TECHNISCHE UNIVERSITÄT MÜNCHEN
Lehrstuhl für Genomorientierte Bioinformatik

Predicting virulence factors in filamentous fungi: Regulation and evolution of secondary metabolism gene clusters

Christian Martin Konrad Sieber

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. K. Jung,
   (Ludwig-Maximilians-Universität München)

Die Dissertation wurde am 13.11.2014 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 20.01.2015 angenommen.

Abstract

Pathogenic filamentous fungi constitute a health risk to humans and animals all over the world. Most of these fungi provide a diverse repertoire of bioactive small molecules like antibiotics and mycotoxins which play a key role in diseases but are also utilized as drugs and growth factors of plants. Especially *Fusarium* species are known to be involved in many plant diseases that lead to large agricultural and economic damage.

Genes that encode enzymes of a secondary metabolism pathway usually are locally clustered on the chromosome. The rapidly increasing number of available fungal genomes enables comparative genomics studies for the identification of host specific virulence factors. Furthermore, large-scale genomic mining for gene clusters of novel bioactive compounds has become feasible.

In this work the genome sequence of the rice pathogen *Fusarium fujikuroi* alongside with an extensive comparative analysis to *Fusarium* species with diverse host specificities is presented. To better understand the regulation of secondary metabolism gene clusters and virulence associated genes, transcriptomic, proteomic and epigenetic measurements taken under virulence inducing conditions are integrated. A significant genome wide correlation between gene expression and protein abundance could be determined that is even higher when focusing on secondary metabolism genes alone. Epigenetic regulation of gene clusters by differential chromatin acetylation at different physiological conditions is discovered. Furthermore, influence of the histone deacetylase Hda1 and the global regulator Sge1 on secondary metabolism could be shown by analyzing experimental data of deletion mutants.

Genome shaping and plasticity is heavily influenced by transposable elements. *In silico* prediction of repetitive interspersed sequences revealed two highly abundant repeat families that occur exclusively in two distinct phylogenetic branches inside the *Fusarium* phylum. Extensive evidence for the repeat induced point mutation (RIP) mechanism that inactivates transposable elements in *F. fujikuroi* is presented. Expression data analysis revealed an inverse correlation between the amount of RIP-specific point mutations and expression intensities of transposable elements.

Recent genome analysis studies on filamentous fungi revealed a diversity of putative secondary metabolism genes. At the same time only a small fraction of synthesized compounds is known. To determine the amount of secondary metabolism genes that is involved in biosynthetic pathways, putative secondary metabolism gene clusters based on statistical overrepresentation of secondary metabolism related functions are predicted. Additionally, evidence in terms of co-expression by available experimental data and computed conserved promoter motifs is presented. In collaboration with experimental biologists it was possible to identify previously unknown compounds and verify the regulatory function of overrepresented promoter motifs.

The presence of gene clusters between even closely related fungi differs considerably in most of the cases and the evolutionary origin of many biosynthetic genes is yet unknown. In an extensive ortholog analysis of predicted clusters of 22 *Fusarium* species different evolutionary processes are observed, such as individual gene loss and horizontal gene transfer, that are possibly responsible for the non-uniform phylogenetic distribution of gene clusters. In addition shuffling of pathway genes between gene clusters, alternative tailoring of a common signature enzyme as well as duplication and divergence of clusters is observed.

The results of this work not only identify species-specific virulence related features and give insight into the regulation of secondary metabolism, but also reveal evolutionary processes that shed light on the origin of secondary metabolism gene clusters.

Zusammenfassung

Pathogene filamentöse Pilze stellen ein Gesundheitsrisiko für Mensch und Tier auf der ganzen Welt dar.

Dazu trägt das vielfältige Repertoire an kleinen bioaktiven Molekülen bei, das von vielen dieser Pilze synthetisiert wird. Die sogenannten Sekundärmetabolite bieten ein Selektionsvorteil für die Organismen, sind aber nicht essentiell für deren Entwicklung. Beispiele sind Antibiotika, Phytohormone und Pilzgifte, die eine wichtige Rolle bei Krankheiten spielen oder als Medikamente in der Medizin sowie als Wachstumsfaktoren für Pflanzen Verwendung finden. Vor allem Pilze der Gattung *Fusarium* sind in vielen Pflanzenkrankheiten involviert, die zu großem landwirtschaftlichen und somit auch wirtschaftlichen Schaden führen.

Gene, die Enzyme eines Sekundeärmetabolismus Pathways kodieren, sind in der Regel in direkter chromosomaler Nachbarschaft zueinander in Form von Genclustern angeordnet. Die schnell wachsende Anzahl an verfügbaren Pilz-Genomsequenzen ermöglicht neue, vergleichende genomische Studien zur Identifizierung von wirtsspezifischen Virulenzfaktoren. Darüber hinaus werden genomweite, großformatige Vorhersagen von Gencluster möglich, die noch unbekannte, bioaktive Stoffe synthetisieren.

In dieser Arbeit wird die Genomsequenz des Reispathogens *Fusarium fujikuroi* zusammen mit einer umfangreichen vergleichenden Analyse mehrerer Fusarium-Arten mit verschiedenen Host-Präferenzen präsentiert. Um die Regulation von Sekundärmetabolismus Genclustern und Virulenz-Genen besser zu verstehen, werden Daten aus Transkriptomik-, Proteomik- und Epigenetik- Experimenten zu Virulenz induzierenden Bedingungen integriert.

Eine genomweite Korrelation zwischen Genexpressionsintensität und quantifizierten Proteinen konnte bestimmt werden, die besonders stark bei Sekundeärmetabolismusgenen zu beobachten ist. Des weiteren wird eine epigenetische Regulation der Gencluster gezeigt, die auf unterschiedliche Chromatin-Acetylierung unter verschiedenen physiologischen Bedingungen zurückgeführt werden konnte. Darüber hinaus konnte ein regulatorischer Einfluss der Histon-Deacetylase Hda1 und des globalen Regulators Sge1 auf den Sekundärmetabolismus durch experimentelle Daten von Deletionsmutanten gezeigt werden.

Transposons und Repeats beeinflussen die Struktur und Plastizität eines Genoms. Die *in silico* Vorhersage von repetitiven Sequenzen resultierte in zwei sehr präsenten Repeat-Familien, die ausschließlich in zwei monophyletischen Gruppen innerhalb der *Fusarium* Phylogenie vorkommen. Hinweise für den aktiven Abwehrmechanismus RIP (repeat-induced point mutation), welcher Transposons inaktiviert, konnten in *F. fujikuroi* gefunden werden. Eine Expressionsdaten-Analyse zeigte eine inverse Korrelation zwischen der Anzahl an RIP-spezifischen Punktmutationen und der Transposon Expressions Intensität.

In aktuellen Analysen von vollständig sequenzierten Pilzgenomen wurde eine Vielzahl von vermeintlichen Sekundärmetabolismusgenen vorhergesagt. Gleichzeitig ist aber nur ein kleiner Teil an synthetisierten Metaboliten bekannt. Um putative Sekundeärmetabolismusgene zu bestimmen, die Teil eines Biosynthese-Pathways darstellen, wurde eine Gencluster Vorhersage durchgeführt, welche auf statistischer Überrerpesentation von Sekundärmetabolismus Funktionen beruht. Zusätzlich wurden Evidenzen zur Coexpression, in Form von experimentellen Daten und konservierten Promotor-Motiven, mit einbezogen. In Kooperation mit experimentellen Biologen konnten bisher unbekannte Metabolite identifiziert und die regulatorische Funktion überrepräsentierter Promotor Motive bestätigt werden. Selbst zwischen nah verwandten Species unterscheidet sich die Präsenz von Genclustern im Genomvergleich teilweise erheblich, wobei die Herkunft vieler Biosynthesegene unbekannt ist. In einer umfangreichen Analyse der vorhergesagten Orthologen Cluster von 22 Fusarium-Arten wurden verschiedene putative evolutionäre Abläufe identifiziert, die möglicherweise verantwortlich für die ungleichmäßige phylogenetische Verteilung der Gencluster sind. Diese umfassen den individuellen Verlust von Genclustern, aber auch horizontalen Gentransfer. Zusätzlich konnte die Fusion von Pathwaygenen zweier Gencluster, unterschiedliche Tailoring Enzyme um ein gemeinsames Signature-Enzym und eine Duplikation mit Divergenz eines Clusters beobachtet werden.

Die Ergebnisse dieser Arbeit identifizieren nicht nur artspezifische Virulenzfaktoren und geben Einblick in die Sekundeärmetabolismus-Regulation, sondern zeigen auch evolutionäre Prozesse, die Licht auf die Entstehung der Sekundärmetabolismus Gencluster werfen.

Acknowledgements

A lot of people accompanied and supported me during the last years. First, I would like to thank my supervisor Prof. Dr. Hans-Werner Mewes for his support, advice and for giving me the opportunity to pursue my research interests.

A very special thank goes to my advisor Dr. Ulrich Güldener for superb supervision, advise and motivation during the whole time of my PhD. I really enjoyed the time at the IBIS and the Fungal Genomics Group and I am thankful for the countless inspiring discussion with the present and past group members including Dr. Martin Münsterkötter, Mathias Walter and Philip Wong.

I also want to thank Prof. Dr. Kirsten Jung for constructive input during the thesis committee meetings.

For the great cooperations in the *F. fujikuroi* genome project and the follow up projects I want to acknowledge the groups of Prof. Dr. Bettina Tudzynski and Prof. Dr. Hans-Ulrich Humpf from the University of Münster. Especially Dr. Eva Niehaus, Dr. Philipp Wiemann, Katharina von Bargen, Dr. Lena Studt, Dr. Karin Kleigrewe, Dr. Sabine Albermann, Dr. José Espino, Dr. Caroline Michielse, Sarah Rösler and Andreas Pfannmüller.

The participants of the *F. graminearum* cluster project especially Dr. Wanseon Lee but also Prof. Dr. Gerhard Adam and his group from the BOKU in Vienna.

For getting insight into computational evolution and the 'kiwiana' way of life down under I thank Prof. Dr. Alexei Drummond, Dr. Tim Vaughan, Alex Popinga, Dr. Sasha Gavryushkina of the CompEvo Group at the University of Auckland.

For intercultural cooperations beside the topics of my thesis I thank Dr. Nafees Ahmad from the Institute of Biomedical and Genetic Engineering in Islamabad.

For support in technical issues I acknowledge the past and present IBIS-admins Dr. Giovani Gomez Estrada, Jonathan Hoser and Sven Duscha.

I also have to thank the DFG and HELENA for funding and support.

Finally, I want to thank my parents, family and friends for their endless support and motivation.

Publications

Most results that are discussed in this thesis are published in peer-reviewed journals:

- **Sieber CMK***, Lee W*, Wong P, Münsterkötter M, Mewes HW, Schmeitzl C, Varga E, Berthiller F, Adam G, Güldener U.
  The *Fusarium graminearum* genome reveals more secondary metabolite gene clusters and hints of horizontal gene transfer.
  PLoS One. 2014, 9, e110311

- Wiemann P*, **Sieber CMK***, von Bargen KW*, Studt L, Niehaus EM, Espino JJ, Huß K, Michielse CB, Albermann S, Wagner D, Bergner SV, Connolly LR, Fischer A, Reuter G, Kleigrewe K, Bald T, Wingfield BD, Ophir R, Freeman S, Hippler M, Smith KM, Brown DW, Proctor RH, Münsterkötter M, Freitag M, Humpf HU, Güldener U, Tudzynski B.
  Deciphering the cryptic genome: genome-wide analyses of the rice pathogen *Fusarium fujikuroi* reveal complex regulation of secondary metabolism and novel metabolites.
  PLoS Pathog. 2013, 9, e1003475

- Studt L, Schmidt FJ, Jahn L, **Sieber CMK**, Connolly LR, Niehaus EM, Freitag M, Humpf HU, Tudzynski B.
  Two histone deacetylases, FfHda1 and FfHda2, are important for *Fusarium fujikuroi* secondary metabolism and virulence.
  Appl Environ Microbiol. 2013, 79, 7719-7734

- Michielse CB, Studt L, Janevska S, **Sieber CMK**, Espino JJ, Arndt B, Humpf HU, Güldener U, Tudzynski B.
  The global regulator FfSge1 is required for expression of secondary metabolite gene clusters, but not for pathogenicity in *Fusarium fujikuroi.*
  Environ Microbiol. 2014 (Epub ahead of print)

*equal contributions

- Niehaus EM, Janevska S, von Bargen KW, **Sieber CMK**, Harrer H, Humpf HU, Tudzynski B.
  Apicidin F: Characterization and genetic manipulation of a new secondary metabolite gene cluster in the rice pathogen *Fusarium fujikuroi*.
  PLoS One. 2014, 9, e103336

- Pfannmüller A, Wagner D, **Sieber CMK**, Schönig B, Boeckstaens M, Marini AM, Tudzynski B.
  The general amino acid permease FfGap1 of *Fusarium fujikuroi* is sorted to the vacuole in a nitrogen-dependent, but Npr1 kinase-independent manner.
  Submitted.

Further publications

During my PhD studies I participated in some side projects that are not discussed in this thesis but have resulted in the following publications:

- Niehaus EM, Kleigrewe K, Wiemann P, Studt L, **Sieber CMK**, Connolly LR, Freitag M, Güldener U, Tudzynski B, Humpf HU.
  Genetic manipulation of the *Fusarium fujikuroi* fusarin gene cluster yields insight into the complex regulation and fusarin biosynthetic pathway.
  Chem Biol. 2013, 20, 1055-1066

- Nasir M, Ahmad N, **Sieber CMK**, Latif A, Malik SA, Hameed A.
  In silico characterization of a novel pathogenic deletion mutation identified in XPA gene in a Pakistani family with severe xeroderma pigmentosum.
  J Biomed Sci. 2013, 20, 70

- Nasir M, Rahman SB, **Sieber CMK**, Mir A, Latif A, Ahmad N, Malik SA, Hameed A.
  Identification of recurrent c.742G>T nonsense mutation in ECM1 in Pakistani families suffering from lipoid proteinosis.
  Mol Biol Rep. 2014, 41, 2085-2092

Contents

List of Figures

List of Tables

List of Abbreviations

AAP	amino acid permease
ACP domain	acyl-carrier domain
AT domain	acyl transferase domain
aLRT	approximate likelihood-ratio test
BBH	bidirectional best hit
CDS	coding DNA sequence
DH domain	dehydratase domain
DMATS	dimethylallyl tryptophan synthase
DNA	deoxyribonucleic acid
DTC	citerpene cyclase
ER	endoplasmatic reticulum
ER domain	enolreductase domain
GA	gibberellic acid
GFC	*Gibberella fujikuroi* species complex
FGC	*Fusarium graminearum* species complex
FOC	*Fusarium oxysporum* species complex
GAP	general amino acid permease
HAT	histone acetyltransferase
HDAC	histone deacetylases
HGT	horizontal gene transfer
HMM	hidden markov model
KR domain	ketoreductase domain
KS domain	ketosynthase domain
LINE	long interspersed nuclear element
LTR	long terminal repeat
mRNA	messenger RNA
NLCS	normalized locus-specific chromatin state
NPS	non-ribosomal peptide synthase
PKS	polyketide synthase
RC	reliability class
RNA	ribonucleic acid

RIP	repeat induced point mutation
SINE	short interspersed nuclear element
SIX	secreted in xylem
SM	secondary metabolism
SP	secreted protein
SRP	signal recognition particle
SSP	small secreted protein
STC	sesquiterpene cyclase
TF	transcription factor
TP	transporter
TPS	terpene synthase
tRNA	transfer RNA
UTR	untranslated region

Chapter 1

Introduction

Filamentous fungi of the genus *Fusarium* are nearly omnipresent. Their geographic distribution comprise America [5, 73, 245], Africa [26, 98, 160], Asia [6, 188, 247], Oceania [88, 118, 146] and Europe [40, 157, 124]. The host range of *Fusarium* species covers many agriculturally important plants which suffer severe plant diseases upon infection. These diseases are often caused by mycotoxins which are synthesized by the pathogenic fungi. The infection of plants and the mycotoxin contamination of food and feed lead to financial damage and represents a health hazard for humans and animals.

The raising number of available *Fusarium* whole genome sequences enables the identification of host specific virulence factors through comparative genomics and bioinformatics approaches.

1.1 *Fusarium* species

Fusarium graminearum is one of the most extensively studied *Fusarium* species with high economical impact world wide. The head blight, crown- and root- rot diseases in cereals such as barley and wheat lead to financial loss of around $3 billion in the United States during the 1990s [235]. Synthesized mycotoxins like deoxynivalenol (DON) cause vomiting, diarrhea and leukocytosis upon consumption [166]. As mycotoxins are not necessary for the fungal growth and development they are categorized as secondary metabolites.

However, not all secondary metabolites produced by filamentous fungi are an object of disutility. Gibberellic acid synthesized by the rice pathogen *F. fujikuroi* for example is a growth hormone and main cause of the 'bakanae' disease in rice. 'Bakanae' is japanese and means foolish seedling because the gibberellic acid leads to an elongation of rice seedlings and causes chlorotic stems and leaves [242]. The plants become infertile and therefore growth of grain is inhibited. On the other hand gibberellins are applied in agriculture to regulate plant growth and development. For this purpose, *F. fujikuroi* is used for the commercial production of these hormones [10, 214].

The *Gibberella fujikuroi* species complex (GFC) is divided into the Asian, African and American geographical clades. At the moment whole genome sequences of five GFC species are available, comprising *F. fujikuroi*, *F. mangiferae* and *F. proliferatum* of the Asian clade. *F. fujikuroi* and the associated 'bakanae' disease were described 100 years ago in Japan [117]. While *F. proliferatum* has a wide host range *F. mangiferae* preferably grows on mango and causes mango malformation [68, 74]. A representative of the African clade is the maize pathogen *F. verticillioides* which is responsible for ear rot and stalk rot diseases [132]. *F. circinatum* a member of the American clade and causal agent of pitch canker of pines was sequenced recently [237].

Beside contaminated grain *F. graminearum* also constitutes a direct thread to humans. The fusariosis called disease describes the infection of immunocompromised patients and is a risk especially during organ transplantations [25, 137]. Another *Fusarium* mediated disease is the eye infection Keratitis which is mainly developed by people wearing contact lenses [3, 41].

1.2 Virulence factors

Virulence factors contribute to the ability of fungi to infect certain hosts and to trigger plant diseases. Their presence or absence allows to distinguish from virulent or avirulent strains [123, 216]. Genes encoding effectors which are small secreted proteins or enzymes that synthesize mycotoxins have impact on the virulent properties of fungi. In the tomato pathogen *F. oxysporum* f. sp. lycopersici seven effector genes are known which are directly involved in host-pathogen interaction. The SIX (secreted in xylem) genes interact directly with immunity genes in tomato. The genes are localized on a supernumerary, lineage specific chromosome that is only present in virulent *F. oxysporum* strains [96, 123, 179]. Interestingly a transfer of the chromosome in vitro between strains lead to the acquisition of virulence in previously avirulent strains [132]. Furthermore, lineage specific regions in the genome of *F. oxysporum* harbor many transposons and genes with distinct codon adaptation index (cai) and codon usage compared to the core-genome which is also a hint of horizontal acquisition [132]. A connection between supernumerary chromosomes and virulence was also determined in the pea pathogen *F. solani* [84]. Recent studies showed that also in *F. graminearum* virulence related small secreted proteins are present. Parts of the candidate effectors have orthologs in other *Fusarium* species and thus are not connected to host specificity [33]. In the closely related *F. pseudograminearum* virulence genes from other fungal cereal pathogens and plant associated bacteria were determined [77].

In general, secreted proteins can be divided into two classes according to the mechanism of secretion. Polypeptides containing a signal sequence in their n-terminal region can be recognized by a signal recognition particle (SRP) and transferred into the endoplasmatic reticulum (ER) in passing the Sec61p translocon complex [45]. Beside this classical secretion pathway proteins without signal peptide can also be transported into the extracellular space using mechanisms that work independently of signal peptide recognition or the ER [153].

Algorithms for the *in silico* prediction of secreted proteins determine sequence based features that are characteristic for secreted proteins. In a neural network approach SignalP 4.0 identifies signal peptides and cleavage sites in amino acid sequences [167]. To distinguish between signal peptides and N-terminal transmembrane domains, which have similar hydrophobic properties, SignalP applies two models trained for the prediction of signal peptides and transmembrane helices, respectively [167]. TargetP integrates SignalP and extends it by predictors of sub cellular locations of proteins. Using a neural network TargetP returns a reliability class (RC) score for compartments such as chloroplast, mitochondrion or in case of secreted proteins the ER and golgi [33]. WolfPSort predicts more compartments in applying a k-nearest neighbor classifier that considers amino acid compositions and functional motifs of the candidate sequences and compares it to the training set [95]. Beside prediction of golgi mediated secretion, SecretomeP focuses on non-classically secreted proteins that lack a N-terminal signal peptide. It integrates the tools SignalP and WolfPSort for signal peptide and compartment prediction and uses TMHMM [114] for the prediction of transmembrane helices. Non classically secreted proteins are predicted by a sequence feature-based neural network [16]. An estimation of the secretome size of *F. graminearum* using a pipeline of *in silico* tools resulted in 574 candidate proteins [33]. However only classically secreted proteins were taken into account. In Section 2.2.4 an improved prediction pipeline that considers all kinds of secreted proteins is described.

In addition to secreted proteins the products of secondary metabolism pathways like mycotoxins and phytohormones have a considerable impact on virulence.

1.3 Secondary metabolism gene clusters

Microbes and plants produce a variety of secondary metabolites (SMs) with diverse bioactive features of ecological and medical impact. The low molecular wight compounds are not crucial for the development of the organism but provide a selective advantage over other species. A prominent example are antibiotics like penicillin which is synthesized by bacteria and fungi as well. The medical important antibacterial compound targets the cell wall of bacteria and prevents their reproduction. Other SMs are applied in medicine or agriculture as immunosuppressant (Cyclosporin A), antitumor (Daunorubicin HCI), antifungal (Amphotericin B) or herbicide (Bialaphos) agents [52]. But many SMs also constitute a threat for humans and animals like the highly toxic and cancerogenic aflatoxins that are synthesized by the fungus *Aspergillus flavus* [91, 198]. In plant pathogenic fungi SMs like mycotoxins and phytohormones are often associated with virulence, abnormal growth and plant diseases of hosts [75, 242].

Initial whole genome analyses revealed a pleiotropy of secondary metabolism genes while only comparable low fraction of SMs are known. The advent of cheaper and faster sequencing techniques increased the available number of fungal genomes dramatically in recent years enabling large scale bioinformatics analysis for putative gene clusters and novel bioactive compounds.

The main synthesis step in fungal secondary metabolism pathways constitutes the creation of the backbone structure of the compound. This step is done by signature enzymes like type I polyketide synthases (PKS), non-ribosomal peptide synthases (NPS), terpene synthases (TPS) or dimethylallyltryptophan synthases (DMATS). Polyketides are the most abundant fungal secondary metabolites. Their synthases are multi-domain proteins and constitute of a minimal set of three functional domains: a ketosynthase (KS domain), an acyl-carrier domain (ACP domain) and an acyltransferase domain (AT domain). Reducing type PKS (R-PKS) additionally contain a ketoreductase- (KR domain), a dehydratase- (DH domain) and enolreductase-domain (ER domain) which are required for ketone reduction in fatty acids [106]. NPS enzymes also consist of multiple functional domains that are required for the backbone assembly of non-ribosomal peptides.

After the main synthesis step additional modifications of the compound can involve tailoring enzymes such as cytochrome P450s, methyltransferases, acyltransferases, oxidoreductases or glycosyltransferases in further pathway steps [163].

In most cases the enzyme encoding genes that are involved in a secondary metabolism pathway are physically clustered on the chromosome [105, 139]. These secondary metabolism gene clusters often have co-localized transcription factors and transporters for pathway regulation and export of the synthesized compound. It can be observed that many clusters are located at the subtelomeric region of chromosomes where the genes are exposed to an increased mutation rate [163].

1.4 Regulation of virulence genes

Many secondary metabolism gene clusters like the aflatoxin cluster in *Aspergillus flavus* [194] contain a transcription factor that specifically regulates the genes and the synthesis of the metabolite. In order to save resources gene regulation ensures that the secondary metabolites and effectors are only synthesized when they are required. Beside specific transcription factors global regulators like AreA, AreB or the velvet complex regulate virulence associated genes and secondary metabolism pathways [97, 143, 229]. The global transcription factor Ryp1 in *Histoplasma* *capsulatum* and its ortholog Wor1 in *Candida albicans* control changes in morphology and lifestyle in both human pathogens. In *H. capsulatum* it switches between saprophytic filamentous growth to yeast like morphology and pathogenic lifestyle at body temperature [152]. Recently Sge1 was identified as an ortholog of Wor1/Ryp1 in *F. oxysporum* where it regulates the virulence associated *SIX* genes [152]. A deletion of the regulator lead to a loss of the photo toxic sequiterpenoid and trichothecene synthesis in *F. graminearum* and revealed a connection to virulence and pathogenicity in the cereal pathogen [101]. However, the impact of Sge1 in *F. fujikuroi* and other *Fusarium* species is still unknown. In Section 2.2.3 the gene expression of a *SGE1* deletion mutant was analyzed in order to predict putative target genes and secondary metabolism pathways that are significantly affected by the deletion in *F. fujikuroi*.

Epigenetic regulation of secondary metabolism in terms of chromatin modifications was observed in several species [21, 233]. Examples of epigenetic controlled secondary metabolism pathways are the fumonisin gene cluster in *F. verticillioides* [220] or the sterigmatocystin cluster in *A. nidulans* [180]. By acetylation, methylation or phosphorylation of specific residues in the unstructured amino acid tails of the histones the conformation of the chromatin can be changed from the loosely packed euchromatin to the condensed heterochromatin and vice versa. This affects the accessibility of the DNA and the regulation of genes as expression is only possible in the loosely packed euchromatin state [69, 79]. Enzymes for the addition and removal of modifying histone marks act as opposing forces. Histone acetyltransferases (HATs) for example are able to add acetyl groups to specific amino acid residues while histone deacetylases (HDACs) remove these marks [69, 79]. In Section 2.2.8 the impact of the histone deacetylase Hda1 in *F. fujikuroi* on secondary metabolism is shown.

Environmental stimuli like the availability of nutrients, light or pH influence gene expression in general but have also an impact on virulence and secondary metabolism synthesis. While mycotoxins are usually induced during plant infection, the production of pigments for the protection of UV radiation is regulated by light stimuli [211, 218]. It was shown that the gene expression of *ipnA* which encodes the isopenicillin N synthase in *A. nidulans* is affected by pH [62]. Especially in plant pathogenic fungi the availability of nitrogen was shown to be involved as a switch for expression induction of infection related genes as shown in *Magnaporthe grisae* and *F. oxysporum* [55, 58]. In Section 2.2.6, I examine the effects of different nitrogen concentrations on the gene expression and protein abundance in *F. fujikuroi.* Alterations in the genome wide chromatin landscape due to nitrogen availability are described in Section 2.2.7.

To understand the nitrogen regulatory network in *F. fujikuroi* which also involves regulation of secondary metabolism, it is important to investigate in the sensory and uptake mechanisms of nitrogen. In *S. cerevisiae* the general amino acid permease Gap1 plays a role in both sensory and transport [57, 217]. Therefore, putative Gap1 homologs among a set of predicted amino acid permeases are identified for further experimental characterization in Section 2.2.5.

Transporter proteins are also important for the eflux of synthesized secondary metabolite compounds. The production of the mycotoxin zearalenone (ZEA) in *F. graminearum* depends on the activity of the putative ABC transporter ZRA1. It was shown that expression of *ZRA1* is significantly different between ZEA producing and non-producing *F. graminearum* strains. And a deletion of the *ZRA1* gene resulted in reduced ZEA production [121]. I predicted transporter proteins based on functional domains and compared the abundance between different *Fusarium* genomes in Section 2.2.5.

1.5 Prediction of secondary metabolism gene clusters

Several strategies for the *de novo* prediction of secondary metabolism gene clusters exist. The prediction tools SMURF [107] and AntiSMASH [19] utilize the characteristic functional enzymatic composition to predict gene clusters based on protein domains. A similar approach with a focus on *Fusarium graminearum* has been performed by Ma *et al.* [132]. 15 novel clusters have been predicted using functional domain information in combination with two microarray experiments of expression quantification during plant infection and sexual development as evidence. This set of predicted clusters was extended with four novel clusters that were identified based on co-expression analysis by Zhang *et al.* using time series microarray experiments of *F. graminearum* growing inside wheat coleoptiles [248]. Utilizing four microarray experiments as co-expression evidence, Lawler *et al.* showed that co-expressed cluster genes in *F. graminearum* often contain transcription associated proteins such as transcription factors and genes involved in biosynthetic pathways like the butenolide gene cluster [120].

In Section 4.2.2, I present a de novo approach that utilizes four sources of evidence to predict novel gene clusters and to validate known ones (Table 4.3). Candidate PKS, NPS, TPS and DMATS clusters are predicted based on functional domain composition and overrepresented promoter motifs which suggest co-regulation are identified. I determined evolutionary conservation of gene clusters by searching a protein similarity database of 381 genomes for orthologous clusters [191]. Finally microarray experiments were analyzed in order to determine co-expression of genes with an emphasis on expression during plant infection (Table 2.1). Besides clusters of known metabolites, the analyses identified a plurality of putative SM gene clusters (Table A.3).

1.6 Evolution of secondary metabolism gene clusters

In general evolutionary processes for the generation and acquisition of novel genetic material include the duplication and divergence of genes, hybridization of DNA between species and horizontal gene transfer. Horizontal gene transfer was observed between fungal species as well as between bacteria and fungi. The amount of the acquired genetic material ranges from single genes to whole chromosomes as observed in *F. oxysporum* [132]. In *F. solani* a supernumerous chromosome was determined which contains genes with different G+C ratio compared to the other genes of *F. solani.* The genes on the chromosome contain the pea pathogenicity (PEP) gene cluster which are involved in plant pathogenicity [84]. Many gene clusters located on the main chromosomes exhibit also a discontinuous phylogenetic distribution among closely related species. The β-lactam (penicillin) cluster is present in bacteria and fungi as well. Protein similarity suggest a transfer of the whole cluster between the two kingdoms. However, phylogenetic analyses could not provide significant support for this hypothesis [34, 195]. In a comparative genomics approach horizontal transfer of the bikaverin gene cluster between the *Fusarium* and *Botrytis* phylum could be shown. The cluster was determined in three *Fusarium* and two *Botrytis* genomes whereas the neighboring genes of the cluster are only conserved in the respective phylum. The orthologous cluster genes showed a similar G+C ratio while it was different to the ratio of the neighboring genes in *Botrytis* [37, 38].

Evolutionary driven creation of novel secondary metabolism pathways is also possible by modification and adaption of existing gene collectives. Re-ordering, adding or removing of genes by genome rearrangements was described in bacterial operons [170]. A comparison of bacterial aminoglycoside gene clusters showed a conservation of signature enzymes but alternating tailoring enzymes. The alternative tailoring leads to a common metabolite backbone synthesized by the signature enzyme with different peripheral sugars added by tailoring enzymes [66, 116].

Duplication and divergence of genes is a way of divergent evolution to create new genes with new functions while keeping the original copy of the gene. The gene family of polyketide synthases is a prominent example were duplication and divergence resulted in a functional diversity of enzymes with a wide distribution among many fungal genomes [115]. Interestingly, also convergent evolution was observed where unrelated clusters independently evolved the same molecule. The growth hormone gibberellic acid for example is synthesized by plants, bacteria and fungi using three different synthesis pathways [148, 214, 92]. The gibberellic acid gene cluster is distributed among the species of the GFC. In *F. proliferatum* traces of divergent evolution in terms of a whole cluster duplication of the synthesis pathway of the convergently evolved molecule were determined (Section 5.2.9).

1.7 Transposable elements

Transposons contribute to genome plasticity, evolution of single genes and exchange of genetic material as well. Therefore prediction of interspersed repeats and transposons in fungal genomes is of major interest. Interspersed repeats are repetitive and putative transposable DNA sequence elements that occur numerous times in most eukaryotic genomes. In inhibiting gene conversion they maintain genetic diversity [27, 185]. According to their sequence features and mechanism of propagation transposable elements can be assigned into two classes. Transposons of class I are also called retrotransposons as a reverse transcriptase is involved in the replication mechanism where a RNA intermediate is created which is then reverse transcribed afterwards. With this copy-paste like mechanism new repeat elements are created and thus the overall repeat content is increased [228]. Retrotransposons can be further categorized in LTR-retrotransposons which have long terminal repeats (LTRs) at their ends like the Gypsy and Copia elements, Long Interspersed Nuclear Elements (LINEs) and Short Interspersed Nuclear Elements (SINEs) that both lack the terminal repeat sequences. LINEs and LTR-retrotransposons are very common in fungal genomes [149, 158]. Class II transposons are DNA-transposons which can be divided into two subclasses according to their transposition. While elements of subclass one transpose by excision and insertion, subclass two also replicate themselves like Retrotransposons but without an RNA intermediate. In both transposon classes elements that lack one or more proteins for the transposition are known. These non-autonomous transposons rely on the proteins encoded on other transposable elements [64]. Although transposable elements are important for genome variability and evolution the excision and insertion of TEs can lead to genome instability and therefore is putatively harmful [187]. In many fungal organisms defense mechanisms can be observed that inactivate transposons and prevent them from further transposition [72, 187]. The mechanism of repeat induced point mutation (RIP) mutates transposons and repetitive DNA. Between mating and meiosis RIP recognizes DNA repeat elements with a minimum length of 400 bp and a minimum sequence identity of 80% and induces C:G to T:A mutations [36, 187, 226]. RIP has been first detected in *Neurospora crassa* [187] but evidence was observed also in other filamentous fungi [43] like *Ustilago hordei* [119], *F. solani* [44] and *F. graminearum* [47]. RIP has an impact on genome evolution as it inactivates novel genes that were created by duplication. However also an acceleration of evolution of existing genes is possible. Dependent on the amount of induced mutations it is possible that a duplicated gene remains functional [72]. Beside accelerating the evolution of the duplicated gene itself it was shown that genes next to RIP affected sequences are also exposed to higher rates of point mutations. In *Dothideomycetes* effector genes were frequently found near TEs where the increased mutation rate putatively promotes the adaption to the host [161]. I predicted repetitive and transposable elements in *Fusarium* species and incorporate available expression data to get evidence for active transposition and investigated for indications of RIP in Chapter 3.

1.8 Research questions

The main goal of this thesis is to identify virulence factors including secondary metabolism gene clusters and to investigate their regulatory mechanisms and evolutionary origin. More precisely, I will predict secreted proteins, effectors and secondary metabolism gene clusters on available fully sequenced *Fusarium* genomes. I will incorporate experimental data to get insight into their regulation and expression conditions. A comparative genomics approach will be used to analyze orthologs and to discuss possible evolutionary origins.

The sequence based comparison of fungal genomes reveals differences in terms of chromosome size and number. The question is how these differences can be linked to lifestyle and host specificity of fungi. Furthermore, the contribution of transposable elements to the genome shape is of interest as well as putative defense mechanisms that inactivate transposition of repetitive sequences.

During the infection process secreted proteins are involved in the host-pathogen interaction between filamentous fungi and the host plant. These effector proteins are of major interest as they belong to the main factors that contribute to virulence of pathogenic fungi. However the prediction of secretion candidates is challenging and previous studies in *Fusarium* were only focusing on the classical, signal peptide mediated secretion pathway. I will ask how large the whole set of secreted proteins is and investigate the fraction of differentially expressed secreted protein coding genes under virulence inducing conditions.

In previous fungal genome studies a high number of secondary metabolism genes was predicted while only a small fraction of these genes could be linked to known secondary metabolites. One question is how many compounds are synthesized by filamentous fungi that are not produced under laboratory conditions and therefore difficult to measure. An estimate to that number gives the amount of putative secondary metabolism gene clusters which are candidates for synthesis pathways. The environmental conditions and stimuli that trigger the expression of the clusters are also of interest in order to characterize the synthesized compound.

Beside mining gene clusters and novel natural products, the clustered organization of secondary metabolism genes still poses many unanswered questions. It is still unclear why secondary metabolism pathways, in contrast to their primary equivalents, are encoded physically linked on the chromosome in terms of gene collectives. Furthermore the origin of some of these clusters is still unknown. Some demonstrated cases of horizontal whole gene cluster transfers rise the question how frequently these events occur and whether gene clusters are popular targets for this mechanism of inheritance. Another interesting question relates to other evolutionary processes that act on gene clusters and putatively support the organism in the host-pathogens arms race. This thesis presents a workflow for the prediction and comprehensive regulatory and evolutionary analysis of secondary metabolism gene clusters.

1.9 Overview of this thesis

In **Chapter 2** I concentrate on the genome analysis of the rice pathogen *F. fujikuroi*. In a comparative genomics approach I predict gene families in *F. fujikuroi* and related *Fusarium* species and align the genome sequences to identify unique sequence features like supernumerary chromosomes that may have relevance to the respective host specificity. In combination with multiple protein classification tools I predict secreted effector proteins and take classical and non-classical secretion pathways into account. I then integrate available experimental data of transcriptomics, proteomics and epigenetic ChIP-seq experiments at virulence inducing conditions to get insight into the regulation of virulence genes and to determine the environmental stimuli that are necessary for their expression. The data is also used as co-expression evidence to identify pathway genes of known secondary metabolites such as gibberellins or fusarins.

In **Chapter 3** I investigate to which extent transposable elements contribute to the genomic differences observed in Chapter 2. Therefore, I predict transposable elements in *Fusarium* genomes and identify their distribution in the respective phylogenetic clades. I also determine evidence of repeat induced point mutation (RIP), a defense mechanism that inactivates transposable elements. Microarray data is used to identify active transposons in *F. fujikuroi*.

Based on the functional composition and observed co-regulation of secondary metabolism gene clusters in Chapter 2 I predict secondary metabolism gene clusters in 20 fungal species in **Chapter 4**. I make use of experimental data in *F. fujikuroi* and *F. graminearum* to identify co-expressed clusters, which presumably play a role in virulence. In order to identify significantly overrepresented, putative transcription factor binding sites I combine three motif prediction algorithms with a genome wide promoter analysis. According to the predictions and experimental evidence I also suggest additional members of previously identified secondary metabolism gene clusters.

In **Chapter 5** I ask for reasons of the discontinuous distribution of the predicted gene clusters in the 20 fungal genomes examined in Chapter 4. Using a protein database with similarity information of all available fully sequenced fungal genomes, I identify orthologous clusters outside the *Fusarium* phylum and determine evidence of horizontally transferred gene clusters. In addition, I identify evolutionary processes that modify gene clusters and may play a role in the adaption process during the host-pathogen arms race.

In the final **Chapter 6**, I discuss the impact of the presented findings on the scientific field and propose possible extensions and future projects.

Chapter 2

Predicting virulence factors by integrative genomics

In this chapter the comparative genome analysis of plant pathogenic *Fusarium* species with focus on the prediction of virulence genes and the identification of species and clade specific features is presented. Additionally, experimental data is integrated in order to analyze gene expression and gene regulation under virulence inducing conditions.

Most results of this chapter are published in:

- Wiemann P*, **Sieber CMK***, von Bargen KW*, Studt L, Niehaus EM, Espino JJ, Huß K, Michielse CB, Albermann S, Wagner D, Bergner SV, Connolly LR, Fischer A, Reuter G, Kleigrewe K, Bald T, Wingfield BD, Ophir R, Freeman S, Hippler M, Smith KM, Brown DW, Proctor RH, Münsterkötter M, Freitag M, Humpf HU, Güldener U, Tudzynski B.
  Deciphering the cryptic genome: genome-wide analyses of the rice pathogen *Fusarium fujikuroi* reveal complex regulation of secondary metabolism and novel metabolites.
  PLoS Pathog. 2013, 9, e1003475
- Studt L, Schmidt FJ, Jahn L, **Sieber CMK**, Connolly LR, Niehaus EM, Freitag M, Humpf HU, Tudzynski B.
  Two histone deacetylases, FfHda1 and FfHda2, are important for *Fusarium fujikuroi* secondary metabolism and virulence.
  Appl Environ Microbiol. 2013, 79, 7719-7734

*equal contributions

- Michielse CB, Studt L, Janevska S, **Sieber CMK**, Espino JJ, Arndt B, Humpf HU, Güldener U, Tudzynski B.
  The global regulator FfSge1 is required for expression of secondary metabolite gene clusters, but not for pathogenicity in *Fusarium fujikuroi.*
  Environ Microbiol. 2014 (Epub ahead of print)

- Pfannmüller A, Wagner D, **Sieber CMK**, Schönig B, Boeckstaens M, Marini AM, Tudzynski B.
  The general amino acid permease FfGap1 of *Fusarium fujikuroi* is sorted to the vacuole in a nitrogen-dependent, but Npr1 kinase-independent manner.
  Submitted.

2.1 Materials and methods

2.1.1 Sequence data

One of the major points of this chapter is the analysis of the *Fusarium fujikuroi* genome. The sequence data and annotation has been submitted to NCBI where it is accessible under the BioProject ID 185772. In addition, publicly available genome data and annotations of the GFC species *F. fujikuroi* strain B14 (BioProject ID: 171493), *F. circinatum* (BioProject ID: 41113) and *F. verticillioides* (BioProject ID: 15553) were used. Furthermore, unpublished genomes of *F. proliferatum* and *F. mangiferae* were utilized. The unpublished *F. proliferatum* strain NRRL62812 and the *F. fujikuroi* strain UCIM 1100 were used only for the phylogenetic analysis of the gibberellic acid cluster in Section 5.2.9. Genomic data of fungi of the Fusarium oxysporum species complex were also obtained from NCBI for *F. oxysporum* f. sp. lycopersici 4287 (BioProject ID: 18813) and 9 additional strains (Table A.2). *F. graminearum* genome data and annotation used are based on FGDB version 3.2 and the corresponding Pedant database [221, 239]. The assembly of *F. pseudograminearum* was received from NCBI (BioProject ID: 66583). In addition, the unpublished genome sequence and annotation of *F. asiaticum* were used. All further genomic and proteomic data used for ortholog analysis is based on Pedant databases represented in SIMAP [7, 175] and listed in Table A.2.

2.1.2 Whole genome alignments and protein comparisons

In order to determine the number of chromosomes and species specific genomic regions I calculated whole genome alignments using the suffix-tree algorithm MUMmer [51] with a cluster length of exact matches of at least 100 nt and at most 500 nt mismatches between two exact matches.

I obtained information of orthologs in terms of bidirectional best hits (BBHs) between two genomes from the protein similarity database SIMAP [7, 175]. To identify the degree of collinearity between two genomes I applied the tool Orthocluster [246] which calculates collinear blocks based on the location of genes and the information of BBHs. A collinear block was defined by at least three consecutive, orthologous genes, allowing one additional or missing gene in between. Unique proteins in species were identified in selecting proteins that do not have a bidirectional best hits to any other genome. Analogously, I define proteins with bidirectional hits to all species of a clade as core proteins for this clade.

2.1.3 Calculation of phylogenetic trees

I determined phylogenetic relationships between 21 *Fusarium* species, *Botrytis fuckeliana* and *Aspergillus nidulans* based on the genes encoding the RNA polymerase II subunits *RPB1* and *RPB2* of the transcription elongation factor gene *TEF1*α. Alignments of the amino acid sequences using Mafft [103] were calculated and a conversion to codon alignments with PAL2NAL [204] was performed. After that conserved blocks of the multiple nucleotide sequence alignments were identified using Gblocks [206] and concatenated for the phylogenetic tree calculation. I applied the maximum likelihood approach PhyML [81] with the HKY85 [89] substitution model and performed a bootstrap test with 1000 replicates to determine support of the resulting phylogenetic tree (Figure 2.1).

2.1.4 Transcriptomics- and proteomics- data

I analyzed microarray data of *F. fujikuroi* from experiments performed under nitrogen sufficient and deficient culture conditions. Two sources of nitrogen with different pH have been used: acidic glutamine (GLN) and alkaline nitrate (NO3).

The varying nitrogen availability and pH aim to simulate virulence inducing conditions in order to identify genes that are expressed during the plant infection process. Beside the wild type strain I analyzed gene expression of mutant strains of the histone deacetylase *Hda1* and the global regulator *Sge1*. All expression data was submitted to Gene Expression Omnibus (GEO) (Table 2.1). The RNA extraction was performed by the lab of Bettina Tudzynski from the University of Münster, the hybridization of the microarrays was done at Arrows Biomedical in Münster.

Gene expression data of *F. graminearum* was obtained from PlexDB [49]. The data comprises five time series experiments measuring gene expression during plant infection or conidiation [82, 131, 190, 200, 248] and seven case control studies investigating effects of transcription factor deletions [101, 130, 189], the impact of different growth conditions [75, 82, 189] and the expression profile of different phenotypes during infection [80] (Table 2.1). These experiments use the *F. graminearum* Affymetrix gene chip [82] which is based on the assembly version 1 and preliminary CDS annotations. In order to get expression values for the latest annotation version (3.2) BLAST [4] was used to map the probes on the current ORF-sequences, whereas only hits with 100% identity were accepted. All ambiguous probe set to ORF hits were removed.

In addition to transcriptomics data, I received quantitative proteomics data from whole cell protein experiments in *F. fujikuroi* that were performed by the lab of Michael Hippler from the University of Münster.

2.1.5 Expression data analysis

For normalization of expression data and summarization of probe-sets I applied the statistical computing environment R [174] and the implementation of the Robust Multi-array Average (RMA) algorithm of the affy R-package [78]. To determine significantly differentially expressed genes I fitted linear models for each gene and computed moderated t-statistics using the empirical Bayes method of the limma R-package [196]. P-value adjustment for multiple testing has been performed in calculating false discovery rates (FDR) using Benjamini-Hochberg procedure [17]. Genes with an absolute fold change above two and an adjusted p-value below 0.05

are classified as differentially expressed. In case of time series without control experiment, the first time point of the measurement has been taken as reference.

Table 2.1: Used expression data of *Fusarium graminearum* and *F. fujikuroi* . FG accession numbers correspond to PlexDB (`www.plexdb.org`), GSE identifier describe experiments submitted to Gene Expression Omnibus (`www.ncbi.nlm.nih.gov/geo`).

Accession No.	Experiment description	Reference
FG1	Fusarium transcript detection on Morex barley spikes using *Fusarium* Affy GeneChips	[82]
FG2	Expression Profiles in Carbon and Nitrogen Starvation Conditions	[82]
FG7	*Fusarium* gene expression profiles during conidia germination stages	[190]
FG10	Response to trichodiene treatment in *Fusarium graminearum*	[189]
FG11	Gene Regulation by *Fusarium* Transcription Factors *Tri6* and *Tri10*	[189]
FG12	*Fusarium graminearum* gene expression during crown rot of wheat	[200]
FG13	The transcription factor *FgStuAp* influences spore development, pathogenicity and secondary metabolism in *Fusarium graminearum*	[130]
FG14	DON induction media	[75]
FG15	*Fusarium graminearum* gene expression during wheat head blight	[131]
FG16	*Fusarium graminearum* gene expression in wheat stems during infection	[80]
FG18	Trichothecene synthesis in a *Fusarium graminearum Fgp1* mutant	[101]
FG19	Stage-specific expression patterns of *Fusarium graminearum* growing inside wheat coleoptiles with laser microdissection	[248]
GSE43745	Genome-wide analyses of *Fusarium fujikuroi* reveal complex regulation of secondary metabolism and new metabolites	[231]
GSE43768	Two histone deacetylases, *FfHda1* and *FfHda2*, are important for secondary metabolism and virulence in *Fusarium fujikuroi*	[202]
GSE53977	The global regulator *FfSge1* is required for expression of secondary metabolite gene clusters but not for pathogenicity in *Fusarium fujikuroi*	[144]

2.1.6 Analysis of ChIP-seq histone modification data

ChIP-seq experiments have been performed by Lena Studt from the University of Münster in cooperation with the lab of Michael Freitag from Oregon State University. Obtained reads have been mapped on the genome with Tophat2 [109]. Read densities and enrichment on gene regions were determined using EpiChip [90]. To quantify epigenetic modification levels I calculated normalized locus-specific

Table 2.2: Details of the used microarray data sets on conditions and strains. Experimental conditions and strains explored in expression data analysis. FG accession numbers correspond to PlexDB (`www.plexdb.org`), GSE identifier describe experiments submitted to Gene Expression Omnibus (`www.ncbi.nlm.nih.gov/geo`). Abbreviations used in heatmaps figures are given in the first column.

Condition Abbreviation	Case Condition	Control Condition	Accession-No
FG1_24h	Barley infection (24h)	Water control	FG1
FG1_48h	Barley infection (48h)	Water control	FG1
FG1_72h	Barley infection (72h)	Water control	FG1
FG1_96h	Barley infection (96h)	Water control	FG1
FG1_144h	Barley infection (144h)	Water control	FG1
FG2_c.starv	C nutrient deficient medium	Complete medium	FG2
FG2_n.starv	N nutrient deficient medium	Complete medium	FG2
FG7_2h	Conidiation (2h)	Conidiation (0h)	FG7
FG7_8h	Conidiation (8h)	Conidiation (0h)	FG7
FG7_24h	Conidiation (24h)	Conidiation (0h)	FG7
FG10_250Tri	Trichodiene medium	Normal medium	FG10
FG11_tri6	Tri6 deletion mutant	Wildtype	FG11
FG11_tri10	Tri10 deletion mutant	Wildtype	FG11
FG12_2dpi	Wheat infection (2d)	Complete medium	FG12
FG12_14dpi	Wheat infection (14d)	Complete medium	FG12
FG12_35dpi	Wheat infection (35d)	Complete medium	FG12
FG13_stua.cmc.24h	FgStuA deletion mutant during spore production (24h)	Wildtype during spore production	FG13
FG13_stua.wheat.72h	FgStuA deletion mutant during wheat infection (72h)	Wildtype during wheat infection (72h)	FG13
FG13_stua.secmet	FgStuA deletion mutant during secondary metabolism inducing conditions	Wildtype during secondary metabolism inducing conditions	FG13
FG14_agmat	Agmatine medium (DON inducing)	Glutamine medium (DON non-inducing)	FG14
FG15_wt.wheat.24h	Wheat infection (24h)	Water control	FG15
FG15_wt.wheat.48h	Wheat infection (48h)	Water control	FG15
FG15_wt.wheat.72h	Wheat infection (72h)	Water control	FG15
FG15_wt.wheat.96h	Wheat infection (96h)	Water control	FG15
FG15_wt.wheat.144h	Wheat infection (144h)	Water control	FG15
FG15_wt.wheat.192h	Wheat infection (192h)	Water control	FG15
FG16_rw	Radial growth	Infection front	FG16
FG16_sw	Senescent wheat	Infection front	FG16
FG16_yp	Perithecium formation	Infection front	FG16
FG18_put.fgp	Fgp1 deletion mutant on putrescine medium	Wildtype on putrescine medium	FG18
FG19_16hpi	Wheat infection (16h)	Wheat infection (0)	FG19
FG19_40hpi	Wheat infection (40h)	Wheat infection (0)	FG19
FG19_64hpi	Wheat infection (46h)	Wheat infection (0)	FG19
FG19_240hpi	Wheat infection (240h)	Wheat infection (0)	FG19
wt high NO3 vs. wt low NO3	Wild type on 120 mM NO3 (100%) medium	Wild type on 6 mM NO3 (5%)	GSE43745
wt high GLN vs. wt low GLN	Wild type on 60 mM Gln (100%) medium	Wild type on 6 mM Gln (10%)	GSE43745
Δ*sge1* lowGLN vs. wt low GLN	Δ*sge1* mutant on 6 mM Gln (10%) medium	Wild type on 6 mM Gln (10%)	GSE53977
Δ*sge1* high GLN vs. wt high GLN	Δ*sge1* mutant on 60 mM Gln (100%) medium	Wild type on 60 mM Gln (100%)	GSE53977
Δ*hda1* low GLN vs. wt low GLN	Δ*hda1* mutant on 6 mM Gln (10%) medium	Wild type on 6 mM Gln (10%)	GSE43768
Δ*hda1* high GLN vs. wt high GLN	Δ*hda1* mutant on 60 mM Gln (100%) medium	Wild type on 60 mM Gln (100%)	GSE43768

chromatin state (NLCS) values from 50 nt upstream and 1000 nt downstream of the start codons. I calculated probabilities of enriched regions to be a signal in applying the implemented curve fitting approach for distinguishing between background and signal. Genes with a signal probability above 0.95 were defined as significantly enriched. To compare chromatin modification states of different experimental conditions I applied a quantile normalization on all NLCS values.

The ChIP-seq data of the wild type is available in NCBI Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo) under the accession numbers GSM1122108, GSM1122109, GSM1122110 and GSM1122111. HDA mutant experiments are available in NCBI Short Read Archive (www.ncbi.nlm.nih.gov/sra) under the accession numbers SRR826542, SRR1011532, SRR1011533, and SRR1011534.

2.1.7 Prediction of secreted proteins

To cover the whole set of secreted proteins I applied five different bioinformatics tools in a pipeline approach to classify protein features.

To determine proteins that are secreted by the classical, signal recognition particle (SRP) mediated pathway, I predict SRPs using SignalP [167] with a cutoff S-score of 0.5. I furthermore determined the subcellular location of the proteins using the neural network approach TargetP. I selected proteins with a reliability class (RC) score below 4 for the compartments golgi and endoplasmatic reticulum (ER). Additionally, a prediction of extracellular target compartments has been done with Wolfpsort [95]. I excluded putative membrane bound proteins in predicting transmembrane domains using the HMM-based algorithm TMHMM [114]. To include non-classically secreted candidate proteins I used SecretomeP [16] which applies a feature based neural network on the predictions of SignalP, WolfPSort and TMHMM. I selected proteins with a neural network score above 0.6 and a SignalP S-score below 0.5 as non-classically secreted proteins.

2.1.8 Functional gene set enrichment analysis

For identifying functional categories of genes I utilized the FunCat catalogue [184] classification system implemented in the PEDANT database [221]. Significantly overrepresented categories in gene sets were identified using Fisher's Exact Test [65] (F-test) and the MGSA R-package [14]. I estimate the false discovery rate (FDR) for multiple hypothesis testing of the F-test by applying the Benjamini-Hochberg procedure [17]. Functional categories with adjusted p-value less than 0.05 were regarded as significantly overrepresented. In the functional analysis of the core proteome sets in Section 2.2.9 only categories with an MGSA estimate above 0.5 were accepted.

2.2 Results

2.2.1 Assembly and comparative genome analysis of *F. fujikuroi*

The rice pathogen *Fusarium fujikuroi* is the causal agent of bakanae disease and producer of many secondary metabolites including gibberellins and harmful mycotoxins [117]. By genome sequencing and comparative analysis I intended to identify secondary metabolism gene clusters and virulence related genes. Sequencing has been done by MWG biotech using 454 technique resulting in 0.94 Gb of raw sequence reads, which could be assembled into 12 scaffolds with a total sequence length of 43.9 Mb and a 19 fold average sequence coverage. Gene prediction resulted in 14,813 protein coding genes. The number of genes corresponds to the average of predicted gene models in other *Fusarium* genomes (Table 2.3). The highest amount of genes can be found in *F. oxysporum* (17458) and the lowest number in *F. graminearum* (13826). Interestingly, the *F. oxysporum* genome exhibits the lowest gene density (284.53 genes/Mb) whereas in *F. graminearum* the highest number of genes per megabase (379.36 genes/Mb) among the compared genomes was found (Table 2.3). I performed a functional prediction analysis of all coding sequences using the FunCat [184] catalogue of protein function which is part of the PEDANT [221] pipeline. I determined proteins related to disease, virulence and defense which is similar to the proportion of other *Fusarium* genomes (3.7% to 4%). The highest absolute numbers were determined in *F. oxysporum* (645, 3.7% of genes) and *F. solani* (637, 4% of genes). The rice pathogen *F. fujikuroi* has 545 virulence classified genes (3.7%) (Table 2.4).

In recent studies a mapping of sequence contigs to physical chromosomes has been performed in *F. verticillioides* using an optical map [132],[241]. I used this information to estimate the amount of chromosomes in the closely related *F. fujikuroi* in aligning the assembled 12 contigs to the experimentally verified chromosome sequence of *F. verticillioides* using Mauve [48] and Mummer [51]. Both methods align eleven *F. fujikuroi* contigs to the eleven *F. verticillioides* chromosomes. The result of the alignment calculated by Mummer is depicted as dot-plot in figure 2.2. The plot illustrates the overall collinearity between the two genomes at which red dots indicates an orthologous alignment and blue dots illustrate sequence inversions. The magnification of supercontig 12 shows, that this smallest contig could not be mapped on a chromosome of *F. verticillioides*, suggesting an additional, putative dispensable chromosome. The size of supercontig 12 in *F. fujikuroi* is 693 kb and contains 173 predicted genes. The vast majority of 139 genes (80%) show no significant sequence similarity (blastp E-value $<$ 1e-25) to annotated Swiss-Prot proteins. However the proportion of proteins with functional annotation exhibits a significant overrepresentation of the functional category "guidance of longitudinal cell extension and cell migration" (P-value < 0.01) according to the FunCat [184] database of annotated protein functions.

Beside the varying number of chromosomes there are further notable differences worth noting. The fourth contig in *F. fujikuroi* is around 1.1 Mb shorter compared to the fourth chromosome in *F. verticillioides.* The dot-plot (Figure 2.2) illustrates missing parts of 285 kb and 820 kb on either side of the contig, respectively. I investigated the functional make-up of the 408 unique genes in *F. verticillioides* in these parts and determined a significant functional overrepresentation of the FunCat categories "secondary metabolism" , "detoxification" and "metabolism of melanin" (P-value < 0.01). Three additional genomic regions are also absent in *F. fujikuroi* but present in *F. verticillioides.* On chromosome VII a region of 70 kb comprises 29 genes, eight unique genes are located on a segment of 22 kb on chromosome III and a segment of 12 kb on chromosome V consist of six genes. No functional category is enriched among the 33 unique genes as the majority has no similarity to a protein with annotated function.

The amount of chromosomes and length of chromosome IX have been experimentally verified by Wiemann *et al.* [231] using PCR and contour-clamped homogeneous electric field (CHEF) gel electrophoresis experiments.

To infer the phylogenetic relationship of *F. fujikuroi* to the other available *Fusarium* species I computed a phylogenetic trees using phyML [81]. For each species I selected genes of the DNA-directed RNA polymerase II subunits *RPB1*, *RPB2* and the translation elongation factor 1 alpha (*TEF1α*) and calculated alignments based on their nucleotide sequence. To root the tree I included *A. nidulans* and *B. fuckeliana* as outgroups. In the resulting tree (Figure 2.1) the *Fusarium* species complexes "*Gibberella fujikuroi* species complex" (GFC), "*Fusarium oxysporum* species complex" (FOC) and "*Fusarium graminearum* species complex" (FGC) group together. Moreover in the GFC the three geographic clades can be identified as monophyletic groups.

2.2.2 A comparative analysis of gene families in *Fusarium* species reveals an increased amount of transcription factors in *F. fujikuroi*

Beside differences in the genome sequence I am interested in variations of gene families between fungi with different host specificities. Specific transcription factors may be involved in regulation genes in involved in virulence and host-interaction. InterProScan [244] is a functional classification tool to predict functional domains in protein sequences. I applied the method for the identification of transcription factors in predicting binding domains in the amino acid sequence of proteins. In *F. fujikuroi*, 950 TFs were determined. This number is significantly higher compared to *F. verticillioides* (640), *F. mangiferae* (643), *F. circinatum* (841), and *F. oxysporum f.sp. lycopersici* (876) but very similar to the closest examined relative *F. proliferatum* (966) (Table 2.4). The higher number of TFs in *F. fujikuroi* is due to the TFs of the group "fungal-specific TF / ZN(2)C6 fungal type DNA binding domain" represented by the Interpro domains IPR007219 and IPR001138 (Table FFUJ-TF-Table). This family of TFs comprises 235 in *F. fujikuroi* and 208 in *F. mangiferae* compared to only 90 in *F. verticillioides* and 144 in *F. graminearum.* Interestingly *F. fujikuroi* has 53 species specific TFs that do not have an ortholog (less than 60% sequence identity) in other *Fusarium* species. 33 among these are characterized as ZN(2)C6 TFs, which are known as pathway specific regulators of secondary metabolism [42, 63, 238].

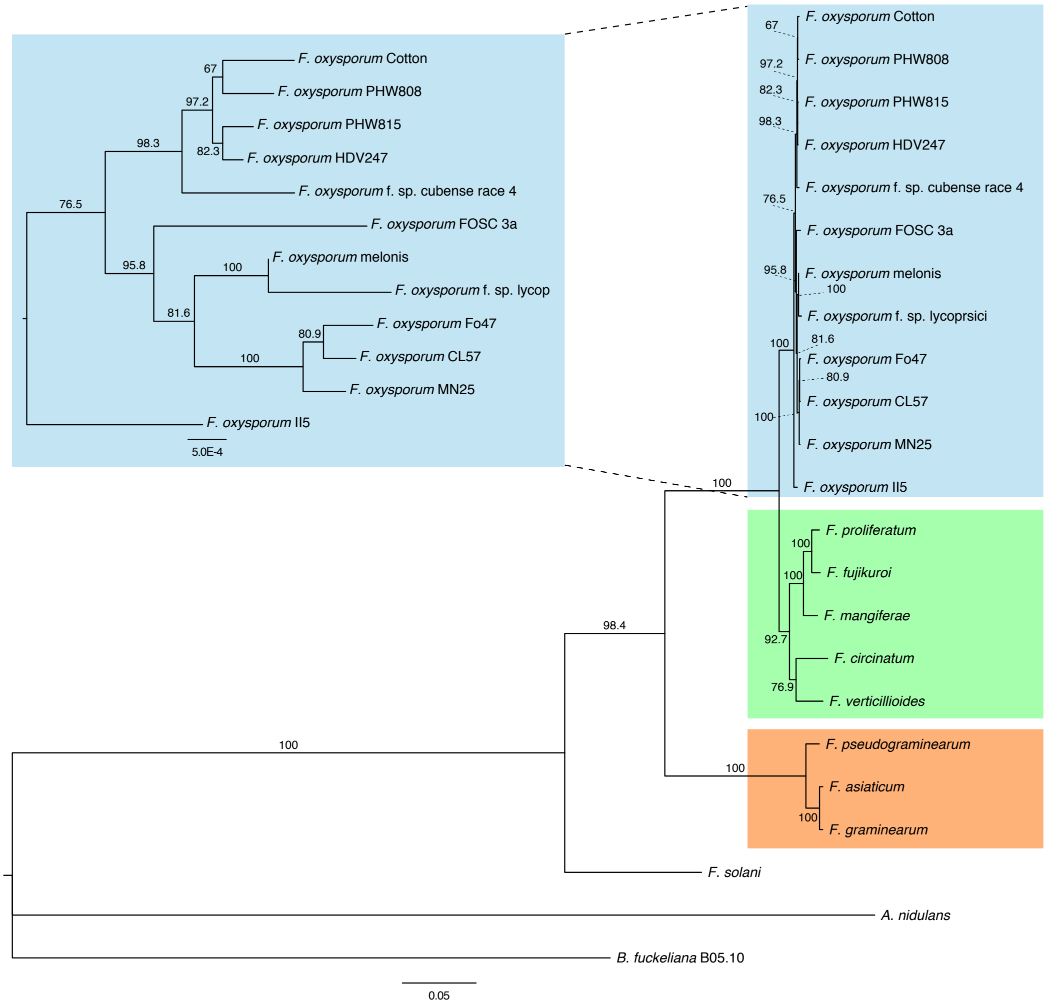

Figure 2.1: Phylogenetic tree of available *Fusarium* genomes. *Gibberella fujikuroi* species complex (GFC) is highlighted in green, *Fusarium graminearum* speceis complex (FGC) in orange and *Fusarium oxysporum* species complex (FOC) in blue. FOC subtree is additionally magnified on the top left. Midpoint rooted tree is calculated using a concatenated alignment of the nucleotide sequences of the DNA-directed RNA polymerase II subunits *RPB1*, *RPB2* and the translation elongation factor 1 alpha (*TEF1α*). Percentage values of 1000 bootstraps are given at each branch. Scale bar indicates average substitutions per site.

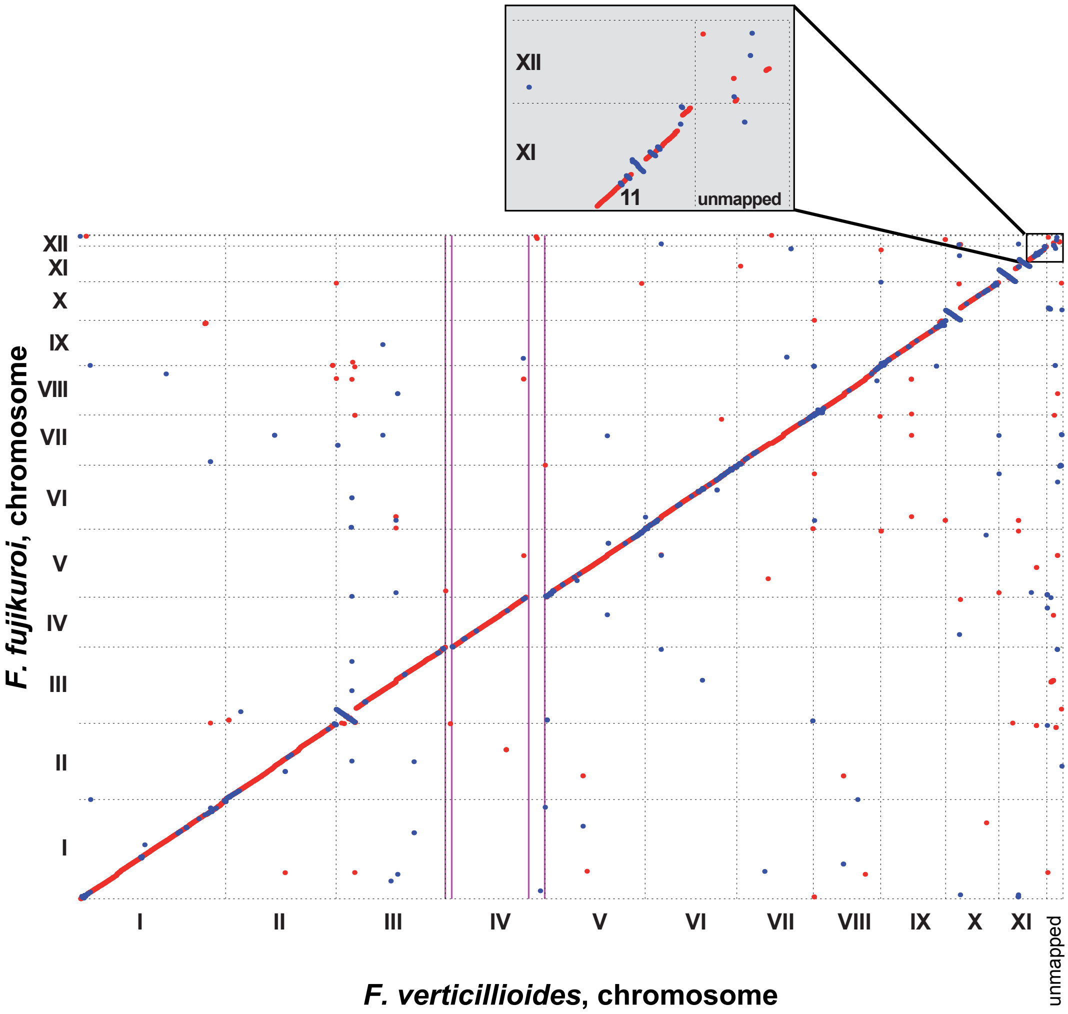

Figure 2.2: Whole genome alignment of *F. fujikuroi* and *F. verticillioides* illustrated as dot plot. A dot represents alignment between the respective sequence areas on the x-axis (*F. verticillioides*) and y-axis (*F. fujikuroi*). Red dots indicate forward alignments, blue dots illustrate inversions relative to the *F. verticilliodes* chromosomes. Gaps between the genomes are emphasized by vertical solid bars. Supernumerary chromosome XII of *F. fujikuroi* is magnificated above. Figure is adopted from Wiemann *et al.* [231].

2.2.3 The global regulator Sge1 influences expression of secondary metabolism gene clusters

Secondary metabolism is regulated not only by specific transcription factors but also by global regulators like Fgp1 in *F. graminearum* [101]. To identify the targets of the Fgp1 ortholog *Sge1* in *F. fujikuroi* I analyzed gene expression data of a Δ*sge1* (FFUJ_07864) mutant strain. Growth conditions of low and high acidic nitrogen medium were chosen as secondary metabolism genes are known to be induced in these conditions according to previous microarray experiments (Section 2.2.6). The regulator *SGE1* itself shows also a significantly increased (1.33 fold change, P-value = 0.004) gene expression under high nitrogen conditions.

Under low nitrogen conditions I determined a significant change in gene expression in 84 genes, whereas most of the genes were down regulated (70) and only a small portion (14) showed a significant increase. However, I found genes involved in the metabolism pathways of diterpenes and isoprenoid that are significantly enriched in the set of repressed genes according to a FunCat [184] analysis.

Under high nitrogen and *SGE1* inducing conditions a significant difference in gene expression could be determined in 1357 genes. Especially ribosomal RNA related functions such as rRNA processing and rRNA synthesis are significantly (P-value $<$ 2.79e-7) enriched and proteins related to heat shock response could be observed (P-value = 8.81e-3) in the repressed set of 805 genes. The 552 up-regulated genes show an enrichment of the FunCat [184] category “Transport of ATPases”.

Beside targets in primary metabolic pathways and housekeeping genes I am interested in the regulation of secondary metabolism gene clusters. Under low nitrogen conditions where the gibberellic acid gene cluster is expressed in the wild type, a significant decrease in gene expression in the Δ*sge1* mutant can be seen (Figure 2.6A, see below). Under acidic high nitrogen conditions where the apcidin (Figure 4.3A, see below), fusaric acid and bikaverin (Figure 2.7A, see below) clusters are expressed a significant lower expression rate is observed in the mutant. Interestingly, the deletion of the *SGE1* gene has not an inhibiting effect in all gene clusters. A significant increase in gene expression in 7 of the 11 genes of the fumonisin cluster can be seen.

Expression data analysis suggest that Sge1 is a positive regulator of the gibberellic acid, apicidin, fusaric acid and bikaverin gene clusters and has an inhibiting effect on the fumonisin gene cluster. The abundance of the respective metabolite under these conditions were experimentally confirmed by Michielse *et al.* [144].

	F. fujikuroi	*F. proliferatum*	*F. mangiferae*	*F. verticillioides*	*F. circinatum*	*F. oxysporum*	*F. graminearum*	*F. solani*
Genome size (Mb)	43.9	45.2	45.6	41.8	44.3	61.4	36.4	51.3
GC-content	47.42%	48.12%	48.82%	48.62%	47.26%	47.28%	48.04%	50.73%
Protein coding genes	14839	16224	16261	14180	15022	17458	13826	15702
Gene density (Number of genes per Mb)	337.97	358.85	356.77	339.43	339.42	284.53	379.36	306.16
Average intergenic distance (kb)	1.4	1.2	1.3	1.5	1.5	2.1	1.1	1.7
Percent coding	53.40%	55.31%	52.59%	48.32%	48.44%	39.79%	56.75%	49.22%
GC-content coding	51.63%	51.59%	51.62%	52.14%	51.75%	52.00%	51.57%	54.49%
Average gene length (kb)	1.5	1.4	1.3	1.3	1.3	1.2	1.4	1.4
Mean protein length (aa)	484.7	474.5	447.8	419.6	436.5	409.7	453.1	479.6
Exons	41652	45107	43543	38477	41023	46670	38453	48203
Average exon length (bp)	518.11	512.13	501.81	463.96	479.2	459.87	488.83	468.56
Exons/gene	2.81	2.78	2.68	2.71	2.73	2.67	2.78	3.07
Average intron length (bp)	69.66	66.05	77.61	96.16	68.52	101.12	76.62	81.73

Table 2.3: Comparison of genome properties in *Fusarium*

2.2.4 Expression of secreted protein genes is affected by nitrogen and pH

Secreted proteins have an impact on plant pathogenicity in terms of effector proteins [209]. In order to determine the complete set of secreted proteins associated with classically and non-classically secretion pathways I applied a pipeline using five bioinformatic approaches on the predicted set proteins of all seven *Fusaria.* Around 9% (1336) of the *F. fujikuroi* proteome were predicted as secreted candidates. 126 of the 1336 proteins are classified as part of a non-classical secretory pathway. The total amount of secreted proteins is similar in the closely related

	F. fujikuroi	*F. proliferatum*	*F. mangiferae*	*F. verticillioides*	*F. circinatum*	*F. oxysporum*	*F. graminearum*	*F. solani*
tRNA genes	289	311	304	293	296	305	319	286
Secreted proteins (SP)	1336	1405	1422	1239	1262	1541	1264	1337
SPs (% of Proteome)	9.00%	8.66%	8.74%	8.74%	8.40%	8.83%	9.14%	8.51%
Unique secreted proteins (simap-ratio < 0.6)	72	91	7	168	203	416	450	756
Small secreted proteins (< 300aa) (SSP)	512	527	586	531	567	694	548	510
SSP (% of SP)	38.32%	37.51%	41.21%	42.86%	44.93%	45.04%	43.35%	38.15%
Non-classically SPs	126	127	165	150	168	208	204	190
Non-classically SSPs (< 300 aa)	50	50	78	87	87	126	75	77
Transporters	857	929	679	840	895	995	673	979
ABC transporters	65	68	33	70	65	77	63	73
Aminoacid permeases	99	105	70	103	108	126	86	126
Ammonium permeases	4	5	5	5	5	5	4	2
Transcription factors	950	966	643	640	841	876	726	933
Unique transcription factors (simap-ratio < 0.6)	53	61	314	77	118	253	726	530
Coverage by repeats	1.8	0.6	0.6	0.5	1.6	13	0.5	3
Coverage by repeats	4.08%	1.39%	1.35%	1.28%	3.68%	21.24%	1.31%	5.85%
Transposable elements	2.20%	0.41%	0.54%	0.47%	1.08%	4.76%	0.33%	1.64%
Disease, virulence and defense genes	545	627	584	534	590	645	515	637
Disease, virulence and defense genes (%)	3.67%	3.86%	3.59%	3.77%	3.93%	3.69%	3.72%	4.06%

Table 2.4: Comparison of predicted gene families in *Fusarium* genomes

species *F. mangiferae* (1422), *F. circinatum* (1262) and *F. verticillioides* (1239), but differs compared to the expanded secretome of *F. oxysporum* which has 15% more predicted secreted proteins in comparison to *F. fujikuroi.* The proportion of secreted proteins of the whole proteome is similar in all species and ranges from 8.4% (*F. circinatum*) to 9.1% (*F. graminearum*). *F. oxysporum* encodes the highest amount of potentially small secreted (694). The highest amount of predicted non-classically secreted proteins can also be found in *F. oxysporum* (208) which is similar to *F. graminearum* (204).

To determine the environmental conditions that promote the expression of secreted proteins I utilized the available microarray expression data and determined the amount of differentially expressed secreted protein genes. In *F. fujikuroi* 136 differentially expressed secreted protein encoding genes were found in the secondary metabolism inducing high nitrogen *in vitro* condition compared to low nitrogen growth medium. Vice versa, 160 predicted secreted proteins were up-regulated in the low nitrogen condition. In *F. graminearum* expression data of *in planta* experiments is available. I found no significantly up-regulated (fold change > 2, P-value < 0.05) secreted protein genes during barley infection (FG1) after 24h and 60 proteins at 48h. The amount increased considerably to 213 after 72h, 235 (96h) and 282 after 144h. Interestingly the amount of down-regulated secreted proteins increase between the last two time points of 72h and 96h from 18 to 125. A similar observation could be made during wheat infection (FG15) where only 10 secreted proteins were significantly up-regulated after 24h but 237 after 48h of infection. A peak was reached after 96h where 384 secreted proteins were up-regulated. At this time point also the amount of down-regulated proteins increased from 47 (72h) to 193 proteins (96h) (Figure 2.3). While a strong correlation between the expression profiles of secreted proteins during the infection of barley (FG1) and wheat (FG15) after 72h was observable (Pearson $= 0.85$, Spearman $= 0.84$, P-value < 0.01), only a moderate correlation can be calculated between these time points and the time point after 64h while growing in wheat coleoptiles (FG19) (Pearson $= 0.40$ and 0.37, Spearman $= 0.40$ and 0.39, P-value < 0.01) (Figure 2.3).

2.2.5 Prediction of transporter proteins and general amino acid permeases

Transporter proteins play a role in the secretion process of proteins and non-ribosomal products such as secondary metabolites. I used InterProScan [244] to determine functional domains and to classify the transporter type. Interestingly *F. oxysporum*, comprising the largest secretome (1541), has also the highest amount of transporter proteins (995). This is comparable to the number of 979 transporters in *F. solani* with a considerably smaller secretome (1337). In *F. fujikuroi* (857), *F. verticillioides* (840) and *F. circinatum* (895) the amount of transporters is similar, *F. graminearum* (673) and *F. mangiferae* (679) exhibit the smallest count of transporter proteins (Table 2.4).

Nitrogen is important for fungal growth and development. All fungi are able to use amino acids as nitrogen source which are imported into the cell via amino

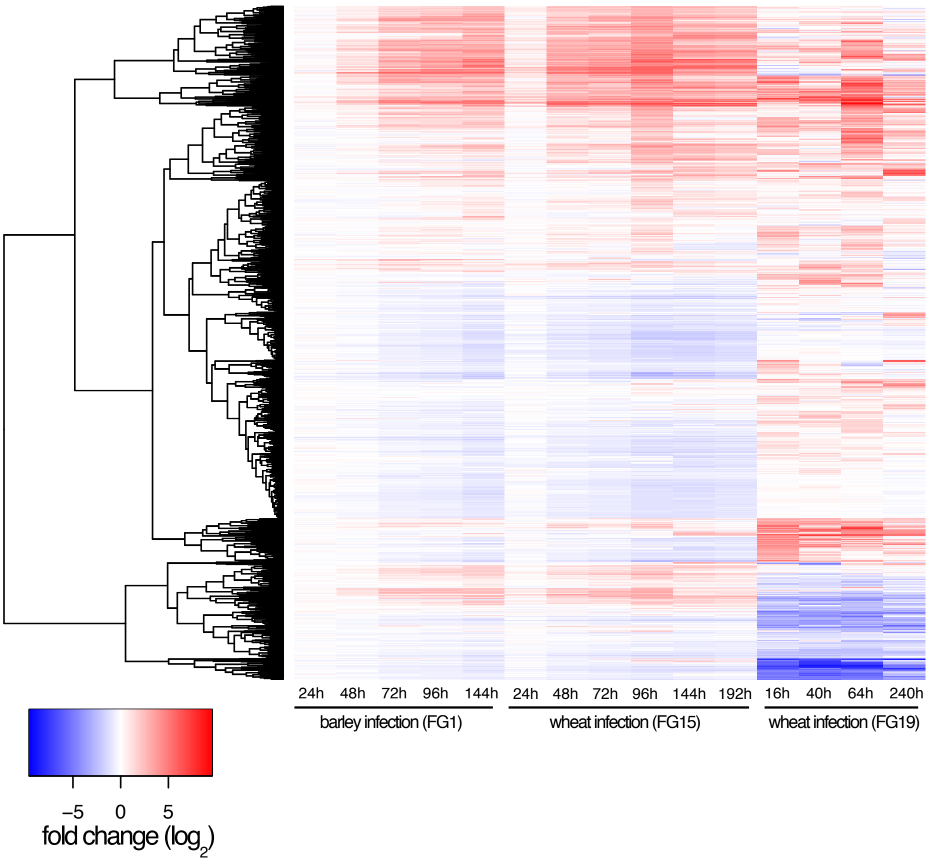

Figure 2.3: Gene expression heatmap of predicted secreted proteins in *F. graminearum* during infection of barley (FG1) and wheat (FG15, FG19). Fold changes of gene expression intensities of the respective time point of measurement and the control condition are depicted in red (increase) and blue (decrease). Dendrogram on the left indicates genes with similar gene expression profile.

acid permeases (AAPs). I predicted the highest amount of AAPs in *F. oxysporum* and *F. solani* (126, respectively) whereas only 70 AAPs could be found in *F. mangiferae*. In *F. fujikuroi* 99 AAPs were found.

In order to identify putative functional general amino acid permeases (GAPs) in *Fusarium* that are able to take up most or all of the 20 common amino acids like the *ScGAP1* in *Saccharomyces cerevisiae* [177] I searched for predicted AAPs in *F. fujikuroi* with similarity to *ScGAP1* and experimentally verified GAPs of *Candida albicans* [113], *Penicillium chrysogenum* [213] and *Neurospora crassa* [136].

A BLAST [4] search revealed 19 of the 99 *F. fujikuroi* AAPs with significant similarity (E-value $<$ 1e-50, identity $>$ 25%) to the four reference GAPs and were selected as putative GAP candidates. To infer the evolutionary relationship between the candidates and references I constructed a maximum likelihood phylogenetic tree (Figure 2.4) and selected closely related proteins in extracting the members of the smallest subtree that contains all four known reference GAPs (Figure 2.4, blue frame). This resulted in a refined candidate set of eleven putative *F. fujikuroi* GAP proteins. Experimental investigation by Pfannmüller *et al.* resulted in fully restored growth of a yeast Δ*GAP1* mutant upon transformation suggesting FFUJ_09118 as Gap1 ortholog (Pfannmüller *et al.*, submitted).

2.2.6 Availability of nitrogen and pH levels influence gene expression and protein abundance

Expression of virulence genes and biosynthesis of secondary metabolites is induced by environmental factors like pH or availability of nitrogen. Recent studies show that global regulators like AreA and PacC induce or inhibit expression of SM genes under different nitrogen and pH conditions [97, 138, 210]. I analyzed data of microarray, proteomics and epigenetic ChIP-seq experiments to get insight into the regulation of genes involved in virulence and SM. In total 3,117 and 3,242 genes are significantly up-regulated under acidic (glutamine, Gln) and alkaline (nitrate, NO3) low-nitrogen conditions, respectively (Fold change $>$ 1, P-value $<$ 0.05). 2,494 of the genes were up-regulated under both nitrogen conditions, whereas 560 genes showed significant higher expression only in acidic conditions and 717 only in the alkaline condition. A different regulation in the two nitrogen conditions could be observed in 63 genes that were up-regulated in the acidic low nitrogen condition but repressed in the alkaline condition and in 31 genes that showed the opposite expression pattern. I performed a FunCat [184] analysis on the set of differentially expressed genes and determined an overrepresentation of genes involved in transport, carbon metabolism and detoxification among nitrogen induced genes. Among the genes that are up-regulated under high-nitrogen conditions I determined a significant overrepresentation (P-value $<$ 0.05) of genes involved in

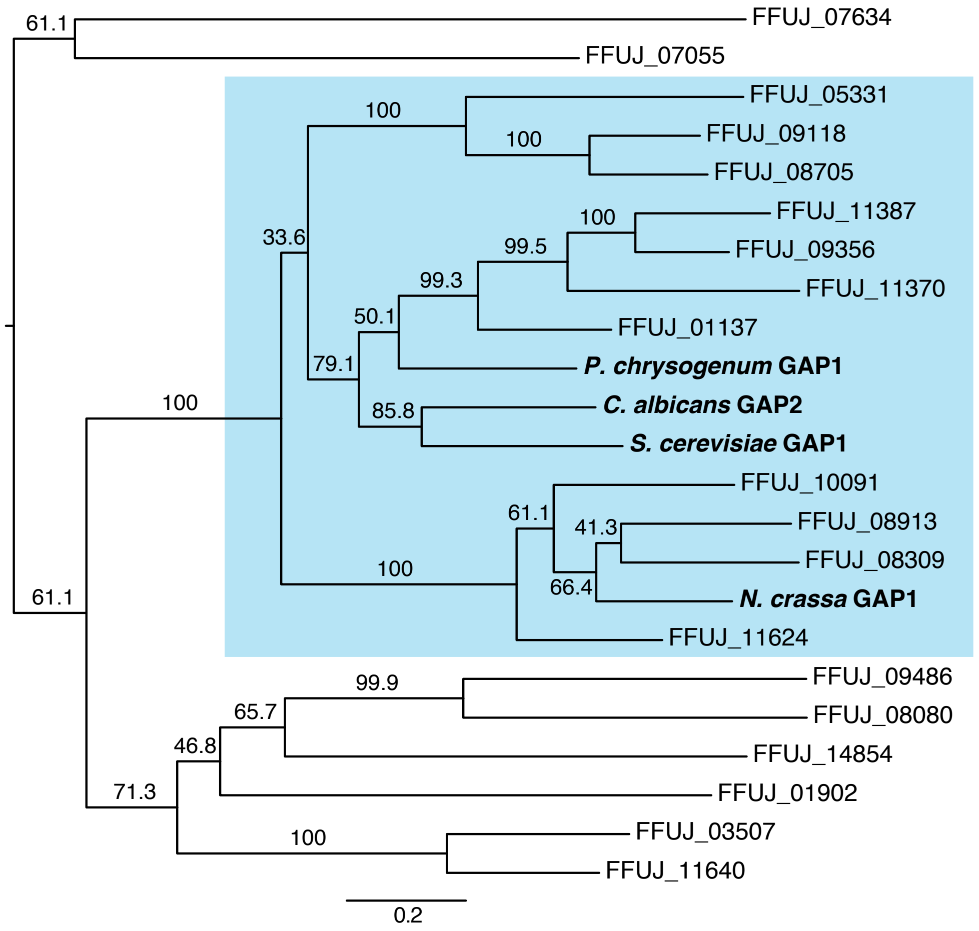

Figure 2.4: Maximum likelihood phylogenetic tree of general amino acid permeases of *S. cerevisiae*, *C. albicans*, *P. chrysogenum*, *N. crassa* and 19 predicted amino acid permeases of *F. fujikuroi*. Percentage values of 1000 bootstraps are given at each branch. Scale bar indicates average substitutions per site. Subtree highlighted by blue box contains putative *F. fujikuroi* GAP target genes that were selected for experimental verification by Pfannmüller *et al.* (submitted). Figure is adopted and modified from Pfannmüller *et al.* (submitted).

primary metabolism, e.g. amino acid metabolism, DNA processing, transcription, transport, protein synthesis and protein folding and modification. The complete set of genes induced under either acidic or alkaline nitrogen conditions comprised 3,860 and 4,192 genes, respectively. 3,021 of those genes were up-regulated under both conditions, whereas 808 were up-regulated only in the acidic and 1,108 only in the alkaline high nitrogen condition. Among the differentially expressed genes 13 polyketide synthases (PKS), 11 non-ribosomal peptide synthase (NPS) and one dimethylallyl tryptophan synthases (DMATS) enzyme were differentially expressed. Further, two diterpene cyclases (DTC), two sesquiterpene cyclases (STC) and the type III-PKS containing gene showed significant changes in their expression pattern.

After showing that the transcriptome is affected by different nitrogen- and pH-levels I am interested if the same observations can be made on the proteomics level. Therefore, I received quantitative proteomics data from whole cell protein experiments performed by the collaborative lab of Michael Hippler from the University of Münster. Two replicates of measurements under acidic low and high conditions have been performed. In total 1,644 proteins could be quantified in both replicates, whereas 261 proteins have been discarded due to a too high divergence between the replicates. Among the remaining 1383 quantified proteins, 347 proteins were up-regulated (fold change > 2) under high acidic nitrogen condition, whereas 418 showed a down-regulation of at least two fold. In comparing both experimental approaches a moderate correlation between the fold changes of transcripts and the protein ratios was observed (Pearson = 0.45, Spearman = 0.36, P-value < 0.01) (Figure 2.5). Among the proteins that are up-regulated in the high acidic nitrogen condition according to transcriptomics (fold change > 2, P-value < 0.05) and proteomics (fold change > 2) an overrepresentation of proteins involved in “nitrogen, sulphur and selenium metabolism” and “secondary metabolism” could be determined (P-value < 0.05). Genes expressed under acidic low nitrogen conditions according to proteomics and transcriptomics were significantly enriched for “virulence, disease and defence” and for “secondary metabolism”. It is notable that genes connected to secondary metabolism are not significantly overrepresented in the set of differentially expressed genes considering transcriptomics data alone.

Interestingly, when concentrating on known and predicted secondary metabolism cluster genes alone, I determined a strong correlation of transcriptomics and proteomics fold changes (Pearson = 0.82, Spearman = 0.81, P-value < 0.01). Especially proteins of the gibberellic acid, fumonisin and bikaverin cluster exhibited up regulation under acidic low nitrogen conditions, apicidin, fusaric acid and fusarin C clusters correlate in their expression in the acidic high nitrogen condition (Figure 2.5).

Among the cluster genes that are exclusively differentially expressed in the alkaline nitrogen condition according to transcriptomics data, no significant differences in proteomics data between different acidic nitrogen concentrations could be determined.

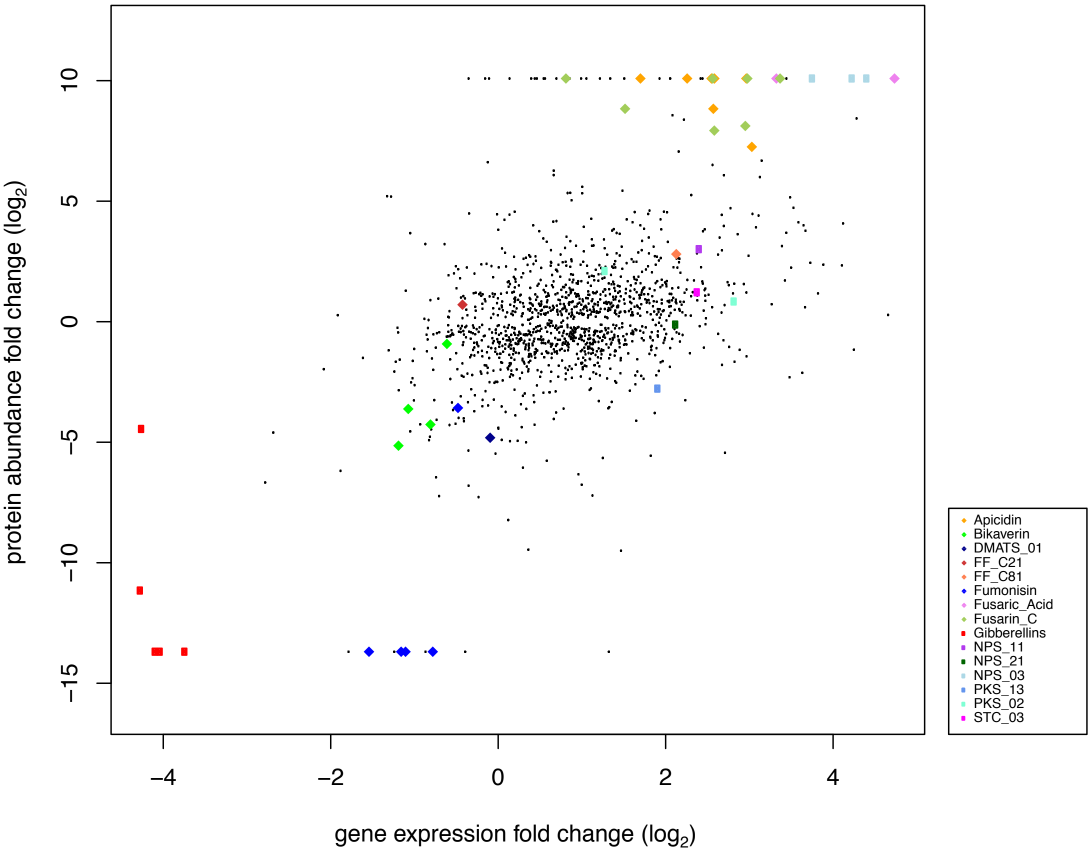

Figure 2.5: Comparison of transcriptomics- and proteomics- fold changes of *F. fujikuroi* grown under secondary metabolism inducing high and low acidic nitrogen conditions. Colored dots represent genes of known and predicted secondary metabolism gene clusters. Correlation between transcriptomics and proteomics is higher among secondary metabolism genes than under the whole measured set.

2.2.7 Nitrogen and pH influence genome wide chromatin landscape

The determined change in protein abundance and gene expression of secondary metabolism genes due to the availability of nitrogen rises the question to which extent epigenetic regulation is involved like previously observed in *F. verticillioides* [180] or *A. fumigatus* [220]. In cooperation with Lena Studt and the group of Michael Freitag I received data of ChIP-seq experiments performed under acidic low and high nitrogen concentrations that are equivalent to the experimental conditions of the microarray and proteomics experiments. In the ChIP-seq experiments two specific antibodies for activating modifications (H3K9ac and H3K4me2) and one for silencing modifications (H3K9me3) have been used. In order to quantify the level of modification, I calculated normalized locus specific chromatin state (NLCS) values [90] of ChIP-seq reads for each gene.

In comparing both nitrogen conditions I determined 1561 genes that were exclusively enriched for H3K9ac acetylation patterns in the nitrogen rich and 486 in the nitrogen starving condition. On the genome wide comparison no correlation between fold changes in NLCS values and gene expression fold changes could be determined. Interestingly, a moderate correlation (Pearson = 0.41, Spearman = 0.27, P-value < 0.01) was calculated on the set of previously predicted secondary metabolism genes. Particularly the genes of the gibberellic acid gene cluster exhibit an enrichment of the histone mark H3K9ac in the low nitrogen condition, while it is almost absent in the high nitrogen condition (Figure 2.6B). This observation can also be made in genes of the bikaverin (Figure 2.7B) and fumonisin gene cluster which both exhibit increased gene expression under low nitrogen conditions. In contrast to the acetylation, activating methylation marks could only be observed for two genes of the gibberellic acid gene cluster. Here two cytochrome P450 genes (*P450-2* and *P450-4*) were enriched for H3K4me2 exclusively in the low nitrogen condition (Figure 2.6B). Considering the whole genome only 1.7% (257 genes) of all genes showed an exclusive enrichment for the H3K4me2 mark under low nitrogen and only 1.3% (194 genes) under the nitrogen rich condition. Comparing NLCS fold changes of the activating H3K4me2 mark to transcriptomics data a correlation neither in all nor in cluster genes alone was determined.

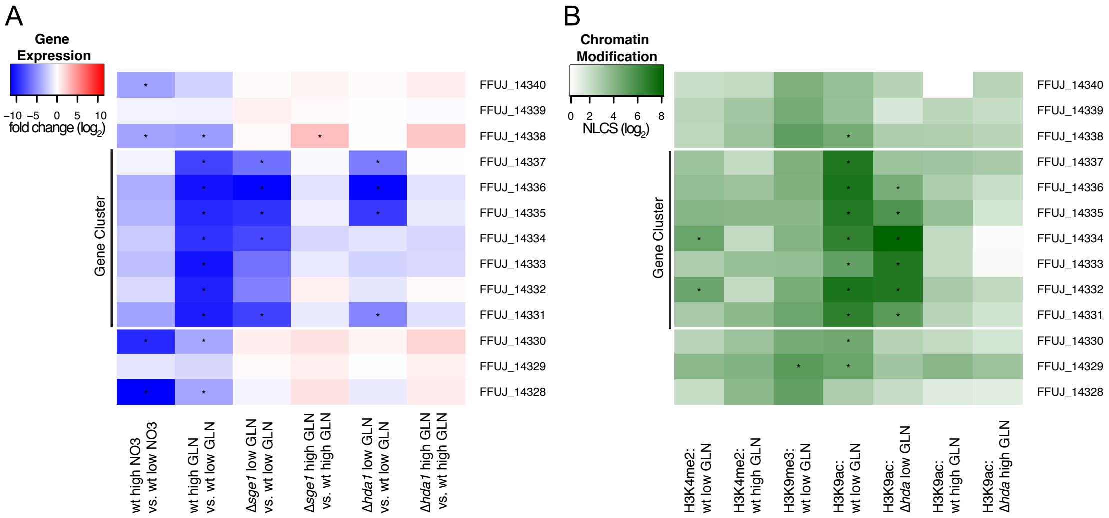

Figure 2.6: Gene expression and epigenetic modification of the gibberellic acid gene cluster and adjacent genes of *F. fujikuroi* wild type (wt), Δ*hda1* and Δ*sge1* mutant under limiting (low) and sufficient (high) glutamine (GLN) and nitrate (NO3) conditions. Genes are listed in chromosomal order on y-axis, gene cluster is indicated by vertical black bar. **A:** Fold changes (log_2 scale) of gene expression between two experimental conditions. Up-regulated genes compared to the control condition are depicted in red, down-regulated genes in blue. Asterisk indicate significant changes in gene expression (fold change > 2, P-value < 0.05). **B:** NLCS values (log_2 scale) indicating degree of chromatin modification for each experimental condition. Significant signals are indicated by asterisk (signal probability > 0.95).

2.2.8 Histone deacetylase Hda1 affects chromatin structure and gene expression of secondary metabolism gene clusters in *F. fujikuroi*

In the previous section it was shown that the chromatin landscape is not static but changes upon external stimuli like pH or the availability of nitrogen. Chromatin modifying enzymes modify histones and change the chromatin structure. Histone acetyltransferases (HATs) and histone deacetylases (HDAs) for example add and remove acetyl groups of histones and therefore influence the accessibility of DNA [11].I am now interested in the regulatory impact of the histone deacetylase Hda1 (FFUJ_09787) on histone acetylation and which particular cellular functions are affected. In collaboration with Lena Studt *et al.* [202], I therefore analyzed both gene expression and ChIP-seq data of the *F. fujikuroi* wild type and Δ*hda1* mutant strain. The strains were grown under secondary metabolism inducing conditions of high and low acidic nitrogen concentrations [202]. To find the differences in the acetylation pattern between the mutant and wild type strain I selected genes with a two fold difference in NLCS values and a signal probability above 95%. In the Δ*hda1* mutant strain I found a significant higher signal of acetylation marks under low nitrogen conditions in 450 genes (above two fold change in NLCS values, signal probability > 0.95) but also a significant lower signal in 411 genes compared to the wild type strain. Among the genes that are exclusively acetylated in the mutant, I found a significant enrichment of proteins related to transport processes (P-value = 0.018, FunCat category 20.03 [184]). An even larger difference between mutant and wild type could be observed in the high nitrogen condition. Here 492 genes exhibit an exclusive acetylation mark in the mutant, whereas 225 genes are acetylated only in the wild type. Among the differentially acetylated genes in the mutant I was able to determine three significantly enriched pathways of the "cellular signaling" FunCat category. Proteins related to transmembrane receptor protein tyrosine kinase signaling pathways (FunCat category 30.05.01.12) as well as genes involved in cAMP/cGMP mediated signal transduction (30.01.09.07) and the MAPKKK cascade (30.01.05.01.03) could be determined. Moreover endocytosis related proteins were significantly enriched (P-value = 3.11e-5). Interestingly, under both nitrogen conditions no significant overrepresentation of FunCat categories could be determined among genes that are exclusively acetylated in the wild type.

In analyzing gene expression data that was measured under equivalent experimental conditions a significant differential expression could be observed in the gibberellic acid gene cluster. In total four out of seven genes (FFUJ_14331, FFUJ_14335, FFUJ_14336, and FFUJ_14337) were significantly repressed (Fold change > 2, P-value < 0.05) in the Δ*hda1* mutant strain under low nitrogen conditions. Simultaneously, I determined an increased H3K9 acetylation signal in two genes (FFUJ_14333 and FFUJ_14334) and a reduced signal in five genes (FFUJ_14331, FFUJ_14332, FFUJ_14335, FFUJ_14336, and FFUJ_14337), whereas no significant signal could be determined only in FFUJ_14337 (signal probability < 0.95). No acetylation was measured under nitrogen sufficiency in both strains (Figure 2.6).

An unexpected expression pattern could be observed in genes of the bikaverin cluster. While the optimal expression conditions of the wild type constitutes nitrogen deficient media the Δ*hda1* mutant strain exhibits a significantly increased gene expression (Fold change > 2, P-value < 0.05) on nitrogen sufficient medium in four cluster genes (FFUJ_06742 to FFUJ_06745) compared to the wild type (Figure 2.7A). Although no significant differences in gene expression could be determined in nitrogen starving conditions, an increased signal of H3K9 acetylation was measured in all cluster genes in the Δ*hda1* mutant (1.67 to 16.91 fold change in NLCS). In the wild type no significant acetylation signal (signal probability < 0.95) could be determined in two genes under this condition (FFUJ_06744 and FFUJ_06747). A slightly increased signal (1.18 to 1.51 fold change in NLCS) was observed in five cluster genes (FFUJ_06742 to FFUJ_06746) while no significant signal was measured in FFUJ_06747 in both strains (Figure 2.7B).

2.2.9 Similarity analysis of the *Fusarium* proteome

Filamentous fungi exhibit a broad host range. Beside the plant pathogenic *Fusarium* species fungi of the phylum *Aspergillus* preferably infect humans and animals [208, 219]. To determine proteins that are characteristic for plant pathogens I utilized the SIMAP protein database [7, 175] and identified in total 5240 proteins that are encoded in all 22 available *Fusarium* genomes. In order to get a comparison value I did the same for eight *Aspergillus* species which resulted in 4149 'core' proteins. The intersection of both sets comprises 2177 proteins, which occur in all 30 species. I used the FunCat [184] protein annotation database to find overrepresented functional categories in the identified protein sets. Beside proteins involved in primary metabolism I found transport mechanisms like mitochondrial, vesicular and vacuolar or lysosomal transport (FunCat-IDs: 20.09.07 and 20.09.13) to be significantly enriched (P-value < 0.01) among the set of shared proteins of *Fusarium* and *Aspergillus.* Furthermore cell cycle and DNA processing (FunCat-ID: 10) relevant proteins as well as functions connected to protein synthesis (FunCat-ID: 12) and protein binding (FunCat-ID: 16.01) were determined. I calculated the set differences to determine the proteins that are characteristic for

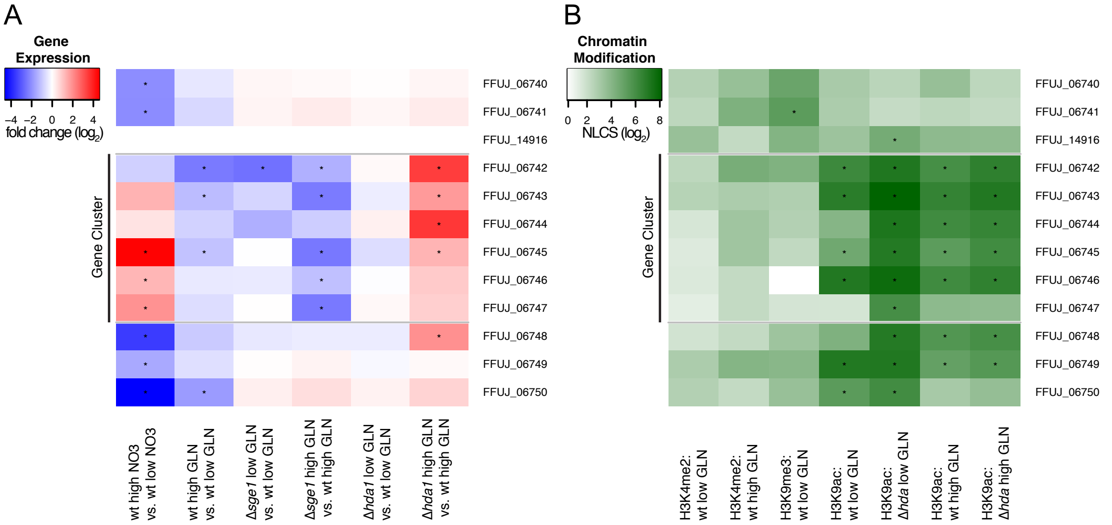

Figure 2.7: Gene expression and epigenetic modification of the bikaverin gene cluster and adjacent genes of *F. fujikuroi* wild type (wt), Δ*hda1* and Δ*sge1* mutant under limiting (low) and sufficient (high) glutamine (GLN) and nitrate (NO3) conditions. Genes are listed in chromosomal order on y-axis, gene cluster is indicated by vertical black bar. **A:** Fold changes (log_2 scale) of gene expression between two experimental conditions. Up-regulated genes compared to the control condition are depicted in red, down-regulated genes in blue. Asterisk indicate significant changes in gene expression (fold change > 2, P-value < 0.05). **B:** NLCS values (log_2 scale) indicating degree of chromatin modification for each experimental condition. Significant signals are indicated by asterisk (signal probability > 0.95).

each phylum. 3063 proteins could be found exclusively in *Fusarium* and 1972 are specific for *Aspergillus*. Interestingly, 43% (1303) of the *Fusarium* core protein set have no functional classification whereas in *Aspergillus* only 32% (622) proteins of the unique set were unclassified.

Beside proteins with unknown function I determined a significant functional overrepresentation (P-value < 0.01) in primary metabolism (FunCat-ID: 01) as well as cell rescue, defense and virulence (FunCat-ID: 32) in the *Fusarium* specific set. In addition I predicted 11% of proteins to be secreted. The unique set in *Aspergillus* exhibited a significant enrichment (P-value < 0.01) for proteins involved in protein folding, modification or destination and proteins connected to stress response (FunCat-IDs: 14 and 32.01) (Figure 2.8).

I am furthermore interested in species specific proteins in the respective *Fusar-*

ium species. To determine the unique genome specific set proteins without ortholog in terms of a bidirectional best hit in any other *Fusarium* genome were selected. *F. solani*, which is the genome with the largest phylogenetic distance to the other *Fusaria*, has the biggest share of unique proteins (3398 proteins, 21.63%). It is remarkable that *F. graminearum* has the second highest proportion of species specific proteins (1570 proteins, 11.36%), while the proportion in the closely related *F. asiaticum* (324 proteins, 2.60%) and *F. pseudograminearum* (650 proteins, 5.27%) is considerably smaller. *F. oxysporum* lycop. has a proportion of 10.80% (1885) unique proteins while the proportion in the *F. oxysporum* strain MN25 comprises 430 proteins (2.66%). Big differences in the amount of species specific proteins can also be observed in the species of the GFC. The genome of *F. fujikuroi* encodes 360 (2.43%) species specific proteins, its close relative *F. verticillioides* of the African clade and *F. circinatum* of the American clade have nearly four times more unique proteins among all *Fusaria* (1412 proteins, 9.96% and 1416 proteins, 9.43%). The highest amount of unique proteins in GFC species among all available genomes can be found in *F. verticillioides* (198 proteins, 1.40%) and *F. circinatum* (330 proteins, 2.20%), while in *F. fujikuroi* only 0.81% of the genome (120 proteins) were detected as unique.

F. solani has the biggest share of unique proteins (1851 proteins, 11.87%) when comparing to all available genomes (Table A.2). Interestingly, only 81 proteins (0.59%) in *F. graminearum* lack a bidirectional best hit to genomes outside the *Fusarium* phylum while the unique proportion in *F. graminearum* compared to other *Fusarium* species was much higher (11.36%).

2.3 Discussion

In this chapter the genome sequence of *Fusarium fujikuroi* was analyzed and associations between different levels of regulation by integrating genome wide multiple 'omics' data were shown. Measurements during secondary metabolism inducing conditions helped to get insight into the regulation of secondary metabolism gene clusters.

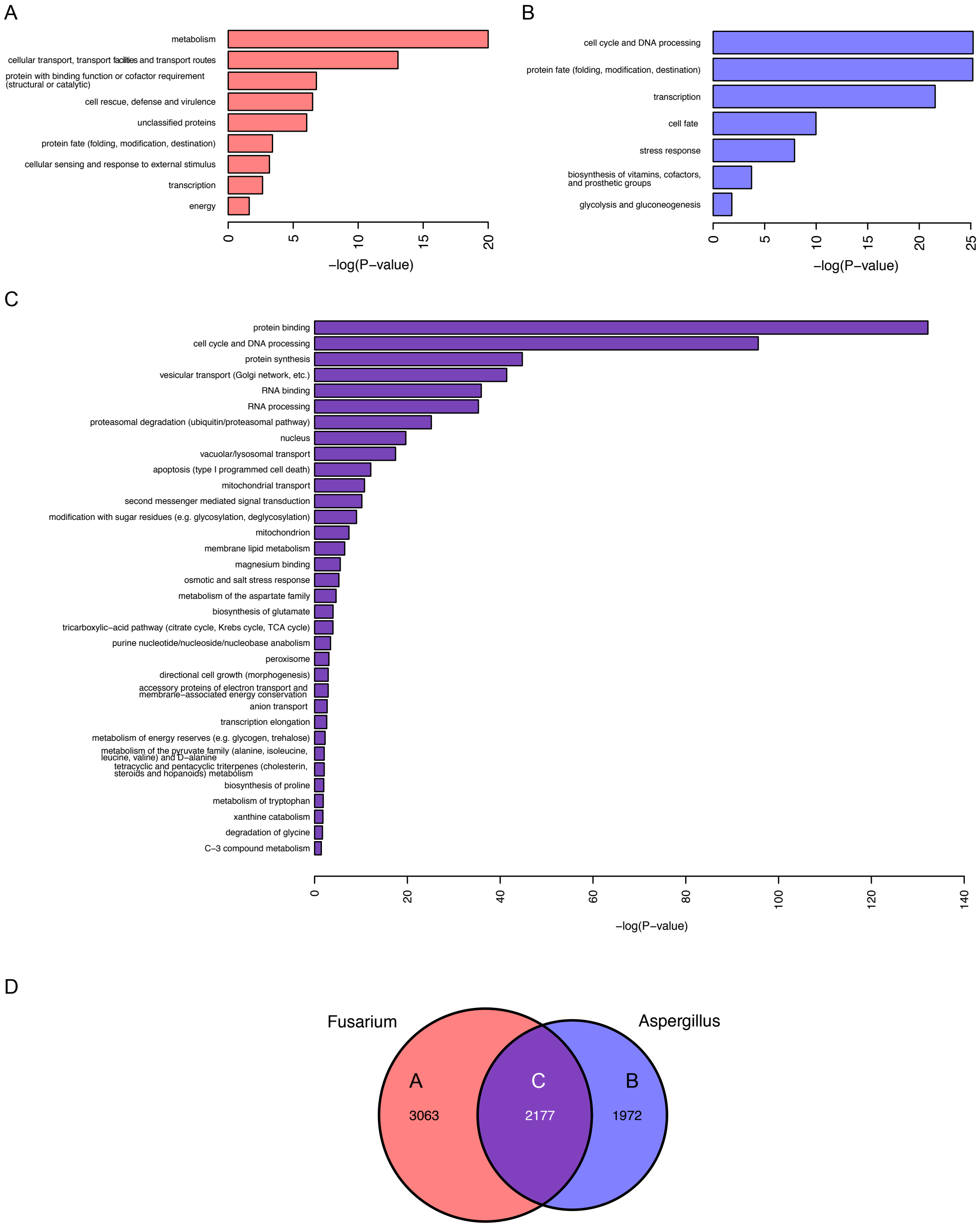

Figure 2.8: Core proteomes of *Fusarium* and *Aspergillus*. Overrepresented functional categories are depicted for the sets of *Fusarium* (A) and *Aspergillus* (B) specific proteins as well as for the intersecting set of both phyla (C). The amount of shared proteins is visualized as Venn diagram (D).

2.3.1 Genome analysis of *F. fujikuroi* reveals a dispensable chromosome

The assembly of the *F. fujikuroi* genome resulted in twelve distinct chromosomes. A comparison to the closely related *F. verticillioides* genome by a whole genome alignment revealed in most instances syntenic regions but also two main differences including a shorter chromosome IV and a unique chromosome XII in *F. fujikuroi.* Subsequent experiments have been done to compare the property of chromosomes of the sequenced strain of *F. fujikuroi* to other strains and species [231]. Nine additional *F. fujikuroi* isolates from different geographic regions were analyzed by PCR to determine strain specificity of the short chromosome IV and the presence of chromosome XII. The results indicate that the shortened chromosome IV is species-specific as parts of the peripheral region of the chromosome are missing in all nine isolates. In contrast the presence of chromosome XII could not be determined in two strains from Japan (m570, C1995) whereas in three italian strains (E289, E292, E325) only some parts of the chromosome are observed [231]. The findings suggest that chromosome XII in *F. fujikuroi* is not essential for pathogenicity as strains without chromosome XII were also able to cause bakanae disease on rice [231]. This is in contrast to the supernumerary chromosomes of *F. oxysporum* [132]. Missing genes on the shorter chromosome IV may be relevant to the life style of the organisms. The unique gene set on the *F. verticillioides* chromosome IV revealed an overrepresentation of secondary metabolism genes and genes related to detoxification which may play a role in host-pathogen interaction.

2.3.2 Sge1 globally regulates secondary metabolism in *F. fujikuroi*

Comparative analysis of gene families revealed some major differences between *Fusarium* species. The high amount of transcription factors (TFs) in *F. fujikuroi* can be linked to an expansion of the Zn(2)C6 zinc finger domain TF family which is often associated with specific secondary metabolite pathway regulation like in the case of the aflatoxin or sterigmatocystin gene cluster in *Aspergillus* [42, 63, 238] or in the compacting biosynthesis in *Penicillium citrinum* [1].

Global regulators like Wor1 are known as morphology and lifestyle switches [152] as well as regulators of pathogenicity and reproduction [101]. In analyzing genome wide transcriptomics data it was shown that the Wor1 ortholog Sge1 in *F. fujikuroi* positively regulates the secondary metabolism gene clusters of gibberellic acid, apicidin F, fusaric acid, bikaverin and two additional cryptic clusters of yet unknown metabolites. Interestingly the fumonisin SM pathway genes were activated in the mutant, suggesting a negative regulation by Sge1. However an increased expression of FUM genes and an elevated production of fumonisin could not be determined by subsequent northern blot and high-performance liquid chromatography (HPLC) experiments [144]. An explanation for this contradiction might be a delayed or premature fumonisin production in combination with different time points of RNA extraction for gene chip and northern blot experiments. Measurements at more time points will be necessary to elucidate the regulatory effect of Sge1 on the fumonisin pathway. Moreover, the mode of action of Sge1 in general is still unknown. In *C. albicans* a recognition sequence for direct binding to DNA was described [127]. However, an overrepresented motif among putative target genes could not be determined in *F. graminearum* [101] and *F. fujikuroi*, as well.

For the assessment of pathogenicity in terms of secreted proteins, I predicted classically secreted, signal-peptide containing candidates, but took also non classically secreted proteins into account. The largest secretome was determined in the tomato pathogen *F. oxysporum* f. sp. lycopersici where direct interactions between secreted effectors and the host plant have been shown in previous studies [123, 179]. The predicted secretome size in *F. graminearum* exceeds the previously suggested set by Brown *et al.* [33] which was focused on the classically secreted protein set only. The different predictions emphasize the difficulty to uncover the fast evolving class of virulence related secreted- and effector-proteins [50, 56, 234]. Still, induced gene expression during plant infection in *F. graminearum* or secondary metabolism inducing conditions in *F. fujikuroi* suggests a role in virulence of the predicted candidate genes.

2.3.3 Secondary metabolism is affected by nitrogen

I investigated regulatory mechanisms of secondary metabolism gene clusters by integrating data of transcriptomics, proteomics and epigenetics experiments under secondary metabolism inducing conditions. Secondary metabolism genes like those of the gibberellic acid gene cluster were known to be affected by nitrogen dependent regulation [145]. Interestingly, these genes not only exhibited significantly increased gene expression under inducing conditions but also increased protein abundance and an enrichment of activating H3K9ac histone marks. Regulation on the epigenetic level is also evident in terms of H3K9 acetylation in SM clusters of apicidin F, bikaverin and when focusing on all predicted secondary metabolism genes in general. On the subset of all SM genes I calculated a moderate correlation between gene expression and degree of histone modification. In contrast to that I was not able to determine a correlation on genome wide level. In *A. nidulans* H3K9me3 methylation marks are involved in regulation of secondary metabolism [180, 201]. Compared to these findings I was not able to determine a significant enrichment of H3K9me3 marks on any of the predicted clusters. Furthermore no correlation between secondary metabolism gene clusters and H3K4me2 methylation marks was found, except in two genes of the gibberellic acid gene cluster. Although considering only three of a pleiotropy of possible histone modifications it can be said that among the tested antibodies especially H3K9 histone acetylation is associated with gene expression in *F. fujikuroi.* In addition, like in *A. parasitikus* and *A. nidulans* a correlation between histone acetylation and gene expression can be observed [159, 183, 197]. The association of distinct patterns in chromatin landscape and secondary metabolism gene clusters will also be useful to find novel SM clusters of yet unknown metabolites.

Correlation between gene expression and proteomics are often low due to post transcriptional regulatory mechanisms and different decay rates of proteins and mRNA [150, 225]. Previously, in *A. fumigatus* correlations on genome wide measurements could be determined [12]. Based on the measured set of proteins I determined a moderate correlation under SM inducing conditions between protein and mRNA abundances in *F. fujikuroi.* Again, the correlation was considerably higher when looking at secondary metabolism genes alone underlining nitrogen as external regulatory stimulus for secondary metabolism in general.

Sensing and up-take mechanisms of environmental nitrogen sources such as amino acids play an important role in the nitrogen regulatory network [67, 94]. In combination of sequence similarity and phylogenetic approaches I predicted putative orthologs of the *Saccaromyces cerevisiae* general amino acid permease (GAP) in *F. fujikuroi*. Subsequent experimental investigations by Pfannmüller *et al.* (submitted) showed that one of the predicted candidates is able to import at least five amino acids suggesting FFUJ_09118 as ortholog of the *S. cerevisiae* Gap1 and key player in the nitrogen regulatory network in *F. fujikuroi*.

2.3.4 Deletion of histone deacetylase *HDA1* reveals unexpected regulatory dependencies of gene clusters

Alterations in the chromatin landscape can be performed by certain proteins that either add or remove modifications of histone proteins. Histone deacetylases (HDAs) remove acetyl groups and lead to a conformation change of the affected histone which in turn leads to change DNA accessibility. Deletion of the histone deacetylase *HDA1* in *F. fujikuroi* lead not to an expected increase but to a decrease of H3K9 histone acetylation of the gibberellic acid (GA) gene cluster and simultaneously to a down-regulation under favorable conditions. This observation suggests that more factors are involved in the regulation of the biosynthesis genes. An interplay with the global regulator AreA, which was shown to be involved in the regulation of GA genes [145], and histone acyl transferases (HATs) is plausible. The lack of histone hyperacetylation of GA genes suggests that transcription factors like AreA or co-activators are main targets of Hda1. On the chromatin landscape of the bikaverin (BIK) cluster an increase in H3K9 acetylation marks on biosynthesis inducing and repressing conditions could be observed. Despite H3K9 acetylation no increase in gene expression was determined under normally inducing low nitrogen conditions. However, an unexpected increase in the high nitrogen condition was observed which is consistent with the enrichment of activating H3K9ac marks. The measurements under the two nitrogen concentrations hint that DNA accessibility by H3K9ac histone marks alone is not sufficient for activation of the pathway genes. Additional histone modifications and/or recruitment of external transcription factors may play a role in regulation of the bikaverin gene cluster. Expression of genes under normally repressing conditions may be explained by the influence of Hda1 on signal transduction pathways of external stimuli which was determined by a gene set enrichment analysis on all differentially acetylated genes of wild type and mutant Δ*hda1* strain. Based on the genome wide approach including microarray and ChIP-seq experiments I conclude that Hda1 is involved in both, activation and silencing of genes.

2.3.5 Estimation of core proteomes and species specific proteins in *Fusarium* and *Aspergillus*

In comparing the proteomes of the plant pathogenic *Fusarium* species to the fungi of the genus *Aspergillus*, which are a thread to human and animals, I identified genus characteristic genes that are candidates to play a role in the respective pathogenicity. Bioinformatic analyses of the unique gene set in *Fusarium* showed a significant enrichment of genes involved in cellular sensing and response to external stimulus which seems to be important for plant pathogens to regulate metabolism and expression of genes involved in defense and virulence during plant infection. Furthermore the overrepresented amount of putative genes coding for transporter and secreted proteins suggest a specialization of *Fusarium* species on plant interaction in terms of secreted effector proteins. The high amount of unclassified proteins hints towards many fast evolving proteins driven by the host pathogen arms race. Interestingly, in *Aspergillus* only a small proportion of unclassified proteins was found. The functional overrepresentation of stress response related genes was the sole category that could be linked to pathogenicity in the unique set of *Aspergillus*. As expected, the union of *Fusarium* and *Aspergillus* proteins basically revealed enrichments of essential cellular functions.

I furthermore compared proteins of *Fusarium* species among each other to determine host specific genes. The low number of unique proteins in *F. fujikuroi* could be due to the host ranges of the species. Until now *F. fujikuroi* is known to infect rice exclusively while *F. graminearum* grows on more host plants as wheat, barley or maize. The correlation holds also for *F. solani* which exhibits the highest amount of unique proteins and has a wide host range [13]. Furthermore, acquisition of horizontally transferred genes in *F. solani* was shown and contributes also to the number of unique proteins compared to other *Fusaria* [126]. Same holds for *F. oxysporum* f. sp. lycopersici which contains pathogenicity related chromosomes that were putatively acquired via HGT [132].

However, the number of putative unique proteins is connected with the quality and correctness of the respective gene annotation as incorrectly predicted genes are likely to have no orthologs in other species. Especially virulence genes like small secreted effector genes are hard to predict by *de novo* algorithms. On the other hand pseudogenes and false positive gene calls will appear as species specific as no orthologs of these genes will be found. The accurate gene prediction in *F. fujikuroi* may be one reason for the relatively low amount of unique proteins in this species. Reliable gene models are crucial for most down stream analyses including prediction of protein functions, designing of probe sets for microarray gene expression measurements and determining orthologous genes. However, for the increasing amount of fungal genome sequences an extensive manual inspection is not applicable for projects like the 1K fungal genomes project. Homology based prediction algorithms profit from well annotated genes when predicting genomes of closely related species. Prediction algorithms such as mGene [15, 186], Augustus [199] or SnowyOwl [178] that take experimental evidence into consideration will provide an alternative to time consuming manual verification. Gene expression data in terms of RNA-seq experiments can be used to determine location of genes, define splice sites and alternative splicing. Recently published RNA-seq data of *F. graminearum* and *F. verticillioides* will help to improve the gene models of these species [192]. However, as only a small part of the genes are expressed at the same time and many genes especially those related to virulence and secondary metabolism are only expressed upon a certain environmental stimulus such as contact with the host plant, many experiments are required to cover also the mRNA of conditionally expressed genes. The advantage of *in planta* experiments is the possibility to identify virulence related genes and to uncover cryptic effectors by capturing their mRNA.

Chapter 3

Interspersed repeats in fungal genomes

Comparative analysis of *Fusarium* genomes in Chapter 2 showed differences even among closely related species. In this chapter I estimate the impact of transposable elements on genome evolution, identify hints of active transposition and defense mechanisms for the inactivation of transposons.

3.1 Methods

3.1.1 Determination of repetitive and noncoding sequences

Transposable elements (TEs) influence genome stability and plasticity. To identify transposons and interspersed repeats I built a pipeline which includes de-novo prediction algorithms and a mapping step of previously published TEs. Therefore the pipeline can roughly be separated into two major parts. First a sequence library containing different classes of noncoding elements like transposable elements, pseudogenes or interspersed repeats is assembled. In order to consider novel families of interspersed repeats I applied RepeatScout [169]. The algorithm finds consensus seeds with high similarity on the genome sequence and extends them until a certain threshold is reached. In doing so the boundaries of repeat elements can be identified clearly. Repeat families with less than 10 repeats and a consensus sequence length shorter than 50 bp are removed. Additionally I filter low complexity and simple sequence repeats which are determined by NSEG [240] and Tandem-Repeat-Finder [18] during the RepeatScout procedure. I used RepBase database [102] to determine known families of transposable elements, pseudogenes and integrated viruses. In the second step the RepBase library and the calculated library of interspersed repeat families are used as input for RepeatMasker [193] in order to determine the exact locations of the repetitive elements on the genome.

3.1.2 Analysis of repeat induced point mutation

All determined interspersed repeats were used as input for the calculation of evidence for repeat induced point mutations (RIP). I used RIPCAL [85] to calculate dinucleotide frequencies of all interspersed repeats and of non-repetitive control sequences. Afterwards I applied the alignment based approach to scan for RIP-like di-nucleotide changes and to calculate RIP dominance in each repeat family. RIP dominance is the ratio of the amount of one CpN $\leftrightarrow$ TpN mutation to the amount of all three other possible CpN $\leftrightarrow$ TpN mutations. I considered repeat families with a RIP dominance above two and a total number of the dominant RIP-mutation above 50 as putatively RIP affected.

3.2 Results

3.2.1 Repeat prediction reveals high abundance of a GFC specific interspersed repeat family

Transposons have impact on stability and plasticity of whole genomes but also affect evolution of single genes in terms of duplication or deletion upon insertion [227]. I identified interspersed repeats in five representatives of the *Gibberella fujikuroi* species complex (GFC), five strains of the *Fusarium oxysporum* species complex (FOC) and three genomes of the *Fusarium graminearum* species complex (FGC). I determined repeat families with a minimal length of the consensus sequence of 50 nt and a minimal occurrence of 10 copies in the genome. The highest diversity of interspersed repeat families could be determined in the strains of *F. oxysporum*. In *F. oxysporum lycopersici* both the highest amount of single interspersed repeat elements (15325) and the highest number of repeat families (463) could be determined. More than 300 families and 10000 elements could be found in the *F. oxysporum* strains PHW815, PHW808, HDV247, Cotton and melonis 26406. In total 353 of the repeat families in *F. oxysporum lycopersici* are longer than 400 nt and therefore classified as long interspersed repeat whereas 110 families are shorter than 400 nt. This is in contrast to *F. oxysporum* PHW808 where the amount of short interspersed repeat families is considerably lower (307) compared to the amount of long interspersed repeat families (88). Interestingly, *F. oxysporum lycopersici* contains six repeat families that could not be found in any other of the analyzed species. The lowest family count among the analyzed species was observed in *F. graminearum* with 37 families consisting of 35 short and two long interspersed repeat families. The closely related *F. pseudograminearum* has slightly more families (45) and is the only organism without long interspersed repeat families. Among the GFC species *F. circinatum* exhibits the highest diversity of interspersed repeats (86 families, 4966 elements). *F. fujikuroi* has with 66 elements per family the highest average family size. *F. proliferatum* contains 14 repeat families that could not be found in other species of the GFC but in some *F. oxysporum* genomes (Table 3.1).

I found two families of short interspersed repeats that are highly abundant in the GFC and FGC, respectively. The repeat family FF_R.0 comprises 536 elements which is around 15% of all interspersed repeats in this species. The consensus sequence of the element with the length of 140 nt occurs in similar high amount in *F. proliferatum* (561 elements), *F. mangiferae* (533), *F. verticillioides* (464) and *F. proliferatum* (453). The element can also be found in the genomes of the FOC in frequencies between 125 and 139 copies. Remarkably the element could not be found in any other species that is more distantly related than the FOC. A repeat element with a similar length of 143 nt but different sequence composition compared to the repeat family FF_R.0 could be found in the FGC. The repeat FG_R.0 occurs 818 times on the genome of *F. pseudograminearum* and constitutes around 30% of all interspersed repeats. In *F. graminearum* and *F. asiaticum* the repeat can also be found in high abundance with 764 (37%) and 741 copies (27%), respectively. Except one copy in *F. oxysporum* II5 the repeat element can be found exclusively inside the FGC. Both elements contain stop codons on every frame and can be found nearly exclusively in intergenic regions.

Table 3.1: Interspersed repeat families of *Fusarium* species, *Neurospora crassa* and *Botrytis fuckeliana*. Table is ordered by amount of determined repeat families (fams) per species. Families are classified into large- ($> 400nt$) and small interspersed repeats ($< 400nt$). In a cross species mapping of families the amount of unique families per species was determined (unique fams). The amount of families that could only be found in the same clade (GFC, FGC, FOC) is also indicated (clade specific fams). No values are given *F. solani*, *N. crassa* and *B. fuckeliana* as they are the only representatives of their clade. Additionally the total amount of repeat elements and the average amount of elements per family is given (avg. fam size).

Species	families	elements	fams <400nt	fams >400nt	clade specific fams	unique fams	avg. fam size
F. oxysporum lycop 4287	463	15325	110	353	6	6	33.10
F. oxysporum PHW815	410	12425	286	124	1	1	30.30
F. oxysporum PHW808	395	12675	307	88	1	0	32.09
F. oxysporum HDV247	391	12317	248	143	1	0	31.50
F. oxysporum Cotton	316	10884	201	115	1	1	34.44
F. oxysporum melonis 26406	302	10313	207	95	0	0	34.15
F. oxysporum Fo47	215	7411	142	73	0	0	34.47
N. crassa	199	5282	144	55	-	154	26.54
F. oxysporum CL57	185	6486	138	47	0	0	35.06
F. oxysporum MN25	168	6144	111	57	0	0	36.57
F. solani	139	5152	93	46	-	29	37.06
F. oxysporum FOSC 3a	125	4976	92	33	0	0	39.81
F. oxysporum II5	97	4568	65	32	0	0	47.09
F. circinatum	86	4966	66	20	2	0	57.74
F. mangiferae	66	2766	58	8	2	1	41.90
F. asiaticum	62	2710	52	10	1	1	43.71
B. fuckeliana B05.01	60	2002	41	19	-	47	33.37
F. proliferatum	56	2428	17	39	14	0	43.36
F. fujikuroi	54	3587	39	15	0	0	66.43
F. verticillioides	45	2245	38	7	0	0	49.90
F. graminearum	37	2054	35	2	0	0	55.51
F. pseudograminearum	45	2678	45	0	0	0	59.51

3.2.2 Microarray data exhibits expression of repeats

In order to determine whether the determined interspersed repeat elements are still active transposable elements, repeat specific probes were included into the chip-design of the custom *F. fujikuroi* microarray. The probes were chosen as unique sequences among all gene models and interspersed repeats. In total it was possible to include probes of 1845 interspersed repeats on the microarray. I utilized measurements under high and low nitrogen growth conditions, that were used for the investigation of nitrogen regulation in section 2.2.6. Among the highest expressed repeats (95% percentile) seven significantly overrepresented repeat families were found. The three repeat families FF_R.12, FF_R.19 and FF_R.26, that were classified as LTR-Retrotransposon and contain open reading frames, were highly expressed in both nitrogen conditions. The retrotransposon family FF_R.29 as well as the DNA transposons FF_R.30 and FF_R.36 were also present under both nitrogen concentrations. In contrast to that, the DNA transposon FF_R.71 is exclusively expressed under low nitrogen conditions. Of the highly abundant repeat family FF_R.0 eight elements under high nitrogen conditions were detected. Six of them could also be found in the highly expressed set under low nitrogen conditions. In order to find differences in the expression of interspersed repeat elements I performed a differential expression analysis in comparing both nitrogen concentration conditions. In total I determined 27 and 23 significantly up-regulated repeat elements under high- and low nitrogen conditions, respectively (Fold change > 2, P-value < 0.05). Beside five elements of the mentioned DNA transposon family FF_R.71 I found two copies of the LTR-retrotransposon family FF_R.29 and another two copies of the DNA-transposon family FF_R.75 to be exclusively up-regulated under low nitrogen concentrations. I also found families that were expressed under high nitrogen conditions exclusively. Three elements of the LTR-retrotransposon family FF_R.65, as well as two SINE elements (FF_R.34) and two members of the unclassified repeat family FF_R.39. Members of families of the remaining differentially expressed repeats were determined under high and low nitrogen conditions as well. For example six members of the biggest interspersed repeat family FF_R.0 could be found in the set of significantly expressed repeats under high nitrogen, whereas two members were expressed under low nitrogen conditions. The significant differences in repeat expression intensities hints towards an active transposition of these repeat elements and may explain the high repeat frequency of the genome.

3.2.3 *F. fujikuroi* shows evidence of repeat induced point mutation

Some fungal species developed a defense mechanism like repeat induced point mutation (RIP) against proliferating transposable sequences. The mechanism induces preferably C:G to T:A point mutations during the sexual cycle in transposons and long repetitive elements. To identify traces of this mechanism in the analyzed *Fusarium* genomes I calculated di-nucleotide frequencies of the predicted interspersed repeat families and compared them to frequencies of non repetitive control sequences. In *F. fujikuroi* and *F. circinatum* repeats with two fold increased di-nucletoides frequencies compared to the control sequence were found (figure 3.1). Interestingly both species also showed the highest overall frequency difference between repeats and control sequences. Especially a reduced amount of the di-nucleotides CpA, CpG and TpG could be measured. As TpA is the only increased (fold change above two) di-nucleotide pair, CpA → TpA are suggested to be the dominant form of CpN → TpN di-nucleotide mutations. In order to identify single families with increased rip specific point mutations I applied an alignment based approach using RIPcal [85]. I determined RIP dominance of CpA ↔ TpA above two with more than 50 di-nucleotide mutations in six of 54 repeat families in *F. fujikuroi* (FF_R.6, FF_R.8, FF_R.12, FF_R.15, FF_R.16, FF_R.26) and in one of the 86 repeat families in *F. circinatum* (FC_R.23). Repeat dominance is the ratio of CpA ↔ TpA mutations to the three other possible CpN ↔ TpN di-nucleotide mutations. The repeat family of the LTR retro transposon (retrotransposon HobS hobase) FF_R.26 exhibits a RIP dominance of 2.2 and therefore has twice as many mutations from CpA ↔ TpA than the sum of the other RIP targeted mutations of CpC, CpG and CpT di-nucleotides. Figure 3.2A shows the alignment of the members of the family and the dominance of CpA ↔ TpA mutations (red curve) compared to the other mutations (figure 3.2B). Seven members of the transposon family FF_R.26 were determined under the high expressed set in section 3.2.2.

Interestingly, five of the seven high expressed elements exhibit a considerable low amount of characteristic RIP-mutations as illustrated by asterisk in figure 3.2A. I performed a gene set enrichment analysis on the neighboring genes of the predicted RIP affected transposons but were not able to determine any significantly overrepresented functional categories or an enrichment of secreted proteins.

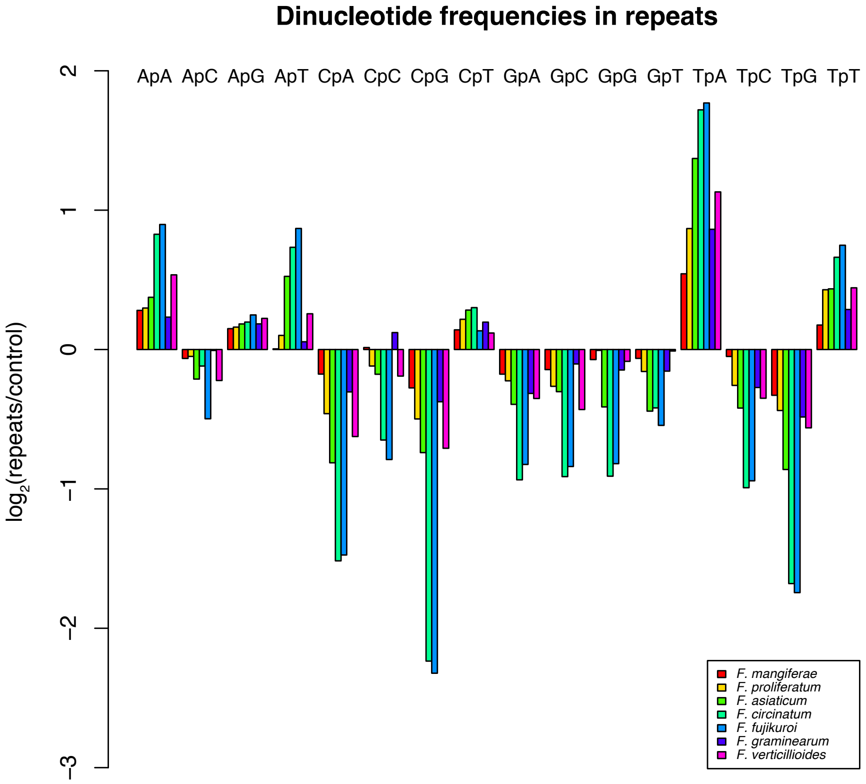

Figure 3.1: Dinucleotide frequency fold changes of all repeat families of seven *Fusarium* genomes compared to non-repetitive control sequences. Turquoise and light blue bar indicate low CpA, CpG and TpG abundance in *F. fujikuroi* and *F. circinatum* but high abundance of TpA di-nucleotides.

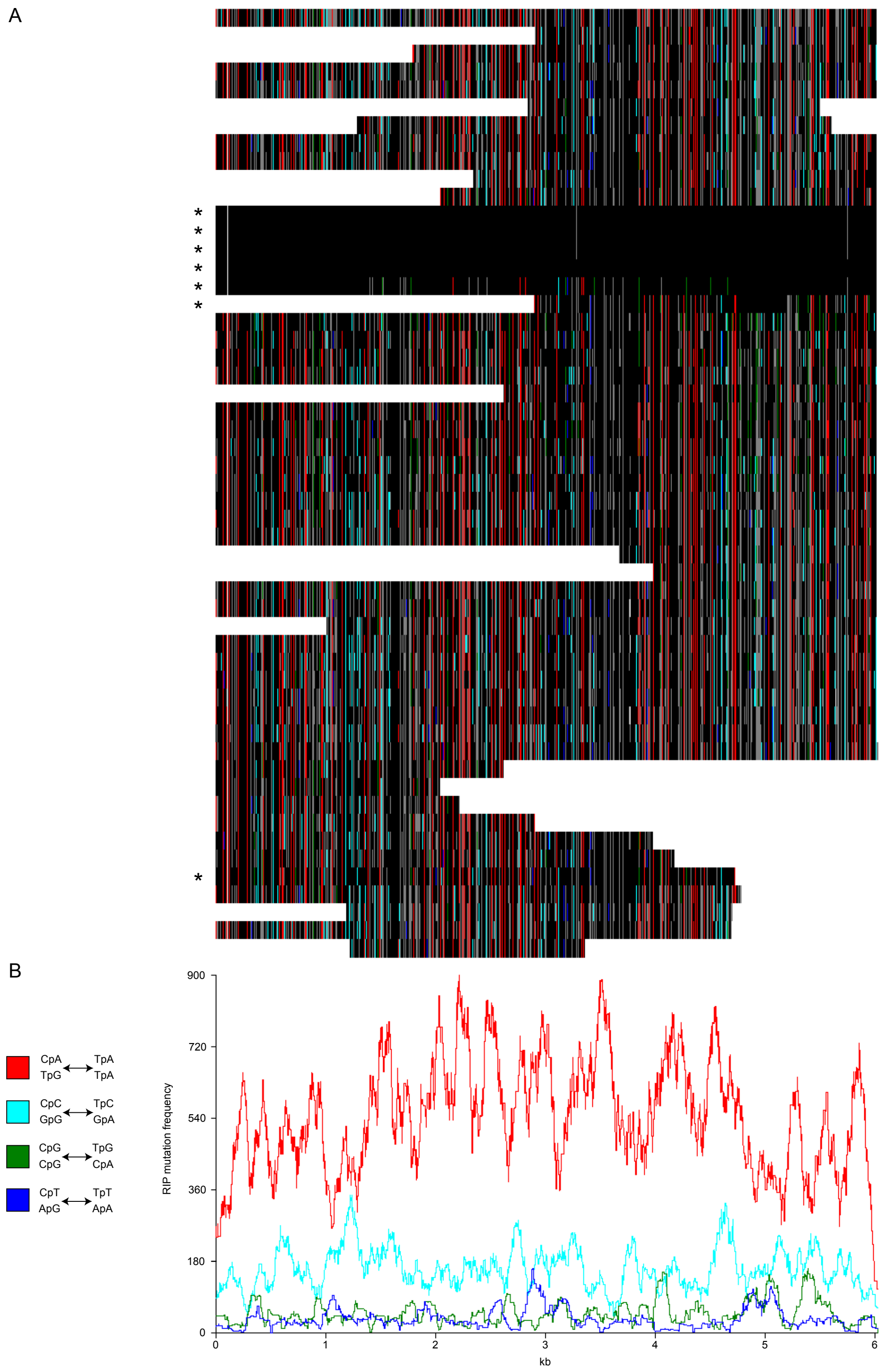

Figure 3.2: Evidence of repeat induced point mutation (RIP) in *F. fujikuroi.* **A:** Multiple alignment of members of the putative transposon repeat family FF_R.26. Matches are indicated by black, gaps by white colors. Mismatches corresponding to RIP specific di-nucleotide mutations are colored as indicated. Asterisk indicate repeat elements with high expression rate according to expression data analysis. **B:** Frequency of RIP-mutations of the alignment in A calculated by a 50 bp sliding window. Red curve illustrates the dominance of CpA ↔ TpA mutations over other CpN ↔ TpN mutations.

3.3 Discussion: Evolutionary impact of transposable elements and repeat induced point mutation

Transposable elements influence genome evolution by duplications and deletions and lead to an increased rate of polymorphisms in the vicinity. They also play a role in horizontal gene transfer and acquisition of genetic material [161, 182]. I found two highly abundant repeat families that occur exclusively inside the FGC and GFC+FOC, respectively. The consensus sequences of both repeat families have a length between 140 and 143 bp and contain stop codons on every frame. Accordingly, the elements are found especially in intergenic regions as gene insertions would lead to a disruption of coding sequences. I encountered the question whether these elements still transpose actively in analyzing expression data of interspersed repeats. Hereby it was possible to define a unique probe set for a subset of the predicted transposons. However the experiments at different secondary metabolism inducing conditions provided evidence of expression of the high abundant repeat family in *F. fujikuroi.* Beside these short repeats I also identified families of LTR-retrotransposons and DNA-transposons with considerable evidence of expression and thus transposition activity. Because the insertion into genes can be deleterious, some fungi developed mechanisms like repeat induced point mutation (RIP) to inactivate transposons and prevent their transposition [72, 187]. Among the analyzed *Fusarium* species, the analysis of RIP signatures in predicted transposons revealed the strongest evidence for RIP in *F. fujikuroi.* The di-nucleotide frequency of all predicted transposable elements as well as the dominance of characteristic CpA ↔ TpA mutations hint towards an active RIP mechanism. This is comparable to *Neurospora crassa* where CpA dinucleotides are also preferably affected by RIP [35]. Furthermore the transposon family with highest CpA ↔ TpA dominance shows high expression in elements with low mutations and low expression in elements with high amount of RIP-mutations. The correlation of transposon expression and amount of RIP mutations suggest a successful inactivation of some but not all transposons of this family. Recent studies postulate that RIP can act as evolutionary accelerator of neighboring genes of affected transposons [161]. However, a gene set enrichment analysis of flanking transposon genes revealed no significant enrichment of virulence related or secreted proteins. As expected it was not possible to determine traces of RIP in the short high abundant repetitive elements given that RIP operates only on transposons that are longer than 400 bp [226]. The measured expression of the high abundant repeat family in *F. fujikuroi* and the inability of RIP to inactivate it raises the hypothesis that the elements still contribute to the shape of genes and the genomes of *F. fujikuroi* and maybe other *Fusarium* species.

Chapter 4

Secondary metabolism gene clusters contribute to virulence

Genes that are part of a secondary metabolism pathway exhibit characteristic features like co-expression, co-regulation and a typical functional composition. In Chapter 2, I showed nitrogen dependent expression of gene clusters in *F. fujikuroi.* Based on preliminary work by Wanseon Lee [122], I will now make use of these features in order to predict putative gene clusters with yet unknown metabolic compound. Most results of this chapter are published and discussed in:

- **Sieber CMK***, Lee W*, Wong P, Münsterkötter M, Mewes HW, Schmeitzl C, Varga E, Berthiller F, Adam G, Güldener U.
  The *Fusarium graminearum* genome reveals more secondary metabolite gene clusters and hints of horizontal gene transfer.
  PLoS One. 2014, 9, e110311
- Niehaus EM, Janevska S, von Bargen KW, **Sieber CMK**, Harrer H, Humpf HU, Tudzynski B.
  Apicidin F: Characterization and genetic manipulation of a new secondary metabolite gene cluster in the rice pathogen *Fusarium fujikuroi.*
  PLoS One. 2014, 9, e103336

*equal contributions

- Wiemann P*, **Sieber CMK***, von Bargen KW*, Studt L, Niehaus EM, Espino JJ, Huß K, Michielse CB, Albermann S, Wagner D, Bergner SV, Connolly LR, Fischer A, Reuter G, Kleigrewe K, Bald T, Wingfield BD, Ophir R, Freeman S, Hippler M, Smith KM, Brown DW, Proctor RH, Münsterkötter M, Freitag M, Humpf HU, Güldener U, Tudzynski B. Deciphering the cryptic genome: genome-wide analyses of the rice pathogen *Fusarium fujikuroi* reveal complex regulation of secondary metabolism and novel metabolites. *PLoS Pathog.* 2013, 9, e1003475

*equal contributions

4.1 Methods

4.1.1 *De novo* prediction of secondary metabolism gene clusters

Secondary metabolism gene clusters usually contain a characteristic enzymatic composition. To predict gene functions I utilized InterProScan [244] to identify functional domains in protein sequences in order to classify proteins as signature enzyme, tailoring enzyme, transcription factor and transporter. After that I searched with a sliding window for local accumulations of these gene classes on the supercontigs, starting with three seed genes and allowing one unclassified gene in between. Statistical significance of the gene clusters to be enriched for functions associated with secondary metabolism was obtained by applying Fisher's Exact Test [65]. Resulting p-values were adjusted for multiple testing using Benjamini-Hochberg procedure [17]. In case of $P < 0.05$ clusters are seen as significantly enriched for secondary metabolism. The results were compared to the output of the secondary metabolism gene cluster prediction tools AntiSMASH [19] and SMURF [107] afterwards and adjusted manually.

4.1.2 Determination of co-expressed gene clusters

For the synthesis of secondary metabolites several enzymes are required. In most cases all these enzymes are encoded as genes of one gene cluster and are expressed at the same time under synthesis favoring conditions. To determine co-regulation of gene clusters I utilized available expression data as evidence in two ways. First of all experimental conditions (Table 2.2) where at least 60% of genes of a cluster are either significantly up- or down-regulated were identified. Differentially expressed genes were determined as described in Section 2.1.5.

Additionally, clusters with correlated gene expression during time-series experiments were identified. For this purpose I determined chromosomally clustered genes with correlated gene expression profile in general. Afterwards, I selected predicted SM gene clusters where at least 60% of genes are part of a co-expression cluster. Predicted SM gene clusters are extended by neighboring co-expressed genes. The mean Pearson correlation coefficient (R) was used as a measure of similarity of expression profiles. For each time-series experiment, a R cutoff (R_{min}) was determined in calculating the 95th percentile of 1000 Rs of randomly sampled sets of three genes. Using a sliding window, three neighboring genes are accepted as co-expression seed in case the mean R of their expression profile is above R_{min} and at least two genes show a significant change in their gene expression profile between two growth conditions (absolute fold change above two, P-value < 0.05). Seeds are extended in calculating Rs of upstream and downstream genes. Genes with $R > R_{min}$ are added successively to the seed, allowing one non-correlating gene in between.

4.1.3 Analysis of cluster specific cis-regulatory motifs

The expression of many secondary metabolism gene clusters is controlled by a cluster specific transcription factor that binds specifically to sequence motifs in promoters of the pathway genes. In order to determine evidence for cis-regulation of gene clusters I developed a pipeline that compares the frequency of predicted *de novo* motifs and published binding sites in the cluster to the distribution in the whole genome. In order to identify new conserved sequence motifs I utilized Meme [9, 8], Weeder [164] and Phylocon [222] on the set of cluster promoter sequences. Additionally, I scanned for known binding sites in aligning the matrices stored in the TRANSFAC-db [236]. The promoter of a gene was defined as the 5' intergenic sequence with a maximum of 1kb of upstream nucleotides. Also promoter sequences of orthologous genes were included into the search space for the de-novo algorithms. All computed de-novo motifs and the matrices of the

TRANSFAC-db were used as query for a genome wide promoter scan. I assessed the significance of determined sequence motifs in applying Fisher's exact test [65], which takes the occurrence of a motif on cluster promoters as well as its distribution on the genome into account. To correct for multiple testing, the resulting p-values have been adjusted using Bonferroni procedure [23, 24]. Sequence motifs with an adjusted p-value below 0.01 which occur on at least least 80% of cluster promoters are accepted as significant, cluster specific motifs.

4.1.4 Phylogenetic distribution of the apicidin cluster

To infer the phylogenetic relationship of species which contain orthologs of the apicidin cluster I calculated a phylogenetic tree based on the RPB2 (GenBank: KF255548.1) gene alone, as the genome sequence of *Fusarium semitectum* is not publicly available, yet. I used Mafft [103], PAL2NAL [204] and Gblocks [206] to prepare the alignment and PhyML [81] to calculate the tree as described in Section 2.1.3.

4.2 Results

4.2.1 Estimating the genetic potential for secondary metabolite production

The main synthesis step in the pathways of many secondary metabolites, including mycotoxins, phytohormones and antibiotics, is catalyzed by signature enzymes. In order to estimate and compare secondary metabolism pathways in *Fusarium* I used InterProScan [244] to identify functional domains of polyketide synthases (PKS), non-ribosomal peptide synthases (NPS), dimethylallyl tryptophan synthases (DMATS) and terpene synthases (TPS). In *F. fujikuroi* 17 putative type I PKSs which contain a canonical ketosynthase (KS domain, Interpro-ID: IPR020841), an acyl-carrier (ACP domain, Interpro-ID: IPR009081) and an acyltransferase domain (AT domain, Interpro-ID: IPR020801) were found. 14 of the 17 predicted PKSs additionally contain ketoreductase (KR domain, Interpro-ID: IPR013968), dehydratase (DH domain, Interpro-ID: IPR020807) and enolreductase (ER domain, Interpro-ID: IPR020843) domains. These reducing type PKSs (R-PKS) catalyze the complete reduction of β-carbonyl during polyketide synthesis [28, 39]. Four of the R-PKS were classified as NPS-PKS hybrid as they contain an additional NPS-module. The remaining three PKSs lack the KR, DH and ER domains and therefore also the reducing feature. Beside the PKSs of type I also one type III PKS enzyme could be identified. The PKS type III enzyme class was assumed to be characteristic for plants and bacteria, but has recently also be detected in filamentous fungi [104, 132, 243]. Furthermore, 23 NPS, 2 DMATS and 17 TPS were predicted in *F. fujikuroi*.

In *F. proliferatum* 19 PKS and 30 NPS were found, which is the highest determined amount of these enzymes among the compared species (Table 4.1). The highest number of 23 TPS was also found in *F. proliferatum* and in *F. asiaticum*. The minimum of PKS was determined in *F. circinatum* (8) but has in turn also the highest number of PKS-like enzymes (23) which are enzymes that lack one of the necessary functional domains. This could be due to the draft gene model in *F. circinatum* [237]. Like in *F. fujkikuroi*, four PKS-NPS hybrids were identified in *F. proliferatum* and *F. oxysporum* lycop. whereas only one enzyme was found in *F. graminearum*, *F. pseudograminearum*, *F. asiaticum* and *F. solani* (Table 4.1 and Table 4.2).

4.2.2 Prediction of secondary metabolism gene clusters

Secondary metabolism synthesis genes are usually tightly clustered on the chromosome. To identify gene clusters of yet unknown secondary metabolites of including putatively harmful mycotoxins, I scanned for local clustering of signature and tailoring enzymes and additional proteins like transcription factors and transporter proteins. Additionally I performed a functional enrichment analysis of secondary metabolism related functions to determine the significance of the found gene clusters. I also considered tentative secondary metabolism enzymes to account for remnants of functional clusters or new evolving pathways.

In *F. fujikuroi* 49 gene clusters were found that contain a significantly overrepresented (P-value < 0.05, see methods Section 4.1.1) amount of secondary metabolism genes presumably involved in secondary metabolite biosynthesis (Ta-

Table 4.1: Overview of secondary metabolism genes in *Fusarium* species. Number of predicted signature- and tailoring- enzymes and amount of predicted secondary metabolism gene clusters based on Interpro domains.

	F. fujikuroi	*F. verticillioides*	*F. proliferatum*	*F. mangiferae*	*F. circinatum*	*F. oxysporum* lycop.	*F. graminearum*	*F. asiaticum*	*F. pseudograminearum*	*F. solani*
PKS	17	11	19	17	8	12	15	16	12	12
PKS-NPS hybrids (of PKS)	4	3	4	3	3	4	1	1	1	1
NPS	23	22	29	23	20	19	23	25	20	16
DMATS	2	0	2	3	3	0	0	0	0	0
TPS	17	16	23	19	18	14	17	20	19	9
PKS-like	6	15	7	5	23	11	5	6	6	8
NPS-like	49	42	51	46	44	52	35	33	36	39
DMATS-like	0	1	1	1	1	2	0	0	0	1
Cytochrome P450	143	130	185	160	145	168	114	119	104	162
Acyltransferases	100	87	103	101	96	110	88	85	86	109
Glycosyltransferases	21	28	22	23	28	39	28	24	22	28
Methyltransferases	57	91	68	63	93	107	88	51	51	64
Oxidoreductases	174	158	175	172	164	197	129	135	120	200
Predicted SM clusters	49	33	56	47	28	30	67	24	27	27

ble A.3). Besides tailoring enzymes and additional proteins like transcription factors or transporter proteins, 42 of the predicted clusters contain at least one signature enzyme. The predictions comprise 10 clusters of known metabolite (Table 4.3) and include all 17 PKS genes as well as 20 of the 23 predicted NPS genes. Furthermore two DMATS, 12 TPCs and 34 P450 are included. The number of cluster in *F. fujikuroi* is similar to *F. maniferae* with 47 clusters but lower compared to *F. proliferatum*, the third species of the asian clade, which contains 56 predicted clusters. In *F. verticillioides* (33 clusters) and *F. circinatum* (28 clusters) the number of predicted gene clusters was considerably lower (Table 4.1).

Compared to the predicted number of gene clusters in *F. fujikuroi* the amount in *F. graminearum* in considerably higher. A total number of 67 statistically significant (F-test, P-value < 0.05) potential gene clusters were identified. At least one signature enzyme was found in 46 clusters which is about 58% of the predicted SM genes including all 15 PKS, 21 of 23 NPS, 14 of 17 TPS and 42% of the cytochrome P450 genes (48 out of 114) (Table 4.1). In particular, the genes with known functions and associated metabolite clusters reported for *F. graminearum* (Table 4.3) are all included in these clusters which may represent also extensions to previously reported functional gene clusters.

Table 4.2: Overview of secondary metabolism genes in *Fusarium oxysporum* strains. Number of predicted signature- and tailoring- enzymes and amount of predicted secondary metabolism gene clusters based on Interpro domains.

	F. oxysporum lycop.	*F. oxysporum* MN25	*F. oxysporum* CL57	*F. oxysporum* II5	*F. oxysporum* Fo47	*F. oxysporum* HDV247	*F. oxysporum* PHW808	*F. oxysporum* melonis	*F. oxysporum* FOSC 3a	*F. oxysporum* cotton	*F. oxysporum* PHW815
PKS	12	13	10	12	12	12	12	12	13	11	11
PKS-NPS hybrids (of PKS)	4	5	3	3	3	5	5	3	4	3	3
NPS	19	22	25	19	23	23	27	26	23	24	21
DMATS	0	5	3	3	4	3	3	4	5	3	2
TPS	14	18	20	20	19	22	23	18	24	16	21
PKS-like	11	9	10	9	8	8	9	13	9	6	7
NPS-like	52	48	46	45	49	54	62	58	52	61	51
DMATS-like	2	0	1	1	1	0	0	0	0	0	2
Cytochrome P450	168	162	172	151	165	193	188	172	163	175	186
Acyltransferases	110	106	107	100	108	112	105	109	104	109	107
Glycosyltransferases	39	30	30	25	32	33	31	30	26	28	29
Methyltransferases	107	70	69	69	70	74	69	70	70	71	68
Oxidoreductases	197	171	184	179	187	199	183	199	173	189	183
Predicted SM clusters	30	39	38	37	40	36	35	41	41	38	35

Table 4.3: Gene clusters of known metabolites of *F. fujikuroi* and *F. graminearum* and their orthologs in other *Fusarium* species. Full circles indicate that orthologs of all genes of the cluster are present, empty circles illustrate that the cluster is missing completely. Partial circles indicate incomplete orthologous clusters. Reference for the proportion of present genes are the published clusters in *F. fujikuroi* and *F. graminearum*.

Metabolite	Reference	*F. fujikuroi*	*F. graminearum*	*F. asiaticum*	*F. prolifeartum*	*F. mangiferae*	*F. cricinatum*	*F. verticillioides*	*F. oxysporum* f. sp. lycop.	*F. solani*	*F. pseudograminearum*
α-acorenol	[29]	●	◑	◑	◕	●	◕	◑	◑	○	◑
apicidin F	[154, 231]	●	○	●	○	○	○	○	○	○	○
beauvericin	[128]	●	○	○	●	◑	◑	◑	●	○	○
bikaverin	[232]	●	○	○	●	◕	◕	●	●	○	○
fumonisin	[173, 171]	●	○	○	◕	○	○	◕	○	○	○
fusaric acid	[31]	●	○	○	◕	●	◕	●	◕	○	○
fusarin C	[155]	●	●	●	●	○	●	●	○	◕	●
fusarubin	[203]	●	●	●	●	●	◔	●	●	◕	●
gibberellins	[215, 125, 22, 133]	●	○	○	●	●	●	◔	○	○	○
neurosporaxanthine	[181]	●	●	●	●	◕	●	◕	●	◕	●
aurofusarin	[59, 129, 110, 111, 134]	○	●	●	○	○	○	○	○	○	●
butenolide	[87, 224]	○	●	●	○	○	○	○	○	○	●
carotenoid	[100]	●	●	●	●	◕	●	●	●	●	●
culmorin	[142]	◕	●	◕	◕	◔	◔	◕	◕	◕	●
ferricrocin	[212]	●	●	●	●	◕	◔	◑	◕	◕	◕
fusarielin	[205]	○	●	◕	○	○	○	○	○	○	◑
malonichrome	[61]	○	●	●	●	○	○	●	●	●	●
triacetylfusarinin	[162]	●	●	●	●	●	●	●	●	●	●
trichothecene	[32]	○	●	●	○	○	○	○	○	○	●
zearalenone	[71, 112]	○	●	●	○	○	○	○	○	○	●
precursor of insoluble perithecial pigment	[172]	◕	●	○	◕	◔	◔	◑	◕	○	◕

4.2.3 Putative gene clusters are supported by expression data

Most genes of secondary metabolism gene clusters are co-regulated as all encoded enzymes have to be present to establish all steps of the pathway that are required for the specific compound synthesis. A co-expression of genes in several gene clusters in *F. fujikuroi* could be observed. For example gene expression of all genes in the gibberellic acid (Figure 2.6) and bikaverin (Figure 2.7) gene cluster changes significantly due to the availability of nitrogen in the media [231] as described in Section 2.2.6.

I utilize available *F. graminearum* expression data (Section 2.1.4) as evidence for predicted cluster genes to be part in one secondary metabolism pathway and to examine additional genes that are potentially part of the functional gene clusters. In addition the environmental conditions are revealed that are necessary for the metabolite biosynthesis as many clusters remain silent under laboratory conditions.

In *F. graminearum* 26 clusters were found where at least 60% of genes show a significant correlation in expression profile during plant infection (see Section 4.1.2) and 42 clusters with more than 60% of genes are differentially expressed in at least one condition. These clusters comprise the known gene clusters of the metabolites trichothecene [32, 53], butenolide [87], fusarin C [70, 135, 176, 155] and aurofusarin [134]. Beside these known, experimentally validated pathways I found correlations in gene expression in the neighboring genes of the biosynthetic enzymes of triacetylfusarinin and malonichrome. Five genes in a cluster with enzymes involved in triacetylfusarinin biosynthesis show differential expression and correlation in their expression profile during infection. Interestingly all genes (FGSG_03747 to FGSG_16212) are significantly down-regulated (Fold change < -2, P-value < 0.05) during C- and N- starving conditions (FG2) but up-regulated during conidiation (FG7) and infection of wheat (FG19) (Figure 4.1).

Four neighboring genes of the malonichrome associated NPS (FGSG_11026) are significantly up-regulated (Fold change > 2, P-value < 0.05) during the infection process of wheat and exhibit correlated expression profiles during barley and wheat infection. Interestingly the genes are down-regulated when forming perithecia (FG16) and during trichodiene treatment (FG10) (Figure 4.1). The promoter analysis resulted in a significantly enriched motif CAGGGATCGGCC (P-value = 9.17e-6), which is present in the promoters of the genes FGSG_11029 to FGSG_11026 but not in the transcripton factor FGSG_11025. As all the pathway genes of both metabolites are not elucidated yet, the results give a hint on the extent of the gene cluster.

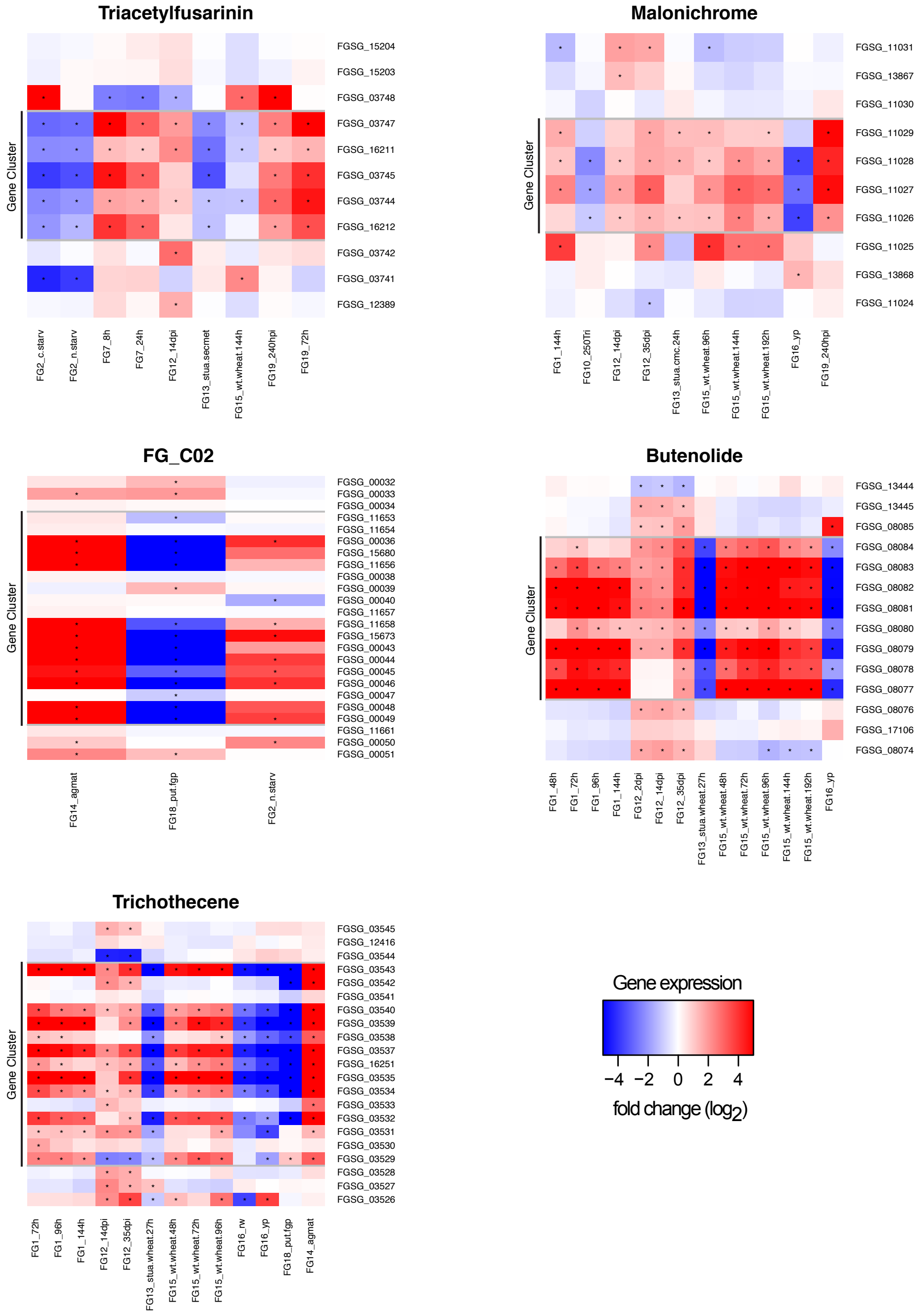

Figure 4.1: Differential gene expression heatmap of clusters and neighboring genes. Heatmaps illustrate fold changes in gene expression (log_2 scale) between experimental conditions. Genes are listed in chromosomal order on y-axis. Abbreviations of experimental conditions on x-axis are according to Table 2.2. Vertical black bars indicate predicted gene clusters. Figure is adopted and modified from Sieber *et al.* [191].

4.2.4 The NPS31 cluster in *F. fujikuroi* shows similarity to the apicidin cluster in *F. semitectum*

A previously unknown cluster in *F. fujikuroi* was predicted that shows similarity to 11 of 12 genes of a previously published cluster in *F. semitectum* (Table 4.4). In *F. semitectum* the pathway genes are responsible for the synthesis of apicidin, a histone deacetylase inhibitor with anti parasitic activity [99]. A comparison of the gene order revealed conserved collinearity between the orthologous clusters with the exceptions of one missing protein in *F. fujikuroi* and an opposite orientation of the genes FFUJ_00012 and FFUJ_00013. Among the available sequenced genomes three further orthologous clusters in *F. asiaticum*, *Setosphaeria turcica* and *Pyrenophora tritici-repentis* could be identified (Figure 4.2). Like in *F. fujikuroi* the cluster is located in the distal part of a supercontig suggesting also a subtelomeric location in these species.

Microarray experiments reveal a co-expression of all 11 cluster genes. A significant increase in expression rate could be observed under high nitrogen conditions, whereas gene expression is significantly repressed in the mutant of the global regulator Sge1. Interestingly no significant difference in gene expression could be determined in the upstream genes of the cluster. In accordance to the differences in gene expression also a different histone acetylation pattern could be monitored between the two nitrogen concentration conditions. An enrichment of the activating acetylation mark H3K9ac was detected in the nitrogen rich condition, while almost no H3K9 acetylation is present under low-nitrogen conditions (Figure 4.3). The proteins of seven apicidin genes were measured in the proteomics experiment. Two genes (FFU_00005 and 00010) were present in significant higher concentration in the nitrogen rich medium (log_2 fold change > 7). Another five genes (FFUJ_00006 to FFUJ_00008, FFUJ_00011 and FFUJ_00013) were detected exclusively in this condition. However, the NPS31 signature enzyme (FFUJ_00003) could not be measured using mass spectrometry [231]. Taken together transcriptomics- and proteomics- data provide evidence for co-expression of the apicidin F cluster genes under high nitrogen conditions. Further, the enrichment of H3K9ac on the cluster region hints that high-nitrogen conditions can activate expression of the apicidin F pathway.

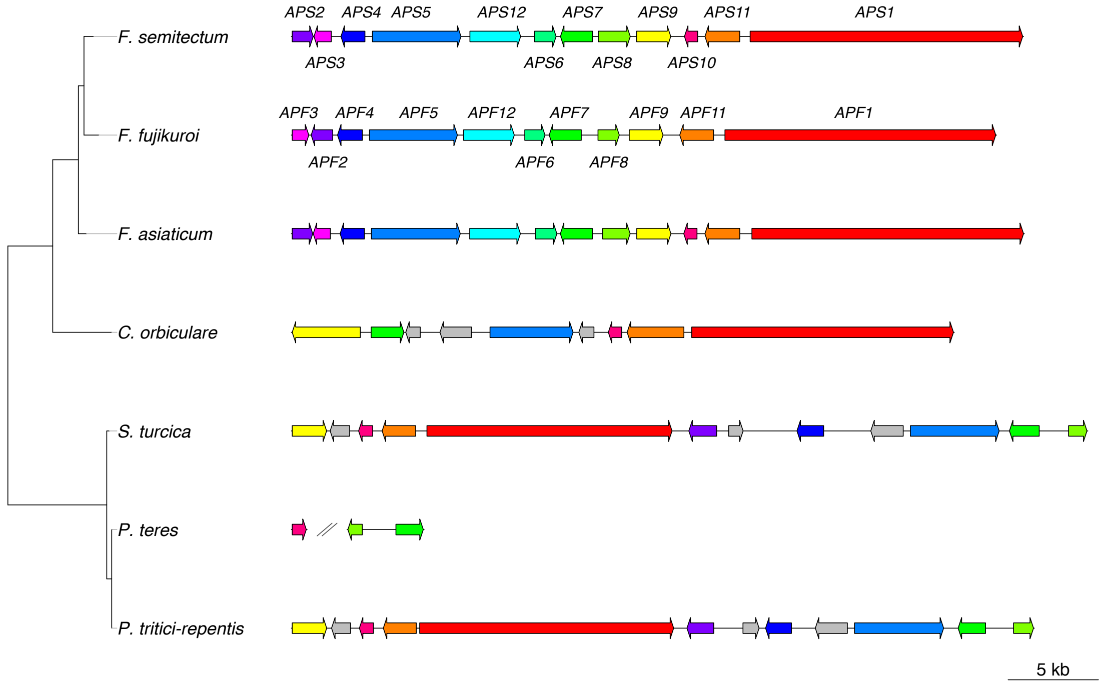

Figure 4.2: Phylogenetic distribution of the apicidin gene cluster. Colored arrows illustrate genes and their orthologs of the apicidin gene cluster in *F. semitectum*. Orthologs of the cluster could not be found in the other sequenced species of the *Gibberella fujikuroi* species complex as well as in the strains of *F. oxysporum* and *F. graminearum*. Orthologs in *Pyrenophora teres* are separated on two supercontigs indicated by dashes. Description of predicted gene functions are listed in table 4.4.

4.2.5 Overrepresented promoter motifs are involved in the regulation of apicidin F biosynthesis

An experimental elucidation of the biosynthesis pathway and the molecular structure of the novel compound apicidin F was performed by Eva Niehaus *et al.* [154]. Additionally, the regulatory role of the transcription factor Apf2 (FFUJ_00012) in the cluster was shown by deletion- and over-expression mutants. I am now interested in the promoter binding site in order to determine further target genes of the transcription factor. In assuming conservation of regulatory elements in orthologous genes I analyzed the promoters of *F. fujikuroi* and *F. semitectum* clusters for overrepresented motifs. The analysis using the Phylocon algorithm [223] resulted in a putative binding motif with the consensus sequence 5-TGACGTGA-3 which is

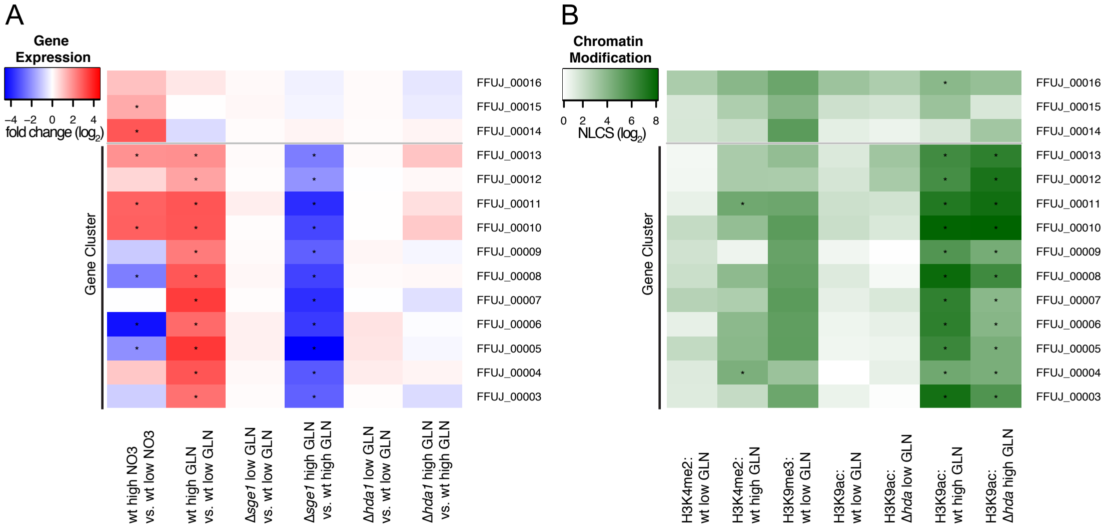

Figure 4.3: Gene expression and epigenetic modification of the apicidin F cluster and adjacent genes of *F. fujikuroi* wild type (wt), Δ*hda1* and Δ*sge1* mutant under limiting (low) and sufficient (high) glutamine (GLN) and nitrate (NO3) conditions. Genes are listed in chromosomal order on y-axis, gene cluster is indicated by vertical black bar. No upstream genes are listed as the cluster is located in the subtelomeric area. **A:** Fold changes (log_2 scale) of gene expression between two experimental conditions. Up-regulated genes compared to the control condition are depicted in red, down-regulated genes in blue. Asterisk indicate significant changes in gene expression (fold change >2, P-value <0.05). Abbreviations of experimental conditions on x-axis are given in table 4.4. **B:** NLCS values (log_2 scale) indicating degree of chromatin modification for each experimental condition. Significant signals are indicated by asterisk (signal probability >0.95).

Table 4.4: Predicted function of apicidin cluster genes in *F. semitectum* and their orthologs in *F. fujikuroi*. No ortholog of APS10 was found in *F. fujikuroi*.

F. semitectum cluster gene	Ortholog in *F. fujikuroi*	Predicted function
APS1	FFUJ_00003 (*APF1*)	Non-ribosomal peptide synthase
APS11	FFUJ_00004 (*APF11*)	Major facilitator superfamily transporter
APS9	FFUJ_00005 (*APF9*)	FAD-dependent monooxygenase
APS8	FFUJ_00006 (*APF8*)	Cytochrome P450
APS7	FFUJ_00007 (*APF7*)	Cytochrome P450
APS6	FFUJ_00008 (*APF6*)	O-Methyltransferase
APS12	FFUJ_00009 (*APF12*)	Cytochrome b5-like
APS5	FFUJ_00010 (*APF5*)	Fatty acid synthase, alpha subunit
APS4	FFUJ_00011 (*APF4*)	Aminotransferase
APS2	FFUJ_00012 (*APF2*)	bANK transcription factor
APS3	FFUJ_00013 (*APF3*)	Δ1-Pyrroline-5-carboxylate reductase
APS10	-	Related to ketoreductase

significantly enriched in the cluster promoters compared to the genome wide distribution of the motif in *F. fujikuroi* (Fisher's exact test [65], P-value < 0.01). This motif can be found in all cluster gene promoters in *F. fujikuroi* except the transcription factor gene *APF2* itself. In *F. semitectum* it is not present in *APS2*, *APS3* and *APS10*. The bidirectional promoters of the genes APF1/APF11, APF4/APF5 and APF7/APF8 contain the motif only once. A query against the TRANSFAC database [236] revealed the similar, but perfect palindromic sequence 5-TGACGTCA-3 which has been shown to be the binding site for the mammalian global bZIP transcription factor CREB (cAMP-response element-binding protein) that is involved in glucose homeostasis [147]. The binding site could be experimentally verified by Niehaus *et al.* [154] to have a regulatory impact on the expression on the apicidin-F pathway. In fact after mutating the binding site upstream of *APF1* a loss of apicidin-F production could be observed [154].

In order to determine the global impact of the TF on gene expression I investigated the amount of putative target genes among the whole set of differentially up-regulated genes. A significant overrepresentation (P-value = 0.00023) of 315 putative target genes among the 1470 significantly up-regulated (fold change > 2, P-value < 0.05) genes could be determined under nitrogen rich conditions.

4.2.6 Palindromic promoter motif correlates with gene expression

To determine which other clusters are regulated by a secondary metabolism specific transcription factor, a scan for conserved binding motifs of the promoter regions of each cluster was performed. I applied Fisher's exact test [65] to determine the significance of motifs found in cluster gene promoters compared to the genome wide distribution of motifs. In total 19 clusters in *F. graminearum* and 13 clusters in *F. fujikuroi* (Table A.3) with significantly overrepresented promoter motif (P-value < 0.01) could be identified. One of the most significant motifs in *F. graminearum* is the palindrome 5'-GTGGtgCCAC-3' in the cluster FG_C02 as previously reported [75]. The cluster consists of 16 genes (FGSG_11653 - FGSG_00049) of which 12 genes carry the putative palindromic binding site in their promoters (Table A.1). As already reported, the gene expression of 11 of the 12 putative target genes is significantly increased in the wildtype when growing on agmatine- compared to glutamine-medium (4.6 to 9.2 fold on log_2 scale, P-Value < 0.05) [75]. Interestingly, the expression of all 12 genes is significantly repressed in the Δfgp mutant growing on putrescine-medium (1 to 8.5 fold on log_2 scale, P-value < 0.05) [101]. Additionally under nitrogen starving conditions a significant increase in gene expression in seven genes takes place (3.4 to 4.8 fold on log_2 scale, P-value < 0.05) [82] (Figure 4.1). The predicted motif occurs in only 4% of all promoters on the genome and is significantly enriched in this cluster (P-value = 6.8e-13).

4.2.7 Significantly overrepresented promoter motifs in clusters of known metabolites

In Section 4.2.3, I postulated four genes that are putatively involved in the malonichrome biosynthesis pathway because of co-expression and shared promoter motifs of the genes that are locally clustered with the NPS signature enzyme. This evidence could also be determined in other clusters of known metabolites in *F. graminearum* like the butenolide or trichothecene cluster. The experimentally elucidated genes of the butenolide cluster [87] exhibit significant differential expression *in planta* during barley (FG1) and wheat infection (FG12 and FG15) whereas the neighboring genes do not correlate in expression profiles (Figure 4.1). The significantly overrepresented sequence motif 5'-[AT]A[AG]T[GT][CG][TA]CCG-3' could be identified in all promoters of differentially expressed genes but not in the adjacent genes that do not correlate in gene expression. Likewise, an overrepresented sequence motif could also be found in the promoters of the ferricrocin gene cluster. Although the cluster genes were not significantly expressed under any experimental condition the overrepresented (P-value < 0.01) putative binding site 5'-TTTGGCAACA-3' was found in the promoters of all nine ferricrocin cluster genes and hints towards a co-regulation of this cluster.

In *F. fujikuroi* and *F. proliferatum* a novel promoter motif in the fumonisin cluster was found. In both species the putative binding site 5'-GTATCCGA-3' is represented in all promoters of the 15 cluster genes. Compared to the genome wide distribution of the motif a significant overrepresentation is determined in the clusters of both species (P-value < 0.01).
In the cluster of the mycotoxin trichothecene in *F. graminearum* an overrepresented (P-value $= 0.0042$) promoter motif was found. This motif matches the known binding site of the Cys2His2 zinc-finger regulatory protein Tri6, which is a positive regulator of the orthologous trichothecene genes in *F. sporotrichioides* [93]. Interestingly, the motif 5'-TnAGGCCT-3' in this cluster is not only located in the promoters of nine of the 12 currently postulated trichothecene cluster genes (FGSG_03543 (*TRI14*) to FGSG_03532) [32] but also in two promoters of adjacent downstream genes (FGSG_03531 to FGSG_03529). However, Nasmith *et al.* experimentally determined the binding affinity of Tri6 to a different motif, which is located in five promoters of the cluster genes but could not be detected by this approach [151]. The three additional genes are also supported by expression data during barley and wheat infection (Figure 4.1). All trichothecene cluster genes, including the three additional genes, are co-expressed with an increase in gene expression until the third day of growth except for the genes cytochrome P450 (FGSG_03542), FGSG_03541 and FGSG_03533. This observation strongly supports the assumption that the new cluster genes might be involved in yet unknown steps of the trichothecene biosynthetic pathway or a trichothecene related function.

In collaboration with the lab of Gerhard Adam we tried to find evidence of possible roles of the additional genes according to their functional characterization. Because the protein encoded by FGSG_03529 is predicted as "1,3-betaglucosidase" we hypothesized that this enzyme may be involved in the reactivation of plant-neutralized DON [156, 168]. Another hypothesis is that the gene FGSG_03530, which encodes a hypothetical protein similar to an acetylesterase, may play a role in the deacetylation of the biosynthetic precursor 3,15-diacetyl-DON [2]. However, the experimental examination in a heterologous yeast setup by the lab of Gerhard Adam did not confirm these hypotheses [191].

4.2.8 Predicted secondary metabolism clusters exhibit characteristic gene expression *in planta*

To select predicted clusters that play a role during host infection, I focus on the gene expression measurements of four experiments *in planta*. The time series data cover time intervals of the first hours after infection up to several days. In calculating the Pearson correlation coefficient of neighboring genes a correlation of gene expression profiles in 26 clusters was found, which is above the 95th percentile of randomly selected genes. Beside the known synthesis genes of aurofusarin, trichothecene, butenolide, triacylfusarinin and malonichrome I determined correlations in 20 predicted clusters of which the associated metabolite is unknown.

The expression profiles of cluster FG_C16 which contains a PKS, a terpenoid synthase and two methyltransferases is significantly increased after 72 hours post inoculation (hip) on barley [82]. During the infection process of wheat, the expression of genes increases significantly after 96h and decreases afterwards [131]. However in a second experiment, gene expression after 35 days post inoculation is still increased compared to the control measurement in complete defined media [200] (Figure 4.4).

Like the gene expression profile of the aurofusarin cluster genes which shows a peak at 64 hpi on wheat seedling coleoptiles, the cluster FG_C64 show a very similar profile. The cluster consists of nine genes comprising two NPS signature enzymes and one cytochrome P450 protein. All genes are up-regulated simultaneously at 64 hpi and down regulated afterwards as already reported by Zhang et al. [248]. Beyond that, an increase in gene expression can also be observed during the infection of barley [82] where the maximum expression is reached at 96 hpi [131] (Figure 4.5).

15 clusters in *F. graminearum* were found that show significant co-expression evidence but consist only of tailoring enzymes and additional proteins. Eight of the 15 exhibit a significant expression correlation during *in planta* infection. For example the cluster FG_C09 contains five cytochrome P450 genes and is significantly down-regulated during the infection of wheat (FG12) but significantly up-regulated in the Δ*fgp1* mutant on putrescine medium (FG18). The cluster FG_C14 shows also an overrepresentation of four P450 genes and is significantly increased in gene expression during wheat infection (FG12).

These results show that the genes of predicted clusters are co-expressed *in planta* and show similar expression profiles like those of the trichothecene or aurofusarin cluster genes. Therefore, it is likely that these clusters code for novel metabolites which have an impact in plant pathogenesis and are targets for further experimental investigation.

4.3 Discussion

Fungal secondary metabolism provides a versatile source of bioactive compounds with importance in medicine and agriculture. Genes involved in fungal secondary metabolism pathways are often locally clustered in fungal genomes. While classical gene cluster studies focused on single gene clusters or individual genes involved in certain specific functions, the increasing number of available fungal genome sequences due to next generation sequencing techniques enables large scale mining for gene clusters of yet unknown compounds. Genome wide high throughput gene expression experiments provide additional hints for the identification of co-regulated secondary metabolism pathway genes. The growing number of available fungal genomes bears also the opportunity to trace the evolutionary history of species specific gene clusters.

In *Fusarium graminearum*, a total of 51 putative signature enzyme encoding genes have been predicted, exceeding the number of currently known secondary

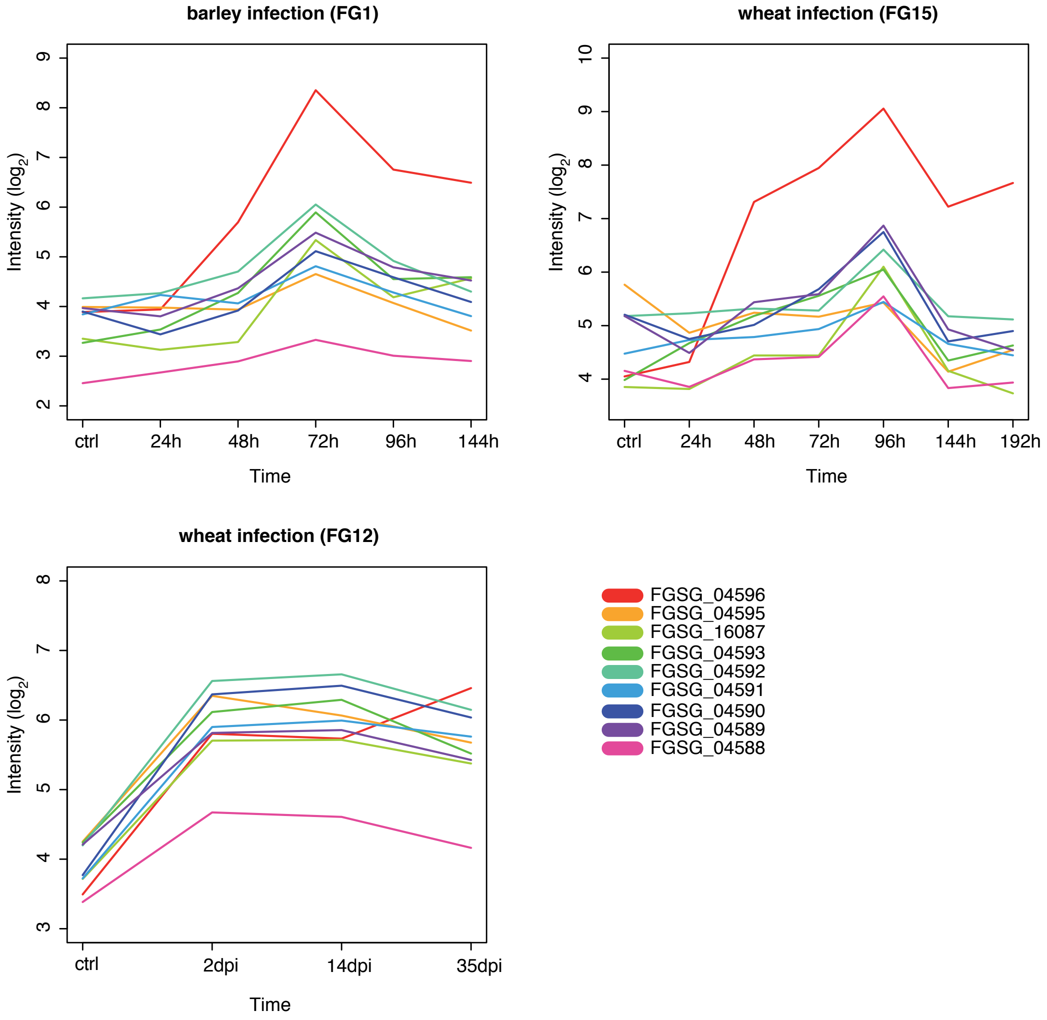

Figure 4.4: Gene expression profiles of the cluster FG_C16 in *F. graminearum* during barley (FG1) and wheat infection (FG12, FG15). X-axis shows time after inoculation on wheat, gene expression intensity is drawn on y-axis. Figure is adopted from Sieber *et al.* [191]

metabolites in this organism [30, 47, 86, 239]. By screening the *F. graminearum* genome for spatially clustered signature and tailoring enzymes, 67 potentially functional gene clusters were identified [191]. Most of the clusters contain signature enzymes with unknown synthesis product and therefore constitute candidates of novel secondary metabolism pathways.

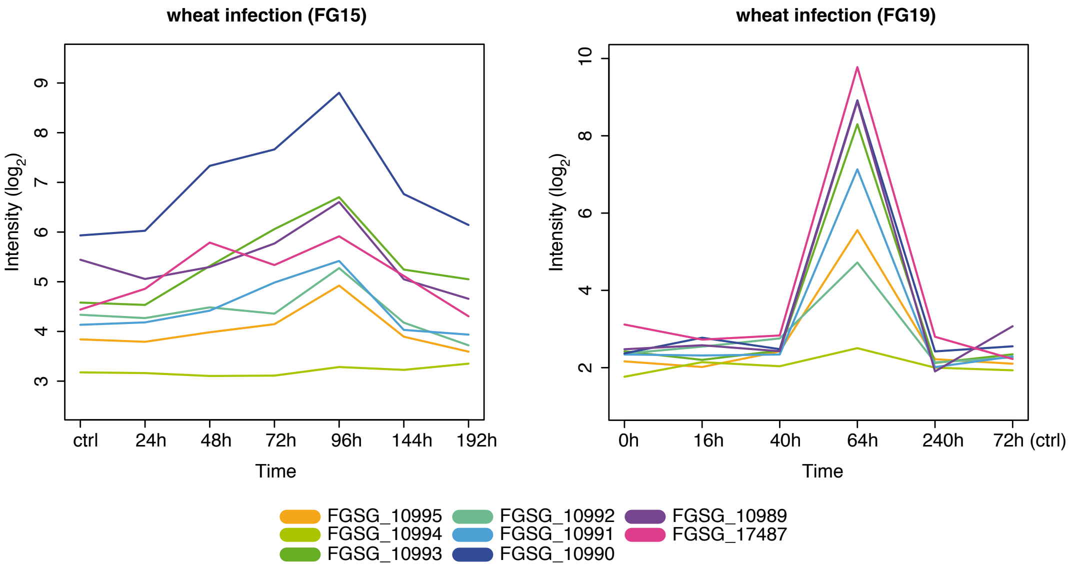

Figure 4.5: Gene expression profiles of cluster FG_C64 in *F. graminearum* during wheat infection (FG15, FG19). X-axis shows time after inoculation on wheat, gene expression intensity is drawn on y-axis. Figure is adopted from Sieber *et al.* [191]

Also clusters that lack a signature enzyme but exhibit an overrepresentation of tailoring enzymes were predicted. An example is the cluster FG_C09 which contains five P450 enzymes. These clusters may also be involved as modifiers in secondary metabolism pathways of other clusters or may be remnants of formerly larger clusters. Vice versa there are also clusters containing more than one signature enzyme. FG_C30 for example consists of a terpene synthase, a NPS and four P450 enzymes.

Large clusters that contain many predicted secondary metabolism enzymes could possibly be the result of the fusion of two clusters which now act as one supercluster like recently shown in *Aspergillus fumigatus* [230]. An example is cluster FG_C15, which comprises the 2 PKS genes (FGSG_17745 and FGSG_15980 - formerly described as *PKS3* and *PKS14*, [112]), the oxidoreductase (FGSG_15979) and the specific transcription factor (FGSG_02398), but additionally also contains other co-regulated genes of still unknown function including the *NPS15* gene.

The predicted clusters also include genes identified as key enzymes for biosynthesis of known compounds (Table 4.3). Particularly, *NPS1*, *NPS2* and

Table 4.5: Functional gene descriptions of the predicted clusters C16 and C64 illustrated in Figures 4.4 and 4.5.

Cluster_ID	Position	Gene_Code	Description
FG_C16	1	FGSG_04596	related to O-methyltransferase
	2	FGSG_04595	related to hydroxylase
	3	FGSG_16087	hypothetical protein
	4	FGSG_16088	related to 3-ketoacyl-acyl carrier protein reductase
	5	FGSG_04593	related to para-hydroxybenzoate polyprenyltransferase precursor
	6	FGSG_04592	related to light induced alcohol dehydrogenase Bli-4
	7	FGSG_04591	probable farnesyltranstransferase (al-3)
	8	FGSG_04590	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	polyketide synthase
FG_C64	1	FGSG_10996	conserved hypothetical protein
	2	FGSG_10995	related to multidrug resistance protein
	3	FGSG_10994	conserved hypothetical protein
	4	FGSG_10993	related to selenocysteine lyase
	5	FGSG_10992	related to polysaccharide deacetylase
	6	FGSG_10991	related to benzoate 4-monooxygenase cytochrome P450
	7	FGSG_10990	related to AM-toxin synthase (AMT)
	8	FGSG_10989	conserved hypothetical protein
	9	FGSG_17487	related to non-ribosomal peptide synthase

NPS6 found in three clusters are the only genes known to be involved in production of malonichrome, ferricrocin and triacetylfusarinin, respectively. The clusters may require additional genes to complete certain biosynthetic pathways. Correlation in expression profiles and the presence of overrepresented promoter motifs in gene clusters provide evidence of putative pathway genes. Deletion analysis and heterologous expression of the gene clusters will help to validate them.

The predicted clusters in *F. graminearum* are comparable to gene clusters recently defined by three previously utilized approaches: 'secondary metabolite biosynthetic (SMB) gene clusters' [132], 'Secondary Metabolite Unique Regions Finder (SMURF)' [107] and 'AntiSMASH' [19]. All three analyses were not able to identify the already known butenolide cluster. However butenolide was detected by a generalized search of co-regulation networks [120]. SMB missed the NPS class secondary metabolite gene clusters and SMURF missed the TPS class gene clusters. The SMB cluster search focused only on two classes (PKSs and terpene synthase (TSs)) and SMURF used four classes of SM (PKSs, NPSs, Hybrid NPS-PKS, and prenyltransferases (DMATSs)). AntiSMASH takes more enzyme classes into consideration and is also able to detect clusters without signature enzymes. My approach in contrast considers four types of signature enzymes (PKS, NPS, TPS and DMATS) as well as five tailoring enzyme classes (methyltransferases, acyltransferases, oxidoreductases, glycosyltransferases and cytochrome P450s) and takes transcription factors and transporter enzymes into account that might contribute to regulation and secretion of the metabolite. Overall this approach results in the most comprehensive set of potential SM clusters containing 30 clusters not found by any of the previous analyses. Vice versa SMURF and AntiSMASH detected nine and 14 clusters not found by my pipeline, respectively. Because of evidence in terms of co-expression and common promoter motifs in total ten additional clusters from the two prediction tools were included in the candidate set. The previously predicted clusters in the SMB approach [132] were all detected by this approach.

4.3.1 Three gene clusters associated with an unknown metabolite are possibly involved in plant infection

Three novel gene clusters (FG_C62, FG_C16, FG_C64) are expected to play important roles during plant infections, supported by co-expression and their collection of predicted functions. All three clusters contain at least one signature enzyme as well as additional tailoring enzymes and exhibit a significant change in gene expression during plant infection. The NPS containing cluster FG_C62 is induced after 64 hpi inside wheat coleoptiles but repressed while growing on the stem base of wheat, which hints towards a specific regulation of these genes dependent on certain plant tissues (Figure 5.3B). The core part of the cluster is co-expressed and conserved in *Cochliobolus heterostrophus* and *Pyrenophora teres*. Because the NPS gene is not co-regulated with the core and orthologs of the NPS are located on separate contigs, it is difficult to say whether it is part of the same biosynthesis pathway. The clusters FG_C16 and FG_C64 exhibit an increase in gene expression on wheat and barley as well (Figures 4.4 and 4.5). Like the expression profiles of the aurofusarin cluster, the profiles of the predicted clusters reach a peak after 64 to 96 hpi followed by a decrease afterwards. The *NPS9* (FGSG_10990) and the transporter gene (FGSG_10995) of cluster FG_C64 were mutated by Zhang *et al.* which resulted in reduced virulence [248].

The cluster FG_C16 containing *PKS15* and a further 10 genes (Figure 4.4 and Table 4.5) is one of the most promising clusters for further analysis. *PKS15* was shown to be expressed during plant infection and has been considered as one of the strong candidates producing a metabolite of unknown function with a role in virulence [70]. However, not much information has been determined for the genes adjacent to *PKS15*: one terpenoid synthase, one cytochrome P450, one secreted protein and six additional enzymes such as a methyltransferase, a dehydrogenase/reductase and a 3-ketoacyl-acyl carrier protein reductase. Further characterization of the enzymes may point to the associated metabolite structures. No pathway-specific transcription factor is found in this cluster. Transcription seems to be controlled by other regulatory proteins affecting chromatin structure, such as a histone methyltransferase [46]. Available evidence for genes involved in a common pathway or function with *PKS15* will promote targeted research on this putative SM cluster.

4.3.2 Gene clusters possibly co-regulated due to shared promoter motifs

Many gene clusters are regulated by secondary metabolism specific transcription factors [165] and global regulators [151] as well. Because of the high frequency of binding sites of global regulators it is difficult to distinguish them from randomly occurring, non-functional motifs. I therefore concentrate on pathway specific transcription factors in assuming that the target binding sites are statistically overrepresented in promoters of cluster genes compared to the distribution of the motifs on the whole genome. I took also promoter sequences of orthologous genes into account with the assumption that regulatory elements are conserved between species. The discovery of conserved promoter motifs as well as orthologous genes in aflatoxin-producing *Aspergillus* species [60, 63] is an example of the possible benefit of such comparisons. The determination of specific motifs helps to identify 'cryptic' gene clusters that are silent under the observed experimental conditions due to a missing required external stimulus.

A statistical overrepresentation of a conserved motif in the promoters of the trichothecene mycotoxin genes was discovered. The identified motif was previously reported by Hohn *et al.* as binding site of the Tri6 transcription factor of the orthologous trichothecene gene cluster in *F. sporotrichioides* [93]. As the promoter motif is also present in two co-regulated downstream genes of the cluster, a role in trichothecene biosynthesis for these genes is suggested (Figure 4.1 and Table A.1). Especially the functional annotation of two genes (FGSG_03529 and FGSG_03530) propose plausible enzymatic activities related to the TRI pathway. However, experimental investigation by the lab of Gerhard Adam could not determine an involvement of the genes in the hypothesized functions of deacetylation of acetyl-DON and removal of glucose from DON-3-glucoside [191]. Further investigation by more sophisticated and direct approaches are required to explore and validate the putative functions of these genes.

Recently, Nasmith *et al.* proposed high binding affinity of Tri6 to another motif which consists of repeats with the pattern GTGA [151]. The 198 Tri6-target genes predicted by ChIP-seq experiments [151] contain five of the TRI-cluster genes but none of our proposed additional genes. However, the overrepresentation of our determined motif and its conservation in *F. sporotrichioides* suggests regulatory importance of the binding site by a second transcription factor.

Besides the motif of the well-studied TRI cluster I determined a putative motif in the butenolide synthesis genes, which is significantly enriched, compared to the genome wide motif distribution. Furthermore, it is supported by the gene expression profile of the cluster genes (Figure 4.1 and Table A.1). Overrepresentation and correlation to expression data hypothesize that the predicted motif might constitute the binding site for the zinc finger transcription factor, which his located in the cluster. So far, there are no transcription factors associated with this binding pattern in the Jaspar or Yeastract database [141, 207]. The palindromic motif in FG_C02, which has been previously determined by Gardiner *et al.* was confirmed by our approach [75]. Due to the amount of the analyzed expression data it could be shown that the putative target genes are differentially expressed in even more environmental conditions than reported before (Figure 4.1 and Table A.1). This adds evidence to the hypothesis that the predicted binding site has a regulatory function.

Structures of transcription factor binding sites can be quite diverse. Especially short or degenerated recognition sequences are difficult to predict and enrichment alone does not necessarily predict functionality of the motifs. Experimental approaches like ChIP-seq or ChIP-chip experiments will be necessary to confirm predicted binding sites.

4.3.3 Discovery and characterization of the apicidin F gene cluster

Prediction of gene clusters and following protein similarity analysis revealed orthologs of the apicidin biosynthetic genes of *F. semitectum* in the genome of *F. fujikuroi* (Figure 4.2). Apicidin is a compound with histone deacetylase inhibitory activity [99], whereupon co-expression of apicidin cluster genes and differential acetylation of histone proteins during secondary metabolism inducing conditions suggests a role in virulence. Furthermore, repression of cluster genes in the Δ*sge1* mutant hints towards a global positive-regulatory control of the cluster genes. Like in *F. semitectum* over-expression of the pathway specific transcription factor *APF2* lead to activation of most of the pathway genes and to an increase in apicidin F production [99, 154]. A promoter analysis for putative pathway specific regulatory binding sites exhibited a significantly overrepresented motif that is also present on the promoters of orthologous genes in *F. asiaticum* and *F. semitectum.* Subsequent *in vitro* mutation of the motif resulted in loss of apicidin F synthesis [154]. Furthermore, a significant overrepresentation of putative target genes with the determined promoter motif among genes with similar expression profile could be determined. A complex regulatory network controlling the synthesis of apicidin F is suggested by the Apf2 dependent gene expression, cluster specific overrepresented promoter motifs, differential expression upon Δ*sge1* deletion and significant enrichment of H3K9ac marks (Figure 4.3).

Plausible scenarios for the discontinuous phylogenetic distribution of the apicidin genes include horizontal gene transfer and individual loss. However, no significant evidence in terms of different GC ratios could be determined that underline the HGT hypothesis. As the clusters are located near chromosome ends at least in *F. fujikuroi* and *F. asiaticum* the increased mutation rate in subtelomeric regions may have lead to a loss of the cluster in closely related species like *F. verticillioides*, *F. proliferatum* or *F. graminearum*.

Chapter 5

Evolution and origin of secondary metabolism gene clusters

In the last chapter I showed that the presence of secondary metabolism gene clusters differs considerably even between closely related *Fusarium* species (Table 4.3). In this chapter I will investigate the phylogenetic distribution of the predicted clusters of 20 fungal species determined in Chapter 4. I will search for mechanisms that influence the evolution and distribution of gene clusters and take predicted transposable elements of Chapter 3 into account. Most results of this chapter are published and discussed in:

> **Sieber CMK***, Lee W*, Wong P, Münsterkötter M, Mewes HW, Schmeitzl C, Varga E, Berthiller F, Adam G, Güldener U.
> The *Fusarium graminearum* genome reveals more secondary metabolite gene clusters and hints of horizontal gene transfer.
> *PLoS One.* 2014, 9, e110311

*equal contributions

5.1 Methods

5.1.1 Determination of orthologous clusters and evidences for horizontal gene transfer

A subset of the SIMAP protein similarity database [7, 175] was used to determine orthologous cluster genes in other species. Proteins of 234 fungal genomes, 150 bacterial reference genomes and the proteins of *Arabidopsis thaliana* were defined as search space (Table A.2). All protein hits that constitute a bidirectional best hit between *F. graminearum* and the target organism with an e-value below 1e-04 and at least 50% overlap in the amino acid sequences were taken into account.

The gene order in orthologous clusters is often not conserved. Thus, collinearity is not an adequate criteria to determine chromosomal aggregations of bidirectional best cluster hits. I developed ClustRFindR (implemented in the statistical computing language R [174]) which queries clustered bidirectional best protein hits that are not necessarily collinear in the target organism but located close to each other. As distance threshold between the matching genes I chose the double length in nt of the gene cluster in the query organism. To respect that some genome assemblies consist of thousands of small contigs, a split of the cluster on more than one contig is allowed in case that the aggregation of orthologs on one contig is at least three. Clusters with more than 50% of conserved genes between two species are selected. To test for evidence of horizontal gene transfer I calculated GC ratios of cluster genes and compared them to the distribution of GC ratios of all genes in the host genome. I applied a two-sided Kolmogorov-Smirnov (KS) test [140] to test for significant differences between the GC ratio distributions.

5.1.2 Phylogeny of gibberellin cluster genes

To infer the phylogenetic relationship of gibberellin cluster genes I calculated a phylogenetic trees based on their codon-aligned nucleotide sequence (Figure 5.7). I used Mafft [103], PAL2NAL [204] and Gblocks [206] to prepare the alignment and PhyML [81] to calculate the tree as described in Section 2.1.3 but with an approximate likelihood-ratio test (aLRT) instead of the bootstrapping to test branches.

5.1.3 Visualization orthologous clusters

For the visualization of orthologous clusters, I developed ClustRPlottR, a method that is based on the GenoPlotR R-package [83]. ClustRPlotR takes the output of the orthologous cluster search function and displays the clustered orthologous genes according to their position on the chromosome. Orthologous groups are indicated by the same color. A phylogenetic tree in newick format enables a mapping of clusters on the tree and an arrangement of the hits according to their phylogenetic relationship as exemplified in Figure 5.5 (see below).

I infered phylogenetic relationships of species which contain orthologous secondary metabolism gene clusters as described in Section 2.1.3. Because annotation quality of some fungal genomes varies and orthologs of *RPB1*, *RPB2* and *TEF1α* could not be identified in all species of interest, I calculated the trees based on 15 putative housekeeping genes that are annotated as FunCat [184] category "Energy" (table 5.1). Bidirectional best hits of the proteins were determined using the SIMAP database [175].

5.2 Results

5.2.1 Ortholog analysis of predicted clusters in *Fusarium*

After predicting secondary metabolism clusters in all *Fusarium* genomes I am interested in the distribution of cluster orthologs in and outside the respective clades and phyla. In the *Gibberella fujikuroi* species complex (GFC) nine clusters that occur only in one GFC species were found. Two of them have orthologs in the *Fusarium graminearum* species complex (FGC) like the apicidin cluster which can also be found in *F. asiaticum* and the FV_C20 cluster in *F. verticillioides* is present in *F. oxysporum.* Five of the nine clusters can not be found in any other species. I found eight clusters in the FGC that can be found in exactly one genome of the three FGC species. Interestingly six of them are present in *F. asiaticum* whereas *F. pseudograminearum* has no unique clusters in the FGC. Two of the FGC-unique clusters in *F. asiaticum* have orthologs in the *Aspergillus* phylum. Among the 13 analyzed *F. oxysporum* species eight clusters that had no orthologs

Gene code	Description
FFUJ_01475	probable ZWF1 - glucose-6-phosphate dehydrogenase
FFUJ_01340	related to vacuolar ATP synthase subunit H
FFUJ_02998	probable ATP3 - F1F0-ATPase complex, F1 gamma subunit
FFUJ_04860	related to 3-hydroxyisobutyrate dehydrogenase
FFUJ_04893	probable ubiquinol-cytochrome-c reductase Rieske iron-sulfur protein
FFUJ_05032	related to kinesin
FFUJ_07582	related to cytochrome-c oxidase chain VIIa
FFUJ_08292	probable H+-transporting ATPase, vacuolar, 41 kDa subunit
FFUJ_08315	probable beta-succinyl CoA synthase precursor
FFUJ_08541	probable oxaloacetate/sulfate carrier
FFUJ_08584	probable pyruvate dehydrogenase (lipoamide) alpha chain precursor
FFUJ_09575	probable NDE1 - mitochondrial cytosolically directed NADH dehydrogenase
FFUJ_09852	related to mitochondrial serine-tRNA ligase
FFUJ_13760	probable atp-specific succinyl-coa synthase alpha subunit
FFUJ_13774	probable histone acetyltransferase

Table 5.1: *Fusarium fujikuroi* genes for phylogenetic tree calculation

among the other *F. oxysporum* strains were determined. Six clusters that could be found only once in *Fusarium* have orthologs outside the *Fusarium* phylum, two of them are present in *Aspergillus.* For 13 clusters no orthologous clusters in any other species could be identified.

5.2.2 Analysis of horizontal gene cluster transfer

Because of the observed non-uniform distribution of some clusters among closely related species I am interested to which extent horizontal gene transfer may play a role in the inheritance of gene clusters. Therefore I explored significant differences in the GC ratio of cluster genes and their host genome in applying a Kolmogorov-Smirnov (KS) test [140] on the distribution of GC ratios. I tested differences in GC ratios for all clusters in *Fusarium* that have orthologs outside the phylum. In 19 clusters a significant difference (P-value <0.01) between the GC ratios of the cluster genes compared to the GC ratio distribution of the whole *Fusarium* genome, indicating a putatively inherited cluster in *Fusarium*. One example of a putatively inherited cluster is described in Section 5.2.4. Vice versa 12 clusters where the orthologous cluster genes differ in their GC ratio significantly from the host species outside the *Fusarium* phylum were found. Among the determined set of candidate HGT clusters are previously demonstrated cases of horizontal cluster transfer like the fumonisin [108] and the bikaverin [37, 38] cluster. In *F. graminearum* with five clusters the highest amount of HGT evidences could be determined. The putatively laterally inherited clusters FG_C47 and FG_C62 are described in the Sections 5.2.3 and 5.2.5.

5.2.3 The PKS23 cluster in *F. graminearum* shows evidence of horizontal gene cluster transfer into the *Botrytis* lineage

In the predicted gene cluster FG_C47 in *F. graminearum* (FGSG_08209 - FGSG_17085) several hints for HGT could be found. The cluster includes the signature enzyme PKS23 (FGSG_08208), a NPS, a methyl transferase and a cytochrome P450 enzyme. All genes are repressed simultaneously during the infection of wheat (Stephens et al., 2008) compared to the expression rate on complete defined medium (2.1 to 4.4 on log_2 scale, P-value < 0.05). Further, the influence of DON-inducing agmatine in growth medium causes also a significant decrease in gene expression of the whole cluster (4.5 to 6.7 on log_2 scale, P-value < 0.05) [75] (Figure 5.1B)). Neither the metabolite synthesized by this cluster nor its function are known so far. A complete orthologous cluster can also be found in the closely related *F. asiaticum* but not in *F. pseudograminearum* where the PKS enzyme (FGSG_08208) is the only conserved cluster member. Anyhow, orthologs of the surrounding genes of the *F. graminearum* cluster constitute a collinear region on a different scaffold than the PKS. Interestingly, an ortholog of this cluster can be found in the two *Botrytis fuckeliana* strains B05.01 and T4 whereas the neighboring genes are not present in both genomes. All other inspected genomes lack a syntenic gene cluster. The clusters in the two *Botrytis fuckeliana* strains contain an additional P450 gene (B05.01: BC1G_09046, T4: BofuT4_059840.1) that is not present in *F. graminearum* and a NPS like enzyme that is unique for the B05.01 strain (BC1G_09041). Additionally a Gypsy transposable element BOTY_I [54] (Repeatmasker SW-score: 40718), consisting of three ORFs, could be identified by the genome wide repeat analysis in Chapter 3 (Figure 5.1A). The GC-contents of the orthologous clusters are very similar (median GCcontent of 50.0% and 52.6% for *F. graminearum* and *B. fuckeliana* respectively) whereas the distributions of genome-wide GC contents differs considerably (median GC content of 51.3% and 46.2% for *F. graminearum* and *B. fuckeliana*, respectively) (Figure 5.1C). In performing a two-sided KS test for the GC distributions a significant (P-value = 2.2e-16) difference between the GC content of the *Botrytis* cluster genes and the genome-wide distribution of *Botrytis* was obtained. On the other hand the null hypothesis could not be rejected in comparing the GC content of the same cluster ORFs to the genome-wide distribution of *F. graminearum* (P-value = 0.4277). These results suggest a potential horizontal gene cluster transfer from the *Fusarium* lineage into *B. fuckeliana*.

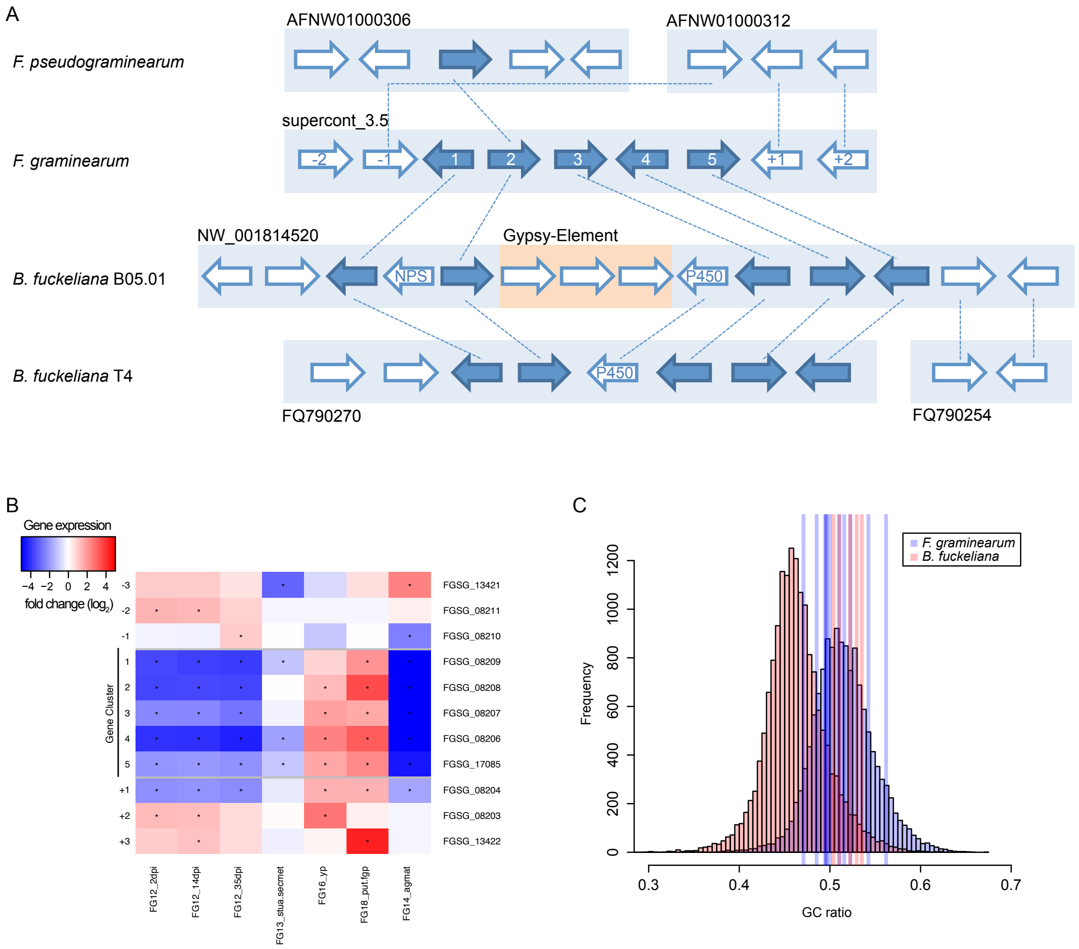

Figure 5.1: A: Predicted gene cluster (FG_C47) in *Fusarium graminearum* and orthologous genes in *F. pseudograminearum* and the *Botrytis fuckeliana* strains B05.01 and T4 (solid dark blue arrows) on their respective supercontigs (light blue boxes). Adjacent genes are illustrated as white arrows, dashed lines depict orthologous groups. The gypsy transposable element in *B. fuckeliana* B05.01 is indicated as orange box. Enumeration in *F. graminearum* is according to Table 5.2. **B:** The heatmap illustrates fold changes in gene expression (log_2 scale) between two experimental conditions. Asterisk indicate significant differences between experimental conditions (fold change >2, P-value <0.05). Genes are listed in chromosomal order on y-axis. Abbreviations of experimental conditions on x-axis are according to Table 2.2. Gene cluster is indicated by vertical black bar. **C:** Histograms show whole genome distributions of open reading frame GC ratios in *F. graminearum* (blue) and *B. fuckeliana* B05.01 (red). Vertical lines illustrate GC ratios of cluster genes. Figure is adopted and modified from Sieber *et al.* [191].

5.2.4 Evidence of an HGT-inherited cluster in two *F. oxysporum* strains

Evidence for an inherited cluster by HGT in two *F. oxysporum* strains was found. The cluster FOP5_C12 in *F. oxysporum* PHW815 consists of five genes including a NPS (FOP5_15561_g), a P450 gene (FOP5_15559_g), a methyltransferase (FOP5_15557_g), a gene with alcohol dehydrogenase and PKS enoylreductase domains (FOP5_15558_g) and a PKS gene (FOP5_15560_g) that is related to the lovastatin nonaketide synthase (ident 35.2%) of *Magnaporthe grisea* (Figure 5.2A). In the *F. oxysporum* strain HDV247 an orthologous cluster that lacks the NPS gene but contains a PKS-NPS hybrid instead can be found. Due to the draft gene models in the *F. oxysporum* strains it might be possible that the PKS and NPS genes in PHW815 are fused to one gene as well. The cluster is not present in any other analyzed *Fusarium* genome but orthologs can be found in *Colletotrichum graminicola*, *Colletotrichum orbiculare* and *Setosphaeria turcica* (Figure 5.2A). Like in *F. oxysporum* HDV247 four cluster genes are conserved in *C. graminicola.* In *C. orbiculare* and *S. turcica* only three genes of the cluster are present, whereas an additional P450 gene (SETTUDRAFT_161316) is contained in the cluster in *S. turcica* (Figure 5.2A).

The median GC ratio of the cluster genes in the *F. oxysporum* strains PHW815 and HDV247 amounts 61.9% and 62.8%, respectively. Compared to the distribution of all open reading frame GC ratios with median 51.3% in the both strains a significant difference was calculated using a two-sided KS test (P-value = 7.34e-4 (PHW815), P-value = 4.0e-3 (HDV247)) (Figure 5.2B). Compared to that no significant difference could be determined in the GC ratios of the three orthologous clusters and the GC distribution of ORFs in their respective genome. While the median GC ratio of the clusters in *C. graminicola* (60.0%), *C. orbiculare* (63.5%) and *F. oxysporum* PHW815 (62.8%) and HDV247 (62.3%) is similar, the cluster genes in *S. turcica* have a much lower median GC ratio (54.4%) (Figure 5.2B). Interestingly, in all species the cluster is embedded in regions with a high amount of predicted interspersed repeat elements and AT-rich regions.

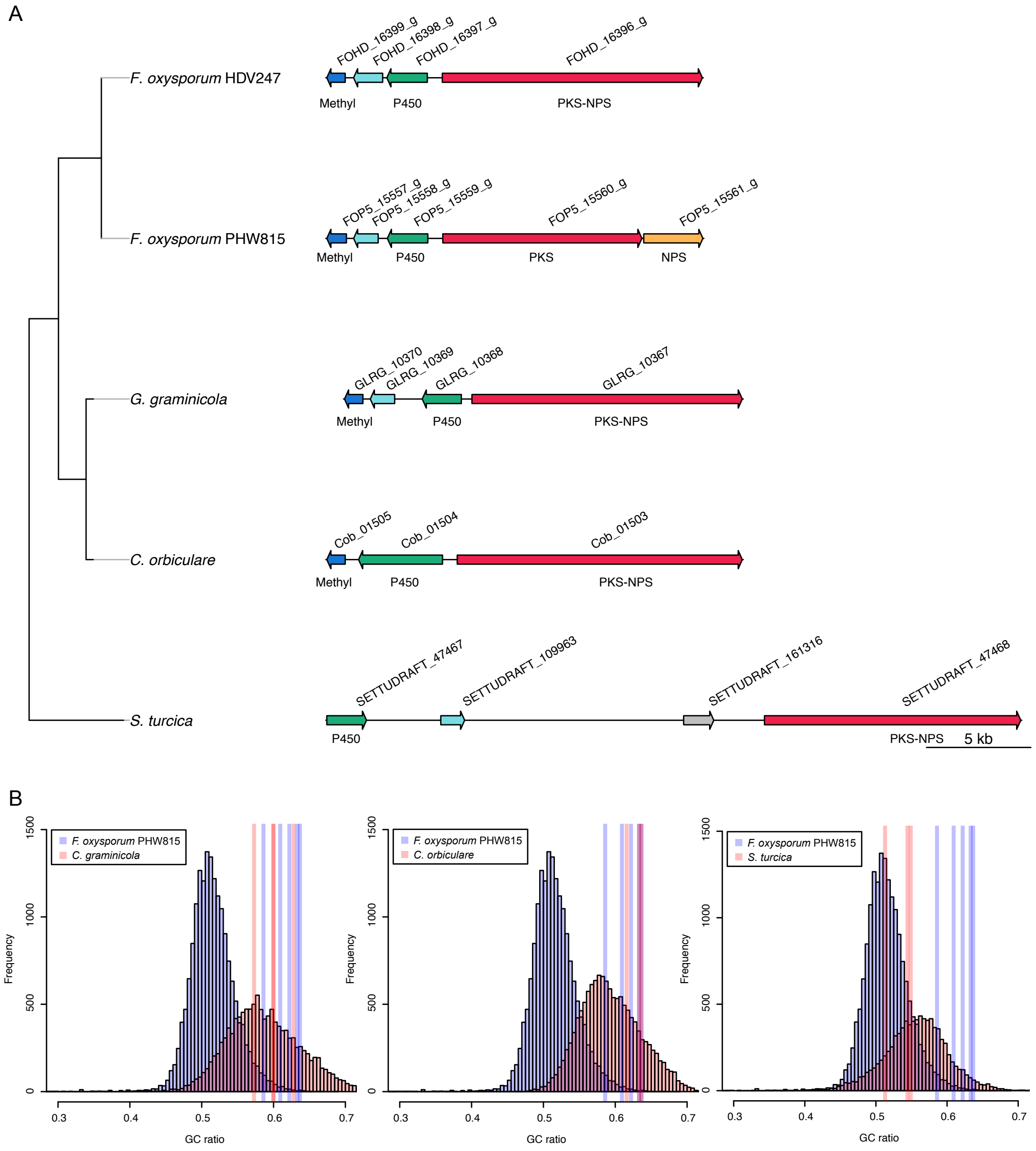

Figure 5.2: Evidence of horizontal gene transfer of a predicted secondary metabolism gene cluster in the *F. oxysporum* strains PHW815 and HDV247. **A:** Orthologous cluster genes and phylogenetic relationships of species containing the predicted cluster. Orthologs in *Colletotrichum graminicola*, *Colletotrichum orbiculare* and *Setosphaeria turcica* are depicted by colored arrows. Orthologous groups are illustrated by arrows of the same color. Grey arrows share no similarity with the predicted clusters in *F. oxysporum*. **B:** Histograms show whole genome distributions of open reading frame GC ratios in *F. oxysporum* PHW815 (blue) and species with orthologous cluster (red). Vertical lines illustrate GC ratios of cluster genes

5.2.5 An unknown NPS containing secondary metabolism gene cluster is conserved in *Cochliobolus heterostrophus* and *Pyrenophora teres*

In the peripheral region of chromosome I in *F. graminearum* (at 267 kb) the putative cluster FG_C62 (FGSG_10608 - FGSG_10617) consist of eleven genes including a NPS and two cytochrome P450 genes. The core part of the cluster (FGSG_10608 - FGSG_10614) shows a co-expression pattern and is not represented in other *Fusaria* by orthologs but is found in *Cochliobolus heterostrophus* and *Pyrenophora teres* (Figure 5.3A). The partially conserved cluster contains the two P450 genes and genes with FAD- and NAD(P)- binding domains but lacks the PKS-NPS enzyme, which is located in the distal part of the cluster. Also, a reverse transcriptase can be found exclusively in *C. heterostrophus* next to the cluster. In order to test for a potential HGT-event I calculated the median ORF GC ratio of *C. heterostrophus* which is slightly higher compared to *F. graminearum* (53.4% vs. 51.3%). The GC ratios of both clusters are in turn rather similar to each other (50.7% *C. heterostrophus*, 50.9% *F. graminearum*) and to the genome wide content of *F. graminearum*. However, in comparing the distributions of GC ratios of the cluster genes and the host genomes using a two-sided KS test I calculated a significant difference in *C. heterostrophus* (P-value = 0.002), but not in *P. teres* and *F. graminearum*. In taking a closer look at the gene expression during host infection, a significant increase in the expression intensity of the NPS at 40 hours post inoculation (hpi) while growing inside wheat coleoptiles is observable. As mentioned before, this gene belongs not to the co-expressed core part of the cluster, therefore no change in expression of the other cluster genes can be observed at this time point (3.85 fold on log_2-scale, P-value < 0.05). However, at 64 hpi the expression rate of the cluster genes that are conserved in *C. heterostrophus* and *P. teres* is significantly increased (1.1 to 3.5 fold on log_2-scale, P-value < 0.05) whereas the NPS is reduced [248] (Figure 5.3B). A similar observation can be made when looking at the gene expression during infection of wheat stems (1.9 to 4 fold on log_2-scale, P-value < 0.05) [80] as well as during crown rot disease of wheat (2.3 to 5.6 fold on log_2-scale, P-value < 0.05) [200]. Further, glutamine enriched medium causes a 4.6 to 7.8 fold (log_2-scale) increase in expression of the core cluster genes compared to the DON-inducing agmatine medium [75] (Figure 5.3B). The co-expression of the genes and the conservation of the cluster in *C. heterostrophus* suggest a functional but yet unknown gene cluster.

5.2.6 Ortholog analysis gives hints towards gene cluster evolution

Beside clusters that are conserved only in distantly related fungi, I also detected clusters that are unique in one species among the 381 analyzed genomes. In *F. verticillioides* a PKS containing cluster was found that could not be determined in any of the other compared species (Table A.2). The PKS15 signature enzyme as well as the seven cluster genes including three P450s have no orthologs in other *Fusarium* species. However, unclustered orthologs of single genes could be found in fungi outside the *Fusarium* phylum. In *Aspergillus terreus*, *A. niger* and *Rhynchosporium orthosporum* the PKS15 and three additional genes are conserved, but located on different contigs and chromosomes. *Neofusicoccum parvum* lacks the PKS15 signature enzyme but contains also four orthologs that are spread on different contigs. The adjacent genes around the unclustered orthologs show no similarity to the cluster or the neighboring genes in *F. verticillioides.* (Figure 5.5).

The cluster FG_C61 of *F. graminearum* consists of eight genes (FGSG_10542 - FGSG_17387) comprising a PKS, a NPS, a serine hydrolase, a transcription factor and four additional genes of unknown function. The genes are significantly repressed (Fold change < -2, P-value < 0.05) during C- and N- starving conditions (FG2) as well as in the FgStuA deletion mutant under secondary metabolism conditions (FG13). Six genes exhibit an increase in expression rate during wheat infection after 64 hours (FG19). Interestingly, *Aspergillus clavatus* is the only fungus where a bidirectional best hit of the PKS and the NPS could be determined. The signature enzymes seem to be part of one secondary metabolism gene cluster in *A. clavatus* as they are clustered with orthologs of the serine hydrolase gene and the ABC transporter, a unique transcription factor and a transporter in *A. clavatus* (Figure 5.4). However, no syntenic cluster can be found in any other fungal genome, although the signature enzymes alone are present in other species. The PKS for example was also found in *Aspergillus nidulans*, *A. niger*, *A. oryzae* and

A. tereus whereas the NPS is not present in these species. Protein similarity hypothesizes that the NPS gene is also conserved in the bacteria *Gordonia bronchialis* and *Bacillus amyloliquefaciens*.

5.2.7 The PKS19 cluster and genes of the PKS8 cluster are fused in some species

The polyketide synthase FFUJ_12239 (PKS19) is unique in *F. fujikuroi* and *F. proliferatum* in the GFC. In both organisms the *PKS19* belongs to a putative gene cluster consisting of six genes which are embedded in an AT-rich region on chromosome VIII. Interestingly the genes next to the AT-rich region have orthologs that are arranged in collinear order in other species of the GFC. Outside the GCF an ortholog of the complete PKS19 cluster can be found in *Chaetomium globosum* and three genes including the PKS gene are present in *Setosphaeria turcica* (arrows with warm colors in Figure 5.6).The predicted cluster of six genes including signature enzyme *PKS8* is located on chromosome XII in *F. fujikuroi*. Orthologs of this cluster can be found also in *F. proliferatum* on chromosome XII except the signature enzyme gene *PKS8* (arrows with cool colors in Figure 5.6). While the PKS19 and PKS8 clusters are clearly separated on two chromosomes in *F. fujikuroi* and *F. proliferatum*, a mixture of genes of both clusters can be observed in other species like *F. circinatum*. Here three genes including the alcohol dehydrogenase like enzyme, a transcription factor and a protein of unknown function, that are orthologs of the PKS19 cluster, are surrounded by four orthologs of the PKS8 cluster including the PKS enzyme. A similar observation can be made in *F. verticillioides*, *F. mangiferae*, and seven *F. oxysporum* strains (Figure 5.6).

Because of the sparse phylogenetic distribution of the cluster I am interested in finding evidence of horizontal gene transfer (HGT). Therefore I compared the GC ratios of the cluster genes to the distribution of GC ratios of the genomes. Although considerable differences in the mean GC ratios among the species *F. fujikuroi*, *F. proliferatum*, *Chaetomium globosum* and *Setosphaeria turcica*, no significant difference between the GC ratio of cluster genes and the respective genome could be determined.

Cluster	Position	Gene_Code	Description	Predicted_Motif
FG_C47	-2	FGSG_08211	conserved hypothetical protein	
	-1	FGSG_08210	conserved hypothetical protein	
	1	FGSG_08209	non-ribosomal peptide synthase	TAGGGACTTTGG
	2	FGSG_08208	polyketide synthase	TAGGGACTTTGG
	3	FGSG_08207	related to cytochrome P450 7B1	TAGGAACTATGG
	4	FGSG_08206	conserved hypothetical protein	TTGGGACTTTGG
	5	FGSG_17085	related to ornithine aminotransferase	TTGGGACTTTGG
	+1	FGSG_08204	conserved hypothetical protein	
	+2	FGSG_08203	conserved hypothetical protein	
FG_C61	-1	FGSG_17385	hypothetical protein	
	1	FGSG_10542	conserved hypothetical protein	
	2	FGSG_13782	putative protein	
	3	FGSG_10543	hypothetical protein	
	4	FGSG_17386	related to non-ribosomal peptide synthase	
	5	FGSG_10545	conserved hypothetical protein	
	6	FGSG_10546	hypothetical protein	
	7	FGSG_10547	related to multidrug resistance protein	
	8	FGSG_17387	probable type I polyketide synthase	
	+1	FGSG_10549	conserved hypothetical protein	
FG_C62	-2	FGSG_10606	probable cytochrome-c peroxidase precursor	
	-1	FGSG_10607	hypothetical protein	
	1	FGSG_10608	conserved hypothetical protein	
	2	FGSG_10609	related to 6-hydroxy-d-nicotine oxidase	
	3	FGSG_17400	related to cytochrome P450 monooxygenase	
	4	FGSG_17401	hypothetical protein	
	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	conserved hypothetical protein	
	9	FGSG_17402	probable beta-glucosidase precursor	
	10	FGSG_10616	related to vegetatible incompatibility protein HET-E-1	
	11	FGSG_10617	related to non-ribosomal peptide synthase MxcG	
	+1	FGSG_10618	hypothetical protein	

Table 5.2: Functional gene descriptions and positions of overrepresented promoter motifs on predicted clusters and neighboring genes. Orthologs of the predicted clusters are shown in Figures 5.1, 5.3 and 5.4.

5.2.8 Clusters with known metabolite and non-uniform phylogenetic distribution

Two clusters that could be linked to known metabolites show also hints of HGT. The genes of the metabolites aurofusarin and fusarielin are conserved in the closely related *F. pseudograminearum*, but cannot be found in other *Fusarium* species like the ones in the *Gibberella fujikuroi* species complex GFC. Ten to seven genes of the aurofusarin cluster can be found in other species outside the *Fusarium* phylum. For example, the genes from FGSG_02320 to FGSG_02329 are conserved in *Trichophyton tonsurans* but the orthologs of the PKS (FGSG_02324) and the adjacent gene of unknown function (FGSG_02325) are located on another scaffold as the rest of the cluster. *Arthroderma benhamie* and *Arthroderma gypseum* have a syntenic cluster of eight genes but totally lack orthologs of the PKS and the genes FGSG_02316 and FGSG_02321. In *A. gypseum* an ortholog of FGSG_02325 can be found on a different scaffold.

Furthermore, the fusarielin cluster and its orthologs in *Aspergillus fumigatus*, *A. niger* and *A. clavatus* were previously described [205] and detected by this. The closely related *F. pseudograminearum* has seven of the eleven cluster genes, comprising the PKS (FGSG_10464) and the putative NPS (FGSG_10459) but lacking the cytochrome P450 enzyme (FGSG_10461). The genes FGSG_10459 to FGSG_10464 are significantly up-regulated during wheat infection in young perithecia (2.3 to 3.8 fold change on log_2-scale, P-value <0.05) [80].

5.2.9 The gibberellic acid gene cluster duplicated and diverged in *F. proliferatum*

Orthologs analysis showed an interesting distribution of the gibberellic acid (GA) gene cluster in related fungi. While only two genes of the cluster can be found in *F. verticillioides*, the full cluster is contained in *F. fujikuroi*, *F. proliferatum*, *F. mangiferae* and *F. circinatum* (Table 4.3). In the available strains of *F. oxysporum* the cluster shows a non homologous distribution as the cluster can be found completely in four strains (*F. oxysporum* HDV247, PHW808, PHW815 and FOSC 3-a), while in the others only remnants are present. No orthologs can be found in the more distantly related *Fusaria F. graminearum* and *F. asiaticum*. However orthologs of four cluster genes could be detected in *Claviceps purpurea*. The genes are located in the peripheral part of a supercontig and grouped together with other genes related to secondary metabolism like a predicted PKS and a protein related to gibberellin 20-oxidase. Interestingly *F. proliferatum* contains a second copy of the cluster on a different supercontig, hinting towards a duplication event. While one cluster is located on chromosome V like in *F. fujikuroi* a second cluster could be found on a separate scaffold that could not be assigned to one of the eleven chromosomes. In order to determine whether the duplication is species specific for *F. proliferatum* or whether it also occurred in other gibberellin producing strains of *F. fujikuroi* a second strain of *F. proliferatum* (NRRL62812) and two additional strains of *F. fujikuroi* (B14, UCIM1100) were analyzed. However, in the additionally analyzed strains only the cluster on chromosome V was found.

In order to determine the phylogenetic relationship of all GA clusters I computed phylogenetic trees of the seven orthologous cluster genes. The resulting trees (Figure 5.7) show that the genes of the *F. oxysporum* and *F. fujikuroi* strains cluster together in terms of monophyletic groups. The same observation can be made for the two *F. proliferatum* strain clusters located on chromosome V, but the genes of the duplicated cluster on scaffold 20 are more similar to the orthologs in *F. circinatum*. Interestingly the distance between the two duplicated clusters in *F. proliferatum* (0.21, 0.23, 0.23, 0.16, 0.15, 0.17 and 0.19 average substitutions per site for *P450-3*, *CPS/KS*, *GGS2*, *P450-2*, *P450-1*, *P450-4* and *DES*, respectively) is bigger or equal compared to the distance of the *F. fujikuroi* group to the *F. proliferatum* cluster on chromosome V for all genes (0.17, 0.20, 0.17, 0.14, 0.15, 0.13 and 0.12 average substitutions per site, respectively). The distance is also bigger for all genes except the *DES* and *P450-2* between *F. fujikuroi* and the cluster on scaffold 20 (0.19, 0.21, 0.19, 0.17, 0.14, 0.15 and 0.20 average substitutions per site, respectively). The four orthologs in *C. purpurea* omit the biggest distance (Figure 5.7).

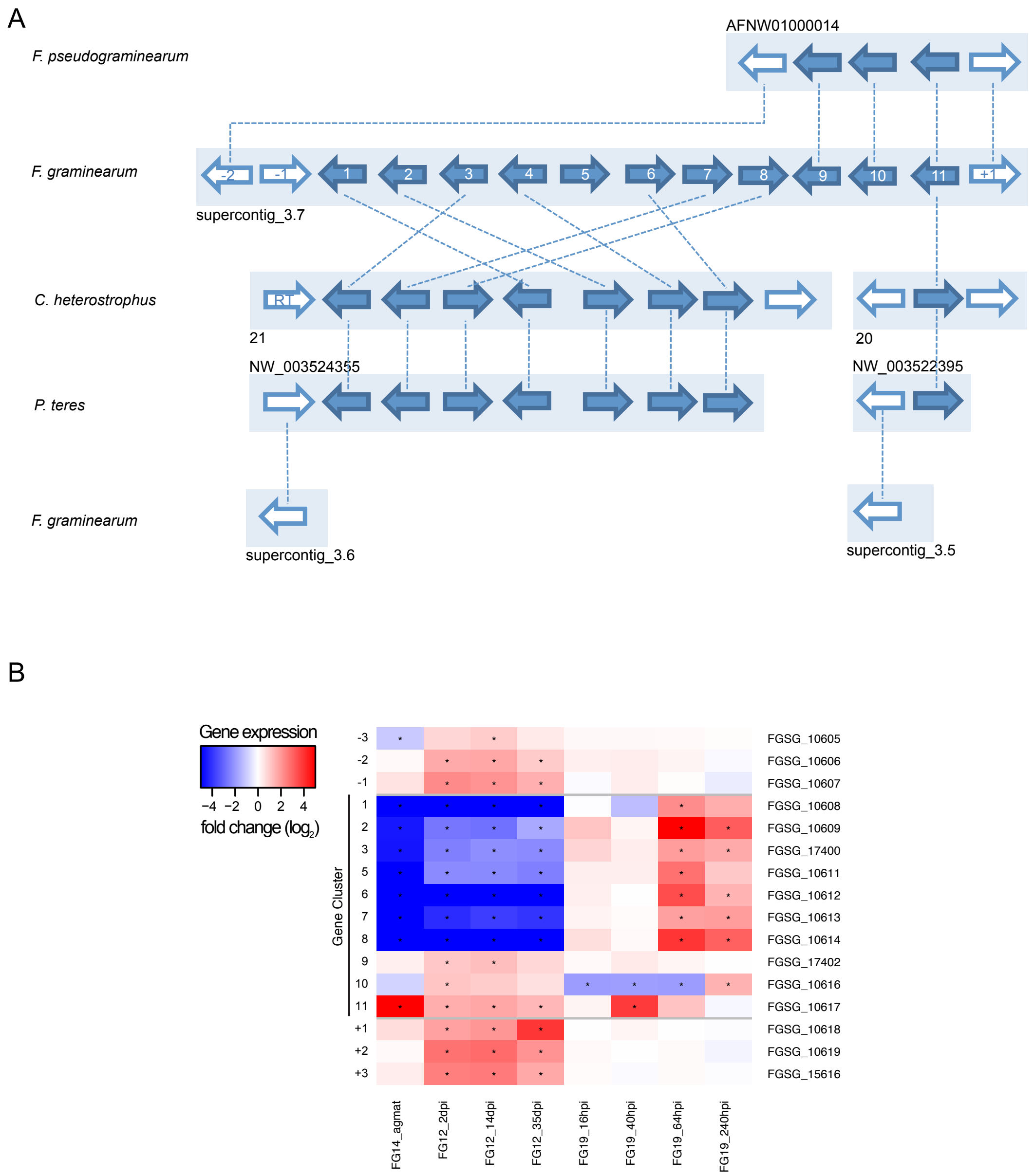

Figure 5.3: **A:** Predicted gene cluster (FG_C62) in *Fusarium graminearum* and orthologous genes in *F. pseudograminearum*, *Cochliobolus heterostrophus* and *Pyrenophora teres* (solid dark blue arrows) on their respective supercontigs (light blue boxes). Adjacent genes are shown in white, dashed lines between genes illustrate orthologous groups. Enumeration in *F. graminearum* is according to Table 5.2. Reverse transcriptase in *C. heterostrophus* is indicated as RT. **B:** Heatmap illustrates fold changes in gene expression (log_2 scale) of cluster and adjacent genes between experimental conditions. Genes are listed in chromosomal order on y-axis. Gene cluster is indicated by vertical black bar. Abbreviations of experimental conditions on x-axis are according to Table 2.2. No expression data is available for FGSG_10610, as a distinct mapping of probes on this gene model was not possible. Figure is adopted and modified from Sieber *et al.* [191].

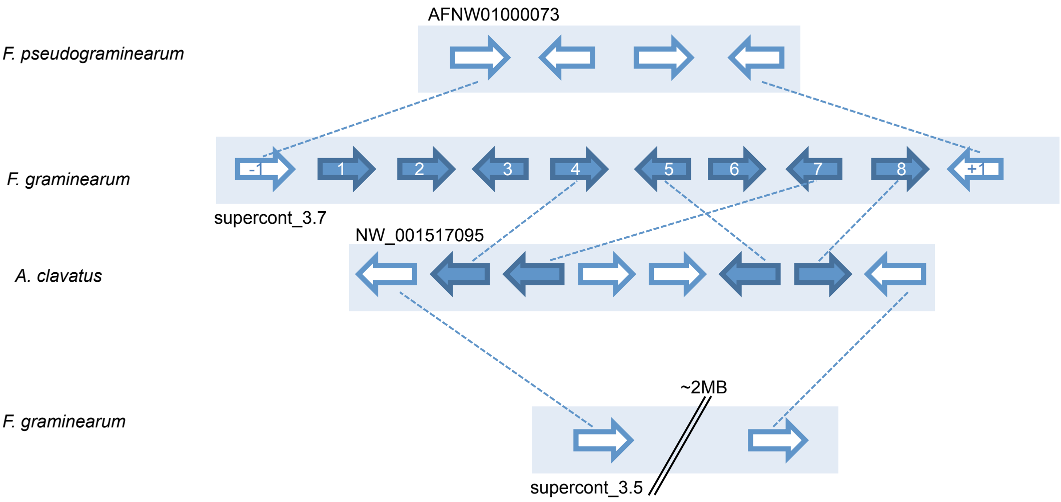

Figure 5.4: Predicted gene cluster FG_C61 in *Fusarium graminearum* and orthologous genes in *F. pseudograminearum* and *Aspergillus clavatus* are depicted in solid dark blue colors, adjacent genes are shown in white. Dashed lines illustrate orthologous groups. Enumeration in *F. graminearum* is according to Table 5.2. Figure is adopted from Sieber *et al.* [191].

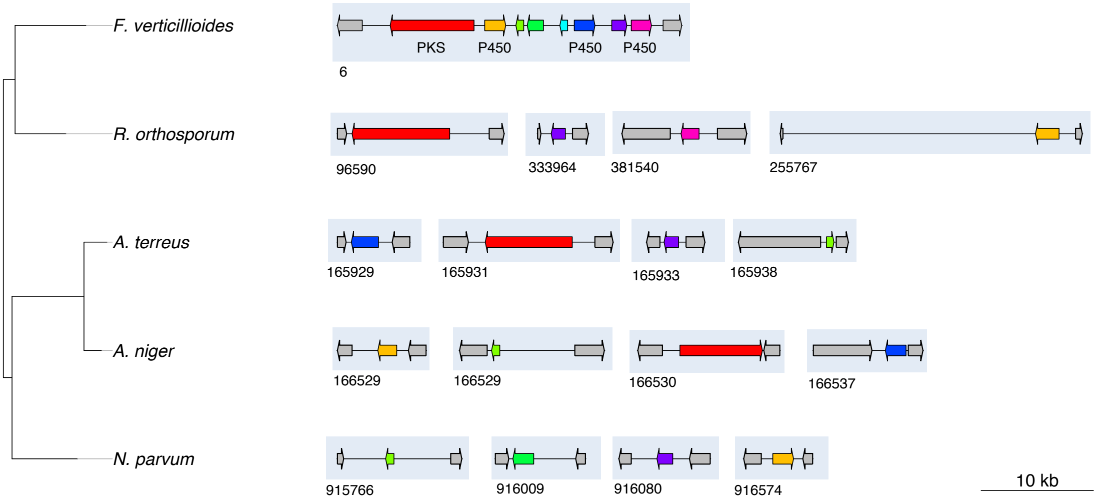

Figure 5.5: Unclustered orthologs of the PKS15 (FV_C31) gene cluster in *F. verticillioides*. Genes of the predicted cluster FV_C31 are indicated by colored arrows. Orthologs of cluster genes (arrows of same color) can be found in the distantly related species *Rhynchosporium orthosporum*, *Neofusicoccum parvum*, *Aspergillus terreus* and *A. niger* (indicated by phylogenetic tree on the left) but are located on different supercontigs, respectively (illustrated by boxes and supercontig-IDs below). Grey arrows indicate adjacent genes that share no similarity with the predicted cluster FV_C31.

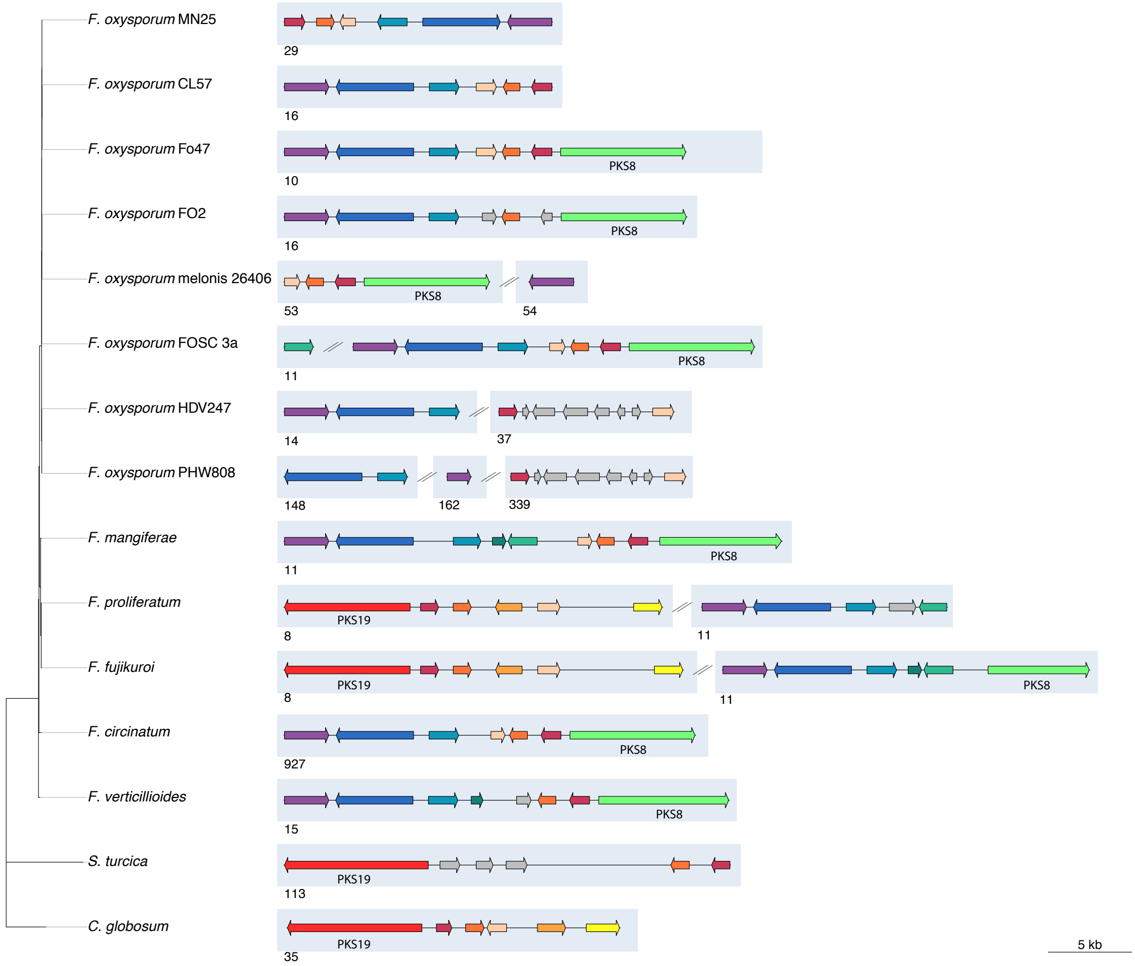

Figure 5.6: Fusion of the PKS19 and PKS8 gene clusters in some species. Genes of the predicted PKS19 (arrows in warm colors) and the PKS8 gene cluster (arrows in cold colors) are separated on two different chromosomes (blue boxes) in *Fusarium fujikuroi* and *F. proliferatum*. Orthologous genes (arrows of same color) of both clusters are part of one cluster in some species. Phylogenetic tree on the left indicates relationship between species.

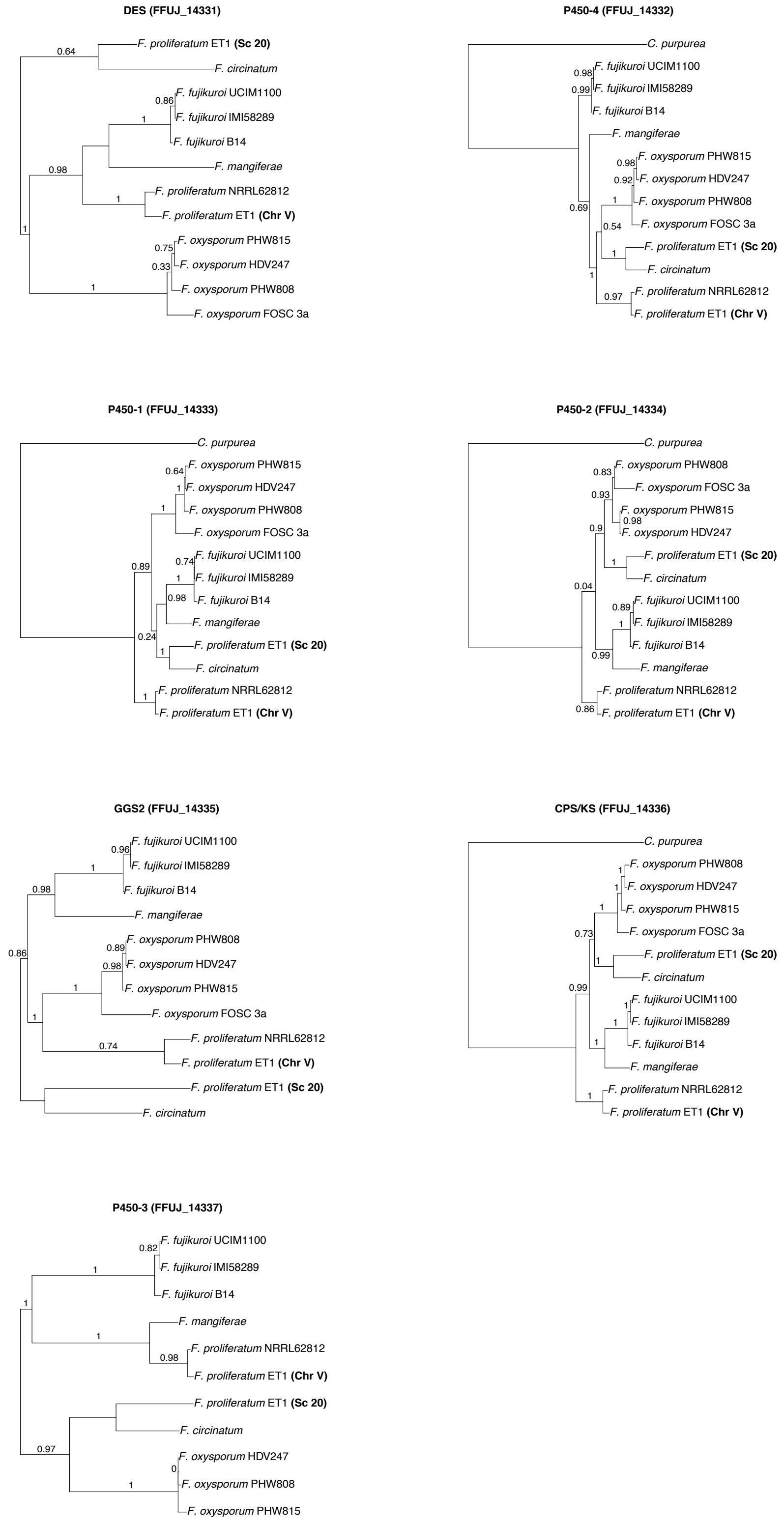

Figure 5.7: Phylogeny of gibberellic acid cluster genes. Maximum likelihood tree calculated based on the nucleotide sequences of the seven cluster genes *DES*, *P450-4*, *P450-1*, *P450-2*, *GGS2*, *CPS/KS* and *P450-3*. Approximate likelihood ratio test values are given at each branch.

5.3 Discussion

5.3.1 Evidence of horizontal gene cluster transfer

Horizontal gene transfer is an evolutionary process for microbes to collect new genetic material. Whereas the exchange between kingdoms including the interaction between fungi and their hosts is mostly limited to single genes [76], evidence of whole gene cluster transfers between fungi could be observed for example between *Fusarium* and *Aspergillus* [108] or *Botrytis* [37, 38].

The increasing number of available sequenced genomes enables a more and more accurate investigation in the phylogenetic distribution of secondary metabolism gene clusters and thus their evolutionary background. I used the genome sequences of 381 species in combination with the pre-calculated similarity network of their protein sequences to identify orthologs of predicted gene clusters of 22 *Fusarium* species. Beside *Fusarium* specific clusters, orthologs of several gene clusters could be found in distantly related species including gene clusters of known metabolites like aurofusarin and fusarielin. Comparison of GC ratio distributions of the genomes to the predicted pathway genes revealed evidence of 19 gene clusters that are putatively inherited by *Fusarium* species and 12 clusters with evidence of a *Fusarium* donor. The presented method also recognized previously reported gene transfers of the bikaverin [37, 38] and fumonisin [108] gene cluster.

While the cluster FG_C62 of *F. graminearum* has no orthologs in other examined *Fusaria* the cluster FG_C47 was only found in the closely related *F. asiaticum* where collinear orthologs of the cluster and neighboring genes are present. One explanation for this observation could be that the respective cluster was present in a common ancestor and due to mutations the genes got lost individually. However in the case of the *PKS23* containing cluster FG_C47, which can be found exclusively in *F. graminearum*, *F. asiaticum* and the *Botrytis fuckeliana* strains B05.01 and T4, I found evidence for horizontal gene inheritance between the three species (Figure 5.1A). The comparison of GC ratios of the orthologous clusters and the genomes supports the hypothesis that the cluster was transferred into the *Botrytis* lineage. In fact the GC ratios of both cluster orthologs are similar to the average ratio of *F. graminearum*, but differ significantly from the whole genome ORF GC ratio of *Botrytis* (Figure 5.1B).

Although the GC ratio of the clusters fits the average ORF GC ratio of *F. graminearum* and *F. asiaticum*, it is unlikely that the cluster originates from these organisms. There is no sequence identity between the neighboring genes of the cluster in *F. graminearum* and the genes adjacent to the *PKS23* gene in *F. pseudograminearum*, which is the only orthologous gene of the cluster in this species. Moreover, the orthologs of the neighboring genes of the cluster in *F. graminearum* constitute a collinear region on a different scaffold compared to *PKS23* in *F. pseudograminearum*. The clusters in *B. fuckeliana* B05.01 and T4 both contain an additional collinear P450 gene that does not exist in the *Fusaria*, but its GC ratio is considerably higher than the average of *Botrytis*. The same holds for the additional NPS-like gene, which is unique for the B05.01 strain. The results favor the hypothesis that the original cluster present in an unknown ancestor has at least seven genes, all present in *B. fuckeliana* B05.01, but retained only partially in T4 and *F. graminearum*. Because of the different cluster sizes in *F. graminearum*/*F. asiaticum* and *Botrytis*, the collinear flanking region in *F. pseudograminearum* and the difference in GC ratios, I assume that the donor organism is related to *Fusarium*.

The average GC ratios of the genomes *Cochliobolus heterostrophus*, *Pyrenophora teres* and *F. graminearum* are very similar. Therefore it is more difficult to determine hints of HGT between the species based on GC ratios of cluster orthologs. Significant differences in GC ratios of orthologs of the predicted NPS clusters FG_C62 and the host genomes could only be determined in *C. heterostrophus*, where also a reverse transcriptase could be found adjacent to the cluster (Figure 5.3). This evidence hints towards an insertion event of the genes.

In two *Fusarium oxysporum* strains strong evidence of an acquired gene cluster was found. The putatively inherited cluster is located in AT-rich genomic regions of the strains HDV247 and PHW815 but was not found in the other eleven examined *F. oxysporum* strains (Figure 5.2A). The considerably higher GC ratios of the putative biosynthetic genes suggest an origin outside the *Fusarium* phylum. As the mean genome wide GC ratio of all determined species with orthologous cluster fits the GC ratio of the clusters in *Fusarium* I assume that the origin of the cluster is located in the *Colletotrichum* or *Setosphaeria* lineage (Figure 5.2B). Like in *Fusarium* all orthologous clusters are embedded in repetitive AT-rich regions indicating a possible insertion into these genomes, as well.

5.3.2 Ortholog analysis gives hints towards evolution of gene clusters

In addition to horizontal gene transfer other evolutionary processes putatively responsible for the creation of novel secondary metabolism gene clusters could be identified by the applied orthologs analysis. I found evidence of enzyme shuffling, alternative tailoring and duplication and divergence of clusters.

Unique clusters suggest sources for an exclusive metabolite that might be beneficial to the lifestyle specific to the fungus. Questions about the evolutionary background and origin of these clusters rise. Possible scenarios comprise an individual loss in all sequenced genomes except the observed one or horizontal gene transfer from a not yet sequenced species. A third possibility however is that the proteins itself are present in other species but the clustering of the encoding genes is exclusive for one genome. In *F. verticilliodes* a unique cluster (FV_C31) with orthologous genes in four distantly related species that are all separated on different contigs was observed (Figure 5.5). A possible scenario would be that the predicted synthesis genes were already clustered in a common ancestor of all five species but due to genome reorganizations the physical linkage was lost in all species except *F. verticillioides.* This hypothesis suggests selection pressure on the predicted pathway genes that might be connected to the ecological niche of the maize pathogen.

In *F. graminearum* the cluster FG_C61 cannot be found in other sequenced fungi except *A. clavatus* where orthologs of four cluster genes, including the two signature enzymes and one neighboring gene also form a cluster (Figure 5.4). Other *Aspergilli* like *A. nidulans* or *A. tereus* contain a putative ortholog of the PKS, *Claviceps purpurea* and the bacterium *Bacillus amyloliquefaciens* contain an orthologous NPS. However, there is no other organism that contains both signature enzymes in terms of a bidirectional best hit, but *F. graminearum* and *A. clavatus.* It is likely that orthologs of the respective signature enzymes act in a different secondary metabolism pathway with different tailoring enzymes. The NPS ortholog in *B. amyloliquefaciens* for example, is part of the iturin A biosynthetic cluster [20] and the PKS in *A. tereus* seems to be part of a cluster with a second neighboring PKS gene. Mutations and genome reorganizations might be the driving force behind shuffling and deletion of pathway genes and creation of putatively novel metabolic products.

Likewise, an exchange of putative pathway genes was observed between the predicted PKS19 and PKS8 gene clusters that are separated on two different chromosomes in *F. fujikuroi* and *F. proliferatum* but are partially merged on one loci in other species of the GFC and *Fusarium oxysporum* species complex (FOC) (Figure 5.6). The *PKS19* gene in *F. fujikuroi* exhibited expression *in planta* and overexpression of the clustered transcription factor resulted in increased production of a yet unknown compound of partially elucidated structure [231]. It is plausible that orthologs of the transcription factor also plays a regulatory role of the putative pathways in the fused clusters for example in *F. mangiferae* or *F. circinatum.*

Beside alternative tailoring and shuffling of pathway genes, duplication and divergence of secondary metabolism genes is another evolutionary tool to generate novel SM pathways as previously reported in case of the gene family of polyketide synthases [115]. In a *F. proliferatum* strain a duplication of the whole gibberellic acid (GA) gene cluster was found. The phytohormone gibberellic acid is synthesized by the rice pathogen *F. fujikuroi* and main cause of the 'bakanae' disease [242]. Orthologous GA synthesis genes are distributed among most species of the GFC but also in more distantly related fungi like *Claviceps purpurea.* Phylogenetic analysis of the seven cluster genes failed in forming a phylogenetic group of the single copy genes in *F. proliferatum* strain NRRL62812 and the two clusters in *F. proliferatum* strain ET1 (Figure 5.7). All seven genes of the duplicated cluster in ET1 are more similar to the orthologs in *F. circinatum* than to the other orthologous clusters in the *F. proliferatum* strains. The high degree of sequence divergence of the duplicated cluster suggests an alteration of enzymatic functionality. However, no gibberellic acid producing *F. proliferatum* strain is known so far, raising the question of the role of the two orthologous clusters as non functional clusters usually are of limited evolutionary lifetime. As no gibberellic acid producing *F. proliferatum* strain is known so far questions of the role of the GA cluster genes in this species remains to be elucidated.

Chapter 6

Conclusion and outlook

In this thesis we analyzed important regulatory and evolutionary aspects of fungal secondary metabolism. By integrating multiple 'omics' data we discovered correlations between protein abundance, gene expression and cluster specific chromatin modifications, at which especially the activating H3K9ac mark is associated with regulation of gene expression in *Fusarium fujikuroi.* In addition, deletion mutants of the histone deacetylase *HDA1* and the global regulator *SGE1* influenced expression of secondary metabolism gene clusters. Taken together these results reveal a complex regulation of secondary metabolism on various levels in *F. fujikuroi.*

Based on these findings we predicted previously unknown putative secondary metabolism gene clusters and reconfirmed clusters of known metabolites based on intrinsic and extrinsic evidences. For the detection of gene cluster specific promoter motifs we combined multiple prediction algorithms in a bioinformatics pipeline. We included promoter sequences of orthologous genes and determined the genome wide occurrence of the motifs to filter significant candidate binding sites. In cooperation with experimental groups our predictions lead to the discovery of two new substances in *F. fujikuroi.* Furthermore, experimental verification by Niehaus *et al.* of a significantly overrepresented predicted promoter motif in the apicidin F cluster suggests regulatory importance. We showed that especially clusters with genes that are co-regulated and expressed under virulence inducing conditions provide targets for future experimental investigations. Moreover, these results demonstrate that regulatory binding sites can be identified from a multiplicity of predicted candidate motifs in taking orthologous promoters and the genome wide motif distribution into account. Significantly overrepresented motifs will be useful in identifying functional, but under standard laboratory condition silent, gene clusters.

Comparative analysis of transposons revealed clade specific transposable elements that might have contributed to genome evolution in *Fusarium*. Analysis of expression data hints towards active transposition of several TE families in *F. fujikuroi*. Additionally, by combining of expression data and an alignment based approach we discovered evidence of defense mechanisms that inactivate propagation of TEs. The observed transposon dynamics suggest impact of TEs on reorganization of genomes and transfer of genetic material between species.

The observed discontinuous phylogenetic distribution of secondary metabolism gene clusters among closely related *Fusarium* species could partially be explained by genome reorganizations and individual loss but also by horizontal gene transfer events. Horizontal transfer of whole secondary metabolism gene clusters may provide advantages in host infection and shape the evolution of fungi. In addition the observed relocation of cluster genes in genomes and duplication events of gene clusters contribute to the overall metabolic diversity in fungi. However, details on the mechanism of HGT and the force behind genomic clustering of secondary metabolism synthesis genes are still speculative. Sequencing initiatives like the 1000 fungal genomes project will enable a more detailed analysis of the evolution of gene clusters. Furthermore, metagenomics projects will provide the opportunity to estimate the amount of horizontal gene transfer of DNA in general and of gene clusters in particular among microbial communities in the great outdoors. This is favored by improvements in sequencing techniques and assembly algorithms which result in increasing read and contig lengths and lead to complete assembled genomes out of metagenomics data.

Applying these approaches will facilitate the discovery of new pathway genes and result in secondary metabolites with major impact on food and feed safety as well as the specification of new bioactive compounds for potential use in medical applications.

Bibliography

[1] Y. Abe, C. Ono, M. Hosobuchi, and H. Yoshikawa. Functional analysis of mlcr, a regulatory gene for ml-236b (compactin) biosynthesis in penicillium citrinum. *Mol Genet Genomics*, 268(3):352–361, Nov 2002.

[2] N. J. Alexander, S. P. McCormick, C. Waalwijk, T. van der Lee, and R. H. Proctor. The genetic basis for 3-adon and 15-adon trichothecene chemotypes in fusarium. *Fungal Genet Biol*, 48(5):485–495, May 2011.

[3] E. C. Alfonso, J. Cantu-Dibildox, W. M. Munir, D. Miller, T. P. O'Brien, C. L. Karp, S. H. Yoo, R. K. Forster, W. W. Culbertson, K. Donaldson, J. Rodila, and Y. Lee. Insurgence of fusarium keratitis associated with contact lens wear. *Arch Ophthalmol*, 124(7):941–947, Jul 2006.

[4] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, Oct 1990.

[5] C. L. Alvarez, S. Somma, R. H. Proctor, G. Stea, G. Mulè, A. F. Logrieco, V. F. Pinto, and A. Moretti. Genetic diversity in fusarium graminearum from a major wheat-producing region of argentina. *Toxins (Basel)*, 3(10):1294–1309, Oct 2011.

[6] T. Aoki, F. Tanaka, H. Suga, M. Hyakumachi, M. M. Scandiani, and K. O'Donnell. Fusarium azukicola sp. nov., an exotic azuki bean root-rot pathogen in hokkaido, japan. *Mycologia*, 104(5):1068–1084, 2012.

[7] R. Arnold, F. Goldenberg, H.-W. Mewes, and T. Rattei. Simap–the database of all-against-all protein sequence similarities and annotations with new interfaces and increased coverage. *Nucleic Acids Res*, 42(Database issue):D279–D284, Jan 2014.

[8] T. Bailey, N. Williams, C. Misleh, and W. Li. Meme: discovering and analyzing dna and protein sequence motifs. *Nucleic acids research*, 34(Web Server issue):W369–373, 2006.

[9] T. L. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, 2:28–36, 1994.

[10] S. Bandelier, R. Renaud, and A. Durand. Production of gibberellic acid by fed-batch solid state fermentation in an aseptic pilot-scale reactor. *Process Biochemistry*, 32(2):141–145, 1997.

[11] A. J. Bannister and T. Kouzarides. Regulation of chromatin by histone modifications. *Cell Res*, 21(3):381–395, Mar 2011.

[12] B. M. Barker, K. Kroll, M. Vödisch, A. Mazurie, O. Kniemeyer, and R. A. Cramer. Transcriptomic and proteomic analyses of the aspergillus fumigatus hypoxia response using an oxygen-controlled fermenter. *BMC Genomics*, 13:62, 2012.

[13] D. Barreto, S. Babbitt, M. Gally, and B. Pérez. Nectria haematococca causing root rot in olive greenhouse plants. *RIA INTA*, 32(1):49–55, 2003.

[14] S. Bauer, P. N. Robinson, and J. Gagneur. Model-based gene set analysis for bioconductor. *Bioinformatics*, 27(13):1882–1883, Jul 2011.

[15] J. Behr, R. Bohnert, G. Zeller, G. Schweikert, L. Hartmann, and G. Rätsch. Next generation genome annotation with mgene. ngs. *BMC Bioinformatics*, 11(Suppl 10):O8, 2010.

[16] J. Bendtsen, L. Jensen, N. Blom, G. Von Heijne, and S. Brunak. Feature-based prediction of non-classical and leaderless protein secretion. *Protein Engineering Design and Selection*, 17(4):349, 2004.

[17] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[18] G. Benson. Tandem repeats finder: a program to analyze dna sequences. *Nucleic acids research*, 27(2):573, 1999.

[19] K. Blin, M. H. Medema, D. Kazempour, M. A. Fischbach, R. Breitling, E. Takano, and T. Weber. antismash 2.0–a versatile platform for genome mining of secondary metabolite producers. *Nucleic Acids Res*, 41(Web Server issue):W204–W212, Jul 2013.

[20] J. Blom, C. Rueckert, B. Niu, Q. Wang, and R. Borriss. The complete genome of bacillus amyloliquefaciens subsp. plantarum cau b946 contains a

gene cluster for nonribosomal synthesis of iturin a. *J Bacteriol*, 194(7):1845–1846, Apr 2012.

[21] J. W. Bok, Y.-M. Chiang, E. Szewczyk, Y. Reyes-Dominguez, A. D. Davidson, J. F. Sanchez, H.-C. Lo, K. Watanabe, J. Strauss, B. R. Oakley, C. C. C. Wang, and N. P. Keller. Chromatin-level regulation of biosynthetic gene clusters. *Nat Chem Biol*, 5(7):462–464, Jul 2009.

[22] C. Bömke and B. Tudzynski. Diversity, regulation, and evolution of the gibberellin biosynthetic pathway in fungi compared to plants and bacteria. *Phytochemistry*, 70(15-16):1876–1893, 2009.

[23] C. E. Bonferroni. Il calcolo delle assicurazioni su gruppi di teste. In *Studi in Onore del Professore Salvatore Ortu Carboni*, pages 13–60. Rome, 1935.

[24] C. E. Bonferroni. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62, 1936.

[25] E. I. Boutati and E. J. Anaissie. Fusarium, a significant emerging pathogen in patients with hematologic malignancy: ten years' experience at a cancer center and implications for management. *Blood*, 90(3):999–1008, Aug 1997.

[26] A.-L. Boutigny, T. J. Ward, G. J. Van Coller, B. Flett, S. C. Lamprecht, K. O'Donnell, and A. Viljoen. Analysis of the fusarium graminearum species complex from wheat, barley and maize in south africa provides evidence of species-specific differences in host preference. *Fungal Genet Biol*, 48(9):914–920, Sep 2011.

[27] N. J. Bowen and I. K. Jordan. Transposable elements and the evolution of eukaryotic complexity. *Curr Issues Mol Biol*, 4(3):65–76, Jul 2002.

[28] A. Brakhage and V. Schroeckh. Fungal secondary metabolites–strategies to activate silent gene clusters. *Fungal Genetics and Biology*, 48(1):15–22, 2011.

[29] N. L. Brock, K. Huss, B. Tudzynski, and J. S. Dickschat. Genetic dissection of sesquiterpene biosynthesis by fusarium fujikuroi. *Chembiochem*, 14(3):311–315, Feb 2013.

[30] D. W. Brown, R. A. E. Butchko, S. E. Baker, and R. H. Proctor. Phylogenomic and functional domain analysis of polyketide synthases in fusarium. *Fungal Biol*, 116(2):318–331, Feb 2012.

[31] D. W. Brown, R. A. E. Butchko, M. Busman, and R. H. Proctor. Identification of gene clusters associated with fusaric acid, fusarin, and perithecial pigment production in fusarium verticillioides. *Fungal Genet Biol*, 49(7):521–532, Jul 2012.

[32] D. W. Brown, R. B. Dyer, S. P. McCormick, D. F. Kendra, and R. D. Plattner. Functional demarcation of the fusarium core trichothecene gene cluster. *Fungal Genet Biol*, 41(4):454–462, Apr 2004.

[33] N. A. Brown, J. Antoniw, and K. E. Hammond-Kosack. The predicted secretome of the plant pathogenic fungus fusarium graminearum: a refined comparative analysis. *PLoS One*, 7(4):e33731, 2012.

[34] C. Buades and A. Moya. Phylogenetic analysis of the isopenicillin-n-synthetase horizontal gene transfer. *J Mol Evol*, 42(5):537–542, May 1996.

[35] E. B. Cambareri, B. C. Jensen, E. Schabtach, and E. U. Selker. Repeat-induced g-c to a-t mutations in neurospora. *Science*, 244(4912):1571–1575, Jun 1989.

[36] E. B. Cambareri, M. J. Singer, and E. U. Selker. Recurrence of repeat-induced point mutation (rip) in neurospora crassa. *Genetics*, 127(4):699–710, Apr 1991.

[37] M. A. Campbell, A. Rokas, and J. C. Slot. Horizontal transfer and death of a fungal secondary metabolic gene cluster. *Genome Biol Evol*, Jan 2012.

[38] M. A. Campbell, M. Staats, J. Van Kan, A. Rokas, and J. C. Slot. Repeated loss of an anciently horizontally transferred gene cluster in botrytis. *Mycologia*, Aug 2013.

[39] D. E. Cane and C. T. Walsh. The parallel and convergent universes of polyketide synthases and nonribosomal peptide synthetases. *Chem Biol*, 6(12):R319–R325, Dec 1999.

[40] A. Cao, R. Santiago, A. J. Ramos, X. C. Souto, O. Aguín, R. A. Malvar, and A. Butrón. Critical environmental and genotypic factors for fusarium verticillioides infection, fungal growth and fumonisin contamination in maize grown in northwestern spain. *Int J Food Microbiol*, 177:63–71, May 2014.

[41] D. C. Chang, G. B. Grant, K. O'Donnell, K. A. Wannemuehler, J. Noble-Wang, C. Y. Rao, L. M. Jacobson, C. S. Crowell, R. S. Sneed, F. M. T. Lewis, J. K. Schaffzin, M. A. Kainer, C. A. Genese, E. C. Alfonso, D. B. Jones, A. Srinivasan, S. K. Fridkin, B. J. Park, and F. K. I. T. . Multistate

outbreak of fusarium keratitis associated with use of a contact lens solution. *JAMA*, 296(8):953–963, Aug 2006.

[42] P. K. Chang, K. C. Ehrlich, J. Yu, D. Bhatnagar, and T. E. Cleveland. Increased expression of aspergillus parasiticus aflr, encoding a sequence-specific dna-binding protein, relieves nitrate inhibition of aflatoxin biosynthesis. *Appl Environ Microbiol*, 61(6):2372–2377, Jun 1995.

[43] A. J. Clutterbuck. Genomic evidence of repeat-induced point mutation (rip) in filamentous ascomycetes. *Fungal Genet Biol*, 48(3):306–326, Mar 2011.

[44] J. J. Coleman, S. D. Rounsley, M. Rodriguez-Carres, A. Kuo, C. C. Wasmann, J. Grimwood, J. Schmutz, M. Taga, G. J. White, S. Zhou, D. C. Schwartz, M. Freitag, L.-J. Ma, E. G. J. Danchin, B. Henrissat, P. M. Coutinho, D. R. Nelson, D. Straney, C. A. Napoli, B. M. Barker, M. Gribskov, M. Rep, S. Kroken, I. Molnár, C. Rensing, J. C. Kennell, J. Zamora, M. L. Farman, E. U. Selker, A. Salamov, H. Shapiro, J. Pangilinan, E. Lindquist, C. Lamers, I. V. Grigoriev, D. M. Geiser, S. F. Covert, E. Temporini, and H. D. Vanetten. The genome of nectria haematococca: contribution of supernumerary chromosomes to gene expansion. *PLoS Genet*, 5(8):e1000618, Aug 2009.

[45] A. Conesa, P. J. Punt, N. van Luijk, and C. A. van den Hondel. The secretion pathway in filamentous fungi: a biotechnological view. *Fungal Genet Biol*, 33(3):155–171, Aug 2001.

[46] L. R. Connolly, K. M. Smith, and M. Freitag. The fusarium graminearum histone h3 k27 methyltransferase kmt6 regulates development and expression of secondary metabolite gene clusters. *PLoS Genet*, 9(10):e1003916, Oct 2013.

[47] C. Cuomo, U. Güldener, J. Xu, F. Trail, B. Turgeon, A. Di Pietro, J. Walton, L. Ma, S. Baker, M. Rep, et al. The fusarium graminearum genome reveals a link between localized polymorphism and pathogen specialization. *Science*, 317(5843):1400, 2007.

[48] A. Darling, B. Mau, F. Blattner, and N. Perna. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome research*, 14(7):1394, 2004.

[49] S. Dash, J. Van Hemert, L. Hong, R. P. Wise, and J. A. Dickerson. Plexdb: gene expression resources for plants and plant pathogens. *Nucleic Acids Res*, 40(Database issue):D1194–D1201, Jan 2012.

[50] R. Dawkins and J. R. Krebs. Arms races between and within species. *Proc R Soc Lond B Biol Sci*, 205(1161):489–511, Sep 1979.

[51] A. Delcher, A. Phillippy, J. Carlton, and S. Salzberg. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Research*, 30(11):2478–2483, 2002.

[52] A. L. Demain. Pharmaceutically active secondary metabolites of microorganisms. *Appl Microbiol Biotechnol*, 52(4):455–463, Oct 1999.

[53] A. E. Desjardins, T. M. Hohn, and S. P. McCormick. Trichothecene biosynthesis in fusarium species: chemistry, genetics, and significance. *Microbiol Rev*, 57(3):595–604, Sep 1993.

[54] A. Diolez, F. Marches, D. Fortini, and Y. Brygoo. Boty, a long-terminal-repeat retroelement in the phytopathogenic fungus botrytis cinerea. *Appl Environ Microbiol*, 61(1):103–108, Jan 1995.

[55] H. H. Divon, B. Rothan-Denoyes, O. Davydov, A. DI Pietro, and R. Fluhr. Nitrogen-responsive genes are differentially regulated in planta during fusarium oxyspsorum f. sp. lycopersici infection. *Mol Plant Pathol*, 6(4):459–470, Jul 2005.

[56] P. N. Dodds, G. J. Lawrence, A.-M. Catanzariti, T. Teh, C.-I. A. Wang, M. A. Ayliffe, B. Kobe, and J. G. Ellis. Direct protein interaction underlies gene-for-gene specificity and coevolution of the flax resistance genes and flax rust avirulence genes. *Proc Natl Acad Sci U S A*, 103(23):8888–8893, Jun 2006.

[57] M. C. V. Donaton, I. Holsbeeks, O. Lagatie, G. Van Zeebroeck, M. Crauwels, J. Winderickx, and J. M. Thevelein. The gap1 general amino acid permease acts as an amino acid sensor for activation of protein kinase a targets in the yeast saccharomyces cerevisiae. *Mol Microbiol*, 50(3):911–929, Nov 2003.

[58] N. M. Donofrio, Y. Oh, R. Lundy, H. Pan, D. E. Brown, J. S. Jeong, S. Coughlan, T. K. Mitchell, and R. A. Dean. Global gene expression during nitrogen starvation in the rice blast fungus, magnaporthe grisea. *Fungal Genet Biol*, 43(9):605–617, Sep 2006.

[59] J. E. Dvorska, P. F. Surai, B. K. Speake, and N. H. Sparks. Effect of the mycotoxin aurofusarin on the antioxidant composition and fatty acid profile of quail eggs. *Br Poult Sci*, 42(5):643–649, Dec 2001.

[60] K. C. Ehrlich, B. G. Montalbano, and J. W. Cary. Binding of the c6-zinc cluster protein, aflr, to the promoters of aflatoxin pathway biosynthesis genes in aspergillus parasiticus. *Gene*, 230(2):249–257, Apr 1999.

[61] T. Emery. Malonichrome, a new iron chelate from fusarium roseum. *Biochim Biophys Acta*, 629(2):382–390, May 1980.

[62] E. A. Espeso, J. Tilburn, H. Arst, Jr, and M. A. Peñalva. ph regulation is a major determinant in expression of a fungal penicillin biosynthetic gene. *EMBO J*, 12(10):3947–3956, Oct 1993.

[63] M. Fernandes, N. P. Keller, and T. H. Adams. Sequence-specific binding by aspergillus nidulans aflr, a c6 zinc cluster protein regulating mycotoxin biosynthesis. *Mol Microbiol*, 28(6):1355–1365, Jun 1998.

[64] C. Feschotte and E. Pritham. Dna transposons and the evolution of eukaryotic genomes. *Annual review of genetics*, 41:331, 2007.

[65] R. A. Fisher. On the interpretation of χ 2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922.

[66] P. M. Flatt and T. Mahmud. Biosynthesis of aminocyclitol-aminoglycoside antibiotics and related compounds. *Nat Prod Rep*, 24(2):358–392, Apr 2007.

[67] H. Forsberg and P. O. Ljungdahl. Sensors of extracellular nutrients in saccharomyces cerevisiae. *Curr Genet*, 40(2):91–109, Sep 2001.

[68] S. Freeman, M. Maimon, and Y. Pinkas. Use of gus transformants of fusarium subglutinans for determining etiology of mango malformation disease. *Phytopathology*, 89(6):456–461, Jun 1999.

[69] A. Gacek and J. Strauss. The chromatin code of fungal secondary metabolite gene clusters. *Appl Microbiol Biotechnol*, 95(6):1389–1404, Sep 2012.

[70] I. Gaffoor, D. W. Brown, R. Plattner, R. H. Proctor, W. Qi, and F. Trail. Functional analysis of the polyketide synthase genes in the filamentous fungus gibberella zeae (anamorph fusarium graminearum). *Eukaryot Cell*, 4(11):1926–1933, Nov 2005.

[71] I. Gaffoor and F. Trail. Characterization of two polyketide synthase genes involved in zearalenone biosynthesis in gibberella zeae. *Appl Environ Microbiol*, 72(3):1793–1799, Mar 2006.

[72] J. E. Galagan and E. U. Selker. Rip: the evolutionary cost of genome defense. *Trends Genet*, 20(9):417–423, Sep 2004.

[73] L. R. Gale, S. A. Harrison, T. J. Ward, K. O'Donnell, E. A. Milus, S. W. Gale, and H. C. Kistler. Nivalenol-type populations of fusarium graminearum and f. asiaticum are prevalent on wheat in southern louisiana. *Phytopathology*, 101(1):124–134, Jan 2011.

[74] E. Gamliel-Atinsky, A. Sztejnberg, M. Maymon, H. Vintal, D. Shtienberg, and S. Freeman. Infection dynamics of fusarium mangiferae, causal agent of mango malformation disease. *Phytopathology*, 99(6):775–781, Jun 2009.

[75] D. M. Gardiner, K. Kazan, and J. M. Manners. Novel genes of fusarium graminearum that negatively regulate deoxynivalenol production and virulence. *Mol Plant Microbe Interact*, 22(12):1588–1600, Dec 2009.

[76] D. M. Gardiner, K. Kazan, and J. M. Manners. Cross-kingdom gene transfer facilitates the evolution of virulence in fungal pathogens. *Plant Sci*, 210:151–158, Sep 2013.

[77] D. M. Gardiner, M. C. McDonald, L. Covarelli, P. S. Solomon, A. G. Rusu, M. Marshall, K. Kazan, S. Chakraborty, B. A. McDonald, and J. M. Manners. Comparative pathogenomics reveals horizontally acquired novel virulence genes in fungi infecting cereal hosts. *PLoS Pathog*, 8(9):e1002952, Sep 2012.

[78] L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry. affy—analysis of affymetrix genechip data at the probe level. *Bioinformatics*, 20(3):307–315, 2004.

[79] P. A. Grant. A tale of histone modifications. *Genome Biol*, 2(4):REVIEWS0003, 2001.

[80] J. C. Guenther, H. E. Hallen-Adams, H. Bcking, Y. Shachar-Hill, and F. Trail. Triacylglyceride metabolism by fusarium graminearum during colonization and sexual development on wheat. *Mol Plant Microbe Interact*, 22(12):1492–1503, Dec 2009.

[81] S. Guindon, J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of phyml 3.0. *Syst Biol*, 59(3):307–321, May 2010.

[82] U. Güldener, K.-Y. Seong, J. Boddu, S. Cho, F. Trail, J.-R. Xu, G. Adam, H.-W. Mewes, G. J. Muehlbauer, and H. C. Kistler. Development of a fusarium graminearum affymetrix genechip for profiling fungal gene expression in vitro and in planta. *Fungal Genet Biol*, 43(5):316–325, May 2006.

[83] L. Guy, J. R. Kultima, and S. G. E. Andersson. genoplotr: comparative gene and genome visualization in r. *Bioinformatics*, 26(18):2334–2335, Sep 2010.

[84] Y. Han, X. Liu, U. Benny, H. C. Kistler, and H. D. VanEtten. Genes determining pathogenicity to pea are clustered on a supernumerary chromosome in the fungal plant pathogen nectria haematococca. *Plant J*, 25(3):305–314, Feb 2001.

[85] J. K. Hane and R. P. Oliver. Ripcal: a tool for alignment-based analysis of repeat-induced point mutations in fungal genomic sequences. *BMC Bioinformatics*, 9:478, 2008.

[86] F. T. Hansen, J. L. Srensen, H. Giese, T. E. Sondergaard, and R. J. N. Frandsen. Quick guide to polyketide synthase and nonribosomal synthetase genes in fusarium. *Int J Food Microbiol*, 155(3):128–136, Apr 2012.

[87] L. J. Harris, N. J. Alexander, A. Saparno, B. Blackwell, S. P. McCormick, A. E. Desjardins, L. S. Robert, N. Tinker, J. Hattori, C. Pich, J. P. Schernthaner, R. Watson, and T. Ouellet. A novel gene cluster in fusarium graminearum contains a gene that contributes to butenolide synthesis. *Fungal Genet Biol*, 44(4):293–306, Apr 2007.

[88] S. A. Harrow, R. Farrokhi-Nejad, A. R. Pitman, I. A. W. Scott, A. Bentley, C. Hide, and M. G. Cromey. Characterisation of new zealand fusarium populations using a polyphasic approach differentiates the f. avenaceum/f. acuminatum/f. tricinctum species complex in cereal and grassland systems. *Fungal Biol*, 114(4):293–311, Apr 2010.

[89] M. Hasegawa, H. Kishino, and T. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *J Mol Evol*, 22(2):160–174, 1985.

[90] D. Hebenstreit, M. Gu, S. Haider, D. J. Turner, P. Li, and S. A. Teichmann. Epichip: gene-by-gene quantification of epigenetic modification levels. *Nucleic Acids Res*, 39(5):e27, Mar 2011.

[91] M. T. Hedayati, A. C. Pasqualotto, P. A. Warn, P. Bowyer, and D. W. Denning. Aspergillus flavus: human pathogen, allergen and mycotoxin producer. *Microbiology*, 153(Pt 6):1677–1692, Jun 2007.

[92] P. Hedden, A. L. Phillips, M. C. Rojas, E. Carrera, and B. Tudzynski. Gibberellin biosynthesis in plants and fungi: A case of convergent evolution? *J Plant Growth Regul*, 20(4):319–331, Dec 2001.

[93] T. M. Hohn, R. Krishna, and R. H. Proctor. Characterization of a transcriptional activator controlling trichothecene toxin biosynthesis. *Fungal Genet Biol*, 26(3):224–235, Apr 1999.

[94] I. Holsbeeks, O. Lagatie, A. Van Nuland, S. Van de Velde, and J. M. Thevelein. The eukaryotic plasma membrane as a nutrient-sensing device. *Trends Biochem Sci*, 29(10):556–564, Oct 2004.

[95] P. Horton, K.-J. Park, T. Obayashi, N. Fujita, H. Harada, C. J. Adams-Collier, and K. Nakai. Wolf psort: protein localization predictor. *Nucleic Acids Res*, 35(Web Server issue):W585–W587, Jul 2007.

[96] P. M. Houterman, D. Speijer, H. L. Dekker, C. G. DE Koster, B. J. C. Cornelissen, and M. Rep. The mixed xylem sap proteome of fusarium oxysporum-infected tomato plants. *Mol Plant Pathol*, 8(2):215–221, Mar 2007.

[97] M. J. Hynes. Studies on the role of the area gene in the regulation of nitrogen catabolism in aspergillus nidulans. *Aust J Biol Sci*, 28(3):301–313, Jun 1975.

[98] A. Jacobs, P. S. Van Wyk, W. F. O. Marasas, B. D. Wingfield, M. J. Wingfield, and T. A. Coutinho. Fusarium ananatum sp. nov. in the gibberella fujikuroi species complex from pineapples with fruit rot in south africa. *Fungal Biol*, 114(7):515–527, Jul 2010.

[99] J. Jin, S. Lee, J. Lee, S. Baek, J. Kim, S. Yun, S. Park, S. Kang, and Y. Lee. Functional characterization and manipulation of the apicidin biosynthetic pathway in fusarium semitectum. *Molecular microbiology*, 76(2):456–466, 2010.

[100] J.-M. Jin, J. Lee, and Y.-W. Lee. Characterization of carotenoid biosynthetic genes in the ascomycete gibberella zeae. *FEMS Microbiol Lett*, 302(2):197–202, Jan 2010.

[101] W. Jonkers, Y. Dong, K. Broz, and H. C. Kistler. The wor1-like protein fgp1 regulates pathogenicity, toxin synthesis and reproduction in the phytopathogenic fungus fusarium graminearum. *PLoS Pathog*, 8(5):e1002724, 2012.

[102] J. Jurka, V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz. Repbase update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research*, 110(1-4):462–467, 2005.

[103] K. Katoh, G. Asimenos, and H. Toh. Multiple alignment of dna sequences with mafft. *Methods Mol Biol*, 537:39–64, 2009.

[104] Y. Katsuyama and Y. Ohnishi. Type iii polyketide synthases in microorganisms. *Methods Enzymol*, 515:359–377, 2012.

[105] N. P. Keller and T. M. Hohn. Metabolic pathway gene clusters in filamentous fungi. *Fungal Genet Biol*, 21(1):17–29, Feb 1997.

[106] N. P. Keller, G. Turner, and J. W. Bennett. Fungal secondary metabolism - from biochemistry to genomics. *Nat Rev Microbiol*, 3(12):937–947, Dec 2005.

[107] N. Khaldi, F. T. Seifuddin, G. Turner, D. Haft, W. C. Nierman, K. H. Wolfe, and N. D. Fedorova. Smurf: Genomic mapping of fungal secondary metabolite clusters. *Fungal Genet Biol*, 47(9):736–741, Sep 2010.

[108] N. Khaldi and K. H. Wolfe. Evolutionary origins of the fumonisin secondary metabolite gene cluster in fusarium verticillioides and aspergillus niger. *Int J Evol Biol*, 2011:423821, 2011.

[109] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg. Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*, 14(4):R36, 2013.

[110] J.-E. Kim, K.-H. Han, J. Jin, H. Kim, J.-C. Kim, S.-H. Yun, and Y.-W. Lee. Putative polyketide synthase and laccase genes for biosynthesis of aurofusarin in gibberella zeae. *Appl Environ Microbiol*, 71(4):1701–1708, Apr 2005.

[111] J.-E. Kim, J. Jin, H. Kim, J.-C. Kim, S.-H. Yun, and Y.-W. Lee. Gip2, a putative transcription factor that regulates the aurofusarin biosynthetic gene cluster in gibberella zeae. *Appl Environ Microbiol*, 72(2):1645–1652, Feb 2006.

[112] Y.-T. Kim, Y.-R. Lee, J. Jin, K.-H. Han, H. Kim, J.-C. Kim, T. Lee, S.-H. Yun, and Y.-W. Lee. Two different polyketide synthase genes are required for synthesis of zearalenone in gibberella zeae. *Mol Microbiol*, 58(4):1102–1113, Nov 2005.

[113] L. Kraidlova, G. Van Zeebroeck, P. Van Dijck, and H. Sychrov. The candida albicans gap gene family encodes permeases involved in general and specific amino acid uptake and sensing. *Eukaryot Cell*, 10(9):1219–1229, Sep 2011.

[114] A. Krogh, B. Larsson, G. Von Heijne, and E. Sonnhammer. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes1. *Journal of molecular biology*, 305(3):567–580, 2001.

[115] S. Kroken, N. L. Glass, J. W. Taylor, O. C. Yoder, and B. G. Turgeon. Phylogenomic analysis of type i polyketide synthase genes in pathogenic and saprobic ascomycetes. *Proc Natl Acad Sci U S A*, 100(26):15670–15675, Dec 2003.

[116] F. Kudo and T. Eguchi. Biosynthetic genes for aminoglycoside antibiotics. *J Antibiot (Tokyo)*, 62(9):471–481, Sep 2009.

[117] M. Kvas, W. Marasas, and B. Wingfield. Diversity and evolution of fusarium species in the gibberella fujikuroi complex. *Fungal Diversity*, 34:1–21, 2009.

[118] M. H. Laurence, L. W. Burgess, B. A. Summerell, and E. C. Y. Liew. High levels of diversity in fusarium oxysporum from non-cultivated ecosystems in australia. *Fungal Biol*, 116(2):289–297, Feb 2012.

[119] J. D. Laurie, S. Ali, R. Linning, G. Mannhaupt, P. Wong, U. Güldener, M. Münsterkötter, R. Moore, R. Kahmann, G. Bakkeren, and J. Schirawski. Genome comparison of barley and maize smut fungi reveals targeted loss of rna silencing components and species-specific presence of transposable elements. *Plant Cell*, 24(5):1733–1745, May 2012.

[120] K. Lawler, K. Hammond-Kosack, A. Brazma, and R. M. Coulson. Genomic clustering and co-regulation of transcriptional networks in the pathogenic fungus fusarium graminearum. *BMC Syst Biol*, 7:52, 2013.

[121] S. Lee, H. Son, J. Lee, Y.-R. Lee, and Y.-W. Lee. A putative abc transporter gene, zra1, is required for zearalenone production in gibberella zeae. *Curr Genet*, 57(5):343–351, Oct 2011.

[122] W. Lee. *Comprehensive Discovery of Fungal Gene Clusters: Unexpected Co-work Reflected at the Genomic Level.* Dissertation, Technische Universität München, München, 2010.

[123] B. Lievens, P. M. Houterman, and M. Rep. Effector gene screening allows unambiguous identification of fusarium oxysporum f. sp. lycopersici races and discrimination from other formae speciales. *FEMS Microbiol Lett*, 300(2):201–215, Nov 2009.

[124] M. Lindblad, A. Gidlund, M. Sulyok, T. Börjesson, R. Krska, M. Olsen, and E. Fredlund. Deoxynivalenol and other selected fusarium toxins in swedish

wheat–occurrence and correlation to specific fusarium species. *Int J Food Microbiol*, 167(2):284–291, Oct 2013.

[125] P. Linnemannstöns, T. Voss, P. Hedden, P. Gaskin, and B. Tudzynski. Deletions in the gibberellin biosynthesis gene cluster of gibberella fujikuroi by restriction enzyme-mediated integration and conventional transformation-mediated mutagenesis. *Appl Environ Microbiol*, 65(6):2558–2564, Jun 1999.

[126] X. Liu, M. Inlow, and H. D. VanEtten. Expression profiles of pea pathogenicity (pep) genes in vivo and in vitro, characterization of the flanking regions of the pep cluster and evidence that the pep cluster region resulted from horizontal gene transfer in the fungal pathogen nectria haematococca. *Curr Genet*, 44(2):95–103, Nov 2003.

[127] M. B. Lohse, R. E. Zordan, C. W. Cain, and A. D. Johnson. Distinct class of dna-binding domains is exemplified by a master regulator of phenotypic switching in candida albicans. *Proc Natl Acad Sci U S A*, 107(32):14105–14110, Aug 2010.

[128] M. S. López-Berges, C. Hera, M. Sulyok, K. Schäfer, J. Capilla, J. Guarro, and A. Di Pietro. The velvet complex governs mycotoxin production and virulence of fusarium oxysporum on plant and mammalian hosts. *Mol Microbiol*, 87(1):49–65, Jan 2013.

[129] E. Lysøe, S. S. Klemsdal, K. R. Bone, R. J. N. Frandsen, T. Johansen, U. Thrane, and H. Giese. The pks4 gene of fusarium graminearum is essential for zearalenone production. *Appl Environ Microbiol*, 72(6):3924–3932, Jun 2006.

[130] E. Lyse, M. Pasquali, A. Breakspear, and H. C. Kistler. The transcription factor fgstuap influences spore development, pathogenicity, and secondary metabolism in fusarium graminearum. *Mol Plant Microbe Interact*, 24(1):54–67, Jan 2011.

[131] E. Lyse, K.-Y. Seong, and H. C. Kistler. The transcriptome of fusarium graminearum during the infection of wheat. *Mol Plant Microbe Interact*, 24(9):995–1000, Sep 2011.

[132] L. Ma, H. van der Does, K. Borkovich, J. Coleman, M. Daboussi, A. Di Pietro, M. Dufresne, M. Freitag, M. Grabherr, B. Henrissat, et al. Comparative genomics reveals mobile pathogenicity chromosomes in fusarium. *Nature*, 464(7287):367–373, 2010.

[133] S. Malonek, C. Bömke, E. Bornberg-Bauer, M. Rojas, P. Hedden, P. Hopkins, and B. Tudzynski. Distribution of gibberellin biosynthetic genes and gibberellin production in the gibberella fujikuroi species complex. *Phytochemistry*, 66(11):1296–1311, 2005.

[134] S. Malz, M. N. Grell, C. Thrane, F. J. Maier, P. Rosager, A. Felk, K. S. Albertsen, S. Salomon, L. Bohn, W. Schfer, and H. Giese. Identification of a gene cluster responsible for the biosynthesis of aurofusarin in the fusarium graminearum species complex. *Fungal Genet Biol*, 42(5):420–433, May 2005.

[135] C. M. Maragos, M. Busman, and R. D. Plattner. Development of monoclonal antibodies for the fusarin mycotoxins. *Food Addit Contam Part A Chem Anal Control Expo Risk Assess*, 25(1):105–114, Jan 2008.

[136] E. Margolis-Clark, I. Hunt, S. Espinosa, and B. J. Bowman. Identification of the gene at the pmg locus, encoding system ii, the general amino acid transporter in neurospora crassa. *Fungal Genet Biol*, 33(2):127–135, Jul 2001.

[137] K. A. Marr, R. A. Carter, F. Crippa, A. Wald, and L. Corey. Epidemiology and outcome of mould infections in hematopoietic stem cell transplant recipients. *Clin Infect Dis*, 34(7):909–917, Apr 2002.

[138] J. F. Martín. Molecular control of expression of penicillin biosynthesis genes in fungi: regulatory proteins interact with a bidirectional promoter region. *J Bacteriol*, 182(9):2355–2362, May 2000.

[139] M. F. Martín and P. Liras. Organization and expression of genes involved in the biosynthesis of antibiotics and other secondary metabolites. *Annu Rev Microbiol*, 43:173–206, 1989.

[140] F. J. Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

[141] A. Mathelier, X. Zhao, A. W. Zhang, F. Parcy, R. Worsley-Hunt, D. J. Arenillas, S. Buchman, C.-y. Chen, A. Chou, H. Ienasescu, J. Lim, C. Shyr, G. Tan, M. Zhou, B. Lenhard, A. Sandelin, and W. W. Wasserman. Jaspar 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res*, 42(Database issue):D142–D147, Jan 2014.

[142] S. P. McCormick, N. J. Alexander, and L. J. Harris. Clm1 of fusarium graminearum encodes a longiborneol synthase required for culmorin production. *Appl Environ Microbiol*, 76(1):136–141, Jan 2010.

[143] C. B. Michielse, A. Pfannmüller, M. Macios, P. Rengers, A. Dzikowska, and B. Tudzynski. The interplay between the gata transcription factors area, the global nitrogen regulator and areb in fusarium fujikuroi. *Mol Microbiol*, 91(3):472–493, Feb 2014.

[144] C. B. Michielse, L. Studt, S. Janevska, C. M. K. Sieber, B. Arndt, J. J. Espino, H.-U. Humpf, U. Güldener, and B. Tudzynski. The global regulator ffsge1 is required for expression of secondary metabolite gene clusters, but not for pathogenicity in fusarium fujikuroi. *Environ Microbiol*, Aug 2014.

[145] M. Mihlan, V. Homann, T.-W. D. Liu, and B. Tudzynski. Area directly mediates nitrogen regulation of gibberellin biosynthesis in gibberella fujikuroi, but its activity is not affected by nmr. *Mol Microbiol*, 47(4):975–991, Feb 2003.

[146] R. D. Monds, M. G. Cromey, D. R. Lauren, M. di Menna, and J. Marshall. Fusarium graminearum, f. cortaderiae and f. pseudograminearum in new zealand: molecular phylogenetic analysis, mycotoxin chemotypes and co-existence of species. *Mycol Res*, 109(Pt 4):410–420, Apr 2005.

[147] M. R. Montminy, K. A. Sevarino, J. A. Wagner, G. Mandel, and R. H. Goodman. Identification of a cyclic-amp-responsive element within the rat somatostatin gene. *Proc Natl Acad Sci U S A*, 83(18):6682–6686, Sep 1986.

[148] D. Morrone, J. Chambers, L. Lowry, G. Kim, A. Anterola, K. Bender, and R. J. Peters. Gibberellin biosynthesis in bacteria: separate ent-copalyl diphosphate and ent-kaurene synthases in bradyrhizobium japonicum. *FEBS Lett*, 583(2):475–480, Jan 2009.

[149] A. Muszewska, M. Hoffman-Sommer, and M. Grynberg. Ltr retrotransposons in fungi. *PLoS One*, 6(12):e29425, 2011.

[150] S. Mller, C. Baldin, M. Groth, R. Guthke, O. Kniemeyer, A. A. Brakhage, and V. Valiante. Comparison of transcriptome technologies in the pathogenic fungus aspergillus fumigatus reveals novel insights into the genome and mpka dependent gene expression. *BMC Genomics*, 13:519, 2012.

[151] C. G. Nasmith, S. Walkowiak, L. Wang, W. W. Y. Leung, Y. Gong, A. Johnston, L. J. Harris, D. S. Guttman, and R. Subramaniam. Tri6 is a global transcription regulator in the phytopathogen fusarium graminearum. *PLoS Pathog*, 7(9):e1002266, Sep 2011.

[152] V. Q. Nguyen and A. Sil. Temperature-induced switch to the pathogenic yeast form of histoplasma capsulatum requires ryp1, a conserved transcriptional regulator. *Proc Natl Acad Sci U S A*, 105(12):4880–4885, Mar 2008.

[153] W. Nickel and M. Seedorf. Unconventional mechanisms of protein transport to the cell surface of eukaryotic cells. *Annu Rev Cell Dev Biol*, 24:287–308, 2008.

[154] E.-M. Niehaus, S. Janevska, K. W. von Bargen, C. M. K. Sieber, H. Harrer, H.-U. Humpf, and B. Tudzynski. Apicidin f: Characterization and genetic manipulation of a new secondary metabolite gene cluster in the rice pathogen fusarium fujikuroi. *PLoS One*, 9(7):e103336, 2014.

[155] E.-M. Niehaus, K. Kleigrewe, P. Wiemann, L. Studt, C. M. K. Sieber, L. R. Connolly, M. Freitag, U. Güldener, B. Tudzynski, and H.-U. Humpf. Genetic manipulation of the fusarium fujikuroi fusarin gene cluster yields insight into the complex regulation and fusarin biosynthetic pathway. *Chem Biol*, 20(8):1055–1066, Aug 2013.

[156] M. Nielen, C. Weijers, J. Peters, L. Weignerová, H. Zuilhof, and M. Franssen. Rapid enzymatic hydrolysis of masked deoxynivalenol and zearalenone prior to liquid chromatography mass spectrometry or immunoassay analysis. *World Mycotoxin Journal*, 7(2):107–113, 2014.

[157] L. K. Nielsen, D. J. Cook, S. G. Edwards, and R. V. Ray. The prevalence and impact of fusarium head blight pathogens and mycotoxins on malting barley quality in uk. *Int J Food Microbiol*, 179:38–49, Jun 2014.

[158] O. Novikova, V. Fet, and A. Blinov. Non-ltr retrotransposons in fungi. *Funct Integr Genomics*, 9(1):27–42, Feb 2009.

[159] H.-W. Nützmann, Y. Reyes-Dominguez, K. Scherlach, V. Schroeckh, F. Horn, A. Gacek, J. Schümann, C. Hertweck, J. Strauss, and A. A. Brakhage. Bacteria-induced natural product formation in the fungus aspergillus nidulans requires saga/ada-mediated histone acetylation. *Proc Natl Acad Sci U S A*, 108(34):14282–14287, Aug 2011.

[160] K. O'Donnell, T. J. Ward, D. Aberra, H. C. Kistler, T. Aoki, N. Orwig, M. Kimura, S. Bjørnstad, and S. S. Klemsdal. Multilocus genotyping and molecular phylogenetics resolve a novel head blight pathogen within the fusarium graminearum species complex from ethiopia. *Fungal Genet Biol*, 45(11):1514–1522, Nov 2008.

[161] R. A. Ohm, N. Feau, B. Henrissat, C. L. Schoch, B. A. Horwitz, K. W. Barry, B. J. Condon, A. C. Copeland, B. Dhillon, F. Glaser, C. N. Hesse, I. Kosti, K. LaButti, E. A. Lindquist, S. Lucas, A. A. Salamov, R. E. Bradshaw, L. Ciuffetti, R. C. Hamelin, G. H. J. Kema, C. Lawrence, J. A. Scott, J. W. Spatafora, B. G. Turgeon, P. J. G. M. de Wit, S. Zhong, S. B. Goodwin, and I. V. Grigoriev. Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen dothideomycetes fungi. *PLoS Pathog*, 8(12):e1003037, 2012.

[162] S. Oide, W. Moeder, S. Krasnoff, D. Gibson, H. Haas, K. Yoshioka, and B. G. Turgeon. Nps6, encoding a nonribosomal peptide synthetase involved in siderophore-mediated iron metabolism, is a conserved virulence determinant of plant pathogenic ascomycetes. *Plant Cell*, 18(10):2836–2853, Oct 2006.

[163] A. Osbourn. Secondary metabolic gene clusters: evolutionary toolkits for chemical innovation. *Trends in Genetics*, 26(10):449–457, 2010.

[164] G. Pavesi, P. Mereghetti, G. Mauri, and G. Pesole. Weeder web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic acids research*, 32(suppl 2):W199, 2004.

[165] K. F. Pedley and J. D. Walton. Regulation of cyclic peptide biosynthesis in a plant pathogenic fungus by a novel transcription factor. *Proc Natl Acad Sci U S A*, 98(24):14174–14179, Nov 2001.

[166] J. J. Pestka and A. T. Smolinski. Deoxynivalenol: toxicology and potential effects on humans. *J Toxicol Environ Health B Crit Rev*, 8(1):39–69, 2005.

[167] T. Petersen, S. Brunak, G. von Heijne, and H. Nielsen. Signalp 4.0: discriminating signal peptides from transmembrane regions. *nature methods*, 8(10):785–786, 2011.

[168] B. Poppenberger, F. Berthiller, D. Lucyshyn, T. Sieberer, R. Schuhmacher, R. Krska, K. Kuchler, J. Glössl, C. Luschnig, and G. Adam. Detoxification of the fusarium mycotoxin deoxynivalenol by a udp-glucosyltransferase from arabidopsis thaliana. *J Biol Chem*, 278(48):47905–47914, Nov 2003.

[169] A. Price, N. Jones, and P. Pevzner. De novo identification of repeat families in large genomes. *Bioinformatics*, 21(Suppl 1):i351, 2005.

[170] M. N. Price, A. P. Arkin, and E. J. Alm. The life-cycle of operons. *PLoS Genet*, 2(6):e96, Jun 2006.

[171] R. H. Proctor, M. Busman, J.-A. Seo, Y. W. Lee, and R. D. Plattner. A fumonisin biosynthetic gene cluster in fusarium oxysporum strain o-1890 and the genetic basis for b versus c fumonisin production. *Fungal Genet Biol*, 45(6):1016–1026, Jun 2008.

[172] R. H. Proctor, R. A. E. Butchko, D. W. Brown, and A. Moretti. Functional characterization, sequence comparisons and distribution of a polyketide synthase gene required for perithecial pigmentation in some fusarium species. *Food Addit Contam*, 24(10):1076–1087, Oct 2007.

[173] R. H. Proctor, R. D. Plattner, D. W. Brown, J.-A. Seo, and Y.-W. Lee. Discontinuous distribution of fumonisin biosynthetic genes in the gibberella fujikuroi species complex. *Mycol Res*, 108(Pt 7):815–822, Jul 2004.

[174] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2013.

[175] T. Rattei, P. Tischler, S. Götz, M. Jehl, J. Hoser, R. Arnold, A. Conesa, and H. Mewes. Simap a comprehensive database of pre-calculated protein sequence similarities, domains, annotations and clusters. *Nucleic acids research*, 38(suppl 1):D223–D226, 2010.

[176] D. O. Rees, N. Bushby, R. J. Cox, J. R. Harding, T. J. Simpson, and C. L. Willis. Synthesis of [1,2-13c2, 15n]-l-homoserine and its incorporation by the pks-nrps system of fusarium moniliforme into the mycotoxin fusarin c. *Chembiochem*, 8(1):46–50, Jan 2007.

[177] B. Regenberg, L. Dring-Olsen, M. C. Kielland-Brandt, and S. Holmberg. Substrate specificity and gene expression of the amino-acid permeases in saccharomyces cerevisiae. *Curr Genet*, 36(6):317–328, Dec 1999.

[178] I. Reid, N. O'Toole, O. Zabaneh, R. Nourzadeh, M. Dahdouli, M. Abdellateef, P. M. K. Gordon, J. Soh, G. Butler, C. W. Sensen, and A. Tsang. Snowyowl: accurate prediction of fungal genes by using rna-seq and homology information to select among ab initio models. *BMC Bioinformatics*, 15:229, 2014.

[179] M. Rep, H. C. van der Does, M. Meijer, R. van Wijk, P. M. Houterman, H. L. Dekker, C. G. de Koster, and B. J. C. Cornelissen. A small, cysteine-rich protein secreted by fusarium oxysporum during colonization of xylem vessels is required for i-3-mediated resistance in tomato. *Mol Microbiol*, 53(5):1373–1383, Sep 2004.

[180] Y. Reyes-Dominguez, J. W. Bok, H. Berger, E. K. Shwab, A. Basheer, A. Gallmetzer, C. Scazzocchio, N. Keller, and J. Strauss. Heterochromatic marks are associated with the repression of secondary metabolism clusters in aspergillus nidulans. *Mol Microbiol*, 76(6):1376–1386, Jun 2010.

[181] R. Rodríguez-Ortiz, M. C. Limón, and J. Avalos. Regulation of carotenogenesis and secondary metabolism by nitrogen in wild-type fusarium fujikuroi and carotenoid-overproducing mutants. *Appl Environ Microbiol*, 75(2):405–413, Jan 2009.

[182] U. L. Rosewich and H. C. Kistler. Role of horizontal gene transfer in the evolution of fungi. *Annu Rev Phytopathol*, 38:325–363, 2000.

[183] L. V. Roze, A. E. Arthur, S.-Y. Hong, A. Chanda, and J. E. Linz. 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(3):713–726, Nov 2007.

[184] A. Ruepp, A. Zollner, D. Maier, K. Albermann, J. Hani, M. Mokrejs, I. Tetko, U. Güldener, G. Mannhaupt, M. Münsterkötter, et al. The funcat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Research*, 32(18):5539, 2004.

[185] J. C. Schimenti and C. H. Duncan. Ruminant globin gene structures suggest an evolutionary role for alu-type repeats. *Nucleic Acids Res*, 12(3):1641–1655, Feb 1984.

[186] G. Schweikert, A. Zien, G. Zeller, J. Behr, C. Dieterich, C. S. Ong, P. Philips, F. De Bona, L. Hartmann, A. Bohlen, N. Krger, S. Sonnenburg, and G. Rtsch. mgene: accurate svm-based gene finding with an application to nematode genomes. *Genome Res*, 19(11):2133–2143, Nov 2009.

[187] E. U. Selker. Premeiotic instability of repeated sequences in neurospora crassa. *Annu Rev Genet*, 24:579–613, 1990.

[188] D.-G. Seo, C. Phat, D.-H. Kim, and C. Lee. Occurrence of fusarium mycotoxin fumonisin b1 and b2 in animal feeds in korea. *Mycotoxin Res*, 29(3):159–167, Aug 2013.

[189] K.-Y. Seong, M. Pasquali, X. Zhou, J. Song, K. Hilburn, S. McCormick, Y. Dong, J.-R. Xu, and H. C. Kistler. Global gene regulation by fusarium transcription factors tri6 and tri10 reveals adaptations for toxin biosynthesis. *Mol Microbiol*, 72(2):354–367, Apr 2009.

[190] K.-Y. Seong, X. Zhao, J.-R. Xu, U. Güldener, and H. C. Kistler. Conidial germination in the filamentous fungus fusarium graminearum. *Fungal Genet Biol*, 45(4):389–399, Apr 2008.

[191] C. M. K. Sieber, W. Lee, P. Wong, M. Münsterkötter, H.-W. Mewes, C. Schmeitzl, E. Varga, F. Berthiller, G. Adam, and U. Güldener. The fusarium graminearum genome reveals more secondary metabolite gene clusters and hints of horizontal gene transfer. *PLoS One*, 9(10):e110311, 2014.

[192] U. R. Sikhakolli, F. López-Giráldez, N. Li, R. Common, J. P. Townsend, and F. Trail. Transcriptome analyses during fruiting body formation in fusarium graminearum and fusarium verticillioides reflect species life history and ecology. *Fungal Genet Biol*, 49(8):663–673, Aug 2012.

[193] A. Smit, R. Hubley, and P. Green. Repeatmasker open-3.0. http://www.repeatmasker.org, 1996-2010.

[194] C. A. Smith, C. P. Woloshuk, D. Robertson, and G. A. Payne. Silencing of the aflatoxin gene cluster in a diploid strain of aspergillus flavus is suppressed by ectopic aflr expression. *Genetics*, 176(4):2077–2086, Aug 2007.

[195] M. W. Smith, D. F. Feng, and R. F. Doolittle. Evolution by acquisition: the case for horizontal gene transfers. *Trends Biochem Sci*, 17(12):489–493, Dec 1992.

[196] G. K. Smyth. Limma: linear models for microarray data. In R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, and W. Huber, editors, *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, pages 397–420. Springer, New York, 2005.

[197] A. A. Soukup, Y.-M. Chiang, J. W. Bok, Y. Reyes-Dominguez, B. R. Oakley, C. C. C. Wang, J. Strauss, and N. P. Keller. Overexpression of the aspergillus nidulans histone 4 acetyltransferase esaa increases activation of secondary metabolite production. *Mol Microbiol*, 86(2):314–330, Oct 2012.

[198] R. A. Squire. Ranking animal carcinogens: a proposed regulatory approach. *Science*, 214(4523):877–880, Nov 1981.

[199] M. Stanke, M. Diekhans, R. Baertsch, and D. Haussler. Using native and syntenically mapped cdna alignments to improve de novo gene finding. *Bioinformatics*, 24(5):637–644, 2008.

[200] A. E. Stephens, D. M. Gardiner, R. G. White, A. L. Munn, and J. M. Manners. Phases of infection and gene expression of fusarium graminearum

during crown rot disease of wheat. *Mol Plant Microbe Interact*, 21(12):1571–1581, Dec 2008.

[201] J. Strauss and Y. Reyes-Dominguez. Regulation of secondary metabolism by chromatin structure and epigenetic codes. *Fungal Genet Biol*, 48(1):62–69, Jan 2011.

[202] L. Studt, F. J. Schmidt, L. Jahn, C. M. K. Sieber, L. R. Connolly, E.-M. Niehaus, M. Freitag, H.-U. Humpf, and B. Tudzynski. Two histone deacetylases, ffhda1 and ffhda2, are important for secondary metabolism and virulence in fusarium fujikuroi. *Appl Environ Microbiol*, 79(24):7719–7734, Dec 2013.

[203] L. Studt, P. Wiemann, K. Kleigrewe, H.-U. Humpf, and B. Tudzynski. Biosynthesis of fusarubins accounts for pigmentation of fusarium fujikuroi perithecia. *Appl Environ Microbiol*, 78(12):4468–4480, Jun 2012.

[204] M. Suyama, D. Torrents, and P. Bork. Pal2nal: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res*, 34(Web Server issue):W609–W612, Jul 2006.

[205] J. L. Srensen, F. T. Hansen, T. E. Sondergaard, D. Staerk, T. V. Lee, R. Wimmer, L. G. Klitgaard, S. Purup, H. Giese, and R. J. N. Frandsen. Production of novel fusarielins by ectopic activation of the polyketide synthase 9 cluster in fusarium graminearum. *Environ Microbiol*, 14(5):1159–1170, May 2012.

[206] G. Talavera and J. Castresana. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol*, 56(4):564–577, Aug 2007.

[207] M. C. Teixeira, P. T. Monteiro, J. F. Guerreiro, J. P. Gonçalves, N. P. Mira, S. C. dos Santos, T. R. Cabrito, M. Palma, C. Costa, A. P. Francisco, S. C. Madeira, A. L. Oliveira, A. T. Freitas, and I. Sá-Correia. The yeastract database: an upgraded information system for the analysis of gene and genomic transcription regulation in saccharomyces cerevisiae. *Nucleic Acids Res*, 42(Database issue):D161–D166, Jan 2014.

[208] L. A. Tell. Aspergillosis in mammals and birds: impact on veterinary medicine. *Med Mycol*, 43 Suppl 1:S71–S73, May 2005.

[209] L. F. Thatcher, D. M. Gardiner, K. Kazan, and J. M. Manners. A highly conserved effector in fusarium oxysporum is required for full virulence on arabidopsis. *Mol Plant Microbe Interact*, 25(2):180–190, Feb 2012.

[210] J. Tilburn, S. Sarkar, D. A. Widdick, E. A. Espeso, M. Orejas, J. Mungroo, M. A. Peñalva, and H. Arst, Jr. The aspergillus pacc zinc finger transcription factor mediates regulation of both acid- and alkaline-expressed genes by ambient ph. *EMBO J*, 14(4):779–790, Feb 1995.

[211] D. Tisch and M. Schmoll. Light regulation of metabolic pathways in fungi. *Appl Microbiol Biotechnol*, 85(5):1259–1277, Feb 2010.

[212] C. Tobiasen, J. Aahman, K. S. Ravnholt, M. J. Bjerrum, M. N. Grell, and H. Giese. 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(1):43–58, Jan 2007.

[213] H. Trip, M. E. Evers, J. A. K. W. Kiel, and A. J. M. Driessen. Uptake of the beta-lactam precursor alpha-aminoadipic acid in penicillium chrysogenum is mediated by the acidic and the general amino acid permease. *Appl Environ Microbiol*, 70(8):4775–4783, Aug 2004.

[214] B. Tudzynski. Gibberellin biosynthesis in fungi: genes, enzymes, evolution, and impact on biotechnology. *Appl Microbiol Biotechnol*, 66(6):597–611, Mar 2005.

[215] B. Tudzynski and K. Hölter. Gibberellin biosynthetic pathway in gibberella fujikuroi: evidence for a gene cluster. *Fungal Genet Biol*, 25(3):157–170, Dec 1998.

[216] H. C. van der Does and M. Rep. Virulence genes and the evolution of host specificity in plant-pathogenic fungi. *Mol Plant Microbe Interact*, 20(10):1175–1182, Oct 2007.

[217] G. Van Zeebroeck, M. Rubio-Texeira, J. Schothorst, and J. M. Thevelein. Specific analogues uncouple transport, signalling, oligo-ubiquitination and endocytosis in the yeast gap1 amino acid transceptor. *Mol Microbiol*, May 2014.

[218] S. Veluchamy and J. A. Rollins. A cry-dash-type photolyase/cryptochrome from sclerotinia sclerotiorum mediates minor uv-a-specific effects on development. *Fungal Genet Biol*, 45(9):1265–1276, Sep 2008.

[219] P. Verweij, M. Brandt, P. Murray, E. Baron, J. Jorgensen, M. Landry, M. Pfaller, et al. Aspergillus, fusarium, and other opportunistic moniliaceous fungi. *Manual of clinical microbiology: Volume 2*, (Ed. 9):1802–1838, 2006.

[220] I. Visentin, V. Montis, K. Dll, C. Alabouvette, G. Tamietti, P. Karlovsky, and F. Cardinale. Transcription of genes in the biosynthetic pathway for fumonisin mycotoxins is epigenetically and differentially regulated in the fungal maize pathogen fusarium verticillioides. *Eukaryot Cell*, 11(3):252–259, Mar 2012.

[221] M. C. Walter, T. Rattei, R. Arnold, U. Güldener, M. Münsterkötter, K. Nenova, G. Kastenmüller, P. Tischler, A. Wölling, A. Volz, N. Pongratz, R. Jost, H. W. Mewes, and D. Frishman. Pedant covers all complete refseq genomes. *Nucleic Acids Res*, 37(Database issue):408–411, Jan 2009.

[222] T. Wang. Using phylocon to identify conserved regulatory motifs. *Curr Protoc Bioinformatics*, Chapter 2:Unit 2.12, Sep 2007.

[223] T. Wang and G. D. Stormo. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics*, 19(18):2369–2380, Dec 2003.

[224] Y.-M. Wang, S.-Q. Peng, Q. Zhou, M.-W. Wang, C.-H. Yan, H.-Y. Yang, and G.-Q. Wang. Depletion of intracellular glutathione mediates butenolide-induced cytotoxicity in hepg2 cells. *Toxicol Lett*, 164(3):231–238, Jul 2006.

[225] K. M. Waters, T. Liu, R. D. Quesenberry, A. R. Willse, S. Bandyopadhyay, L. E. Kathmann, T. J. Weber, R. D. Smith, H. S. Wiley, and B. D. Thrall. Network analysis of epidermal growth factor signaling using integrated genomic, proteomic and phosphorylation data. *PLoS One*, 7(3):e34515, 2012.

[226] M. K. Watters, T. A. Randall, B. S. Margolin, E. U. Selker, and D. R. Stadler. Action of repeat-induced point mutation on both strands of a duplex and on tandem duplications of various sizes in neurospora. *Genetics*, 153(2):705–714, Oct 1999.

[227] S. R. Wessler. Transposable elements and the evolution of eukaryotic genomes. *Proc Natl Acad Sci U S A*, 103(47):17600–17601, Nov 2006.

[228] T. Wicker, F. Sabot, A. Hua-Van, J. L. Bennetzen, P. Capy, B. Chalhoub, A. Flavell, P. Leroy, M. Morgante, O. Panaud, E. Paux, P. SanMiguel, and A. H. Schulman. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet*, 8(12):973–982, Dec 2007.

[229] P. Wiemann, D. W. Brown, K. Kleigrewe, J. W. Bok, N. P. Keller, H.-U. Humpf, and B. Tudzynski. Ffvel1 and fflae1, components of a velvet-like complex in fusarium fujikuroi, affect differentiation, secondary metabolism and virulence. *Mol Microbiol*, Jun 2010.

[230] P. Wiemann, C.-J. Guo, J. M. Palmer, R. Sekonyela, C. C. C. Wang, and N. P. Keller. Prototype of an intertwined secondary-metabolite supercluster. *Proc Natl Acad Sci U S A*, 110(42):17065–17070, Oct 2013.

[231] P. Wiemann, C. M. K. Sieber, K. W. von Bargen, L. Studt, E.-M. Niehaus, J. J. Espino, K. Huß, C. B. Michielse, S. Albermann, D. Wagner, S. V. Bergner, L. R. Connolly, A. Fischer, G. Reuter, K. Kleigrewe, T. Bald, B. D. Wingfield, R. Ophir, S. Freeman, M. Hippler, K. M. Smith, D. W. Brown, R. H. Proctor, M. Münsterkötter, M. Freitag, H.-U. Humpf, U. Güldener, and B. Tudzynski. Deciphering the cryptic genome: Genome-wide analyses of the rice pathogen fusarium fujikuroi reveal complex regulation of secondary metabolism and novel metabolites. *PLoS Pathog*, 9(6):e1003475, Jun 2013.

[232] P. Wiemann, A. Willmann, M. Straeten, K. Kleigrewe, M. Beyer, H.-U. Humpf, and B. Tudzynski. Biosynthesis of the red pigment bikaverin in fusarium fujikuroi: genes, their function and regulation. *Mol Microbiol*, 72(4):931–946, May 2009.

[233] R. B. Williams, J. C. Henrikson, A. R. Hoover, A. E. Lee, and R. H. Cichewicz. Epigenetic remodeling of the fungal secondary metabolome. *Org Biomol Chem*, 6(11):1895–1897, Jun 2008.

[234] J. Win, W. Morgan, J. Bos, K. V. Krasileva, L. M. Cano, A. Chaparro-Garcia, R. Ammar, B. J. Staskawicz, and S. Kamoun. Adaptive evolution has targeted the c-terminal domain of the rxlr effectors of plant pathogenic oomycetes. *Plant Cell*, 19(8):2349–2369, Aug 2007.

[235] C. E. Windels. Economic and social impacts of fusarium head blight: changing farms and rural communities in the northern great plains. *Phytopathology*, 90(1):17–21, Jan 2000.

[236] E. Wingender, X. Chen, E. Fricke, R. Geffers, R. Hehl, I. Liebich, M. Krull, V. Matys, H. Michael, R. Ohnhäuser, et al. The transfac system on gene expression regulation. *Nucleic acids research*, 29(1):281–283, 2001.

[237] B. D. Wingfield, E. T. Steenkamp, Q. C. Santana, M. Coetzee, S. Bam, I. Barnes, C. W. Beukes, W. Yin Chan, L. De Vos, G. Fourie, et al. First fungal genome sequence from africa: a preliminary analysis. *South African Journal of Science*, 108(1-2):01–09, 2012.

[238] C. P. Woloshuk, G. L. Yousibova, J. A. Rollins, D. Bhatnagar, and G. A. Payne. Molecular characterization of the afl-1 locus in aspergillus flavus. *Appl Environ Microbiol*, 61(8):3019–3023, Aug 1995.

[239] P. Wong, M. Walter, W. Lee, G. Mannhaupt, M. Münsterkötter, H.-W. Mewes, G. Adam, and U. Güldener. Fgdb: revisiting the genome annotation of the plant pathogen fusarium graminearum. *Nucleic Acids Res*, 39(Database issue):D637–D639, Jan 2011.

[240] J. Wootton and S. Federhen. Statistics of local complexity in amino acid sequences and sequence databases. *Computers & chemistry*, 17(2):149–163, 1993.

[241] J. R. Xu and J. F. Leslie. A genetic map of Gibberella fujikuroi mating population A (Fusarium moniliforme). *Genetics*, 143(1):175–189, May 1996.

[242] T. Yabuta, T. Hayasi, et al. Biochemical studies of'bakanae'fungus of rice. *Journal of the Imperial Agricultural Experimental Station, Nisigahara, Tokyo*, 3(3):365–400, 1940.

[243] D. Yu, F. Xu, J. Zeng, and J. Zhan. Type iii polyketide synthases in natural product biosynthesis. *IUBMB Life*, 64(4):285–295, Apr 2012.

[244] E. Zdobnov and R. Apweiler. Interproscan – an integration platform for the signature-recognition methods in interpro. *Bioinformatics*, 17(9):847–848, 2001.

[245] K. A. Zeller, R. L. Bowden, and J. F. Leslie. Diversity of epidemic populations of gibberella zeae from small quadrats in kansas and north dakota. *Phytopathology*, 93(7):874–880, Jul 2003.

[246] X. Zeng, M. Nesbitt, J. Pei, K. Wang, I. Vergara, and N. Chen. Orthocluster: a new tool for mining synteny blocks and applications in comparative genomics. In *Proceedings of the 11th international conference on Extending database technology: Advances in database technology*, pages 656–667. ACM, 2008.

[247] L. Zhang, J. Wang, C. Zhang, and Q. Wang. Analysis of potential fumonisin-producing fusarium species in corn products from three main maize-producing areas in eastern china. *J Sci Food Agric*, 93(3):693–701, Feb 2013.

[248] X.-W. Zhang, L.-J. Jia, Y. Zhang, G. Jiang, X. Li, D. Zhang, and W.-H. Tang. In planta stage-specific fungal gene profiling elucidates the molecular strategies of fusarium graminearum growing inside wheat coleoptiles. *Plant Cell*, 24(12):5159–5176, Dec 2012.

Appendix A

Tables

Table A.1: Functional gene descriptions and positions of overrepresented promoter motifs on predicted clusters and neighboring genes in *F. graminearum*. Heatmaps of co-regulated predicted gene clusters are shown in Figure 4.1.

Cluster_ID	Position	Gene_Code	Description	Predicted_Motif
Triacetylfusarinin	-3	FGSG_15204	hypothetical protein	
	-2	FGSG_15203	hypothetical protein	
	-1	FGSG_03748	conserved hypothetical protein	
	1	FGSG_03747	related to AM-toxin synthetase (AMT)	
	2	FGSG_03745	related to aerobactin siderophore biosynthesis protein iucB	
	3	FGSG_03744	related to major facilitator MirA	
	4	FGSG_03742	related to cellobiose dehydrogenase	
	5	FGSG_03741	related to O-methylsterigmatocystin oxidoreductase	
	+1	FGSG_12389	conserved hypothetical protein	
	+2	FGSG_16211	related to enoyl-CoA hydratase	
	+3	FGSG_16212	hypothetical protein	
Malonichrome	-3	FGSG_11031	hypothetical protein	
	-2	FGSG_13867	hypothetical protein	
	-1	FGSG_11030	related to ferric reductase Fre2p	
	1	FGSG_11029	related to major facilitator MirA	TAGGGATCGGCG
	2	FGSG_11028	related to ATP-binding cassette transporter protein YOR1	CAGGGATCGGCC
	3	FGSG_11027	conserved hypothetical protein	CAGGGATCGGCC
	4	FGSG_11026	non-ribosomal peptide synthetase	CAGGGATCGGCA
	5	FGSG_11025	putative C2H2 zinc finger transcription factor	
	+1	FGSG_13868	conserved hypothetical protein	
	+2	FGSG_11024	probable cytochrome P450 51 (eburicol 14 alpha-demethylase)	
	+3	FGSG_11023	conserved hypothetical protein	
C02	-3	FGSG_00032	related to non-heme chloroperoxidase	
	-2	FGSG_00033	conserved hypothetical protein	
	-1	FGSG_00034	related to alpha-glucoside transport protein	
	1	FGSG_11653	probable sulfatase	
	2	FGSG_11654	related to nitrate assimilation regulatory protein	
	3	FGSG_00036	probable fatty acid synthase, alpha subunit	GTGGtgCCAC
	4	FGSG_11656	related to FAS1 - fatty-acyl-CoA synthase, beta chain	GTGGtgCCAC
	5	FGSG_00038	hypothetical protein	GTGGtgCCAC
	6	FGSG_00039	conserved hypothetical protein	
	7	FGSG_00040	conserved hypothetical protein	
	8	FGSG_11657	conserved hypothetical protein	
	9	FGSG_11658	hypothetical protein	
	10	FGSG_00043	conserved hypothetical protein	
	11	FGSG_00044	conserved hypothetical protein	GTGGtgCCAC
	12	FGSG_00045	conserved hypothetical protein	GTGGtgCCAC
	13	FGSG_00046	related to multidrug resistance protein	GTGGtgCCAC
	14	FGSG_00047	conserved hypothetical protein	GTGGtgCCAC
	15	FGSG_00048	related to flavonol synthase-like protein	GTGGtgCCAC
	16	FGSG_00049	related to branched-chain amino acid aminotransferase	GTGGtaCCAC
	17	FGSG_11661	conserved hypothetical protein	GTGGtgCCAC
	18	FGSG_00050	conserved hypothetical protein	GTGGtgCCAC
	+1	FGSG_00051	related to aliphatic nitrilase	
	+2	FGSG_15673	non-ribosomal peptide synthetase	
	+3	FGSG_15680	related to benzoate-para-hydroxylase (cytochrome P450)	
Butenolide	-3	FGSG_13444	related to allantoate transporter	
	-2	FGSG_13445	probable benzoate 4-monooxygenase cytochrome P450	

Continued on next page

Table A.1 – continued from previous page

Cluster_ID	Position	Gene_Code	Description	Predicted_Motif
	-1	FGSG_08085	conserved hypothetical protein	
	1	FGSG_08084	related to monocarboxylate transporter 4	TAATGCTCCG
	2	FGSG_08083	related to glutamic acid decarboxylase	AAATGGACCG
	3	FGSG_08082	conserved hypothetical protein	AAATGGACCG
	4	FGSG_08081	related to gibberellin 20-oxidase	AAATTGTCCG
	5	FGSG_08080	conserved hypothetical protein	AAGTGCTCCG
	6	FGSG_08079	probable benzoate 4-monooxygenase cytochrome P450	TAATGCTCCG
	7	FGSG_08078	related to general amidase	AAATGCTCCG
	8	FGSG_08077	related to flavin oxidoreductase	AAATGCTCCG
	+1	FGSG_08076	hypothetical protein	
	+2	FGSG_17106	hypothetical protein	
	+3	FGSG_08074	conserved hypothetical protein	
Trichothecenes	-3	FGSG_03545	related to OrfH - unknown, trichothecene gene cluster	
	-2	FGSG_12416	conserved hypothetical protein	
	-1	FGSG_03544	deacetylase	
	1	FGSG_03543	putative trichothecene biosynthesis gene	TCAGGCCT
	2	FGSG_03542	probable cytochrome P450	
	3	FGSG_03541	trichothecene efflux pump	TCAGGCCT
	4	FGSG_03540	isotrichodermin C-15 hydroxylase	TTAGGCCT
	5	FGSG_03539	hypothetical protein	TCAGGCCT
	6	FGSG_03538	regulatory protein	
	7	FGSG_03537	trichodiene synthase [sesquiterpene cyclase]	TAAGGCCT
	8	FGSG_16251	trichothecene biosynthesis positive transcription factor	TCAGGCCT
	9	FGSG_03535	trichodiene oxygenase [cytochrome P450]	TCAGGCCT
	10	FGSG_03534	trichothecene 15-O-acetyltransferase	
	11	FGSG_03533	related to TRI7 - trichothecene biosynthesis gene cluster	TCAGGCCT
	12	FGSG_03532	trichothecene 3-O-esterase	TCAGGCCT
	13	FGSG_03531	monooxygenase	
	14	FGSG_03530	acetylesterase, trichothecene gene cluster	TCAGGCCT
	15	FGSG_03529	related to glucan 1,3-beta-glucosidase	TCAGGCCT
	+1	FGSG_03528	conserved hypothetical protein	
	+2	FGSG_03527	conserved hypothetical protein	
	+3	FGSG_03526	unknown, trichothecene gene cluster	

Table A.2: List of publicly available genomes used for ortholog analysis. Extract of SIMAP protein similarity database used for the ortholog analysis of predicted gene clusters, listing species / strain, Pedant database name and corresponding NCBI bioproject-ID (PRJNA). Genomes of *F. oxysporum* strains were obtained from Broad institute and partially have no bioproject-ID, yet. These genomes are indicated by an asterisk and carry an internal accession number.

Organism	Pedant-DB	Accession
Acidithiobacillus ferrooxidans ATCC 23270	p3_r57649_Aci_ferro	57649
Acinetobacter baumannii ATCC 17978	p3_r58731_Aci_bauma	58731
Actinobacillus pleuropneumoniae L20	p3_r58789_Act_pleur	58789
Agaricus bisporus var. burnettii JB137-S8	p3_r61007_Aga_bispo	61007
Aggregatibacter actinomycetemcomitans D11S-1	p3_r41333_Agg_actin	41333
Agrobacterium radiobacter K84	p3_r58269_Agr_radio	58269
Agrobacterium tumefaciens str. C58	p3_r57865_Agr_fabru	57865
Ajellomyces capsulatus G186AR	p3_r12635_Aje_capsu	12635
Ajellomyces capsulatus NAm1	p3_p12654_His_capsu	12654
Ajellomyces dermatitidis SLH14081	p3_r41099_Aje_derma	41099
Amycolatopsis mediterranei U32	p3_r50565_Amy_medit	50565
Anaplasma marginale str. St. Maries	p3_r57629_Ana_margi	57629
Arabidopsis thaliana	p3_p116_Ara_thali	116
Arcobacter butzleri RM4018	p3_r58557_Arc_butzl	58557
Arthroderma benhamiae CBS 112371	p3_r51431_Art_benha	51431
Arthroderma gypseum CBS 118893	p3_r61175_Art_gypse	61175
Arthroderma otae CBS 113480	p3_r49289_Art_otae	49289
Ashbya gossypii ATCC 10895	p3_r10623_Ash_gossy	10623
Aspergillus clavatus NRRL 1	p3_r18467_Asp_clava	18467
Aspergillus flavus NRRL3357	p3_r38227_Asp_flavu	38227
Aspergillus fumigatus A1163	p3_r18733_Asp_fumig	18733
Aspergillus fumigatus AF210	p3_r52783_Asp_fumig	52783
Aspergillus fumigatus Af293	p3_r14003_Asp_fumig	14003
Aspergillus nidulans FGSC A4	p3_r40559_Asp_nidul	40559
Aspergillus niger ATCC 1015	p3_r15785_Asp_niger	15785
Aspergillus niger CBS 513.88	p3_r19263_Asp_niger	19263
Aspergillus oryzae	p3_t5062_Asp_oryza_RIB40_NITE	5062
Aspergillus oryzae RIB40	p3_r28175_Asp_oryza	28175

Continued on next page

Table A.2 – continued from previous page

Organism	Pedant-DB	Accession
Aspergillus terreus NIH2624	p3_r17637_Asp_terre	17637
Bacillus amyloliquefaciens DSM 7	p3_r53535_Bac_amylo	53535
Bacillus anthracis str. Ames	p3_r57909_Bac_anthr	57909
Bacillus cereus ATCC 14579	p3_r57975_Bac_cereu	57975
Bacillus megaterium DSM319	p3_r48371_Bac_megat	48371
Bacillus subtilis subsp. subtilis str. 168	p3_r57675_Bac_subti	57675
Bacteroides fragilis YCH46	p3_r58195_Bac_fragi	58195
Batrachochytrium dendrobatidis JAM81	p3_r41157_Bat_dendr	41157
Batrachochytrium dendrobatidis JEL423	p3_p13653_Bat_dendr	13653
Baudoinia compniacensis UAMH 10762	p3_r53579_Bau_compn	53579
Beauveria bassiana ARSEF 2860	p3_r38719_Bea_bassi	38719
Bifidobacterium longum NCC2705	p3_r57939_Bif_longu	57939
Bipolaris maydis C5	p3_r42739_Bip_maydi	42739
Bipolaris sorokiniana ND90Pr	p3_r53923_Bip_sorok	53923
Bordetella pertussis Tohama I	p3_r57617_Bor_pertu	57617
Botryotinia fuckeliana	p3_i2_p16118_Bot_ciner_T4	16118
Botryotinia fuckeliana B05.10	p3_r20061_Bot_fucke	20061
Botryotinia fuckeliana BcDW1	p3_r188482_Bot_fucke	188482
Botryotinia fuckeliana T4	p3_r64593_Bot_fucke	64593
Brachybacterium faecium DSM 4810	p3_r58649_Bra_faeci	58649
Bradyrhizobium japonicum USDA 110	p3_r57599_Bra_diazo	57599
Brucella abortus biovar 1 str. 9-941	p3_r58019_Bru_abort	58019
Brucella abortus S19	p3_r58873_Bru_abort	58873
Brucella melitensis 16M	p3_r57735_Bru_melit	57735
Brucella suis 1330	p3_r57927_Bru_suis	57927
Buchnera aphidicola str. APS (Acyrthosiphon pisum)	p3_r57805_Buc_aphid	57805
Burkholderia ambifaria AMMD	p3_r58303_Bur_ambif	58303
Burkholderia mallei ATCC 23344	p3_r57725_Bur_malle	57725
Burkholderia multivorans ATCC 17616	p3_r58697_Bur_multi	58697
Burkholderia pseudomallei K96243	p3_r57733_Bur_pseud	57733
Campylobacter jejuni subsp. jejuni NCTC 11168	p3_r57587_Cam_jejun	57587
Candida albicans SC5314	p3_r14005_Can_albic	14005
Candida albicans WO-1	p3_r16373_Can_albic	16373
Candida dubliniensis CD36	p3_r38659_Can_dubli	38659
Candida glabrata CBS 138	p3_r12376_Can_glabr	12376
Candida tenuis ATCC 10573	p3_r33673_Can_tenui	33673
Candida tropicalis MYA-3404	p3_r39569_Can_tropi	39569
Caulobacter crescentus CB15	p3_r57891_Cau_cresc	57891
Ceriporiopsis subvermispora B	p3_r60419_Cer_subve	60419
Chaetomium globosum CBS 148.51	p3_r16821_Cha_globo	16821
Chaetomium thermophilum var. thermophilum DSM 1495	p3_r47065_Cha_therm	47065
Chlamydophila pneumoniae CWL029	p3_r57811_Chl_pneum	57811
Chlorobium phaeobacteroides DSM 266	p3_r58133_Chl_phaeo	58133
Claviceps purpurea 20.1	p3_p76493_Cla_purpu	76493
Clavispora lusitaniae ATCC 42720	p3_r41079_Cla_lusit	41079
Clostridium acetobutylicum ATCC 824	p3_r57677_Clo_aceto	57677
Clostridium difficile 630	p3_r57679_Clo_diffi	57679
Clostridium kluyveri DSM 555	p3_r58885_Clo_kluyv	58885
Clostridium perfringens str. 13	p3_r57681_Clo_perfr	57681
Clostridium saccharolyticum WM1	p3_r51419_Clo_sacch	51419
Clostridium thermocellum ATCC 27405	p3_r57917_Clo_therm	57917
Coccidioides immitis RS	p3_r16822_Coc_immit	16822
Coccidioides posadasii str. Silveira	p3_r17787_Coc_posad	17787
Cochliobolus heterostrophus	p3_t5016_Coc_heter	5016
Colletotrichum graminicola M1.001	p3_r37879_Glo_grami	37879
Colletotrichum higginsianum IMI 349063	p3_r47061_Col_higgi	47061
Colletotrichum orbiculare MAFF 240422	p3_r171217_Col_orbic	171217
Coniophora puteana RWD-64-598 SS2	p3_r53977_Con_putea	53977
Coniosporium apollinis CBS 100218	p3_r157061_Con_apoll	157061
Coprinopsis cinerea okayama7#130	p3_r29797_Cop_ciner	29797
Cordyceps militaris CM01	p3_r41129_Cor_milit	41129
Corynebacterium diphtheriae NCTC 13129	p3_r57691_Cor_dipht	57691
Corynebacterium glutamicum ATCC 13032	p3_r57905_Cor_gluta	57905
Corynebacterium jeikeium K411	p3_r58399_Cor_jeike	58399
Corynebacterium pseudotuberculosis FRC41	p3_r50585_Cor_pseud	50585
Cryptococcus bacillisporus WM276	p3_r62089_Cry_gatti	62089
Cryptococcus neoformans var. neoformans B-3501A	p3_r19119_Cry_neofo	19119
Cryptococcus neoformans var. neoformans JEC21	p3_r10698_Cry_neofo	10698
Dacryopinax sp. DJM-731 SS1	p3_r62091_Dac_DJM731	62091
Debaryomyces hansenii CBS767	p3_r12410_Deb_hanse	12410
Desulfitobacterium hafniense Y51	p3_r58605_Des_hafni	58605
Desulfovibrio vulgaris subsp. vulgaris str. Hildenborough	p3_r57645_Des_vulga	57645
Dichomitus squalens LYAD-421 SS1	p3_r53511_Dic_squal	53511
Dothistroma septosporum NZE10	p3_r74753_Dot_septo	74753
Edhazardia aedis USNM 41457	p3_r65125_Edh_aedis	65125
Edwardsiella tarda EIB202	p3_r41819_Edw_tarda	41819
Ehrlichia ruminantium str. Welgevonden	p3_r58013_Ehr_rumin	58013
Encephalitozoon cuniculi GB-M1	p3_r155_Enc_cunic	155
Enterobacter cloacae subsp. cloacae ATCC 13047	p3_r48363_Ent_cloac	48363
Enterobacter sakazakii ATCC BAA-894	p3_r58145_Cro_sakaz	58145
Enterococcus faecalis V583	p3_r57669_Ent_faeca	57669

Continued on next page

Table A.2 – continued from previous page

Organism	Pedant-DB	Accession
Enterocytozoon bieneusi H348	p3_r28163_Ent_biene	28163
Eremothecium cymbalariae DBVPG#7215	p3_r78153_Ere_cymba	78153
Erwinia amylovora ATCC 49946	p3_r46943_Erw_amylo	46943
Erwinia chrysanthemi str. 3937	p3_r52537_Dic_dadan	52537
Escherichia coli O157:H7 str. Sakai	p3_r57781_Esc_coli	57781
Escherichia coli str. K12 substr. MG1655	p3_r57779_Esc_coli	57779
Eutypa lata UCREL1	p3_r187490_Eut_lata	187490
Exophiala dermatitidis NIH/UT8656	p3_r64935_Exo_derma	64935
Fibroporia radiculosa TFFH 294	p3_r72357_Fib_radic	72357
Fomitiporia mediterranea MF3/22	p3_r56107_Fom_medit	56107
Francisella novicida U112	p3_r58499_Fra_novic	58499
Francisella tularensis subsp. holarctica LVS	p3_r58595_Fra_tular	58595
Francisella tularensis subsp. tularensis SCHU S4	p3_r57589_Fra_tular	57589
Fusarium fujikuroi IMI 58289	p3_p185772_Fus_fujik_v21	185772
Fusarium graminearum PH-1	p3_p13839_Fus_grami_v32	13839
Fusarium oxysporum CL57	p3_i2_t660032_Fus_oxysp_CL57	2147483418 *
Fusarium oxysporum Cotton	p3_i2_t909454_Fus_oxysp_Cotton	2147483417 *
Fusarium oxysporum f. sp. cubense race 1	p3_r174274_Fus_oxysp	174274
Fusarium oxysporum f. sp. cubense race 4	p3_r174275_Fus_oxysp	174275
Fusarium oxysporum f. sp. lycopersici 4286	p3_p18813_Fus_oxysp_FO2	18813
Fusarium oxysporum f. sp. melonis 26406	p3_i2_t1089452_Fus_oxysp_melonis_26406	73541
Fusarium oxysporum Fo47	p3_i2_t660027_Fus_oxysp_Fo47	2147483416 *
Fusarium oxysporum FOSC 3-a	p3_i2_t909455_Fus_oxysp_FOSC_3a	2147483422 *
Fusarium oxysporum HDV247	p3_i2_t909453_Fus_oxysp_HDV247	72771
Fusarium oxysporum II5	p3_i2_t660034_Fus_oxysp_II5	2147483415 *
Fusarium oxysporum MN25	p3_i2_t660028_Fus_oxysp_MN25	72769
Fusarium oxysporum PHW808	p3_i2_t660031_Fus_oxysp_PHW808	73543
Fusarium oxysporum PHW815	p3_i2_t660033_Fus_oxysp_PHW815	73545
Fusarium pseudograminearum CS3096	p3_r66583_Fus_pseud	66583
Fusarium solani f. pisi 77-13-4	p3_r51499_Nec_haema	51499
Fusarium verticillioides 7600	p3_p15553_Fus_verti_v31	15553
Fusobacterium nucleatum subsp. nucleatum ATCC 25586	p3_r57885_Fus_nucle	57885
Gaeumannomyces graminis var. tritici R3-111a-1	p3_r37931_Gae_grami	37931
Gardnerella vaginalis 409-05	p3_r43211_Gar_vagin	43211
Gardnerella vaginalis ATCC 14019	p3_r55487_Gar_vagin	55487
Geobacillus thermoglucosidasius C56-YS93	p3_r48129_Geo_therm	48129
Geobacter sulfurreducens PCA	p3_r57743_Geo_sulfu	57743
Geomyces destructans 20631-21	p3_r39257_Geo_destr	39257
Glarea lozoyensis 74030	p3_r74639_Gla_lozoy	74639
Gordonia bronchialis DSM 43247	p3_r41403_Gor_bronc	41403
Grosmannia clavigera kw1407	p3_r39837_Gro_clavi	39837
Haemophilus influenzae Rd KW20	p3_r57771_Hae_influ	57771
Haemophilus somnus 2336	p3_r57979_Hae_somnu	57979
Halobacterium sp. NRC-1	p3_r57769_Hal_NRC1	57769
Helicobacter pylori 26695	p3_r57787_Hel_pylor	57787
Kazachstania africana CBS 2517	p3_r178246_Kaz_afric	178246
Kazachstania naganishii CBS 8797	p3_r70969_Kaz_nagan	70969
Kluyveromyces lactis NRRL Y-1140	p3_r12377_Klu_lacti	12377
Kluyveromyces waltii NCYC 2644	p3_p10734_Klu_walti	10734
Komagataella pastoris CBS 7435	p3_r62483_Kom_pasto	62483
Laccaria bicolor	p3_t29883_Lac_bico	29883
Laccaria bicolor S238N-H82	p3_r29019_Lac_bicol	29019
Lachancea thermotolerans CBS 6340	p3_r39575_Lac_therm	39575
Lactobacillus acidophilus NCFM	p3_r57685_Lac_acido	57685
Lactobacillus casei ATCC 334	p3_r57985_Lac_casei	57985
Lactobacillus delbrueckii subsp. bulgaricus ATCC 11842	p3_r58647_Lac_delbr	58647
Lactobacillus fermentum IFO 3956	p3_r58865_Lac_ferme	58865
Lactobacillus helveticus DPC 4571	p3_r58761_Lac_helve	58761
Lactobacillus johnsonii NCC 533	p3_r58029_Lac_johns	58029
Lactobacillus reuteri DSM 20016	p3_r58471_Lac_reute	58471
Lactobacillus salivarius subsp. salivarius UCC118	p3_r58233_Lac_saliv	58233
Lactococcus lactis subsp. lactis Il1403	p3_r57671_Lac_lacti	57671
Legionella pneumophila subsp. pneumophila str. Philadelphia 1	p3_r57609_Leg_pneum	57609
Leptosphaeria maculans JN3	p3_r171003_Lep_macul	171003
Leptospira biflexa serovar Patoc strain Patoc 1 (Ames)	p3_r58511_Lep_bifle	58511
Leptospira borgpetersenii serovar Hardjo-bovis L550	p3_r58507_Lep_borgp	58507
Leptospira interrogans serovar Lai str. 56601	p3_r57881_Lep_inter	57881
Leuconostoc mesenteroides subsp. mesenteroides ATCC 8293	p3_r57919_Leu_mesen	57919
Listeria seeligeri serovar 1/2b str. SLCC3954	p3_r46215_Lis_seeli	46215
Lodderomyces elongisporus NRRL YB-4239	p3_r19611_Lod_elong	19611
Magnaporthe grisea 70-15	p3_r1433_Mag_oryza	1433
Malassezia globosa CBS 7966	p3_r27973_Mal_globo	27973
Marssonina brunnea f. sp. multigermtubi MB_m1	p3_r66127_Mar_brunn	66127
Melampsora larici-populina 98AG31	p3_r46711_Mel_laric	46711
Metarhizium acridum CQMa 102	p3_r38715_Met_acrid	38715
Metarhizium anisopliae ARSEF 23	p3_r38717_Met_aniso	38717
Methanococcus maripaludis S2	p3_r58035_Met_marip	58035
Methylobacterium extorquens PA1	p3_r58821_Met_extor	58821
Mixia osmundae IAM 14324	p3_r48573_Mix_osmun	48573
Mucor circinelloides f. circinelloides 1006PhL	p3_r172437_Muc_circi	172437
Myceliophthora thermophila ATCC 42464	p3_r79339_Myc_therm	79339

Continued on next page

Table A.2 – continued from previous page

Organism	Pedant-DB	Accession
Mycobacterium bovis AF2122/97	p3_r57695_Myc_bovis	57695
Mycobacterium leprae TN	p3_r57697_Myc_lepra	57697
Mycobacterium tuberculosis H37Rv	p3_r57777_Myc_tuber	57777
Mycoplasma capricolum subsp. capricolum ATCC 27343	p3_r58525_Myc_capri	58525
Mycoplasma gallisepticum str. R(low)	p3_r57993_Myc_galli	57993
Mycoplasma hyopneumoniae 232	p3_r58205_Myc_hyopn	58205
Mycoplasma mycoides subsp. mycoides SC str. PG1	p3_r58031_Myc_mycoi	58031
Mycoplasma pneumoniae M129	p3_r57709_Myc_pneum	57709
Mycosphaerella fijiensis CIRAD86	p3_r19049_Pse_fijie	19049
Mycosphaerella graminicola IPO323	p3_i2_p19047_Myc_grami	19047
Mycosphaerella graminicola IPO323	p3_r170847_Zym_triti	170847
Naumovozyma castellii CBS 4309	p3_r79343_Nau_caste	79343
Naumovozyma dairenensis CBS 421	p3_r79341_Nau_daire	79341
Neisseria gonorrhoeae FA 1090	p3_r57611_Nei_gonor	57611
Neisseria meningitidis MC58	p3_r57817_Nei_menin	57817
Nematocida parisii ERTm1	p3_r51843_Nem_paris	51843
Neofusicoccum parvum UCRNP2	p3_r187491_Neo_parvu	187491
Neosartorya fischeri NRRL 181	p3_r18475_Neo_fisch	18475
Neurospora crassa OR74A	p3_r132_Neu_crass	132
Neurospora crassa OR74A	p3_p13841_Neu_crass_MIPS	13841
Neurospora tetrasperma FGSC 2508	p3_r65273_Neu_tetra	65273
Nosema bombycis CQ1	p3_r30919_Nos_bomby	30919
Paenibacillus polymyxa E681	p3_r53477_Pae_polym	53477
Pantoea ananatis LMG 20103	p3_r46807_Pan_anana	46807
Paracoccidioides brasiliensis Pb01	p3_r48325_Par_lutzi	48325
Paracoccidioides brasiliensis Pb03	p3_p27779_Par_brasi_Pb03	27779
Paracoccidioides brasiliensis Pb18	p3_p28733_Par_brasi_Pb18	28733
Pasteurella multocida subsp. multocida str. Pm70	p3_r57627_Pas_multo	57627
Penicillium chrysogenum Wisconsin 54-1255	p3_r39879_Pen_chrys	39879
Phaeosphaeria nodorum SN15	p3_r21049_Pha_nodor	21049
Phanerochaete carnosa HHB-10118-sp	p3_r38425_Pha_carno	38425
Phanerochaete chrysosporium RP-78	p3_p135_Pha_chrys	135
Phycomyces blakesleeanus	p3_t4837_Phy_blake	4837
Pichia guilliermondii ATCC 6260	p3_r19593_Mey_guill	19593
Pichia pastoris GS115	p3_r39439_Pic_pasto	39439
Pichia stipitis CBS 6054	p3_r18881_Sch_stipi	18881
Piriformospora indica	p3_t65672_Pir_indic	65672
Podospora anserina DSM 980	p3_t5145_Pod_anser	5145
Podospora anserina DSM 980	p3_r29799_Pod_anser	29799
Porphyromonas gingivalis W83	p3_r57641_Por_gingi	57641
Postia placenta	p3_p19789_Pos_place	19789
Postia placenta Mad-698-R	p3_r38699_Pos_place	38699
Prochlorococcus marinus subsp. marinus str. CCMP1375	p3_r57995_Pro_marin	57995
Propionibacterium acnes KPA171202	p3_r58101_Pro_acnes	58101
Pseudomonas aeruginosa PAO1	p3_r57945_Pse_aerug	57945
Pseudomonas fluorescens Pf-5	p3_r57937_Pse_prote	57937
Pseudomonas mendocina ymp	p3_r58723_Pse_mendo	58723
Pseudomonas putida KT2440	p3_r57843_Pse_putid	57843
Pseudozyma antarctica T-34	p3_r186736_Pse_antar	186736
Pseudozyma hubeiensis SY62	p3_r203274_Pse_hubei	203274
Puccinia graminis f. sp. tritici CRL 75-36-700-3	p3_r66375_Puc_grami	66375
Punctularia strigosozonata HHB-11173 SS5	p3_r52407_Pun_strig	52407
Pyrenophora teres f. teres 0-1	p3_r66337_Pyr_teres	66337
Pyrenophora tritici-repentis Pt-1C-BFP	p3_r29813_Pyr_triti	29813
Pyrococcus furiosus DSM 3638	p3_r57873_Pyr_furio	57873
Ralstonia pickettii 12J	p3_r58737_Ral_picke	58737
Ralstonia solanacearum GMI1000	p3_r57593_Ral_solan	57593
Rhizobium etli CFN 42	p3_r58377_Rhi_etli	58377
Rhizobium leguminosarum bv. viciae 3841	p3_r57955_Rhi_legum	57955
Rhizopus oryzae RA 99-880	p3_r13066_Rhi_delem	13066
Rhodobacter sphaeroides 2.4.1	p3_r57653_Rho_sphae	57653
Rhodospirillum rubrum ATCC 11170	p3_r57655_Rho_rubru	57655
Rhodothermus marinus DSM 4252	p3_r41729_Rho_marin	41729
Rickettsia bellii RML369-C	p3_r58405_Ric_belli	58405
Rickettsia canadensis str. McKiel	p3_r58159_Ric_canad	58159
Rickettsia massiliae MTU5	p3_r58801_Ric_massi	58801
Rickettsia typhi str. Wilmington	p3_r58063_Ric_typhi	58063
Saccharomyces arboricola H-6	p3_r88533_Sac_arbor	88533
Saccharomyces bayanus	p3_t4931_Sac_bayan	4931
Saccharomyces castellii NRRL Y-12630	p3_t226301_Sac_caste_NRRL_Y12630	226301
Saccharomyces cerevisiae CEN.PK113-7D	p3_r52955_Sac_cerev	52955
Saccharomyces kluyveri NRRL Y-12651	p3_t226302_Sac_kluyv_NRRL_Y12651	226302
Saccharomyces kudriavzevii IFO 1802	p3_t226230_Sac_kudri_IFO_1802	226230
Saccharomyces mikatae IFO 1815	p3_t226126_Sac_mikat_IFO_1815	226126
Saccharomyces paradoxus NRRL Y-17217	p3_t226125_Sac_parad_NRRL_Y17217	226125
Salinibacter ruber DSM 13855	p3_r58513_Sal_ruber	58513
Salmonella typhimurium LT2	p3_r57799_Sal_enter	57799
Schizophyllum commune H4-8	p3_r51487_Sch_commu	51487
Schizosaccharomyces japonicus	p3_p13640_Sch_japon	13640
Schizosaccharomyces japonicus yFS275	p3_r32667_Sch_japon	32667
Schizosaccharomyces pombe	p3_r127_Sch_pombe	127

Continued on next page

Table A.2 – continued from previous page

Organism	Pedant-DB	Accession
Schizosaccharomyces pombe 972h-	p3_p13836_Sch_pombe	13836
Sclerotinia sclerotiorum 1980 UF-70	p3_r20263_Scl_scler	20263
Serpula lacrymans var. lacrymans S7.9	p3_r32885_Ser_lacry	32885
Setosphaeria turcica Et28A	p3_r82947_Set_turci	82947
Shewanella baltica OS155	p3_r58259_She_balti	58259
Shewanella putrefaciens CN-32	p3_r58267_She_putre	58267
Shigella boydii Sb227	p3_r58215_Shi_boydi	58215
Shigella sonnei Ss046	p3_r58217_Shi_sonne	58217
Sinorhizobium meliloti 1021	p3_r57603_Sin_melil	57603
Sordaria macrospora k-hell	p3_r51569_Sor_macro	51569
Spathaspora passalidarum NRRL Y-27907	p3_r53891_Spa_passa	53891
Sporobolomyces roseus IAM 13481	p3_t365493_Spo_roseu	365493
Staphylococcus aureus subsp. aureus N315	p3_r57837_Sta_aureu	57837
Staphylococcus epidermidis RP62A	p3_r57663_Sta_epide	57663
Staphylococcus lugdunensis HKU09-01	p3_r46233_Sta_lugdu	46233
Stereum hirsutum FP-91666 SS1	p3_r52843_Ste_hirsu	52843
Streptococcus gallolyticus UCN34	p3_r46061_Str_gallo	46061
Streptococcus mutans UA159	p3_r57947_Str_mutan	57947
Streptococcus parasanguinis ATCC 15912	p3_r49313_Str_paras	49313
Streptococcus pneumoniae TIGR4	p3_r57857_Str_pneum	57857
Streptococcus pyogenes M1 GAS	p3_r57845_Str_pyoge	57845
Streptococcus thermophilus CNRZ1066	p3_r58221_Str_therm	58221
Sulfolobus islandicus M.14.25	p3_r58849_Sul_islan	58849
Sulfolobus solfataricus P2	p3_r57721_Sul_solfa	57721
Synechococcus elongatus PCC 6301	p3_r58235_Syn_elong	58235
Synechocystis sp. PCC 6803	p3_r57659_Syn_PCC6803	57659
Talaromyces marneffei ATCC 18224	p3_r32665_Tal_marne	32665
Talaromyces stipitatus ATCC 10500	p3_r38857_Tal_stipi	38857
Tetrapisispora blattae CBS 6284	p3_r188088_Tet_blatt	188088
Tetrapisispora phaffii CBS 4417	p3_r79335_Tet_phaff	79335
Thermus thermophilus HB27	p3_r58033_The_therm	58033
Thielavia terrestris NRRL 8126	p3_r79337_Thi_terre	79337
Togninia minima UCRPA7	p3_r188116_Tog_minim	188116
Torulaspora delbrueckii CBS 1146	p3_r79345_Tor_delbr	79345
Trachipleistophora hominis	p3_r84343_Tra_homin	84343
Trametes versicolor FP-101664 SS1	p3_r56097_Tra_versi	56097
Tremella mesenterica DSM 1558	p3_r32829_Tre_mesen	32829
Treponema pallidum subsp. pallidum str. Nichols	p3_r57585_Tre_palli	57585
Trichoderma atroviride IMI 206040	p3_r19867_Tri_atrov	19867
Trichoderma reesei QM6a	p3_r15571_Tri_reese	15571
Trichoderma virens Gv29-8	p3_r19983_Tri_viren	19983
Trichophyton equinum CBS 127.97	p3_r20577_Tri_equin	20577
Trichophyton rubrum CBS 118892	p3_r65025_Tri_rubru	65025
Trichophyton tonsurans CBS 112818	p3_r38223_Tri_tonsu	38223
Trichophyton verrucosum HKI 0517	p3_r51485_Tri_verru	51485
Tropheryma whipplei str. Twist	p3_r57705_Tro_whipp	57705
Tuber melanosporum Mel28	p3_r49017_Tub_melan	49017
Uncinocarpus reesii 1704	p3_r39807_Unc_reesi	39807
Ureaplasma parvum serovar 3 str. ATCC 700970	p3_r57711_Ure_parvu	57711
Vanderwaltozyma polyspora DSM 70294	p3_r20539_Van_polys	20539
Verticillium alfalfae VaMs.102	p3_r51263_Ver_albo	51263
Verticillium dahliae VdLs.17	p3_r28529_Ver_dahli	28529
Vibrio cholerae O1 biovar eltor str. N16961	p3_r57623_Vib_chole	57623
Vibrio cholerae O395	p3_r58425_Vib_chole	58425
Vibrio fischeri ES114	p3_r58163_Vib_fisch	58163
Wallemia ichthyophaga EXF-994	p3_r193177_Wal_ichth	193177
Wallemia sebi CBS 633.66	p3_r64975_Wal_sebi	64975
Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis	p3_r57853_Wig_gloss	57853
Xanthomonas campestris pv. campestris str. ATCC 33913	p3_r57887_Xan_campe	57887
Xanthomonas oryzae pv. oryzae KACC10331	p3_r58155_Xan_oryza	58155
Xylella fastidiosa 9a5c	p3_r57849_Xyl_fasti	57849
Yarrowia lipolytica CLIB122	p3_r12414_Yar_lipol	12414
Yersinia enterocolitica subsp. enterocolitica 8081	p3_r57741_Yer_enter	57741
Yersinia pestis CO92	p3_r57621_Yer_pesti	57621
Yersinia pseudotuberculosis IP 32953	p3_r58157_Yer_pseud	58157
Zygosaccharomyces rouxii CBS 732	p3_r39573_Zyg_rouxi	39573
Zymomonas mobilis subsp. mobilis ATCC 10988	p3_r55403_Zym_mobil	55403

Cluster_ID	Gene range	Size	Signature enzymes	Tailoring enzymes	Diff_Up	Diff_Down	Motifs
PKS20	FFUJ_12705 - FFUJ_12717	13	PKS-NRPS (1)	Methyl (1) , P450 (4)		A1_wt.100NO3	CACTCCGGTC (10)
PKS17-PKS18	FFUJ_12060 - FFUJ_12079	20	NPS-like (1), PKS (2)	Methyl (3) , Oxido (2) , P450 (1)			AAGAGATAAA (18
gibberellic acid [31]	FFUJ_14337 - FFUJ_14331	7	TPS (2)	P450 (4)		A3_wt.100GLN, A3_sge1.10GLN	CCCAGGGGTC (6)
STC5	FFUJ_11736 - FFUJ_11743	8	NPS (1), TPS (1)	P450 (3)			
FF_C26	FFUJ_14692 - FFUJ_14698	7	PKS-NRPS (1)	Methyl (1) , Oxido (1) , P450 (1)			AGATCTGCACTC (5)
NPS25	FFUJ_05344 - FFUJ_05351	8	NPS (1)	Methyl (1) , Oxido (1) , P450 (2)			GCGGGCCGAA (8)
fumonisin [173, 171]	FFUJ_09254 - FFUJ_09240	15	NPS (1), NPS-like (2), PKS (1)	Methyl (1) , Oxido (1) , P450 (3)	A3_sge1.10GLN	A3_wt.100GLN	GTATCCGA (15), TATCCGAGA (13)
FF_C81	FFUJ_10894 - FFUJ_10903	10	NPS (1)	Oxido (3)			CCTGAATAAA (9)
STC2	FFUJ_01365 - FFUJ_00969	4	TPS (1)				
α-acorenol [29]	FFUJ_10346 - FFUJ_10353	8	PKS (1), TPS (1)	Methyl (1)			
NPS03	FFUJ_02112 - FFUJ_02166	10	NPS (1), NPS-like (1)		A1_wt.100NO3, A3_wt.100GLN	A3_sge1.100GLN	
fusarin C [155]	FFUJ_10058 - FFUJ_10050	9	PKS-NRPS (1)	Methyl (1) , P450 (1)	A1_wt.100NO3, A3_wt.100GLN		
NPS13	FFUJ_02436 - FFUJ_02440	5	NPS-like (1)			A1_wt.100NO3	TGGGNTCGAACT-CACG (4), GGGATCGAACTCA (4)
PKS02	FFUJ_00128 - FFUJ_00115	14	NPS (1), PKS (1)	Methyl (1)			CTATTCTCGT (12)
PKS19	FFUJ_12239 - FFUJ_12244	6	PKS (1)	Methyl (1) , P450 (1)			
FF_C44	FFUJ_05796 - FFUJ_05806	11	NPS-like (1)	P450 (2)			
STC01	FFUJ_00049 - FFUJ_00027	23	TPS (1)	P450 (2)			
apicidin F [231]	FFUJ_00013 - FFUJ_00003	11	NPS (1)	Methyl (1) , P450 (2)	A3_wt.100GLN	A3_sge1.100GLN	AT?TCACGTCA (9)
PKS01	FFUJ_02213 - FFUJ_02228	16	PKS-NRPS (1)	Methyl (1)		A1_wt.100NO3	
PKS13	FFUJ_12015 - FFUJ_12027	13	PKS (1), TPS (1)	Methyl (1)			
PKS07	FFUJ_06259 - FFUJ_06263	5	PKS (1)	Methyl (1)			
NPS11	FFUJ_10934 - FFUJ_10938	5	NPS-like (1)	Oxido (1)	A1_wt.100NO3		
NPS20	FFUJ_06716 - FFUJ_06723	8	NPS (1)				
STC03	FFUJ_04072 - FFUJ_04062	11	Anti (1), TPS (1)	P450 (1)			
NPS23	FFUJ_12005 - FFUJ_12010	6	NPS (1), NPS-like (1)	P450 (2)		A1_wt.100NO3	
fusarubin [203]	FFUJ_03989 - FFUJ_03984	6	PKS (1)	Methyl (1)		A1_wt.100NO3	TCGGACTCCG (6)
bikaverin [232]	FFUJ_06742 - FFUJ_06747	6	PKS (1)	Methyl (1)		A3_sge1.100GLN	
FF_C30	FFUJ_14829 - FFUJ_14834	6	TPS (1)	P450 (1)			
DMATS1	FFUJ_09179 - FFUJ_09173	7	DMATS (1)	P450 (1)			
FF_C47	FFUJ_06645 - FFUJ_06649	5	Anti (1)			A1_wt.100NO3	
FF_C21	FFUJ_03506 - FFUJ_03510	5	NPS (1)		A1_wt.100NO3		
FF_C22	FFUJ_03598 - FFUJ_03602	5	NPS (1)			A1_wt.100NO3	
NPS2	FFUJ_04614 - FFUJ_04610	5	NPS (1)		A1_wt.100NO3		
beauvericin [128]	FFUJ_09298 - FFUJ_09294	5	NPS (1)				CCGTGTCTCGGT (5)
PKS16	FFUJ_11198 - FFUJ_11201	4	PKS (1)	Methyl (1)		A1_wt.100NO3	
neurosporaxanthine [181]	FFUJ_11800 - FFUJ_11805	6	TPS (1)				
NPS6	FFUJ_10732 - FFUJ_10736	5	NPS (1)	Acyl (1)	A1_wt.100NO3, A3_wt.100GLN	A3_sge1.100GLN	
NPS4	FFUJ_08120 - FFUJ_08113	8	NPS (1)				
FF_C23	FFUJ_03624 - FFUJ_03627	4		P450 (1)			
FF_C79	FFUJ_10711 - FFUJ_10725	15	NPS (1)	Oxido (1)			CCGAACCGAA (12)
FF_C42	FFUJ_05864 - FFUJ_05868	5	PKSIII (1)	P450 (1)			
FF_C24	FFUJ_14588 - FFUJ_14590	3	NPS (1)				
NPS21	FFUJ_02028 - FFUJ_02017	12	NPS (1)				AGACGGGGAC (10)
NPS17	FFUJ_03641 - FFUJ_03645	4	NPS (1)				
PKS14	FFUJ_11023 - FFUJ_11036	13	PKS (1)	P450 (1)		A1_wt.100NO3	AGCAATGTCC (12)
fusaric acid [31]	FFUJ_02105 - FFUJ_02109	5	PKS (1)		A1_wt.100NO3, A3_wt.100GLN	A3_sge1.100GLN	
FF_C46	FFUJ_06605 - FFUJ_06611	7		Oxido (1)		A1_wt.100NO3	
NPS12	FFUJ_14786 - FFUJ_14791	6	NPS-like (1)	P450 (1)			TCAGATGCATAA (5)
FF_C100	FFUJ_12090 - FFUJ_12093	4	PKS-like (1)				

Table A.3: Predicted gene clusters and clusters of known metabolites in *F. fujikuroi*