

Evolution and diversity of biosynthetic gene clusters in *Fusarium*

Koen Hoogendoorn

Registration number: 941124359110

Course code: BIF-80336

Supervisors: Marnix Medema, Theo van der Lee

Abstract

Plant pathogenic fungi in the *Fusarium* genus cause severe damage to crops resulting in great financial losses and health hazards. Secondary metabolites synthesized by these fungi are known to be involved in the infections of the host plant, and might give insight into the pathogenicity of a species. Recent developments in the field of bioinformatics allow for the detection of gene clusters involved in the production of these secondary metabolites using annotated genomes. Here, we demonstrate that through the combination of several bioinformatics tools, the evolutionary history and diversity of secondary metabolites can be reconstructed. We validated this approach using previously identified gene clusters and identified several new gene clusters that may contribute to *Fusarium* pathogenicity or host specificity. Our analysis indicates that additional genomic data, especially on non-pathogenic *Fusarium* species would allow more elaborate and more accurate predictions of which BGCs are involved in pathogenicity.

Introduction

The *Fusarium* genus is an extensive fungal genus of ascomycetes consisting of mostly saprotrophic soil-borne species. However, some *Fusarium* species are important plant pathogens. Two *Fusarium* species can be found in the top five of plant pathogens (Dean *et al.*, 2012). In addition, some *Fusarium* species cause animal and human infections (Nucci and Anaissie, 2007). So while most of the species in the genus are harmless, the ones that actually show pathogenic abilities cause alarming amounts of financial as well as medical damage. Infection of wheat and barley by *Fusarium graminearum* for example causes

the disease fusarium ear blight, also known as fusarium head blight (Windels, 2000; McMullen and Stack, 2011). This disease eventually causes kernels to shrivel. When infected kernels are used as seeds for new generations of wheat plants, growth is severely hindered, and final production gravely decreased. Reports have shown that, although difficult to determine exact sums, financial damage from *F. graminearum* in the U.S. alone is easily estimated over tens of millions of dollars *per annum* (Windels, 2000; Wu, 2007). Apart from the decreased kernel quality, these kernels often contain high amounts of mycotoxins. When consumed, these mycotoxins cause serious health hazards in

humans and other animals ranging from irritations to death depending on the types and concentrations of mycotoxins

Mycotoxins, are secondary metabolites produced by fungi. The secondary metabolism, also called specialized metabolism, is part of the metabolism of fungi which is not deemed essential for direct survival. Its functions include, but are not limited to host infection, interspecies competition and defense against predators. The complexity of secondary metabolites, synthesized by the secondary metabolism, can vary greatly (**Figure 1**). To achieve the production of such complex molecules several proteins cooperate, each slightly modifying an existing backbone to form the final product. In the secondary metabolite hyoscyine hydrobromide, produced by plants in the genus of the *Solanaceae*, some

elements of the amino acid phenylalanine, although heavily modified, can still be recognized (O'Hagan, 2000). In vancomycin, produced by *Amycolatopsis orientalis*, the backbone consists of seven amino acids which are all heavily modified and crosslinked to form the final product (Samel *et al.*, 2008). The complexity of these compounds makes it nearly impossible for them to be synthesized by a single gene; therefore, many genes are often involved in the production of a secondary metabolite.

Genes encoding the production for such a metabolite are often located in close proximity to each other, forming a secondary metabolite gene cluster (Keller *et al.*, 2005). Along with genes responsible for the biosynthesis of the secondary metabolite, genes with regulatory and transport functions are usually also present

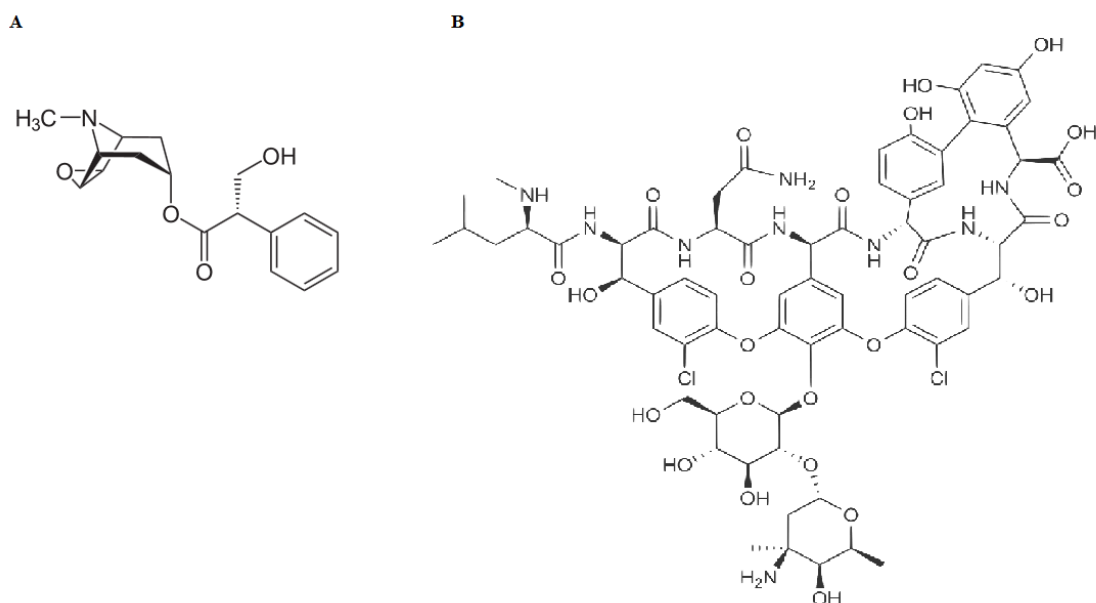


Figure 1. Varying complexity of secondary metabolites with A) the relatively simple alkaloid Hyoscyine hydrobromide and B) the complex nonribosomal peptide Vancomycin.

in these clusters (Brown and Proctor, 2016). Identifying orthologous gene clusters, especially when the product of the gene cluster is not known, has shown to be a difficult task. Since most secondary metabolites are synthesized from amino-acids, isoprene units or derivatives from malonic acid, the protein domains involved in their modifications share high sequence similarity across gene clusters. This makes it possible to identify gene clusters based on domain presence (Medema *et al.*, 2011). The modular nature of clusters causes issues when distinguishing gene clusters that produce compounds that are similar in structure due to high sequence and domain content similarity (Wolf *et al.*, 2015). To successfully identify orthologous gene clusters, BiG-SCAPE was developed (Yeong, 2016). By combining three different distance indices, a final distance metric is created, which can be used to calculate pairwise distances of all biosynthetic gene clusters. Another bio-informatics based strategy relies on the detection of co-expressed (and therefore co-regulated) genes: As the production of a secondary metabolite requires all of the genes in the cluster to be active at the same time, a method to identify gene clusters based on shared regulatory motifs was devised (Wolf *et al.*, 2015). Guided by these bioinformatics-based approaches, discovery of BGCs has become more accessible to researchers.

Whereas previously, time-consuming and expensive mutation studies based on mutant library screening would have to be conducted to identify genes involved in the

production of a compound, nowadays the structurally annotated genome is sufficient to identify clusters. Targeted functional analysis studies, guided by these predictions, to validate the identified clusters are more efficient and cheaper than the mutant library screening method. Also, with the vast amount of genomic data that is available now, secondary metabolites that are produced under specific conditions, and are generally not found during mutant library screening, can be identified. With clusters now identifiable with the push of a button, it is interesting to see how these clusters developed over time, to find out whether their biological function corresponds with ecological events throughout history. Several tools and methods are available that are suited for this so-called ancestral state reconstruction.

The goal of this project was (i) to devise a pipeline concatenating existing bioinformatics tools to identify and organize BGCs, (ii) to validate this pipeline with known BGCs, (iii) to identify all BGCs in a set of well annotated *Fusarium* species, (iv) to trace the history of secondary metabolite gene clusters in several *Fusarium* species, and (v) to identify which BGCs are most likely to be important in host specificity or pathogenicity. Knowing which clusters contribute to pathogenicity would allow for a specifically targeted way of battling these pathogens through means of genetic manipulation or by means of specific fungicides that are capable of inhibiting the pathogen, thereby increasing yields and reducing health hazards worldwide.

Results

Assembly and annotation quality

To assess the covered gene space of the genomes and their corresponding structural annotations that were used

the assembly is the presence of certain gene clusters in a corresponding genome annotation. In *Fusarium* species, three gene clusters are considered to be conserved in all species of the genus (Wiemann *et al.*, 2013). For all species in the analysis, all three of these BGCs were identified.

Table 1. Summary of assembly statistics containing N50 and BUSCO scores along with general genome statistics.

Species	Genome size (bp)	Scaffolds	Scaffold N50 (bp)	Contigs	Contig N50	BUSCO Score (Complete/Fragmented/Missing)
<i>F. graminearum</i>	36,667,552	199	8,791,613	424	258,133	1415/22/1
<i>F. culmorum</i>	37,688,228	207	8,831,140	2,274	39,999	1418/17/3
<i>F. pseudograminearum</i>	36,973,259	281	8,840,934	685	186,303	1427/10/1
<i>F. poae</i>	46,476,831	181	8,783,590	182	9,631,096	1425/11/2
<i>F. fujikuroi</i>	43,832,314	12	4,234,805	65	1,182,607	1419/18/1
<i>F. verticillioides</i>	41,885,085	39	1,959,799	213	392,397	1422/13/3
<i>F. oxysporum</i> Fo47	49,664,628	124	3,884,136	419	762,152	1423/13/2
<i>F. oxysporum</i> f. sp. <i>lycopersici</i>	61,471,697	117	1,976,106	1,371	95,416	1396/33/9

during this project, BUSCO was used (Simão *et al.*, 2015). All genomes had scores of at least 97% complete genes found (Error! Reference source not found.).

One gene (BUSCOfEOG7MH16D) in the BUSCO set, which encodes a cytosolic dynein heavy chain, was not detected in any of the tested genomes, which indicates this gene could be absent in *Fusarium* species in general. BUSCO scores were determined, and assembly statistics were retrieved from the NCBI database. Although neither the BUSCO score nor N50 statistics can give a definitive representation of the quality of an assembly, they do give an indication of its quality. For each assembly both scores were assessed, and deemed sufficient to continue the analysis with this dataset. Another indication of the completeness of

Host specificity

For every species, the main host was determined from literature (Alabouvette *et al.*, 1993; Benfield and Gardiner, 2015; Cuomo *et al.*, 2007; Gardiner *et al.*, 2014; Ma *et al.*, 2010; Moolhuijzen *et al.*, 2013; Wiemann *et al.*, 2013). When no main host could be selected, the group of host plants the pathogen infected was chosen as “host” instead. Six out of eight species used in the analysis are known to infect cereals. The other two, both *Fusarium oxysporum* strains have different hosts, where *F.oxysporum* f.sp. *lycopersici* infects tomato and *F.oxysporum* Fo47 is a biological control strain that, while also infecting tomato, does not cause symptoms of infection, and

protects against infection by other *Fusarium* species (Alabouvette *et al.*, 1993).

antiSMASH

In the eight genomes, a total of 392 biosynthetic gene clusters were found, averaging 49 clusters per genome. Previous studies on fungi found similar numbers of gene clusters per genome (**Figure 2A**) (Wiemann *et al.*, 2013). When compared to literature, clusters predicted by antiSMASH often came close to the experimentally established gene clusters, but almost never completely matched, and predictions generally differed in length in the order of a few genes. AntiSMASH often extended the BGC because of its ‘greedy’ approach (Medema *et al.*, 2011). This is exemplified by the fusarubin gene cluster in *Fusarium fujikuroi*. This cluster consists of six genes (FFUJ_03984 to FFUJ_03989) (Studt *et al.*, 2012). AntiSMASH however, predicts the fusarubin cluster to run from FFUJ_03975 to FFUJ_03993, resulting in a difference of prediction of nine genes upstream and four genes downstream of the cluster. It is worth noting that in the set of experimentally established gene clusters, antiSMASH never missed genes that were part of the gene cluster, and only added genes that were not part of the cluster.

The most common type of gene cluster found in the dataset, and perhaps one of the most studied types in general, is the non-ribosomal peptide synthase cluster, and although the terpenes come in as a close second, the third most common type

in this analysis is the “other” class (**Figure 2B-C**). These “other”-type classes are expected consist mostly of discrete nrps systems containing at least one A-T domain (M.H. Medema, personal communication, July 26, 2016). This once again reveals that much about the secondary metabolism is still unknown. Hybrid clusters occur less often in the dataset, although their actual numbers may be higher than is shown here due to “mispredictions” from antiSMASH. Especially when detecting larger hybrid clusters, the distance between the two main biosynthetic genes can be too large for antiSMASH to detect them as one cluster, and will instead detect two clusters of different types which often overlap at one end. However, these “mispredictions” could easily be identified by comparing the end coordinates one cluster with the start coordinates from the next. Furthermore, several clusters were detected on the supernumerary chromosomes of *F. poae*. Because of the fragmentary assembly of these chromosomes, these clusters however only consisted of a few genes, often spanning the entire contig. The supernumerary chromosomes of *F. poae*. Possible biosynthetic gene clusters could have ended up being fragmented over multiple contigs, with their associated domains spread across these contigs as well. While it is likely that some clusters are present on these supernumerary chromosomes, it is not possible to identify these clusters in their entirety until a better assembly becomes available.

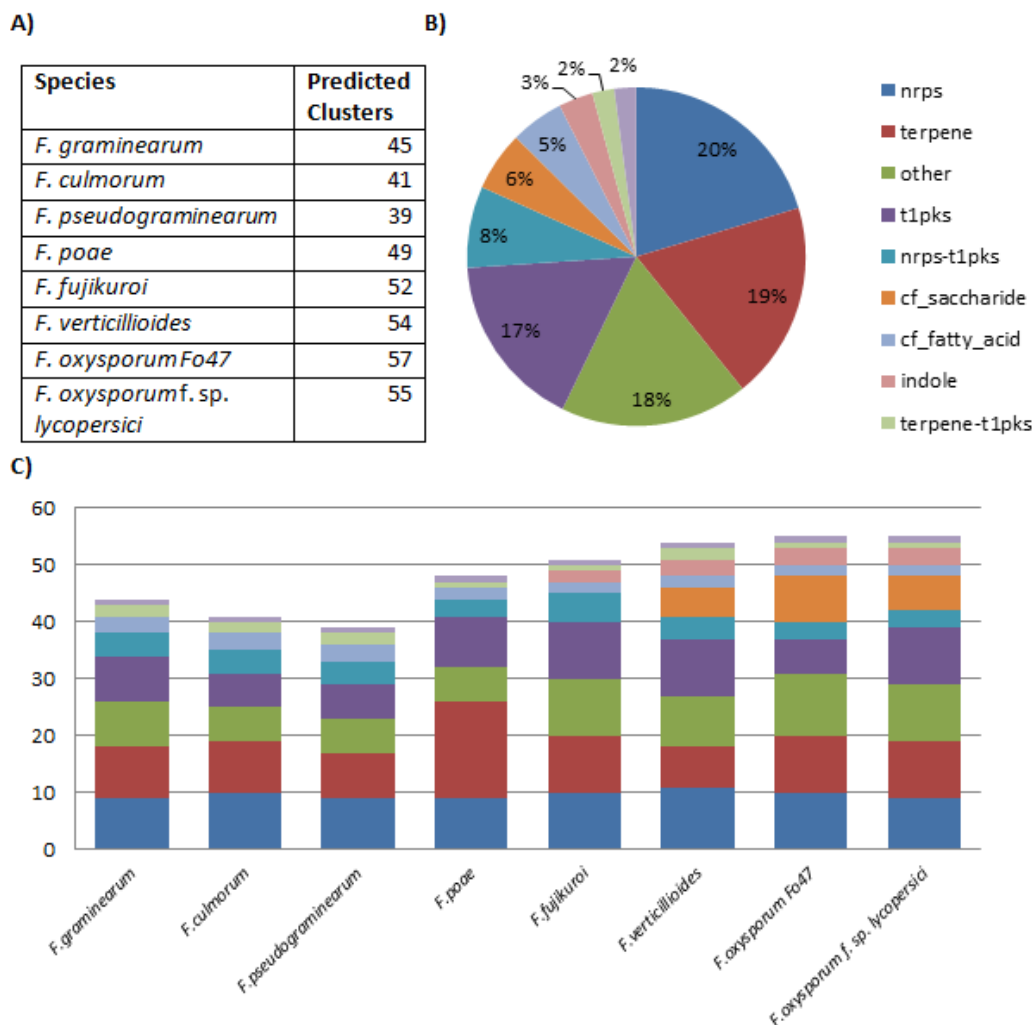


Figure 2. A) Predicted number of clusters and distribution of the ten most common types of biosynthetic gene clusters B) summarized (n=334) showing that while the nrps, pks and terpenes make up the largest part of the clusters, the amount of “other-type” BGCs is still the third most common type of cluster. C) Per species composition of the ten most common BGC types.

Cassis vs. antiSMASH

In order to validate BGCs predicted by antiSMASH, a comparison was made with a different tool that is used to predict biosynthetic gene clusters; CASSIS (Wolf *et al.*, 2015). While similar results were obtained by the two different tools, CASSIS generally predicts clusters to be smaller, trying to prevent prediction of extra genes at the cost of sometimes missing genes that are supposed to be included in a cluster (Figure 3)(T. Wolf, personal communication,

matching prediction was observed only 4/608 (0.6%) of times, while matching predictions of either start or end positions were more common, being 48/628 (7.6%) and 62/610 (10.2%) respectively. Of 155 of the 786 (19.7%) anchor genes for which clusters were detected by antiSMASH no cluster was detected by CASSIS. This indicates that regulation of the expression of a BGC may be more complex than CASSIS anticipates. CASSIS and antiSMASH rarely predicted the cluster to start and stop at

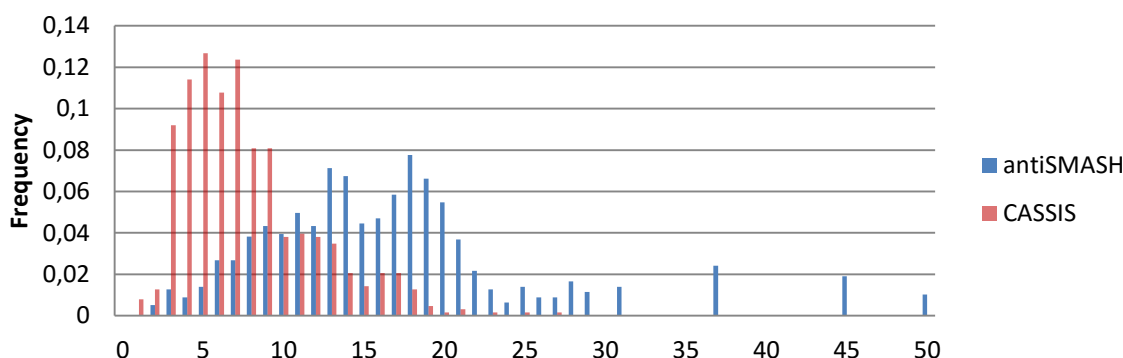


Figure 3. Distribution of secondary metabolite gene cluster sizes predicted by both antiSMASH (n=786) and CASSIS (n=631). Sample size varies between tools since CASSIS did not always predict a cluster where antiSMASH did.

March 31, 2016). AntiSMASH does the opposite, predicting extra genes to be part of the cluster to make sure no genes are excluded that should be in the cluster. For every anchor gene, start and end genes of the predicted associated BGC were compared, and the difference between the two summarized (Table 2). A perfectly

the same gene, and predictions by both tools still differed from experimentally established clusters. To ensure potentially important data was not lost through the use of CASSIS, antiSMASH clusters were chosen to conduct further analyses on.

Table 2. Summarized differences in prediction between antiSMASH and CASSIS.

Raw data is available on https://git.wageningenur.nl/hoooge096/MSc_Thesis/

difference in prediction	0	1-3	4-6	7-9	10-12	13-15	16+
Start (n=628)	48	228	171	109	44	12	16
Stop (n=610)	62	222	130	109	39	38	10
Both (n=608)	4	74	109	124	108	93	96

BiG-SCAPE & Multigeneblast

Gene clusters generated by BiG-SCAPE were visualized in Cytoscape, where every gene cluster was represented by a node, and significantly similar nodes were represented by edges between the two nodes (**Figure 4**) (Smoot *et al.*, 2011). A raw distance threshold of 0.7 was chosen after testing 0.1 step increments in the range 0.6-

0.9. With this threshold, all except one connected component in the network contained a maximum number of nodes equal to the amount of species used for the analysis. Apart from a special case with BGCs on the supernumerary chromosomes of *F. poae*, connected components contained a maximum of one node per species, indicating that whole cluster duplications were generally absent.

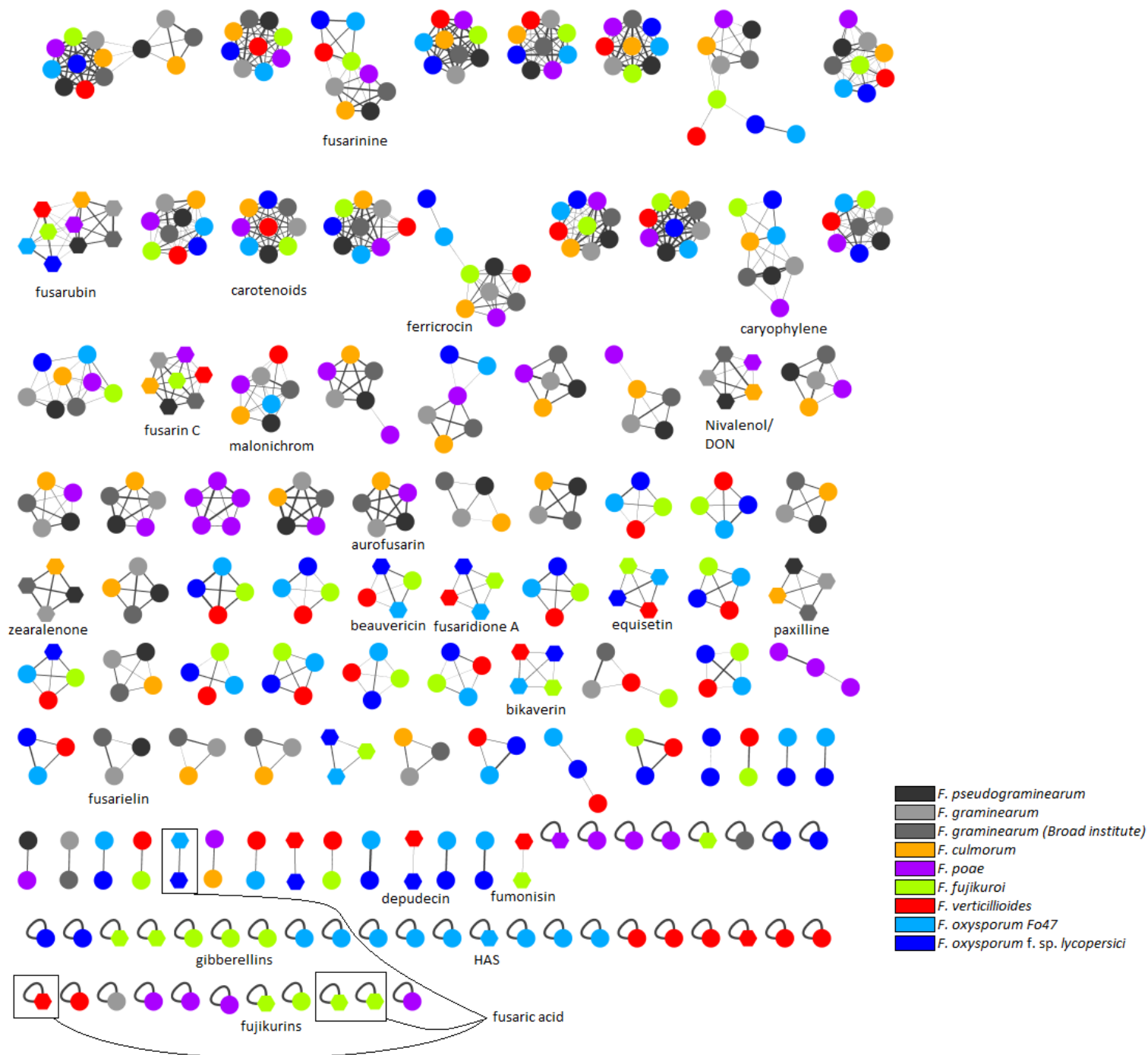


Figure 4. Overview of the BGC network generated by BiG-SCAPE. Hexagonal nodes have accessions in the MIBiG database. BGCs with known compounds used in following analyses have been marked down in the image. Source file including raw data can be downloaded at https://git.wageningenur.nl/hooge096/MSc_Thesis/. Visualisation generated in Cytoscape v3.4.0

Two distinct subgroups of smaller order clusters could be identified; The ones that were present in *F. graminearum*, *F. pseudograminearum*, *F. culmorum* and *F. poae* but absent in the other species, and the ones which were present in the exact opposite manner. Increasing the raw distance threshold did not combine these types of clusters, indicating that there is a considerable diversity in biosynthetic gene cluster content within a species. Even within higher order clusters, this split was often clearly visible by the difference in raw distance between and within the two subgroups (**Figure 5A**). The BGC producing fusaric acid can be observed to be spread across several independent nodes. While all

the species containing the BGC can be correctly identified, the corresponding nodes were not considered sufficiently homologous by BiG-SCAPE (Niehaus *et al.*, 2014).

An extra step was taken to validate the network produced by BiG-SCAPE. Multigeneblast runs were conducted on all biosynthetic gene clusters, with a database consisting of all the biosynthetic gene clusters across all samples (Medema *et al.*, 2013). While the method is less sophisticated than the analysis conducted by BiG-SCAPE, it provided an extra level of visual information that could be manually checked since Multigeneblast matches are returned as images showing homology to

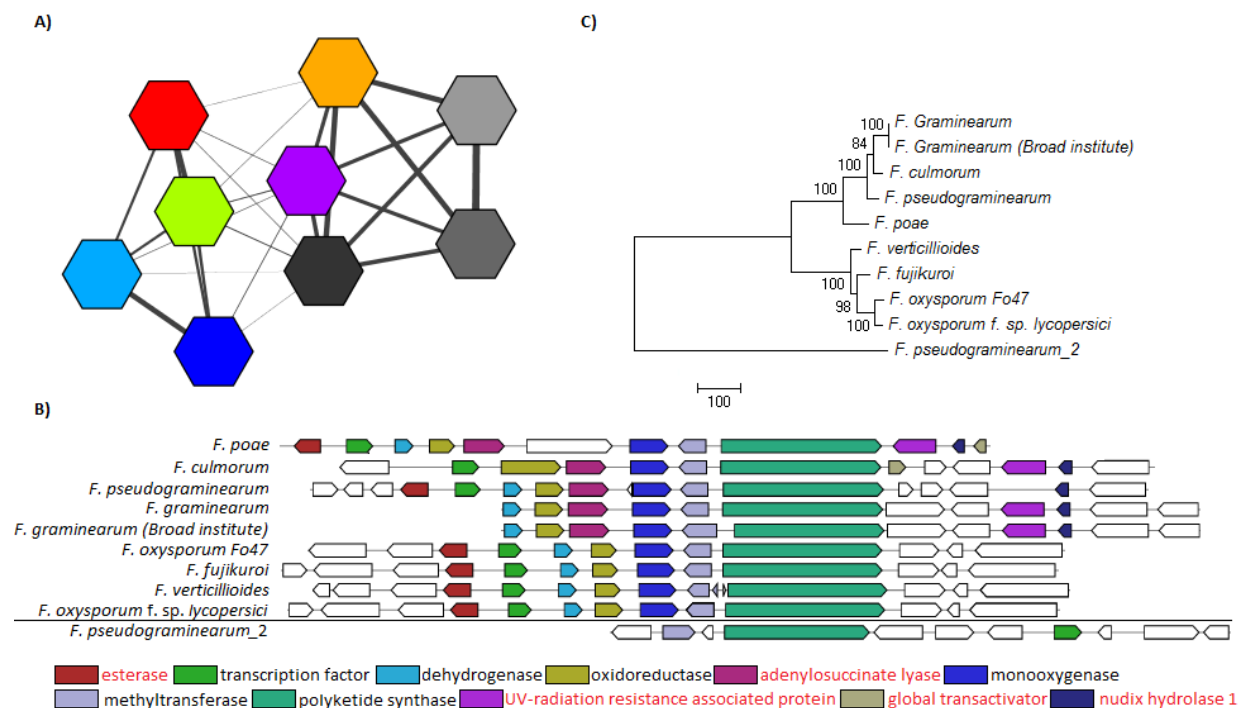


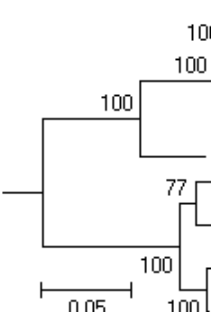
Figure 5. The fusarubin BGC as visualized by A) BiG-SCAPE. Colors correspond with Figure 4. A smaller line width indicates a larger raw distance between two nodes. B) Multigeneblast. The next most homologous BGC is indicated on the bottom. Genes in red are not known to be associated with fusarubin production. C) Phylogeny based on the polyketide synthase gene shows a large distance between *F. pseudograminearum_2* indicating that its BGC is indeed not involved in fusarubin production. Bootstrap consensus percentages are shown at nodes in the tree.

the user in descending order of identity (**Figure 5B**). From these images, it was apparent when a hit was no longer the same cluster as the query. The homology percentage dropped rapidly at these hits, as opposed to those hits that were still homologous to the query (**Figure 5C**). All clusters were checked, and corrected if necessary. To find out where clusters arose or disappeared during evolution of the *Fusarium* genus, gene clusters with experimentally established products along with the nodes present in the same cluster as these gene clusters were selected for ancestral state reconstruction. Five of these clusters, HAS, depudecin, bikaverin, beauvericin and paxillin were inferred from orthology with other taxa according to their entries in the MIBiG database. By creating a presence/absence matrix of the biosynthetic gene clusters and species, losses and duplications could be visualized.

Ancestral state reconstruction

Ancestral state reconstruction, the inference of mutations, genome duplications, BGC births and deaths in the past from genomic data obtained from organisms in the present has been an essential technique to increase our understanding of the evolutionary history of a species. Reconstructing the ancestral states of biosynthetic gene clusters gives insight into the development of clusters through time. Gains and losses of clusters were inferred using a maximum parsimony approach. Maximum parsimony reconstructions are based on the assumption that evolution will take the “easiest” route, with the minimal amount of change throughout history. The RBP2 gene sequence was used to construct a phylogenetic tree, which matched a previous more extensive phylogeny (O'Donnell *et al.*, 2010). Using this phylogeny along with the presence/absence

Table 3. Binary matrix containing information on the presence or absence of secondary metabolite gene clusters with known products in several *Fusarium* species. Color of the products correspond to various secondary metabolite types indicated above these colors.

		nrps-indole					tlpks			nrps-tlpks				nrps			terpene			terpene-tlpks				
		HAS	Depudecin	Fujikurins	Bikaverin	Fusarinic Acid	Fusarinol	Aurofusarin	Fumonisin	Equisetin	Zearalenone	Fusarin C	Beauvericin	Fusaridinone A	Malonic brom	Ferrirocin	Fusarinine	Nivalenol/DON	Gibberellins	Fusarubin	Caryophyllols	Paxillin	pyrrolizidine	
	100	<i>F. graminearum</i>	0	0	0	0	0	1	1	0	0	1	1	0	0	1	1	1	1	0	1	1	1	1
	100	<i>F. culmorum</i>	0	0	0	0	0	0	1	0	0	1	1	0	0	1	1	1	1	0	1	1	1	1
	100	<i>F. pseudograminearum</i>	0	0	0	0	0	1	1	0	0	1	1	0	0	1	1	1	1	0	1	1	1	1
		<i>F. poae</i>	0	0	0	0	0	0	1	0	0	0	1	0	0	1	1	1	1	0	1	1	1	0
	77	<i>F. fujikuroi</i>	0	0	1	1	1	0	0	1	1	0	1	1	1	0	1	1	1	0	1	1	1	0
		<i>F. verticillioides</i>	0	1	1	1	1	0	0	1	1	0	1	1	1	1	1	1	0	0	1	1	1	0
	100	<i>F. oxysporum</i> Fo47	1	0	0	1	1	0	0	0	1	0	0	1	1	1	1	1	0	0	1	1	1	0
0.05	100	<i>F. oxysporum</i> f. sp. <i>lycopersici</i>	0	1	0	1	1	0	0	0	1	0	0	1	1	0	1	1	0	0	1	1	1	0

matrix of clusters with known products allowed for reconstruction of the ancestral states of the gene clusters (**Table 3**). Interesting patterns of cluster development are already apparent from the table.

Several gene clusters of both terpene, nrps and t1pks types are conserved throughout all samples, whereas others seem limited to four out of the eight samples used. Evolution of these gene clusters can be visualized on a phylogenetic tree, which shows that the split of gene cluster presence coincides with a split in the phylogeny early during evolution (Figure 6). In the case of bikaverin and aurofusarin, which are both pigments, one can clearly observe this split (Schumacher *et al.*, 2013; Malz *et al.*, 2005). There are other interesting patterns found as well. When looking at depudecin, which is known to be a virulence factor in green cabbage

(*Brassica oleracea*), an interesting pattern can be observed (Wight *et al.*, 2009). Only two species, *F. oxysporum* f. sp. *Lycopersici* and *F. verticilliioides*, contain this cluster. Using maximum parsimony analysis, the ancestral states reconstruction indicates that the cluster was independently acquired twice during evolution. While this is one of the possibilities, a second possibility is horizontal gene transfer (Ma *et al.*, 2010).

In order to establish whether horizontal gene transfer occurs, a phylogenetic comparison was made. Using multiple sequence alignments of the anchor genes of BGCs and comparing the resulting phylogenies with the RBP2 phylogeny, it is possible to predict horizontal gene transfer events. Three BGCs with interesting occurrence patterns (fusarubin, fusarin C and malonichrom) were selected for this analysis.

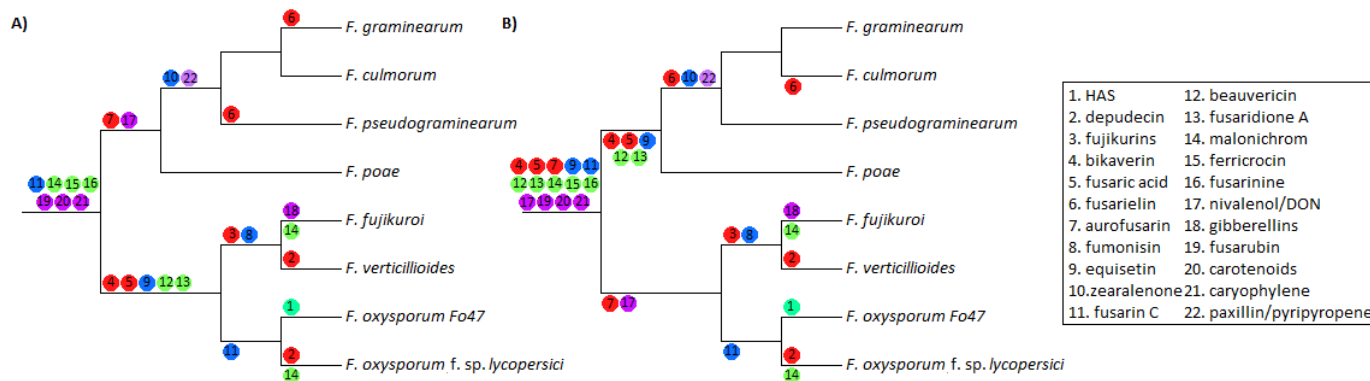


Figure 6. Maximum parsimony ancestral state reconstruction of BGCs with an experimentally established product, mapped on the RBP2 tree. Numbers above the branches visualize BGC births, whereas numbers beneath branches visualize BGC deaths. To account for unresolvable cases with maximum parsimony, two trees were made. A) maximizing BGC births and B) maximizing BGC deaths. The true ancestral states can be a combination of both trees. Colors correspond to BGC types from Table 3.

The fusarubin BGC, which is considered conserved across all *Fusarium* species, shows a phylogeny closely resembling the RBP2 phylogeny, only deviating slightly. Less conserved BGCs like those producing fusarin C and malonichrom, which cannot be found in all *Fusarium* species show to generally follow the RBP2 phylogeny as well (Figure 7), only slightly deviating from it. The evolutionary distance between the two subgroups in the RBP2 phylogeny is perfectly mimicked by the BGC phylogenies, indicating that horizontal gene transfer of these BGCs did not occur between the *Fusarium* species used in this project. This does not exclude the possibility that other

BGCs could have been acquired through horizontal gene transfer but shows that generally, BGC evolution is in line with the evolution of the species.

Discussion

BGCs fulfill diverse roles in *Fusarium*

From the 22 secondary metabolite gene clusters with known products, only a few genes are known to be involved in pathogenicity. Most BGCs produce metabolites not involved in pathogenicity like pigments or antibiotics, and are

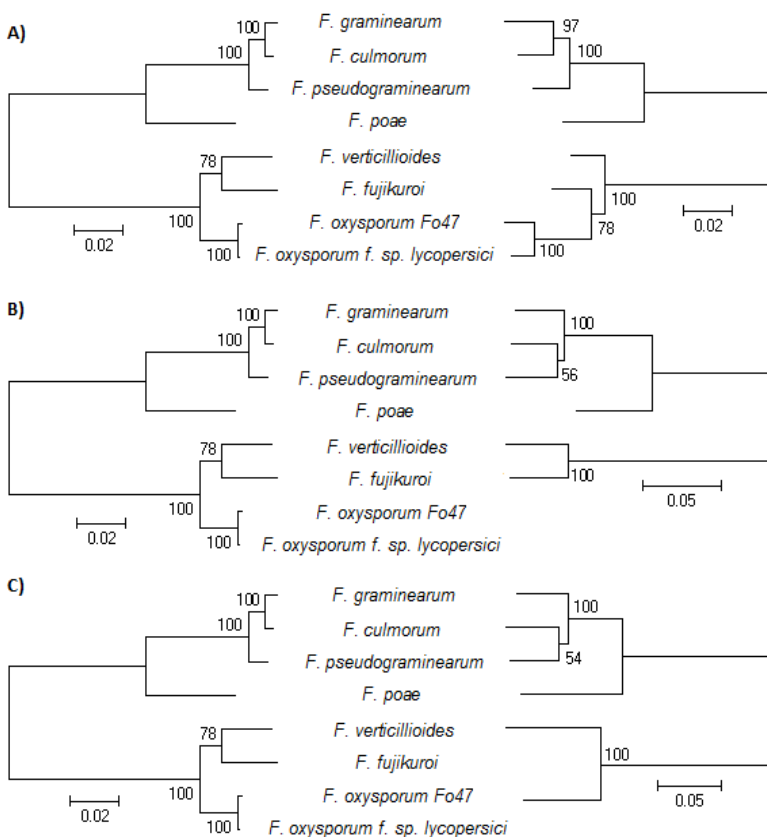


Figure 7. Comparison of RBP2 phylogenies with those of A) fusarubin, B) fusarin C and C) malonichrom BGC anchor genes. BGC phylogenies mostly follow the RBP2 phylogeny, and always show the same split between the two *Fusarium* subgroups.

involved in other important processes in fungal cells. Several mycotoxins are also produced, and while some of these mycotoxins are involved in pathogenicity towards the host, most are not. Depudecin and fumonisin have both been shown to be virulence factors (Cruz *et al.*, 2013; Wight *et al.*, 2009). Fumonisin, a strong virulence factor for maize, produced by *F. verticillioides*, is not directly related to disease in rice when infected by *F. fujikuroi*. This shows that much about its role in virulence is still unknown. While in this case, the compounds created are directly responsible for infection of the plant, there are also cases known where indirect effects occur. When synthesis of aurofusarin, a red pigment, is disrupted in *F. pseudograminearum*, production of zearalenone, a mycotoxin causing infertility, has been detected, while none was detected in strains with functional aurofusarin biosynthesis (Malz *et al.*, 2005; Schoevers *et al.*, 2012; Kuiper-Goodman *et al.*, 1987).

Current analyses allow for quick screening for toxins

By applying the analyses conducted during this project to newly sequenced genomes, quick and easy information can be acquired concerning the potential for the production of toxins in a species. *F. oxysporum* Fo47, an oxysporum strain that is widely used as a biological control strain to prevent other *Fusaria* from infecting a plant for example, still shows that it has the potential to produce the compound beauvericin. This is a worrisome discovery, since beauvericin is a mycotoxin capable of

triggering apoptosis in several human cell lines (Logrieco *et al.*, 1998). While this compound is not necessarily synthesized by *F. oxysporum* Fo47 under normal conditions, it is possible that through alteration of gene expression through either mutations or changing ecological conditions, production of beauvericin is resumed. The hexadehydro-astechrome (HAS) BGC is also detected in this *oxysporum* strain, showing 87% homology with the BGC identified in *Aspergillus fumigatus*. This compound is associated with increased mortality in mice when infected by *A. fumigatus* (Yin *et al.*, 2013). While it is thought to only be produced by a small group of human opportunistic fungi, indications are there that some form of this BGC is actually present in *F. oxysporum* Fo47. This should be a very good reason to for concern regarding the use of this strain for biological control purposes. Many more of these cases might arise in the future, and preliminary screening of genomes for toxin production capabilities can assist with making decisions concerning widespread use of an organism for either agricultural or medical purposes.

BGCs generally follow species phylogeny

Horizontal gene transfer is a phenomenon commonly observed in bacteria. Genetic material is exchanged, often conferring antibiotic resistance or toxins production capability (Koonin *et al.*, 2001; Gyles and Boerlin, 2014). While horizontal gene transfer is most common in bacteria, it is also observed in eukaryotes (Does and Rep, 2007; Gallagher and Jensen,

2015; Wisecaver and Rokas, 2015). Even though no horizontal gene transfer events have been observed during this analysis, it is likely that some of these events have actually taken place. The fumonisin cluster found in *F. verticillioides* and *F. fujikuroi* has also been found in the distantly related fungus *Aspergillus niger*, and was likely transferred through horizontal gene transfer (Wiemann *et al.*, 2013). Conserved BGCs like the fusarubin BGC however, have shown to precisely follow the RBP2 phylogeny, and are therefore likely to have been present in the *Fusarium* lineage for a very long time.

Supernumerary chromosomes could harbor developing BGCs

The genome of *F. poae* consists of several core chromosomes which are highly conserved across strains, along with several supernumerary chromosomes. These supernumerary chromosomes are less conserved and differ from strain to strain, comparable to those found in the *F. oxysporum* species complex (Ma *et al.*, 2010). Several terpene class BGCs located on the supernumerary chromosomes of *F. poae* were detected by antiSMASH. Most of these BGCs spanned entire contigs and are most likely incomplete, but were clustered together by BiG-SCAPE nonetheless. This indicates that even though no duplications of BGCs were found in other *Fusarium* species, the *F. poae* genome seems to contain several duplicated BGCs. Even though these BGCs might not actually be functional producers of secondary metabolites, it is possible that the supernumerary chromosomes act as a

nursery for developing BGCs. In order to find out whether this is actually the case, the quality of the assembly of *F. poae* will have to be improved after which a more in depth analysis can be conducted on these supernumerary chromosomes.

Distribution of depudecin hints at ecological importance

Through ancestral state reconstruction, several interesting patterns of BGC occurrences were recovered. The depudecin BGC is found in *F. oxysporum* f. sp. *lycopersici* and *F. verticillioides*, pathogens to tomato and maize respectively, while it is absent from the closely related *F. fujikuroi* and *F. oxysporum* Fo47. Both of these hosts find their origin of domestication in Mesoamerica several thousands of years ago (Bai and Lindhout, 2007; Mangelsdorf, 1958). It is possible that the birth of the depudecin BGC originated there, and that only those *Fusarium* species infecting plants originating from that place and time contain this BGC. It is therefore interesting to look at *Fusarium* species infecting other plants that are known to have their origins of cultivation in Mesoamerica. *Fusarium solani*, known to infect squash, beans and potato which all originate from the Andes mountain range, spanning Meso- and South-America. *Fusarium oxysporum* strains have also shown to infect chili, potato and cacao which are also thought to originate from the Andean region (Mangelsdorf, 1958; Smith, 1997; Delgado-Salinas *et al.*, 2011; Rosmana, 2014). If the depudecin BGC is detected in these additional species, it indicates that besides being a virulence

factor, depudecin might have, besides contributing to virulence, an additional ecological function as well.

More data is required for increased efficiency and better predictions

While it is known that there are many *Fusarium* species which are saprotrophic, not many of these species have been studied intensively. From an agricultural point of view, this is because pathogenic *Fusarium* species are more important to study when trying to fight the infection of host plants. For this project however, it is of great importance to have data on the presence or absence of secondary metabolite gene clusters in these saprotrophic species, since differences in presence/absence patterns can then pinpoint even secondary metabolites of which the product is yet unknown as likely candidates for involvement in pathogenicity. The genome of *F. langsethiae* that was not included in this analysis because of its draft status, although pathogenic, can be a good candidate to include in the analysis. Especially since not much is known about this species, while it is a suspected plant pathogen strongly resembling *F. poae* (Imathiu *et al.*, 2013). By increasing the amount of data, based on observed absence/presence patterns, candidate BGCs for pathogenicity can more easily be selected, and increase our understanding of the role of secondary metabolites in pathogenic interactions. In the future, this may lead to new methods to aid in the combat against pathogens,

reducing health risks and improving crop yields.

Materials and methods

Dataset

The initial dataset consisted of seven *Fusarium* genomes (**Table 4**). Due to a missing structural annotation, the genome of *F. langsethiae* was dropped. On the other hand, two genomes were added to the analysis. The genome of *F. oxysporum* Fo47, for it was known that this strain was not pathogenic in known *Fusarium* hosts, and even provided host plants with resistance against further *Fusarium* infections. Lastly, a newer version of the *F. graminearum* genome by the Broad Institute was added to see whether differences were observed between the two versions that were used.

Table 4. All genomes used during this project. Accession numbers coincide with the accession/version system of the NCBI database, and correspond to the first chromosome of the used genome annotation. * Added later in the experiment, ** removed from experiment.

Species	Source & accession
<i>F. graminearum</i>	NCBI (HG970330.2)
<i>F. graminearum</i> *	Broad institute (CM000574.1)
<i>F. culmorum</i>	NCBI (HG323944.1)
<i>F. pseudograminearum</i>	NCBI (CM003198.1)
<i>F. poae</i>	In house/NCBI (LYXU01000001.1)
<i>F. fujikuroi</i>	NCBI (HF679024.1)
<i>F. verticillioides</i>	NCBI (CM000578.1)
<i>F. oxysporum</i> f. sp. <i>lycopersici</i>	NCBI (CM000589.1)
<i>F. oxysporum</i> Fo47*	NCBI (JH717896.1)
<i>F. langsethiae</i> **	In house (No accession available)

Genome annotations were acquired in GenBank format (.gb/.gbk) and, when necessary, merged into one multiple-records GenBank file to fit antiSMASH's required input format. The *F. poae* GenBank file, as well as the *F. culmorum* GenBank file had to be manually enhanced because deviations from the official GenBank format were present in these files. The *F. poae* GenBank file that was used, although now available via NCBI, was not yet published at the moment of the analysis and was still in a draft stage. Several improvements have been made since, but results should be mostly unaffected. From the GenBank file, fasta files as well as GFF3 files with gene features were created to be used as input for BUSCO and CASSIS. CASSIS uses only gene features, so leaving out CDS, mRNA, intron and exon features should not interfere with any results acquired from CASSIS (T. Wolf, personal communication, March 31, 2016). All input required by following tools was derived from the aforementioned files by running them through the tools in the pipeline in order.

Phylogeny construction

The RBP2 gene was used as the sole gene to base the phylogeny of this *Fusarium* subset on. Using MEGA6, a bootstrapped tree was constructed using the multiple sequence alignment of the BLAST results from the RBP2 gene in *F. oxysporum* f. sp. *lycopersici* (chromosome 7, location: 3840192-3844693, locus: FOXG_10639). Default settings were used, with 100 bootstrap replicates.

Used tools & parameters

For this project, several bioinformatics tools were used, along with some python packages. To successfully run the script (run.py), all tools and their prerequisites should be installed on the system.

BUSCO

To assess gene space, BUSCO v1.1b1 was used. The tool along with the fungi dataset were downloaded from <http://busco.ezlab.org/>. Default parameters were used for all input files, and no additional parameters were used. In the final script, users can enable or disable arguments by use of a configuration file to maintain usability even for less advanced command prompt users. This configuration file contains options for all tools used in the script.

antiSMASH

For identification of gene clusters, antiSMASH was used. A development version was used, although any version higher than 3.0.5 should be compatible with the script including development versions. Official releases are available at <http://antismash.secondarymetabolites.org> and development versions are available at <https://bitbucket.org/antismash/antismash>. To more accurately predict cluster borders, integrate MIBiG database blast results, and conduct additional analyses on the genes present in the cluster, antiSMASH was run with the following additional parameters:

```
--borderpredict_only  
--knownclusterblast  
--smcogs
```

To visualise progress, and to decrease runtime, the following parameters were used:

--verbose

--cpus 8

These do not influence the results in any way, but enable for easier troubleshooting when something goes wrong.

CASSIS

Additional testing of cluster borders was done using CASSIS (v. April 2016, available

from <https://sbi.hki-jena.de/cassis/>)

was used. Due to the large amounts of data that had to be tested, and the relatively long runtime from CASSIS, the following (default) settings were used:

--predict

--gap-length 2

Verbose output was enabled as well as the increase of CPUs used. Even with eight CPUs allocated to CASSIS, runtimes for the dataset easily exceeded a day using default settings, which were quite stringent. Allowing for more lenient settings, runtimes increased even more.

BiG-SCAPE

antiSMASH output was used as input for BiG-SCAPE, which was run with the following extra parameter along with the default parameters.

--include_disc_nodes

This parameter was included to ensure that unique secondary metabolite gene clusters would not be left out in the final network. BiG-SCAPE is available at https://git.wageningenur.nl/yeong001/BGC_networks

Cytoscape

In order to visualise the networks generated by BiG-SCAPE, Cytoscape v.3.4.0 was used. Cytoscape is available from <http://www.cytoscape.org/>.

Two linkout entries were created to link nodes to, if available, their corresponding MIBiG database accession and the Multigeneblast results. The various samples were distinguished by different node colours, and the shape of the node was changed when the node had an accession in the MIBiG database. Edge width was set to decrease with an increasing raw distance, to better visualise internodal relations.

Multigeneblast

To visualise homology between secondary metabolite gene clusters in different species, an adapted version of the BLAST algorithm was used. The tool that was used for that during this project was Multigeneblast. antiSMASH cluster output files were modified slightly to contain the species name and cluster number in the version/accession feature to make results easier to interpret. The BLAST database used for the analysis comprised all predicted clusters across all samples. Predicted clusters were blasted one by one against this database to acquire visual homology results that could be used to validate results produced by BiG-SCAPE. Default settings were used when running multigeneblast.

(availability: <http://multigeneblast.sourceforge.net/>)

Mesquite

The ancestral state reconstruction of secondary metabolite gene clusters with

known compounds was conducted using mesquite. Mesquite was chosen for its great flexibility in visualization as well as analysis settings (Maddison and Maddison, 2016). The tool allowed for multiple reconstruction methods to be used. Maximum parsimony and maximum likelihood analyses are the most commonly used methods. Since the dataset was small, and maximum likelihood trees can be unreliable and biased in certain cases, a maximum parsimony approach was taken while reconstructing ancestral states (Kolaczowski and Thornton, 2004). With the use of a phylogenetic tree based on the RBP2 gene, emergence, disappearance and eventual duplication of gene clusters can be traced back through evolutionary history. A binary matrix containing species and gene clusters with known compounds was created to be used in mesquite.

References

- Alabouvette, C. *et al.* (1993) Recent advances in the biological control of fusarium wilts. *Pestic. Sci.*, **37**, 365–373.
- Bai, Y. and Lindhout, P. (2007) Domestication and breeding of tomatoes: What have we gained and what can we gain in the future? *Ann. Bot.*, **100**, 1085–1094.
- Benfield, A.H. and Gardiner, D.M. (2015) A genetic map of *Fusarium pseudograminearum*. *Unpublished*.
- Brown, D.W. and Proctor, R.H. (2016) Insights into natural products biosynthesis from analysis of 490 polyketide synthases from *Fusarium*. *Fungal Genet. Biol.*, **89**, 37–51.
- Cruz, A. *et al.* (2013) Phylogenetic analysis, fumonisin production and pathogenicity of *Fusarium fujikuroi* strains isolated from rice in the Philippines. *J. Sci. Food Agric.*, **93**, 3032–9.
- Cuomo, C.A. *et al.* (2007) The *Fusarium graminearum* genome reveals a link between localized polymorphism and pathogen specialization. *Science*, **317**, 1400–2.
- Dean, R. *et al.* (2012) The Top 10 fungal pathogens in molecular plant pathology. *Mol. Plant Pathol.*, **13**, 414–430.
- Delgado-Salinas, A. *et al.* (2011) *Vigna* (Leguminosae) sensu lato: The names and identities of the American segregate genera. *Am. J. Bot.*, **98**, 1694–1715.
- Does, H.C. van der and Rep, M. (2007) Virulence Genes and the Evolution of Host Specificity in Plant-Pathogenic Fungi. <http://dx.doi.org/10.1094/MPMI-20-10-1175>.
- Gallagher, K.A. and Jensen, P.R. (2015) Genomic insights into the evolution of hybrid isoprenoid biosynthetic gene clusters in the MAR4 marine streptomycete clade. *BMC Genomics*, **16**, 960.
- Gardiner, D.M. *et al.* (2014) Genome Sequence of *Fusarium graminearum* Isolate CS3005. *Genome Announc.*, **2**.
- Gyles, C. and Boerlin, P. (2014) Horizontally transferred

(availability:

<https://mesquiteproject.wikispaces.com/home>)

Script availability

All scripts that were created during this project were developed in Python 2.7. Scripts, along with their documentation and data files have been made publicly available on

<http://git.wageningenur.nl/hooge096/MScThesis.git>.

Acknowledgements

Big thanks go out to my supervisors Marnix Medema and Theo van der Lee, who provided me with advice and assistance along the entire way. Cees Waalwijk's constructive criticism was always useful and insightful. Lastly I would like to thank every fellow student who proofread the draft version of this thesis, and helped make it so much better.

- genetic elements and their role in pathogenesis of bacterial disease. *Vet. Pathol.*, **51**, 328–40.
- Imathiu, S.M. *et al.* (2013) Fusarium langsethiae - a HT-2 and T-2 Toxins Producer that Needs More Attention. *J. Phytopathol.*, **161**, 1–10.
- Keller, N.P. *et al.* (2005) Fungal secondary metabolism — from biochemistry to genomics. *Nat. Rev. Microbiol.*, **3**, 937–947.
- Kolaczowski, B. and Thornton, J.W. (2004) Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature*, **431**, 980–984.
- Koonin, E. V. *et al.* (2001) Horizontal Gene Transfer in Prokaryotes: Quantification and Classification ¹. *Annu. Rev. Microbiol.*, **55**, 709–742.
- Kuiper-Goodman, T. *et al.* (1987) Risk assessment of the mycotoxin zearalenone. *Regul. Toxicol. Pharmacol.*, **7**, 253–306.
- Logrieco, A. *et al.* (1998) Beauvericin production by Fusarium species. *Appl. Environ. Microbiol.*, **64**, 3084–8.
- Ma, L.-J. *et al.* (2010) Comparative genomics reveals mobile pathogenicity chromosomes in Fusarium. *Nature*, **464**, 367–73.
- Maddison, W.P. and Maddison, D.R. (2016) Mesquite: a modular system for evolutionary analysis.
- Malz, S. *et al.* (2005) Identification of a gene cluster responsible for the biosynthesis of aurofusarin in the Fusarium graminearum species complex. *Fungal Genet. Biol.*, **42**, 420–433.
- Mangelsdorf, P.C. (1958) Ancestor of Corn. *Science (80-.)*, **128**, 1313–1320.
- McMullen, M. and Stack, R. (2011) Fusarium head blight (scab) of small grains. *NDSU Ext. Serv.*, **804**, 6.
- Medema, M.H. *et al.* (2011) antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res.*, **39**, W339–46.
- Medema, M.H. *et al.* (2013) Detecting sequence homology at the gene cluster level with MultiGeneBlast. *Mol. Biol. Evol.*, **30**, 1218–23.
- Moolhuijzen, P.M. *et al.* (2013) Draft genome sequences of six wheat associated Fusarium spp. isolates. *Unpublished*.
- Niehaus, E.-M. *et al.* (2014) Characterization of the fusaric acid gene cluster in Fusarium fujikuroi. *Appl. Microbiol. Biotechnol.*, **98**, 1749–1762.
- Nucci, M. and Anaissie, E. (2007) Fusarium Infections in Immunocompromised Patients. *Clin. Microbiol. Rev.*, **20**, 695–704.
- O'Donnell, K. *et al.* (2010) Internet-accessible DNA sequence database for identifying fusaria from human and animal infections. *J. Clin. Microbiol.*, **48**, 3708–3718.
- O'Hagan, D. (2000) Pyrrole, pyrrolidine, pyridine, piperidine and tropane alkaloids (1998 to 1999). *Nat. Prod. Rep.*, **17**, 435–446.
- Rosmana, A. (2014) Cultural and pathogenic characterization of Fusarium fungi isolated from dieback branches of cacao.
- Samei, S.A. *et al.* (2008) How to tailor non-ribosomal peptide products--new clues about the structures and mechanisms of modifying enzymes. *Mol. Biosyst.*, **4**, 387–93.
- Schoevers, E.J. *et al.* (2012) Transgenerational toxicity of Zearalenone in pigs. *Reprod. Toxicol.*, **34**, 110–119.
- Schumacher, J. *et al.* (2013) A Functional Bikaverin Biosynthesis Gene Cluster in Rare Strains of Botrytis cinerea Is Positively Controlled by VELVET. *PLoS One*, **8**, e53729.
- Simão, F.A. *et al.* (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.
- Smith, B.D. (1997) The Initial Domestication of Cucurbita pepo in the Americas 10,000 Years Ago. *Science (80-.)*, **276**, 932–934.
- Smoot, M.E. *et al.* (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, **27**, 431–2.
- Studt, L. *et al.* (2012) Biosynthesis of fusarubins accounts for pigmentation of fusarium fujikuroi perithecia. *Appl. Environ. Microbiol.*, **78**, 4468–4480.
- Wiemann, P. *et al.* (2013) Deciphering the Cryptic Genome: Genome-wide Analyses of the Rice Pathogen Fusarium fujikuroi Reveal Complex Regulation of Secondary Metabolism and Novel Metabolites. *PLoS Pathog.*, **9**, e1003475.
- Wight, W.D. *et al.* (2009) Biosynthesis and Role in Virulence of the Histone Deacetylase Inhibitor Depudecin from Alternaria brassicicola. *Mol. Plant-Microbe Interact. MPMI*, **22**, 1258–1267.
- Windels, C.E. (2000) Economic and social impacts of fusarium head blight: changing farms and rural communities in the northern great plains. *Phytopathology*, **90**, 17–21.
- Wisecaver, J.H. and Rokas, A. (2015) Fungal metabolic gene clusters “caravans” traveling across genomes and environments. *Front. Microbiol.*, **6**.
- Wolf, T. *et al.* (2015) CASSIS and SMIPS: promoter-based prediction of secondary metabolite gene clusters in eukaryotic genomes. *Bioinformatics*, **btv713**–.
- Wu, F. (2007) Measuring the economic impacts of Fusarium toxins in animal feeds. *Anim. Feed Sci. Technol.*, **137**, 363–374.
- Yeong, M. (2016) BiG-SCAPE: exploring biosynthetic diversity through gene cluster similarity networks.
- Yin, W.-B. *et al.* (2013) A Nonribosomal Peptide Synthetase-Derived Iron(III) Complex from the Pathogenic Fungus Aspergillus fumigatus.