Genomic clustering within functionally related gene families in Ascomycota fungi

Danielle Hagee, Ahmad Abu Hardan, Juan Botero, James T. Arnone*

Department of Biology, William Paterson University, Wayne, NJ 07470, USA

Abstract

Multiple mechanisms collaborate for proper regulation of gene expression. One layer of this regulation is through the clustering of functionally related genes at discrete loci throughout the genome. This phenomenon occurs extensively throughout Ascomycota fungi and is an organizing principle for many gene families whose proteins participate in diverse molecular functions throughout the cell. Members of this phylum include organisms that serve as model systems and those of interest medically, pharmaceutically, and for industrial and biotechnological applications. In this review, we discuss the prevalence of functional clustering through a broad range of organisms within the phylum. Position effects on transcription, genomic locations of clusters, transcriptional regulation of clusters, and selective pressures contributing to the formation and maintenance of clusters are addressed, as are common methods to identify and characterize clusters.

© 2020 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Contents

1. Introduction ................................................. 3268
2. Position effects due to the proximity of heterochromatin. ......................................................... 3268
3. Functional clustering of genes within complex biosynthetic pathways ........................................ 3270
   3.1. Secondary metabolite gene clusters ................................................................. 3270
   3.1.1. Pathogenic toxins and defense ............................................................... 3270
   3.1.2. Detoxification clusters ........................................................................ 3271
   3.2. Primary metabolite gene clusters ................................................................. 3271
   3.3. Superclusters .......................................................................................... 3272
   3.4. Regulons .............................................................................................. 3272
4. Transcriptional co-regulation within a neighborhood may drive functional clustering ................. 3272
5. Computational approaches to identify and characterize clusters .............................................. 3273
   5.1. Identification of functional clustering relationships using Euclidean distance from gene expression datasets .......................................................... 3273
   5.2. Pearson's correlation coefficient analysis of gene expression data ......................... 3273
   5.3. Spearman's correlation coefficient analysis of genomic neighborhoods ................... 3274
   5.4. The multivariate copula model for analysis of gene expression data ....................... 3274
   5.5. Determining the statistical significance of genomic arrangement through the with a hypergeometric distribution ........................................ 3274
6. Summary and outlook ........................................ 3274
Declaration of Competing Interest .................................. 3274
Acknowledgements ............................................. 3274
References ...................................................... 3275

* Corresponding author at: Department of Biology, William Paterson University, 300 Pompton Road, Wayne, NJ 07470, USA.
E-mail address: arnonej@wpunj.edu (J.T. Arnone).

https://doi.org/10.1016/j.csbj.2020.10.020
2001-0370 © 2020 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
1. Introduction

Transcriptional regulation is essential to ensure cellular and organismal survival. Cellular cues, both intracellularly initiated and extracellularly recognized, trigger comprehensive changes to the transcriptome. Such changes establish cellular identity and maintain homeostasis during stress [1–6]. Transcriptional changes can also be much smaller in scope, allowing for fine-tuning expression of individual genes as needed. At the core of these transcriptional changes is the production of a transient messenger RNA (mRNA) that serves as a template for protein synthesis via the ribosome, commonly referred to as the Central Dogma of Molecular Biology [7,8]. While the Central Dogma is a bit of an oversimplification, it remains a foundation for an understanding of gene expression [9,10]. Extensive study over the decades has yielded a wealth of knowledge about the myriad processes that collate to regulate proper gene expression in countless organisms – from the simplest single celled to staggeringly complex [11–15]. Gene positioning and arrangement throughout the genome can profoundly influence transcription, with the cis regulatory ‘logic’ effecting expression across a locus to silence or activate neighboring genes [16,17]. The influence of genomic arrangement on transcription has yielded significant insights, with implications for all organisms.

One of the major clades of living organisms is fungi, with estimates that there may be up to 1.5 million species within this group. There is incredible diversity within the fungal kingdom and throughout the phylum Ascomycota, with members exhibiting vast differences in morphology and lifestyles [18,19]. Ascomycetes form a characteristic sac-like structure called an ascus, that forms around their meiotic spores and leads to their common name, sac fungus (Fig. 1) [20]. Representative members are ideal as model organisms for researchers, providing many insights into eukaryotic biology – with multiple Nobel prizes awarded to researchers studying the budding yeast and the fission yeast. It includes filamentous fungi, bread molds, and the causative pathogens for powdery mildew, black rot, and anthracnose [21–25]. There are also a number of opportunistic human pathogenic Ascomycete organisms, including several emerging pathogens [26].

Members of this phylum can have significant metabolic flexibility, which makes them useful for a variety of biotechnological applications, including the production of fatty alcohols, fatty acids, biofuels, reduction and degradation of chemicals and solvents [27]. This phylum includes a broad range of well-studied model organisms, plant pathogens, animal pathogens, and a number of interests to the pharmaceutical and biotechnology industries (Table 1). Due to the many applications for Ascomycetes fungi in academia, industry, and medicine, a thorough understanding of genomic organization and the potential implications on gene expression is vital. There have been many advances in the identification and characterization of Ascomycetes on a genomic and transcriptional level. In this paper, we review these advances, with a focus on genomic organization and implications upon transcriptional regulation.

2. Position effects due to the proximity of heterochromatin

One of the advantages for the usage of members of Ascomycota for a wide range of applications is that they are amenable to genetic manipulations [22,28]. Classical genetic approaches allow for the characterization of molecular functions and frequently utilize reporter genes, or selectable markers, typically utilizing nutritional or drug resistance for selection. The telomere proximal effect (TPE) is a phenomenon well characterized in the budding yeast, *Saccharomyces cerevisiae*, initially identified during the integration of the *URA3*, *TRP1*, and *ADE2* reporter genes. Regardless of genomic orientation, when integrated adjacent to certain telomeres transcriptional repression of the reporter constructs is observed. The TPE is distance dependent; as the region between the telomere and a reporter increased, so does expression of the reporter gene [29]. Transcriptional repression by the TPE is due, in part, to the assembly of transcriptionally inactive heterochromatin that forms at the telomeres and spreads outwards in a continuous fashion mediated by the silent information regulator (SIR) histone deacetylase proteins and the structural maintenance of chromosome (SMC) protein complexes [30,31].

Further study demonstrated that silencing was not universal at all native yeast telomeres equally, while there are a number of telomeric sites where integration of a reporter gene is heavily silenced there are other sites with little silencing observed [32]. The major indicator of silencing at telomeres appears to be the proximity of the telomere to an autonomously replicating sequence contained within one of the telomeric repetitive elements (the core X element) [32]. A recent study provides the most definitive understanding of TPE via global characterization of transcription utilizing a highly sensitive RNA sequencing approach. There is widely seen transcription at many endogenous yeast telomeres, although at lower absolute levels of transcription compared to non-telomeric regions. Likewise, the SIR proteins play a role in silencing genes at telomeric regions adjacent to known SIR protein binding sites. The majority of telomeric genes are unaffected by the loss of SIR protein function, indicating other mechanisms limiting their expression [33].

The overwhelming majority of eukaryotic organisms maintain heterochromatin at the telomeres. This is true for many Ascomycetes, including *Schizosaccharomyces pombe*, *Aspergillus fumigatus*, and *Candida species* [34,35]. Heterochromatin is a feature of repet-

![Fig. 1. Penicillium visualized by light microscopy. At lower power (A, 400× magnification) the ascocarp (a completely closed fruit body called a cleistothecium) is visible. At higher magnifications (B, 1000× magnification) the ascus is visible with the individual ascospores (three of the spores are highlighted with an arrow). Images courtesy of Kelley Healey.](image-url)
itive genomic regions, which include the telomeres, centromeres, and mating loci [36]. While oftentimes heterochromatin conjures images of a static, silent structure, it can be dynamic, exhibiting plasticity under specific growth conditions. *S. pombe* and *C. albicans* alters heterochromatin at the telomere during elevated temperatures in a SIR dependent manner, indicating that this is dynamic – undergoing significant changes depending on the growth conditions [37,38]. This could result in the clustering of specific genes at

Table 1

| Model Organisms | Note about species |
|-----------------|--------------------|
| Aspergillus nidulans | Filamentous fungus |
| Neurospora crassa | Bread mold |
| Saccharomyces cerevisiae | Budding yeast |
| Schizosaccharomyces japonicas | Fission yeast |
| Schizosaccharomyces pombe | Fission yeast |
| Sordaria macrospora | Model of fruiting body development |

| Pathogenic and Medically Relevant | Note about species |
|----------------------------------|--------------------|
| Aspergillus fumigatus | Aspergillosis |
| Aspergillus terreus | Potato blight |
| Aspergillus viridinutans | Aspergillosis |
| Blastomyces dermatitidis | Blastomycosis |
| Botrytis cinerea | Necrotrophic fungus of grapes |
| Candida albicans | Candidiasis |
| Candida auris | Candidiasis, high incidence of drug resistance |
| Candida glabrata | Candidiasis |
| Candida krusei | Opportunistic pathogen |
| Candida lusitaniae | Fungemia |
| Candida parapsilosis | Candidiasis |
| Candida rugosa | Emerging human fungal pathogen |
| Candida tropicalis | Candidiasis |
| Cladophialaphora bantianum | Cerebral phaeohyphomycosis |
| Claviceps purpurea | Erogot fungus that infects rye and cereal and forage plants |
| Coccidioides immitis | Valley fever (coccidioidomycosis) |
| Coccidioides posadasii | Valley fever (coccidioidomycosis) |
| Cordyceps militaris | Entomopathogenic fungus |
| Cryphonectria parasitica | Chestnut blight (Stajich et al 2009) |
| Döthistroma septosporum | Red band needle blight |
| Exophiala dermatitidis | Phaeohyphomycosis |
| Exophiala jeaneselmei | Skin infections |
| Fusarium fujikuroi | Banaae disease |
| Fusarium oxysporum | Animal and plant pathogen |
| Fusarium oxysporum | Plant and animal pathogen |
| Fusarium solanum | Onychomycosis, keratomycosis, plant pathogen |
| Gibberella moniliformis | Maize ear and stalk rot |
| Histoplasma capsulatum | Pulmonary mycosis |
| Laccaria lobo | Lobo's disease |
| Leptosphaeria maculans | Blackleg disease |
| Ochroconis gallopava | Pathogen to fowls, turkeys, poults, and humans |
| Ophiostoma ulmi | Dutch elm disease (Stajich et al 2009) |
| Paecilomyces variotii | Common environmental mold |
| Paracoccidioides brasiliensis | Paracoccidioidomycosis |
| Paracoccidioides lutzii | Paracoccidioidomycosis |
| Penicillium expansum | Psychrophilic blue mold, infects apples |
| Pneumocystis jiroveci | Pneumocystis pneumonia (PCP) |
| Pseudallescheria boydii | Eumycetoma; maduromycosis; pseudallescheriasis |
| Ramichloridium musae | Banana pathogen |
| Sporothrix schenckii | Sporotrichosis |
| Talaraomyces marneffei | Talaromyces (penicilliosis) |
| Aureobasidium pullulans | Reductase |
| Candida boidini | Phenylalanine dehydrogenase |
| Candida cylindracea | Lipoas (Lip) hydrolyze triglycerides into fatty acids and glycerol |
| Candida maltosa | Resolving D- and L-racemic mixtures of amino acids |
| Candida sorbophila | α-oxidizing yeast |
| Debaryomyces hansenii | Oleaginous yeast; osmotolerant |
| Geotrichum candidum | Dehydrogenase |
| Rhuyveromyces lactis | Assimilate lactose and convert it into lactic acid |
| Rhuyveromyces marxianus | Aerobic yeast capable of respiro-fermentative metabolism |
| Komagataella pastoris | Glycolate oxidase; Phytase |
| Ogaetea polymorpha | Protein expression; use methanol consumption |
| Pichia pastoris | Methanol metabolism |
| Podospora anserina | Pentaketides |
| Saccharomyces bouardii | Treatment of intestinal diseases |
| Scheffersomyces stipitis | Ferment xylose, converting it to ethanol |
| Schwanomyces occidentalis | Myolytic enzymes, including α-amylase and glucoamylase |
| Trichoderma reesei | Cellulolytic enzymes (cellulases and hemicellulases) |
| Trigonopsis variabilis | D-amino acid oxidase |
| Yarrowia lipolytica | Specialty lipids; lipolytic enzymes |
| Zygosaccharomyces rouxii | L-Glutaminase |

Information from the table compiled and adapted [22,25–28,126].
3. Functional clustering of genes within complex biosynthetic pathways

Many prokaryotic organisms contain a streamlined genome, with functionally related, co-regulated genes organized in a linear arrangement under the transcriptional regulation of a single promoter region. This organization, called an operon, represents an efficient mechanism to balance production of multiple components within a metabolic process, playing a critical role in gene expression and organismal survival [40–43]. Operon-like gene clusters are present in at least one eukaryote, although the canonical operon structure is largely absent in eukaryotes on the whole [44].

One characteristic feature of eukaryotic chromosomes is that they exhibit domains, or neighborhoods, of correlated gene expression throughout the genome [45–48]. In S. cerevisiae, there is a broad incidence of locally correlated gene expression, which is distance dependent (the closer any two genes are located the higher their average transcriptional correlation throughout the genome) [49]. This, coupled with the fact that many functionally related genes cluster together non-randomly across the genome, supports the hypothesis that functional clustering represents a fundamental layer of transcriptional regulation for many of these genes [39,45].

Approximately 25% of functionally related gene families are organized into such clusters, and this arrangement increases the transcriptional similarity (as quantified by the Pearson’s correlation coefficient) of these clusters compared to the non-clustered members within the same co-regulated gene family [39]. This phenomenon is not unique to S. cerevisiae – clustering of gene families occurs throughout the phylum for gene families participating in a variety of different molecular processes. In this section, we discuss the types and the incidence of these clusters identified in and observed across Ascomycetes species.

3.1. Secondary metabolite gene clusters

Fungi produce a number of bioactive compounds collectively called secondary metabolites (SMs), molecules that are not required during normal growth. These molecules typically confer a survival advantage exhibiting properties that are antibiotic, anti-proliferative, and catabolic [50]. Many different Ascomycetes produce SMs that are mycotoxins, phytotoxins, and compounds that enhance virulence and pathogenesis [51,52]. The production of SMs by pathogenic fungi can facilitate fungal cooption of a host’s cells, triggering apoptosis and the absorption of host nutrients [52]. Due to the bioactive effects of many SMs, these molecules are of broad interest pharmaceutically and medically for their potentially therapeutic effects [53,54].

3.1.1. Pathogenic toxins and defense

Aspergillus species are non-specific pathogens that infect plants, animals, insects, and immunocompromised people. They cause diseases that cause significant economic impacts in the agricultural industry and produce the potent carcinogen aflatoxin [55]. In A. nidulans the pathway for production of sterigmatocystin, a highly toxic metabolite and a precursor to the aflatoxins, is located as a 23-gene cluster that spans a 54 kilobase genomic region. In A. fumigatus the six genes which are necessary conidial pigment biosynthesis – known to increase the virulence in this species – form a cluster spanning 19 kb (Fig. 2A) [56]. The biosynthetic pathway for the synthesis of the meroterpenoids austinol and dehydroausti- nol in A. nidulans are found as a split cluster of four and 10 genes [57].

Gliotoxin is an ETP (epipolythiodioxopiperaziné)-type fungal toxin produced by A. fumigatus that controls the immune response and induces apoptosis in specific cell types [58]. The biosynthesis of gliotoxin arises from a 16 gene cluster – which includes the synthetic enzymes and the proteins to detoxify and protect A. fumigatus from the effects of gliotoxin [59,60]. This ETP cluster is conserved in many Aspergillus lineages. Although there are differences in the clusters between species, this cluster was most likely the result of an ancestral relationship. The current model is that following assembly of this gene cluster they diverged and diversified during the course of evolution of Ascomycetes [61].

The clustering of functionally related genes across the genome is widespread within this phylum for genes that are necessary for virulence. The fungal pine pathogen, Dothistroma septosporum, produces the potent phytotoxic dothistromin – which is chemically similar to aflatoxin. Biosynthesis of dothistromin requires 19 genes, found in six distinct clusters along the same chromosome [62]. This clustering appears to be the result of a dispersal of an ancestral cluster that exhibited a tighter clustering organizational relationship [62]. Gibberella moniliformis, a plant pathogen that causes ‘bakanee’ disease in rice, contains a cluster of 18 putative genes spanning a 75 kB region for the production of the toxin fumonisin [63].

The rye and grass pathogen, Claviceps purpurea, produces ergot alkaloids – potent bioactive mycotoxins – from a four-gene cluster [64]. The antibiotic viriditoxin exhibits anti-proliferative effects in bacteria (via interactions with FtsZ during division) and in cancer cells [65,66]. This toxin is produced in Paeclomyces variotii and Aspergillus viridinutans via a nine- and eight- gene cluster respectively [67]. Penicillium expansum is a blue mold that causes apple decay and produces the cytotoxic SM patulin, which can be immunosuppressive and carcinogenic [68]. This biosynthetic pathway is encoded by a 15-gene cluster that includes the biosynthetic enzymes and detoxification proteins [69]. Aspergillus terreus has a gene cluster for lovastatin biosynthesis, which is used medically to treat high cholesterol and triglyceride levels in patients [70]. Bikaverin is a reddish pigment produced by Fusarium fungi with reported antibiotic and antitumor properties. This molecule depends, in part, on a six-gene cluster as seen in Fusarium fujikuroi, where the genes for biosynthesis, regulation, and transport are found in a contiguous stretch [71].

Horizontal gene transfer (HGT) refers to the passage of genes between organisms by means other than parent to offspring transmission. While HGT is quite frequent in prokaryotic organisms, it is significantly less common in eukaryotic organisms. The production of toxins confers a growth and a survival advantage, making SM pathways excellent candidates for HGT in eukaryotes. The 23-gene cluster necessary for sterigmatocystin production in A. nidulans, was adopted by the filamentous fungus, Podospora anserine, via horizontal gene transfer from Aspergillus [72]. Likewise, the biosynthetic clusters for gliotoxin production were acquired by horizontal gene transfer in several species [61]. Interestingly, five of the clustered genes involved in the synthesis of bikaverin in Fusarium fungi are identified in the distantly related Botrytis cinerea. Although the regulatory genes maintained their functionality, the others did not. This presents an example of horizontal gene transfer of the entire cluster and suggests this mechanism
might contribute to the incorporation of novel regulators in addition to metabolic pathways [73].

### 3.1.2. Detoxification clusters

Arsenic is a naturally occurring element found in a variety of compounds, many of which can be toxic. In budding yeast, S. cerevisiae, the three genes that confer resistance to arsenic containing compounds, \( ARR1 \), \( ARR2 \), and \( ARR3 \), are clustered together with a 4.2 kilobase region on chromosome XVI [74]. The expression of these genes results in the production of a basic helix-turn-helix transcription factor, arsenate reductase, and a metalloid-proton antiporter that collaborate to protect the cell from the toxin. Fusarium oxysporum contains a two-gene cluster that allows for the detoxification of cyanate (CNO−), a defense compound that is produced by a wide range of organisms that inhibits oxidative phosphorylation through the interference with cytochrome \( C \) function [75,76].

In addition to protection from exogenous toxins, there is an inherent risk to produce SMs that are nonspecific toxins. A form of protection for the host cell can involve a gene (or genes) that offers protection from the toxin, termed the ‘resistance hypothesis’. This can include efflux pumps for export and modification or detoxification enzymes [77]. Leptosphaeria maculans contains a sirodesmin biosynthesis gene cluster, which is a nonspecific mycotoxin. This gene cluster is predicted to contain 18 genes, including several genes that code for P450 cytochrome proteins that serve to detoxify this compound [78]. The necrotrophic fungus, Botrytis cinerea, contains the five-gene cluster for the synthesis of the phytotoxin sequesterterpine and several members of the p450 monoxygenase gene family [79]. It also produces the plant hormone abscisic acid (ABA), coded by a cluster of four genes that include two members of the p450 monoxygenase gene family [80].

### 3.2. Primary metabolite gene clusters

Primary metabolite (PM) biosynthetic genes are those whose products participate in basic metabolic processes that are widespread in many organisms, whereas SMs biosynthetic pathways occur in a limited number of organisms. One of the best characterized functional clustering of a gene family are those of the galactose metabolism genes, which have been extensively characterized in S. cerevisiae. One of the genes is co-localized together on chromosome II: GAL7-GAL10-GAL1 which encode the enzymes galactose-1-phosphate uridyl transferase, UDP-glucose-4-epimerase, and galaktokinase, respectively [81]. Clustering of the galactose metabolism genes in the Saccharomycotina and Candida ancestors, with the S. pombe and S. japonicas acquiring the cluster via HGT from a common ancestor from the Candida cluster [82].

Proline catabolism depends on the activity of a permease, an oxidase, and a P5C (pyrroline-5-carboxylate) dehydrogenase to produce glutamate. In A. nidulans, proline metabolism is dependent on a four-gene cluster not conserved in S. cerevisiae. The biosynthesis of biotin is dependent on a three-gene cluster (Fig. 2B) [83–85]. Nicotinate metabolism is the result of a six-gene cluster that occurs over a 14.4kB region along chromosome VI in A. nidulans [86]. The utilization of alcohol as a carbon source depends on a five-gene cluster found on chromosome VII [87]. Interestingly, there are two additional genes seen in this cluster when compared to the cluster as seen in A. fumigatus, representing a species specific expansion of this cluster [87].

Pichia stipites contains the genes to catabolize the sugar L-rhamnose, of which four of the five genes are clustered [88]. There at least partial conservation of this grouping in members of the subphyla Pezizomycotina and Saccharomycotina [88]. Sulfate assimilation depends on the function of an ATP-sulfursase and PAPS (3-...
phosphoadenine-5'-phosphosulfate reductase, which are found as a gene pair in *A. terreus* [89].

*C. albicans* catabolize phenolic compounds into acetyl coA via the 3-oxoadipate pathway. The genes involved in this are localized into two distinct clusters and are conserved in the *Candida* and ‘CTG’ lineages (species that translate CUG as serine), although with significant evolutionary divergence [90]. The three genes responsible for GlcNAc metabolism into fructose-6-phosphate are found clustered in *C. albicans*. This arrangement is conserved in *Trichoderma reesei*, along with the transcription factor, *RON1* that regulates expression of the cluster [91].

In the case of PMs, there is an advantage to the clustering of genes; however, it is not a ubiquitous phenomenon. The production of SMs demonstrates an advantage to the maintenance of clusters to maintain integrity of the product and to deal with the potential ramifications of toxicity. This selection does not appear to be the case for PMs which exhibit more variation between species.

### 3.3. Superclusters

The descriptor of superclusters refers to the functional clustering of genes that code for the production of multiple SMs or PMs that have a complex relationship. COR (corydinepep, or 3'-deoxynadenosine) is produced by *Cordyceps militaris* from a four-gene cluster that also includes the enzyme to produce PTN (pentostatin, or 2'-deoxyxycinomycin), an inhibitor of deoxynucleosine deaminases (Fig. 2C). Both of these are bioactive molecules – COR has antibiotic and anti-inflammatory properties and PTN has chemotherapeutic effects [92,93]. Interestingly, PTN production prevents the deamination of COR to 3'-deoxynosine by endogenous enzymes in *C. militaris* [94]. COR and PTN biosynthesis is mediated by a single gene cluster and PTN prevents COR deamination, potentially through enzyme inhibition of endogenous deoxynucleosine deaminases present in *C. militaris*. Another example of a supercluster is seen in *A. fumigatus*, where the genes for the production of fumagillin and pseudotin are localized to the subtelomeric region of chromosome VIII. There are fifteen genes localized to this cluster, presenting a complex intertwined relationship for these two seemingly unrelated chemicals [95]. This genomic arrangement is conserved, albeit with significant rearrangements in related species [95].

### 3.4. Regulons

While operons are primarily a prokaryotic phenomenon, eukaryotic organisms contain regulons; functionally related gene families that are co-regulated and spread throughout the genome. Canonical examples include the genes involved with the biogenesis of the ribosome – both the ribosomal protein (RP) gene family and the rRNA and ribosome biosynthesis (RBB, or *Ribi*) gene family – are found clustered together throughout the genome in *S. cerevisiae* [96]. This distribution is non-random and highly unlikely to occur by chance [96,97]. This distribution is not limited to budding yeast, the clustering of both gene families is conserved in both *C. albicans*, and the distantly related *S. pombe*. The identity of the individual paired genes differed among species, however the absolute numbers of clustered genes (e.g. the overall number of pairings that exist) is similar [96,98]. Systematic analysis revealed that approximately 25% of all functionally related gene families exhibit a non-random genomic distribution as clustered pairings in *S. cerevisiae*. This genomic distribution results in a tighter transcriptional response during cellular changes when compared to the singleton (non-clustered) members within the same family [39].

### 4. Transcriptional co-regulation within a neighborhood may drive functional clustering

In addition to the repressive effects that can occur based on proximity to heterochromatic regions, certain genomic regions are more susceptible to transcriptional disruption during the course of genomic manipulations. The advent of genomic libraries allows for high throughput genetic screening in many organisms, and a wealth of resources exist for *S. cerevisiae*. Genetic manipulation alters expression of genes surrounding the site of manipulation at 7–15% of targeted regions. This led to the misannotation of over nine-thousand genetic interactions in systematic screens [99,100]. Furthermore, genomic integration sites are responsible for 13-fold variation in protein levels [101]. The disruption of gene expression via the integration of highly expressed reporters may lead to this phenomena through transcriptional interference and repression of the adjacent gene(s) [102,103].

Aside from transcriptional interference, there is evidence that the non-random genomic distribution of functionally related genes is essential to their proper transcriptional regulation via shared promoter elements. Using a RBB gene pair as a test, functional dissection of the *MPP10-MRX12* pairing identified both genes share a common promoter region. What is notable about this pairing is that the genes are oriented in a convergent manner (→←), meaning that the canonical promoter elements, termed PAC and RP, exert influence over a 40kB genomic region [104]. Both promoter elements are localized upstream of *MPP10*, there are no identifiable motifs in the promoter region of *MRX12*, and their mutation uncouples transcription of this pair from the rest of the regulon. Transcription of the pair is disrupted by their separation, and proper expression depends on chromatin remodeling and transcription factor binding to the promoter of *MPP10* [105]. This phenomenon is termed ‘adjacent gene co-regulation’ (AgcoR).

There are several conformations for functionally related gene clusters, including a divergent orientation (Fig. 3A). This confirmation results in cis regulatory elements between the pair of genes, where it can act as a bidirectional promoter, which is quite common in yeast [106]. Much less common are the tandem (→←) and convergent orientations (Fig. 3B and C). In these arrangements

---

**Fig. 3.** Representative clustering arrangements observed in *S. cerevisiae* at three RBB paired loci. (A) The *CDC10-NOP2* locus, (B) the *RRP15-NOC4* locus, and (C) the *RPF1-GAR1* locus are shown relative to the PAC and RP cis-regulatory promoter motifs shown in purple and red, respectively. All loci are to approximate scale, although the relative size of the promoter motifs has been exaggerated to highlight their location and arrangement. Image is . (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.) adapted from [97]
co-expression and transcriptional regulation must occur across a larger distance – which is not as common in budding yeast [107]. Interestingly, the convergent gene pair RRP15-NOC4 have the PAC and RRPE elements localized upstream of one member of the pair, as in the case with MPP10-MRX12. The pairing RPF1-GAR1 have the two elements split, with each gene containing either the PAC or the RRPE promoter motif. Though not functionally dissected, it would be quite interesting to observe the transcriptional regulation across a broad genomic distance once it is completed.

Transcriptional regulation at a distance depends, in part, on alterations to the chromatin underlying a genomic neighborhood extending across a broad region [108]. The constraints that these could impose would result in certain portions of the genome being more permissive to local influencing regulators. This is a particularly attractive model based on the diverse nature of gene families that are clustered and do not necessarily share the same promoter motifs. In support of this model, the ribosomal proteins, the nitrogen metabolism, and the toxin response genes, cluster into genomic regions that are more susceptible to transcription at a distance during induction of the stress response [108]. Data was extracted and the pairwise Spearman’s correlation coefficient was plotted as a function of genomic distance (between the transcription start sites for each gene) for every member of each family. The members of each family were divided into plots for the singleton members (Fig. 4A) and the clustered members (Fig. 4B). Both plots reveal a positive correlation in expression regardless of genomic confirmation across the stress response. These data fit a logarithmic decay curve, which was overlaid as a separate plot (Fig. 4C). The transcriptional similarity is greater and spreads across a longer genomic distance in the clustered gene members – indicating that regardless of function, clusters are localizing to transcriptionally permissive regions of the genome (at least for these gene families). This would potentially explain how there are similar numbers of clusters within the RP and RRB gene families throughout Ascomycota, though the identities of the pairings differ greatly [96].

5. Computational approaches to identify and characterize clusters

The identification and characterization of co-regulated functional clusters as described throughout this work is currently an active area of research. This primarily involves two interrelated components – the first is the identification of co-expressed genes and the second is the characterization of the co-localization of two genes to the same genomic locus. In this section, we describe several tools employed to identify correlated expression from gene expression datasets.

5.1. Identification of functional clustering relationships using Euclidean distance from gene expression datasets

Identification of the complete membership of co-regulated gene families is a fundamental challenge for researchers, but can yield rewarding applications. One approach that successfully identified the membership of the ribosome biogenesis regulon is through the utilization of the function daisy from the S library [109]. Through the analysis of the budding yeast stress-response datasets, the Euclidean distance for all budding yeast genes (from the composite average gene expression response of the RRB gene family) was determined. This resulted in the expansion of the gene family to 188 members – many of which were clustered throughout the genome and had yet to be annotated [97]. The metric utilized is defined as:

$$d(i,j) = \sqrt{\sum_{k=0}^{n} (x_{ik} - x_{jk})^2}$$

where $x_{ik}$ and $x_{jk}$ denote the expression levels of genes $i$ and $j$ at time point $k$ [97]. This approach represents a straightforward method to mine gene expression data when the transcriptional behavior of a gene family is known beforehand.

5.2. Pearson’s correlation coefficient analysis of gene expression data

The transcriptional similarity between two genes can be calculated by the Pearson’s correlation coefficient (PCC). This approach is straightforward and represents a standard analysis that can be applied to gene expression data to identify linear relationships among genes, yet powerful enough determine correlated networks and similarities [39]. To calculate the PCC between two genes, $X$ and $Y$, across a series of $N$ conditions:

$$S(X, Y) = \frac{1}{N} \sum_{i=1}^{N} \left( \frac{X_i - X_{\text{offset}}}{\Phi X} \right) \left( \frac{Y_i - Y_{\text{offset}}}{\Phi Y} \right)$$

where:

$$\Phi G = \frac{1}{N} \sum_{i=1}^{N} (G_i - G_{\text{offset}})^2$$

where $G_{\text{offset}}$ is set to the reference state in each data set.

Fig. 4. Functionally clustered genes localize to genomic loci that are more conducive to regulation at a distance. The pairwise Spearman’s correlation coefficient was determined for the nitrogen metabolism, ribosomal protein, and toxin response genes in S. cerevisiae as described [108]. SCC versus genomic distance is plotted for (A) the non-clustered, singleton members of each set and (B) the clustered members of each set and the data was fit to a logarithmic decay. For clarity and ease of comparison, the decay curves are overlaid in (C).
5.3. Spearman’s correlation coefficient analysis of genomic neighborhoods

The Spearman’s correlation coefficient (SCC) is similar to the PCC, although without the magnitude component utilized in the PCC. The use of the SCC is ideal for characterizing gene expression throughout a broad region and uncovering more subtle transcriptional effects within a genomic locus or neighborhood [49]. The SCC is computed utilizing the formula:

\[ \rho = \frac{\text{cov}(g_1, g_2)}{\sigma_{g1}\sigma_{g2}} \]

where \( g_1 \) and \( g_2 \) are the corresponding genes for comparison, \( \text{cov} \) is their covariance, and \( \sigma \) is the standard deviation of expression. Such analysis can also give an accurate representation of the transcriptional effects that ripple outward as the result of transcriptional activation and that may be indirect as a function of proximity [16,49].

5.4. The multivariate copula model for analysis of gene expression data

An addition to the computational toolkit that allows for modeling the directional dependence of two genes is through the application of a multivariate copula model (MCM) [110]. This represents a powerful computational approach that can help to construct the gene interaction networks and identify co-regulated gene families. This approach, when applied to gene expression datasets, allows for the reconstruction of regulatory relationships even when the data are non-linear [111]. The MCM is an enhancement over the use of the Pearson’s and Spearman’s correlation coefficients that can allow of the characterization of complex networks and relationships – as evidenced by the identification of network of chromosomal maintenance and cell-cycle regulating genes [112].

5.5. Determining the statistical significance of genomic arrangement through the with a hypergeometric distribution

Characterization of functional clustering observed in co-expressed gene families has been primarily done via manual curat. While this can be laborious at time, the systematic nomenclature adopted by many organisms allow for rapid searching and querying for identification of clusters. Once the genomic distribution of a gene set has been characterized, the probability of this distribution can be calculated. The statistical significance for the genomic distribution of a functionally related gene family is determined by calculating the binomial probability for the arrangement. The chance probability that there would be \( j \) adjacent genes within a regulon of size \( M \) genes is:

\[ 1 - \sum_{k=0}^{j} \binom{M!}{k!(M-k)!} p^k (1-p)^{M-k} \]

where \( N \) is the total number of genes present within \( S. \text{cerevisiae} \) (total number of genes after deduction of dubious open reading frames).

6. Summary and outlook

The physical linkage of genes into functional clusters throughout the genome can buffer the effects of stochastic noise in gene expression. Even in isogenic, clonal populations of cells there can be considerable variation in the levels of expression from similar genetic constructs [113,114]. One model for selective pressures that favor clustering is that this arrangement would minimize the effects of stochastic noise. This is advantageous when dealing with biological pathways that produce potentially toxic intermediates, as seen with the galactose metabolism and the tyrosine biosynthetic pathways. The clustering of the genes within those pathways can limit variations in the levels of component enzymes – minimizing the risk of toxin buildup [115]. The risk of toxin production is significant, when the GAL genes are not clustered in \( S. \text{cerevisiae} \) there is significantly higher levels of the toxic galactose-1-phosphate and reduced cell viability [116]. Gene linkage via functional clustering would also minimize copy number variations that would occur during DNA replication, synchronizing the timing together.

A clustering arrangement would also be advantageous when dealing the production of large macromolecular complexes where stoichiometric levels of proteins are necessary. Some of these complexes, such as the ribosome, consume significant energetic reserves [117]. Transcription of the component genes to balance the production and abundance of mRNA as needed, allocates the limited cellular resources to minimize waste [118]. In addition to efficient energetic expenditures, the unbalanced production of the components to produce a ribosome results in orphan ribosomal proteins (RPs). The presence of these extra-ribosomal RPs perturbs cellular proteostasis, activates the expression of the heat shock activated transcription factor, \( HSF1 \), and reduces cellular fitness [119,120].

There are thousands of species in the fungal kingdom – and in the phylum Ascomycota – yet to be discovered and are still being characterized. There will undoubtedly be many novel SMs identified. Characterization of biosynthetic gene clusters that produce these SMs will provide valuable insight to the target and mechanisms of these SMs, and the defense and protective genes present to ensure secondary metabolites do not harm the host. [121]. This could offer therapeutic targets for emerging pathogens, provide genetic modifications for the protection of food crops, and offer novel pathways for industrial and pharmaceutica biosynthesis. The farther the levels of transcription deviate from the optimal expression level for a gene, the more advantageous the tolerance for transcriptional noise [122]. This is particularly germane for researchers modifying organisms to recreate biosynthetic pathways in organisms that are easier to cultivate, including \( S. \text{cerevisiae} \). The site of manipulation and modification should be chosen carefully in order to maximize organismal fitness, biosynthetic output, and metabolite yield [123].

The functional clustering of co-regulated genes – and the effects of positional expression – are not limited to Ascomycetes or to fungi. It has long been known that the effects of heterochromatin, as in the TPE, are conserved in more complex eukaryotic organisms, including humans [124]. There is conservation of the molecular mechanisms as well [125]. In addition to these repressive effects, there is a global correlation between proximity of two genes and transcriptional similarity [49]. Transcriptional activation of a gene can alter transcription across genomic ‘neighborhoods’, activating expression of the neighboring genes across a large distance [16]. Comprehensive analysis and further elucidation of the mechanisms underlying this phenomena are essential to fully understanding the relationship between the clustering of related genes and co-regulation – and a systems level understanding of transcription within the cell.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
Acknowledgements

The authors would like to acknowledge and thank Kelley Kealey (William Paterson University, NJ) for providing the Penicillium images in Fig. 1. The authors would like to acknowledge David White (Denison University, OH) for his helpful discussions on the manuscript. The authors acknowledge support for this work from the Office of the Provost for Assigned Release Time (ART) and the Office of the Dean of the College of Science and Health Center’s Research (CIR) at William Paterson University. The authors would also like to acknowledge the reviewers of this manuscript, who provided thoughtful suggestions that enhanced this work.

References

[1] Sekhon RS, Lin H, Childs KL, Hansey CN, Bueler CR, et al. Genome-wide atlas of transcription during maize development. Plant J 2011;66:553–63.

[2] Mohammed H, Hernando-Herrera I, Savino A, Scaldone A, Macaulay I, Mulas C, Chandra T, Voet T, Dean W, Nichols J, Marioni JC, Reik W. Single-cell landscape of transcriptional heterogeneity and cell fate decisions during mouse early gastrulation. Cell Reports 2017;20:5:1215–28. https://doi.org/10.1016/j.celrep.2017.07.009.

[3] Vork PJ, Ohm RA. The role of homoeodomain transcription factors in fungal development. Fungal Biol Rev 2018;32(4):219–30. https://doi.org/10.1016/j.fbr.2018.04.002.

[4] Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO, Silver PA. Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes. MBio 2000;11(12):4241–57. https://doi.org/10.1128/mBio.01225-05.

[5] Storz G, Hengen R. editors. Bacterial Stress Responses. Washington, DC: ASM Press; 2010.

[6] Pakos-Zebrucka K, Koryga I, Mnich K, Ljujic M, Samali A, Gorman AM. The evolutionary significance of CpG methylation in the control of transcription and cellular stress responses in fungi. Cell Biol Rep 1997;13:1–5.

[7] Gottschling DE, Aparicio OM, Billington BL, Zakian VA. Position effect at S. cerevisiae telomeres: reversible repression of Pol II transcription. Cell 1990;63(4):751–62. https://doi.org/10.1016/0092-8674(90)90141-Z.

[8] Renaud H, Aparicio OM, Zierath PD, Billington BL, Chhablani SK, Gottschling DE. Silent domains are assembled continuously from the telomere and are defined by promoter distance and strength, and by SIR3 dosage. Genes Dev 1993;7(11):1331–45. https://doi.org/10.1101/gr.7.11.1333.

[9] Moradi-Fard S, Sarthi J, Tittel-Elmer M, Lalonde M, Cusaneii E, et al. Smc5/6 is a telomere-associated complex that regulates Sir4 binding and TPE. PLoS Genetics 2016;12:e1006068.

[10] Murphy D, Louis EJ. Limitations of silencing at native yeast telomeres. EMBO J 1999;18:258–50.

[11] Ellahi A, Thurtle DM, Rine J. The Chromatin and Transcriptional Landscape of Native Saccharomyces cerevisiae Telomeres and Subtelomeric Domains. Genetics 2005;200(2):505–21. https://doi.org/10.1534/genetics.105.175711.

[12] De ALP, Júarez-Cepeda J, López-Fuentes E, Briones-Martin-Del-Campo M, Gutiérrez-Escobedo G, et al. Local and regional chromatin silencing in Candida glabrata: consequences for adhesion and the response to stress. FEMS Yeast research 2015;15:2018.

[13] Buscaino A. Chromatin-Mediated Regulation of Genome Plasticity in Human Fungal Pathogens. Genes 2019;10:855.

[14] Bühler M, Caster SM. Silent chromatin at the middle and ends: lessons from yeasts. EMBO J 2009;28:2149–61.

[15] Freire-Benétrez V, Price RJ, Tarrant D, Berman J, Buscaino A. Candida albicans repetitive elements display epigenetic diversity and plasticity. Sci Rep 2016;6(1). https://doi.org/10.1038/srep29895.

[16] Ayoub N, Goldshmidt I, Cohen A. Position effect variegation at the mating-type locus of fission yeast: a cis-acting element inhibits coexpressed genes in the silent and expressed domains. Genetics 1999;152:495–508.

[17] Eldabagh RS, Mejia NG, Barrett RL, Moncoro CR, So MK, et al. Systematic identification, characterization, and conservation of adjacent-gene coexpression in the budding yeast Saccharomyces cerevisiae. Mpsphere 2018:3.

[18] Jacob F, Monod J. Genetic mapping of the elements of the lactose region in Escherichia coli. Biochemical and Biophysical Research Communications 1965;18:693–701.

[19] Kolter R, Yanovsky S. Attenuation in Amino Acid Biosynthetic Operons. Annu. Rev. Genet. 1982;16:113–34. https://doi.org/10.1146/annurev.ge.16.120182.000553.

[20] Osborn AM, Bruce KD, Strike P, Ritchie DA. Distribution, diversity and evolution of the bacterial mercury resistance (mer) operon. FEMS Microb Rev 1997;19(4):239–62. https://doi.org/10.1016/S0168-6445(97)00030-0.

[21] Ren Fekih I, Zhang C, Li YP, Zhao Y, Alwathnani HA, et al. Distribution of arsenic resistance genes in prokaryotes. Frontiers in microbiology 2018:9:2473.

[22] Blumenthal T, Davis PS, Carrido-Lecca A. Operon and non-operand gene clusters in the C. elegans genome. WormBook: The Online Review of C. elegans Biology [Internet]. WormBook, 2018.

[23] Cohen BA, Mitra RD, Hughes JD, Church GM. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. Nat Genet 2000;26(2):183–6. https://doi.org/10.1038/79896.

[24] Schmid M, Davison TS, Hett SR, Pape U, Demar M, Vingron M, Schölkopf B, Weigel D, Lohmann JU. A gene expression map of Arabidopsis thaliana development. Nat Genet 2005;37(5):501–6. https://doi.org/10.1038/ng1542.

[25] Spellman PT, Rubin GM. A large dataset of similarly expressed genes in the Drosophila genome. J Biol 2002;1:1–8.

[26] Williams EJ, Bowles DJ. Coexpression of neighboring genes in the genome of Candida tropicalis. Genome Biology 2004;5:1406–7.

[27] Quintero-Cadena P, Sternberg PW. Enhancer Sharing Promotes Neighborhood of Transcriptional Regulation Across Eukaryotes. G3 2016;6(12):4167–74. https://doi.org/10.1534/g3.116.036228.

[28] Vining LC. Functions of Secondary Metabolites. Annu Rev Microbiol 1990;44(1):395–427. https://doi.org/10.1146/annurev.mi.44.110190.002143.

[29] Fox EM, Howlett BJ. Secondary metabolism: regulation and role in fungal biology.Curr Opin Microbiol 2008;11(6):481–7. https://doi.org/10.1016/j.mib.2008.10.007.
Martens JA, Laprade L, Winston F. Intergenic transcription is required to repress the Saccharomyces cerevisiae SER3 gene. Nature 2004;429:571–4.

Martens JA, Wu P-Y-J, Winston F. Regulation of an intergenic transcript controls adjacent gene transcription in Saccharomyces cerevisiae. Genes Dev 2005;19:2695–704.

Bosio MC, Negri R, Dieci G. Promoter architectures in the yeast ribosomal expression program. Transcription 2011;2:71–7.

Arnone JT, Arace JR, Soorneedi AR, Citino TT, Kamitaki TL, et al. Dissecting the cis and trans elements that regulate adjacent-gene coregulation in Saccharomyces cerevisiae. Eukaryot Cell 2014;13:738–48.

Xu Z, Wei W, Gagneur J, Perocchi F, Clauder-Münster S, et al. Bidirectional promoters generate pervasive transcription in yeast. Nature 2009;457:1033–7.

Dobi KC, Winston F. Analysis of transcriptional activation at a distance in Saccharomyces cerevisiae. Mol Cell Biol 2007;27:5575–86.

Cera A, Holganza MK, Hardan AA, Eldadgh RS, et al. Functionally Related Genes Cluster into Genomic Regions That Coordinate Transcription at a Distance in Saccharomyces cerevisiae. Msphere 2019;4.

Xu H, Liu J-J, Liu Z, Li Y, Jin Y-S, et al. Synchronization of stochastic expressions drives the clustering of functionally related genes. Sci Adv 2019;5:eaaax6525.

McGary KL, Slot JC, Rokas A. Physical linkage of metabolic genes in fungi is an adaptation against the accumulation of toxic intermediate compounds. Proc Natl Acad Sci 2013;110:11481–6.

Xu H, Liu J-J, Liu Z, Li Y, Jin Y-S, et al. Synchronization of stochastic expressions drives the clustering of functionally related genes. Sci Adv 2019;5:eaaax6525.

Warner JR. The economics of ribosome biosynthesis in yeast. Trends Biochem Sci 1999;24:437–40.

Li G-W, Burkhart D, Gross C, Weissman JS. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. Cell 2014;157:624–35.

Sorger PK, Pelham H. Purification and characterization of a heat-shock element binding protein from yeast. EMBO J 1987;6:3035–41.

Tye BW, Commins N, Ryazanova LV, Wuhr M, Springer M, et al. Proteotoxicity from aberrant ribosome biogenesis compromises cell fitness. Elife 2019;8:e43002.

Keller NP. Translating biosynthetic gene clusters into fungal armor and weaponry. Nat Chem Biol 2015;11:671–7.

Duvau F, Hodgins-Davis A, Metzger BP, Yang B, Tryban S, et al. Fitness effects of altering gene expression noise in Saccharomyces cerevisiae. Elife 2018;7:e37272.

Kim J-M. Analysis of directional dependence using asymmetric copula-based regression models. J Stat Comput Simul 2014;84:1990–2010.

Kim J-M, Jung Y-S, Sungur EA, Han K-H, Park C, et al. A copula method for modeling directional dependence of genes. BMC Bioinf 2008;9:225.

Elowitz MB, Levine AJ, Sigga ED, Swain PS. Stochastic gene expression in a single cell. Science 2002;297:1183–6.

Blake WJ, Katz M, Cantor CR, Collins JJ. Noise in eukaryotic gene expression. Nature 2003;422:633–7.

McGary KL, Slot JC, Rokas A. Physical linkage of metabolic genes in fungi is an adaptation against the accumulation of toxic intermediate compounds. Proc Natl Acad Sci 2013;110:11481–6.

Xu H, Liu J-J, Liu Z, Li Y, Jin Y-S, et al. Synchronization of stochastic expressions drives the clustering of functionally related genes. Sci Adv 2019;5:eaaax6525.

Warner JR. The economics of ribosome biosynthesis in yeast. Trends Biochem Sci 1999;24:437–40.