TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Genomorientierte Bioinformatik

Comprehensive Discovery of Fungal Gene Clusters: Unexpected Co-work Reflected at the Genomic Level

Wanseon Lee

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender:		Univ.-Prof. Dr. W. Liebl
Prüfer der Dissertation:	1.	Univ.-Prof. Dr. H.-W. Mewes
	2.	Univ.-Prof. Dr. D. Frischmann

Die Dissertation wurde am 14.04.2010 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 26.08.2010 angenommen.

Acknowledgement

It's a great pleasure to thank people who made my thesis possible. My path to doctorate works has been neither easy nor straightforward, but I have met many good people and gained a lot of scientific and life experiences. Those years would have been much more difficult without the assistance of those who gave their support in different ways.

First I would like to thank my supervisor Prof. Hans-Werner Mewes for giving me the opportunity to work in MIPS. I feel very lucky that I had the opportunity. Thank you, Professor. I would also like to express my appreciation to the other members of my thesis committee, Prof. Frishman, my second supervisor and Prof. Liebl, my committee chair for accepting my asking at short notice.

I would like to especially thank members of fungal group. Very special thank goes to Dr. Ulrich Güldener, a leader of fungal group, for his guidance and support throughout my research. He was a wise adviser who knows my points for my often stupid questions. I am deeply indebted to Philip for his constant helps and useful discussion. I am also greatly thankful to Dr. Gertrud Mannhaupt and Dr. Martin Münsterkötter for reviewing my works and helping to makee them so much smoother. I would like to thank to all members for evaluating and giving their suggestions to improve my manuscript.

Many colleagues have contributed in diverse manners. Brigitte, Elisabeth, and Gabi gave me practical information to draw up many documents and manage official procedures for PhD works as former PhD students during my urgent or sudden asking. Thank you, Brigitte, Elisabeth, and Gabi. I also appreciate useful discussions and cheerful conversation with Yu, Alexey, and my office mate, Pawel. I would like to thank Giovani for solving many machine troubles and interest of Korean foods, Igor for statistical advice, Igor in Canada for technical support for the Netbean application, and Mathias for helping in Pedant-data access.

I would like to express my gratitude to Filka for helping me to make an adjustment to a new environment. Specially, I won't forget the first German sauna experience with her. I should also express my appreciation of Goar's encouragement to take some exercise for a lazy girl. I am grateful to the secretaries, Elisabeth and Petra for assisting me in many different ways.

I am grateful to Prof. Cho for introducing MIPS to me. Prof. Cho at Busan National Uni., my advisor of Master thesis and Dr. Kim, CEO of ISTECH Corp. deserve special thanks. Thanks to them, my careers and skills in both academia and industry broadened my perspective on the practical aspects in my PhD works.

Also outside of science, there were people whose help and their support was very important. I like to express further greatest thanks to two Yangs for a patient hearing to my whining as well as encouraging words. It was a pleasure to know Korean sisters (Jiyeon, Yukyong, Seehie, Hyunjeong, and Sunjung) in Munich. All have been truly amazing friends, and I am so thankful for all their concerns about me. My sincere thanks go to my friends, Mi-hyang, Kwanseon, Dr. Oh, Dr. Um, Dr. Kang, Dr. Song, Minyong, and Minjung in Korea and Alesia in Australia for continuous support like encouragement, alcohol etc.

Last but by no means least, I owe to my parents and my brother for their immense support, love and belief in me throughout my life. 엄마, 아빠 감사합니다.

Again, I offer my regards and blessings to all of those who always encouraged and supported me in any respect during the completion of my PhD work. Vielen Dank!

Contents

1 Zusammenfassung

In Pilz-Genomen sind Gene des Sekundär-Metabolismus oft räumlich auf den Chromosomen in Clustern angeordnet. In der Forschung sind sie von großem Interesse wegen ihrer oft toxischen Produkte (Mycotoxine) aber auch, wie etwa bei Penicillin, wegen möglicher pharmazeutischer Anwendungen. Die Identifizierung von bislang unbekannten Clustern ist allerdings schwierig, da die Gene unter Laborbedingungen meist nicht exprimiert werden. Die in silico Analyse ist daher ein vielversprechender Weg um neue potentielle Cluster vorherzusagen, die dann experimentell untersucht werden können.

Die vorliegende Arbeit hat die Identifizierung neuer Pilz-Gen-Cluster in zwei Pflanzenpathogenen zum Ziel, der necrotrophen Species *Fusarium graminearum* und dem biotrophen Basidiomyzeten Ustilago maydis. Hierzu wurde basierend auf Eigenschaften von bekannten Gen-Clustern ein Screening-Verfahren entwickelt, das auf drei Ebenen ansetzt. Es wurde nach speziellen funktionellen Eigenschaften der Gene sowie nach Co-Expression und gemeinsamen, möglicherweise regulatorischen Motiven in den Promotoren benachbarter Gene gesucht. Die Identifizierung der fünf bekannten Cluster aus F. graminearum bestätigte die Funktion des entwickelten Screening-Verfahrens. Insgesamt wurden basierend auf Schlüsselenzymen (PKS, NRPS, Cytochrome P450s) 77 vorläufige, funktionelle Gen-Cluster (TFCs) gefunden, die wahrscheinlich im Sekundärmetabolismus eine Rolle spielen. Die Gene von 23 dieser Cluster zeigten zudem Co-Expression und/oder wiesen eine gemeinsame, putativ regulatorische Bindestelle in ihren Promotoren auf. Die Suche nach Bindestellen fand neben einer ganzen Reihe von möglichen Clustern auch das Trichothecen Cluster mit einer experimentell validierten Transkriptionsfaktor-Bindestelle. Die Integration aller drei Ansätze in das Screening-Verfahren bestätigt bzw. erweitert bekannte Mycotoxin Gen-Cluster und zeigt putative, bislang nicht bekannte Gen-Cluster auf.

Die Analyse des *Ustilago maydis* Genoms ergab unerwartete Hinweise auf ein potentielles Kluster mit necrotrophen Attributen in dem normalerweise biotroph lebenden Pilz. Zudem wurden gemeinsame, regulatorische Elemente in Promotoren der zuvor beschriebenen Gen-Cluster gefunden, die für sekretierte Proteine kodieren. Die durch verschiedene Evidenzen gestützten Ergebnisse ermöglichen somit weitere Studien noch unbekannter Funktionen und ihrer Produkte in *F. graminearum* und *U. maydis*. Darüber hinaus können die entwickelten Methoden generalisiert und für die Suche nach funktionellen Modulen, repräsentiert durch Gen-Cluster, in jedem Pilzgenom angewendet werden.

2 Abstract

Fungal genes which are involved in pathways synthesizing secondary metabolites are often found to be physically clustered on chromosomes. These genes are of major interest due to toxic properties of their products like mycotoxins and/or pharmaceutical applications like penicillin. However, it is difficult to identify novel fungal gene clusters as most of them appear to be silent under laboratory conditions. Thus, efficient screening by in silico analysis is a promising approach to find new candidates of biosynthetic pathway gene clusters for subsequent studies.

This thesis focuses on the identification of known and novel fungal gene clusters in two plant pathogens, the necrotroph *Fusarium graminearum* and the biotroph *Ustilago maydis*. For this purpose, a systematic screening through a comprehensive approach was developed based on three properties investigated from known fungal gene clusters: particular compositions of gene functions, co-expression, and the presence of common putative promoter motifs in promoters of neighboring genes. Applying the screen to *F. graminearum*, five known mycotoxin gene clusters were re-discovered as a proof of concept. 77 tentative functional gene clusters (TFCs), probably acting in secondary metabolism were deduced based on key enzymes (e.g. secondary metabolites, cytochrome P450s). 23 TFCs are supported by two crucial evidences for co-regulation of fungal gene clusters, either specific co-expression patterns or conserved promoter motifs. Exploration of promoters for common putative regulatory motifs detected among numerous putative secondary metabolite gene clusters also the trichothecene mycotoxin genes with an experimentally verified common regulatory motif. Integration of the three independent result sets supports re-demarcation of mycotoxin gene clusters as well as finding putative new gene clusters.

In *U. maydis*, our analysis revealed potential necrotrophic gene clusters which may provide an unprecedented insight into the biotrophic fungus. Moreover, a possible mechanism of co-expression via a common regulatory element for the secreted protein clusters is proposed. Our results, highlighted by different evidences, will lead towards further studies of so far unknown products and/or biological functions in *F. graminearum* and *U. maydis*. In addition, the procedures presented are general and may present a new paradigm for the discovery of functional modules based on locally clustered genes in fungal genomes.

3 Introduction

It is no longer rational to suppose that gene order is randomly organized in eukaryotic genomes. In well-studied genomes, genes of similar and/or coordinated expression tend to be linked [1]. Clustered genes have been found to share diverse functional relations such as having membership in the same metabolic pathway and protein complexes [1,2]. In fungal genomes, genes involved in particular metabolic pathways or common functions have been observed to form tightly linked physical clusters on the chromosome maintaining functional clusters [3,4,5,6]. These clustered genes perform a broad range of functional roles that are of major interest by virtue of their toxic properties and/or pharmaceutical application: production of mycotoxins such as aflatoxin, ochratoxin and trichothecene; synthesis of penicillin and cyclosporine; pigmentation by melanin; nutrient utilization including nitrogen and carbon. Despite importance to our health and economic impact, it is difficult to identify fungal gene clusters because most clustered genes are not constitutively expressed and some appear to be silent under laboratory conditions [7,8].

A new approach which is more fungal gene specific in terms allowing the discovery of rapidly evolved species specific gene clusters would be of great value for the study of fungal gene organization and evolution. In this study, three properties of known fungal gene clusters were investigated and the insight gained was used to design an approach for comprehensive discovery of fungal gene clusters at the genomic level. Our analyses identified five known mycotoxin gene clusters from *F. graminearum* and secreted protein clusters from *U. maydis* serving as a proof of concept. We can further propose several putative novel gene clusters classified by diverse evidences, suggesting new functional relationships such as metabolic pathways.

3.1 Hallmark and importance of fungal gene clusters

3.1.1 Secondary metabolism

Fungi produce a huge number of secondary metabolites, low-molecular-weight natural products that are not necessary for normal growth. Genes that synthesize common secondary metabolites have been identified as groups of genes that are co-regulated and physically linked on chromosomes. Products of fungal secondary metabolites are often toxic; antibiotics to bacteria, phytotoxins to plants, mycotoxins

to vertebrates and other animal groups [9]. Many fungal toxins produced by secondary metabolism have multiple effects with overlapping toxicities to diverse hosts, invertebrates, plants, and microorganisms [10].

The most well known fungal secondary metabolites are the β-lactam antibiotics such as penicillins by several *Penicillium* species and *Aspergillus nidulans* and cephalosporins by *Acremonium chrysogenum* [11,12,13,14,15,16,17]. Penicillin biosynthesis is catalyzed by three enzymes encoded by pcbAB, pcbC, and penDE that are organized into a cluster at a single locus [11,12,13,14,15]. Cephalosporin biosynthesis genes are separated in at least two clusters that contain cefEF and cefG at one loci in addition to pcbAB and pcbC at another loci [15,16,17]. The fungal β-lactam biosynthesis genes are controlled by intricate regulatory mechanisms. Diverse environmental influences such as carbon source, light, and pH elicit production of differently β-lactams [18]. The pH dependant regulator (PacC) and the FadA G-protein α-subunit have positive effects on penicillin production whereas light (VeA) mediating regulator and carbon catabolites are associated with negative regulation of penicillin [18,19].

Plant pathogenic fungi produce a range of secondary metabolites that often cause diseases in target hosts. According to fungal lifestyles during host infection, plant pathogenic fungi can be divided into three groups; necrotrophs that kill plant cells to derive nutrition, biotrophs that require a living host to complete their life cycle, and hemi-biotrophs that usually have an initial biotrophic growth phase and switch to a necrotroph phase to kill the host [20]. Diverse secondary metabolites producing toxins and causing diseases have been reported mostly in studies of necrotrophs that are highly destructive to host tissues and which obtain their nutrients from dead plant tissues. These fungi establish inside the host tissue by releasing diverse enzymes to macerate, which disrupt tissue integrity and cause death of hosts. Diverse toxins are usually classified as host-selective or non-selective. Host-selective (specific) toxins (HSTs) have limited host range and kill plant cells by targeting specific enzymes or components [21,22]. Most known HSTs are produced by species of the genera *Alternaria* and *Cochliobolus* [22,23]. They are some of the most notorious fast acting compounds that produce disease symptoms in susceptible plants. Host non-selective toxins have activity on the host as well as on non-host plants and target diverse cellular components. A well studied example of a non-host toxin is cercosporin produced by many species of the genus *Cercospora*. The toxin has been reported to have effect on several hundred plant species as well as on mice and bacteria [24,25].

Secondary metabolites are also important in mediating interactions between fungus and host. In the rice blast fungus *Magnaporthe grisea*, an ACE1 secondary metabolite gene encoding a hybrid

polyketide synthase/non-ribosomal peptide synthetase is involved in recognition of particular rice cultivars [26]. The ACE1 gene is located in a cluster containing 15 genes that is specifically expressed during penetration into the host and defined as infection-specific gene cluster involved in secondary metabolism [26,27].

Secondary metabolite genes often involve particular types of key enzymes, mostly secondary metabolite genes (Polyketide synthase (PKS), Non-ribosomal peptide synthetase (NPS), and Terpenoid synthase) and cytochrome P450 genes. These key enzymes are clustered along with various combinations of additional enzymes for further metabolite catalyzing and with transporters and transcription factors that are essential for the regulation of most of the clustered genes. Some metabolic clusters contain all three functional gene types, in others it is unclear whether regulatory proteins and transporters are closely located or if they even exist (Table 3-1).

Two types of transcription factors have been reported to be critical for co-regulation of secondary metabolite genes; pathway-specific regulators and global regulators. Pathway-specific regulators positively regulate expression of associated genes in a pathway, each of which has its own promoter binding sites. Pathway-specific regulatory proteins often include the Zn2Cys6 zinc binuclear domain, which are a class of proteins so far only found in fungi. The typical protein in this class is AflR, which is required for biosynthetic gene activation of aflatoxin and sterigmatocystin [28]. AflR in *Aspergillus parasiticus* binds to the palindromic sequence 5'-TCGn{5}CGA-3' in the promoters of 8 out of 11 genes of the aflatoxin cluster and activates their expression [29]. Another type of regulatory protein found in biosynthetic gene clusters is the Cys2His2 zinc-finger. TRI6, a Cys2His2 zinc-finger protein, recognizes and binds to the motif 5'-TnAGGCCT-3' in the promoter regions of the trichothecene cluster in *Fusarium sporotrichioides* [30,31]. Other pathway-specific regulator genes acting on individual pathways have been described in Table 3-2.

Secondary metabolite biosynthesis is also coordinated at an upper hierarchic level by global regulators encoded by genes unlinked to the genes in the pathway. Such genes regulate multiple physiological processes and generally respond to environmental signals such as nitrogen and carbon sources, temperature, light and pH signalling [32,33,34,35]. They can positively or negatively regulate metabolite production. PacC, a pH dependent transcriptional regulator has a positive role in regulation of penicillin gene expression under alkaline conditions but is also associated with negative regulation of the sterigmatocystin and aflatoxin gene clusters in *Aspergillus* species [36,37,38].

3.1.2 Host-fungus interaction

Host-fungus interactions can range from the elimination of the fungus to the death and/or disease of the host by the states of colonization, commensalism, infection, and persistence [39]. In many cases, the capacity to cause various outcomes for particular hosts depends on specific genes encoding host-determining 'virulence factors'. These virulence factors include diverse enzymes involved in toxin synthesis, signal cascade components such as G proteins, and secreted proteins. Some of the virulence factors have been observed to form co-regulated gene clusters in pathogenic fungi. These clusters mostly contain genes synthesizing secondary metabolites and secreted proteins. Secondary metabolite clusters acting as mediators of interactions with particular hosts are already described in 3.1.1.

Secreted proteins are one of the key features of host-fungus interaction. Some act as inducers of diverse diseases. Many known secreted proteins have been individually characterized as host-selective toxins (HSTs) or genes belonged to secondary metabolite clusters. By analysis of the complete genome sequence of *Ustilago maydis*, secreted protein clusters were first revealed. Gene clusters comprised of only secreted proteins are unusual gene arrangements, which have not yet been found in other fungi. About 20% of its 426 secreted proteins are organized in 12 clusters comprising 3-26 genes, and are co-induced during infection [6]. Gene disruption experiments revealed that five of the gene clusters encoding secreted proteins play a role in disease on the maize host. Deletion of four of the gene clusters reduced virulence. These studies demonstrate the importance of secreted protein clusters in the interaction with the host.

3.1.3 Gene clusters in yeast

The budding yeast *Saccharomyces cerevisiae* genome contains several gene clusters that are essential for growth under certain conditions. These include the DAL gene cluster for the use of allantoin as a nitrogen source, the GAL gene cluster for utilization of carbon source, and gene clusters for biotin synthesis [40,41,42,43,44,45]. These gene clusters are well-studied and have provided information about mechanisms of genome rearrangement and adaptation of gene clusters during evolution. First, the DAL cluster consists of six adjacent genes encoding proteins that enable yeast to use allantoin as a nitrogen source. The DAL cluster is completely conserved in four closely related yeast species (the *Saccharomyces sensu stricto* group) [40]. In any of the more distantly related hemiascomycetes, homologs of the six DAL genes are present but they are found individually at dispersed chromosomal locations [40]. Phylogenetic analysis of the DAL genes has showed that the DAL cluster in *S. cerevisiae*

was formed relatively recently by a series of near-simultaneous relocations of genes that were previously scattered around the genome [40]. This result provides evidence for rapid genomic rearrangement and recruitment of dispersed genes into clusters during short evolutionary time periods. Second, the GAL genes that enable cells to use galactose as a carbon source are physically clustered in the yeast genomes [42]. Galactose utilization is widespread among the yeasts but several yeast species have lost the ability to use galactose. Investigation of the inability to use galactose has found that three out of the four non-utilizing species (*Saccharomyces kudriavzevii*, *Candida glabrata*, *Kluyveromyces waltii*, and *Eremothecium gossypii*) have lost most or all of the genes of the GAL pathway [43]. One out of the four non-utilizing species, *Saccharomyces kudriavzevii* is a close relative of *S. cerevisiae* and retains all seven GAL loci as syntenic pseudogenes, providing very recent gene losses and hence a recent change in the metabolic capacity of this species [43]. Besides, *S. kudriavzevii* showed additional divergent physiological properties that are associated with a shift in ecological niche [43]. These results suggest that adaptation to new ecological niches may be associated with gene inactivation. Lastly, biotin is an essential vitamin required for many carboxylation reactions and needs three genes (BIO3, BIO4, and BIO5) for synthesis [44]. Many wild isolsates of *S. cerevisiae* cannot synthesize biotin but retain the genes from this biosynthetic pathway. The pathway for biotin synthesis was lost in a yeast ancestor and then regained in the *S. cerevisiae* lineage by mechanisms of gene loss and gene gain by both a horizontal gene transfer (HGT) from diverse bacteria and gene duplication with neo-functionalization [46]. These three cases suggest that selective advantage driven by the need to adapt for growth under different environments promotes the formation of physical gene clusters during evolution even though it is not clear what ecological change permitted gene loss or duplication in each case.

3.1.4 Difficulty to identify fungal gene clusters

Particular genes encoding virulence proteins or secondary metabolites are often found in fungal gene clusters which are difficult to identify under laboratory conditions. The virulence genes are frequently absent or transform function into avirulence even in closely related species and isolates from different hosts. For example, PWL2 identified in a *Magnaporthe grisea* isolate from rice plays a major role in blocking pathogenicity toward weeping lovegrass without altering pathogenicity in rice and barley [47]. The avirulence gene, AVR1-CO39, in *Magnaporthe oryzae* is absent in most isolates from rice [48], whereas it is present in isolates from perennial ryegrass [49]. Moreover, secondary metabolite genes are often expressed under very specific environmental or developmental conditions. For instance,

homologs of the aflatoxin cluster of strains of *Aspergillus oryzae* and *A. sojae* were not expressed even under the conditions that favor aflatoxin synthesis [7,8].

The spectrum of present and the capacity to synthesize secondary metabolites differ even in closely related species or among strains of one species. For instance, closely related species of the *Gibberella fujikuroi* species complex showed that only some are able to produce gibberellins (GAs), while the others have lost this ability due to mutations in the GA gene cluster including losses of one or more genes [50,51,52,53]. These specificities of fungal genes and gene clusters may result from the limited number of sequenced strains. It may be overcome by sequencing of diverse strains as is currently on the way for *Fusarium verticillioides*.

3.2 Paradigms of study of gene clusters

3.2.1 Functional genomics of plant pathogenic fungi

Recent advances in sequencing and genomic technologies have led to a remarkable increase in the number of sequenced fungal genomes. The next major challenge for plant pathogenic fungal research is to translate genome sequence information into biological function. A well-known effective way to study the disease-causing mechanisms of plant pathogenic fungi is to disrupt their genes and isolate mutants exhibiting altered virulence. For gene manipulation at the genomic level, insertional mutagenesis techniques have been mostly applied in filamentous fungi including random restriction enzyme-mediated integration (REMI), *Agrobacterium tumefaciens*-mediated transformation (ATMT), and transposon tagging [54,55]. REMI is based on the integration of a defined DNA carrying a selectable marker by the action of a specific restriction enzyme [56]. ATMT is used to produce large-scale T-DNA insertions [57]. This technique has the possibility to transform intact fungal cells and generates a high percentage of transformants with single-copy integrated DNA [58,59]. Transposable elements (TEs) are ubiquitous DNA segments in prokaryotic and eukaryotic organisms which have the ability to move and replicate within genomes [60] and have been used for insertional mutagenesis in diverse fungi [55,61,62,63]. Insertion mutants by transposon tagging can be obtained rapidly without the need for repeated transformation [55]. The main advantage of insertional mutagenesis is that the mutated gene is tagged by transforming DNA and can subsequently be cloned. However, integration of the transforming DNA into the target genomes have detected a bias toward highly transcribed genomic regions [56,57,59]. Moreover, the efficiency to identify genes associated in virulence by insertional mutagenesis is very

low. Recent studies based on transformation techniques conducted in several *Fusarium* species showed that a small portion of insertions are linked to mutant phenotypes for pathogenicity [64,65,66,67]. For instance, only eleven pathogenicity mutants (0.17%) were identified by screening 6,500 transformants by the REMI approach in *F. graminearum*. Additionally, assays which determine the functional roles of the found target genes are still time-consuming and labor-intensive. As a crucial part of functional genomics, development of efficient gene knockout and modification tools are also required to promote systematic characterization of individual genes. Despite these limitations, these techniques have been used successfully to detect novel pathogenicity factors in fungi and can provide evidences to assign putative functions to fungal gene clusters.

3.2.2 Global regulator-based identification

Genes synthesizing secondary metabolites can be identified by global transcription regulators. LaeA, a nuclear protein with homology to arginine and histone methyltransferases in *Aspergillus* species showed a possible mechanism of global regulation of secondary metabolite gene clusters and provided a mean for the discovery of new metabolites. LaeA was first identified in a screen for complementation of a sterigmatocystin mycotoxin developmental mutant [68]. Additional deletion studies lead to the identification of LaeA's role as a global regulator of secondary metabolism in *Asparagillus* genus. Loss of LaeA silenced production of sterigmatocystin (carcinogen), penicillin (antibiotic) and lovastatin (antihypercholesterolemic agent) in *A. nidulans* and production of gliotoxin (toxin with immunosuppressive activity) in *A. fumigatus* [68,69]. Furthermore, a genomics-based approach using LaeA was exploited to identify new metabolite genes. Transcriptional profiling of a LaeA-deletion and a LaeA-overexpression strain by microarrays revealed that LaeA transcriptionally regulated numerous secondary metabolite clusters [70]. Deletion of a gene in one of the clusters results in loss of production of the antitumor compound terrequinone A, which had not previously been reported in *A. nidulans*. Recently, a gene with strong similarity to the LaeA regulator in *Penicillium chrysogenum* (PcLaeA) was found and shown to control not only biosynthesis of secondary metabolism but also pigmentation and sporulation in *P. chrysogenum* [71]. As exemplified by the use of the LaeA regulator as a metabolite-mining tool, such investigations by other known global regulatory proteins may also elicit hints to the signals leading to the production of other secondary metabolites.

3.2.3 Chromatin level identification

Chromatin structures are important for regulation of secondary metabolite (SM) gene clusters. Chromatin regulation of gene expression can be directed by modifications of histones such as methylation and acetylation. Histone modification patterns control the interaction of histones with transcription activators or repressors [72]. The characterization of a global transcriptional factor, LaeA described in 3.2.2, demonstrated several cases for chromatin-based regulation of SM gene clusters. LaeA appears to occur via methylation of proteins that regulate chromatin structure [68,73]. The LaeA protein initiates conversion of heterochromatin to euchromatin by interfering with methylases or deacetylases, and induces expression of cluster genes which are silenced in heterochromatin [73]. In *Aspergillus* species, some subtelomeric SM gene clusters were located in heterochromatic regions in which suppression of expression was relieved by deletion of a key histone deacetylase [74]. Links between chromatin modification and regulation of SM gene clusters suggests that silencing of SM gene clusters can be reversed by repressing genes important for the establishment of repressive chromatin configurations [75]. SM gene clusters are often silent under laboratory conditions so investigation of silent gene clusters is a substantial challenge in fungi. A recent study revealed that the loss-of-function CclA, ortholog of *S. cerevisiae* Bre2 involved in histone H3 lysine 4 methylation, activated the expression of cryptic SM gene clusters in *Aspergillus nidulans* [75]. As shown in this approach, genetically manipulated chromatin can help one to discover silenced fungal secondary metabolite genes at the genomic level [75].

3.2.4 Genes with common attributes

Examining physical locations of genes that are linked by common attributes is straightforward to identify fungal gene clusters. This approach has been already conducted in diverse sequenced eukaryote genomes by exploiting co-expressions, co-functions, or common metabolic pathways. Microarray data has been mostly used for genomic level studies of gene order in relation to gene expression or protein functions. Analysis of local co-expression on chromosomes is typically done by measuring correlations among expression profiles of genes positioned close to each other. Genomic clustering of co-expressed fungal genes was first identified in the *S. cerevisiae* genome [76]. About 25 percent of yeast genes with cell cycle-dependant expression patterns were directly adjacent to genes induced during the same phase of the cell cycle [76]. Adjacent pair of genes in budding yeast exhibit highly correlated expression patterns [77]. In addition, co-expressed neighboring genes have been observed in diverse eukarytes[1]. However, it includes not only functionally related genes such as tissue-

specific genes in humans or the same functional category but also apparently unrelated genes. Besides, co-expressed gene clusters were closer to each other than expected by chance but genes in many clusters were sparse over genomes. These indicate that additional observations are required to apply co-expression studies for prediction of fungal gene clusters that contains distinctive characteristics such as comprising of sets of adjacent genes and specific functional organizations as described in 3.1.

During analysis of expression by microarray data, it is necessary to consider two limitations. First, gene expression inferred from microarray data strongly depend on experimental conditions. In yeast it was shown that adjacent gene pairs were correlated in one experiment, but the same genes were frequently not detected as correlated under different conditions [77]. Second, levels of gene expression can be biased according to different microarray technologies such as cDNA, oligo, and Affymetrix arrays. Several discrepancies have been reported across different platforms. Gene expression patterns in 56 human cancer cell lines from the National Cancer Institute (NCI 60) showed poor correlation between data sets from cDNA and Affymetrix arrays [78]. Comparison of data sets for human neuroblastoma cells also revealed distinct gene expression patterns in two array systems; Affymetrix and long cDNA arrays [79].

Genome-wide analysis of metabolic pathway gene clustering showed significant tendency of metabolic pathway genes to be locally clustered in eukaryote genomes. In one study, metabolic genes assigned to the same pathway as defined in the Kyoto Encyclopedia for Genes and Genomes (KEGG), with missing enzymes filled in by homology [80]. The average distances of gene pairs within the same pathway were compared to the distances between genes relocated at randomly generated gene positions. In five sequenced eukaryote genomes (*S. cerevisiae*, *Homo sapiens*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, and *Drosophila melanogaster*) investigated, all genomes contain KEGG pathways with significant gene clustering. However, the fractions of significant KEGG pathway gene clusters in these genomes are highly variable, ranging between 30 percent and 98 percent. This suggests that genes of these KEGG pathways evolved independent among species.

The KEGG metabolic pathways provide evidences that gene clusters are widespread in eukaryotes but it has limited utility for the identification of metabolite pathway gene clusters in fungal genomes. Except the pathways of the primary metabolism most often derived from yeast species, most fungal genes are not assigned in known KEGG pathways. For instance, in *F. graminearum* over half of the predicted genes have only similarity to unknown proteins or no similarity to genes of any up to now

sequenced organism [81]. Besides, fungal secondary metabolite gene clusters are often species-specific and of unknown function, which lack of course any KEGG pathway information.

3.2.5 Comparative genome analysis

To predict gene clusters, comparative genome analysis is a limited but valuable approach as highlighted by epipolythiodioxopiperazine (ETP) which is a class of secondary metabolite toxins produced by various ascomycete fungi [82]. A member of the ETP gliotoxin cluster was identified in the animal pathogen *A. fumigatus* by homolog search using genes from the ETP sirodesmin cluster of the plant pathogen *Leptosphaeria maculans* [83,84]. This case is an example which shows that known gene cluster allows finding gene clusters in other fungal genomes. However, gene clusters without sufficient sequence similarity to known gene clusters like the secreted protein clusters of *U. maydis* [6] will remain concealed from this method.

Comparative genome analysis to study the differences between pathogenic and non-pathogenic fungi can allow detection of a broad range of pathogenicity-related genes which often occur in fungal gene clusters. Comparison of thirty-six fungal genomes including seven plant pathogenic species using Pfam motifs uncovered several protein families (e.g. plant cell wall degradation, toxin biosynthesis) specific to or relatively more common in phytopathogenic genomes [85]. However, this analysis also suggested that the presence of novel, universal pathogenicity factors is improbable in phytopathogenic fungi because there was no orthologous protein set that was completely specific to phytopathogenic species [85]. A fungal phylogeny based on 42 complete genomes also showed that phytopathogenic species are found in all taxonomic divisons of fungi and are often closely related to non-pathogenic fungi [86]. These results suggest that more comprehensive studies are necessary to interpret the large flow of genome sequence data and to explore fungal gene clusters.

Pathway	N. of gene	Functions in the pathway				SM	Species
		TF	TP	Enzyme (CYP)	Others / unknown		
Aflatoxin	25	2	1?	21 (4)	0 / 1	PKS	*Aspergillus parasiticus* [87]
Aurofusarin	11	2	1	6 (0)	1 / 1	PKS	*Fusarium graminearum* [88]
Asperfuranone	7	1	1	5 (0)	0 / 0	PKS	*A. nidulans* [89]
Butenolide	8	1	1	5 (1)	0 / 1	-	*F. graminearum* [90]
Citrinin	6	1	1	4 (0)	0 / 0	PKS	*Monascus purpureus* [91], *A. oryzae* [92]
Conidial pigment	6	0	0	5 (0)	0 / 1	PKS	*A. fumigatus* [93]
Equisetin	11					NPS-PKS	*F. heterosporum* [94]
Ergot Alkaloids	17	0	0	12 (1)	0 / 5	NPS	*Claviceps purpurea* [95]
Fumonisin	27	3	2	13	9 / 0	PKS	*F. verticillioides* [96]
Gibberellin	7	0	0	7 (4)	0 / 0	TPS	*F. fujikuroi* [97]
Gliotoxin	12	1	1	9 (2)	0 / 1	NPS	*A. fumigatus* [98, 99]
Lovastatin	18	2	3	5 (2)	2 / 6	PKS	*A. terreus* [100]
Penicillin	3	0	0	3	0 / 0	NPS	*Penicillium chrysogenum* [11], *P. nalgiovense* [12]
Sterigmatocystin	25	1	0	21 (4)	2 / 1	PKS	*A. nidulans* [28]
Terrequinone	> 5	0	0	4 (0)	0 / 1	NPS	*A. nidulans* [101]
Trichothecene	12	2	1	7 (3)	0 / 2	TPS	*F. sporotrichioides* [102]
Zearalenone	>4	1	0	3	0 / 0	PKS	*F. graminearum* [103]

Table 3-1: Functional organization in diverse metabolite pathways

Clustered genes involved in common pathways show special functional organizations. Pathway gene clusters include genes encoding pathway specific transcription factors, transporters, and enzymes. Some pathway gene clusters contain all three types of genes; in others it is unclear whether transcription factors and transporters are closely positioned.

TF: transcription factor, TP: transporter, CYP: cytochrome P450, PKS: Polyketide synthase, NPS: Non-ribosomal peptide synthetase, TPS: Terpenoid synthases, NPS-PKS: NPS-PKS hybrid gene

Secondary metabolic pathways	Transcription factors	Protein domain	Species
Aflatoxin	AflR	Zn_2Cys_6	*Aspergillus flavus*[104], *A. parasiticus* [29,104,105]
Aurofusarin	GIP2	Zn_2Cys_6	*Fusarium graminearum* [106]
Fumonisin	ZFR1	Zn_2Cys_6	*F. verticillioides* [107]
Gliotoxin	GliZ	Zn_2Cys_6	*A. fumigatus* [108]
Sterigmatocystin	AflR	Zn_2Cys_6	*A. nidulans* [109]
Trichothecene	Tri6	Zn_2His_2	*F. sporotrichioides* [110]

Table 3-2: Pathway specific transcription factors in fungi

Pathway specific-regulatory proteins containing the Zn2Cys6 zinc binuclear or the Cys2His2 zinc-finger domain have been reported in some fungal metabolic pathway gene clusters. The pathway specific-regulatory proteins are known to recognize cis-regulatory elements in the promoters of most of the corresponding pathway genes.

3.3 The necrotrophic plant pathogen *Fusarium graminearum*

The genome sequence of the plant pathogenic fungus *Fusarium graminearum* (anamorph: *Gibberella zeae*) was publicly released by the Broad Institute in 2003. The genome of 36.1 Mb contains 13,937 genes and is organized in four chromosomes (The MIPS *F. graminearum* Genome Database (FGDB): http://mips.helmholtz-muenchen.de/genre/proj/FGDB/) [81]. Of these genes, 2,004 don't have any similarity to genes of any up to now sequenced organism and 5,812 have similarity to unknown proteins (classification of proteins in FGDB) [81]. Very few repetitive sequences and a low number of paralogous genes were detected in the genome [111].

As a major pathogen of cultivated cereals, *F. graminearum* is a valuable target organism for the identification of gene clusters at the genomic level and serves as a test case for this work. Five mycotoxin biosynthetic genes or gene clusters have been already identified to date; aurofusarin, butenolide, fusarin C, trichothecene, and zearalenone (Table 3-3). In addition, 51 secondary metabolite genes are predicted, including 15 polyketide synthases (PKS), 20 nonribosomal peptide synthetases (NPS), and 16 terpenoid synthases.

Some of the secondary metabolite genes have been functionally analyzed. Specific disruption of each PKS gene independently was accomplished and the mutants were characterized [112]. Of the 15 PKS genes disrupted, five of these genes are involved in the synthesis of aurofusarin, fusarin C, zearalenone, and the black perithecial pigment. The functions and products of the other 10 PKS genes still remain obscure. Related functional groups of PKS genes were designated by expression analysis under a range of developmental, nutritional and pathogenic conditions (Table 3-4) [112]. Out of the 20 predicted NPS genes, 3 NPS genes have been functionally determined (Table 3-5). NPS6 is responsible for the biosynthesis of extracellular siderophores as a virulence determinant [113]. NPS1 and NPS2 were found to be related to genes involved in siderophore synthesis and not affected in sexual development [114,115]. One terpenoid synthase (TRI5) is found in the co-regulated trichothecene cluster [102]. About 65% of the predicted secondary metabolite genes are functionally not described with a potential to elucidate new surrounding functional gene clusters.

Mycotoxin	(Number of gene), gene range	Chr.	SM gene	Proved regulatory element	Host	Phenotypes
Aurofusarin	(11), FGSG_02320 ~ FGSG_02330 [116,117,118]	1	PKS12		Barley, wheat, poultry etc	Yellow/red pigmentation according to pH
Butenolide	(8), FGSG_08077 ~ FGSG_08084 [90]	2	-		Grazing cattle	Fescue foot
Fusarin C	FGSG_07798 [112]	4	PKS10		Con, soybean, etc by *F. moniliforme* [119, 120]	Immunosuppressant effect
Trichothecene	(12), FGSG_03543 ~ FGSG_03532 [102]	2	TRI5	5'-TnAGGCCT-3' in *F. sporotrichioides*[31]	Barley, wheat, etc.	Red rash, blotch etc.
Zearalenone	(5), FGSG_02395, ~ FGSG_02398 [103,121]	1	PKS4, PKS13		Rice, wheat, etc [103]	Hormonal disturbance

Table 3-3: Mycotoxin biosynthetic genes or gene clusters produced by *F. graminearum*

Five mycotoxin biosynthetic genes or gene clusters identified in *F. graminearum*. The genes are expressed in different hosts and influence diverse phenotypes. Except butenolide mycotoxin, secondary metabolite genes are involved in synthesizing mycotoxins. Chr.: chromosome, SM: Secondary metabolite, PKS: polyketide synthases, NPS: Non-ribosomal peptide synthetase

Expression specificity	Gene name	FGDB ID
Plant specific	PKS15	FGSG_04588
Gran specific	PKS14	FGSG_03964
	PKS13	FGSG_02395
	PKS4	FGSG_12126
Sexual development related, Expressed during grain colonization	PKS3	FGSG_09182
	PKS7	FGSG_08795
	PKS9	FGSG_10464
	PKS11	FGSG_01790
Mycelial growth related	PKS10	FGSG_07798
	PKS1	FGSG_10548
	PKS2	FGSG_04694
	PKS8	FGSG_03340
Not expressed during infection	PKS12	FGSG_02324
Constitutively expressed	PKS6	FGSG_08208

Table 3-4: Expression specificity of PKS genes in *F. graminearum*.

PKS genes in *F. graminearum* were designated to six functional groups of genes with similar expression patterns under grain colonization, plant colonization, sexual development, and mycelial growth [112].

Related function	Gene name	FGDB ID
Siderophore synthesis [115], Not affect sexual development [114, 115]	NPS1	FGSG_11026
	NPS2	FGSG_05372
Extracellular siderophores [113]	NPS6	FGSG_03747

Table 3-5: Functions of NPS genes in *F. graminearum*.

3 out of 19 NPS genes in *F. graminearum* are currently functionally classified.

3.4 The biotrophic plant pathogen *Ustilago maydis*

Ustilago maydis is a biotrophic basidiomycete fungus that causes smut disease in *Zea mays* (corn). The fungus induces tumours filled with mass of dark diploid teliospores and can result in stunted plant growth, leading to necrosis, hyperplasia and hypertrophy of infected organs [122]. *U. maydis* has been established as a model organism for the study of plant-fungus interactions and fungal virulence because it is amenable to genetic analysis and molecular manipulation [123, 124]. The genome is about 20.5 Mb and contains 6,782 genes distributed on 23 chromosomes (The MIPS *Ustilago maydis* Database (MUMDB): http://mips.helmholtz-muenchen.de/genre/proj/ustilago/).

Studies of gene clusters in the *U. maydis* genome can help to unravel biotrophy clusters because it has unexpected features which differ from those in a necrotrophic pathogen. First, the biotrophic pathogen, *U. maydis* possesses a strongly reduced set of genes known to be involved in pathogenesis found in necrotrophic fungi. For instance, the fungus has only 33 hydrolytic enzymes (polysaccharide hydrolases, polysaccharide lyases, and pectin esterases) that degrade living and dead plant cell walls, in contrast to 103 for *F. graminearum* [6]. In addition, the genome contains 12 genes encoding 3 types of secondary metabolites, whereas *F. graminearum* has 51 genes. Second, analysis of the genome sequence demonstrated that about 20% of its 426 predicted secreted proteins are positioned in 12 regions containing 3 to 26 genes of unknown function [6]. The expression of most clustered genes was significantly up-regulated during the tumour stage. The clustered genes encoding secreted proteins are a novel type of gene cluster which can be involved in important roles during pathogenesis. Unlike gene clusters responsible for secondary metabolite biosynthesis, which often contain regulatory proteins and diverse enzymes, no such functional genes have been observed within the clustered secreted protein genes. It is also unknown how expressions of these genes are coordinated.

4 Methods

4.1 Identification of tentative functional gene clusters (TFCs)

4.1.1 Combinatorial functions *in F. graminearum*

In *F. graminearum*, 51 secondary metabolite genes and 117 cytochrome P450 genes were defined as base genes to find gene clusters with genes that can be grouped according to common functions. Centering on base genes, functional information of neighboring genes in both upstream and downstream directions were observed whether they have secreted proteins, transcription factors and/or transporters, secreted proteins using Interpro [125], TargetP [126], and annotation information in FGDB [81] (Table 4-1, s1-s5). Statistical significance was calculated by comparing to the occurrence of the same set of functions in gene neighborhoods found in random genomes.

To create a random genome, the relative positions of genes were randomized by swapping each gene with a random partner in the genome. Swapping was done twice for each gene. Neighboring genes are defined as those genes with five types of functional descriptors (s1-s5 in Table 4-1) within a maximum gap of 3 genes without such functional descriptors. The occurrence of neighboring genes with particular functional classifications was counted. A genome was considered enriched with clusters of neighboring genes performing a set of functions, if less than 10/1000 random genomes had more of such clusters compared with the real genome (P-value < 0.01). Overlapping clusters from different base genes were combined and defined as one tentative functional gene cluster (TFC) with specific composition of gene functions. TFCs were each denoted as the tfc$species_$number. Abbreviations are used for name of species.

4.1.2 Functional enrichments

For genes encoding for secondary metabolites, cytochrome P450 and secreted proteins (Table 4-1, s1, s2 and s5), statistical significance of functional enrichments were calculated by the same procedure as describes in 3.1.1. The probability of obtaining by chance the previously identified clusters is additionally calculated using a hypergeometric distribution test. P-values < 0.01 were defined as a threshold for significant functional enrichment in the clusters. Gene clusters enriched for secreted protein genes were named as sp$species_$number.

ID	Source	N. of gene	
		F. graminearum	*U. maydis*
s1	Polyketide synthase (manual annotation)	15	6
	Non-ribosomal peptide synthetase (manual annotation)	20	1
	Terpenoid synthase (IPR008949)	16	5
s2	Cytochrome P450 (IPR001128)	117	22
s3	Fungal specific transcription factor (IPR007219)	166	36
	Transcription factors and binding proteins (IPR001138, IPR004827, IPR007087, IPR015880, FGDB description ('*transcription factor*', '*regulatory protein*'))	367	139
s4	Transporters (IPR002293, IPR003439, IPR005828, IPR005829, IPR010573, IPR011527, IPR011701)	558	210
s5	Secreted proteins (TargetP[126]: reliability class 1,2)	1413	656

Table 4-1: Features of known gene clusters used to identify new gene clusters

Five types of functional descriptors were derived from known gene clusters based on information from InterPro [125], TargetP [126], and manual annotation.

4.2 Co-expressed neighboring genes in *F. graminearum*

4.2.1 Data pre-processing

The gene chip expression data of two independent experiments were analyzed: The first experiment (FG1) monitored the expression of the fungal genes during growth on barley spikes [127]. The second experiment (FG5) examined gene expression during sexual development [128].

Log_2-transformed expression values were obtained using the Robust Multichip Analysis (RMA) algorithm[129] from the Bioconductor package in R. All expression values from the replication arrays were averaged for each time point. Affymetrix detection (Present/Marginal/Absent) calls were used to exclude genes with unreliable data. We picked probe sets that have 'Present' calls for at least 2 replicas at any time point from the test arrays for analysis. Probe sets with all expression values < |0.5| were regarded as belonging to non-expressed genes and filtered out.

4.2.2 Expression coherence of neighboring genes

The mean Pearson correlation coefficient (r) was used as a measure of similarity of expression profiles from each gene group. We calculated mean R for the expression profiles of all possible pairs of 3 to 25 neighboring genes in the genome of *F. graminearum*. To determine whether neighboring genes were significantly co-expressed, we assessed statistical significance by comparing with the 95th percentile of the distribution of mean Rs from 10,000 randomly generated probe sets containing an equal number of genes. For example, for 10 neighboring genes within a group G, correlation coefficients of a total of 45 pairs were calculated and the mean R was used as a value of co-expression of the group G. The group G was selected as a cluster of co-expressed neighboring genes when the mean R of group G is higher than the significant threshold (the 95th percentile of the mean r distribution for a random set containing 10 genes). Clusters of co-expressed neighboring genes were combined permitting one gene gap between genes and maximum of 20 percent gene gaps in one cluster if the mean R of the combined clusters was also significant. Co-expressed gene clusters were each denoted as the fg1_$number cluster and the fg5_$number cluster for FG1 and FG5 experiments, respectively.

4.2.3 Clustering of expression profiles

To observe patterns of expression profiles from clusters of co-expressed neighboring genes, representative cluster profiles were generated by averaging expression values for all genes in a cluster for each time point. The representative cluster profiles were clustered using hierarchical clustering with centered Pearson correlation and complete linkage.

4.2.4 Regulatory motifs in co-expressed gene clusters

Overrepresented regulatory motifs for each of the co-expressed neighboring genes were examined using the same procedures as described in 4.3.1.

4.3 Regulatory motifs enriched in promoters of neighboring genes

4.3.1 Screening neighboring gene promoters for conserved regulatory motifs

Contig sequences and gene details were obtained from the following databases: FGDB (http://mips.helmholtz-muenchen.de/genre/proj/FGDB) [81] for *Fusarium graminearum*, MUMDB (http://mips.helmholtz-muenchen.de/genre/proj/ustilago/) for *Ustilago maydis*, and PEDANT 3 (http://pedant.helmholtz-muenchen.de) [130] for *F. verticillioides*, *F. oxysporum*, *Ustilago hordei*, and *Sporisorium reilianum*. The *Fusarium* genome compilations are based on the genome assemblies of the Broad Institute (http://www.broadinstitute.org). The *Ustilago* species and *S. reilianum* genomes were assembled at MIPS in collaboration with the MPI Marburg. The genome data of *U. hordei* and *S. reilianum* are not published at the time of submission of this thesis and any data is considered as private.

Up to 1kb of upstream sequence from the start codon of each gene was extracted from the intergenic regions. Motif seeds (MSs) were explored in the 5' UTR sequences of a maximum of 25 neighboring genes by a sliding window. The MSs are hexa-mer sequences having the structure ABC-gap-DEF, where A, B, C, D, E, F can be any nucleotide and gap is 0 to 9 non-specified bases. The neighboring genes were defined as genes that were immediately adjacent in the genome regardless of sequence strand.

$$P-value = \sum_{m}^{n} \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}} \qquad (1)$$

The probability that a certain MS occurs m or more times in neighboring genes was calculated using the hypergeometric distribution. In the formula (1), m is the number of genes that contain a MS in a group of n selected neighboring genes, relative to M genes that contain the MS in all N genes of a genome. The p-values were adjusted applying Bonferroni corrections [131] by multiplying each p-value by the total number of possible MSs. Bonferroni corrected p-values < 0.01 were considered significant.

To specify clusters more exactly significant MSs were re-scanned in all intergenic regions of gene clusters extended by 5 genes in both directions, allowing gaps of 5 genes. Overlapping gene clusters were combined into one gene cluster if they have co-occurring MSs. These clusters were ranked according to the adjusted p-values of MSs as described above and named ms$species_$rank of MS. Abbreviations are used for name of species; fg for *F. graminearum* and um for *U. maydis*. For instance, the name of the cluster with the most significantly ranked MS in *F. graminearum* is msfg_1.

SeqVista (A graphical tool for sequence feature visualization and comparison) was used to present a holistic, graphical view of features identified gene clusters [132].

4.4 Synteny analysis

Orthologs were first searched for as reciprocal best hits (RBH) within an all-vs-all comparison of proteins in each pair of genomes using the Similarity Matrix of Proteins (SIMAP) [133]. *F. graminearum* was used as reference genome and compared with 2 other *Fusarium* species (*F. verticillioides* and *F. oxysporum*). From genomic maps of orthologous gene, synteny and non-syntenic blocks were defined as follows; syntenic blocks (SBs) for each pair of genomes are genomic regions containing a minimum of 5 contiguous orthologous genes with conserved gene order and non-syntenic blocks (NSBs) are genomic regions consisting of more than 5 genes with no orthologs in target genomes.

5 Results: Gene clusters in *Fusarium graminearum*

Our goal is to identify known and new, putative gene clusters throughout the whole genome. The properties of known fungal gene clusters were investigated and the insight gained was used to design an approach for comprehensive discovery of fungal gene clusters. Three types of properties were examined to identify gene clusters: the frequent functional organization of genes (combinatorial functions and enrichment of specific functions), co-expression, and the presence of promoter motifs. The gene clusters were identified using each of these properties independently and compared with known gene or gene clusters from diverse references.

A fungal gene cluster can be broadly defined as a set of neighboring genes that have at least one property described above observed from known gene clusters. Gene clusters identified by the three individual properties are as follows (Figure 5-1): 223 co-expressed gene clusters (76 co-expressed gene clusters *in planta*, 147 co-expressed gene clusters during sexual development), 81 gene clusters with specific compositions of gene functions and 95 gene clusters with co-occurring motif seeds. Comparing different types of gene clusters, 12 clusters with co-occurring regulatory motifs and 17 co-expressed gene clusters were found which overlap clusters with TFCs. 4 clusters with co-occurring regulatory motifs overlap with cluster of co-expressed neighboring genes.

As a proof of concept, our analyses re-discovered five known mycotoxin gene clusters with independent properties as described in Table 5-1. The trichothecene cluster was identified in all 3 different analyses. The four other remaining mycotoxin clusters are matched by one or two of the analyses. The identified mycotoxin clusters were re-examined for their additional features to re-define the screen for new gene clusters. Among 51 secondary metabolite (SM) genes predicted in *F. graminearum*, 49 SM genes are identified in at least one type of gene cluster at the genomic level and arranged with diverse evidences (Table 5-23).

A total of 24 gene clusters with multiple properties at the genomic level were primarily screened to select novel gene clusters (TFC ID with a red color in Table 5-7). According to different evidences, 6 gene clusters are suggested as novel gene clusters, which contain multiple features found in 3 mycotoxin clusters (trichothecene, aurofusarin, and butenolide clusters). Each of the novel gene clusters and their evidences are described in Table 5-16.

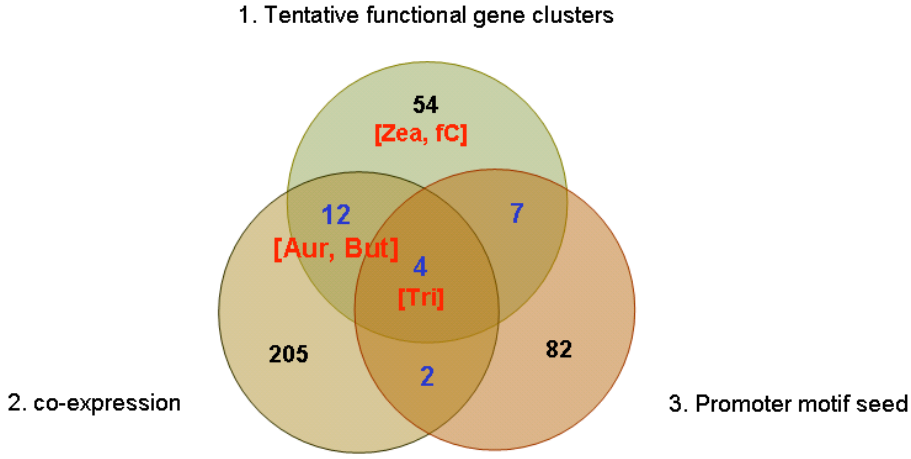


Figure 5-1: Number of gene clusters identified by three independent analyses

1. Tentative functional gene clusters (TFCs): 77 TFCs were identified based on five types of functional descriptors, genes encoding secondary metabolites, cytochrome P450, transcription factors, transporters, and secreted proteins (Table 4-1).

2. Co-expression: 223 clusters of co-expressed neighboring genes. 76 and 147 co-expressed gene clusters were identified from two different microarray analyses during growth in planta and sexual development, respectively.

3. Promoter motif seed: 95 gene clusters with co-occurring promoter motif seeds.

Gene clusters with multiple properties: 25 gene clusters clusters with multiple properties (blue number) were primarily screened and re-examined comparing with characteristics of 3 mycotoxin gene clusters identified by our analyses. Mycotoxins identified by our analyses are represented in red brackets. Tri: Trichothecene, Aur: Aurofusarin, But: Butenoide, Zea: Zearalenone, fC: fusarin C

5.1 Mycotoxin gene clusters

Five known mycotoxin gene clusters were re-discovered by our analyses (Table 5-1): fusarin C, trichothecene, aurofusarin, butenolide, and zearalenone clusters. Independent evidences derived from different analyses for the mycotoxin clusters were integrated and compared with validated data from diverse references. Only the trichothecene mycotoxin has known to have a regulatory element as proved in *F. sporotrichioides*, whereas conserved regulatory elements for the other three mycotoxins have not been reported. We identified regulatory motifs for the trichothecene cluster by our analysis for *F. graminearum*. Additionally, scanning promoter regions based on co-expressed neighboring genes and gene clusters with particular functional organization, motif seeds are provided as candidate promoter motifs for each of the mycotoxin gene clusters. For each cluster region map, gene symbols filled with blue belong to clusters with co-occurring promoter motifs, filled with yellow belong to the two other types of gene clusters, respectively. Details of each of the identified mycotoxin clusters are explained below.

Analysis / Mycotoxins	1. Specific compositions of gene functions				2. Co-expression		3. Promoter motif (Genomic level)
	Combination	Functional enrichment		SM	FG1	FG5	
		CYP	SP				
Trichothecene	•	•	•	•	•		•
Aurofusarin	•		•	•		•	
Butenolide	•				•		
Zearalenone	•			•			
Fusarin C	•			•			

Table 5-1: Properties of five mycotoxin clusters

Five mycotoxin gene clusters in *F. graminearum* were identified in at least one type of cluster analysis. The trichothecene cluster was discovered in all three analyses.

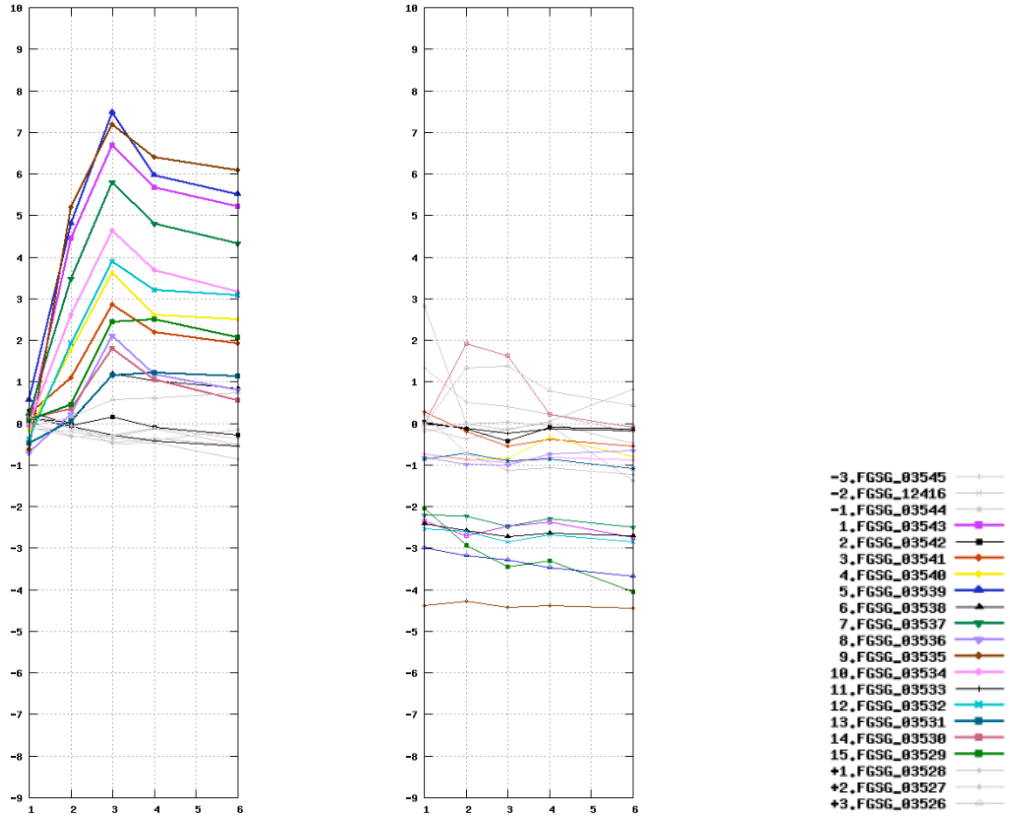
Combination: gene cluster with specific compositions of gene functions based on 5 types of functional descriptiors, CYP: cytochrome P450, SP: secreted protein, SM: secondary metabolite, Co-expression: Co-expressed neighboring genes, FG1: microarray experiment in planta, FG5: microarray experiment during sexual development, Promoter motifs: gene clusters with conserved regulatory elements at the genomic level

5.1.1 The trichothecene cluster

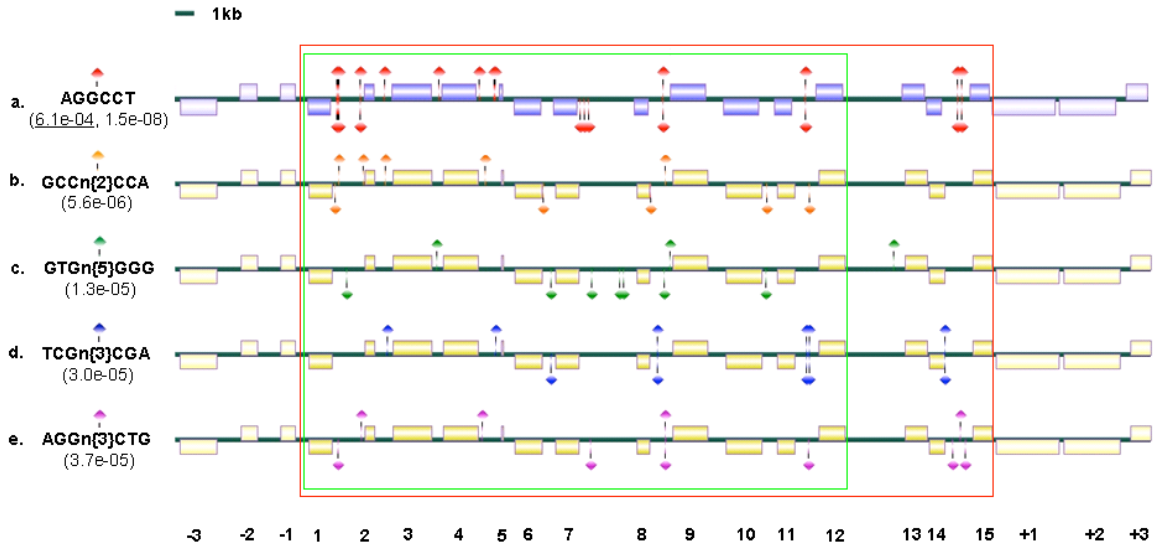
Trichothecenes comprise a large group of sesquiterpenoid toxins produced by diverse fungi including several *Fusarium* species [134]. Various chemotypes of the trichothecenes are described in different environments and species [135,136,137,138]. Most abundant of the naturally occurring trichothecenes are deoxynivalenol (DON) or nivalenol (NIV) in *F. graminearum* and T-2 toxin in *F. sporotrichioides*. Trichothecenes are primarily found as contaminants in human food and animal meals including maize, wheat, barley, and rice. Notably, Fusarium head blight (scab) is an economically devastating disease of wheat and barley, affecting the developing heads of small grains directly. Consumption of the trichothecenes-contaminated crops by humans or animals results in diverse side effects such as vomiting and dermatitis [139]. Trichothecenes can cause a variety of eukaryotic mycotoxicoses through the ability of the toxins to inhibit protein synthesis and induce apoptosis [140,141].

12 genes are currently referred to as the core or main trichothecene gene cluster in *F. graminearum* and *F. sporotrichioides* by a number of observations after gene deletion or disruption [102]. In *F. sporotrichioides*, it has been expected that additional genes may be required for trichothecene biosynthesis because there are several un-assigned steps in the T-2 toxin biosynthetic pathway [142].

By two independent analyses, three additional genes which can be involved in trichothecene biosynthesis were observed downstream of TRI8 (Figure 5-2, Table 5-2). First, gene cluster msfg_14 has co-occurring conserved regulatory motifs showing that three more genes are included in the range of genes with core trichothecene genes. Motif seed 5'-AGGCCT-3' in the cluster msfg_14 is significantly enriched in 15 genes (FGSG_03543 (TRI14) ~ FGSG_03529).This motif seed overlaps with the DNA-binding site 5'-TNAGGCCT-3' previously established for the Cys2His2 zinc-finger regulatory protein TRI6 which acts as a positive regulator of trichothecene biosynthesis in *F. sporotrichioides*. The other striking evidence for the 3 additional genes is the co-expressed gene cluster (cluster fg1_39). The co-expressed genes have an increasing expression pattern until the 3rd day except for one gene, cytochrome P450 (FGSG_03542). These two gene clusters, msfg_14 and fg1_39, were identified in exactly the same range of genes, which strongly support that the three genes can be involved in yet un-assigned steps of the trichothecene biosynthetic pathway or a trichothecene related function.



A. A co-expressed gene cluster in planta (cluster fg1_39, left) and profiles of the corresponding genes during sexual development (experiment FG5, right)



B. Gene clusters with co-occurring regulatory motifs (msfg_14 (a))

Figure 5-2: The trichothecene cluster identified by analysis of promoter motifs and co-expression

A. Co-expressed gene cluster in planta (fg1_39): 15 genes, FGSG_03543 ~ FGSG_03529, on chromosome 2, were co-expressed during growth in planta. The cluster fg1_39 was identified with 3 additional genes (#13~#15) downstream of core trichothecene gene cluster.

B. Gene clusters with co-occurring regulatory motifs (msfg_14 (a)): Gene coding regions are represented by boxes. Gene maps filled with blue and yellow colors refer to gene cluster with co-occurring promoter motifs at the genomic level and based on co-expressed neighboring genes, respectively. Diamond shapes indicate motif seeds. Genes and motif seeds in the forward strand are displayed above the line; genes and motif seeds in the reverse strand are displayed below. The number in parenthesis is p-values of each of motif seeds (The number with underline is an adjusted P-value). Numbers underneath each box indicate gene designations (see Table 5-2). Core trichothecene genes previously reported in *F. graminearum* and *F. sporotrichioides*[102] are highlighted in the green box. The red box encompasses the gene cluster identified by our two analyses, co-expression and conserved motif seeds.

First motif seed 5'-AGGCCT-3' (a) is identified in the analysis on co-occurring regulatory motifs (msfg_14) and overlapped with DNA-binding site 5'-TNAGGCCT-3' previously established for the Cys2His2 zinc-finger regulatory protein TRI6 which acts as a positive regulator of the trichothecene biosynthesis in *F. sporotrichioides*. The other 4 motif seeds (b~e) are suggested by our analysis as putative binding sites for specific TFs in promoters of genes synthesizing trichothecene. Number in parenthesis is a p-value of each of motif seeds.

Gene order	FGDB ID	Gene	Gene feature	Gene description
-3	FGSG_03545			related to OrfH – unknown, trichothecene gene cluster
-2	FGSG_12416			conserved hypothetical protein
-1	FGSG_03544	OrfG[102]		deacetylase
1	FGSG_03543	TRI14		putative trichothecene biosynthesis gene
2	FGSG_03542		CYP	probable cytochrome P450
3	FGSG_03541	TRI12	TP	trichothecene efflux pump
4	FGSG_03540	TRI11	CYP,SP	isotrichodermin C-15 hydroxylase
5	FGSG_03539	TRI9		hypothetical protein
6	FGSG_03538	TRI10	TF	regulatory protein
7	FGSG_03537	TRI5	SM (TPS)	trichodiene synthase [sesquiterpene cyclase]
8	FGSG_03536	TRI6	TF	trichothecene biosynthesis positive transcription factor
9	FGSG_03535	TRI4	CYP	trichodiene oxygenase [cytochrome P450]
10	FGSG_03534	TRI3		trichothecene 15-O-acetyltransferase
11	FGSG_03533			related to TRI7 – trichothecene biosynthesis gene cluster
12	FGSG_03532	TRI8	SP	trichothecene 3-O-esterase
13	FGSG_03531	OrfA[102]		monooxygenase
14	FGSG_03530	OrfB[102]	SP	acetylesterase, trichothecene gene cluster
15	FGSG_03529		SP	related to glucan 1,3-beta-glucosidase
+1	FGSG_03528			conserved hypothetical protein
+2	FGSG_03527			conserved hypothetical protein
+3	FGSG_03526	OrfE[102]	SP	OrfE – unknown, trichothecene gene cluster

Table 5-2: Gene functions and features of the trichothecene cluster

The trichothecene mycotoxin cluster was identified by two independent analyses (based on co-occurring motifs and co-expression), which contains 15 genes, 12 genes (#1~12) currently referred to as the core or main trichothecene gene cluster in *F. graminearum* and additional 3 genes (#13~15, red box) newly identified by two independent analyses. The trichothecene mycotoxin genes shows typical arrangement of fungal gene clusters, which contains genes encoding one secondary metabolite (terpenoid synthase), three cytochrome P450, two transcription factors and one transporter. The trichothecene gene cluster is also enriched in genes encoding cytochrome P450 and secreted proteins.

TPS: Terpenoid synthase, SP: Secreted protein, TF: Transcription factor, TP: Transporter, CYP: Cytochrome P450, SM: Secondary metabolite

In several *Fusarium* species, genes required for trichothecene biosynthesis have been found outside of the core trichothecene cluster [143,144,145,142]. To identify additional genes which may be involved in trichothecene biosynthesis, the FGDB [81] was first screened for genes related to trichothecene regardless of chromosomal order and their expression profiles. Exclusive of genes in the core trichothecene cluster, 26 more genes dispersed on different chromosomes are annotated as gene related to trichothecene/trichodiene. 4 out of 25 genes (one gene FGSG_13797 does not have any matching GeneChip probe) were shown to have similar expression patterns with profiles of core trichothecene genes (Figure 5-3, bold lines). One gene, acetylesterase (FGSG_03530), adjacent to the core trichothecene cluster was already identified by presence of conservation of motif seeds and co-expression in neighboring genes (cluster fg1_39 and msfg_14, Figure 5-2). Another three genes are scattered on 3 different chromosomes, which are cytochrome P450 monooxygenase (FGSG_00071, TRI1, chr.1), oxygenase cytochrome P450 (FGSG_08809, chr.2), and 3-O-acetyltransferase (FGSG_07896, chr.4). In addition these 4 genes have the same conserved motif seed (5'-AGGCCT-3') as the core trichothecene genes.

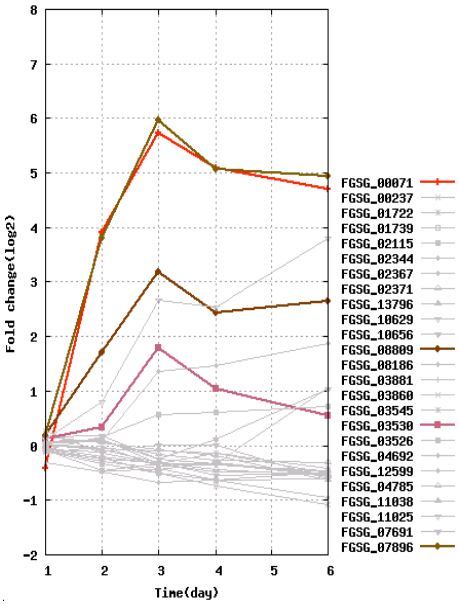


Figure 5-3: Expression profiles of genes related to trichothecene biosynthesis

Expression profiles of 25 genes found outside of the core trichothecene cluster were investigated to identify potential undiscovered trichothecene biosynthesis genes. Four genes, FGSG_00071, FGSG_08809, FGSG_03530, and FGSG_07896 showed similar expression patterns with the core trichothecene genes which have a peak at the time point of the day 3. Interestingly, the four genes have the same conserved motif seed (5'-AGGCCT-3') with ones of core trichothecene genes.

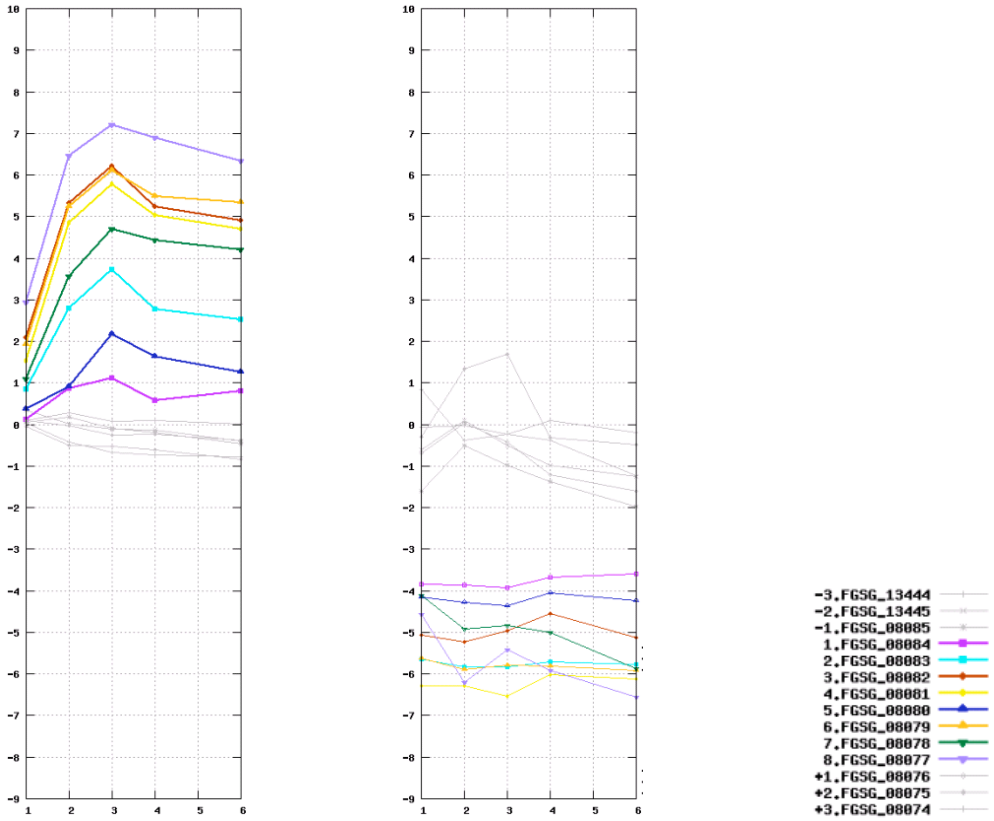
5.1.2 The butenolide cluster

Butenolide was discovered as a water-soluble toxic substance produced by *Fusarium* species isolated from toxic tall fescue [146]. It has been reported to be toxic to mice and rat and can cause fescue foot symptoms in grazing cattle [147,148,149]. Butenolide from plant derived smoke is reported to be a highly effective germination and seedling growth stimulant for a range of arable weeds [150]. It also appears to have no negative impact on seedling morphology and have wide applicability as a germination and growth stimulant [150].

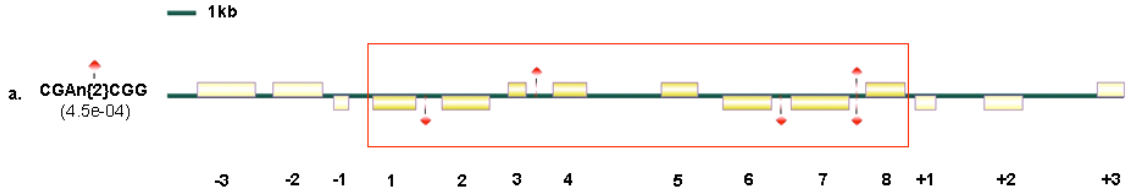
In *F. graminearum*, eight genes involved in butenolide synthesis have been determined by gene disruption and complementation experiments [90]. By co-expression analysis, the same eight genes, FGSG_08084 ~ FGSG_08077, on chromosome 2 were identified to be co-regulated in planta (cluster fg1_32, Figure 5-4(A), and Table 5-3). Expression profiles of the eight genes were clustered in the same COP (COP1.1, Figure 5-8 (A)) as the trichothecene cluster. This result is consistent with previous studies. Northern blot analysis showed that seven out of the eight genes had similar expression and significant amounts of butenolide were produced in trichothecene-producing *F. graminearum* liquid culture [90].

The butenolide was reported not to play a major role in the spread of head blight in wheat [90], contrary to trichothecenes which help spread *F. graminearum* on several crops [151]. Considering very similar expression profiles of both clusters in planta, butenolide may have an important role in other aspects of plant infection or associate with other metabolite genes to cause fescue foot symptoms.

Regulatory elements in the butenolide cluster have not been reported at present. By screening promoter regions of the co-expressed butenolide genes, the motif seed 5'-CGAn{2}CGG-3' is identified and suggested as a putative promoter motif of the butenolide mycotoxin genes (Figure 5-4 (B)).



A. A co-expressed gene cluster in planta (fg1_32, left) and profiles of the corresponding genes during sexual development (right)



B. Gene clusters with co-occurring regulatory motifs

Figure 5-4: The butenolide cluster identified by co-expression analysis and their promoter motif

A. Co-expressed gene cluster in planta (fg1_32): The same eight butenolide related genes previously reported, FGSG_08084 ~ FGSG_08077, on chromosome 2 were identified to be co-expressed *in planta* (cluster fg1_32). Expression profiles of the butenolide synthesis genes showed similar pattern with the trichothecene cluster even though the two mycotoxins have different functions. Butenolide may have an important role in other aspects of plant infection or act with other metabolite genes to cause fescue foot symptoms.

B. Gene clusters with co-occurring regulatory motifs: Gene coding regions are represented by boxes: dark yellow refers to co-expressed genes and light yellow are flanking genes. Diamond shapes indicate motif seeds. Number in parenthesis is a p-value of a motif seeds. Genes and motif seeds in the forward strand are displayed above the line; genes and motif seeds in the reverse strand are displayed below. Numbers underneath each box indicate gene designations (see Table 5-3).

Gene order	FGDB ID	Gene	Gene feature	Gene description
-3	FGSG_13444		TP	related to allantoate transporter
-2	FGSG_13445		CYP, SP	probable benzoate 4-monooxygenase cytochrome P450
-1	FGSG_08085			conserved hypothetical protein
1	FGSG_08084		TP	related to monocarboxylate transporter 4
2	FGSG_08083			related to glutamic acid decarboxylase
3	FGSG_08082			conserved hypothetical protein
4	FGSG_08081			related to gibberellin 20-oxidase
5	FGSG_08080		TF	conserved hypothetical protein
6	FGSG_08079		CYP, SP	probable benzoate 4-monooxygenase cytochrome P450
7	FGSG_08078			related to general amidase
8	FGSG_08077			related to flavin oxidoreductase
+1	FGSG_08076			hypothetical protein
+2	FGSG_08075			conserved hypothetical protein
+3	FGSG_08074			conserved hypothetical protein

Table 5-3: Gene functions and features of the butenolide cluster

8 genes are co-expressed in planta, which contains genes encoding one cytochrome P450, one transcription factor, and one transporter.

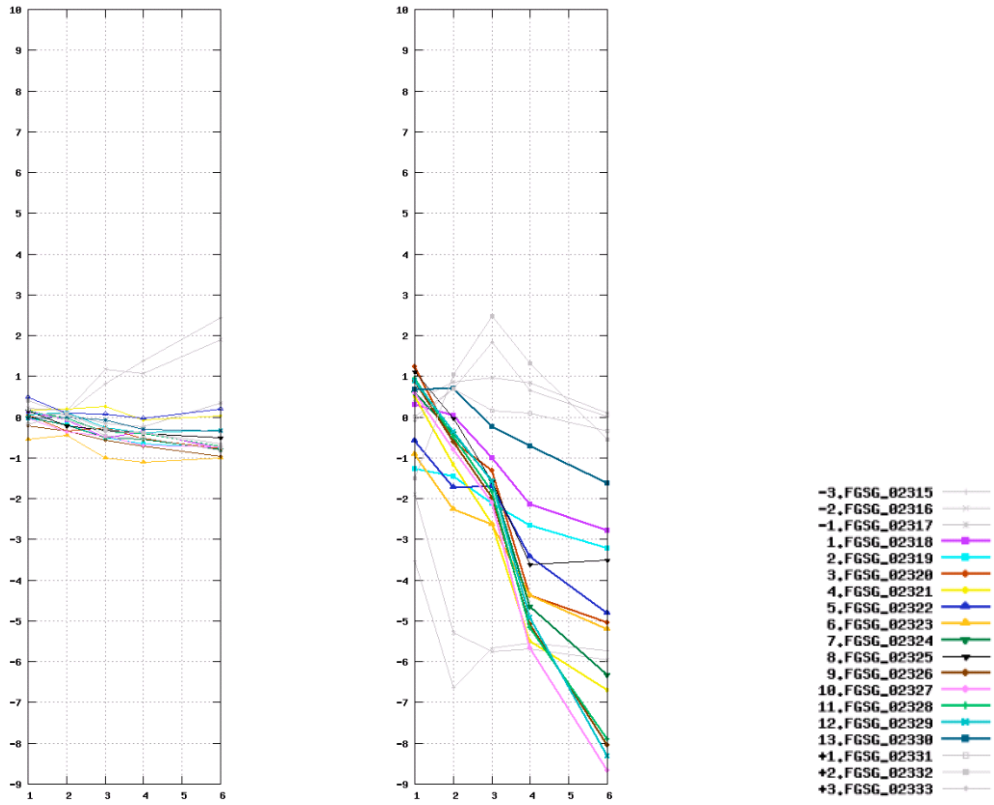
SP: Secreted protein, TF: Transcription factor, TP: Transporter, CYP: Cytochrome P450

5.1.3 The aurofusarin cluster

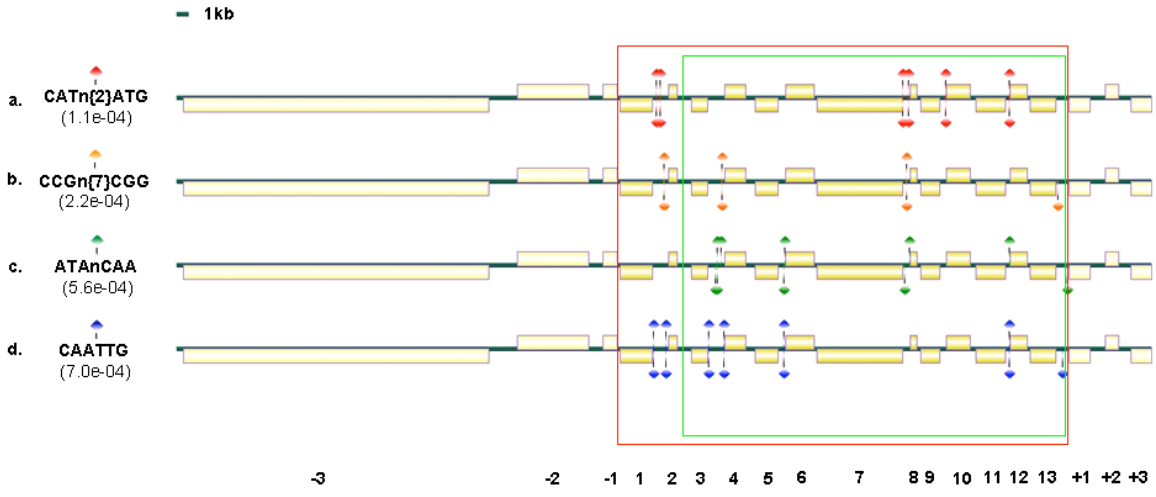
Aurofusarins are red or yellow pigments that cause stem and head blight of wheat and other small grains. The pigments are produced by several *Fusarium* species including *F. graminearum*, *F. culmorum*, and *F. pseudograminearum* [152]. The aurofusarin synthesis genes did not appear to be important for pathogenesis in barley roots and wheat heads [116]. Production of aurofusarin was described to be negatively correlated with vegetative growth and reduce mycelial growth [117]. The aurofusarin mycotoxin has been also reported to compromise the immune system of several poultry [153,154].

Currently, 11 genes (FGSG_02320(aurR1) ~ FGSG_02330 (aurL2) on chromosome 1) are known to be involved in aurofusarin synthesis in *F. graminearum* by several analyses [116,117,118]. By analysis of co-expressed neighbouring genes, 2 more genes (FGSG_02318 and FGSG_02319) upstream of aurR1 were newly identified. The 13 neighbouring genes show sharply decreasing expression patterns from day 1 until day 6 (Figure 5-5 (A) and Table 5-4). It was reported that PKS12 and aurR1 genes which are essential for the aurofusarin synthesis were not transcribed during sexual development [117], which is consistent with our result that the aurofusarin gene cluster was continuously down-regulated during sexual development. Regulatory motifs analysis based on the co-expressed gene cluster revealed that the two newly found genes share the motif seeds, 5'-CATn{2}ATG-3', 5'-CCGn{7}CGG-3', and 5'-

CAATTG-3', with some genes of the described aurofusarin cluster genes (Figure 5-5, B). Experimental validation of these promoter motifs has not been described.



A. A co-expressed gene cluster during sexual development (fg5_27, right) and profiles of the corresponding genes in planta (left)



B. Gene clusters with co-occurring regulatory motifs

Figure 5-5: The aurofusarin cluster identified by co-expression analysis and their promoter motifs

A. A co-expressed gene cluster during sexual development (fg5_27): 13 genes including 11 genes previously reported to produce aurofusarin, FGSG_02318 ~ FGSG_02330, on chromosome 1 were identified to be co-expressed during sexual development.

B. Four significant motif seeds are suggested as putative regulatory elements of the aurofusarin mycotoxin cluster by screening promoter regions of the co-expressed aurofusarin genes. Gene coding regions are represented by boxes: dark yellow refers to co-expressed genes and light yellow are flanking genes. Diamond shapes indicate motif seeds. The number in parenthesis is a p-value for each motif seed. Genes and motif seeds in the forward strand are displayed above the line; genes and motif seeds in the reverse strand are displayed below. Numbers underneath each box indicate gene designations (see Table 5-4).The 11 aurofusarin synthesis genes previously reported are highlighted in the green box. The red box encompasses the co-expressed genes during sexual development.

Gene order	FGDB ID	Gene	Gene feature	Gene description
-3	FGSG_02315	NPS4	SM	related to non-ribosomal peptide synthetase
-2	FGSG_02316		TP	related to multidrug resistance protein
-1	FGSG_02317			conserved hypothetical protein
1	FGSG_02318			conserved hypothetical protein
2	FGSG_02319			conserved hypothetical protein
3	FGSG_02320	aurR1	TF	pathway specific binuclear zinc cluster transcription factor for the aurofusarin gene cluster
4	FGSG_02321	aurO		oxidoreductase that catalyses the conversion of dimeric 9-hydroxyrubrofusarin to aurofusarin
5	FGSG_02322	aurT	TP	aurofusarin/rubrofusarin efflux pump AFLT
6	FGSG_02323	aurR2	FSTF	binuclear zinc cluster transcription factor that regulates the ratio between aurofusarin and rubr
7	FGSG_02324	PKS12	SM (PKS)	polyketide synthase that catalyse the condensation of one acetyl-CoA and six malonyl-CoA resultin
8	FGSG_02325			conserved hypothetical protein
9	FGSG_02326	aurJ		o-methyltransferase that catalyse the methylation of nor-rubrofusarin resulting in formation of rubrofusarin
10	FGSG_02327	aurF	SP	flavin depend monooxygenase that catalyses the oxidation of rubrofusarin to 9-hydroxyrubrofusarin
11	FGSG_02328	gip1	SP	laccase that catalyse the dimerization of two 9-hydroxyrubrofusarin in C7 positions
12	FGSG_02329		SP	conserved hypothetical protein
13	FGSG_02330	aurL2	SP	related to laccase precursor
+1	FGSG_02331		TP	related to monocarboxylate transporter
+2	FGSG_02332		SP	conserved hypothetical protein
+3	FGSG_02333			hypothetical protein

Table 5-4: Gene functions and features of the aurofusarin cluster

The aurofusarin mycotoxin cluster was identified by co-expression analysis. It contains 13 genes; 11 genes (# 3~13) currently referred to as the aurofusarin cluster in *F. graminearum* and 2 additional genes (#1~2, red box) newly identified by co-expression analysis. The aurofusarin cluster is enriched for secreted proteins and includes genes encoding one polyketide synthase gene (PKS12), two transcription factors and one transporter.

PKS: Polyketide synthase, SP: Secreted protein, FSTF: Fungal specific transcription factor, TF: Transcription factor, TP: Transporter, CYP: Cytochrome P450, SM: Secondary metabolite

5.1.4 The zearalenone cluster

Zearalenone (ZEA or ZON) is a nonsteroidal estrogenic mycotoxin produced by several *Fusarium* species [155]. It can be found primarily on maize but can occur at lower concentration in barley, wheat, and other cereals [155,156,157]. Zon can cause serious contamination of grain and feeds with synergistic effects if ZON occurs in combination with other mycotoxins [158,159]. ZON has effects on sexual development in fungi; it enhances as well as inhibits sexual reproduction depending on concentration and its activities are time dependant [160]. Most ZON-producing *Fusarium* species were observed to produce trichothecene mycotoxins. It has been also found to cause reproductive problems in animals. In particular it has genotoxic and carcinogenic effects in mice [161,162].

Studies by gene disruption and Real-Time Quantitative Expression have shown that four tightly linked genes are required for zearalenone biosynthesis [121,103]. Two polyketide synthases genes (PKS4 and PKS13) are known to be essential for the production of ZON. Two other genes encode a transcriptional regulator and an alcohol oxidase which controls oxidization of zearalenol to zearalenone. The analysis of the functional gene organization revealed 8 genes as zearalenone mycotoxin genes (Table 5-5). Three additional genes are positioned close to the upstream region of the previously reported ZON cluster; a non-ribosomal peptide synthetase gene (NPS15, FGSG_02394), a transporter gene (FGSG_12124) and a putative K+ channel beta subunit (FGSG_12125). It is not clear yet whether these genes are part of the same cluster. Only one gene encoding the putative K+ channel beta subunit (FGSG_12125) is known to be expressed in a manner similar to that of two PKS genes in the ZON cluster [163] and also reported to be clearly up-regulated in wheat [103], suggesting that it may have some related function with ZON production and a function during the infection process.

Regulatory motifs were examined based on two different sets of genes, the ZON cluster newly identified by our analysis and the previously reported ZON cluster genes (Figure 5-6). Regulatory motifs predicted from the two different sets of genes are different but both motif sets include the short promoter motif, 5'-CCG-3', which needs to be further investigated.

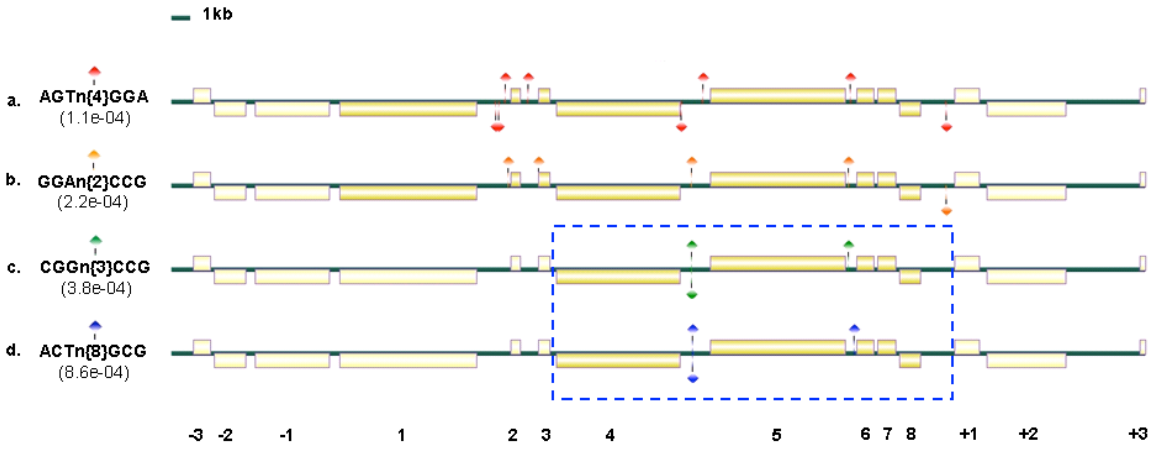


Figure 5-6: The zearalenone cluster identified by a particular composition of gene functions (tfcfg_15) and their promoter motifs

Regulatory motifs were examined based on two different sets of genes, the ZON cluster newly identified by our analysis (a and b, see Table 5-7) and the previously reported ZON cluster genes (highlighted in the blue dot line in c and d). Gene coding regions are represented by boxes: dark yellow refers to clustered genes and light yellow is flanking genes. Diamond shapes indicate motif seeds. The number in parenthesis is a p-value for each motif seed. Genes and motif seeds in the forward strand are displayed above the line; genes and motif seeds in the reverse strand are displayed below. Numbers underneath each box indicate gene designations (see Table 5-5).

Gene order	FGDB ID	Gene	Gene feature	Gene description
-3	FGSG_02391			hypothetical protein
-2	FGSG_02392			probable aldehyde dehydrogenase
-1	FGSG_02393			related to heterokaryon incompatibility protein het-6
1	FGSG_02394	NPS15	SM (NPS)	related to AM-toxin synthetase (AMT)
2	FGSG_12124			related to monocarboxylate transporter
3	FGSG_12125			probable potassium channel beta subunit
4	FGSG_02395	PKS13	SM (PKS)	polyketide synthase
5	FGSG_12126	PKS4	SM (PKS)	polyketide synthase
6	FGSG_12127		SP	probable isoamyl alcohol oxidase
7	FGSG_02397			probable isoamyl alcohol oxidase
8	FGSG_02398		TF	conserved hypothetical protein
+1	FGSG_02399			probable protein kinase Eg22
+2	FGSG_02400			probable calcium P-type ATPase NCA-3 [Ca2+-transporting ATPase]
+3	FGSG_12128			hypothetical protein

Table 5-5: Gene functions and features of the zearalenone cluster

8 genes are suggested as zearalenone mycotoxin genes by analysis of the particular composition of gene functions (tfcfg_15, Table 5-7). It contains three secondary metabolite genes (PKS4, PKS13 and NPS15) and one transcription factor. Two polyketide synthases genes are known to be essential for synthesis of zearalenone.

NPS: Non-ribosomal peptide synthetase, PKS: Polyketide synthase, SM: Secondary metabolite, TF: Transcription factor

5.1.5 The fusarin C cluster

Fusarin C has been found to occur naturally in both visibly *Fusarium*-infected corn kernels and healthy looking corn kernels [119,164,165]. Fusarin C is a highly mutagenic and potentially carcinogenic mycotoxin produced by several *Fusarium* species including *F. moniliforme* and *F. venenatum*, and *F. graminearum* and can cause chromosomal aberrations in mammalian cells [119,120,166,167]. Consumption of grains infected with fusarin-producing *Fusarium* species has been associated epidemiologically with human diseases.

Only one gene encoding PKS10 (FGSG_07798, a hybrid polyketide synthase-non-ribosomal peptide synthetase) is reported to be required for biosynthesis of fusarin C [112]. The PKS10 mutant was unable to synthesize fusarin C in *F. graminearum* and disruption of PKS10 reduced production of fusarin C in *F. venenatum* [112,166].

8 genes possibly involved in production of fusarin C (Table 5-6) are identified by analysis of the particular composition of gene functions (tfcfg_73, see Table 5-7). It contains genes encoding a cytochrome P450 and a transporter as well as PKS10. Based on the 8 genes, five motif seeds are suggested as putative promoter motifs of the fusarin C mycotoxin genes (Figure 5-7). Roles of each of gene in fusarin C biosynthesis remain to be characterized.

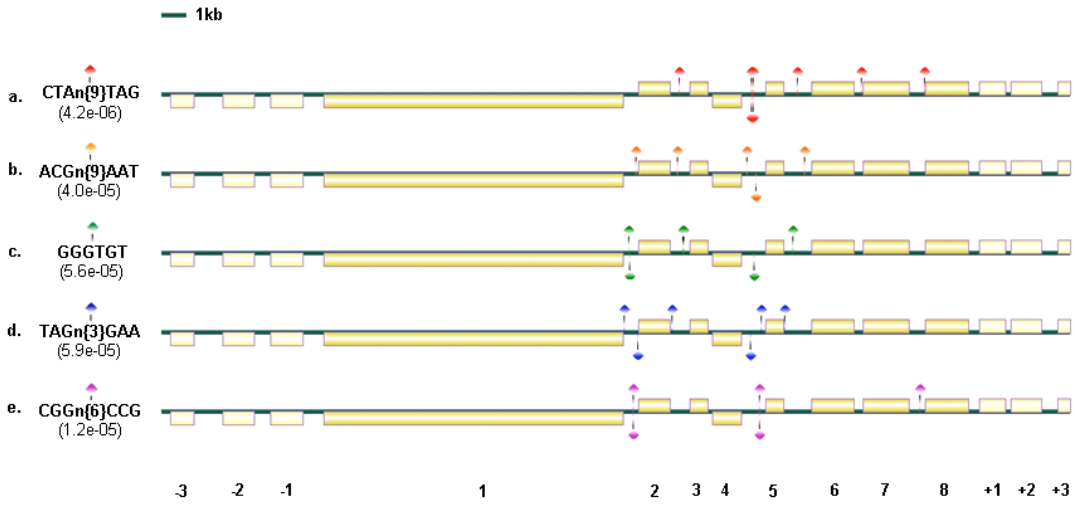


Figure 5-7: The fusarin C cluster identified by a particular composition of gene functions (tfcfg_73) and promoter motifs

Gene coding regions are represented by boxes: dark yellow refers to a gene cluster identified by a particular composition of gene functions (tfcfg_73, see Table 5-7) and light yellow is flanking genes. Diamond shapes indicate motif seeds. Genes and motif seeds in the forward strand are displayed above the line; genes and motif seeds in the reverse strand are displayed below. Numbers underneath each box indicate gene designations (see Table 5-6).

Gene order	FGDB ID	Gene	Gene feature	Gene description
-3	FGSG_07795			conserved hypothetical protein
-2	FGSG_07796			related to flavin-containing monooxygenase
-1	FGSG_07797			conserved hypothetical protein
1	FGSG_07798	PKS10	SM (PKS)	probable polyketide synthase
2	FGSG_13222			conserved hypothetical protein
3	FGSG_13223			related to elongation factor 1-gamma
4	FGSG_07800			related to pepsin A-1 precursor
5	FGSG_07801			conserved hypothetical protein
6	FGSG_07802		TP	related to putative multidrug transporter Mfs1.1 (major facilitator family protein)
7	FGSG_07803			related to aldehyde dehydrogenase (NAD+), mitochondrial
8	FGSG_07804		CYP	related to benzoate 4-monooxygenase cytochrome P450
+1	FGSG_07805			conserved hypothetical protein
+2	FGSG_13224			related to epoxide hydrolase
+3	FGSG_13225			conserved hypothetical protein

Table 5-6: Gene functions and features of the fusarin C cluster

8 genes possibly involved in the production of fusarin C are identified by analysis of the particular compostion of gene functions (tfcfg_73, see Table 5-7). Currently, only one gene encoding PKS10 (FGSG_07798, a hybrid polyketide synthase-non-ribosomal peptide synthetase) is reported to be required for biosynthesis of fusarin C. Roles for the 7 other genes remain to be established.

PKS: Polyketide synthase, SM: Secondary metabolite, TP: Transporter, CYP: Cytochrome P450

5.2 Tentative functional gene clusters (TFCs)

The major aim of the analysis is to screen sets of neighboring genes that have gene compositions similar to known gene clusters. Products of clustered genes in fungi, which carry functional parts in common pathways (e.g. secondary metabolic pathways) often show a specific composition of functions. Particular types of key enzymes, mostly secondary metabolite (SM) genes (i.e. polyketide synthase (PKS), non-ribosomal peptide synthetase (NPS), and terpenoid synthase (TPS)) and cytochrome P450 genes have been found to be clustered with various combinations of additional enzymes for further metabolite catalyzing. Also transporters, and/or transcription factors that are essential for the regulation of most of the clustered genes are often part of the cluster. Other types of clusters enriched in genes encoding cytochrome P450 were found in mycotoxin gene clusters (e.g. gibberellin) [97]. Thus, enrichment of cytochrome P450 enzymes can be a useful indicator to predict gene clusters that do not contain any other gene. Unlike gene clusters responsible for secondary metabolite biosynthesis, which often contain regulatory proteins and diverse enzymes, no such functional genes have been observed within the clusters of secreted protein genes in *U. maydis*. Although these gene clusters are not known to code for a mycotoxin pathway, they have been shown to modulate the fungal-plant interaction, thus the peptides most likely act as effectors [6]. From these observations, five appropriate protein functions, SMs, cytochrome P450s, secreted proteins, transcription factors, and transporters, (s1–s5, Table 4-1) were selected using data from InterPro [125], TargetP [126], and PKS/NPS annotation and applied in generating tentative functional gene clusters (TFCs). In these TFCs different functions are combined or an amplification of one function occurs.

The statistical significance to identify TFCs was calculated by comparison to the occurrence of the same set of functional descriptors in random genomes. Overlapping TFCs were combined into one tentative functional gene cluster. Up to 3 genes without such functions assigned were allowed between genes with five types of functional descriptors in each TFC. A TFC can be composed of an arbitrary number of genes as long as this condition is satisfied.

A total of 81 significant TFCs presumably involved in known or even new pathways or functions were identified at the genomic level (Table 5-7). Notably, the 5 known mycotoxin gene clusters reported in *F. graminearum* are all included in these TFCs (green-shaded boxes in Table 5-7). They contained at least one key enzyme (SM gene and/or cytochrome P450) together with secreted proteins, transcription factors and/or transporters. The aurofusarin gene cluster (tfcfg_73) comprises enrichment

of secreted proteins and the trichothecene gene cluster (tfcfg_41) combines enrichment of cytochrome P450 genes and secreted proteins.

About 92% of the SM genes including 14 PKS, 19 NPS and 14 terpenoid synthase (47 out of the 51) and 70% of the cytochrome P450 genes (79 out of 117) belong to TFCs. 22 SM genes and 47 cytochrome P450 genes are co-localized with genes with three functional descriptors (transcription factors, transporters, and screted proteins).

About 24 % of the cytochrome P450 genes (28 out of 117) are clustered in 10 TFCs (blue-shaded box in Table 5-7). Interestingly, the most significant gene cluster enriched for cytochrome P450 genes (tfcfg_10) was identified as a co-expressed gene cluster during sexual development (fg5_22, Table 5-10, B). Another gene cluster enriched for cytochrome P450 genes (tfcfg_43) is also discovered under the same condition (fg5_43, Table 5-10, B). Both gene clusters overlapped with co-expressed neighboring genes have similar expression profiles with the aurofusarin mycotoxin genes but their biological functions are unknown.

Gene regions identified as TFCs provide an important basis for further research. However, genes included in the TFCs do not exactly determine the boundaries of gene clusters because this screening basically predicts gene ranges possessing only genes encoding the applied five types of function. To more precisely predict clustered genes, additional evidences to characterize functions or functional categories are required. Inclusion of further features into the screening may consolidate or enlarge the found clusters or may identify potential novel ones. Additional support for the biological relevance of the 107 TFCs will arise from the discovery of two further features, based on co-expression and co-occurring of regulatory elements, crucial factors in co-regulation of fungal gene clusters. The set of 107 TFCs was extended by including eight single SM genes, not yet found in TFCs, because co-expression or co-regulation of SM genes and cluster genes can indicate a possible role of this cluster in secondary metabolism.

TFC ID (tfcfg_#)	Chr.	(N.ofGene),Gene range	N. of gene					SPC ID (spfg_#)
			SM	CYP	TF	TP	SP	
1	1	(14), FGSG_00006 ~ FGSG_00015		2	1	2	6	1
2	1	(17), FGSG_00023 ~ FGSG_11655		1	1	2	5	2
3	1	(2), FGSG_11659 ~ FGSG_11660	2 (NPS8, NPS-FGSG_11660)	0	0	0	0	
4	1	(5), FGSG_00067 ~ FGSG_00071		1	1	2	1	
5	1	(6), FGSG_00451 ~ FGSG_00456	1 (TPS_FGSG_00451)	0	0	0	2	
6	1	(12), FGSG_01680 ~ FGSG_01688	2 (NPS16, NPS19)	0	0	1	3	
7	1	(21), FGSG_01722 ~ FGSG_01740	1 (TPS-FGSG_01738)	3	4	1	5	3
8	1	(15), FGSG_01760 ~ FGSG_01771		1	1	1	6	
9	1	(24), FGSG_01776 ~ FGSG_01799	2 (TPS_FGSG_01783, PKS11)	1	1	1	5	
10	1	(13), FGSG_12067 ~ FGSG_02118		5	0	0	9	4
11	1	(25), FGSG_02136 ~ FGSG_02158		1	1	4	7	
12	1	(32), FGSG_02307 ~ FGSG_02338	2 (NPS4, PKS12)	0	2	5	10	5
13	1	(5), FGSG_02343 ~ FGSG_02348		1	0	1	1	
14	1	(25), FGSG_02366 ~ FGSG_02389		3	2	1	8	
15	1	(8), FGSG_02394 ~ FGSG_02398	3 (NPS15, PKS13, PKS4)	0	1	1	1	
16	1	(2), FGSG_12141 ~ FGSG_02458		1	0	1	1	
17	1	(23), FGSG_02668 ~ FGSG_02688		2	1	4	7	6
18	1	(4), FGSG_09913 ~ FGSG_09916		1	0	1	1	
19	1	(15), FGSG_10384 ~ FGSG_10397	1 (TPS-FGSG_10397)	0	1	1	3	
20	1	(18), FGSG_10458 ~ FGSG_10474	1 (PKS9)	1	4	2	2	
21	1	(4), FGSG_10523 ~ FGSG_10525	1 (NPS3)	0	0	0	1	
22	1	(4), FGSG_13782 ~ FGSG_13783	1 (NPS18)	0	0	0	2	
23	1	(8), FGSG_10547 ~ FGSG_10554	1 (PKS1)	0	0	1	3	
24	1	(2), FGSG_10570 ~ FGSG_10571		1	0	1	2	
25	1	(2), FGSG_13796 ~ FGSG_13797		2	0	0	0	
26	2	(20), FGSG_08810 ~ FGSG_08791	1 (PKS7)	1	4	5	1	
27	2	(3), FGSG_08412 ~ FGSG_08410		1	0	1	1	
28	2	(13), FGSG_08380 ~ FGSG_08369		1	2	1	2	
29	2	(4), FGSG_08210 ~ FGSG_08207	2 (NPS7, PKS6)	1	0	0	1	
30	2	(19), FGSG_08193 ~ FGSG_08177	1 (TPS-FGSG_08181)	4	2	1	4	
31	2	(17), FGSG_08093 ~ FGSG_08079		2	1	3	6	7
32	2	(31), FGSG_08030 ~ FGSG_13459		2	4	4	6	
33	2	(5), FGSG_04592 ~ FGSG_04588	2 (TPS-FGSG_04591, PKS15)	1	0	0	1	
34	2	(2), FGSG_12283 ~ FGSG_12284		2	0	0	0	
35	2	(37), FGSG_03987 ~ FGSG_03956	1 (PKS14)	1	0	6	11	
36	2	(10), FGSG_03862 ~ FGSG_15209		1	2	2	1	
37	2	(16), FGSG_12369 ~ FGSG_03840		1	0	1	4	
38	2	(7), FGSG_03801 ~ FGSG_03795		1	0	1	2	
39	2	(28), FGSG_12385 ~ FGSG_03732	1 (NPS6)	1	1	5	6	8
40	2	(10), FGSG_03702 ~ FGSG_03695		1	1	1	4	9
41	2	(17), FGSG_03542 ~ FGSG_03526	1 (TPS-TRI5)	3	2	1	5	10
42	2	(6), FGSG_03498 ~ FGSG_03494	1 (TPS-FGSG_03494)	1	0	0	2	
43	2	(9), FGSG_03348 ~ FGSG_03340	1 (PKS8)	0	0	1	3	
44	2	(19), FGSG_03275 ~ FGSG_03260		2	0	1	6	
45	2	(6), FGSG_03246 ~ FGSG_03242	1 (NPS11)	0	1	0	2	
46	2	(13), FGSG_03009 ~ FGSG_02998		1	0	1	5	
47	2	(6), FGSG_02982 ~ FGSG_02978		1	0	1	1	
48	2	(13), FGSG_02874 ~ FGSG_02862		1	1	3	1	
49	2	(30), FGSG_11513 ~ FGSG_11540		1	1	2	8	
50	3	(11), FGSG_04717 ~ FGSG_12599		2	1	1	1	
51	3	(7), FGSG_05366 ~ FGSG_05372	1 (NPS2)	0	1	1	1	

52	3	(26), FGSG_12813 ~ FGSG_12816	1 (PKS5)	1	3	2	9	11
53	3	(2), FGSG_06068 ~ FGSG_06069		1	0	1	1	
54	3	(7), FGSG_11113 ~ FGSG_11108		1	0	1	2	
55	3	(22), FGSG_13863 ~ FGSG_11024	1 (NPS1)	2	1	3	6	12
56	3	(40), FGSG_11011 ~ FGSG_10977	2 (NPS9, NPS5)	2	2	1	11	13
57	3	(6), FGSG_10938 ~ FGSG_10933	1 (TPS-FGSG_10933)	0	0	2	1	
58	3	(13), FGSG_10921 ~ FGSG_10910		1	1	3	3	
59	3	(6), FGSG_11499 ~ FGSG_11495		1	0	1	2	
60	3	(8), FGSG_11465 ~ FGSG_11458		1	1	1	1	
61	3	(5), FGSG_11399 ~ FGSG_11395	1 (NPS14)	0	0	0	1	
62	3	(8), FGSG_13971 ~ FGSG_11385		1	0	1	3	
63	3	(10), FGSG_11327 ~ FGSG_11318	1 (TPS-FGSG_11327)	0	0	1	2	
64	3	(17), FGSG_11310 ~ FGSG_11294	1 (NPS12)	1	0	0	7	14
65	3	(13), FGSG_11282 ~ FGSG_11271		1	1	2	2	
66	3	(7), FGSG_14006 ~ FGSG_11621		2	0	1	2	
67	4	(13), FGSG_06442 ~ FGSG_06453	1 (TPS-FGSG_06444)	1	2	1	4	15
68	4	(8), FGSG_06503 ~ FGSG_06509	1 (NPS10)	0	1	2	1	
69	4	(6), FGSG_15003 ~ FGSG_06784	1 (TPS-FGSG_06784)	0	0	0	2	
70	4	(13), FGSG_13153 ~ FGSG_07491	1 (NPS13)	0	1	0	3	
71	4	(17), FGSG_07582 ~ FGSG_07596		1	2	4	1	
72	4	(14), FGSG_07661 ~ FGSG_07673	1 (TPS-FGSG_07673)	0	0	3	3	
73	4	(8), FGSG_07792 ~ FGSG_07798	1 (PKS10)	0	0	0	2	
74	4	(8), FGSG_07802 ~ FGSG_07808		1	0	1	2	
75	4	(33), FGSG_09060 ~ FGSG_09090		1	4	4	7	
76	4	(7), FGSG_09177 ~ FGSG_09182	1 (PKS3)	0	2	0	1	
77	4	(31), FGSG_09367 ~ FGSG_09396	1 (TPS-FGSG_09381)	0	3	2	8	
			1 (TPS-FGSG_10097)			1 (+1)	1 (-5)	
			1 (NPS17)			1 (+5)		
			1 (TPS-FGSG_03066)		1 (-2)			
			1 (PKS2)			1 (+5)	1 (+4)	

Table 5-7: 77 tentative functional gene clusters (TFCs) deduced by 5 types of functional descripters in *F. graminearum*

Using five types of functional descriptors (Table 4-1), 77 tentative functional gene clusters (TFCs) were identified at genomic level. It includes 11 TFCs enriched with the function of cytochrome P450, and 15 TFCs enriched with secreted proteins. TFCs in green-shaded boxes are overlapped with known mycotoxin gene or gene clusters in *F. graminearum*: tfcfg_12 - Aurofusarin, tfcfg_15 - Zearalenone, tfcfg_31- Butenolide, tfcfg_41 - Trichothecene, tfcfg_73 - Fusarin C. Detailed gene information in all TFCs are provided in FGDB.

23 TFCs including 3 mycotoxin gene clusters (aurofusarin, butenolide, and trichothecene) are supported by two important factors in co-regulation of fungal gene clusters, co-expression and/or the presence of putative regulatory elements in promoter regions (TFC ID in red, see Table 5-10 and Table 5-11). TFCs with blue-shaded boxes in SM and P450 columns include clusters enriched for genes encoding secondary metabolites and cytochrome P450s, respectively. SM genes with underlines have information of related functions and expression specificity under various conditions (Table 3-4 and Table 3-5).

The 4 single SM genes (gray text) are also co-located to genes with interesting functional features within a maximum of 10 neighboring genes, but they were not statistically significant at the genomic level. The number in parenthesis refers to gene position information based on SM genes (e.g. -2 means the second adjacent gene upstream to the terpenoid synthase gene (FGSG_03066)).

TFC: tentative functional gene cluster, SPC: secreted proteins cluster, NPS: Non-ribosomal peptide synthetase, PKS: Polyketide synthase, TPS: Terpenoid synthase, Chr.: Chromosome, SM: Secondary metabolite, SP: Secreted protein, TF: Transcription factor, TP: Transporter, CYP: Cytochrome P450

5.3 Secreted protein gene clusters

Of 10.1% (1,413) proteins predicted to be secreted (Reliability class 1 and 2 of TargetP [126]), about 18% (257) of secreted proteins are arranged in 67 clusters with 3-16 spanned genes by analysis of functional enrichment (Table 5-8). One gene without secreted protein function assigned was allowed between genes with the function in each cluster. 15 secreted protein clusters overlap with TFCs (spfg_1 ~ spfg_15 in Table 5-8).

SPC ID (spfg_#)	Chr.	(N.ofGene),Gene range	N. of SP gene	TFC ID (tfcfg_#)
1	1	(11), FGSG_00006 ~ FGSG_11648	6	1
2	1	(4), FGSG_00027 ~ FGSG_00029	3	2
3	1	(4), FGSG_01727 ~ FGSG_01730	3	7
4	1	(13), FGSG_12067 ~ FGSG_02118	9	10
5	1	(8), FGSG_02327 ~ FGSG_02334	6	12
6	1	(6), FGSG_02683 ~ FGSG_02687	5	17
7	2	(7), FGSG_08090 ~ FGSG_13445	5	31
8	2	(4), FGSG_03742 ~ FGSG_12390	3	39
9	2	(4), FGSG_03700 ~ FGSG_03698	3	40
10	2	(4), FGSG_03532 ~ FGSG_03529	3	41
11	3	(9), FGSG_05803 ~ FGSG_12816	6	52
12	3	(3), FGSG_11033 ~ FGSG_13866	3	55
13	3	(5), FGSG_10986 ~ FGSG_10982	5	56
14	3	(5), FGSG_11306 ~ FGSG_11302	4	64
15	4	(3), FGSG_06450 ~ FGSG_06452	3	67
16	1	(5), FGSG_00059 ~ FGSG_00063	4	
17	1	(4), FGSG_00111 ~ FGSG_00114	3	
18	1	(3), FGSG_01169 ~ FGSG_01170	3	
19	1	(3), FGSG_11973 ~ FGSG_01595	3	
20	1	(6), FGSG_01828 ~ FGSG_01833	4	
21	1	(3), FGSG_01993 ~ FGSG_01995	3	
22	1	(3), FGSG_02189 ~ FGSG_02191	3	
23	1	(6), FGSG_02255 ~ FGSG_02258	4	
24	1	(3), FGSG_02527 ~ FGSG_02529	3	
25	1	(4), FGSG_10167 ~ FGSG_10170	3	
26	1	(3), FGSG_10560 ~ FGSG_10562	3	
27	2	(4), FGSG_08496 ~ FGSG_08493	3	
28	2	(3), FGSG_08122 ~ FGSG_08120	3	
29	2	(5), FGSG_07996 ~ FGSG_07993	3	
30	2	(5), FGSG_04550 ~ FGSG_04546	3	
31	2	(8), FGSG_03911 ~ FGSG_03904	6	
32	2	(4), FGSG_03897 ~ FGSG_03894	3	
33	2	(7), FGSG_03691 ~ FGSG_03685	5	
34	2	(7), FGSG_03629 ~ FGSG_03624	5	
35	2	(5), FGSG_03601 ~ FGSG_03598	4	
36	2	(5), FGSG_03585 ~ FGSG_03581	4	
37	2	(4), FGSG_03334 ~ FGSG_15182	3	
38	2	(4), FGSG_03309 ~ FGSG_03307	3	
39	2	(5), FGSG_15174 ~ FGSG_03209	3	

40	2	(6), FGSG_15175 ~ FGSG_03191	4
41	2	(3), FGSG_03167 ~ FGSG_03165	3
42	2	(8), FGSG_12502 ~ FGSG_03129	5
43	2	(4), FGSG_03124 ~ FGSG_03121	4
44	2	(5), FGSG_02914 ~ FGSG_02910	3
45	2	(4), FGSG_11546 ~ FGSG_15661	3
46	3	(16), FGSG_04732 ~ FGSG_04745	11
47	3	(4), FGSG_04780 ~ FGSG_04783	3
48	3	(4), FGSG_04856 ~ FGSG_04858	3
49	3	(4), FGSG_05059 ~ FGSG_05061	3
50	3	(5), FGSG_05080 ~ FGSG_05084	4
51	3	(10), FGSG_11233 ~ FGSG_11225	6
52	3	(6), FGSG_11208 ~ FGSG_11204	5
53	3	(3), FGSG_11171 ~ FGSG_11169	3
54	3	(5), FGSG_11129 ~ FGSG_11125	3
55	3	(4), FGSG_11088 ~ FGSG_11085	3
56	3	(4), FGSG_11049 ~ FGSG_11046	4
57	3	(3), FGSG_10959 ~ FGSG_10957	3
58	3	(5), FGSG_10787 ~ FGSG_10784	3
59	4	(5), FGSG_06463 ~ FGSG_06467	4
60	4	(4), FGSG_06610 ~ FGSG_06613	3
61	4	(6), FGSG_13092 ~ FGSG_07208	4
62	4	(3), FGSG_13189 ~ FGSG_07696	3
63	4	(5), FGSG_15448 ~ FGSG_07721	3
64	4	(3), FGSG_07859 ~ FGSG_07861	3
65	4	(3), FGSG_09132 ~ FGSG_09134	3
66	4	(3), FGSG_09141 ~ FGSG_09143	3
67	4	(5), FGSG_09646 ~ FGSG_09650	4

Table 5-8: 67 gene clusters enriched with secreted proteins in *F. graminearum*

Using TargetP[126] information (Reliability class 1 and 2, Table 4-1), 67 secreted proteins clusters (SPCs) were identified at the genomic level. 15 out of the 67 SPCs overlap with tentative functional gene clusters (TFCs) identified by specific compositions of gene functions (Table 5-7). SPCs in green-shaded boxes are overlapped with known mycotoxin gene or gene clusters in *F. graminearum*: spfg_5 - Aurofusarin, spfg_10 - Trichothecene.

TFC: tentative functional gene cluster, SPC: secreted proteins cluster, Chr.: Chromosome

5.4 Co-expressed neighboring genes

5.4.1 Identification of co-expressed neighboring genes

Gene expression values were obtained from two independent analyses of differential expression profiles during growth in planta (FG1) [127] and sexual development (FG5) [128]. For each experiment, the mean Pearson correlation coefficient of all pairs of 5 to 25 neighboring genes was calculated to provide a measure of the similarity in their expression profiles. The significance of this value was determined by the 95th percentile from the distribution for random mean R from the same set of data. In total, 405 genes and 655 genes were found in 76 (FG1) and 147 (FG5) co-expressed neighboring genes (clusters), respectively (Table 5-9).

Co-expressed neighboring genes in planta (FG1)			
Gene number in cluster	Number of clusters	Gene number	Proportion (%) in all genes
3	22	66	0.47
4	11	44	0.32
5	15	75	0.54
6	7	42	0.30
7	7	49	0.35
8	8	64	0.46
9	1	9	0.06
10	3	30	0.22
11	1	11	0.08
15	1	15	0.11
Total	76	405	2.91
Co-expressed neighboring genes during sexual development (FG5)			
Gene number in cluster	Number of clusters	Gene number	Proportion (%) in all genes
3	48	144	1.03
4	41	164	1.18
5	30	150	1.08
6	15	90	0.64
7	8	56	0.40
8	1	8	0.06
9	1	9	0.06
10	1	10	0.07
11	1	11	0.08
13	1	13	0.09
Total	147	655	4.70

Table 5-9: Clusters identified as co-expressed neighboring genes

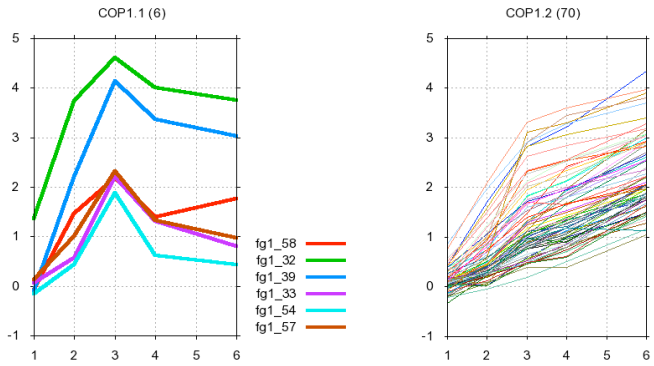
Gene expression values were obtained from two independent analyses of differential transcript accumulation during growth in planta (FG1) and sexual development (FG5). In total, 405 genes and 655 genes were arranged in 76 (FG1) and 147 (FG5) co-expressed gene clusters, respectively. Expression profiles and detailed gene information of all clusters of co-expressed neighboring genes are provided in FGDB.

5.4.2 Characteristics of co-expressed gene clusters

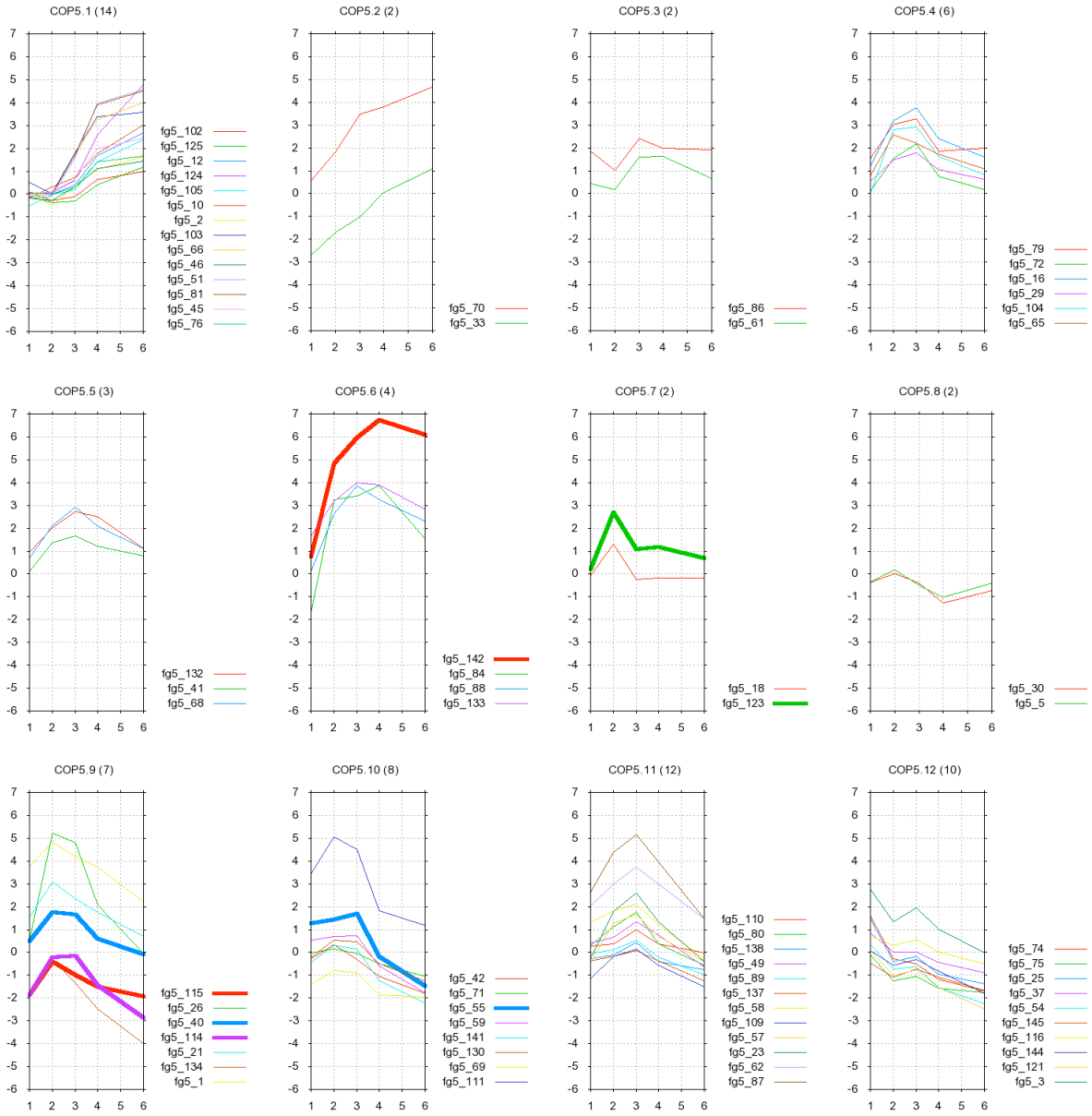
To support analysis of patterns of expression profiles for the two experiments, average values for each time point were generated to create representative profiles for the clusters. For 76 and 147 representative profiles in two experiments, co-expressed patterns (COP) were defined by 0.75 similarity of hierarchical clustering. Characteristics of each of the co-expressed gene clusters were examined by using TFCs (Table 5-7). By co-expression analysis, three mycotoxin gene clusters were identified; butenolide [90] (fg1_32) and trichothecene [102,134] (fg1_39) in planta and aurofusarin [168] (fg5_27) during sexual development.

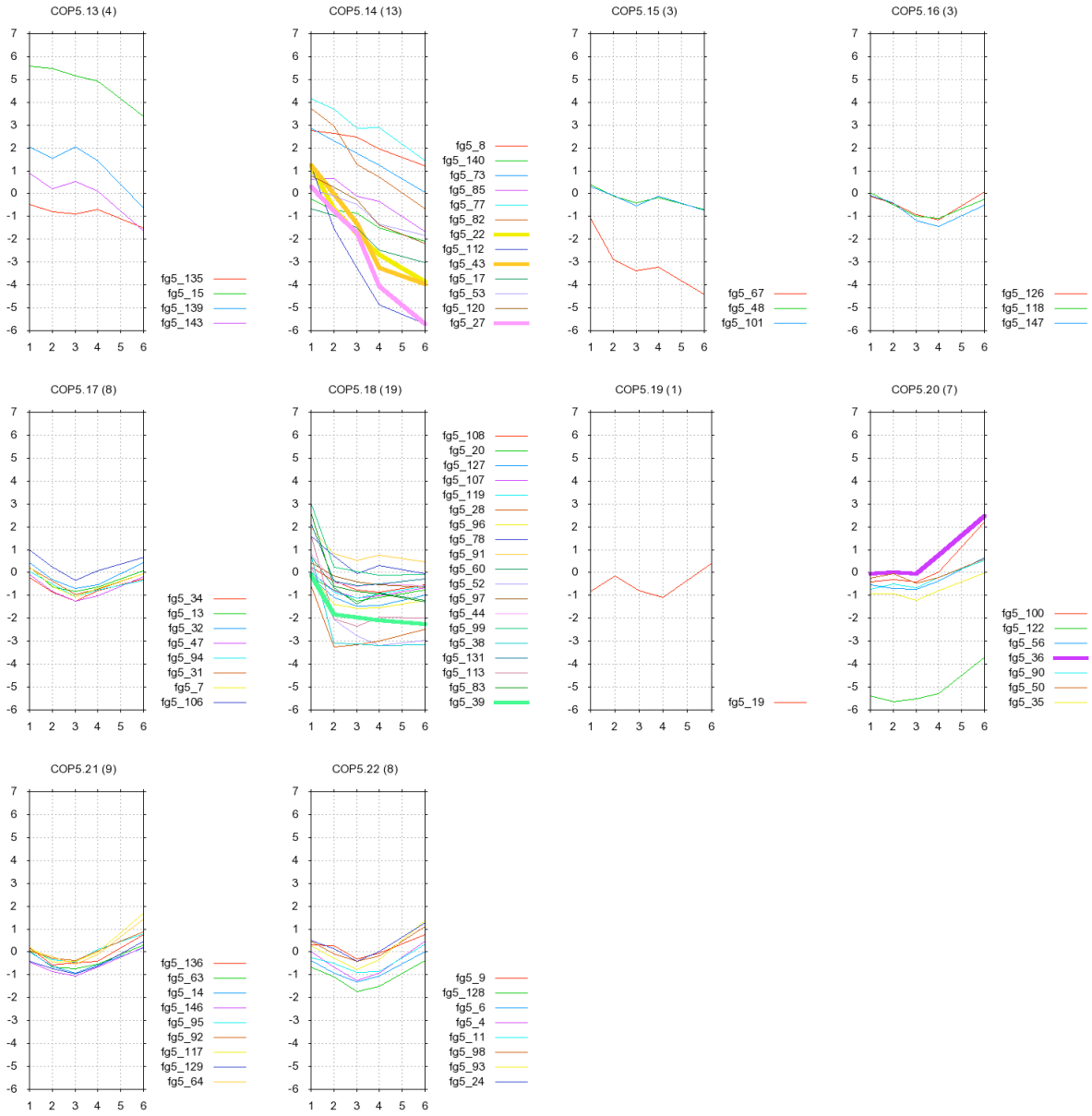
Two major co-expressed patterns (COP) were observed in gene clusters co-expressed in planta (FG1). 6 gene clusters (COP1.1) have peaks in their expression patterns at the time point of the 3rd day and the other 70 clusters (COP1.2) have continuously increased expression (Figure 5-8, A). All 6 clusters with peaks on the 3rd day have known composition of gene functions as shown in many secondary metabolite gene clusters and/or secondary metabolite (SM) genes (Table 5-10). Two gene clusters coincide with known mycotoxin gene clusters, butenolide [90] (fg1_32) and trichothecene [102,134] (fg1_39). The four other clusters contain SM genes or enrichment of secreted proteins; PKS15 and terpenoid synthase (fg1_33), NPS9 and NPS5 (fg1_54), NPS14 (fg1_57), and enrichment of secreted proteins (fg1_58). The PKS15 is reported to be exclusively expressed during plant infection [112] and others are functionally not reported.

Gene clusters co-expressed during sexual development showed diverse expression patterns (Figure 5-8, B). 11 gene clusters were found to have features of known gene clusters and 6 of those including SM genes belong all to different COPs. Cluster fg5_27 including the aurofusarin cluster [168] has a steeply decreasing expression pattern (COP5.14). Two PKS genes, PKS9 and PKS3, reported to be involved in sexual development also showed different expression patterns: cluster fg5_36 including PKS9 gene showed gradually increased expression profiles after the 3rd day (COP5.20) and cluster fg5_142 with the PKS3 gene was highly increased until the 4th day (COP5.6). Cluster fg5_39 (COP5.18) with decreasing pattern until the 2nd day contains PKS1 known to be related to mycelial growth. Cluster fg115 (COP5.9) with NPS1 has a peak in their expression patterns at the time point of the 2nd day. Cluster fg5_55 (COP5.10) with PKS6 and NPS7 has a peak at the time point of the 2nd day and sharply decreasing after the 2nd day.



A: expression patterns of clusters of co-expressed neighboring genes in planta (FG1)





B: expression patterns of clusters of co-expressed neighboring genes during sexual development (FG5)

Figure 5-8: Expression patterns of clusters of co-expressed neighboring genes

Co-expressed patterns (COP): the x-axis indicates the time points (day) and the y-axis indicates $\log_2$ fold change of expression values. Each line represents one cluster of co-expressed neighboring genes. The number in parenthesis refers to number of co-expressed gene clusters grouped in the same COP.Bold lines in graphs indicate clusters of co-expressed neighboring genes with functional features of interest described in Table 5-10.

Cluster ID	Chr.	(N. of gene), gene range	COP	N. of gene SM gene	CYP	TF	TP	SP	TFC ID
fg1_32	2	(8), FGSG_08084 ~ FGSG_08077	COP1.1		1	1	1	1	tfcfg_31
fg1_33	2	(10), FGSG_04596 ~ FGSG_04588	COP1.1	2 (TPS-FGSG_04591, PKS15)	1			1	tfcfg_33
fg1_39	2	(15), FGSG_03543 ~ FGSG_03529	COP1.1	1 (TRI5)	3	2	1	4	tfcfg_41
fg1_54	3	(8), FGSG_10995 ~ FGSG_13878	COP1.1	2 (NPS9, NPS5)	1		1	1	tfcfg_56
fg1_57	3	(5), FGSG_11399 ~ FGSG_11395	COP1.1	1 (NPS14)				1	tfcfg_61
fg1_58	3	(10), FGSG_11307 ~ FGSG_11298	COP1.1		1			5	tfcfg_64

A. 6 co-expressed gene clusters with known features in FG1

Cluster ID	Chr.	(N. of gene), gene range	COP	N. of gene SM gene	CYP	TF	TP	SP	TFC ID
fg5_22	1	(9), FGSG_02111 ~ FGSG_02119	COP5.14		5			5	tfcfg_10
fg5_27	1	(13), FGSG_02318 ~ FGSG_02330	COP5.14	1 (PKS12)		2	1	4	tfcfg_12
fg5_36	1	(4), FGSG_10461 ~ FGSG_10464	COP5.20	1 (PKS9)	1			1	tfcfg_20
fg5_39	1	(5), FGSG_10545 ~ FGSG_10549	COP5.18	1 (PKS1)			1	1	tfcfg_23
fg5_40	1	(5), FGSG_10569 ~ FGSG_10573	COP5.9		1		1	2	tfcfg_24
fg5_43	1	(8), FGSG_10608 ~ FGSG_10614	COP5.14		2			1	tfcfg_25
fg5_55	2	(5), FGSG_08209 ~ FGSG_08205	COP5.10	2 (NPS7, PKS6)	1				tfcfg_29
fg5_114	3	(3), FGSG_11033 ~ FGSG_13866	COP5.9				1	3	tfcfg_55
fg5_115	3	(4), FGSG_11029 ~ FGSG_11026	COP5.9	1 (NPS1)			2		
fg5_123	4	(5), FGSG_06447 ~ FGSG_06450	COP5.7		1	1		1	tfcfg_67
fg5_142	4	(6), FGSG_09182 ~ FGSG_09187	COP5.6	1 (PKS3)					tfcfg_76

B. 11 co-expressed gene clusters with known features in FG5

Table 5-10: Co-expressed gene clusters with functional features of interest

Characteristics of each co-expressed gene cluster were examined by comparing with TFCs (Table 5-7). Gene clusters in green-shaded boxes refer to known mycotoxin clusters in *F. graminearum*; butenolide [90] (fg1_32), trichothecene [102,134] (fg1_39), and aurofusarin [116,117,118] (fg5_27). SM genes with underlines have information of related functions and expression specificity under various conditions (Table 3-4 and Table 3-5). A. 6 gene clusters with peaks in their expression patterns at the time point of the 3rd day contain features of known gene clusters. Other 70 co-expressed gene clusters with continuously increasing expression patterns have not been found features of interest at present. B. 11 co-expressed gene clusters with features of interest show diverse expression patterns.

Chr.: Chromosome. COP: Co-expressed pattern. SM: Secondary metabolite. PKS: Polyketide synthase, NPS: Non-ribosomal peptide synthetase, TPS: Terpenoid synthase, TFC: Tentative functional gene cluster

5.5 Gene clusters with co-occurring promoter motifs

Clusters of co-regulated genes are often found to have identical or consensus regulatory motifs in the promoter region. To search for gene clusters that might be co-regulated, the promoters of neighboring genes spanning 3-25 genes were scanned for co-occurring motifs without prior knowledge of biological functions. Regulatory elements of neighboring genes were identified by probabilistic over-representation at the genomic level. In total, about 8.5 percent (1182/13935) of predicted genes in *F. graminearum* are arranged in 95 clusters having 6 to 25 genes with co-occurring regulatory motifs (see Table S 8-1). Gene clusters were ranked according to the adjusted p-values of the motifs. The found 95 gene clusters were examined with respect to the specific compositions of gene functions identified as TFCs (Table 5-7) and expression data from two microarray experiments. In total 19 gene clusters including the trichothecene cluster were assigned to tentative functional gene clusters (TFCs, Table 5-7), secreted protein gene clusters (SPCs, Table 5-8) or overlapped with co-expressed gene clusters (Table 5-11).

The trichothecene mycotoxin cluster was identified in cluster msfg_14 (Table 5-11, Figure 5-2, and Table 5-2). The core genes in the trichothecene cluster [102] require the zinc-finger containing transcription factor TRI6 for pathway activation. In *Fusarium sporotrichioides*, the TRI6 binding promoter motif, 5'-TnAGGCCT-3', was found in promoter regions of trichothecene genes [31]. The motif 5'-AGGCCT-3' found in the *F. graminearum* cluster msfg_14 overlaps the *F. sporotrichioides* promoter motif. The cluster msfg_14 include 3 additional genes, which is consistent with co-expressed neighboring genes identified in chapter 5.4. The comprehensive results of the trichothecene cluster identified in different analyses are described in 5.1.1.

Notably, the first two significant gene clusters with co-occurring motif seeds have the typical structure of functional organizations observed in fungal mycotoxin gene clusters. The most significant gene cluster (msfg_1) contains two non-ribosomal peptide synthetases (NPS8 and FGSG_11660 'related to NPS8') genes, one cytochrome P450 and one transporter (Figure 5-9, Table 5-12). A regulatory protein (FGSG_11654, #-1 in Table 5-12) is also positioned directly upstream of this cluster but its promoter region does not contain the same motif seeds. It is possible that the regulatory protein may have an independent binding site or other regulatory proteins which control gene transcription of cluster msfg_1. The second significant gene cluster (msfg_2) includes genes encoding one polyketide synthase (PKS11), one cytochrome P450, two secreted proteins, one transporter and one pathway-specific transcription factor. The PKS11 gene is known to be related to sexual development but it has

not been reported whether PKS11 is associated with other neighboring genes. Notably, the cluster msfg_2 is conserved in two other *Fusarium* species (*F. verticillioides* and *F. oxysporum*) (ID 3f_1 in Table 5-15). Functional relations of genes in the two gene clusters, msfg_1 and msfg_2, remain to be determined.

11 out of 51 predicted secondary metabolite genes (15 PKS, 20 NPS and 16 TPS) were arranged in 10 gene clusters with co-occurring motif seeds (Table 5-11). 6 out of the 11 secondary metabolite genes are functionally described (Table 3-4 and Table 3-5): TRI5 (trichothecene cluster), PKS11, PKS9, NPS1, NPS2 and NPS10. PKS11 of cluster msfg_2, mentioned before, and PKS9 of msfg_67 are reported to be preferentially expressed during sexual development [112]. NPS1 and NPS2 are known to be involved in iron metabolism [115]. NPS10 of msfg_47 plays a role in gliotoxin production in *A. fumigatus* [169] but it has not been reported if NPS10 also has the same role in *F. graminearum*. The other 5 SM genes of the 4 clusters (Table 5-11) have unknown functions.

ID	Chr.	(N. of gene), gene range	(N. of motif seed), motif seed	Adjusted P-value	N. of gene					Cluster of co-expressed neighboring genes	TFC and/or SPC ID
					SM gene	CYP	TF	TP	SP		
msfg_1	1	(17), FGSG_00036 ~ FGSG_00049	(8), TGGn{3}CAC, TGGn{2}CCA, GGTn{3}ACC, GTGn{4}CAC, GGTn{2}CAC, GTGn{2}ACC, GTGnCAC, CCAn{9}CAG	7.7e-08	2 (NPS8, NPS-FGSG_11660)	1		1			tfcfg_3
msfg_2	1	(14), FGSG_01786 ~ FGSG_01799	(6), CCGCGG, CGGn{9}GCG, CGGn{8}CGC, CGGn{7}CCG, TCCGCG, GCGn{8}CCG	1.0e-06	1 (PKS11)	1	1	1	2		tfcfg_9
msfg_4	2	(21), FGSG_11532 ~ FGSG_11551	(2), AGCn{3}ACT, TAAn{6}TTA	1.5e-04		1	2		6		tfcfg_49
msfg_5	1	(8), FGSG_01807 ~ FGSG_01814	(1), CGGn{8}CCG	2.18e-04				1		fg5_18	
msfg_12	1	(14), FGSG_01986 ~ FGSG_01997	(1), AGTn{2}ACT	4.6e-04				3	4		spfg_21
msfg_14	2	(15), FGSG_03543 ~ FGSG_03529	(1), AGGCCT	6.1e-04	1 (TPS-TRI5)	3	2	1	4	fg1_39	tfcfg_41, spfg_10
msfg_23	2	(19), FGSG_13352 ~ FGSG_08482	(1), CCAnGCT	1.4e-03					3		spfg_27
msfg_26	1	(12), FGSG_01761 ~ FGSG_01769	(1), CGAGAC	1.6e-03		1		1	4		tfcfg_8
msfg_35	2	(13), FGSG_03177 ~ FGSG_03166	(1), TTTn{3}TTA	2.6e-03				2	3		spfg_41
msfg_37	2	(10), FGSG_03068 ~ FGSG_03060	(1), CTTAGA	2.8e-03	1 (TPS-FGSG_03066)		1				tfcfg_80
msfg_38	3	(22), FGSG_05353 ~ FGSG_05372	(2), AAGn{9}CTT, CACn{3}GTG	3.0e-03	1 (NPS2)		2	1	3		tfcfg_51
msfg_47	4	(16), FGSG_06499 ~ FGSG_06512	(1), CGGn{7}TCT	4.0e-03	1 (NPS10)		1	2	1		tfcfg_68
msfg_54	3	(17), FGSG_11039 ~ FGSG_11026	(1), TAAn{4}CGA	4.4e-03	1 (NPS1)	1		3	5	fg5_114, fg5_115	tfcfg_55, spfg_12
msfg_67	1	(6), FGSG_10459 ~ FGSG_10464	(1), CGGn{3}CCG	5.8e-03	1 (PKS9)	1			1	fg5_36	tfcfg_20
msfg_69	3	(13), FGSG_04728 ~ FGSG_15123	(1), ATGn{3}AGC	5.9e-03					7		spfg_46
msfg_84	3	(18), FGSG_10968 ~ FGSG_10952	(1), AGTAAT	8.3e-03		1			5		spfg_57
msfg_89	3	(12), FGSG_11000 ~ FGSG_10989	(1), GTAn{6}ACT	9.0e-03	1 (NPS9)	1		1	3	fg1_54	tfcfg_56, spfg_13
msfg_92	1	(12), FGSG_00445 ~ FGSG_00456	(1), GTGnCAC	9.4e-03	1 (TPS-FGSG_00451)				2		tfcfg_5

Table 5-11: Gene clusters with co-occurring motif seeds with features of interest

19 clusters including the trichothecene mycotoxin cluster (msfg_14) were assigned to at least one TFC or SPC described in Table 5-7 and Table 5-8 or overlapped with clusters of co-expressed neighboring genes. Gene clusters in blue-shaded lines overlap with clusters of co-expressed neighboring genes. Dtails of all clusters of co-occurring motif seeds in *F. graminearum* are provided in Appendix A (Table S 8-1).

NPS: Non-ribosomal peptide synthetase, PKS: Polyketide synthase, Chr.: Chromosome, SM: Secondary metabolite, SP: Secreted protein, TF: Transcription factor, TP: Transporter, CYP: Cytochrome P450, TFC: tentative functional gene cluster, SPC: secreted proteins cluster

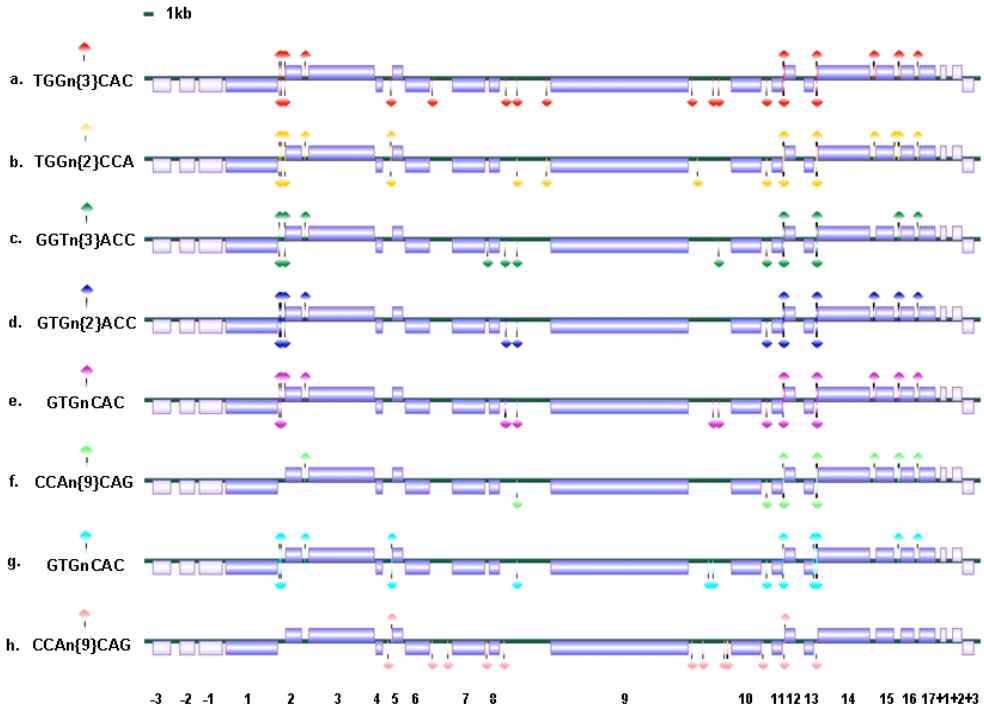


Figure 5-9: The most significant gene cluster with co-occurring motif seeds (msfg_1)

The most significant gene cluster with co-occurring motif seeds (msfg_1) contains 17 genes. 8 significant motif seeds were identified. Gene coding regions are represented by boxes. Diamond shapes indicate motif seeds. Genes and motif seeds in the forward strand are displayed above the line; genes and motif seeds in the reverse strand are displayed below. Numbers underneath each box indicate gene designations (see Table 5-12).

Gene order	FGDB ID	Gene	Gene feature	Gene description
-3	FGSG_00034		TP	related to alpha-glucoside transport protein
-2	FGSG_11653			probable sulfatase
-1	FGSG_11654		TF	related to nitrate assimilation regulatory protein
1	FGSG_00036			probable fatty acid synthase, alpha subunit
2	FGSG_11655		CYP	related to cytochrom P450
3	FGSG_11656			related to FAS1 – fatty-acyl-CoA synthase, beta chain
4	FGSG_00038			hypothetical protein
5	FGSG_00039			conserved hypothetical protein
6	FGSG_00040			conserved hypothetical protein
7	FGSG_11657			conserved hypothetical protein
8	FGSG_11658			hypothetical protein
9	FGSG_11659	NPS8	NPS	non-ribosomal peptide synthetase
10	FGSG_11660		NPS	NPS8 related to non-ribosomal peptide synthetase
11	FGSG_00043			conserved hypothetical protein
12	FGSG_00044			conserved hypothetical protein
13	FGSG_00045			conserved hypothetical protein
14	FGSG_00046		TP	related to multidrug resistance protein
15	FGSG_00047			conserved hypothetical protein
16	FGSG_00048			related to flavonol synthase-like protein
17	FGSG_00049			related to branched-chain amino acid aminotransferase
+1	FGSG_11661			conserved hypothetical protein
+2	FGSG_00050			conserved hypothetical protein
+3	FGSG_00051			related to aliphatic nitrilase

Table 5-12: Gene functions and features of the most significant gene cluster with co-occuring motif seeds (msfg_1)

The most significant gene cluster (msfg_1) contains genes encoding two non-ribosomal peptide synthetase (NPS), one cytochrome P450 and one transporter. Functions of two NPS genes have not been reported.

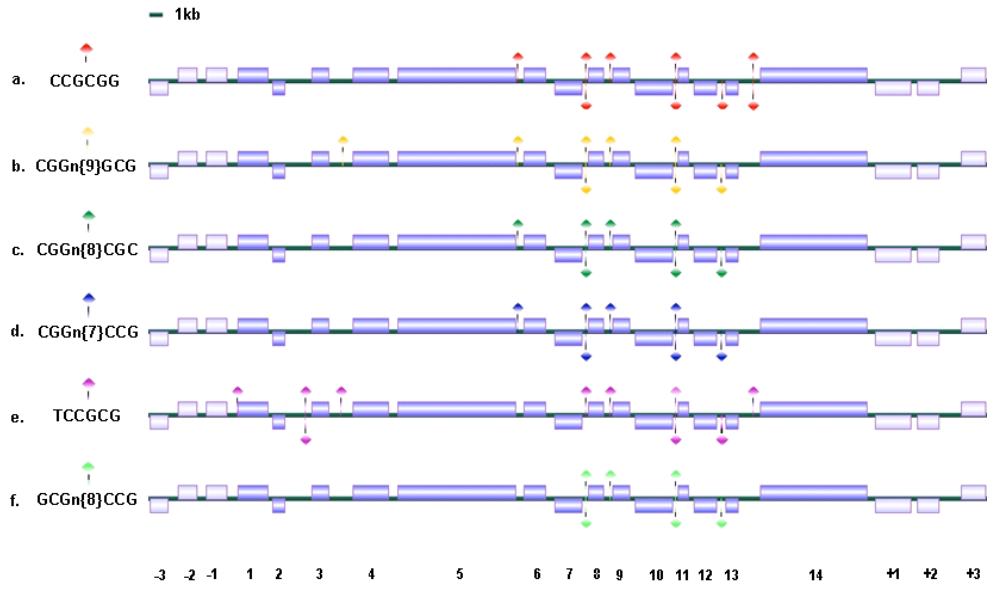


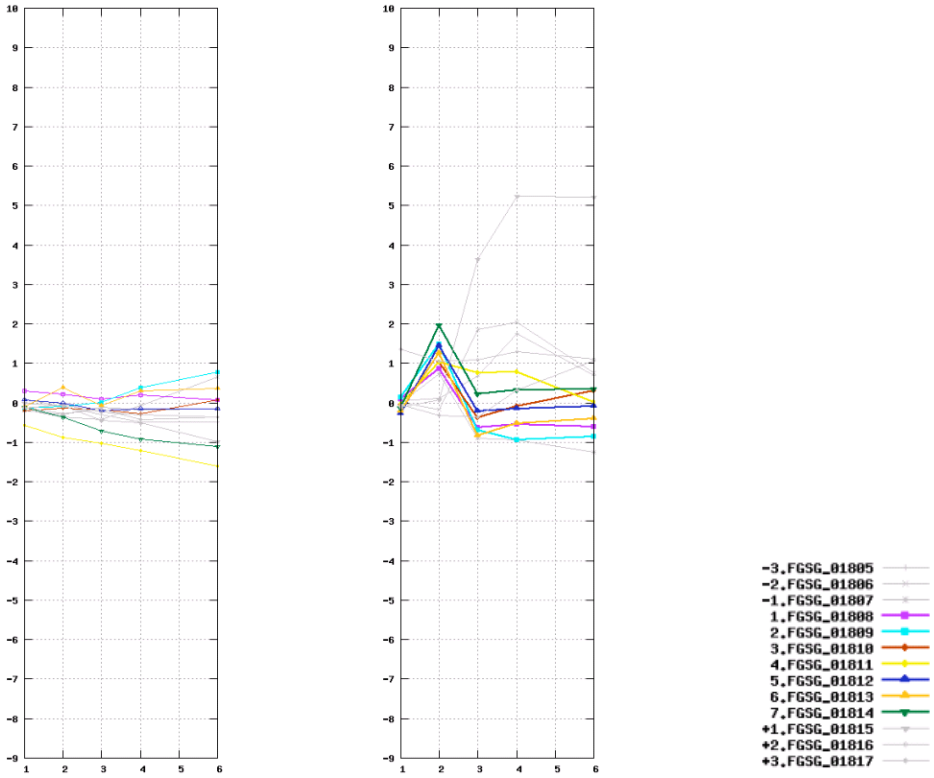
Figure 5-10: The second significant gene cluster with co-occurring motif seeds (msfg_2)

The second significant gene cluster with co-occurring motif seeds (msfg_1) contains 14 genes. 6 significant motif seeds were identified. Gene coding regions are represented by boxes. Diamond shapes indicate motif seeds. Genes and motif seeds in the forward strand are displayed above the line; genes and motif seeds in the reverse strand are displayed below. Numbers underneath each box indicate gene designations (see Table 5-13).

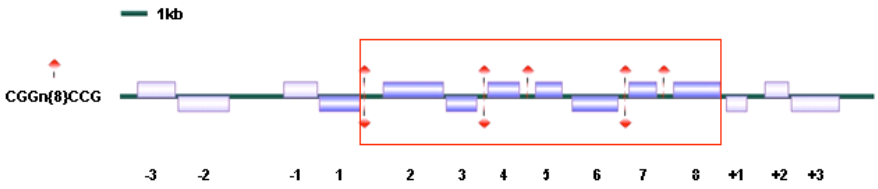
Gene order	FGDB ID	Gene	Gene feature	Gene description
-3	FGSG_01783		TPS	related to aristolochene synthase
-2	FGSG_01784			related to phosphatidylinositol/phosphatidylcholine transfer protein
-1	FGSG_01785			related to C4-dicarboxylate transport protein mae1
1	FGSG_01786		CYP	related to cytochrome P450 monooxygenase
2	FGSG_01787			conserved hypothetical protein
3	FGSG_01788			conserved hypothetical protein
4	FGSG_01789		SP	related to ferric reductase Fre2p
5	FGSG_01790	PKS11	SM (PKS)	polyketide synthase
6	FGSG_01791			conserved hypothetical protein
7	FGSG_01792		TP	related to DAL5 – allantoate and ureidosuccinate permease
8	FGSG_01793			conserved hypothetical protein
9	FGSG_01794			related to short-chain alcohol dehydrogenase
10	FGSG_01795		FSTF	related to pathway-specific regulatory protein
11	FGSG_01796			conserved hypothetical protein
12	FGSG_01797			related to muconate cycloisomerase I
13	FGSG_01798			conserved hypothetical protein
14	FGSG_01799		SP	conserved hypothetical protein
+1	FGSG_01800			conserved hypothetical protein
+2	FGSG_01801			related to glutaminase, kidney isoform, mitochondrial precursor
+3	FGSG_01802			related to vacuolar Ca2+/H+ antiporter

Table 5-13: Gene functions and features of the second significant gene cluster with co-occurring motif seeds (msfg_2)

The second significant gene cluster with co-occurring motif seeds (msfg_1) includes genes encoding one secondary metabolite, one cytochrome P450, one fungal-specific transcription factor, and one transporter. Products and related functions of this cluster have not been reported at present.



A. A co-expressed gene cluster during sexual development (fg5_18, right) and profiles of the corresponding genes in planta (left)



B. Gene clusters with co-occurring regulatory motifs (msfg_5)

Figure 5-11: Gene cluster with co-occurring motif seeds (msfg_5) overlapping with a cluster of co-expressed neighboring genes during sexual development (fg5_18)

7 neighboring genes on chromosome 1 were identified as a cluster of co-expressed neighboring genes (fg5_18, A). The cluster fg5_18 overlap with the the fifth significant gene cluster with co-occurring motif seeds (msfg_5, B).Gene coding regions are represented by boxes. Diamond shapes indicate motif seeds. Genes and motif seeds in the forward strand are displayed above the line; genes and motif seeds in the reverse strand are displayed below. Numbers underneath each box indicate gene positions (see Table 5-14).

Gene order	FGDB ID	Gene	Gene feature	Gene description
-3	FGSG_15001			related to cytosine deaminase
-2	FGSG_01805		SP	related to isoamyl alcohol oxidase
-1	FGSG_01806			conserved hypothetical protein
1	FGSG_01807			conserved hypothetical protein
2	FGSG_01808			conserved hypothetical protein
3	FGSG_01809			related to aryl-alcohol dehydrogenases
4	FGSG_01810			conserved hypothetical protein
5	FGSG_01811			conserved hypothetical protein
6	FGSG_01812			probable CYB2 - lactate dehydrogenase cytochrome b2
7	FGSG_01813			related to dihydrodipicolinate synthetase
8	FGSG_01814		TP	related to putative tartrate transporter
+1	FGSG_01815		SP	conserved hypothetical protein
+2	FGSG_01816			related to theta class glutathione S-transferase
+3	FGSG_01817		TP	conserved hypothetical protein

Table 5-14: Gene functions and features of the fifth significant gene cluster with co-occurring motif seeds (msfg_5)

The fifth significant gene cluster contains no gene encoding key enzymes (e.g. secondary metabolite and cytochrome P450) which are known to be involved in secondary metabolism. Genes in the cluster msfg_5 were significantly co-expressed during sexual development (fg5_18). Products and related functions of this cluster have not been reported at present. SP: Secreted protein, TP: Transporter

5.6 Synteny in 3 *Fusarium* species

To predict candidate genomic regions which may contain genes involved in common pathways, syntenic blocks were first screened for in 3 *Fusarium* genomes. *F. graminearum* was used as reference genome and was compared with 2 other *Fusarium* species (*F. verticillioides* and *F. oxysporum*). Orthologous genes were identified by means of reciprocal best hit (RBH) using the Similarity Matrix of Proteins (SIMAP) [133]. About 77 (10,709) and 66 (9,230) percent of 13,935 genes predicted in *F. graminearum* had orthologous genes in at least one and both *Fusarium* species, respectively. The numbers of orthologous genes between *F. graminearum* and the two *Fusarium* species are as follows; 9,901 (71%) with *F. verticillioides* and 10,038 (72%) with *F. oxysporum*. From the genomic maps of orthologous genes based on the *F. graminearum* genome, syntenic blocks (SBs) were additionally examined to compare their regulatory motifs to predict conserved gene clusters in 3 *Fusarium* species.

5.6.1 Conserved gene clusters in 3 *Fusarium* species

Synteny blocks (SBs) can be one source of information useful for predicting conserved fungal gene clusters. However, it is difficult to distinguish from diverse mechanisms driving the conservation of synteny. One possibility to refine conserved fungal gene clusters from synteny blocks is suggested by an observation of promoter motifs of orthologous genes as shown in aflatoxin-producing *Aspergillus* species. Comparative genomic study of aflatoxin gene clusters in *Aspergillus* species demonstrated that gene order of the clusters and binding sites of a pathway-specific transcription factor were well conserved for more than 25 million years of divergence experienced by the *Aspergillus* species [170].

For the construction of conserved gene cluster in 3 *Fusarium* species, the set of SBs having the same motif seeds among orthologous genes were additionally examined. Significant motif seeds of genes from each of synteny blocks of 3 *Fusarium* species were calculated by the same procedures as described in 3.3.1. SBs containing motif seeds with P-values < 0.01 in orthologous genes of all three *Fusarium* species were defined as significantly conserved gene clusters. Using this strategy, 25 conserved gene clusters within 3 *Fusarium* species were identified and ranked according to the p-values of the motif seeds of *F. graminearum* (Table 5-15). Remarkably, three conserved gene clusters (3f_1, 3f_2, and 3f_11 in Table 5-15) overlap with gene clusters identified by analyses of conserved motif seeds and co-expression in neighboring genes in *F. graminearum*. Moreover, these three conserved gene clusters have a common functional feature related to sexual development. The most significantly conserved gene cluster (3f_1) is positioned adjacent to a PKS11 gene which is reported to

be related to sexual development [112]. Two other conserved gene clusters (3f_2, and 3f_11) were found to be co-expressed neighboring genes during sexual development. The 25 conserved gene clusters found containing orthologous genes and their corresponding regulatory elements may reflect a conserved feature of co-regulation across evolution. To determine if physically conserved gene clusters are functionally conserved, futher study of conserved expression patterns can be a strong criterion.

ID	Chr.	(N. of gene), gene range	(N. of motif seed), motif seed	P-value of motif seed			Gene cluster ID*	N. of gene			
				Fg	Fv	Fo		SM	SP	TF	TP
3f_1	1	(7), FGSG_01791 ~ FGSG_01797	(5), CCGCGG, CGGn{8}CGC, CGGn{9}GCG, CGGn{7}CCG, GCGn{8}CCG	5.82E-09	5.93E-08	7.78E-08	msfg_2	PKS11		1	1
3f_2	1	(6), FGSG_01809 ~ FGSG_01814	(1), CGGn{8}CCG	6.26E-07	2.52E-06	2.04E-06	msfg_5, fg5_18		1		1
3f_3	2	(5), FGSG_03732 ~ FGSG_03728	(6), CGGn{2}CGG, CGGnTCG, GGTnCGG, GTTCGG, GGTTCG, TCGn{2}TCG	2.52E-06	9.05E-06	9.78E-06					1
3f_4	1	(7), FGSG_02761 ~ FGSG_12198	(2), GGAn{9}TCC, GCTn{4}AGC	2.33E-05	5.88E-05	5.44E-05					1
3f_5	2	(5), FGSG_08944 ~ FGSG_08940	(1), ATCnCGG	3.23E-05	1.07E-04	8.14E-05					
3f_6	3	(8), FGSG_05269 ~ FGSG_05276	(1),CACGTG	7.25E-05	1.03E-04	8.75E-05					
3f_7	1	(6), FGSG_01325 ~ FGSG_01330	(1), TAGGTA	9.43E-05	1.44E-04	2.58E-05					
3f_8	1	(6), FGSG_13728 ~ FGSG_10276	(1), AGAn{6}CGC	1.05E-04	2.86E-04	3.21E-04			1		
3f_9	4	(8), FGSG_09718 ~ FGSG_09725	(2), AACn{8}GTT, ATGCAT	1.31E-04	1.37E-04	1.48E-04				1	
3f_10	1	(5), FGSG_02477 ~ FGSG_02481	(1), CCAn{9}TGG	1.31E-04	1.99E-04	2.32E-04					
3f_11	2	(6), FGSG_03649 ~ FGSG_03644	(1), CGGnAAA	1.59E-04	3.21E-04	3.61E-04	fg5_76			1	1
3f_12	3	(7), FGSG_12839 ~ FGSG_05906	(1), GCCn{8}GGG	1.75E-04	3.91E-04	3.83E-04			1		
3f_13	2	(6), FGSG_08477 ~ FGSG_08473	(1), CTCnGGC	1.98E-04	7.43E-04	5.64E-04			1		
3f_14	1	(5), FGSG_10162 ~ FGSG_10166	(1), CTTn{5}TCT	2.10E-04	3.56E-04	2.69E-04			1	3	
3f_15	1	(6), FGSG_01227 ~ FGSG_01232	(1),TCAn{5}ACC	2.16E-04	3.65E-04	2.96E-04			2		
3f_16	2	(6), FGSG_08689 ~ FGSG_08684	(1),CTGGGG	2.32E-04	5.29E-04	6.49E-04					
3f_17	3	(5), FGSG_06367 ~ FGSG_06371	(1), TCCnACT	2.66E-04	3.22E-04	2.59E-04					1
3f_18	1	(7), FGSG_02525 ~ FGSG_02530	(1), CCAn{5}GAT	2.75E-04	4.11E-04	3.57E-04			3	1	
3f_19	1	(8), FGSG_01925 ~ FGSG_01931	(1), GTCn{3}GAC	3.10E-04	4.87E-04	3.25E-04			1		1
3f_20	1	(6), FGSG_01349 ~ FGSG_01354	(2), TGCn{3}TTC, GTTn{2}CTT	3.59E-04	5.59E-04	5.54E-04				1	1
3f_21	1	(5), FGSG_10074 ~ FGSG_10078	(1), CCCnCCA	4.52E-04	8.42E-04	6.73E-04			1		
3f_22	2	(6), FGSG_04108 ~ FGSG_04103	(1), TTTCGC	5.79E-04	9.57E-04	9.74E-04					
3f_23	4	(5), FGSG_07418 ~ FGSG_07422	(1), TTCCCC	5.86E-04	8.85E-04	8.39E-04			1		
3f_24	1	(5), FGSG_02677 ~ FGSG_02681	(1), ATGCAT	6.90E-04	6.70E-04	6.87E-04			1	1	
3f_25	4	(5), FGSG_07948 ~ FGSG_07952	(1), TTTn{2}AAA	7.68E-04	6.27E-04	8.74E-04			1	1	

Table 5-15: Conserved gene clusters of 3 *Fusarium* species

By comparison of motifs seeds of genes in synteny blocks of 3 *Fusarium* species, 25 conserved gene clusters were identified. The first two significantly conserved gene clusters overlap with gene clusters with co-occurring motif seeds identified in *F. graminearum* as described in 5.5. Two conserved gene cluster (3f_2 and 3f_11) are identified in clusters of co-expressed neighboring genes during sexual development. Gene cluster ID*: ID of of gene cluster identified in *F. graminearum* by motif seeds and/or co-expression analysis, Chr.: Chromosome, SM: Secondary metabolite, CYP: Cytochrome P450, SP: Secreted protein, TF: Transcription factor, TP: Transporter, Fg: *F. graminearum*, Fv: *F. verticillioides*, Fo: *F. oxysporum*

5.7 6 novel gene clusters of *F. graminearum*

Each of the gene clusters with even a single property can be a biologically relevant gene cluster. Due to lack of biological information, gene clusters with multiple properties were primarily screened and re-examined for their characteristics to select the most potential novel gene clusters. On the basis of multiple evidences, 6 gene clusters were selected as novel gene clusters that have common properties with mycotoxin gene clusters in *F. graminearum*. Four gene clusters (ID 1~4) showed parallel expression profiles to two mycotoxin gene clusters, the trichothecene and the butenolide cluster, in planta. The two remaining gene clusters (ID 5~6) were clustered in the same COP (COP5.16, Table 5-10) with the aurofusarin mycotoxin cluster during sexual development. Detailed evidences of each of the novel gene clusters are described below.

Putative function	ID	(N. of gene), gene range	Evidences					
			1. Co-expression	2. Particular compostion of gene functions				3. Promoter motif
				SM gene	CYP	SP	TF	TP	
	1	(10), FGSG_11307 ~ FGSG_11298	fg1_58		1	5			TAGn{2}TGC
Plant Infection	2	(10), FGSG_04596 ~ FGSG_04588	fg1_33	TPS (FGSG_04591), PKS15	1	1			CGTn{3}ACG
(5.7.1)	3	(13), FGSG_11000 ~ FGSG_13878	fg1_54	NPS9, NPS5	1	1		1	GTAn{6}ACT (msfg_89)
	4	(5), FGSG_11399 ~ FGSG_11395	fg1_57	NPS14		1			AGTn{5}CCG
Fungal Development	5	(9), FGSG_02111 ~ FGSG_02119	fg5_22		5	5			CCTn{9}GGG
(5.7.2)	6	(8), FGSG_10608 ~ FGSG_10614	fg5_43		2	1			TATnGCC

Table 5-16: 6 novel gene clusters and their evidences

6 gene clusters were selected as novel gene clusters based on multiple evidences. Novel gene clusters are basically co-expressed profiles with mycotoxin gene clusters under the same conditions. Besides, their compositions of gene functions contain known features observed in other fungal gene clusters.

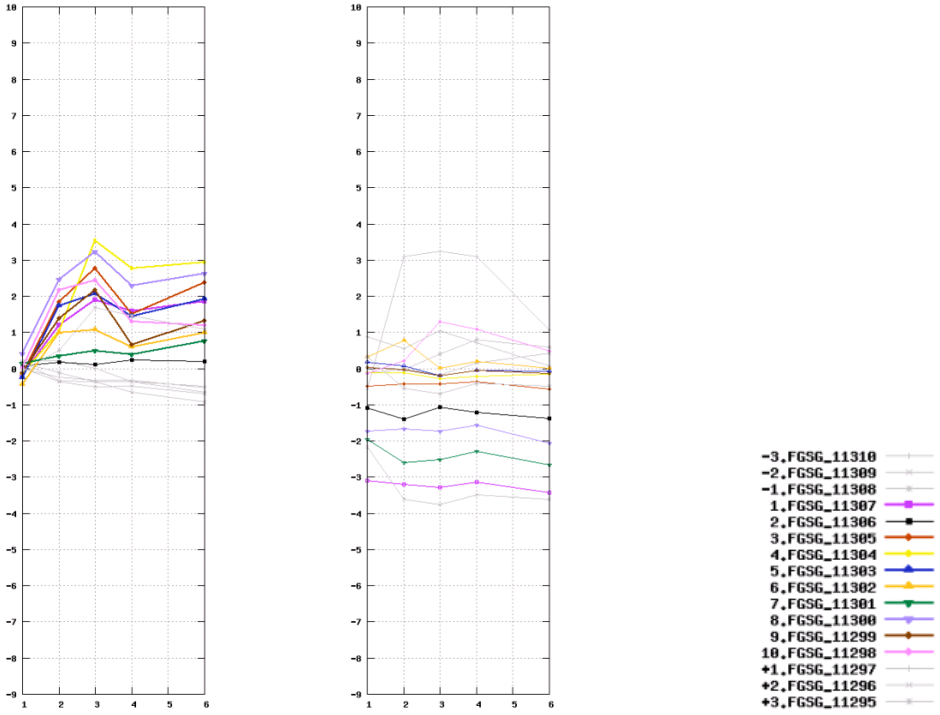
SM: secondary metabolite, TPS: Terpenoid synthase, CYP: cytochrome P450, SP: secreted protein, TF: transcription factor, TP: transporter

5.7.1 4 novel gene clusters involved plant infection

Two mycotoxin clusters, the trichothecene and the butenolide cluster, significantly co-expressed in planta (FG1) showed similar expression patterns with four clusters of unknown function (Figure 5-8 and Table 5-10). The clusters have a highly up-regulated peak at the time point of the 3rd day while the remaining 70 clusters showed continuously increasing expression patterns. The four gene clusters also have interesting features examined in real gene clusters in fungi. No noticeable characteristic have been found in the other 70 clusters at present. These observations can be good indicators that the four gene clusters may be novel clusters possibly related to plant infection.

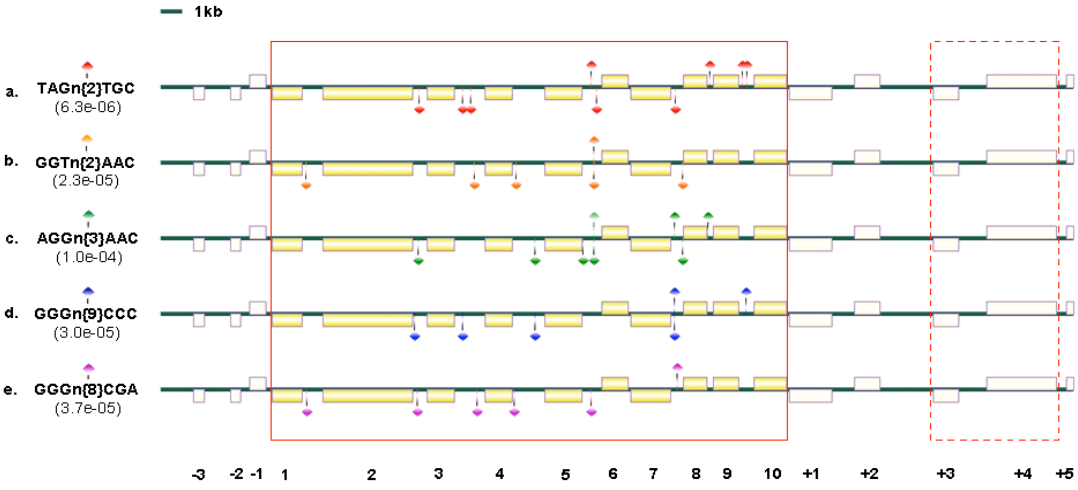
5.7.1.1 Novel gene cluster 1: fg1_58, enrichment of secreted proteins and cytochrome P450 genes

Novel gene cluster 1 co-expressed in planta (cluster fg1_58) is characterized as a group enriched in genes coding for secreted proteins. It also contains one cytochrome P450 gene (FGSG_11303) and one gene involved in lipid/fatty acid transport (FunCat 20.01.13, FGSG_11305). Interestingly, two genes with features of interest were observed downstream of the novel gene cluster 1, encoding a non-ribosomal peptide synthetase (NPS12, FGSG_11294) and an enoyl-CoA hydratase precursor (FGSG_11295) predicted to be involved in lipid and fatty metabolism. The two genes were not identified in cluster fg1_58 because their expression patterns are different at the 6th day (Figure 5-12). However, the transcripts of the two genes were increased until day 3 as in cluster fg1_58. The novel gene cluster 1 and the two adjacent genes may be involved in different functions but also have a possibility that they are functionally associated because the trichothecene and the aurofusarin clusters show similar functional organization; they both have SM genes and enrichment of genes coding for secreted proteins.



A1. A co-expressed gene cluster in planta (fg1_58, left) and profiles of the corresponding genes during sexual development (right)

A2. Expression profiles of NPS12 and flanking gene



B. Gene clusters with co-occurring regulatory motifs

Figure 5-12: Expression profiles and conserved motif seeds of novel gene cluster 1 (fg1_58)

A1. Novel gene cluster 1 containing 10 genes is clustered in the same COP (COP1.1, Figure 5-8) with two mycotoxin gene clusters, the trichothecene and the butenolide cluster, during growth in planta. 10 genes co-expressed in planta did not show co-expressed profiles during sexual development.

A2. Two genes (NPS12 and enoyl-CoA hydratase precursor) with features of interest were observed downstream of the novel gene cluster 1. The transcripts of the two genes were increased until day 3 as in cluster fg1_58. However, the two genes are not selected as part of the co-expressed gene cluster fg1_58 because expression profiles from 4^{th} to 6^{th} day are different.

B. Five motif seeds examined based on a cluster of co-expressed neighboring genes (fg1_58, in the red box) are suggested as putative regulatory motifs of the novel gene cluster 1. Gene coding regions are represented by boxes. Diamond shapes indicate motif seeds. Number in parenthesis is a p-value of a motif seeds. Genes and motif seeds in the forward strand are displayed above the line; genes and motif seeds in the reverse strand are displayed below. Numbers underneath each box indicate gene designations (see Table 5-17).

Gene order	FGDB ID	Gene	Gene feature	Gene description
-3	FGSG_11310			putative protein [EST hit]
-2	FGSG_11309			conserved hypothetical protein
-1	FGSG_11308			hypothetical protein
1	FGSG_11307			conserved hypothetical protein
2	FGSG_11306		SP	conserved hypothetical protein
3	FGSG_11305			related to CSR1 – phosphatidylinositol transfer protein
4	FGSG_11304		SP	related to endo-1,4-beta-xylanase
5	FGSG_11303		CYP, SP	related to isotrichodermin C-15 hydroxylase (cytochrome P-450 monooxygenase CYP65A1)
6	FGSG_11302		SP	conserved hypothetical protein
7	FGSG_11301			conserved hypothetical protein
8	FGSG_11300			related to a retinal short-chain dehydrogenase/reductase
9	FGSG_11299		SP	related to acetyl-hydrolase
10	FGSG_11298			related to monooxygenase
+1	FGSG_11297			related to beta transducin-like protein
+2	FGSG_11296		SP	conserved hypothetical protein
+3	FGSG_11295			related to enoyl-CoA hydratase precursor, mitochondrial
+4	FGSG_11294	NPS12	SM (NPS)	non-ribosomal peptide synthetase
+5	FGSG_15643			hypothetical protein

Table 5-17 Gene functions and features of novel gene cluster 1 (fg1_58)

Novel gene cluster 1 has a typical composition of gene functions observed in fungal gene clusters. The novel gene cluster 1 is characterized as a group enriched in genes coding for secreted proteins. It also contains one cytochrome P450 gene (FGSG_11303) and one gene involved in lipid/fatty acid transport (FunCat 20.01.13, FGSG_11305). Two genes (NPS12 (#+4) and enoyl-CoA hydratase precursor (#+3)) with features of interest were observed in downstream of the novel gene cluster 1 but their functional relation is unknown.

SM: Secondary metabolite, NPS: Non-ribosomal peptide synthetase, SP: Secreted protein, CYP: Cytochrome P450

5.7.1.2 Novel cluster 2: fg1_33, PKS15, Terpenoid (FGSG_04591), cytochrome P450

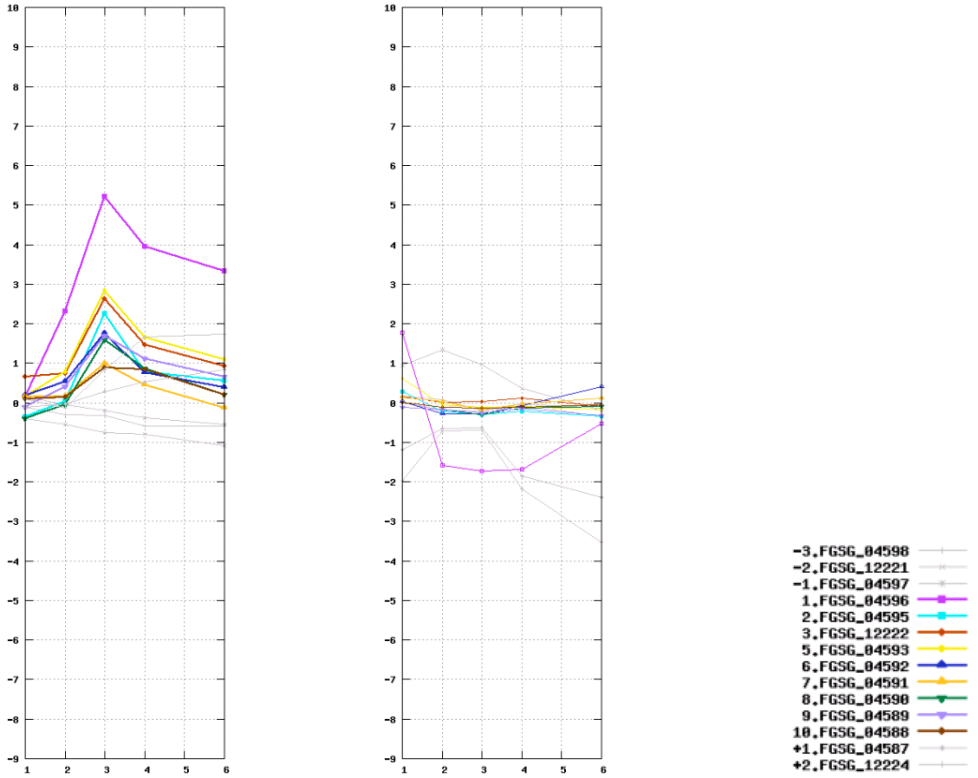
Novel gene cluster 2 contains 10 co-expressed genes during growth in planta. These genes encode two secondary metabolite(SM) genes (PKS15 and terpenoid synthase), one cytochrome P450 and one secreted protein (Table 5-18 and Figure 5-13). In particular, PKS15 has been regarded as one of the strong candidates producing a metabolite of unknown function because PKS15 is reported to be expressed exclusively during plant infection [112]. No transcription factor was found in this cluster. Other regulatory proteins may control transcription of these genes. The novel gene cluster 2 provides genes which are involved in a common pathway or do have a co-function with PKS15, which reinforces motivation for advanced analysis on this cluster.

Gene order	FGDB ID	Gene	Gene feature	Gene description
-3	FGSG_04598			related to CLC chloride channel protein
-2	FGSG_12221			hypothetical protein
-1	FGSG_04597		SP	conserved hypothetical protein
1	FGSG_04596			related to O-methyltransferase
2	FGSG_04595			related to hydroxylase
3	FGSG_12222			related to 3-ketoacyl-acyl carrier protein reductase
4	FGSG_12223			related to 3-ketoacyl-acyl carrier protein reductase
5	FGSG_04593			related to para-hydroxybenzoate polyprenyltransferase precursor
6	FGSG_04592		SP	related to light induced alcohol dehydrogenase Bli-4
7	FGSG_04591		SM (TPS)	probable farnesyltranstransferase (al-3)
8	FGSG_04590		CYP	related to isotrichodermin C-15 hydroxylase (cytochrome P-450 monooxygenase CYP65A1)
9	FGSG_04589			related to tetracenomycin polyketide synthesis O-methyltransferase tcmP
10	FGSG_04588	PKS15	SM (PKS)	polyketide synthase
+1	FGSG_04587			related to WD40-repeat protein (notchless protein)
+2	FGSG_12224			hypothetical protein
+3	FGSG_12225			related to microbial serine proteinase

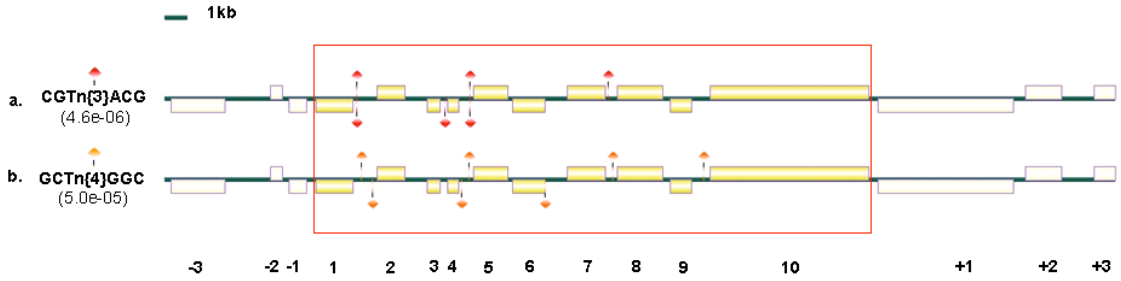
Table 5-18: Gene functions and features of novel gene cluster 2 (fg1_33)

Novel gene cluster 2 comprises 10 genes including genes encoding 2 secondary metabolites (PKS15 (#10) and Terpenoid synthase (#7)), one cytochrome P450, and one secreted protein. The PKS15 has been considered as one of strong candidates producing a metabolite of unknown function because it is expressed exclusively during plant infection. Functional relations among co-expressed genes have not been reported.

PKS: Polyketide synthase, TPS: Terpenoid synthase, SM: Secondary metabolite, SP: Secreted protein, CYP: Cytochrome P450



A. A co-expressed gene cluster in planta (fg1_33, left) and profiles of the corresponding genes during sexual development (right)



B. Gene clusters with co-occurring regulatory motifs

Figure 5-13: Expression profiles and conserved motif seeds of novel gene cluster 2 (fg1_33)

10 genes co-expressed during growth in planta are selected to belong to novel gene cluster 2. The clustered 10 genes did not show co-expressed profiles during sexual development. B. Two motif seeds examined based on a cluster of co-expressed neighboring genes (fg1_33, in the red box) are suggested as putative regulatory motifs of the novel gene cluster 2. Gene coding regions are represented by boxes. Diamond shapes indicate motif seeds. Number in parenthesis is a p-value of a motif seeds. Genes and motif seeds in the forward strand are displayed above the line; genes and motif seeds in the reverse strand are displayed below. Numbers underneath each box indicate gene designations (see Table 5-18).

5.7.1.3 Novel cluster 3: fg1_54, NPS9, NPS5, cytochrome P450, and transport

Novel gene cluster 3 consists of 13 genes co-expressed in planta. It contains genes encoding 2 nonribosomal peptide synthetases (NPS9 and NPS5), one cytochrome P450, and one transporter. NPS9 is related to the synthetase of AM-toxin, a host-specific phytotoxin of the *Alternaria alternata* apple pathotype which induces blotch and necrosis on apple leaves [171,172]. It has not been reported whether NPS9 has the same role in *F. graminearum*. The other SM gene, NPS5, has not been functionally described. The transporter gene in this cluster is annotated to be related to the multidrug resistance protein and may have roles mediating cellular resistance to toxins. Considering the gene functions in novel gene cluster 3, the cluster probably mediates interactions between fungi and hosts.

Gene order	FGDB ID	Gene	Gene feature	Gene description
-3	FGSG_11003			related to TOB3 (member of AAA-ATPase family)
-2	FGSG_11002			probable cytochrome P450 monooxygenase (lovA)
-1	FGSG_11001			related to GUK1 – guanylate kinase
1	FGSG_11000			conserved hypothetical protein
2	FGSG_10999	xylA	SP	endo-1,4-beta-xylanase
3	FGSG_10998		SP	related to 6-hydroxy-D-nicotine oxidase
4	FGSG_10997			conserved hypothetical protein
5	FGSG_10996			conserved hypothetical protein
6	FGSG_10995		TP	related to multidrug resistance protein
7	FGSG_10994			conserved hypothetical protein
8	FGSG_10993			related to selenocysteine lyase
9	FGSG_10992			related to polysaccharide deacetylase
10	FGSG_10991		CYP, SP	related to benzoate 4-monooxygenase cytochrome P450
11	FGSG_10990	NPS9	SM (NPS)	related to AM-toxin synthetase (AMT)
12	FGSG_10989			conserved hypothetical protein
13	FGSG_13878	NPS5	SM (NPS)	related to non-ribosomal peptide synthetase
+1	FGSG_13879			probable tyrocidine synthetase
+2	FGSG_10987			hypothetical protein
+3	FGSG_10986		SP	related to alcohol oxidase

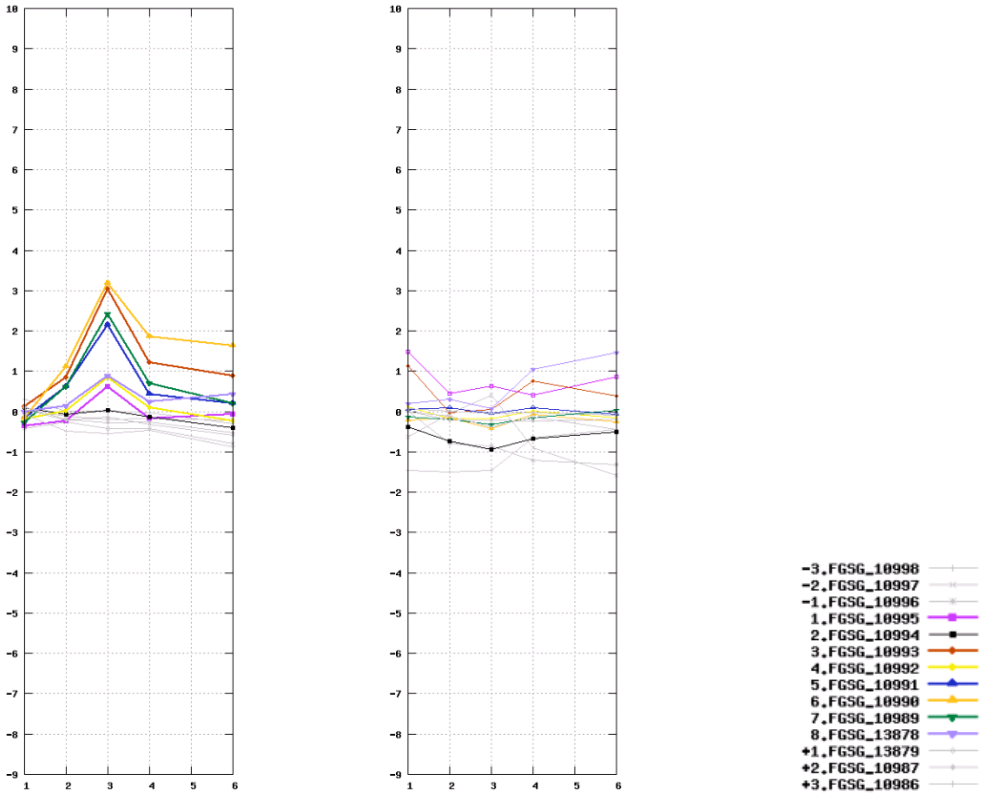
Table 5-19: Gene functions and features of novel gene cluster 3 (fg1_54)

Novel gene cluster 3 comprises 13 genes including 2 nonribosomal peptide synthetase (NPS9 and NPS5), one cytochrome P450 gene, and one transporter gene. NPS9 is related to AM-toxin synthetase but it has not been reported whether NPS9 has the same role in *F. graminearum*. Functional relations of genes in the novel gene cluster 3 remain to be determined. NPS: Non-ribosomal peptide synthetase, SM: Secondary metabolite, SP: Secreted protein, TP: Transporter, CYP: Cytochrome P450

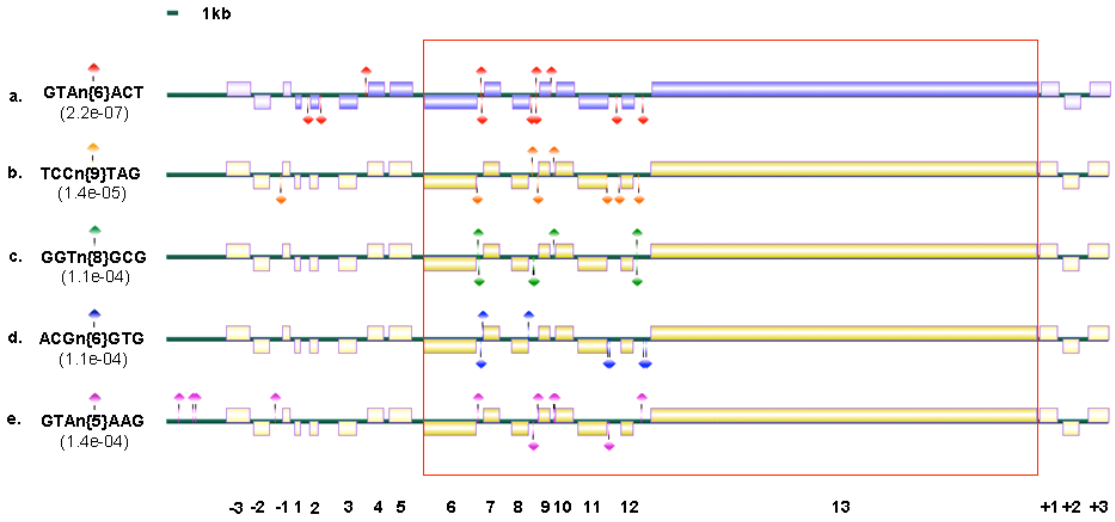
Figure 5-14: Expression profiles and conserved motif seeds of novel gene cluster 3 (fg1_54)

13 genes co-expressed during growth in planta are selected to belong to novel gene cluster 3. The clustered 13 genes did not show co-expressed profiles during sexual development. B. Co-expressed gene cluster fg1_54 overlaps with gene cluster with co-occurring gene cluster (msfg_89). Four additional motif seeds examined based on co-expressed neighboring genes (fg1_33) are also suggested as putative regulatory motifs of novel gene cluster 3. Gene coding regions are represented by boxes. Gene maps filled with blue and yellow colors refer to a gene cluster with co-occurring promoter motif (a, msfg_89) and based on a cluster of co-expressed neighboring genes (b~e, fg1_54), respectively. Diamond shapes indicate motif seeds. The number in

parenthesis is a p-value of each motif seed. Genes and motif seeds in the forward strand are displayed above the line; genes and motif seeds in the reverse strand are displayed below. Numbers underneath each box indicate gene designations (see Table 5-19).



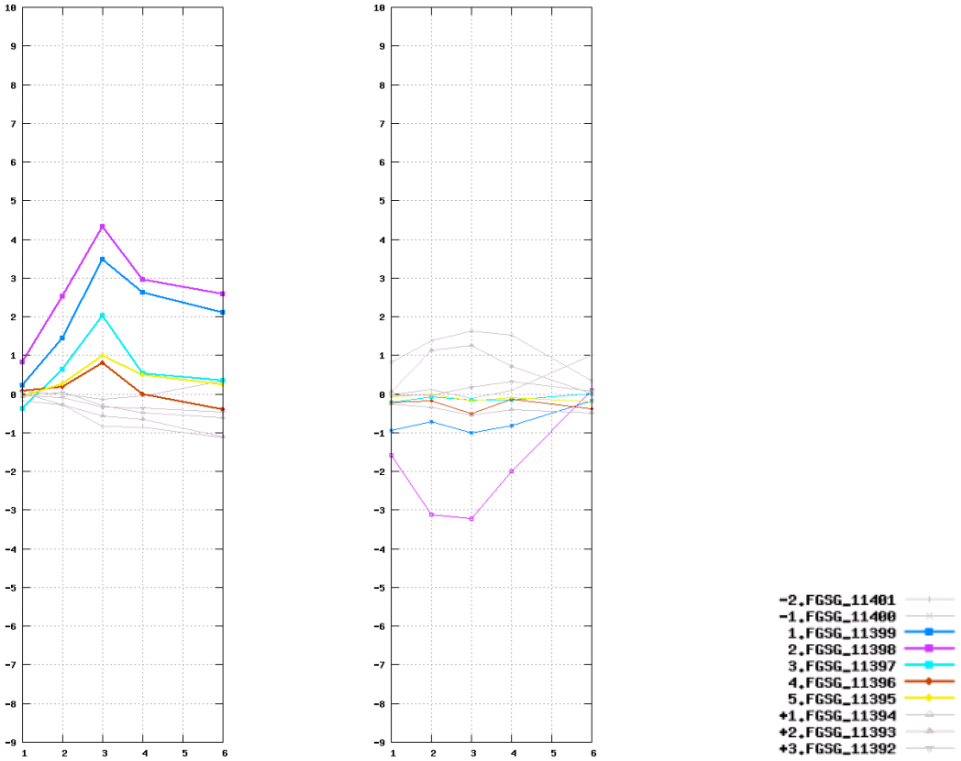
A. A co-expressed gene cluster in planta (fg1_54, left) and profiles of the corresponding genes during sexual development (right)



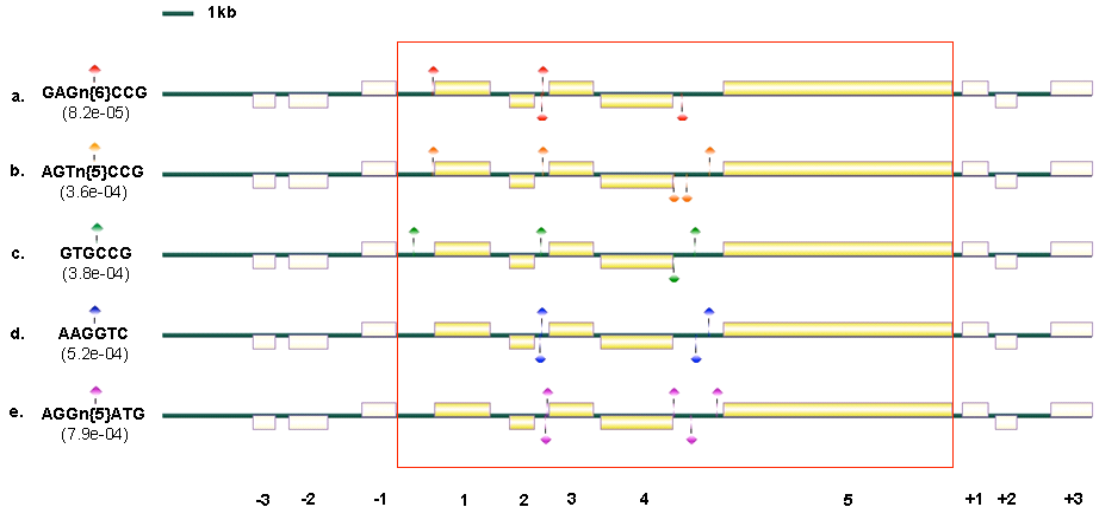
B. Gene clusters with co-occurring regulatory motifs

5.7.1.4 Novel cluster 4: fg1_57, NPS14

Novel gene cluster 4 contains 5 genes co-expressed during growth in planta. These genes encode one nonribosomal peptide synthetase (NPS14) and one secreted protein. The NPS14 is reported to be related to the synthetase of AM-toxin but functional studies have not reported. Neither transcription factors nor transporters have been found in novel cluster 4.



A. A co-expressed gene cluster in planta (fg1_57, left) and profiles of the corresponding genes during sexual development (right)



B. Gene clusters with co-occurring regulatory motifs

Figure 5-15: Expression profiles and conserved motif seeds of novel gene cluster 4 (fg1_57)

5 genes co-expressed during growth in planta are selected to belong to novel gene cluster 4. The clustered 5 genes did not show co-expressed profiles during sexual development. B. Five motif seeds examined based on a cluster of co-expressed neighboring genes (fg1_57, in the red box) are suggested as putative regulatory motifs of the novel gene cluster 4. Gene coding regions are represented by boxes. Diamond shapes indicate motif seeds. The number in parenthesis is a p-value of each motif seed. Genes and motif seeds in the forward strand are displayed above the line; genes and motif seeds in the reverse strand are displayed below. Numbers underneath each box indicate gene designations (see Table 5-20).

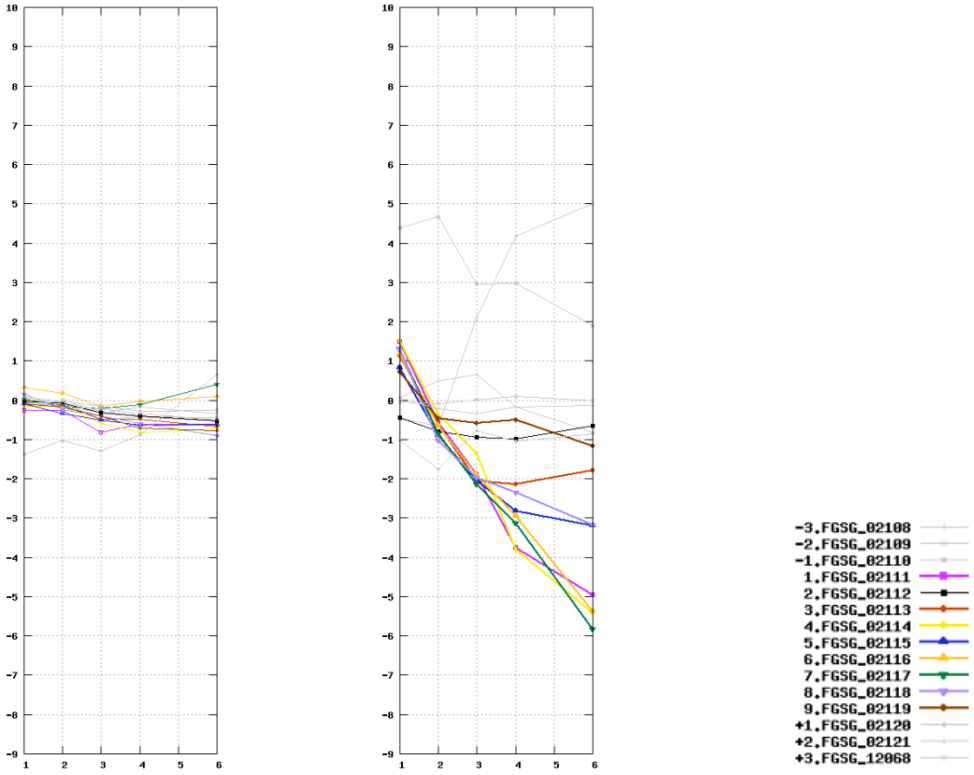
Gene order	FGDB ID	Gene	Gene feature	Gene description
-3	FGSG_13970			related to ankyrin
-2	FGSG_11401			conserved hypothetical protein
-1	FGSG_11400			conserved hypothetical protein
1	FGSG_11399		SP	related to oxidoreductase
2	FGSG_11398			conserved hypothetical protein
3	FGSG_11397			related to desaturase
4	FGSG_11396			related to ASN2 – asparagine synthetase
5	FGSG_11395	NPS14	SM (NPS)	related to AM-toxin synthetase (AMT)
+1	FGSG_11394			related to P.aeruginosa anthranilate synthase component II
+2	FGSG_11393			hypothetical protein
+3	FGSG_11392			conserved hypothetical protein

Table 5-20: Gene functions and features of novel gene cluster 4 (fg1_57)

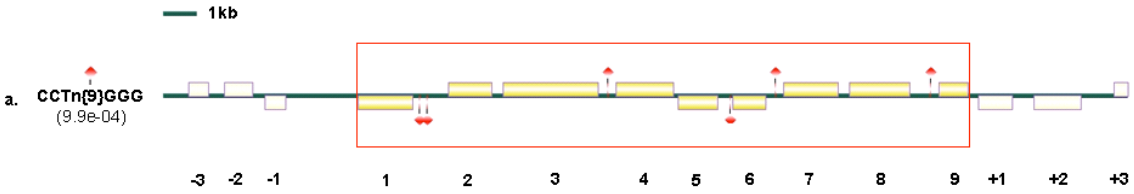
Novel gene cluster 4 contains one nonribosomal peptide synthetase (NPS14) and one secreted protein. NPS14 is related to AM-toxin synthetase but it has not been reported whether NPS14 has the same role in *F. graminearum*. NPS: Non-ribosomal peptide synthetase, PKS: Polyketide synthase, SM: Secondary metabolite, SP: Secreted protein

5.7.2 2 novel gene clusters possibly associated with fungal development

Two novel gene clusters were highly suggested to have functions associated with fungal development; novel gene cluster 5 (Figure 5-16and Table 5-21) and novel gene cluster 6 (Figure 5-17 and Table 5-22). Both novel gene clusters were discovered to have similar expression profiles with the aurofusarin mycotoxin cluster during sexual development and enriched with genes encoding cytochrome P450 enzymes. These two evidences are commonly associated with functions of fungal development. First, the aurofusarin mycotoxin cluster was continuously down-regulated during sexual development, which is consistent with previous reports that aurofusarin accumulation adversely leads to conidia production and mycelial growth of *F. graminearum* [116,117]. The aurofusarin mycotoxin is a red pigment that cause stem and head blight of cereals. Although influences of pigments on fungal development are different, fungal pigments in other species also have been reported to be associated with particular developmental stages [173,174]. Second, cytochrome P450 genes in fungi are known to be involved in diverse biological processes: cell wall biosynthesis, developmental regulation, mycotoxin biosynthesis, pathogenesis, etc [175]. Notably, fungal gene clusters enriched for cytochrome P450 genes have been observed in biosynthetic pathways of secondary metabolites such as gibberellin [97], sterigmatocystin [28], aflatoxin [176], and trichothecene [102]. Considering the steeply decreased expression patterns, genes in the 2 novel gene clusters are inhibited or reduced during sexual development, probably the spores are sensitive to extend levels of aurofusarin.



A. A co-expressed gene cluster during sexual development (fg5_22, right) and profiles of the corresponding genes in planta (left)



B. Gene clusters with co-occurring regulatory motifs

Figure 5-16: Expression profiles and conserved motif seeds of novel gene cluster 5 (fg5_22)

9 genes co-expressed during sexual development are selected as novel gene cluster 5. The clustered 9 genes did not show co-expressed profiles during growth in planta. B. One motif seed examined based on a cluster of co-expressed neighboring genes (fg1_33, in the red box) is suggested as putative regulatory motifs of the novel gene cluster 5. Gene coding regions are represented by boxes. Diamond shapes indicate motif seeds.Number in parenthesis is a p-value of each of motif seeds. Genes and motif seeds in the forward strand are displayed above the line; genes and motif seeds in the reverse strand are displayed below. Numbers underneath each box indicate gene designations (see Table 5-21).

Gene order	FGDB ID	Gene	Gene feature	Gene description
-3	FGSG_02108			conserved hypothetical protein
-2	FGSG_02109		SP	conserved hypothetical protein
-1	FGSG_02110		SP	conserved hypothetical protein
1	FGSG_02111		CYP, SP	related to cytochrome P450 7A1
2	FGSG_02112			conserved hypothetical protein
3	FGSG_02113		CYP, SP	related to cytochrome P450 3A7
4	FGSG_02114		CYP	related to cytochrome P450 7B1
5	FGSG_02115		SP	related to TRI7 – trichothecene biosynthesis gene cluster
6	FGSG_02116			conserved hypothetical protein
7	FGSG_02117		CYP	related to cytochrome P450 monooxygenase (lovA)
8	FGSG_02118		CYP	related to pisatin demethylase / cytochrome P450 monooxygenase
9	FGSG_02119			related to YER185w, Rta1p
+1	FGSG_02120			hypothetical protein
+2	FGSG_02121			related to amidohydrolase AmhX
+3	FGSG_12068			hypothetical protein

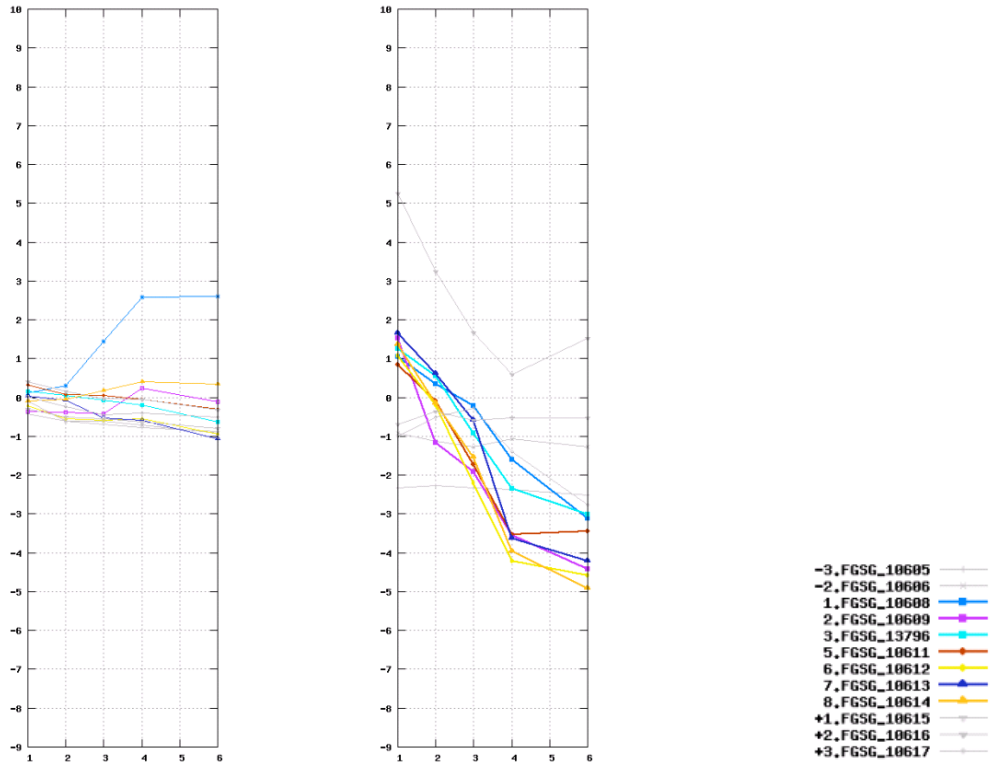
Table 5-21: Gene functions and features of novel gene cluster 5 (fg5_22)

Novel gene cluster 5 consisting of 9 genes is most significantly enriched for genes encoding cytochrome P450 in *F. graminearum*. It is also enriched for genes encoding secreted proteins. SP: Secreted protein, CYP: Cytochrome P450

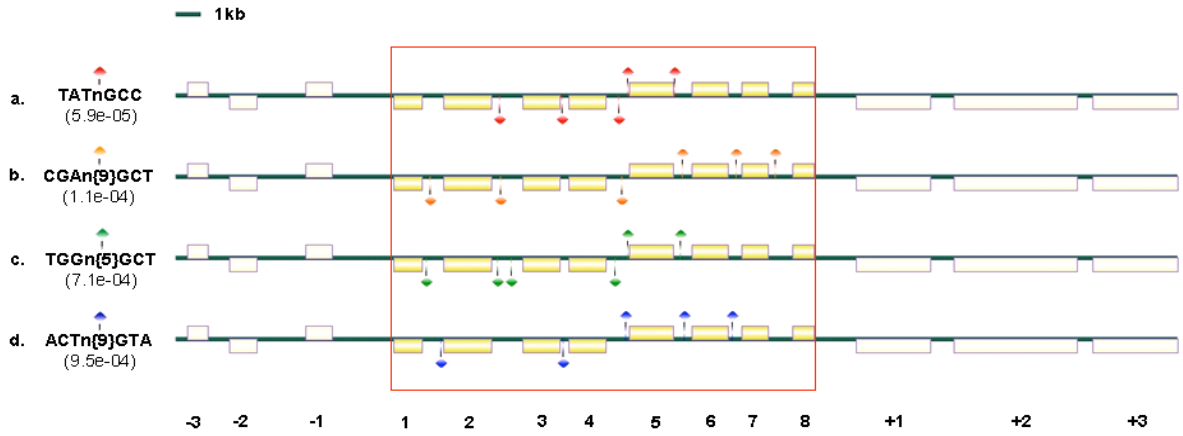
Gene order	FGDB ID	Gene	Gene feature	Gene description
-3	FGSG_10605			conserved hypothetical protein
-2	FGSG_10606			probable cytochrome-c raminearu precursor
-1	FGSG_10607			hypothetical protein
1	FGSG_10608			Glucose/ribitol dehydrogenase (IPR002347)
2	FGSG_10609		SP	related to 6-hydroxy-d-nicotine oxidase
3	FGSG_13796		SP, CYP	related to Tri13 – putative cytochrome p450 monooxygenase
4	FGSG_13797		CYP	related to Tri13 – putative cytochrome p450 monooxygenase
5	FGSG_10611			related to 6-hydroxy-d-nicotine oxidase
6	FGSG_10612			related to salicylate hydroxylase
7	FGSG_10613			related to para-hydroxybenzoate polyprenyltransferase precursor
8	FGSG_10614			Cupin, RmlC-type (IPR011051)
+1	FGSG_10615		SP	related to beta-glucosidase precursor
+2	FGSG_10616			related to vegetatible incompatibility protein HET-E-1
+3	FGSG_10617			related to nonribosomal peptide synthetase MxcG

Table 5-22: Gene functions and features of novel gene cluster 6 (fg5_43)

Novel gene cluster 6 comprises 8 genes and is characterized as a group of enrichment for genes encoding cytochrome P450. SP: Secreted protein, CYP: Cytochrome P450



A. Co-expressed gene cluster during sexual development (fg5_43, right) and profiles of corresponding genes in planta (left)



B. Gene clusters with co-occurring regulatory motifs

Figure 5-17: Expression profiles and conserved motif seeds of novel gene cluster 6 (fg5_43)

8 genes co-expressed during sexual development are selected as novel gene cluster 6. The clustered 8 genes did not show co-expressed profiles during growth in planta. B. Four motif seeds examined based on a cluster of co-expressed neighboring genes (fg5_43, in the red box) are suggested as putative regulatory motifs of the novel gene cluster 6. Gene coding regions are represented by boxes. Diamond shapes indicate motif seeds.The number in parenthesis is a p-value of each of motif seeds. Genes and motif seeds in the forward strand are displayed above the line; genes and motif seeds in the reverse strand are displayed below. Numbers underneath each box indicate gene designations (see Table 5-22).

5.8 Multiple properties of 51 secondary metabolite genes

Genes controlling fungal secondary metabolites (SM) have been observed to be organized in clusters [4,3]. The presence of secondary metabolite genes can be an indication that the cluster is involved in production of secondary metabolites though they are not always necessary for synthesis of the secondary metabolites. Currently, 18 out of 51 SM genes in *F. graminearum* are functionally described; 5 genes are essential for production of mycotoxins (TRI5, PKS12, PKS13, PKS4 and PKS10) and 13 genes have functional annotation and expression specificity (Table 3-3, Table 3-4, and Table 3-5).

By screening for gene clusters with the 3 methods described in detail above, 49 out of 51 SM genes are identified in at least one type of gene cluster (Table 5-23). It is practical to compare with results from other studies focused on single SM genes for advanced research. For instance, 4 PKS genes (PKS3, PKS7, PKS9, and PKS11) were reported to be related to sexual development by expression analysis of individual genes [112]. In our analyses by different properties, these 4 PKS genes showed diverse characteristics. 2 PKS genes, PKS3 and PKS9, are identified as belonging to different co-expressed gene clusters (fg5_142 and fg5_36) during sexual development. These two co-expressed gene clusters showed different expression profiles, indicating that they are involved in sexual development in different aspects. Two other PKS genes, PKS7 and PKS11, are discovered to belong to gene clusters with particular compositions of gene functions (tfcfg_26 and tfcfg_9). PKS11 is also identified in a gene cluster with co-occurring motif seeds (msfg_2, Table 5-7). Characteristics observed in individual SM genes will further our understanding to study detailed functions and/or products of SM gene clusters.

Chr.	FGDB ID	SM gene	TFC ID	Gene cluster with co-occurring MS	Cluster of co-expressed neighboring genes		Orthologs & synteny		Reference	
					FG1 or FG5	COP	Fv	Fo	Products or involved functional process	Species
1	FGSG_11659	NPS8	tfcfg_3	msfg_1						
	FGSG_11660	NPS								
	FGSG_00451	TPS	tfcfg_5	msfg_92			●			
	FGSG_01680	NPS16	tfcfg_6							
	FGSG_11989	NPS19					●			
	FGSG_01738	TPS	tfcfg_7							
	FGSG_01783	TPS	tfcfg_9							
	FGSG_01790	PKS11		msfg_2			●	●	Sexual development related	*F. graminearum* [112]
	FGSG_02315	NPS4	tfcfg_12				●	●		
	FGSG_02324	PKS12			fg5_27	COP5.14	●		Aurofusarin	*F. graminearum* [116,117,118]
	FGSG_02394	NPS15								
	FGSG_02395	PKS13	tfcfg_15						Zearalenon, Grain colonization related	*F. graminearum* [112,103, 121]
	FGSG_12126	PKS4					●		Red pigment bikaverin	*Gibberella fujikuroi* [177]
	FGSG_10097	TPS	-				●	●		
	FGSG_10397	TPS	tfcfg_19							
	FGSG_10464	PKS9	tfcfg_20	msfg_67	fg5_36	COP5.20	●	●	Sexual development related	*F. graminearum* [112]
	FGSG_10523	NPS3	tfcfg_21				●	●		
	FGSG_13783	NPS18	tfcfg_22							
									Mycelial growth related	*F. graminearum* [112]
	FGSG_10548	PKS1	tfcfg_23		fg5_39	COP5.18			Melanin	*Colletorichum lagenarium*, *Magnaporthe grisea*
									T-toxin	*Cochliobolus heterostrophus*
	FGSG_10702	NPS17	-				●			
2	FGSG_08795	PKS7	tfcfg_26				●	●	Sexual development related	*F. graminearum* [112]
	FGSG_08209	NPS7	tfcfg_29		fg5_55	COP5.10	●	●		
	FGSG_08208	PKS6						●		
	FGSG_08181	TPS	tfcfg_30					●		
	FGSG_04591	TPS	tfcfg_33		fg1_33	COP1.1				
	FGSG_04588	PKS15							Expressed during plant infection	*F. graminearum* [112]
	FGSG_03964	PKS14	tfcfg_35						Grain colonization related	*F. graminearum* [112]
									Extracellular siderophores (virulence to plant)	*F. graminearum* etc. [113]
	FGSG_03747	NPS6	tfcfg_39				●	●	Virulence and resistance to oxidative Stress	*Cochliobolus heterostrophus* [178]

	FGSG_03537	TRI5	tfcfg_41	msfg_14	fg1_39	COP1.1			Trichothecenes	*F. graminearum* [102]
	FGSG_03494	TPS	tfcfg_42				●	●		
	FGSG_03340	PKS8	tfcfg_43				●	●	Mycelial growth related	*F. graminearum* [112]
	FGSG_03245	NPS11	tfcfg_45				●	●		
	FGSG_03066	TPS	-	msfg_37			●	●		
	FGSG_04694	PKS2	tfcfg_51				●		Mycelial growth related	*F. graminearum* [112]
	FGSG_05372	NPS2	tfcfg_51	msfg_38				●	Ferricrocin, ascospore development Siderophores	*F. graminearum* [114,115] *F. graminearum* [115]
	FGSG_05794	PKS5	tfcfg_52					●		
	FGSG_11026	NPS1	tfcfg_55	msfg_54	fg5_115	COP5.9	●	●	Siderophores	*F. graminearum* [115]
3	FGSG_10990	NPS9	tfcfg_56	msfg_89	fg1_54	COP1.1				
	FGSG_13878	NPS5								
	FGSG_10933	TPS	tfcfg_57				●	●		
	FGSG_11395	NPS14	tfcfg_61		fg1_57	COP1.1				
	FGSG_11327	TPS	tfcfg_63				●	●		
	FGSG_11294	NPS12	tfcfg_64				●	●		
	FGSG_06444	TPS	tfcfg_67				●			
	FGSG_06507	NPS10	tfcfg_68	msfg_47			●	●	Gliotoxin	*A. fumigatus* [169]
	FGSG_06784	TPS	tfcfg_69				●	●		
	FGSG_13153	NPS13	tfcfg_70							
4	FGSG_07673	TPS	tfcfg_72				●	●		
	FGSG_07798	PKS10	tfcfg_73				●	●	Fusarin C, mycelial growth related	*F. graminearum* [112]
	FGSG_09182	PKS3	tfcfg_76		fg5_142	COP5.6	●	●	Sexual development related Perithecium pigment	*F. graminearum* [112]
	FGSG_09381	TPS	tfcfg_77				●	●		

Table 5-23: Multiple properties of 51 secondary metabolite genes in *F. graminearum*

F. graminearum contains 51 secondary metabolite (SM) genes predicted based on three classes of SM; 15 polyketide synthases (PKS), 20 nonribosomal peptide synthetases (NPS), and 16 terpenoid synthases. Currently, 18 out of the 51 SM genes are functionally described; 5 genes essential for production of mycotoxins (TRI5, PKS12, PKS13, and PKS4 in green-shaded boxes, see Table 3-3) and 17 genes with information of related functions and expression specificity (Table 3-3, Table 3-4, and Table 3-5). By screening three types of gene clusters, 49 out of 51 SM genes are identified in at least one type of gene clusters.

Chr.: Chromosome, SM: Secondary metabolite, TFC: tentative functional gene cluster, MS: motif seed, NPS: Non-ribosomal peptide synthetase, PKS: Polyketide synthase, TPS: Terpenoid synthase, COP: Co-expressed pattern, Fv: *F. verticillioides*, Fo*: F. oxysporum*, FG1: microarray experiment during growth in planta, FG5: microarray experiment sexual development

6 Results: Gene clusters in *Ustilago maydis*

6.1 Secreted protein gene clusters

Sequence analysis of the *Ustilago maydis* genome showed unexpected genomic features, secreted protein clusters with unknown functions [6]. The published secreted proteins were predicted based on a combination of SignalP and ProtCom [6,179,180]. This method yielded 426 candidate secreted proteins. About 20% of its 426 secreted proteins are encoded in 12 gene clusters comprising 3-26 genes of unknown function.

Using TargetP [126] (reliability class 1 and 2) information, secreted protein clusters were re-examined to generate data that are comparable to *F. graminearum*. For analysis of functional enrichment for secreted proteins, about 18 % (115 genes) of 656 predicted secreted proteins are positioned in 25 clusters with 3-10 genes (Table 6-1). The newly identified secreted protein clusters overlap with 10 out of 12 previously predicted gene clusters. The remaining two gene clusters which were not found in our analysis consist of genes with low reliability TargetP (class 3 or 4) information. Our result also includes the Mig1 and Mig2 clusters (spum_15 and spum_25, Table 6-1) that were experimentally shown to be secreted, and especially expressed in maize [181,182]. The clusters were not predicted by previous analyses.

spum ID	Chr.	(N. of gene), gene range	P-value	msum ID	Expression analysis		[c] CID	[d] Phenotype of cluster deletion strain
					[a] 3 fold	[b] No probe		
1	1	(3), um10115~um00446	1.4e-03		2	1	1A	virulence unaffected
2	1	(3), um11443~um11444	1.4e-03		2	0		
3	2	(6), um01235~um01240	1.9e-06		5	1	2A	virulence increased
4	2	(6), um01297~um01302	1.9e-06	msum_8	4	2	2B	virulence unaffected
5	3	(3), um01886~um01888	1.4e-03		0	0	3A	virulence unaffected
6	5	(3), um02137~um02139	1.4e-03	msum_52	0	0		
7	5	(5), um02192~um02196	1.7e-05	msum_51	3	1	5A	virulence unaffected
8	5	(4), um10076~um02231	1.5e-04		1	0		
9	5	(3), um02285~um11403	1.4e-03		1	0		
10	5	(7), um02293~um02299	2.1e-07		7	0		
11	5	(3), um02473~um02475	1.4e-03		3	0	5B	non pathogenic
12	6	(8), um02533~um02540	2.3e-08		4	2	6A	virulence reduced
13	7	(5), um02851~um11484	1.7e-05		4	0		
14	8	(3), um03201~um10403	1.4e-03		2	1	8A	virulence unaffected
15	8	(4), um03223~um12217	1.5e-04		2	2		
16	10	(10), um03744~um03753	2.8e-10		4	5	10A	virulence reduced
17	14	(3), um04353~um04355	1.4e-03		2	0		
18	19	(4), um05294~um10553	1.5e-04		4	0		
19	19	(10), um05299~um05308	2.8e-10	msum_34	8	1	19A	virulence dramatically reduced
20	19	(3), um05310~um05312	1.4e-03		2	1		
21	19	(5), um05314~um05319	1.7e-05		3	0		
22	20	(3), um05926~um05928	1.4e-03		2	0		
23	20	(3), um05930~um05932	1.4e-03		3	0		
24	21	(3), um06126~um06128	1.4e-03		3	0		
25	22	(5), um11250~um06181	1.7e-05		3	1		

Table 6-1: Gene clusters for secreted proteins and their features in *U. maydis*

25 gene clusters enriched for secreted proteins were identified using TargetP (reliability class 1 and 2) information. The 25 clusters contain 10 out of 12 previously defined secreted protein clusters [6].

4 out of 25 gene clusters overlap with gene clusters with co-occurring motif seeds. The largest secreted protein cluster (cluster spum_19) was identified by the conserved motif seed, 5'-ATGn{3}GAC-3' (cluster msum_34, Table 6-1). It is observed that virulence was markedly reduced in phenotypes of the deletion strain of gene cluster spum_19. Transcripts of secreted protein clusters may be coordinated by the common regulatory element.

Spum ID: ID of secreted protein clusters identified by analysis of functional enrichment. Msum ID: ID of gene clusters with co-occurring motif seeds in *U. maydis*. Expression analysis: DNA array expression analysis of the depicted genes in fungal cells from tumor tissue and in cells grown in axenic culture [6]. [a]: Number of genes with a larger than 3 fold change in expression. [b]: Number of genes which is not present on the *U. maydis* Affymetrix microarray used for the analysis. [c]: Cluster ID (CID) previously published secreted protein clusters identified by a combination of SignalP and ProtCom [6]. [d]: Phenotypes of plants infected by deletion mutants of the individual secreted protein clusters[6].

6.2 Gene clusters with co-occurring promoter motifs

In total, about 13 percent (910/6782) of predicted genes in *U. maydis* are arranged in 67 clusters having 6-34 genes with co-occurring motif seeds (Table S 8-2). The 67 gene clusters found were examined with respect to known functional features. Two types of interesting features were discovered; secreted protein clusters and putative necrotrophic gene clusters.

Clusters enriched for secreted proteins were identified in gene clusters with co-occurring motif seeds. 4 out of 25 secreted protein clusters overlap with gene clusters with co-occurring motif seeds at the genomic level. The largest secreted protein cluster (cluster spum_19, Table 6-1) of *U. maydis* overlap with gene cluster msum_34 which has the conserved promoter motif seed, 5'-ATGn{3}GAC-3' (Figure 6-1). Genes in the cluster msum_34 were highly up-regulated during tumor stage[6], suggesting that transcripts of the secreted protein cluster spum_19 can be coordinated by the same regulatory element (5'-ATGn{3}GAC-3). In addition, virulence was markedly reduced in phenotypes of the gene cluster spum_19 deletion strain [6].

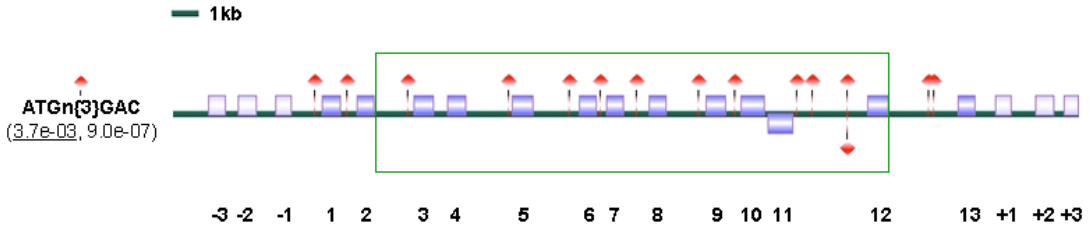


Figure 6-1: The largest secreted protein cluster with co-occurring motif seed in *U. maydis*

The largest secreted protein cluster (cluster spum_19 highlighted in the green box) in *U. maydis* is identified by analysis of clusters with co-occurring motif seeds. Virulence was dramatically reduced in the spum_19 deletion strain[6]. Gene coding regions are represented by boxes: dark blue refers to co-expressed genes and light blue represents flanking genes. Diamond shapes indicate motif seeds. The number in parenthesis is p-values of motif seed (The number with underline is an adjusted P-value). Genes and motif seeds in the forward strand are displayed above the line; genes and motif seeds in the reverse strand are displayed below. Numbers underneath each box indicate gene designations.

Putative necrotrophic gene clusters were identified in *U. maydis* by analysis of conserved motif seeds in neighboring genes. The biotrophic pathogen, *U. maydis* lacks many characteristics observed in necrotrophic fungi. No mycotoxin cluster has been reported at present. There are a smaller number of genes known to be involved in pathogenesis compared to the many known and predicted in necrotrophic fungi. For instance, *U. maydis* contains 6 genes predicted to be related to polyketide synthases (PKS) whereas the necrotrophic pathogen, *F. graminearum* has 15 PKS genes. Interestingly, 4 out of these 6 predicted PKS genes in *U. maydis* were discovered in two gene clusters with co-occurring motif seeds (msum_11 and msum_10). One cluster msum_11, comprising 17 genes, was identified by three co-occuring motif seeds, 5'-GGGTAA-3', 5'-TTACCC-3', and 5'-GTAn{3}GTT-3' (Figure 6-2). The cluster msum_11 includes 4 kinds of functional features of interest mostly found in mycotoxin gene clusters; there are genes encoding 3 PKS, 2 transcription factors, 1 cytochrome P450 and 5 secreted proteins (Table 6-2). Notably, the gene cluster msum_11 is located in the telomere of chromosome 12 and a species-specific gene cluster which is likely absent in two other close species, *Sporisorium reilianum* and *Ustilago hordei*, (preliminary sequencing data) supposing that genes in the cluster msum_11 are maybe in continuous rearrangement or gained/lost due to evolutionary constraints like environmental conditions and host specificity. 2 out of 17 genes in the cluster msum_11 have expression values from an experiment during tumour stage and were highly up-regulated (Table 6-2), suggesting that it is also possible that genes of the cluster may have an independent mechanism to induce tumours or an associated function with one of the secreted protein clusters although expression values of other genes are not present and a result for phenotype analysis is absent. Another gene cluster msum_10 consists of 10 genes which have a conserved common motif seed 5'-GACn{3}GTC-3' in their promoter sequences. It contains one PKS gene and one secreted protein. Expression values of genes in the cluster msum_10 are diverse during tumour stage.

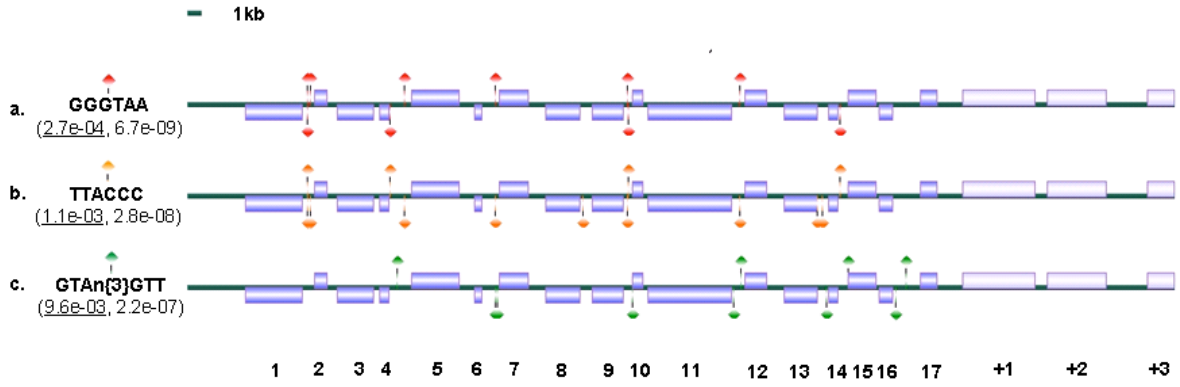


Figure 6-2: Potential necrotrophic gene cluster with co-occurring motif seeds in *U. maydis*

Gene cluster msum_11 with a typical composition of gene functions as observed in mycotoxin clusters, identified in *U. maydis* by analysis of conserved motif seeds in neighboring genes. Genes in the cluster contain four types of functional features described in Table 6-2. Gene coding regions are represented by boxes: dark blue refers to co-expressed genes and light blue represents flanking genes. Diamond shapes indicate motif seeds. The number in parenthesis is p-values of each of motif seeds (The number with underline is an adjusted P-value). Genes and motif seeds in the forward strand are displayed above the line; genes and motif seeds in the reverse strand are displayed below. Numbers underneath each box indicate gene designations (see Table 6-2).

Gene order	MUMDB ID	Exp.	Gene feature	Gene description
1	um04095	-	SM (PKS)	related to polyketide synthase
2	um04096	-	SP	hypothetical protein
3	um04097	-	SM (PKS)	related to polyketide synthase
4	um04098	-		conserved hypothetical Ustilago-specific protein
5	um11110	-	TF	conserved hypothetical protein
6	um04100	-		hypothetical protein
7	um04101	-	TF	related to BAS1 – transcription factor
8	um11111	-	SP	related to Ascorbate oxidase precursor
9	um11112	6.2		related to Versicolorin B synthase
10	um04104	-	SP	conserved hypothetical Ustilago-specific protein
11	um04105	-	SM (PKS)	related to Polyketide synthase
12	um04106	6.2		related to O-methyltransferase B
13	um04107	-		related to Phenol 2-monooxygenase
14	um12253	-	SP	conserved hypothetical Ustilago-specific protein
15	um04109	-	CYP	related to Cytochrome P450
16	um11113	-		conserved hypothetical protein
17	um04111	-	SP	conserved hypothetical protein
+1	um11114	1.4		conserved hypothetical protein
+2	um11115	-1.3		myosin I
+3	um04114	-1.0	SP	probable PHO8 – repressible alkaline phosphatase vacuolar

Table 6-2: Gene functions of potential necrotrophic gene cluster in *U. maydis*

Cluster smum_11 contains typical composition of gene functions as observed in mycotoxin gene clusters. It contains genes encoding 3 polyketide synthase, 2 transcription factors, 1 cytochrome P450, and 5 secreted proteins. Only 2 out of 17 genes in the cluster msum_11 have expression values from an experiment during tumor stage and were highly up-regulated.

Exp.: Expression value of microarray experiments for the depicted genes in fungal cells from tumor tissue and in cells grown in axenic culture [6]. (-) denotes that respected gene is not present on the *U. maydis* Affymetrix microarray used for the analysis.

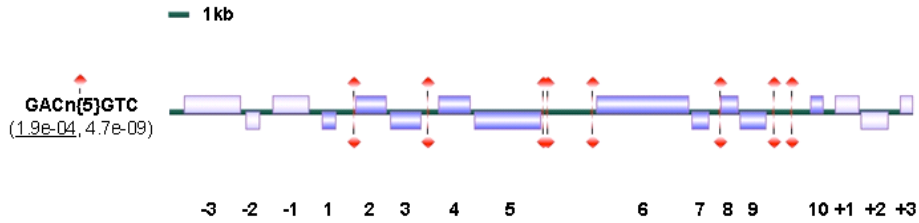


Figure 6-3: Putative secondary metabolite gene cluster with co-occurring motif seeds in *U. maydis*

Putative secondary metabolite gene cluster msum_10 with conserved motif seeds in neighboring genes. Gene coding regions are represented by boxes: dark blue refers to co-expressed genes and light blue is flanking genes. Diamond shapes indicate motif seeds. The number in parenthesis is the p-value of the motif seed (The number with underline is an adjusted P-value). Genes and the motif seed in the forward strand are displayed above the line; genes and the motif seed in the reverse strand are displayed below. Numbers underneath each box indicate gene designations (see Table 6-3).

Gene order	MUMDB ID	Exp.	Gene feature	Gene description
-3	um12267	3.5		related to IST2 – Plasma membrane protein that may be involved in osmotolerance
-2	um04446	-1.1		related to U6 snRNA-associated Sm-like protein LSm4
-1	um04447	-1.6		conserved hypothetical protein
1	um04448	1.1		related to Myosin regulatory light chain 2-A, smooth muscle isoform
2	um10531	2.2		putative protein
3	um10532	-1.7	SM (PKS)	related to polyketide synthase required for biosynthesis of fumonisin mycotoxins
4	um04451	-1.3		conserved hypothetical protein
5	um10533	1.3		probable KGD1 – alpha-ketoglutarate dehydrogenase
6	um10534	2.3		related to Protein-tyrosine raminearum, receptor type 1
7	um10535	-2.4		related to cell cycle control protein cwf15
8	um10536	1.4	SP	related to FPR2 – FK506/rapamycin-binding protein of the ER
9	um04456	-		protein kinase A, catalytic subunit
10	um04457	-1.1		conserved hypothetical protein
+1	um10538	-1.3		conserved hypothetical protein
+2	um10539	-1.7	SM (TPS)	related to farnesyltranstransferase (al-3)
+3	um04460	-1.9		related to RPB8 – DNA-directed RNA polymerase I, II, III 16 KD subunit

Table 6-3: Gene functions of putative secondary metabolite gene cluster in *U. maydis*

Cluster smum_10 comprising 10 genes including one PKS gene and one secreted protein. Expression values of genes are diverse during tumor stage.

Exp.: Expression value of microarray experiments for the depicted genes in fungal cells from tumor tissue and in cells grown in axenic culture [6]. (-) denotes that respected gene is not present on the *U. maydis* Affymetrix microarray used for the analysis.

7 Discussion

Genes which are involved in fungal secondary metabolites synthesis and are involved in the assimilation of certain nutrients have been found to be clustered in diverse fungi. These clustered genes are of major interest by virtue of both their beneficial (pharmaceutical applications) and detrimental (toxic agents) properties. Classic studies focused on single gene clusters or individual genes involved in certain specific functions are now complemented by studies at the genomic level as complete fungal genomes and high-throughput technologies are available. The growing number of fungal genomes available bears another opportunity to find and explore the obviously wide range of unknown fungal secondary metabolites. In *Fusarium graminearum*, a total of 51 genes involved in secondary metabolite synthesis are predicted, which exceeds the number of known secondary metabolite products in the organism. Beside this, analysis of whole genome sequences also revealed new types of clustered genes such as secreted protein clusters of *Ustilago maydis*, probably effectors of pathogen-host interaction. Thus, screening of fungal genomes for gene clusters through diverse approaches at the genomic level is crucial to facilitate unveiling of undiscovered pathway products and/or the understanding of functional relationships.

Five mycotoxin gene clusters of *F. graminearum* are re-discovered

Our investigation of three properties, co-expression, the particular compositions of gene functions, and the presence of consensus promoter motifs, allowed the identification of all 5 mycotoxin gene clusters (aurofusarin, butenolide, fusarin C, trichothecene, and zearalenone) reported in *F. graminearum* at the genomic level (Table 5-1). The trichothecene gene cluster was identified using any of the three properties for analyses, and includes a regulatory element overlapped with one proven in the trichothecene genes of *F. sporotrichioides*. The four other remaining mycotoxin gene clusters are discovered using one or two properties for analyses. In addition, our analyses (re-)demarcated currently known mycotoxin gene clusters except the butenolide gene cluster. In particular, a gene cluster boundary for the fusarin C mycotoxin cluster, which has the only one gene (PKS10) known to synthesize the fusarin C, was first proposed. This result is supported by a preliminary experiment conducted in *F. verticilliodies*, which shows that the conserved syntenic genes of 8 fusarin C genes identified by our analysis were co-regulated and expressed under nitrogen sufficient conditions (unpublished data, and further deletion studies will be conducted in an upcoming project). Not

surprisingly, co-expression studies by two microarray experiments revealed a characteristic of mycotoxin genes, condition-specific co-regulation. For instance, 12 core trichothecene genes were co-expressed with a pattern clearly distinguishable from expression patterns of flanking genes in both directions during growth in planta. However, these genes were not shown to have significant expression during sexual development and were not discernible from other flanking genes (Figure 5-2). These results shown in re-discovered mycotoin gene clusters present that any of three properties alone may not be enough to determine unequivocal mycotoxin gene clusters, but can be instrumental in screening promising mycotoxin gene clusters. Additional standards for describing mycotoxin gene clusters are still required to be developed.

Clusters of co-expressed neighboring genes in *F. graminearum*

Co-regulation is one of the representative characteristics of clustered fungal genes on chromosomes that are linked by common attributes. Co-regulation can be partly explained via co-expression by coordinated transcriptional control. Indeed, the co-regulation of transcripts is seen for secondary metabolite clusters and the DAL cluster in yeast, as well as the clusters encoding secreted proteins in *U. maydis*. Therefore, observation on co-expressed neighboring genes can help to determine fungal gene clusters.

In *F. graminearum*, co-expressed neighboring genes were explored using Affymetrix microarray data of two different conditions for the first time. For microarray experiments containing time-series during growth in planta (FG1) and during sexual development (FG5), about 3 and 5 percent of neighboring genes, respectively, lie in groups containing 3 to 15 genes that have significant correlations in expression (Table 5-9). By the analysis, three mycotoxin gene clusters were identified; butenolide [90] (fg1_32) and trichothecene [102,134] (fg1_39) in planta and aurofusarin [168] (fg5_27) during sexual development.

About 80 % of the clustered genes are organized in large clusters containing more than four genes. Although the clusters are based only on two sets of expression data, the structure of the clusters is closer to the ones observed in higher eukaryotes rather than yeast. In yeast, clusters of co-expressed neighboring genes consist mostly of two or three genes [77]. Conversely, large clusters of co-expressed genes have been found in higher eukaryotes such as *Caenorhabditis elegans* [183], *Arabidopsis thaliana* [184], Drosophila [185], and mammals [186]. Several possible mechanisms involved in the formation and maintenance of large co-expressed gene clusters were observed in secondary metabolite gene clusters

in fungi. First, chromatin-based regulatory mechanisms can be supported. E. g. several secondary metabolite gene clusters in *Aspergillus nidulans* showed that localized modification of chromatin structure might play an important role in their co-regulation [187,188]. It has also been reported in the higher eukaryote Drosophila that regulation of an independent chromatin domain encompasses non-homologous co-expressed neighboring genes [189]. Second, horizontal gene transfer of advantageous gene clusters has been suggested to force keeping genes clustered on the chromosome. Secondary metabolite gene clusters such as penicillin (beta-lactam) from *Penicillium* species [190,191] and the ACE1 cluster from *Magnaporthe grisea* [192] are absent even in closely related species but present in more distantly related species. Moreover, isolates of a species from different environments have been shown lack of genes involved in secondary metabolism. All this emphasizes on some evolutionary pressure to keep the genes of beneficial pathways together in clusters.

Limitation of microarray data to screen fungal gene clusters

Fungal gene clusters can be difficult to monitor by comparing gene expression profiles from microarray experiments alone because of strongly differing and selective influences on transcription of fungal genes under different conditions. The zearalenone (ZON) mycotoxin is a good example to show the difficulty of identifying fungal gene clusters. ZON is known to affect sexual development as a sex hormone in *F. graminearum*. However, ZON mycotoxin genes were not discovered from our analysis of co-expression during sexual development. This result is probable because ZON controls diverse regulatory mechanisms involved in reproduction according to different conditions [193,158]. ZON enhances as well as inhibits sexual reproduction by different amounts of ZON and its activities are time dependant. Another functional study of ZON for infection of hosts have shown that ZON does not appear to be important for infection of wheat and barley [121,163,194]. Our microarray result also showed that ZON mycotoxin genes were not significantly co-expressed during growth in planta. However, a recent study suggested that *F. graminearum* could produce ZON during infection of wheat. ZON genes were active in wheat having clear disease symptoms of infection by 10 and 14 days after inoculation, thus not monitored by the data we used [103]. Another example is that 3 mycotoxin clusters identified by co-expression studies showed expression profiles indistinguishable from ones associated with their flanking genes under the 2 conditions monitored. The trichothecene mycotoxin genes co-expressed in planta were concealed under sexual development (Figure 5-2, A). Although analysis of microarray data is useful to find co-expressed gene clusters at the genomic level, efficient screenings

are additionally required to detect fungal gene clusters considering fungal gene clusters that frequently remaining silent under unfavourable environments.

Tentative functional gene clusters (TFCs) deduced by particular compositions of gene functions

Physical gene clusters are deducible on the basis of the particular compositions of gene functions of known fungal gene clusters based on five types of functional descriptors (Table 4-1). Many genes involved in metabolite production are clustered on chromosomes. Particular enzymes, such as SM genes and cytochrome P450, are often found to reside within clusters of closely located genes that encode transcription factors, transporters and additional enzymes required for the modification of the final products. By comparing with the occurrence of the same sets of genes in random genomes, 77 gene clusters are identified in *F. graminearum* (Table 5-7). This includes all five mycotoxin gene clusters, the aurofusarin, butenolide, trichothecene, zearalenone and fusarin C gene clusters.

Gene clusters identified by particular compositions of gene functions provide important indications for further research at the level of gene clusters by proposing interesting functional features located closely together. For instance, PKS10 is the only gene known to produce the fusarin C mycotoxin, an immunosuppressant of various crops. PKS10 was not identified by the expression data analysis of two microarrays. Other genes which may be involved in synthesis of the fusarin C have not been determined yet. By examining compositions of gene functions in neighboring genes, PKS10 was found in a significant gene cluster with 7 other genes that are predicted to encode a cytochrome P450 and a transporter, suggesting that the fusarin C mycotoxin may be associated with a gene cluster responsible for its synthesis (Table 5-6). Another 76 gene clusters (Table 5-7) in addition to the four mycotoxin gene clusters are also suggested to have statistically significant composition of gene functions, and their roles remain to be established.

The approach is superior to SMURF

Gene clusters identified by analysis of particular compositions of gene functions are comparable with ones from the program 'Secondary Metabolite Unique Regions Finder' (SMURF, http://www.jcvi.org/smurf/, unpublished). This program finds secondary metabolite biosynthesis genes and pathways in fungal genomes based on PFAM and TIGRFAM domain. Comparing results between our analysis and the SMURF program with known mycotoxin gene clusters, the SMURF program didn't find a representative mycotoxin gene cluster of *F. graminearum*, trichothecene but our analysis found it. A crucial difference between the two methods is that the SMURF uses two classes of secondary metabolite (SM) genes (PKS and NPS) as backbone genes whereas our analysis of functional organizations apply 4 kinds of base genes, 3 classes of SM genes (PKS, NPS, and TPS) and cytochrome P450 genes. Additionally, our analysis utilizes three other functional categories (transcription factors, transporters and secreted proteins), which often co-occurred in fungal gene clusters. Our method also predicts cluster boundaries more precisely and calculates a statistical significance. Though ascertaining the number of genes involved in production of diverse metabolites is still difficult, an examination of additional diagnostic enzymes will improve the capacity to deduce fungal gene clusters.

Discovery of gene clusters of secondary metabolites and secreted proteins without prior knowledge of gene functions

Exploration of promoter motifs of neighboring genes helps to identify gene clusters which may be co-regulated on the transcription level. Some fungal metabolic pathway gene clusters have been reported in coordinated gene transcription through the action of cis-regulatory elements [195,196]. Their structure differs considerably from operons found in prokaryotic genomes. Each of the fungal genes of a common pathway or function can have consensus promoter motifs which are recognized by fungal specific-regulatory proteins containing the Zn2Cys6 zinc binuclear or the Cys2His2 zinc-finger domain. Here, promoter regions were scanned for co-occurring promoter motifs in neighboring genes by a sliding window without regarding any functional information. In *F. graminearum* and *U. maydis*, 95 and 67 gene clusters showed co-occurring promoter motifs which are overrepresented at the genomic level and were ranked by their statistical significance, respectively. Different types of gene clusters were characterized from the found gene clusters by examination with respect to compositions of gene functions, which include gene clusters of secondary metabolites and secreted proteins.

In *F. graminearum*, our analysis discovered the trichothecene mycotoxin genes having the conserved promoter motif 5'-AGGCCT-3', ranked 14th according to significance at the genome (Figure 5-2 and Table 5-2). The motif overlaps the TRI6 binding promoter motif, 5'-TNAGGCCT-3', which was validated among the trichothecene genes of *F. sporotrichioides* [110]. Other discovered motifs lack experimental evidences but clusters can be characterized by specific composition of gene functions. 11 out of 51 SM genes belong to 10 of the identified regulatory motif gene clusters. The presence of SM genes is an indicator that the motifs may be involved in production of certain secondary metabolite with adjacent genes though they may not be exclusively involved in secondary metabolism. Moreover, it is notable that the first two significant gene clusters with co-occurring promoter motifs have typical structures of cluster-specific composition of gene functions (Table 5-11, msfg_1 and msfg_2). The most significant gene cluster (msfg_1) contains two non-ribosomal peptide synthetase (NPS) genes and one transporter. The second significant gene cluster (msfg_2) includes the PKS11 gene which is known to be related to sexual development. This gene cluster and their promoter motifs are conserved in two other *Fusarium* species (Table 5-15) but it has not been reported whether the clustered genes has the same role. Functional relations of genes in the two gene clusters, msfg_1 and msfg_2, have not been reported. For these gene clusters identified by co-occurring promoter motifs, studies on their expressions, as well as a detailed analysis of the mechanisms of co-regulation are promising to understand the regulatory systems and biological functions of fungal gene clusters.

In *U.maydis*, gene clusters enriched for secreted proteins were identified by analysis of regulatory elements in neighbouring genes. These secreted protein clusters were reported to be a unique feature first revealed in *U. maydis* [6]. The expression of most clustered genes was significantly up-regulated during the tumour stage [6]. However, the mechanism of co-ordinated expression for the secreted protein clusters is not known. Interestingly, the largest secreted protein cluster (cluster spum_19, Table 6-1) of *U. maydis* was discovered in gene cluster msum_34 which has a conserved promoter motif seed, 5'-ATGn{3}GAC-3' (Figure 6-1). It has been observed that virulence was markedly reduced in phenotypes of the gene cluster spum_19 deletion strain [6]. The presence of the motif seed suggests a possible mechanism where by transcripts of the secreted protein cluster can be coordinated by the same regulatory elements.

A genome-wide approach using conservation of promoter motifs represent a new paradigm to identify fungal gene clusters. However, technical limitations to predict promoter motifs at the genomic level still remain. Our methodology for genome-wide promoter motif discovery involves sequences of the structure 'ABCn{0-9}DEF', consisting of two triplets of specified nucleotide bases separated by a

fixed number of unspecified nucleotide bases. The basic structure is constructed from the well-studied promoter motifs, 5'-TCGn{5}CGA-3', and 5'-TnAGGCCT-3', which are found in the aflatoxin and trichothecene mycotoxin gene cluster of several *Aspergillus* species and *F. sporotrichioides*, respectively [29,110]. However, structures of promoter motifs can be diverse some of which are shorter than a hexamer or are degenerated. Besides, promoter motifs can occur frequently by chance at the genomic level, and hence the enrichment observed is not always sufficient to detect functional motifs with high sensitivity and/or specificity. Comparative study of promoter regions of conserved gene clusters in different species may help to eliminate and select motif seeds with high sensitivity and specificity as exemplified by conserved promoter motifs as well as orthologous genes in aflatoxin-producing *Aspergillus* species [29,109].

Re-demarcation of trichothecene gene cluster supported by two evidences

Examination of multiple properties suggests that additional genes are possibly associated with trichothecene mycotoxin biosynthesis. Trichothecene is a well-studied mycotoxin which causes head blight disease in wheat and barley and is also strongly associated with chronic and fatal toxicity in humans and animals. Trichothecene biosynthetic genes well characterized in 2 *Fusarium* species (*F. graminearum* and *F. sporotrichioides*) have been reported at three loci on different chromosomes; a core cluster containing 12 genes and two miniclusters [102,197,198,199,145,200]. 12 genes are responsible for synthesis of the core trichothecene molecule with several modifications. The one mini cluster includes a single gene (acyl transferase, TRI101) [199]. Other mini cluster consists of two genes (cytochrome P450 monooxygenase (TRI1) and acyl transferase (TRI16)) [145]. For trichothecene biosynthesis, it has been expected that additional genes may belong to this cluster because there are un-assigned steps in the biosynthetic pathway [142]. However, a number of observations from gene deletion, disruption, and EST-based expression analysis indicated that none of genes adjacent to the previously identified trichothecene genes are required for trichothecene biosynthesis even though some ORFs adjacent to the trichothecene cluster showed similarity to proteins with enzymatic activities of unassigned biochemical reactions in two *Fusarium* species [102]. Strikingly, our analyses detected 6 additional genes that can be linked to unassigned biochemical steps by having two common properties with the 12 core trichothecene genes. The 3 genes (FGSG_03531, FGSG_03530, and FGSG_03529) flanking directly the trichothecene genes most probably enlarge this well known cluster according to co-expression (cluster fg1_39) and co-occurring regulatory motifs (cluster msfg_14). Another 3 genes (FGSG_00071 (TRI1, cytochrome P450 monooxygenase), FGSG_08809 (trichodiene oxygenase cytochrome P450),

and FGSG_07896 (trichothecene 3-O-acetyltransferase)) dispersed on different chromosomes were detected from 26 genes that has been annotated as related to trichothecene synthesis in FGDB. These 3 genes were co-expressed in planta with the trichothecenes cluster and showed the same promoter motifs as found with the 12 core trichothecene genes whereas the other 23 genes had neither similar expression patterns nor a significant common promoter motif detected. One (FGSG_00071, TRI1) out of 6 genes is reported to be invovled in trichothecene production in several *Fusarium* species [143,145,201]. The detailed roles of the 6 genes in *F. graminearum* remain to be determined in terms of pathway mechanisms of trichothecene synthesis and their evolutionary changes. This result indicates that integration of multi-properties may provide important insight into prediction of dispersed gene clusters and gene organizations of secondary metabolite genes.

Limitation to predict dynamic fungal gene clusters

In many cases, genes belonging to common secondary metabolite pathway are clustered at a single genomic locus. This feature makes prediction of fungal gene clusters relatively straightforward. However, genes for fungal secondary metabolite biosynthesis can also be located in separated genomic regions. Examples of separted gene clusters include the trichothecene gene clusters induced by diverse mechanisms such as evolutionary rearrangements by ecological adaptations or unknown reasons. Production of T-toxin, a causal agent of southern corn leaf blight, of *Cochliobolus heterostrophus* is controlled by two unlinked loci; Tox1A containing two PKSs (PKS1 and PKS2) and Tox1B containing a decarboxylase (DEC1) [202,203,204]. Mutation of genes at both loci leaded to loss of toxin production and reduced virulence towards maize [203,204]. Recently identified genes for biosynthesis of aflaterm (ATM), a potent tremorgenic toxin known to lead to neurological disorders, are also positioned in two loci in *Aspergillus flavus* and *Aspergillus oryzae*; the ATM1 locus containing 3 genes on chromosome 5 and the ATM2 locus containing 5 genes on chromosome 7 [205]. Reverse transcriptase PCR in *A. flavus* verified that the ATM biosynthesis transcript levels increased with the onset of ATM production [205]. The existence of the ATM gene clusters, separated into two miniclusters, in the two *Aspergillus* speices was first discovered by homology searches of complete genome sequences with the paxilline biosynthesis genes of *Penicillium paxilli*. The paxilline gene cluster is located at a single genomic locus [206], suggesting that the physical arrangement of ATM genes in two *Aspergillus* species may have arisen by fragmentation of a single ancestral cluster [205].

To investigate genes involved in the same pathway and positioned in separate clusters, additional application is required such as the integration of properties derived from experimental evidences. For instance, for our analyses by integration of two types of evidences, 6 additional genes which are possibly associated with production of trichothecene in *F. graminearum* as already described above are suggested. This prediction was accomplished by 2 steps; first, a core gene cluster was identified by analyses of neighboring genes with co-expressed profiles and co-occuring promoter motifs. Second, genes on the genome were re-examined and selected using common features found in the core gene cluster. However, it is still limited to predict defined fungal gene clusters because secondary metabolite clusters have grown by gene reorganization during evolution [192,207,200]. Comparative analysis of gene arrangement in both trichothecene-producing and non-producing *Fusarium* species revealed that TRI1 and TRI101 are located in the core gene cluster in some species but not in others, in addition TRI101 appears to have evolved separately [200,208]. Some of genes involved in trichothecene production have been reported to have functional differences in different species. For instance, the TRI1 enzyme is responsible for important structural differences between trichothecenes produced by different *Fusarium* species. TRI16 is non-functional in *F. graminearum*, whereas the TRI16 enzyme catalyses esterification of a five-carbon carboxylic acid in *F. sporotrichioides*. Differences between trichothecene-related genes can be particular to the chemotype of the strains being used for expression analysis. Various chemotypes of the trichothecenes have been described in different environments, diverse species and strains [135,136,137,138]. A species or strain typically produces only a limited number of trichothecenes under certain conditions. For instance, *F. graminearum* produce most abundantly deoxynivalenol (DON) or nivalenol (NIV) types but *F. sporotrichioides* is known to produce primarily the T-2 toxin chemotype. It is unclear what evolutionary forces induce and maintain reorganization of genes, and cause various chemotypes in different species. This means that to fully understand secondary metabolites it is necessary to understand not only specific properties such as co-expression analysis but also the mechanisms of evolution and regulation for gene clusters.

Six novel gene clusters in *F. graminearum*

In *F. graminearum*, 6 novel gene clusters are strongly suggested based on evidences observed from mycotoxin gene clusters; 4 gene clusters possibly involved in plant infection (ID 1~4, Table 5-16) and 2 gene clusters probably associated in fungal development (ID 5~6, Table 5-16). Characteristics from mycotoxin gene clusters identified by our analyses were examined and re-compared with other gene clusters. The 6 novel gene clusters have specific composition of gene functions of genes and similar expression profiles with 3 mycotoxin gene clusters in two independent microarray experiments.

The former 4 novel gene clusters are expected to have important roles related to plant infections or produce unknown mycotoxins, supporting by two types of remarkably distinguishable features, expression profiles and specific composition of gene functions. Among 76 co-expressed gene clusters identified in planta, only 4 novel gene clusters have a highly up-regulated peak at the time point of the 3rd day similar to those of the two mycotoxin clusters (the trichothecene and the butenolide) and compositions of gene functions of interest, whereas the other 70 co-expressed gene clusters show continuously increasing expression patterns and have no genes with function observed in known gene clusters. The compostion of gene functions of the 4 novel gene clusters are that one cluster (fg1_58) has enriched secreted proteins and the other three clusters (fg1_54, fg1_33, and fg1_57) contain SM genes.

Highest priority for further study can be given to the cluster fg1_33 containing PKS15 and further 9 genes (Figure 5-13 and Table 5-18). PKS15 is reported to be expressed exclusively during plant infection, and has been considered as one of the strong candidates producing a metabolite of unknown function [112]. However, no additional information has been determined for adjacent genes to PKS15. Interestingly, the other 9 genes were found to encode proteins which often occur in other mycotoxin gene clusters; one terpenoid synthase, one cytochrome P450, two secreted proteins, and five diverse enzymes. No pathway-specific transcription factor is found in the cluster. It is possible because transcription of genes in the cluster may be controlled by other regulatory proteins like LaeA, which is an archetypal global regulator of secondary metabolism in fungi [68]. By providing genes possibly involved in a common pathway or function with the PKS15 it can be expected to accelerate further research extensively.

The latter 2 novel gene clusters (ID 5~6, Table 5-16) are suggested to have functions associated with fungal development. This is supported by two evidences; co-expressed profiles with genes

synthesizing aurofusarin mycotoxin during sexual development and enrichment of genes encoding cytochrome P450. First, 2 novel gene clusters were co-expressed with the aurofusarin mycotoxin, a red pigment that cause stem and head blight of cereal. Fungal pigments in other species have been reported to be associated with particular developmental stages [173,174]. Some pigments are known to be required for sporulation structures, the most common being melanins that contribute to virulence and the survival of the fungal spore by protecting fungal cells against immune effector cells and UV damage [209,210,211]. In contrast to the influence of melanins that assist sporulation, our result showed that the aurofusarin mycotoxin genes were continuously down-regulated during sexual development, which is consistent with previous reports that the aurofusarin accumulation adversely lead to the conidia production and mycelial growth of *F. graminearum* [116,117]. There has been described no obvious reason of the negative correlation between aurofusarin production and fungal growth rate. It can be speculated that spores are sensitive to extend levels of aurofusarin and/or other pigments or secondary metabolites necessary for the meiotic spores. Second, the two novel gene clusters are enriched with genes encoding cytochrome P450 enzymes. In *F. graminearum*, over a quarter of cytochrome P450 genes (30 out of 117, 25.6%) are enriched in 11 genomic regions with 2-11 spanned genes (Table 5-7). 2 out of the 11 gene clusters are selected as novel gene clusters. One novel gene cluster 5 (fg5_22) overlaps with gene cluster most significantly enriched for cytochrome P450 genes. Enrichment of cytochrome P450 genes is a useful source to screen gene clusters because they have been observed in biosynthetic pathways of secondary metabolites such as the gibberellin [97], sterigmatocystin [28], aflatoxin [176], and trichothecene. Other types of interesting feature such as transcription factors or transporters were not found in the both novel gene clusters but a similar composition of gene functions was also observed in a secondary metabolite gibberellin in *F. fujikuroi* synthesized by at least four cytochrome P450-catalyzed steps [97,212]. Environmental conditions required for secondary metabolism and sporulation were often similar in filamentous fungi. Considering the steeply decreased expression patterns during sexual development, the 2 novel gene clusters as well as aurofusarin mycotoxin may have a selective advantage for protection against adverse environments.

Three properties associated with diverse biological mechanisms

The three properties, co-expression, particular functions, and common regulatory motifs, are important factors related to co-regulation of physically clustered fungal genes. However, these three factors do not always correspond to the same mechanism because each of the properties is associated with various biological processes.

For a relation between co-expression and biological function, only 6 out of 76 (FG1) and 11 out of 147 (FG5) co-expressed gene clusters were assigned to functional characteristics. It may be mostly because of incomplete functional annotation but probably also uncertain relationships between them. Clusters of genes independent of their expression are often used to infer memberships in the same metabolic pathways, stable protein-protein complexes [213], protein interaction [214,215], or other types of co-functionality [77,216,217]. In yeast, many genes in co-expression clusters are functionally related, either belonging to the same FunCat [77] or the same gene ontology (GO) [216]. Inconsistent cases, on the contrary, also reported in other eukaryotes. Drosophila gene expression profiles determined from over 80 experimental conditions showed that it is also possible that neighboring genes which are not functionally related in any obvious way can be identified by expression profiling [218]. Moreover, discrepant cases were reported even in the same species, for example, *Arabidopsis thaliana*. Co-expression analysis of neighboring genes from 233 experiments including various tissues and conditions showed that common functionality is not the main cause for co-expression of neighboring genes in the Arabidopsis genome [219], but other results of co-expressed neighboring genes analyzed from 128 Affymetrix arrays showed shared functional annotations [184].

Co-expressed genes need not always imply co-regulation by the same cis-regulatory motifs. Co-expression of neighboring genes can be attributable to several different reasons such as mechanisms involving gene regulation by the same and/or different transcription factors, chromatin remodelling and transcriptional ripple effect. Chromatin opening to facilitate gene transcription can simultaneously allow transcription of neighboring genes without any known relationship [218,220,221]. A recent study showed that intensive transcription at one locus frequently spills over into its physical neighboring loci, suggesting that transcriptional activation has a ripple effect, which may have advantage for coordinated expression [222]. Conversely, a commonly shared cis-regulatory system may not completely contribute to co-expression of genes. In yeast, analysis of co-expression for adjacent gene pairs and shared transcription factors showed an indistinct relationship [223].

Different arrangement of secreted proteins in *F. graminearum* and *U. maydis*

The arrangement of secreted protein genes in *F. graminearum* differs from the ones found in the biotrophic pathogen *Ustilago maydis* where many clusters have rather short genes coding exclusively for secreted proteins. In *U. maydis*, about 18 percent of secreted proteins are entirely arranged in 25 groups containing 3 to 10 adjacent genes whereas 7 percent of secreted proteins are grouped with only

3 or 4 contiguous genes in *F. graminearum*. The clustered genes of *U. maydis* show no homology to anything outside the *Ustilaginaceae* but are similar to each other implying duplication events inside the clusters. For some clusters it is shown that they are involved in host-pathogen interaction and are proposed to play a role in circumventing host immune response [6]. As the lifestyle of *F. graminearum* is rather necrotrophic, the pathogen doesn't need these emphasized stealth mechanisms and perhaps this is a reason why extended clusters of this type are missing in this organism. Another dissimilar observation between the two species is that secreted protein clusters in *F. graminearum* are often co-localized with genes encoding transcription factors, transporters, and/or diverse enzymes as often shown in metabolite pathway gene clusters. Indeed, two secreted protein clusters overlap with the two mycotoxin gene clusters, aurofusarin (tfcfg_12) and trichothecene (tfcfg_41) (Table 5-7). On the contrary, no such factors are present in neighboring regions of secreted protein clusters in *U. madis*. These differences for arrangement of genes by comparative analyses will help to elucidate evolutionary changes and to predict new types of gene clusters.

Potential necrotrophic gene clusters discovered in *U. maydis*

In the biotrophic pathogen *U. maydis*, the potential necrotrophic gene cluster (msum_11) with the typical composition of gene functions observed in mycotoxin clusters was identified by analysis of conserved motif seeds in neighbouring genes. Gene cluster msum_11 comprising of 17 genes is identified by three motif seeds, 5'-GGGTAA-3', 5'-TTACCC-3', and 5'-GTAn{3}GTT-3', conserved in their promoter regions (Figure 6-2, Table 6-2). The gene cluster msum_11 includes genes encoding 3 PKS, 2 transcription factors, 1 cytochrome P450 and 5 secreted proteins, representing a combinatorial function of genes mostly found in mycotoxin gene clusters (Table 6-2). Occurrence of the cluster msum_11 gives an unprecedented insight into *U. maydis* known to be biotroph. In contrast to necrotrophs, biotrophs are not reported to produce toxins in plants, which rely on living plant tissue. However, our observation by particular composition of gene functions of genes and their regulatory motifs suggests that *U. maydis* may also cause toxicity towards their hosts. Indeed, it was reported possible toxicological effects of *U. mayids* on animals except virally encoded fungal toxins [224,225,226]. *U. maydis* showed neurotoxicity in rats and has possible synergism with Fumonisin B_1 mycotoxin produced mainly by *Fusarium verticillioides* [227]. *U. maydis* may use different nutritional strategies towards different hosts, biotrophic effects in plants and necrotrophic effects in animals.

Notably, the gene cluster msum_11 is located in the telomere of chromosome 12 and a species-specific gene cluster which is probably absent in two other closely related species, *Sporisorium reilianum* and *Ustilago hordei*. Considering simultaneously the flexibility and specialisation of fungi, the cluster msum_11 is probable to be gained by the rapid genomic reorganization caused by diverse reasons such as environmental conditions. *U. maydis* has been served as an instructive model for biotrophic fungal plant pathogens. However, possible mechanisms as a necrotroph are poorly investigated. Here, in silico analyses revealed a potential necrotrophic gene cluster, which is strongly suggested for further functional analysis.

Gene clusters positioned in diverse chromosomal locations

Gene clusters are positioned in diverse chromosomal locations as shown in *F. graminearum* and *U. maydis*. A previous study for *Aspergillus fumigatus* genome described that the secondary metabolite gene clusters unique to the species are dispersed in the genome with a bias towards telomeric locations [228]. Additionally, in several microbial eukaryotes, subtelomeres are highly enriched in genes with roles in niche adaptation. Several fungal species also showed their host and cultivar specificity genes are located near telomere. The subtelomeres of *Saccharomyces cerevisiae* contain the RTM genes conferring resistance to the toxicity of molasses and varying according to the strain's ecology [229]. Host specific genes of pathogenic fungus *Magnaporthe oryzae* are dispersed in heterochromatin with a bias towards telomeric locations [228,230]. In *F. graminearum*, single-nucleotide polymorphism (SNP) analysis from two different strains provided that highly polymorphic regions (HPR) were frequently located near telomeres. HPR contained sets of genes implicated in plant-fungus interactions, which may allow the fungus to adapt rapidly to changing environments or hosts. However, including five known mycotoxin gene clusters, secondary metabolite gene clusters of *F. graminearum* which have no-orthologues in two other *Fusarium* species do not show any position bias on the chromosomes. Also the secreted protein clusters found in the *U. maydis* as well as in the *F. graminearum* are not located towards the telomeres, they are randomly distributed in the genome [6].

Summary

In this study, three types of properties were investigated from known gene clusters in diverse fungi and applied to discover novel gene clusters in *F. graminearum* and *U. maydis*: co-expression, particular composition of gene functions, and conserved regulatory elements in neighboring genes. Co-expression analysis provides experimental evidences of gene regulations and activities under particular conditions. The last two approaches presented here have the advantage that one can increase effective analyses by increasing the number of species studied. Especially, the regulatory motifs-based approach can be very useful when trying to infer computational indications of co-regulation for genes not-yet annotated and/or not detected in co-expression studies as shown in secondary metabolite gene clusters of *F. graminearum* and necrotrophic gene clusters of *U. maydis*. It is also efficient because it can be directly computed once a genome is available whereas the other two approaches previously described need to be conducted as often as new information like updated functional annotation and new set of microarray data are available. As sequencing costs lower and sequencing capacity increases, additional fungal genomes and/or strains from diverse enviroments will become available soon. Although new technologies based on genetic engineering and chromatin-level regulation have been developed to induce silent genes [75,231], in addition efficient in silico screenings of fungal gene clusters suggested here will save time and efforts for functional validations.

From our analyses, beside the re-discovery of five previously identified mycotoxin gene clusters in *F. graminearum*, the screens identified 6 additional genes possibly involved in the synthesis of trichothecene mycotoxin by integration of different properties. Also clusters supported by single evidence might be potential targets for advanced experiments. Taken together, our finding strongly suggest 6 novel gene clusters of *F. graminearum* based on multiple evidences from real gene clusters in fungi. Furthermore, a possible mechanism of co-expression for the secreted protein clusters in *U. maydis* is proposed, which can be co-ordinated by the same regulatory elements. We hope that these specific results will provide convincing and accurate leads towards further studies to discover so far unknown products and/or biological functions in *F. graminearum* and *U. maydis*.

8 Appendix

Appendix A

ID*	Chr.	(N. of gene), gene range	(N. of motif seed), motif seed	Adjusted P-value	N. of gene SM gene	CYP	TF	TP	SP
1	1	(17), FGSG_00036 ~ FGSG_00049	(8), TGGn{3}CAC, TGGn{2}CCA, GTn{3}ACC, GTGn{4}CAC, GGTn{2}CAC, GTGn{2}ACC, GTGnCAC, CCAn{9}CAG	7.7e-08	2 (NPS8, NPS-FGSG_1166)	1		1	
2	1	(14), FGSG_01786 ~ FGSG_01799	(6), CCGCGG, CGGn{9}GCG, CGGn{8}CGC, CGGn{7}CCG, TCCGCG, GCGn{8}CCG	1.0e-06	1 (PKS11)	1	1	1	2
3	3	(12), FGSG_05239 ~ FGSG_05250	(1), GGCn{5}GCC	5.4e-05					2
4	2	(21), FGSG_11532 ~ FGSG_11551	(2), AGCn{3}ACT, TAAn{6}TTA	1.5e-04		1	2		6
5	1	(8), FGSG_01807 ~ FGSG_01814	(1), CGGn{8}CCG	2.2e-04				1	
6	2	(16), FGSG_08916 ~ FGSG_13271	(1), TGTn{4}CGC	2.5e-04			1		2
7	2	(9), FGSG_04238 ~ FGSG_04230	(1), GTGnGAA	3.0e-04					1
8	4	(16), FGSG_15407 ~ FGSG_07020	(1), ACAn{9}CCC	3.4e-04					1
9	3	(13), FGSG_10763 ~ FGSG_15618	(2), CTCn{8}GAG, GCAn{6}TGT	3.8e-04					1
10	1	(12), FGSG_01209 ~ FGSG_01220	(1), CGCn{2}GCG	4.5e-04			1		1
11	3	(8), FGSG_10732 ~ FGSG_13930	(1), GCCn{4}GGC	4.5e-04					1
12	1	(14), FGSG_01986 ~ FGSG_01997	(1), AGTn{2}ACT	4.6e-04				3	4
13	1	(10), FGSG_01901 ~ FGSG_01910	(1), CGCn{8}GCG	4.9e-04				1	
14	2	(15), FGSG_03543 ~ FGSG_03529	(1), AGGCCT	6.1e-04	1 (TRI5-FGSG_03537)	3	2	1	4
15	2	(13), FGSG_04663 ~ FGSG_04653	(1), GTTn{7}AAC	7.5e-04				1	4
16	3	(14), FGSG_05828 ~ FGSG_12823	(1), TCTn{4}CAG	8.4e-04					1
17	1	(10), FGSG_10265 ~ FGSG_13729	(1), GCTn{9}CGC	8.5e-04			1		
18	3	(9), FGSG_05959 ~ FGSG_05966	(1), GCGn{9}AGG	8.8e-04		1			1
19	3	(10), FGSG_10724 ~ FGSG_10716	(1), TAGn{2}GTG	9.0e-04					
20	4	(13), FGSG_09417 ~ FGSG_13578	(1), CGCn{4}GGG	1.0e-03			1		
21	3	(20), FGSG_05031 ~ FGSG_05049	(2), ATAnCAT, GTCTAT	1.1e-03			1	1	1
22	4	(8), FGSG_09195 ~ FGSG_09200	(1), GGCn{9}TAA	1.2e-03		1			
23	2	(19), FGSG_13352 ~ FGSG_08482	(1), CCAnGCT	1.4e-03					3
24	3	(12), FGSG_12914 ~ FGSG_06406	(1), GGGn{2}GCC	1.4e-03					
25	4	(12), FGSG_13604 ~ FGSG_09558	(1), CGGn{5}GCC	1.5e-03					
26	1	(12), FGSG_01761 ~ FGSG_01769	(1), CGAGAC	1.6e-03		1		1	4
27	4	(10), FGSG_07946 ~ FGSG_07953	(1), GTTn{8}AAC	1.6e-03			1		1
28	1	(9), FGSG_02269 ~ FGSG_02275	(1), GTCn{7}GAC	1.7e-03				1	2
29	4	(10), FGSG_06432 ~ FGSG_06441	(2), TTCCGG, CCGGAA	2.1e-03			1		1
30	1	(7), FGSG_00366 ~ FGSG_00370	(1), GCGn{2}CAG	2.1e-03					
31	1	(12), FGSG_10414 ~ FGSG_10425	(1), GGAn{8}TCC	2.2e-03				1	2
32	2	(9), FGSG_04560 ~ FGSG_04553	(1), CGTn{6}CGT	2.3e-03					2

ID*	Chr.	(N. of gene), gene range	(N. of motif seed), motif seed	Adjusted P-value	N. of gene				
					SM gene	CYP	TF	TP	SP
33	2	(11), FGSG_08289 ~ FGSG_08281	(1), ATCn{8}GAT	2.4e-03					2
34	2	(9), FGSG_13295 ~ FGSG_08786	(1), GACn{6}AGC	2.6e-03					1
35	2	(13), FGSG_03177 ~ FGSG_03166	(1), TTTn{3}TTA	2.6e-03				2	3
36	1	(15), FGSG_10623 ~ FGSG_10636	(1), CCAATA	2.7e-03		1	1		3
37	2	(10), FGSG_03068 ~ FGSG_03060	(1), CTTAGA	2.8e-03	1 (TPS-FGSG_3066)		1		
38	3	(22), FGSG_05353 ~ FGSG_05372	(2), AAGn{9}CTT, CACn{3}GTG	3.0e-03	1 (NPS2)		2	1	3
39	2	(19), FGSG_12572 ~ FGSG_02815	(1), ATTn{5}CAT	3.0e-03			1	1	1
40	1	(14), FGSG_02768 ~ FGSG_02779	(1), CCGn{2}CAA	3.1e-03					1
41	4	(16), FGSG_06483 ~ FGSG_15364	(1), ATCn{4}GCG	3.2e-03					
42	1	(21), FGSG_02598 ~ FGSG_02617	(2), TCCTAG, GGCn{9}GCC	3.2e-03			2	1	1
43	2	(10), FGSG_15509 ~ FGSG_08701	(1), GACnGCG	3.5e-03			1		1
44	1	(11), FGSG_01442 ~ FGSG_01452	(2), AGCn{8}GCT, CGCn{6}GCG	3.7e-03					
45	3	(25), FGSG_06092 ~ FGSG_06116	(1), TTTn{8}AAA	3.7e-03					3
46	3	(9), FGSG_05289 ~ FGSG_05297	(1), ACTn{9}GAC	3.9e-03					2
47	4	(16), FGSG_06499 ~ FGSG_06512	(1), CGGn{7}TCT	4.0e-03	1 (NPS1)		1	2	1
48	4	(6), FGSG_06515 ~ FGSG_06520	(1), GGGn{4}CCC	4.1e-03			1		
49	1	(12), FGSG_01272 ~ FGSG_01282	(1), CCGn{7}AAA	4.2e-03					
50	1	(17), FGSG_02505 ~ FGSG_02519	(2), CAGnACG, CTTn{6}AAG	4.3e-03					2
51	4	(7), FGSG_07716 ~ FGSG_07718	(1), GCGn{8}TAT	4.3e-03					
52	2	(14), FGSG_04117 ~ FGSG_04105	(1), CGTnACG	4.3e-03			1		
53	1	(8), FGSG_10401 ~ FGSG_10406	(1), GACn{9}CTA	4.3e-03				1	1
54	3	(17), FGSG_11039 ~ FGSG_11026	(1), TAAn{4}CGA	4.4e-03	1 (NPS1)	1		3	5
55	1	(13), FGSG_01125 ~ FGSG_01137	(1), AAGn{3}CTA	4.5e-03					
56	4	(13), FGSG_06706 ~ FGSG_06715	(1), CCTnACC	4.6e-03			1		1
57	2	(11), FGSG_03940 ~ FGSG_03931	(1), GTATAC	4.7e-03				2	
58	3	(11), FGSG_12588 ~ FGSG_04719	(1), TAGn{2}CTA	4.7e-03		1			
59	1	(24), FGSG_02412 ~ FGSG_02431	(3), GGATTT, CCGCGG, GCAn{6}GTC	4.9e-03		1	2	1	1
60	2	(19), FGSG_04687 ~ FGSG_04672	(1), TATn{5}TTA	4.9e-03			2	2	2
61	4	(13), FGSG_09335 ~ FGSG_09343	(1), GTCn{9}ATC	5.1e-03			1	3	1
62	4	(13), FGSG_09017 ~ FGSG_13502	(1), CGCnTTT	5.2e-03				1	
63	1	(15), FGSG_01004 ~ FGSG_01014	(1), ACGnCGT	5.3e-03				1	1
64	3	(8), FGSG_06152 ~ FGSG_06159	(1), CCAn{7}CGC	5.5e-03			1		
65	3	(9), FGSG_06168 ~ FGSG_06174	(1), GGCn{2}GCC	5.7e-03			1		1
66	1	(13), FGSG_02572 ~ FGSG_15149	(1), GGCn{3}GCC	5.7e-03				1	1
67	1	(6), FGSG_10459 ~ FGSG_10464	(1), CGGn{3}CCG	5.8e-03	1 (PKS9)	1			1
68	3	(12), FGSG_05004 ~ FGSG_12668	(1), CCTn{5}TTT	5.8e-03				1	
69	3	(13), FGSG_04728 ~ FGSG_15123	(1), ATGn{3}AGC	5.9e-03					7
70	1	(11), FGSG_01488 ~ FGSG_01498	(1), CACn{6}TAG	6.1e-03					
71	4	(7), FGSG_13644 ~ FGSG_09782	(1), CGCn{2}GCG	6.4e-03					
72	4	(14), FGSG_09448 ~ FGSG_13591	(1), GCGn{3}CGC	6.4e-03					

ID*	Chr.	(N. of gene), gene range	(N. of motif seed), motif seed	Adjusted P-value	N. of gene SM gene	CYP	TF	TP	SP
73	4	(10), FGSG_07723 ~ FGSG_07729	(1), ATCn{8}AGT	6.6e-03			1		1
74	2	(6), FGSG_11560 ~ FGSG_11564	(1), CGCn{9}GCG	6.9e-03			1		1
75	1	(6), FGSG_02658 ~ FGSG_02663	(1), GCGn{5}CGC	7.0e-03					2
76	1	(8), FGSG_00760 ~ FGSG_00767	(1), GGGn{8}GGG	7.1e-03			1		
77	1	(8), FGSG_00496 ~ FGSG_00502	(1), CAGn{8}TCG	7.1e-03					1
78	1	(17), FGSG_00346 ~ FGSG_00360	(1), GCCCAT	7.2e-03					1
79	4	(12), FGSG_07601 ~ FGSG_07610	(1), ACAn{6}TGT	7.5e-03			2	1	1
80	1	(16), FGSG_00973 ~ FGSG_00983	(1), GGGn{9}GGT	7.7e-03					
81	1	(10), FGSG_00311 ~ FGSG_00319	(1), GGGn{8}ACC	7.7e-03			1		2
82	3	(11), FGSG_06177 ~ FGSG_06184	(1), GTGn{2}CAC	8.2e-03				1	
83	4	(9), FGSG_09268 ~ FGSG_09275	(1), GCTn{6}CGC	8.3e-03					1
84	3	(18), FGSG_10968 ~ FGSG_10952	(1), AGTAAT	8.3e-03		1			5
85	2	(17), FGSG_03849 ~ FGSG_03835	(1), TGTn{5}ACA	8.3e-03				1	3
86	4	(8), FGSG_08977 ~ FGSG_08983	(1), CGGn{5}CCG	8.7e-03					1
87	3	(11), FGSG_13889 ~ FGSG_10920	(1), ACCn{6}AGA	8.9e-03				1	1
88	2	(10), FGSG_08925 ~ FGSG_08923	(1), GCCGCA	8.9e-03			1		
89	3	(12), FGSG_11000 ~ FGSG_10989	(1), GTAn{6}ACT	9.0e-03	1 (NPS9)	1		1	3
90	2	(9), FGSG_02920 ~ FGSG_12554	(1), GGCn{9}GCC	9.1e-03			1		4
91	3	(14), FGSG_05467 ~ FGSG_15300	(1), TCAn{4}GCC	9.2e-03			1		2
92	1	(12), FGSG_00445 ~ FGSG_00456	(1), GTGnCAC	9.4e-03	1 (TPS-FGSG_0451)				2
93	2	(10), FGSG_08595 ~ FGSG_08586	(1), CGGnGAT	9.6e-03					2
94	1	(16), FGSG_02062 ~ FGSG_02075	(1), AAAn{5}CGG	9.8e-03			1	3	2
95	4	(11), FGSG_09850 ~ FGSG_11626	(1), CAGGCA	1.0e-02			1	1	

Table S 8-1: 95 gene clusters with co-occurring motif seeds in *F. graminearum*

In total, about 8.5 percent (1182/13935) of predicted genes in *F. graminearum* are arranged in 95 clusters having 6 to 25 genes with co-occurring regulatory motifs. Detailed gene information of all clusters of co-occurring motif seeds is provided in FGDB.

ID*: msfg_#, NPS: Non-ribosomal peptide synthetase, PKS: Polyketide synthase, TPS: Terpenoid synthase, Chr.: Chromosome, SM: Secondary metabolite, SP: Secreted protein, TF: Transcription factor, TP: Transporter, CYP: Cytochrome P450

ID*	Chr.	(N. of gene), gene range	(N. of motif seed), motif seed	Adjusted P-value	N. of gene				
					SM gene	CYP	TF	TP	SP
1	13	(17), um11549 ~ um04671	(1), ACGnCGT	1.5e-06				1	
2	2	(26), um10242 ~ um11313	(2), ATCn{9}GAT, TTAn{8}TAA	5.1e-05			1	1	
3	17	(10), um04720 ~ um04731	(1), TAGn{6}CTA	7.5e-05					
4	7	(13), um02900 ~ um02912	(1), ACGn{6}CGT	8.3e-05				2	1
5	8	(34), um03143 ~ um10391	(4), TGTn{4}GTG, CGAn{5}CGA, TGTn{4}GTG, GATn{9}CTG	9.6e-05			3	2	2
6	14	(9), um04345 ~ um04353	(2), TTAnGTT, TTAnGTT	1.3e-04				1	1
7	6	(19), um02802 ~ um10949	(1), TCCn{3}GCC	1.6e-04			1	3	2
8	2	(10), um01298.2 ~ um01306	(1), TAGn{3}CTA	1.7e-04					5
9	15	(18), um05042.2 ~ um05059	(3), ACGn{8}CGT, ACGn{8}CGT, GAGnCTC	1.8e-04			1		3
10	14	(10), um04448 ~ um04457	(2), GACn{5}GTC, GACn{5}GTC	1.9e-04	1 (PKS-um10532)				1
11	12	(17), um04095 ~ um04111	(5), GGGTAA, TTACCC, GGGTAA, TTACCC, GTAn{3}GTT	2.8e-04	3 (PKS-um04095, PKS-m04097, PKS-um04105)	1	2		5
12	5	(9), um10165 ~ um02493	(2), TACn{5}CTA, CCCATG	2.8e-04				1	
13	21	(15), um06049 ~ um06063	(2), GGAn{9}TCC, CAGn{2}TGG	3.0e-04					1
14	1	(12), um00013 ~ um10005	(1), TTATAA	3.2e-04			1		3
15	6	(15), um02651 ~ um02665	(1), GTAn{9}TAC	3.3e-04					1
16	4	(17), um05180 ~ um11567.2	(1), GGAn{5}AAG	5.9e-04			1		1
17	17	(19), um04870 ~ um04888	(1), CCAn{8}TGG	6.3e-04			2	1	1
18	2	(16), um01198 ~ um01213	(1), TTTn{9}AAA	8.2e-04					4
19	1	(7), um11434 ~ um00515	(1), TTAn{9}TAA	9.9e-04					
20	11	(24), um11744 ~ um04019	(2), TGGn{6}TCA, CGCn{5}GCG	1.0e-03					
21	2	(25), um11336 ~ um01442	(6), GGTnACC, AGGn{4}ATT, AGCn{8}GGG, TTCn{3}GAG, CTGnTAA, TTTn{3}TGA	1.1e-03		1		3	5
22	4	(11), um11773 ~ um05076	(1), ATGn{8}TCT	1.2e-03		1		1	1
23	2	(13), um01011 ~ um01024	(1), AGGCCT	1.3e-03					3
24	16	(7), um05745 ~ um05750	(1), ATAn{9}TAA	1.5e-03					
25	17	(7), um04808 ~ um10693	(1), CCACTA	1.7e-03				1	
26	2	(12), um12120 ~ um00958	(1), CTCn{7}GGG	1.8e-03			1		
27	6	(10), um02752 ~ um11931	(1), TAGnCTA	1.9e-03			1		3
28	5	(16), um02412 ~ um02426	(1), GCTn{9}CTC	2.2e-03					
29	12	(21), um04184 ~ um04203	(2), TATn{7}CTG, GGTn{5}ACC	2.2e-03		1		1	3
30	2	(8), um10237 ~ um01328	(1), TAAn{9}AAA	2.4e-03					
31	1	(21), um11355 ~ um00392	(2), CCTn{8}ACA, AGCn{7}CTC	2.6e-03			1	1	
32	1	(13), um00620 ~ um00630	(1), AACn{9}ATC	2.7e-03					2
33	21	(12), um06112 ~ um11212	(2), TGTn{5}AGG, TGTn{5}AGG	3.1e-03					6
34	19	(13), um10553 ~ um05309	(1), ATGn{3}GAC	3.7e-03					11
35	19	(7), um05247 ~ um05253	(1), TAAn{7}TGC	3.8e-03				2	
36	12	(7), um12257 ~ um04260	(1), GTGn{6}CTA	3.8e-03					2

ID*	Chr.	(N. of gene), gene range	(N. of motif seed), motif seed	Adjusted P-value	N. of gene				
					SM gene	CYP	TF	TP	SP
37	3	(16), um11685 ~ um11695	(1), TTGn{2}GAG	4.5e-03					1
38	9	(21), um03474 ~ um03491	(3), GGGn{4}CTC, ACACCT, ACACCT	4.5e-03				2	2
39	3	(16), um15093 ~ um01462	(2), AGGn{7}GGA, TCAn{9}TGA	4.6e-03			1		
40	10	(10), um10473 ~ um10477	(1), TACn{8}CAT	4.8e-03		1			1
41	1	(12), um00528 ~ um11443	(1), CCTn{4}CTG	4.9e-03			2	1	3
42	14	(7), um10843 ~ um10845	(1), TTAn{6}TAA	5.0e-03					1
43	3	(12), um01794 ~ um01805	(1), ACCn{6}TTA	5.4e-03					4
44	5	(10), um02323 ~ um11707	(1), GCGn{5}CGC	5.6e-03					
45	2	(20), um11244 ~ um11257.2	(2), GCTAAC, CTCn{9}GAG	5.7e-03			1	2	1
46	9	(7), um03557 ~ um03563	(1), GTCn{6}GTA	5.7e-03			1		
47	23	(20), um06388 ~ um06406	(2), TCCn{3}TGT, GCAn{9}GCA	6.0e-03				1	
48	16	(11), um05728 ~ um05741	(1), AAAn{6}TCT	6.2e-03					3
49	3	(12), um10870 ~ um10872	(1), AACn{2}GTT	6.6e-03				2	
50	20	(15), um05805.2 ~ um11192	(2), AATGCG, ATTACG	6.7e-03					2
51	5	(16), um02189 ~ um02203	(1), AGTn{2}AAG	6.7e-03		1	1		6
52	5	(7), um02134 ~ um02140	(1), CTAn{8}TGA	6.9e-03					5
53	7	(12), um11941 ~ um03023	(1), CGGn{6}CCG	7.0e-03					1
54	1	(16), um00727 ~ um00741	(1), TGCn{7}CGA	7.0e-03					2
55	3	(6), um01671 ~ um11894	(1), TAAn{6}GTC	7.2e-03					
56	15	(11), um04910 ~ um04920	(1), ACGCGT	7.4e-03			1		1
57	22	(14), um06186 ~ um12327	(2), AACn{5}ACG, ACGn{5}ACT	7.6e-03			1		1
58	10	(15), um10484 ~ um10488	(1), AGAn{4}TCT	7.8e-03			1	1	2
59	6	(10), um02596 ~ um02604	(1), ATGn{8}TCT	7.8e-03			2	1	1
60	9	(21), um11064 ~ um03409	(2), CAGn{7}TAT, GGTn{7}TTT	8.0e-03			2	1	2
61	3	(12), um01900 ~ um12174	(1), GTTn{5}CAA	8.2e-03			1		2
62	11	(7), um03881 ~ um03888	(1), ACCCCA	8.5e-03				1	
63	3	(13), um01639 ~ um01651	(1), TCCn{3}CGA	8.5e-03			1		
64	10	(12), um03638 ~ um03649	(1), ATTTTA	8.6e-03					1
65	15	(12), um11771 ~ um04958	(1), GGAnTCC	9.0e-03					1
66	2	(8), um01054 ~ um01061	(1), GATATC	9.3e-03					2
67	21	(9), um05979 ~ um05987	(1), GCAn{6}CCT	9.5e-03					

Table S 8-2: 67 gene clusters with co-occurring motif seeds in *U. maydis*

In total, about 13 percent (910/6782) of predicted genes in *U. maydis* are arranged in 67 clusters having 6 to 34 genes with co-occurring regulatory motifs. Detailed gene information of all clusters of co-occurring motif seeds is provided in MUMDB.

ID*: msum_#, Chr.: Chromosome, SM: Secondary metabolite, PKS: Polyketide synthase, SP: Secreted protein, TF: Transcription factor, TP: Transporter, CYP: Cytochrome P450

Appendix B: Abbreviations

ATM	Aflaterm
ATMT	*Agrobacterium tumefaciens*-mediated transformation
BUT	Butenolide
Chr.	Chromosome
COP	Co-expressed pattern
DON	Deoxynivalenol
EST	Expressed Sequence Tags
Fg	*Fusarium graminearum*
FGDB	*Fusarium graminearum* Genome Database
Fo	*Fusarium oxysporum*
FSTF	Fungal specific transcription factor
FunCat	Functional catalogue
Fv	*Fusarium verticillioides*
GO	Gene ontology
HGT	Horizontal gene transfer
HPR	Highly polymorphic regions
HSTs	Host-selective toxins
MS	Motif seed
MUMDB	MIPS *Ustilago maydis* Database
NIV	Nivalenol
NSBs	non-syntenic blocks
NPS	Non-ribosomal peptide synthetase
CYP	Cytochrome P450
PKS	Polyketide synthase
RBH	Reciprocal Best Hits
REMI	Random restriction enzyme-mediated integration
SBs	syntenic blocks
SIMAP	Similarity Matrix of Proteins
SM	Secondary metabolite
SNP	Single-nucleotide polymorphism
SP	Secreted protein, Secretory protein
SPC	Secreted proteins cluster
TFC	Tentative functional gene cluster
TF	Transcription factor
TP	Transporter
ZEA, ZON	Zearalenone

Appendix C

List of Figures

List of tables

9 References

1. Hurst, L.D., Pál, C. & Lercher, M.J. The evolutionary dynamics of eukaryotic gene order. *Nat Rev Genet* **5**, 299-310 (2004).

2. Lee, J.M. & Sonnhammer, E.L.L. Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res* **13**, 875-82 (2003).

3. Martín, M.F. & Liras, P. Organization and expression of genes involved in the biosynthesis of antibiotics and other secondary metabolites. *Annu. Rev. Microbiol* **43**, 173-206 (1989).

4. Keller & Hohn Metabolic Pathway Gene Clusters in Filamentous Fungi. *Fungal Genet Biol* **21**, 17-29 (1997).

5. Fraser, J.A. u. a. Convergent evolution of chromosomal sex-determining regions in the animal and fungal kingdoms. *PLoS Biol* **2**, e384 (2004).

6. Kämper, J. u. a. Insights from the genome of the biotrophic fungal plant pathogen Ustilago maydis. *Nature* **444**, 97-101 (2006).

7. Watson, A.J., Fuller, L.J., Jeenes, D.J. & Archer, D.B. Homologs of aflatoxin biosynthesis genes and sequence of aflR in Aspergillus oryzae and Aspergillus sojae. *Appl. Environ. Microbiol* **65**, 307-310 (1999).

8. Tominaga, M. u. a. Molecular analysis of an inactive aflatoxin biosynthesis gene cluster in Aspergillus oryzae RIB strains. *Appl. Environ. Microbiol* **72**, 484-490 (2006).

9. Bennett, J.W. & Klich, M. Mycotoxins. *Clin. Microbiol. Rev.* **16**, 497-516 (2003).

10. Bennett, J.W. Mycotoxins, mycotoxicoses, mycotoxicology andMycopathologia. *Mycopathologia* **100**, 3-5 (1987).

11. Díez, B. u. a. The cluster of penicillin biosynthetic genes. Identification and characterization of the pcbAB gene encoding the alpha-aminoadipyl-cysteinyl-valine synthetase and linkage to the pcbC and penDE genes. *J. Biol. Chem* **265**, 16358-16365 (1990).

12. Laich, F., Fierro, F., Cardoza, R.E. & Martin, J.F. Organization of the gene cluster for biosynthesis of penicillin in Penicillium nalgiovense and antibiotic production in cured dry sausages. *Appl. Environ. Microbiol* **65**, 1236-1240 (1999).

13. MacCabe, A.P., Riach, M.B., Unkles, S.E. & Kinghorn, J.R. The Aspergillus nidulans npeA locus consists of three contiguous genes required for penicillin biosynthesis. *EMBO J* **9**, 279-287 (1990).

14. Brakhage, A.A. Molecular regulation of penicillin biosynthesis in Aspergillus (Emericella) nidulans. *FEMS Microbiol. Lett* **148**, 1-10 (1997).

15. Gutiérrez, S., Fierro, F., Casqueiro, J. & Martín, J.F. Gene organization and plasticity of the beta-lactam genes in different filamentous fungi. *Antonie Van Leeuwenhoek* **75**, 81-94 (1999).

16. Abraham, E. Selective reminiscences of beta-lactam antibiotics: early research on penicillin and cephalosporins. *Bioessays* **12**, 601-606 (1990).

17. Mathison, L., Soliday, C., Stepan, T., Aldrich, T. & Rambosek, J. Cloning, characterization, and use in strain improvement of the Cephalosporium acremonium gene cefG encoding acetyl transferase. *Curr. Genet* **23**, 33-41 (1993).

18. Brakhage, A.A. u. a. Regulation of Penicillin Biosynthesis in Filamentous Fungi. *Molecular Biotechnolgy of Fungal beta-Lactam Antibiotics and Related Peptide Synthetases* 45-90 (2004).

19. Tag, A. u. a. G-protein signalling mediates differential production of toxic secondary metabolites. *Mol. Microbiol* **38**, 658-665 (2000).

20. Agrios, G.N. *Plant Pathology*. (Academic Press: 2005).

21. Otani, H., Kohmoto, K. & Kodama, M. Alternaria toxins and their effects on host plants. *Bot.* **73**, 453-458 (1995).

22. Walton, J.D. Host-selective toxins: agents of compatibility. *Plant Cell* **8**, 1723-1733 (1996).

23. Nishimura, S. & Kohmoto, K. Host-Specific Toxins and Chemical Structures from Alternaria Species. *Annu. Rev. Phytopathol.* **21**, 87-116 (1983).

24. Daub, M.E. & Ehrenshaft, M. THE PHOTOACTIVATED CERCOSPORA TOXIN CERCOSPORIN: Contributions to Plant Disease and Fundamental Biology. *Annu Rev Phytopathol* **38**, 461-490 (2000).

25. Daub, M.E., Herrero, S. & Chung, K. Photoactivated perylenequinone toxins in fungal pathogenesis of plants. *FEMS Microbiol. Lett* **252**, 197-206 (2005).

26. Böhnert, H.U. u. a. A putative polyketide synthase/peptide synthetase from Magnaporthe grisea signals pathogen attack to resistant rice. *Plant Cell* **16**, 2499-2513 (2004).

27. Collemare, J. u. a. Magnaporthe grisea avirulence gene ACE1 belongs to an infection-specific gene cluster involved in secondary metabolism. *New Phytol* **179**, 196-208 (2008).

28. Brown, D.W. u. a. Twenty-five coregulated transcripts define a sterigmatocystin gene cluster in Aspergillus nidulans. *Proc Natl Acad Sci U S A* **93**, 1418-22 (1996).

29. Ehrlich, K.C., Montalbano, B.G. & Cary, J.W. Binding of the C6-zinc cluster protein, AFLR, to the promoters of aflatoxin pathway biosynthesis genes in Aspergillus parasiticus. *Gene* **230**, 249-57 (1999).

30. Proctor, R.H., Hohn, T.M., McCormick, S.P. & Desjardins, A.E. Tri6 encodes an unusual zinc finger protein involved in regulation of trichothecene biosynthesis in Fusarium sporotrichioides. *Appl. Environ. Microbiol* **61**, 1923-1930 (1995).

31. Hohn, T.M., Krishna, R. & Proctor, R.H. Characterization of a transcriptional activator controlling trichothecene toxin biosynthesis. *Fungal Genet. Biol* **26**, 224-235 (1999).

32. Dowzer, C.E. & Kelly, J.M. Cloning of the creA gene from Aspergillus nidulans: a gene involved in carbon catabolite repression. *Curr. Genet* **15**, 457-459 (1989).

33. Kudla, B. u. a. The regulatory gene areA mediating nitrogen metabolite repression in Aspergillus nidulans. Mutations affecting specificity of gene activation alter a loop residue of a putative zinc finger. *EMBO J* **9**, 1355-1364 (1990).

34. Tudzynski, B., Homann, V., Feng, B. & Marzluf, G.A. Isolation, characterization and

disruption of the areA nitrogen regulatory gene of Gibberella fujikuroi. *Mol. Gen. Genet* **261**, 106-114 (1999).

35. Tilburn, J. u. a. The Aspergillus PacC zinc finger transcription factor mediates regulation of both acid- and alkaline-expressed genes by ambient pH. *EMBO J* **14**, 779-790 (1995).

36. Espeso, E.A., Tilburn, J., Arst, H.N. & Peñalva, M.A. pH regulation is a major determinant in expression of a fungal penicillin biosynthetic gene. *EMBO J* **12**, 3947-3956 (1993).

37. Tilburn, J. u. a. The Aspergillus PacC zinc finger transcription factor mediates regulation of both acid- and alkaline-expressed genes by ambient pH. *EMBO J* **14**, 779-790 (1995).

38. Keller, N.P., Nesbitt, C., Sarr, B., Phillips, T.D. & Burow, G.B. pH Regulation of Sterigmatocystin and Aflatoxin Biosynthesis in Aspergillus spp. *Phytopathology* **87**, 643-648 (1997).

39. Casadevall, A. & Pirofski, L.A. Host-pathogen interactions: basic concepts of microbial commensalism, colonization, infection, and disease. *Infect. Immun* **68**, 6511-6518 (2000).

40. Wong, S. & Wolfe, K.H. Birth of a metabolic gene cluster in yeast by adaptive gene relocation. *Nat. Genet* **37**, 777-782 (2005).

41. Lohr, D., Venkov, P. & Zlatanova, J. Transcriptional regulation in the yeast GAL gene family: a complex genetic network. *FASEB J* **9**, 777-787 (1995).

42. Johnston, M. A model fungal gene regulatory mechanism: the GAL genes of Saccharomyces cerevisiae. *Microbiol. Rev* **51**, 458-476 (1987).

43. Hittinger, C.T., Rokas, A. & Carroll, S.B. Parallel inactivation of multiple GAL pathway genes and ecological diversification in yeasts. *Proc. Natl. Acad. Sci. U.S.A* **101**, 14144-14149 (2004).

44. Phalip, V., Kuhn, I., Lemoine, Y. & Jeltsch, J.M. Characterization of the biotin biosynthesis pathway in Saccharomyces cerevisiae and evidence for a cluster containing BIO5, a novel gene involved in vitamer uptake. *Gene* **232**, 43-51 (1999).

45. Wu, H., Ito, K. & Shimoi, H. Identification and characterization of a novel biotin biosynthesis gene in Saccharomyces cerevisiae. *Appl. Environ. Microbiol* **71**, 6845-6855 (2005).

46. Hall, C. & Dietrich, F.S. The reacquisition of biotin prototrophy in Saccharomyces cerevisiae involved horizontal gene transfer, gene duplication and gene clustering. *Genetics* **177**, 2293-2307 (2007).

47. Sweigard, J.A. u. a. Identification, cloning, and characterization of PWL2, a gene for host species specificity in the rice blast fungus. *Plant Cell* **7**, 1221-1233 (1995).

48. Tosa, Y. u. a. Evolution of an avirulence gene, AVR1-CO39, concomitant with the evolution and differentiation of Magnaporthe oryzae. *Mol. Plant Microbe Interact* **18**, 1148-1160 (2005).

49. Peyyala, R. & Farman, M.L. Magnaporthe oryzae isolates causing gray leaf spot of perennial ryegrass possess functional copy of the AVR1-CO39 avirulence gene. *Mol. Plant Pathol.* **7**, 157-165 (2006).

50. Malonek, S. u. a. Functional Characterization of Two Cytochrome P450 Monooxygenase Genes, P450-1 and P450-4, of the Gibberellic Acid Gene Cluster in Fusarium proliferatum (Gibberella fujikuroi MP-D). *Appl. Environ. Microbiol.* **71**, 1462-1472 (2005).

51. Malonek, S. u. a. Distribution of gibberellin biosynthetic genes and gibberellin production in the Gibberella fujikuroi species complex. *Phytochemistry* **66**, 1296-1311 (2005).

52. Malonek, S., Rojas, M.C., Hedden, P., Hopkins, P. & Tudzynski, B. Restoration of Gibberellin Production in Fusarium proliferatum by Functional Complementation of Enzymatic Blocks. *Appl. Environ. Microbiol.* **71**, 6014-6025 (2005).

53. Bomke, C., Rojas, M.C., Hedden, P. & Tudzynski, B. Loss of Gibberellin Production in Fusarium verticillioides (Gibberella fujikuroi MP-A) Is Due to a Deletion in the Gibberellic Acid Gene Cluster. *Appl. Environ. Microbiol.* **74**, 7790-7801 (2008).

54. Mullins, E.D. & Kang, S. Transformation: a tool for studying fungal pathogens of plants. *Cell. Mol. Life Sci* **58**, 2043-2052 (2001).

55. Daboussi, M. & Capy, P. Transposable elements in filamentous fungi. *Annu. Rev. Microbiol* **57**, 275-299 (2003).

56. Brown, J.S. & Holden, D.W. Insertional mutagenesis of pathogenic fungi. *Curr. Opin. Microbiol* **1**, 390-394 (1998).

57. Choi, J. u. a. Genome-wide analysis of T-DNA integration into the chromosomes of Magnaporthe oryzae. *Mol. Microbiol* **66**, 371-382 (2007).

58. Mullins, E.D. u. a. Agrobacterium-Mediated Transformation of Fusarium oxysporum: An Efficient Tool for Insertional Mutagenesis and Gene Transfer. *Phytopathology* **91**, 173-180 (2001).

59. Meng, Y. u. a. A systematic analysis of T-DNA insertion events in Magnaporthe oryzae. *Fungal Genet. Biol* **44**, 1050-1064 (2007).

60. Kidwell, M.G. & Lisch, D. Transposable elements as sources of variation in animals and plants. *Proc. Natl. Acad. Sci. U.S.A* **94**, 7704-7711 (1997).

61. Kempken, F. & Kück, U. Tagging of a nitrogen pathway-specific regulator gene in Tolypocladium inflatum by the transposon Restless. *Mol. Gen. Genet* **263**, 302-308 (2000).

62. Weil, C.F. & Kunze, R. Transposition of maize Ac/Ds transposable elements in the yeast Saccharomyces cerevisiae. *Nat. Genet* **26**, 187-190 (2000).

63. Villalba, F., Lebrun, M.H., Hua-Van, A., Daboussi, M.J. & Grosjean-Cournoyer, M.C. Transposon impala, a novel tool for gene tagging in the rice blast fungus Magnaporthe grisea. *Mol. Plant Microbe Interact* **14**, 308-315 (2001).

64. Seong, K., Hou, Z., Tracy, M., Kistler, H.C. & Xu, J. Random Insertional Mutagenesis Identifies Genes Associated with Virulence in the Wheat Scab Fungus Fusarium graminearum. *Phytopathology* **95**, 744-750 (2005).

65. Dufresne, M. u. a. Transposon-tagging identifies novel pathogenicity genes in Fusarium graminearum. *Fungal Genet. Biol* **45**, 1552-1561 (2008).

66. Michielse, C.B., van Wijk, R., Reijnen, L., Cornelissen, B.J.C. & Rep, M. Insight into the

molecular requirements for pathogenicity of Fusarium oxysporum f. sp. lycopersici through large-scale insertional mutagenesis. *Genome Biol* **10**, R4 (2009).

67. López-Berges, M.S. u. a. Identification of virulence genes in Fusarium oxysporum f. sp. lycopersici by large-scale transposon tagging. *Mol. Plant Pathol* **10**, 95-107 (2009).

68. Bok, J.W. & Keller, N.P. LaeA, a regulator of secondary metabolism in Aspergillus spp. *Eukaryotic Cell* **3**, 527-535 (2004).

69. Nancy Keller, Jinwoo Bok, Dawoon Chung, Robyn M. Perrin & Elliot Keats Shwab LaeA, a global regulator of Aspergillus toxins. (2009).

70. Bok, J.W. u. a. Genomic mining for Aspergillus natural products. *Chem. Biol* **13**, 31-37 (2006).

71. Kosalková, K. u. a. The global regulator LaeA controls penicillin biosynthesis, pigmentation and sporulation, but not roquefortine C synthesis in Penicillium chrysogenum. *Biochimie* **91**, 214-225 (2009).

72. Jenuwein, T. & Allis, C.D. Translating the histone code. *Science* **293**, 1074-1080 (2001).

73. Keller, N.P., Turner, G. & Bennett, J.W. Fungal secondary metabolism - from biochemistry to genomics. *Nat. Rev. Microbiol* **3**, 937-947 (2005).

74. Shwab, E.K. u. a. Histone deacetylase activity regulates chemical diversity in Aspergillus. *Eukaryotic Cell* **6**, 1656-1664 (2007).

75. Bok, J.W. u. a. Chromatin-level regulation of biosynthetic gene clusters. *Nat. Chem. Biol* **5**, 462-464 (2009).

76. Cho, R.J. u. a. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* **2**, 65-73 (1998).

77. Cohen, B.A., Mitra, R.D., Hughes, J.D. & Church, G.M. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat. Genet* **26**, 183-186 (2000).

78. Kuo, W.P., Jenssen, T., Butte, A.J., Ohno-Machado, L. & Kohane, I.S. Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics* **18**, 405-412 (2002).

79. Li, J., Pankratz, M. & Johnson, J.A. Differential gene expression patterns revealed by oligonucleotide versus long cDNA arrays. *Toxicol. Sci* **69**, 383-390 (2002).

80. Lee, J.M. & Sonnhammer, E.L.L. Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res* **13**, 875-882 (2003).

81. Güldener, U. u. a. FGDB: a comprehensive fungal genome resource on the plant pathogen Fusarium graminearum. *Nucleic Acids Res* **34**, D456-458 (2006).

82. Gardiner, D.M., Waring, P. & Howlett, B.J. The epipolythiodioxopiperazine (ETP) class of fungal toxins: distribution, mode of action, functions and biosynthesis. *Microbiology (Reading, Engl.)* **151**, 1021-1032 (2005).

83. Gardiner, D.M., Cozijnsen, A.J., Wilson, L.M., Pedras, M.S.C. & Howlett, B.J. The sirodesmin biosynthetic gene cluster of the plant pathogenic fungus Leptosphaeria maculans.

Mol. Microbiol **53**, 1307-1318 (2004).

84. Gardiner, D.M. & Howlett, B.J. Bioinformatic and expression analysis of the putative gliotoxin biosynthetic gene cluster of Aspergillus fumigatus. *FEMS Microbiol. Lett* **248**, 241-248 (2005).

85. Soanes, D.M. u. a. Comparative genome analysis of filamentous fungi reveals gene family expansions associated with fungal pathogenesis. *PLoS ONE* **3**, e2300 (2008).

86. Fitzpatrick, D.A., Logue, M.E., Stajich, J.E. & Butler, G. A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. *BMC Evol. Biol* **6**, 99 (2006).

87. Yu, J., Bhatnagar, D. & Cleveland, T.E. Completed sequence of aflatoxin pathway gene cluster in Aspergillus parasiticus. *FEBS Lett* **564**, 126-30 (2004).

88. Frandsen, R.J.N. u. a. The biosynthetic pathway for aurofusarin in Fusarium graminearum reveals a close link between the naphthoquinones and naphthopyrones. *Mol Microbiol* **61**, 1069-80 (2006).

89. Chiang, Y. u. a. A Gene Cluster Containing Two Fungal Polyketide Synthases Encodes the Biosynthetic Pathway for a Polyketide, Asperfuranone, in Aspergillus nidulans. *J Am Chem Soc* (2009).doi:10.1021/ja8088185

90. Harris, L.J. u. a. A novel gene cluster in Fusarium graminearum contains a gene that contributes to butenolide synthesis. *Fungal Genet Biol* **44**, 293-306 (2007).

91. Chen, Y. u. a. Exploring the distribution of citrinin biosynthesis related genes among Monascus species. *J. Agric. Food Chem* **56**, 11767-11772 (2008).

92. Sakai, K., Kinoshita, H., Shimizu, T. & Nihira, T. Construction of a citrinin gene cluster expression system in heterologous Aspergillus oryzae. *J Biosci Bioeng* **106**, 466-72 (2008).

93. Tsai, H.F., Wheeler, M.H., Chang, Y.C. & Kwon-Chung, K.J. A developmentally regulated gene cluster involved in conidial pigment biosynthesis in Aspergillus fumigatus. *J. Bacteriol* **181**, 6469-6477 (1999).

94. Sims, J.W., Fillmore, J.P., Warner, D.D. & Schmidt, E.W. Equisetin biosynthesis in Fusarium heterosporum. *Chem. Commun. (Camb.)* 186-188 (2005).doi:10.1039/b413523g

95. Haarmann, T. u. a. The ergot alkaloid gene cluster in Claviceps purpurea: Extension of the cluster sequence and intra species evolution. *Phytochemistry* **66**, 1312-1320 (2005).

96. Proctor, R.H., Brown, D.W., Plattner, R.D. & Desjardins, A.E. Co-expression of 15 contiguous genes delineates a fumonisin biosynthetic gene cluster in Gibberella moniliformis. *Fungal Genet Biol* **38**, 237-49 (2003).

97. Tudzynski, B. Gibberellin biosynthesis in fungi: genes, enzymes, evolution, and impact on biotechnology. *Appl Microbiol Biotechnol* **66**, 597-611 (2005).

98. Fox, E.M. & Howlett, B.J. Biosynthetic gene clusters for epipolythiodioxopiperazines in filamentous fungi. *Mycol. Res* **112**, 162-169 (2008).

99. Gardiner, D.M., Waring, P. & Howlett, B.J. The epipolythiodioxopiperazine (ETP) class of fungal toxins: distribution, mode of action, functions and biosynthesis. *Microbiology*

(Reading, Engl.) **151**, 1021-1032 (2005).

100. Kennedy, J. u. a. Modulation of polyketide synthase activity by accessory proteins during lovastatin biosynthesis. *Science* **284**, 1368-1372 (1999).

101. Bouhired, S., Weber, M., Kempf-Sontag, A., Keller, N.P. & Hoffmeister, D. Accurate prediction of the Aspergillus nidulans terrequinone gene cluster boundaries using the transcriptional regulator LaeA. *Fungal Genet. Biol* **44**, 1134-1145 (2007).

102. Brown, D.W., Dyer, R.B., McCormick, S.P., Kendra, D.F. & Plattner, R.D. Functional demarcation of the Fusarium core trichothecene gene cluster. *Fungal Genet Biol* **41**, 454-62 (2004).

103. Lysøe, E., Bone, K.R. & Klemsdal, S.S. Real-time quantitative expression studies of the zearalenone biosynthetic gene cluster in Fusarium graminearum. *Phytopathology* **99**, 176-184 (2009).

104. Chang, P.K., Ehrlich, K.C., Yu, J., Bhatnagar, D. & Cleveland, T.E. Increased expression of Aspergillus parasiticus aflR, encoding a sequence-specific DNA-binding protein, relieves nitrate inhibition of aflatoxin biosynthesis. *Appl Environ Microbiol* **61**, 2372-7 (1995).

105. Price, M.S. u. a. The aflatoxin pathway regulator AflR induces gene transcription inside and outside of the aflatoxin biosynthetic cluster. *FEMS Microbiol Lett* **255**, 275-9 (2006).

106. Kim, J. u. a. GIP2, a putative transcription factor that regulates the aurofusarin biosynthetic gene cluster in Gibberella zeae. *Appl Environ Microbiol* **72**, 1645-52 (2006).

107. Flaherty, J.E. & Woloshuk, C.P. Regulation of fumonisin biosynthesis in Fusarium verticillioides by a zinc binuclear cluster-type gene, ZFR1. *Appl Environ Microbiol* **70**, 2653-9 (2004).

108. Bok, J.W. u. a. GliZ, a transcriptional regulator of gliotoxin biosynthesis, contributes to Aspergillus fumigatus virulence. *Infect Immun* **74**, 6761-8 (2006).

109. Fernandes, M., Keller, N.P. & Adams, T.H. Sequence-specific binding by Aspergillus nidulans AflR, a C6 zinc cluster protein regulating mycotoxin biosynthesis. *Mol Microbiol* **28**, 1355-65 (1998).

110. Hohn, T.M., Krishna, R. & Proctor, R.H. Characterization of a transcriptional activator controlling trichothecene toxin biosynthesis. *Fungal Genet Biol* **26**, 224-35 (1999).

111. Cuomo, C.A. u. a. The Fusarium graminearum genome reveals a link between localized polymorphism and pathogen specialization. *Science* **317**, 1400-2 (2007).

112. Gaffoor, I. u. a. Functional analysis of the polyketide synthase genes in the filamentous fungus Gibberella zeae (anamorph Fusarium graminearum). *Eukaryotic Cell* **4**, 1926-1933 (2005).

113. Oide, S. u. a. 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 (2006).

114. Tobiasen, C. u. a. Nonribosomal peptide synthetase (NPS) genes in Fusarium graminearum, F. culmorum and F. pseudograminearium and identification of NPS2 as the producer of ferricrocin. *Curr. Genet* **51**, 43-58 (2007).

115. Oide, S., Krasnoff, S.B., Gibson, D.M. & Turgeon, B.G. Intracellular siderophores are essential for ascomycete sexual development in heterothallic Cochliobolus heterostrophus and homothallic Gibberella zeae. *Eukaryotic Cell* **6**, 1339-1353 (2007).

116. Malz, S. u. a. Identification of a gene cluster responsible for the biosynthesis of aurofusarin in the Fusarium graminearum species complex. *Fungal Genet. Biol* **42**, 420-433 (2005).

117. Kim, J. u. a. GIP2, a putative transcription factor that regulates the aurofusarin biosynthetic gene cluster in Gibberella zeae. *Appl. Environ. Microbiol* **72**, 1645-1652 (2006).

118. Frandsen, R.J.N. u. a. The biosynthetic pathway for aurofusarin in Fusarium graminearum reveals a close link between the naphthoquinones and naphthopyrones. *Mol. Microbiol* **61**, 1069-1080 (2006).

119. Gelderblom, W.C.A., Thiel, P.G., Marasas, W.F.O. & Van der Merwe, K.J. Natural occurrence of fusarin C, a mutagen produced by Fusarium moniliforme, in corn. *Journal of Agricultural and Food Chemistry* **32**, 1064-1067 (1984).

120. Bacon, C.W., Marijanovic, D.R., Norred, W.P. & Hinton, D.M. Production of fusarin C on cereal and soybean by Fusarium moniliforme. *Appl. Environ. Microbiol* **55**, 2745-2748 (1989).

121. Kim, Y. u. a. Two different polyketide synthase genes are required for synthesis of zearalenone in Gibberella zeae. *Mol. Microbiol* **58**, 1102-1113 (2005).

122. Basse C.W. & Steinberg G. Ustilago maydis, model system for analysis of the molecular basis of fungal pathogenicity. *Molecular Plant Pathology* **5**, 83-92 (2004).

123. Bölker, M. Ustilago maydis--a valuable model system for the study of fungal dimorphism and virulence. *Microbiology (Reading, Engl.)* **147**, 1395-1401 (2001).

124. Martínez-Espinoza, A.D., García-Pedrajas, M.D. & Gold, S.E. The Ustilaginales as plant pests and model systems. *Fungal Genet. Biol* **35**, 1-20 (2002).

125. Hunter, S. u. a. InterPro: the integrative protein signature database. *Nucleic Acids Res* **37**, D211-215 (2009).

126. Emanuelsson, O., Nielsen, H., Brunak, S. & von Heijne, G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol* **300**, 1005-1016 (2000).

127. Güldener, U. u. a. Development of a Fusarium graminearum Affymetrix GeneChip for profiling fungal gene expression in vitro and in planta. *Fungal Genet Biol* **43**, 316-25 (2006).

128. Hallen, H.E., Huebner, M., Shiu, S., Güldener, U. & Trail, F. Gene expression shifts during perithecium development in Gibberella zeae (anamorph Fusarium graminearum), with particular emphasis on ion transport proteins. *Fungal Genet Biol* **44**, 1146-56 (2007).

129. Irizarry, R.A. u. a. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249-64 (2003).

130. Walter, M.C. u. a. PEDANT covers all complete RefSeq genomes. *Nucleic Acids Res* **37**, D408-11 (2009).

131. Manly, K.F., Nettleton, D. & Hwang, J.T.G. Genomics, prior probability, and statistical tests

of multiple hypotheses. *Genome Res* **14**, 997-1001 (2004).

132. Hu, Z., Frith, M., Niu, T. & Weng, Z. SeqVISTA: a graphical tool for sequence feature visualization and comparison. *BMC Bioinformatics* **4**, 1 (2003).

133. Rattei, T. u. a. SIMAP--structuring the network of protein similarities. *Nucleic Acids Res* **36**, D289-92 (2008).

134. Desjardins, A.E., Hohn, T.M. & McCormick, S.P. Trichothecene biosynthesis in Fusarium species: chemistry, genetics, and significance. *Microbiol Rev* **57**, 595-604 (1993).

135. Yoshizawa, T. & Jin, Y.Z. Natural occurrence of acetylated derivatives of deoxynivalenol and nivalenol in wheat and barley in Japan. *Food Addit Contam* **12**, 689-694 (1995).

136. Quarta, A. u. a. Assessment of trichothecene chemotypes of Fusarium culmorum occurring in Europe. *Food Addit Contam* **22**, 309-315 (2005).

137. Quarta, A. u. a. Multiplex PCR assay for the identification of nivalenol, 3- and 15-acetyl-deoxynivalenol chemotypes in Fusarium. *FEMS Microbiol. Lett* **259**, 7-13 (2006).

138. Zhang, J. u. a. Determination of the trichothecene mycotoxin chemotypes and associated geographical distribution and phylogenetic species of the Fusarium graminearum clade from China. *Mycol. Res* **111**, 967-975 (2007).

139. W. F. O. Marasas *Toxigenic Fusarium species, identity and mycotoxicology*. (Pennsylvania State University Press: University Park, 1984).

140. Ueno, Y., Nakajima, M., Sakai, K., Ishii, K. & Sato, N. Comparative toxicology of trichothec mycotoxins: inhibition of protein synthesis in animal cells. *J. Biochem* **74**, 285-296 (1973).

141. Cundliffe, E., Cannon, M. & Davies, J. Mechanism of inhibition of eukaryotic protein synthesis by trichothecene fungal toxins. *Proc. Natl. Acad. Sci. U.S.A* **71**, 30-34 (1974).

142. Kimura, M. u. a. The trichothecene biosynthesis gene cluster of Fusarium graminearum F15 contains a limited number of essential pathway genes and expressed non-essential genes. *FEBS Lett* **539**, 105-110 (2003).

143. Meek, I.B., Peplow, A.W., Ake, C., Phillips, T.D. & Beremand, M.N. Tri1 encodes the cytochrome P450 monooxygenase for C-8 hydroxylation during trichothecene biosynthesis in Fusarium sporotrichioides and resides upstream of another new Tri gene. *Appl. Environ. Microbiol* **69**, 1607-1613 (2003).

144. McCormick, S.P. u. a. Tri1 in Fusarium graminearum encodes a P450 oxygenase. *Appl. Environ. Microbiol* **70**, 2044-2051 (2004).

145. Brown, D.W., Proctor, R.H., Dyer, R.B. & Plattner, R.D. Characterization of a fusarium 2-gene cluster involved in trichothecene C-8 modification. *J. Agric. Food Chem* **51**, 7936-7944 (2003).

146. Yates, S.G., Tookey, H.L., Ellis, J.J. & Burkhardt, H.J. Toxic butenolide produced by (fries) cesati isolated from tall fescue (schreb.). *Tetrahedron Letters* **8**, 621-625 (1967).

147. Tookey, H.L., Yates, S.G., Ellis, J.J., Grove, M.D. & Nichols, R.E. Toxic effects of a butenolide mycotoxin and of Fusarium tricinctum cultures in cattle. *J. Am. Vet. Med. Assoc*

160, 1522-1526 (1972).

148. Wang, H., Wang, Y. & Peng, S. Repeated administration of a Fusarium mycotoxin butenolide to rats induces hepatic lipid peroxidation and antioxidant defense impairment. *Food and Chemical Toxicology* **47**, 633-637 (2009).

149. Wang, Y.[., Liu, J.[. & Peng, S.[. Effects of Fusarium Mycotoxin Butenolide on Myocardial Mitochondria In Vitro. *Toxicology Mechanisms and Methods* **19**, 79-85 (2009).

150. Daws, M., Davies, J., Pritchard, H., Brown, N. & Van Staden, J. Butenolide from plant-derived smoke enhances germination and seedling growth of arable weed species. *Plant Growth Regulation* **51**, 73-82 (2007).

151. Harris, L.J. u. a. Possible Role of Trichothecene Mycotoxins in Virulence of Fusarium graminearum on Maize. *Plant Disease* **83**, 954-960 (1999).

152. Medentsev, A.G. & Akimenko, V.K. Naphthoquinone metabolites of the fungi. *Phytochemistry* **47**, 935-959 (1998).

153. Dvorska, J.E., Surai, P.F., Speake, B.K. & Sparks, N.H. Effect of the mycotoxin aurofusarin on the antioxidant composition and fatty acid profile of quail eggs. *Br. Poult. Sci* **42**, 643-649 (2001).

154. Dvorska, J.E., Surai, P.F., Speake, B.K. & Sparks, N.H.C. Antioxidant systems of the developing quail embryo are compromised by mycotoxin aurofusarin. *Comp. Biochem. Physiol. C Toxicol. Pharmacol* **131**, 197-205 (2002).

155. Caldwell, R.W., Tuite, J., Stob, M. & Baldwin, R. Zearalenone production by Fusarium species. *Appl Microbiol* **20**, 31-34 (1970).

156. Kuiper-Goodman, T., Scott, P.M. & Watanabe, H. Risk assessment of the mycotoxin zearalenone. *Regul. Toxicol. Pharmacol* **7**, 253-306 (1987).

157. STOB, M., BALDWIN, R.S., TUITE, J., ANDREWS, F.N. & GILLETTE, K.G. Isolation of an anabolic, uterotrophic compound from corn infected with Gibberella zeae. *Nature* **196**, 1318 (1962).

158. Wolf, J.C. & Mirocha, C.J. Control of Sexual Reproduction in Gibberella zeae (Fusarium roseum "Graminearum"). *Appl. Environ. Microbiol* **33**, 546-550 (1977).

159. Placinta, C.M., D'Mello, J.P.F. & Macdonald, A.M.C. A review of worldwide contamination of cereal grains and animal feed with Fusarium mycotoxins. *Animal Feed Science and Technology* **78**, 21-37 (1999).

160. Wolf, J.C. & Mirocha, C.J. Regulation of sexual reproduction in Gibberella zeae (Fusarium roxeum "graminearum") by F-2 (Zearalenone). *Can. J. Microbiol* **19**, 725-734 (1973).

161. El-Nezami, H., Polychronaki, N., Salminen, S. & Mykkänen, H. Binding rather than metabolism may explain the interaction of two food-Grade Lactobacillus strains with zearalenone and its derivative (')alpha-earalenol. *Appl. Environ. Microbiol* **68**, 3545-3549 (2002).

162. Pfohl-Leszkowicz, A., Chekir-Ghedira, L. & Bacha, H. Genotoxicity of zearalenone, an estrogenic mycotoxin: DNA adduct formation in female mouse tissues. *Carcinogenesis* **16**, 2315-2320 (1995).

163. Gaffoor, I. & Trail, F. Characterization of two polyketide synthase genes involved in zearalenone biosynthesis in Gibberella zeae. *Appl. Environ. Microbiol* **72**, 1793-1799 (2006).

164. Gelderblom, W.C.A. u. a. Structure elucidation of fusarin C, a mutagen produced by Fusarium moniliforme. *J. Chem. Soc., Chem. Commun.* 122-124 (1984).

165. Thiel, P.G., Gelderblom, W.C.A., Marasas, W.F.O., Nelson, P.E. & Wilson, T.M. Natural occurrence of moniliformin and fusarin C in corn screenings known to be hepatocarcinogenic in rats. *Journal of Agricultural and Food Chemistry* **34**, 773-775 (1986).

166. Song, Z., Cox, R.J., Lazarus, C.M. & Simpson TJ, T.J. Fusarin C biosynthesis in Fusarium moniliforme and Fusarium venenatum. *Chembiochem* **5**, 1196-1203 (2004).

167. Cheng, S.J., Jiang, Y.Z., Li, M.H. & Lo, H.Z. A mutagenic metabolite produced by Fusarium moniliforme isolated from Linxian county, China. *Carcinogenesis* **6**, 903-905 (1985).

168. Malz, S. u. a. Identification of a gene cluster responsible for the biosynthesis of aurofusarin in the Fusarium graminearum species complex. *Fungal Genet Biol* **42**, 420-33 (2005).

169. Cramer, R.A. u. a. Disruption of a nonribosomal peptide synthetase in Aspergillus fumigatus eliminates gliotoxin production. *Eukaryotic Cell* **5**, 972-980 (2006).

170. Ehrlich, K.C., Yu, J. & Cotty, P.J. Aflatoxin biosynthesis gene clusters and flanking regions. *J. Appl. Microbiol* **99**, 518-527 (2005).

171. Ueno, T. u. a. Isolation of AM-toxin I, a new phytotoxic metabolite from Alternaria mali. *Phytopathology* **65**, 82-83 (1975).

172. Miyashita, M., Nakamori, T., Miyagawa, H., Akamatsu, M. & Ueno, T. Inhibitory activity of analogs of AM-toxin, a host-specific phytotoxin from the Alternaria alternata apple pathotype, on photosynthetic O2 evolution in apple leaves. *Biosci. Biotechnol. Biochem* **67**, 635-638 (2003).

173. Tsai, H.F., Wheeler, M.H., Chang, Y.C. & Kwon-Chung, K.J. A developmentally regulated gene cluster involved in conidial pigment biosynthesis in Aspergillus fumigatus. *J. Bacteriol* **181**, 6469-6477 (1999).

174. Calvo, A.M., Wilson, R.A., Bok, J.W. & Keller, N.P. Relationship between secondary metabolism and fungal development. *Microbiol. Mol. Biol. Rev* **66**, 447-459, table of contents (2002).

175. van den Brink, H.M., van Gorcom, R.F., van den Hondel, C.A. & Punt, P.J. Cytochrome P450 enzyme systems in fungi. *Fungal Genet. Biol* **23**, 1-17 (1998).

176. Yu, J., Chang, P.K., Cary, J.W., Bhatnagar, D. & Cleveland, T.E. avnA, a gene encoding a cytochrome P-450 monooxygenase, is involved in the conversion of averantin to averufin in aflatoxin biosynthesis in Aspergillus parasiticus. *Appl. Environ. Microbiol* **63**, 1349-1356 (1997).

177. Linnemannstöns, P. u. a. The polyketide synthase gene pks4 from Gibberella fujikuroi encodes a key enzyme in the biosynthesis of the red pigment bikaverin. *Fungal Genetics and Biology* **37**, 134-148 (2002).

178. Lee, B. u. a. Functional analysis of all nonribosomal peptide synthetases in Cochliobolus

heterostrophus reveals a factor, NPS6, involved in virulence and resistance to oxidative stress. *Eukaryotic Cell* **4**, 545-555 (2005).

179. Bendtsen, J.D., Nielsen, H., von Heijne, G. & Brunak, S. Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol* **340**, 783-795 (2004).

180. Kundrotas, P.J. & Alexov, E. PROTCOM: searchable database of protein complexes enhanced with domain–domain structures. *Nucleic Acids Res* **35**, D575-D579 (2007).

181. Basse, C.W., Kolb, S. & Kahmann, R. A maize-specifically expressed gene cluster in Ustilago maydis. *Mol. Microbiol* **43**, 75-93 (2002).

182. Basse, C.W., Stumpferl, S. & Kahmann, R. Characterization of a Ustilago maydis gene specifically induced during the biotrophic phase: evidence for negative as well as positive regulation. *Mol. Cell. Biol* **20**, 329-339 (2000).

183. Roy, P.J., Stuart, J.M., Lund, J. & Kim, S.K. Chromosomal clustering of muscle-expressed genes in Caenorhabditis elegans. *Nature* **418**, 975-979 (2002).

184. Zhan, S., Horrocks, J. & Lukens, L.N. Islands of co-expressed neighbouring genes in Arabidopsis thaliana suggest higher-order chromosome domains. *Plant J* **45**, 347-357 (2006).

185. Boutanaev, A.M., Kalmykova, A.I., Shevelyov, Y.Y. & Nurminsky, D.I. Large clusters of co-expressed genes in the Drosophila genome. *Nature* **420**, 666-669 (2002).

186. Soury, E. u. a. Chromosomal assignments of mammalian genes with an acute inflammation-regulated expression in liver. *Immunogenetics* **53**, 634-642 (2001).

187. Shwab, E.K. u. a. Histone deacetylase activity regulates chemical diversity in Aspergillus. *Eukaryotic Cell* **6**, 1656-1664 (2007).

188. Roze, L.V., Arthur, A.E., Hong, S., Chanda, A. & Linz, J.E. The initiation and pattern of spread of histone H4 acetylation parallel the order of transcriptional activation of genes in the aflatoxin cluster. *Mol. Microbiol* **66**, 713-726 (2007).

189. Kalmykova, A.I., Nurminsky, D.I., Ryzhov, D.V. & Shevelyov, Y.Y. Regulated chromatin domain comprising cluster of co-expressed genes in Drosophila melanogaster. *Nucleic Acids Res* **33**, 1435-1444 (2005).

190. Aharonowitz, Y., Cohen, G. & Martin, J.F. Penicillin and cephalosporin biosynthetic genes: structure, organization, regulation, and evolution. *Annu. Rev. Microbiol* **46**, 461-495 (1992).

191. Liras, P. & Martín, J.F. Gene clusters for beta-lactam antibiotics and control of their expression: why have clusters evolved, and from where did they originate? *Int. Microbiol* **9**, 9-19 (2006).

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

193. Wolf, J.C. & Mirocha, C.J. Regulation of sexual reproduction in Gibberella zeae (Fusarium roxeum "graminearum") by F-2 (Zearalenone). *Can. J. Microbiol* **19**, 725-734 (1973).

194. Lysøe, E. u. a. The PKS4 gene of Fusarium graminearum is essential for zearalenone production. *Appl. Environ. Microbiol* **72**, 3924-3932 (2006).

195. Arst, H.N. & MacDonald, D.W. A gene cluster in Aspergillus nidulans with an internally

located cis-acting regulatory region. *Nature* **254**, 26-31 (1975).

196. Sophianopoulou, V., Suárez, T., Diallinas, G. & Scazzocchio, C. Operator derepressed mutations in the proline utilisation gene cluster of Aspergillus nidulans. *Mol. Gen. Genet* **236**, 209-213 (1993).

197. Brown, D.W., McCormick, S.P., Alexander, N.J., Proctor, R.H. & Desjardins, A.E. A genetic and biochemical approach to study trichothecene diversity in Fusarium sporotrichioides and Fusarium graminearum. *Fungal Genet. Biol* **32**, 121-133 (2001).

198. Lee, T., Han, Y., Kim, K., Yun, S. & Lee, Y. Tri13 and Tri7 determine deoxynivalenol- and nivalenol-producing chemotypes of Gibberella zeae. *Appl. Environ. Microbiol* **68**, 2148-2154 (2002).

199. Kimura, M. u. a. Trichothecene 3-O-acetyltransferase protects both the producing organism and transformed yeast from related mycotoxins. Cloning and characterization of Tri101. *J. Biol. Chem* **273**, 1654-1661 (1998).

200. Proctor, R.H., McCormick, S.P., Alexander, N.J. & Desjardins, A.E. Evidence that a secondary metabolic biosynthetic gene cluster has grown by gene relocation during evolution of the filamentous fungus Fusarium. *Mol. Microbiol* (2009).doi:10.1111/j.1365-2958.2009.06927.x

201. McCormick, S.P., Alexander, N.J. & Proctor, R.H. Heterologous expression of two trichothecene P450 genes in Fusarium verticillioides. *Can. J. Microbiol* **52**, 220-226 (2006).

202. Kodama, M. u. a. The translocation-associated tox1 locus of Cochliobolus heterostrophus is two genetic elements on two different chromosomes. *Genetics* **151**, 585-596 (1999).

203. Rose, M.S. u. a. A decarboxylase encoded at the Cochliobolus heterostrophus translocation-associated Tox1B locus is required for polyketide (T-toxin) biosynthesis and high virulence on T-cytoplasm maize. *Mol. Plant Microbe Interact* **15**, 883-893 (2002).

204. Baker, S.E. u. a. Two polyketide synthase-encoding genes are required for biosynthesis of the polyketide virulence factor, T-toxin, by Cochliobolus heterostrophus. *Mol. Plant Microbe Interact* **19**, 139-149 (2006).

205. Nicholson, M.J. u. a. Identification of two aflatrem biosynthesis gene loci in Aspergillus flavus and metabolic engineering of Penicillium paxilli to elucidate their function. *Appl. Environ. Microbiol* **75**, 7469-7481 (2009).

206. Saikia, S., Parker, E.J., Koulman, A. & Scott, B. Defining paxilline biosynthesis in Penicillium paxilli: functional characterization of two cytochrome P450 monooxygenases. *J. Biol. Chem* **282**, 16829-16837 (2007).

207. Cary, J.W. & Ehrlich, K.C. Aflatoxigenicity in Aspergillus: molecular genetics, phylogenetic relationships and evolutionary implications. *Mycopathologia* **162**, 167-177 (2006).

208. Kimura, M., Tokai, T., Takahashi-Ando, N., Ohsato, S. & Fujimura, M. Molecular and genetic studies of fusarium trichothecene biosynthesis: pathways, genes, and evolution. *Biosci. Biotechnol. Biochem* **71**, 2105-2123 (2007).

209. Alspaugh, J.A., Perfect, J.R. & Heitman, J. Cryptococcus neoformans mating and virulence are regulated by the G-protein alpha subunit GPA1 and cAMP. *Genes Dev* **11**, 3206-3217

(1997).

210. Langfelder, K., Streibel, M., Jahn, B., Haase, G. & Brakhage, A.A. Biosynthesis of fungal melanins and their importance for human pathogenic fungi. *Fungal Genet. Biol* **38**, 143-158 (2003).

211. Wang, Y., Aisen, P. & Casadevall, A. Cryptococcus neoformans melanin and virulence: mechanism of action. *Infect. Immun* **63**, 3131-3136 (1995).

212. Kawaide, H. Biochemical and molecular analyses of gibberellin biosynthesis in fungi. *Biosci. Biotechnol. Biochem* **70**, 583-590 (2006).

213. Teichmann, S.A. & Veitia, R.A. Genes encoding subunits of stable complexes are clustered on the yeast chromosomes: an interpretation from a dosage balance perspective. *Genetics* **167**, 2121-2125 (2004).

214. Ge, H., Liu, Z., Church, G.M. & Vidal, M. Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nat. Genet* **29**, 482-486 (2001).

215. Grigoriev, A. A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae. *Nucl. Acids Res.* **29**, 3513-3519 (2001).

216. Fukuoka, Y., Inaoka, H. & Kohane, I.S. Inter-species differences of co-expression of neighboring genes in eukaryotic genomes. *BMC Genomics* **5**, 4 (2004).

217. Purmann, A. u. a. Genomic organization of transcriptomes in mammals: Coregulation and cofunctionality. *Genomics* **89**, 580-587 (2007).

218. Spellman, P.T. & Rubin, G.M. Evidence for large domains of similarly expressed genes in the Drosophila genome. *J. Biol* **1**, 5 (2002).

219. Williams, E.J.B. & Bowles, D.J. Coexpression of neighboring genes in the genome of Arabidopsis thaliana. *Genome Res* **14**, 1060-1067 (2004).

220. Hebbes, T.R., Clayton, A.L., Thorne, A.W. & Crane-Robinson, C. Core histone hyperacetylation co-maps with generalized DNase I sensitivity in the chicken beta-globin chromosomal domain. *EMBO J* **13**, 1823-1830 (1994).

221. Oliver, B., Parisi, M. & Clark, D. Gene expression neighborhoods. *J. Biol* **1**, 4 (2002).

222. Ebisuya, M., Yamamoto, T., Nakajima, M. & Nishida, E. Ripples from neighbouring transcription. *Nat. Cell Biol* **10**, 1106-1113 (2008).

223. Tsai, H., Su, C.P.C., Lu, M.J., Shih, C. & Wang, D. Co-expression of adjacent genes in yeast cannot be simply attributed to shared regulatory system. *BMC Genomics* **8**, 352 (2007).

224. Gu, F. u. a. Structure and function of a virally encoded fungal toxin from Ustilago maydis: a fungal and mammalian Ca2+ channel inhibitor. *Structure* **3**, 805-814 (1995).

225. Park, C.M., Banerjee, N., Koltin, Y. & Bruenn, J.A. The Ustilago maydis virally encoded KP1 killer toxin. *Mol. Microbiol* **20**, 957-963 (1996).

226. Gage, M.J., Bruenn, J., Fischer, M., Sanders, D. & Smith, T.J. KP4 fungal toxin inhibits growth in Ustilago maydis by blocking calcium uptake. *Mol. Microbiol* **41**, 775-785 (2001).

227. Pepeljnjak, S., Petrik, J. & Klarić, M.S. Toxic effects of Ustilago maydis and fumonisin B1 in rats. *Acta Pharm* **55**, 339-348 (2005).

228. Nierman, W.C. u. a. Genomic sequence of the pathogenic and allergenic filamentous fungus Aspergillus fumigatus. *Nature* **438**, 1151-1156 (2005).

229. Denayrolles, M., de Villechenon, E.P., Lonvaud-Funel, A. & Aigle, M. Incidence of SUC-RTM telomeric repeated genes in brewing and wild wine strains of Saccharomyces. *Curr. Genet* **31**, 457-461 (1997).

230. Farman, M.L. Telomeres in the rice blast fungus Magnaporthe oryzae: the world of the end as we know it. *FEMS Microbiol. Lett* **273**, 125-132 (2007).

231. Brakhage, A.A. u. a. Activation of fungal silent gene clusters: a new avenue to drug discovery. *Prog Drug Res* **66**, 1, 3-12 (2008).