microbial genomes. Nucleic Acids Res 29:1216–1221

30. Lathe WC, Snel B, Bork P (2000) Gene context conservation of a higher order than operons. Trends Biochem 25:474–479

31. Rogozin IB, Makarova KS, Murvai J, Czabarka E, Wolf YI, Tatusov RL, Szekely LA, Koonin EV (2002) Connected gene neighborhoods in prokaryotic genomes. Nucleic Acids Res 30:2212–2223

32. Keller NP, Turner G, Bennett JW (2005) Fungal secondary metabolism—from biochemistry to genomics. Nat Rev Microbiol 3:937–947

33. Fox EM, Howlett BJ (2008) Secondary metabolism: regulation and role in fungal biology. Curr Opin Microbiol 11:481–487

34. Turgeon BG, Bushley KE (2009) Secondary metabolism. In: Borkovich K, Ebbole D (eds) Cellular and molecular biology of filamentous fungi. American Society of Microbiology, Washington D.C. (in press)

35. Bohnert HU, Fudal I, Dioh W, Tharreau D, Notteghem JL, Lebrun MH (2004) A putative polyketide synthase/peptide synthetase from *Magnaporthe grisea* signals pathogen attack to resistant rice. Plant Cell 16:2499–2513

36. Collemare J, Pianfetti M, Houlle AE, Morin D, Camborde L, Gagey MJ, Barbisan C, Fudal I, Lebrun MH, Bohnert HU (2008) *Magnaporthe grisea* avirulence gene ACE1 belongs to an infection-specific gene cluster involved in secondary metabolism. New Phytol 179:196–208

37. Brosch G, Ransom R, Lechner T, Walton JD, Loidl P (1995) Inhibition of maize histone deacetylases by HC toxin, the host-selective toxin of *Cochliobolus carbonum*. Plant Cell 7:1941–1950

38. Pedley KF, Walton JD (2001) Regulation of cyclic peptide biosynthesis in a plant pathogenic fungus by a novel transcription factor. Proc Natl Acad Sci USA 98:14174–14179

39. Walton JD (2006) HC-toxin. Phytochem Anal 67:1406–1413

40. Nierman WC, Pain A, Anderson MJ, Wortman JR, Kim HS, Arroyo J, Berriman M, Abe K, Archer DB, Bermejo C, Bennett J, Bowyer P, Chen D, Collins M, Coulsen R, Davies R, Dyer P, Farman M, Fedorova N, Fedorova N, Feldblyum TV, Fischer G, Fosker N, Fraser A, Garcia JL, Garcia MJ, Goble A, Goldman GH, Gomi K, Griffith-Jones S, Gwilliam R, Haas B, Haas H, Harris D, Horiuchi H, Huang J, Humphray S, Jimenez J, Keller N, Khouri H, Kitamoto K, Kobayashi T, Konzack S, Kulkarni R, Kumagai T, Lafton A, Latge JP, Li W, Lord A, Lu C, Majoros WH, May GS, Miller B, Mohamoud Y, Molina M, Monod M, Mouyna I, Mulligan S, Murphy L, O'Neil S, Paulsen I, Penalva MA, Pertea M, Price C, Pritchard BL, Quail MA, Rabbinowitsch E, Rawlins N, Rajandream MA, Reichard U, Renauld H, Robson GD, de Cordoba SR, Rodriguez-Pena JM, Ronning CM, Rutter S, Salzberg SL, Sanchez M, Saunders D, Seeger K, Squares R, Squares S, Takeuchi M, Tekaia F, Turner G, Carlos Weidman J, White O, Woodward J, Yu JH, Fraser C, Galagan JE, Asai K, Machida M, Hall C, Barrell B, Denning DW (2005) Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*. Nature 438:1151–1156

41. Pelaez F (2005) Biological activities of fungal metabolites. In: An Z (ed) Handbook of industrial mycology. Marcel Dekker, New York, pp 49–92

42. Oide S, Moeder W, Krasnoff S, Gibson D, Haas H, Yoshioka K, Turgeon BG (2006) NPS6, encoding a nonribosomal peptide synthetase involved in siderophore-mediated iron metabolism, is a conserved virulence determinant of plant pathogenic ascomycetes. Plant Cell 18:2836–2853

43. Elliott CE, Gardiner DM, Thomas G, Cozijnsen A, Van de Wouw A, Howlett BJ (2007) Production of the toxin sirodesmin PL by *Leptosphaeria maculans* during infection of *Brassica napus*. Mol Plant Pathol 8:791–802

44. Rohlfs M, Albert M, Keller NP, Kempken F (2007) Secondary chemicals protect mould from fungivory. Biol Lett 3:523–525

45. Tanaka A, Tapper BA, Popay A, Parker EJ, Scott B (2005) A symbiosis expressed non-ribosomal peptide synthetase from a mutualistic fungal endophyte of perennial ryegrass confers protection to the symbiotum from insect herbivory. Mol Microbiol 57:1036–1050

46. Woloshuk CP, Foutz KR, Brewer JF, Bhatnagar D, Cleveland TE, Payne GA (1994) Molecular characterization of *aflR*, a regulatory locus for aflatoxin biosynthesis. Appl Environ Microbiol 60:2408–2414

47. Georgianna DR, Payne GA (2009) Genetic regulation of aflatoxin biosynthesis: from gene to genome. Fungal Genet Biol 46:113–125

48. Proctor R, Hohn T, McCormick S, Desjardins A (1995) Tri6 encodes an unusual zinc finger protein involved in regulation of trichothecene biosynthesis in *Fusarium sporotrichiodes*. Appl Environ Microbiol 61:1923–1930

49. Young C, Bryant M, Christensen M, Tapper B, Bryan G, Scott B (2005) Molecular cloning and genetic analysis of a symbiosis-expressed gene cluster for lolitrem biosynthesis from a mutualistic endophyte of perennial ryegrass. Mol Genet Genomics 274:13–29

50. Young CA, Felitti S, Shields K, Spangenberg G, Johnson RD, Bryan GT, Saikia S, Scott B (2006) A complex gene cluster for indole-diterpene biosynthesis in the grass endophyte *Neotyphodium lolii*. Fungal Genet Biol 43:679–693

51. Young CA, Tapper BA, May K, Moon CD, Schardl CL, Scott B (2009) Indole-diterpene biosynthetic capability of Epichloe endophytes as predicted by *ltm* gene analysis. Appl Environ Microbiol 75(7):2200–2211

52. Fleetwood DJ, Scott B, Lane GA, Tanaka A, Johnson RD (2007) A complex ergovaline gene cluster in Epichloe endophytes of grasses. Appl Environ Microbiol 73:2571–2579

53. Bok JW, Keller NP (2004) LaeA, a regulator of secondary metabolism in *Aspergillus* spp. Eukaryot Cell 3:527–535

54. Yu JH, Keller N (2005) Regulation of secondary metabolism in filamentous fungi. Annu Rev Phytopathol 43:437–458

55. Bok JW, Hoffmeister D, Maggio-Hall LA, Murillo R, Glasner JD, Keller NP (2006) Genomic mining for *Aspergillus* natural products. Chem Biol 13:31–37

56. Hoffmeister D, Keller NP (2007) Natural products of filamentous fungi: enzymes, genes, and their regulation. Nat Prod Rep 24:393–416

57. Keller NP, Bok JW, Chung D, Perrin RM, Shwab EK (2006) LaeA, a global regulator of *Aspergillus* toxins. Med Mycol 44:83–85

58. Perrin RM, Fedorova ND, Bok JW, Cramer RA, Wortman JR, Kim HS, Nierman WC, Keller NP (2007) Transcriptional regulation of chemical diversity in *Aspergillus fumigatus* by LaeA. PLoS Pathog 3:e50

59. Shwab EK, Bok JW, Tribus M, Galehr J, Graessle S, Keller NP (2007) Histone deacetylase activity regulates chemical diversity in *Aspergillus*. Eukaryot Cell 6:1656–1664

60. Williams RB, Henrikson JC, Hoover AR, Lee AE, Cichewicz RH (2008) Epigenetic remodeling of the fungal secondary metabolome. Org Biomol Chem 6:1895–1897

61. Henrikson JC, Hoover AR, Joyner PM, Cichewicz RH (2009) A chemical epigenetics approach for engineering the in situ biosynthesis of a cryptic natural product from *Aspergillus niger*. Org Biomol Chem 7:435–438

62. Bergmann S, Schumann J, Scherlach K, Lange C, Brakhage AA, Hertweck C (2007) Genomics-driven discovery of PKS-NRPS hybrid metabolites from *Aspergillus nidulans*. Nat Chem Biol 3:213–217

63. Bayram O, Krappmann S, Ni M, Bok JW, Helmstaedt K, Valerius O, Braus-Stromeyer S, Kwon NJ, Keller NP, Yu JH (2008) VelB/VeA/LaeA complex coordinates light signal with fungal development and secondary metabolism. Science 320:1504–1506

64. Kosalková K, Garcia-Estrada C, Ullán RV, Godio RP, Feltrer R, Teijeira F, Mauriz E, Martin JF (2009) The global regulator LaeA controls penicillin biosynthesis, pigmentation and sporulation, but not roquefortine C synthesis in *Penicillium chrysogenum*. Biochimie 91:214–225

65. Rosewich U, Kistler H (2000) Role of horizontal gene transfer in the evolution of fungi. Annu Rev Phytopathol 38:325–363

66. Walton JD (2000) Horizontal gene transfer and the evolution of secondary metabolite gene clusters in fungi: an hypothesis. Fungal Genet Biol 30:167–171

67. Smith MW, Feng DF, Doolittle RF (1992) Evolution by acquisition: the case for horizontal gene transfers. Trends Biochem Sci 17:489–493

68. Patron NJ, Waller RF, Cozijnsen AJ, Straney DC, Gardiner DM, Nierman WC, Howlett BJ (2007) Origin and distribution of epipolythiodioxopiperazine (ETP) gene clusters in filamentous ascomycetes. BMC Evol Biol 7:174

69. Khaldi N, Collemare J, Lebrun MH, Wolfe K (2008) Evidence for horizontal transfer of a secondary metabolite gene cluster between fungi. Genome Biol 9:R18

70. Carbone I, Ramirez-Prado J, Jakobek J, Horn B (2007) Gene duplication, modularity and adaptation in the evolution of the aflatoxin gene cluster. BMC Evol Biol 7:111

71. Ehrlich KC, Yu J, Cotty PJ (2005) Aflatoxin biosynthesis gene clusters and flanking regions. J Appl Microbiol 99:518–527

72. Kusumoto K, Nogata Y, Ohta H (2000) Directed deletions in the aflatoxin biosynthesis gene homolog cluster of *Aspergillus oryzae*. Curr Genet 37:104–111

73. Chang PK, Horn BW, Dorner JW (2005) Sequence breakpoints in the aflatoxin biosynthesis gene cluster and flanking regions in nonaflatoxigenic *Aspergillus flavus* isolates. Fungal Genet Biol 42:914–923

74. Kämper J, Kahmann R, Böller M, Ma L-J, Brefort T, Saville BJ, Banuett F, Kronstad JW, Müller O (2006) Insights from the genome of the biotrophic funal plant pathogen *Ustilago maydis*. Nature 444:97–101

75. Hittinger CT, Rokas A, Carroll SB (2004) Parallel inactivation of multiple *GAL* pathway genes and ecological diversification in yeasts. Proc Natl Acad Sci USA 101:14144–14149

76. Wong S, Wolfe KH (2005) Birth of a metabolic gene cluster in yeast by adaptive gene relocation. Nat Genet 37:777–782

77. Hall C, Dietrich FS (2007) The reacquisition of biotin prototrophy in *Saccharomyces cerevisiae* involved horizontal gene transfer, gene duplication and gene clustering. Genetics 177:2293–2307

78. Cooper TG (1996) Regulation of allantoin catabolism in *Saccharomyces cerevisiae*. In: Marzluf GA (ed) The mycota III: biochemistry and molecular biology. Springer, Berlin, pp 139–169

79. Peoples MB, Gifford RM (1997) Regulation of the transport of nitrogen and carbon in higher plants. In: Dennis DT, Layzell DB, Lefebvre DD, Turpin DH (eds) Plant metabolism. Longman, Singapore, pp 525–538

80. Bursell E (1967) The excretion of nitrogen in insects. Adv Insect Physiol 4:33–67

81. Pal C, Hurst LD (2003) Evidence for co-evolution of gene order and recombination rate. Nat Genet 33:392–395

82. Hartig A, Simon MM, Schuster T, Gaugherty JR, Yoo HS, Cooper TG (1992) Differentially regulated malate synthase genes participate in carbon and nitrogen metabolism of *S.cerevisiae*. Nucleic Acids Res 20:5677–5686

83. Meneghini MD, Wu M, Madhani HD (2003) Conserved histone variant H2A.Z protects euchromatin from the ectopic spread of heterochromatin. Cell 112:725–736

84. Guillemette B, Bataille AR, Gévry N, Adam M, Blanchette M, Robert F, Gaudreau L (2005) Variant histone H2A.Z is globally localized to the promoters of inactive yeast genes and regulates nucleosome positioning. PLoS Biol 3:2100–2110

85. Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, Lafontaine I, de Montigny J, Marck C, Neuveglise C, Talla E, Goffard N, Frangeul L, Aigle M, Anthouard V, Babour A, Barbe V, Barnay S, Blanchin S, Beckerich JM, Beyne E, Bleykasten C, Boisrame A, Boyer J, Cattolico L, Confanioleri F, de Daruvar A, Despons L, Fabre E, Fairhed C, Ferry-Dumazet H, Groppi A, Hantraye F, Hennequink C, Jaunizux N, Joyet P, Kachouri R, Kerrest A, Koszul R, Lemaire M, Lesur IMLMH, Nicaud JM, Nikolski M, Oztas S, Ozier-Kalogeropoulos O, Pellenz S, Potier S, Richrd GF, Straub ML, Suleau A, Swennen D, Kekaia F, Wesolowski LM, Westhof E, Wikrth B, Zeniou-Meyer M, Zivanovic I, Bolotin-Fukuhara M, Thierry A, Boucher C, Caudron B, Scarpelli C, Gaillardin C, Weissenbach J, Wincker P, Souciet JL (2004) Genome evolution in yeasts. Nature 430:35–44

86. Andersson SG, Kurland CG (1998) Reductive evolution of resident genomes. Trends Microbiol 6:263–268

87. Lawrence JG, Hendrix RW, Casjens S (2001) Where are the pseudogenes in bacterial genomes? Trends Microbiol 9:535–540

88. Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, Wheeler PR, Honore N, Garnier T, Churcher C, Harris D, Mungall K, Basham D, Brown D, Chillingworth T, Connor R, Davies RM, Devlin K, Duthoy S, Feltwell T, Fraser A, Hamlin N, Holroyd S, Hornsby S, Hornsby T, Jagels K, Lacroix C, Maclean J, Moule S, Murphy L, Oliver K, Quail MA, Rajandream MA, Rutherford KM, Rutter S, Seeger K, Simon S, Simmonds M, Skelton J, Squares R, Squares S, Stevens K, Taylor K, Whitehead S, Woodward JR, Barrell BG (2001) Massive gene decay in the leprosy bacillus. Nature 409:1007–1011

89. Xie G, Bonner CA, Jensen RA (2002) Dynamic diversity of the tryptophan pathway in chlamydiae: reductive evolution and a novel operon for tryptophan recapture. Genome Biol 3

90. Harrison PM, Gerstein M (2002) Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. J Mol Biol 318:1155–1174

91. Moran NA (2003) Tracing the evolution of gene loss in obligate bacterial symbionts. Curr Opin Microbiol 6:512–518

92. Bungard RA (2004) Photosynthetic evolution in parasitic plants: insight from the chloroplast genome. BioEssays 26:235–247

93. Bell G (1997) The basics of selection. Chapman & Hall, New York

94. Kassen R (2002) The experimental evolution of specialists, generalists, and the maintenance of diversity. J Evol Biol 15:173–190

95. MacLean RG, Bell G (2002) Experimental adaptive radiation in *Pseudomonas*. Am Nat 160:569–581

96. Frey M, Kliem R, Saedler H, Gierl A (1995) Expression of a cytochrome P450 gene family in maize. Mol Gen Genet 246:100–109

97. Frey M, Chomet P, Glawischnig E, Stettner C, Grun S, Winklmair A, Eisenreich W, Bacher A, Meeley RB, Briggs SP, Simcox K, Gierl A (1997) Analysis of a chemical plant defense mechanism in grasses. Science 277:696–699

98. Gierl A, Frey M (2001) Evolution of benzoxazinone biosynthesis and indole production in maize. Planta 213:493–498

99. Qi X, Bakht S, Leggett M, Maxwell C, Melton R, Osbourn A (2004) 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

100. Qi X, Bakht S, Qin B, Leggett M, Hemmings A, Mellon F, Eagles J, Werck-Reichart D, Schaller H, Lesot A, Melton R, Osbourn A (2006) A different function for a member of an ancient and highly conserved cytochrome P450 family: from essential sterol to plant defense. Proc Natl Acad Sci USA 103:18848–18853

101. Field B, Osbourn AE (2008) Metabolic diversification—independent assembly of operon-like gene clusters in different plants. Science 320:543–547

102. Wilderman PR, Xu M, Jin Y, Coates RM, Peters RJ (2004) Identification of syn-imara-7,15-diene synthase reveals functional clustering of terpene synthases involved in rice phytoalexin/allelochemical biosynthesis. Plant Physiol 135:2098–2105

103. Shimura K, Okada A, Okada K, Jikumaru Y, Ko KW, Toyomasu T, Sassa T, Hasegawa M, Kodama O, Shibuya N, Koga J, Nojiri H, Yamane H (2007) Identification of a biosynthetic gene cluster in rice for momilactones. J Biol Chem 282:34013–34018

104. Papadopoulou K, Melton RE, Leggett M, Daniels MJ, Osbourn AE (1999) Compromised disease resistance in saponin-deficient plants. Proc Natl Acad Sci USA 96:12923–12928

105. Barnes JP, Putnam AR (1987) Role of benzoxazinones in allelopathy by rye (Secale cereale L.). J Chem Ecol 13:889–906

106. Niemeyer HM (1988) Hydroxamic acids (4-hydroxy-1,4-benzoxazin-3-ones), defense chemicals in the Gramineae. Phytochem Anal 27:3349–3358

107. Wu H, Haig T, Pratley J, Deidre L, An M (2001) Allelochemicals in wheat (Triticum aestivum L.): production and exudation of 2,4-dihydroxy-7-methoxy-1,4-benzoxazin-3-one. J Chem Ecol 27:1691–1700

108. Sicker D, Frey M, Schulz M, Gierl A (2000) Role of natural benzoxazinones in the survival strategy of plants. Int Rev Cytol 198:319–346

109. Moraes MCB, Birkett MA, Gordon-Weeks R, Smart LB, Martin JL, Pye BJ, Bromilow R, Pickett JA (2008) cis-Jasmone induces accumulation of defence compounds in wheat, Triticum aestivum. Phytochem Anal 69:9–17

110. Frey M, Stettner C, Paré PW, Schmelz EA, Tumlinson JH, Gierl A (2000) An herbivore elicitor activates the gene for indole emission in maize. Proc Natl Acad Sci USA 26:14801–14806

111. von Rad U, Hüttl R, Lottspeich F, Gierl A, Frey M (2001) Two glucosyltransferases are involved in detoxification of benzoxazinoids in maize. Plant J 28:633–642

112. Frey M, Huber K, Woong JP, Sicker D, Lindberg P, Meeley RB, Simmons CR, Yalpani N, Gierl A (2003) A 2-oxoglutarate-dependent dioxygenase is integrated in DIMBOA-biosynthesis. Phytochem Anal 62:371–376

113. Jonczyk R, Schmidt H, Osterrieder A, Fiesselmann A, Schullehner K, Haslbeck M, Sicker D, Hofmann D, Yalprani N, Simmons C, Frey M, Gierl A (2008) Elucidation of the final reactions of DIMBOA-glucoside biosynthesis in maize: characterization of Bx6 and Bx7. Plant Physiol 146:1053–1063

114. Nomura T, Ishihara A, Imaishi H, Endo T, Ohkawa H, Iwamura H (2002) Molecular characterization and chromosomal localization of cytochrome P450 genes involved in the biosynthesis of cyclic hydroxamic acids in hexaploid wheat. Mol Genet Genomics 267:210–217

115. Nomura T, Ishihara A, Imaishi H, Ohkawa H, Endo T, Iwamura H (2003) Rearrangement of the genes for the biosynthesis of benzoxazinones in the evolution of Triticeae species. Planta 77:6–782

116. Nomura T, Ishihara A, Iwamura H, Endo T (2007) Molecular characterization of benzoxazinone-deficient mutation in diploid wheat. Phytochem Anal 68:1008–1016

117. Grün S, Frey M, Gierl A (2005) Evolution of the indole alkaloid biosynthesis in the genus Hordeum: distribution of gramine and DIBOA and isolation of the benzoxazinoid biosynthesis genes from Hordeum lechleri. Phytochem Anal 66:1264–1272

118. Sandhu D, Gill KS (2002) Gene-containing regions of wheat and the other grass genomes. Plant Physiol 128:803–811

119. Nomura T, Nasuda S, Kawaura K, Ogihara Y, Kato N, Sato F, Kojima T, Toyoda A, Iwamura H, Endo T (2008) Structures of the three homoeologous loci of wheat benzoxazinone biosynthetic genes TaBx3 and TaBx4 and characterization of their promoter sequences. Theor Appl Genet 116:373–381

120. Schullehner K, Dick R, Vitzthum F, Schwab W, Brandt W, Frey M, Gierl A (2008) Benzoxazinoid biosynthesis in dicot plants. Phytochem Anal 69:2668–2677

121. Hostettmann KA, Marston A (1995) Saponins. Chemistry and pharmacology of natural products. Cambridge University Press, Cambridge

122. Haralampidis K, Bryan G, Qi X, Papadopoulou K, Bakht S, Melton R, Osbourn AE (2001) A new class of oxidosqualene cyclases directs synthesis of antimicrobial phytoprotectants in monocots. Proc Natl Acad Sci USA 98:13431–13436

123. Mylona P, Owatworakit A, Papadopoulou K, Jenner H, Qin B, Findlay K, Hill L, Qi X, Bakht S, Melton R, Osbourn A (2008) Sad3 and Sad4 are required for saponin biosynthesis and root development in oat. Plant Cell 20:201–212

124. Chappell J (2002) The genetics and molecular genetics of terpene and sterol origami. Curr Opin Plant Biol 5:151–157

125. Amoutzias G, Van de Peer Y (2008) Together we stand: genes cluster to coordinate regulation. Dev Cell 14:640–642

126. Sproul D, Gilbert N, Bickmore WA (2005) The role of chromatin structure in regulating the expression of clustered genes. Nat Rev Genet 6:775–781

127. Michalak P (2008) Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. Genomics 91:243–248

128. Horton R, Wilming L, Rand V, Lovering RC, Bruford EA, Khodiyar VK, Lush MJ, Povey S, Talbot CC, Wright MW, Wain HM, Trowsdale J, Ziegler A, Beck S (2004) Gene map of the extended human MHC. Nat Rev Genet 5:889–899

129. Kosak ST, Groudine M (2004) Gene order and dynamic domains. Science 306:644–647

130. Singer DS, Mozes E, Kirshner S, Kohn LD (1997) Role of MHC class I molecules in autoimmune disease. Crit Rev Immunol 17:463–468

131. Drozina G, Kohoutek J, Jabrane-Ferrat N, Peterlin BM (2005) Expression of MHC II genes. Curr Top Microbiol Immunol 290:147–170

132. Dean A (2006) On a chromosome far, far away: LCRs and gene expression. Trends Genet 22:38–45

133. Krawczyk M, Seguin-Estevez Q, Leimgruber E, Sperisen P, Schmid C, Bucher P, Reith W (2008) Identification of CIITA regulated genetic module dedicated for antigen presentation. PLoS Genet 4:e1000058

134. Galande S, Purbey PK, Notani D, Kumar PP (2007) The third dimension of gene regulation: organization of dynamic chromatin loopscape by SATB1. Curr Opin Genet Dev 17:408–414

135. Ottaviani D, Lever E, Mitter R, Jones T, Forshew T, Christova R, Tomazou EM, Rakyan VK, Krawetz SA, Platts AE, Segarane B, Beck S, Sheer D (2008) Reconfiguration of genomic anchors upon transcriptional activation of the human major histocompatibility complex. Genome Res 18:1778–1786

136. Kumar P, Bischof O, Purbey PK, Notani D, Urlaub H, Dejean A, Galande S (2007) Functional interaction between PML and SATB1 regulates chromatin-loop architecture and transcription of the MHC class I locus. Nat Cell Biol 9:U45–U57

137. Wang J, Shiels C, Sasieni P, Wu PJ, Islam SA, Freemont PS, Sheer D (2004) Promyelocytic leukemia nuclear bodies associate with transcriptionally active genomic regions. J Cell Biol 164:515–526

138. Kelley J, Walter L, Trowsdale J (2005) Comparative genomics of major histocompatibility complexes. Immunogenetics 56:683–695

139. Yunis EJ, Larsen CE, Fernandez-Vina M, Awdeh ZL, Romero T, Hansen JA, Alper CA (2003) Inheritable variable sizes of DNA stretches in the human MHC: conserved extended haplotypes and their fragments or blocks. Tissue Antigens 62:1–20

140. Kasahara M (2007) The 2R hypothesis: an update. Curr Opin Immunol 19:547–552

141. Kaufman TC, Lewis R, Wakimoto B (1980) Cytogenetic analysis of chromosome 3 in Drosophila melanogaster: the homoeotic gene complex in polytene chromosome interval 84a-B. Genetics 94:115–133

142. Lewis EB (1978) A gene complex controlling segmentation in Drosophila. Nature 276:565–570

143. McGinnis W, Garber RL, Wirz J, Kuroiwa A, Gehring WJ (1984) A homologous protein-coding sequence in Drosophila homeotic genes and its conservation in other metazoans. Cell 37:403–408

144. Scott MP, Weiner AJ (1984) Structural relationships among genes that control development: sequence homology between the Antennapedia, Ultrabithorax, and fushi tarazu loci of Drosophila. Proc Natl Acad Sci USA 411:5–4119

145. Chourrout D, Delsuc F, Chourrout P, Edvardsen RB, Rentzsch F, Renfer E, Jensen MF, Zhu B, de Jong P, Steele RE, Technau U (2006) Minimal ProtoHox cluster inferred from bilaterian and cnidarian Hox complements. Nature 442:684–687

146. Seo HC, Edvardsen RB, Maeland AD, Bjordal M, Jensen MF, Hansen A, Flaat M, Weissenbach J, Lehrach H, Wincker P, Reinhardt R, Chourrout D (2004) Hox cluster disintegration with persistent anteroposterior order of expression in Oikopleura dioica. Nature 431:67–71

147. Deschamps J (2007) Ancestral and recently recruited global control of the Hox genes in development. Curr Opin Genet Dev 17:422–427

148. Marletaz F, Holland LZ, Laudet V, Schubert M (2006) Retinoic acid signaling and the evolution of chordates. Int J Biol Sci 2:38–47

149. Yekta S, Tabin CJ, Bartel DP (2008) MicroRNAs in the Hox network: an apparent link to posterior prevalence. Nat Rev Genet 9:789–796

150. Li Q, Peterson KR, Fang X, Stamatoyannopoulos G (2002) Locus control regions. Blood 100:3077–3086

151. Wozniak RJ, Bresnick EH (2008) Epigenetic control of complex loci during erythropoiesis. Curr Top Dev Biol 82:55–83

152. Gross DS, Garrard WT (1988) Nuclease hypersensitive sites in chromatin. Annu Rev Biochem 57:159–197

153. Spieth J, Brooke G, Kuersten S, Lea K, Blumenthal T (1993) Operons in C. elegans: polycistronic mRNA precursors are processed by trans-splicing of SL2 to downstream coding regions. Cell 73:521–532

154. Zorio DAR, Cheng NSN, Blumenthal T, Spieth J (1994) Operons as a common form of chromosomal organization in C. elegans. Nature 372:270–272

155. Blumenthal T, Evans D, Link CD, Guffanti A, Lawson D, Thierry-Mieg J, Thierry-Mieg D, Chiu WL, Duke K, Kiraly M, Kim SK (2002) A global analysis of Caenorhabditis elegans operons. Nature 417:851–854

156. Blumenthal T, Gleason KS (2003) Caenorhabditis elegans operons: form and function. Nat Rev Genet 4:110–118

157. Lercher MJ, Blumenthal T, Hurst LD (2003) Coexpression of neighboring genes in Caenorhabditis elegans is mostly due to operons and duplicate genes. Genome Res 13:238–243

158. Lawrence J (1999) Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes. Curr Opin Genet Dev 9:642–648

159. Blumenthal T (2005) Trans-splicing and operons. In: The C. elegans Research Community (eds) WormBook, pp 1–9. http://www.wormbook.org

160. Qian WF, Zhang JZ (2008) Evolutionary dynamics of nematode operons: easy come, slow go. Genome Res 18:412–421

161. Ganot P, Kallesoe T, Reinhardt R, Chourrout D, Thompson EM (2004) Spliced-leader RNA trans splicing in a chordate, Oikopleura dioica, with a compact genome. Mol Cell Biol 24:7795–7805

162. Satou Y, Hamaguchi M, Takeuchi K, Hastings KEM, Satoh N (2006) Genomic overview of mRNA 5′-leader trans-splicing in the ascidian Ciona intestinalis. Nucleic Acids Res 34:3378–3388

163. Satou Y, Mineta K, Ogasawara M, Sasakura Y, Shoguchi E, Ueno K, Yamada L, Matsumoto J, Wasserscheid J, Dewar K, Wiley GB, Macmil SL, Roe BA, Zeller RW, Hastings KE, Lemaire P, Lindquist E, Endo T, Hotta K, Inaba K (2008) Improved genome assembly and evidence-based global gene model set for the chordate Ciona intestinalis: new insight into intron and operon populations. Genome Biol 9:R152

164. Marletaz F, Gilles A, Caubit X, Perez Y, Dossat C, Samain S, Gyapay G, Wincker P, Le Parco Y (2008) Chaetognath transcriptome reveals ancestral and unique features among bilaterians. Genome Biol 9:R152

165. Blumenthal T (2004) Operons in eukaryotes. Brief Funct Genomic Proteomic 3:199–211

166. Pouchkina-Stantcheva NN, Tunnacliffe A (2005) Spliced leader RNA-mediated trans-splicing in phylum Rotifera. Mol Biol Evol 22:1482–1489

167. Andrews J, Smith M, Merakovsky J, Coulson M, Hannan F, Kelly LE (1996) The stoned locus of Drosophila melanogaster produces a dicistronic transcript and encodes two distinct polypeptides. Genetics 143:1699–1711

168. Ramanathan P, Guo J, Whitehead RN, Brogna S (2008) The intergenic spacer of the Drosophila Adh-Adhr dicistronic mRNA stimulates internal translation initiation. RNA Biol 5:149–156

169. Brogna S, Ashburner M (1997) The Adh-related gene of Drosophila melanogaster is expressed as a functional dicistronic messenger RNA: multigenic transcription in higher organisms. EMBO J 16:2023–2031

170. Ben-Shahar Y, Nannapaneni K, Casavant TL, Scheetz TE, Welsh MJ (2007) Eukaryotic operon-like transcription of functionally related genes in Drosophila. Proc Natl Acad Sci USA 104:222–227

171. Lee SJ (1991) Expression of growth/differentiation factor 1 in the nervous system: conservation of a bicistronic structure. Proc Natl Acad Sci USA 88:4250–4254

172. Eiden LE (1998) The cholinergic gene locus. J Neurochem 70:2227–2240

173. Reiss J, Cohen N, Dorche C, Mandel H, Mendel RR, Stallmeyer B, Zabot MT, Dierks T (1998) Mutations in a polycistronic nuclear gene associated with molybdenum cofactor deficiency. Nat Genet 20:51–53

174. Wang BH, Shi GZ, Fu YC, Xu XH (2007) Cloning and characterization of a *LASS1-GDF1* transcript in rat cerebral cortex: conservation of a bicistronic structure. DNA Seq 18:92–103

175. Garcia-Rios M, Fujita T, LaRosa PC, Locy RD, Clithero JM, Bressan RA, Csonka LN (1997) Cloning of a polycistronic cDNA from tomato encoding gamma-glutamyl kinase and gamma-glutamyl phosphate reductase. Proc Natl Acad Sci USA 94:8249–8254

176. Muralla R, Chen E, Sweeney C, Gray JA, Dickerman A, Nikolau BJ, Meinke D (2008) A bifunctional locus (*BIO3-BIO1*) required for biotin biosynthesis in Arabidopsis. Plant Physiol 146:60–73

177. Thimmapuram J, Duan H, Liu L, Schuler MA (2005) Bicistronic and fused monocistronic transcripts are derived from adjacent loci in the Arabidopsis genome. RNA 11:128–138

178. Chung KR, Daub ME, Ehrenshaft M (2003) Expression of the cercosporin toxin resistance gene (CRG1) as a dicistronic mRNA in the filamentous fungus *Cercospora nicotianae*. Curr Genet 43:415–424

179. Stahl FW, Murray NE (1966) The evolution of gene clusters and genetic circularity in microorganisms. Genetics 53:569–576

180. Nei M (1967) Modification of linkage intensity by natural selection. Genetics 57:625–641

181. Nei M (2003) Let's stick together. Heredity 90:411–412

182. Krogan NJ (2003) A Snf2 family ATPase complex required for recruitment of the histone H2A variant Htz1. Cell 12:1565–1576

183. Lemons D, McGinnis W (2006) Genomic evolution of *Hox* gene clusters. Science 313:1918–1922

184. Kim SI, Bultman SJ, Kiefer CM, Dean A, Bresnick EH (2009) BRG1 requirement for long-range interaction of a locus control region with a downstream promoter. Proc Natl Acad Sci USA 106:2259–2264

185. Usdin TB, Eiden LE, Bonner TI, Erickson JD (1995) Molecular biology of the vesicular ACh transporter. Trends Neurosci 18:218–222

186. Mugford ST, Qi X, Bakht S, Hill L, Wegel E, Hughes RK, Papadopoulou K, Melton R, Philo M, Sainsbury F, Lomonossoff GP, Deb Roy A, Goss RJM, Osbourn A (2009) A serine carboxypeptidase-like acyltransferase is required for synthesis of antimicrobial compounds and disease resistance in oats. Plant Cell. http://www.plantcell.org/cgi/doi/10.1105/tpc.109.065870