Protein type and lifestyle contribute to the evolution of low complexity regions in proteins of plant colonizing fungi

Gabriel Schweizer (gabriel.schweizer@ieu.uzh.ch)

Research article

Keywords: candidate effector proteins, low complexity regions, plant colonizing fungi, lifestyle, comparative genomics, evolution

Posted Date: March 13th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-17208/v1

Abstract

Background: Interactions between plants and fungi range from mutual symbiosis to parasitism. Fungi secrete proteins (termed effectors) to establish different forms of interactions with host plants. Such proteins are thought to coevolve with their molecular plant targets, and this can favor the emergence of novel alleles of fungal effector proteins. Low complexity regions in protein sequences are abundant in eukaryotes and were shown to contribute to the formation of novel protein sequences. This suggests that low complexity regions may play a role in the evolution of effector proteins. Several effector proteins with low complexity regions were functionally characterized in plant colonizing fungi that showed diverse lifestyles and belonged to different taxonomic groups. To investigate if low complexity regions in fungal effector proteins could contribute to the evolution of different plant-fungus interaction types, I employed publicly available genomic data from 121 species of plant colonizing fungi representing six different lifestyles and three phyla. I classified proteins in each species as cytoplasmic, secreted-non effector or effector protein and predicted low complexity regions in all protein sequences.

Results: I found that the fraction of proteins that contain a low complexity region differs between cytoplasmic and secreted proteins. Moreover, the fraction of a protein sequence that spans a low complexity region differed on average between cytoplasmic and secreted proteins. Inferring homologous relationships between effector proteins revealed that this fraction is higher in recent compared to ancestral proteins, suggesting that low complexity regions contribute to the formation of novel effector alleles. Furthermore, a principal component analysis and the results of a generalized linear model showed that the lifestyle of different fungi contributes to the evolution of low complexity regions. Likewise, the relative position of low complexity regions differed between cytoplasmic and secreted proteins, between ancestral and recent effector proteins, and between effectors of different lifestyles.

Conclusions: Protein type and lifestyle contribute to the evolution of low complexity regions in proteins of plant colonizing fungi, but molecular and evolutionary mechanisms explaining the differences between different protein types and proteins of different lifestyles remain to be elucidated.

Background

Over the last 400 million years, plants and fungi have shared a long history of coevolution, and mutualistic symbiotic interactions with fungi might have already supported the origin of the first land plants [1–3]. Today, diverse types of interactions can be found between plants and fungi, ranging from parasitism to mutualistic symbiosis [4–6]. Pathogenic fungi employ distinct strategies to colonize host plants and to obtain nutrients from them. Necrotrophic fungi kill their host plant and feed on dead plant tissue, whereas biotrophic fungi depend on the survival of the host plant to complete their life cycle. Some biotrophic species strictly depend on their host plant for survival (obligate biotrophs), while others can also grow as free-living organisms on artificial media (facultative biotrophs). Finally, hemibiotrophic fungi switch from an initial biotrophic to a later necrotrophic feeding strategy during plant colonization [7]. Understanding the molecular basis of pathogenic and symbiotic interactions is important, because they have a great influence on natural and agricultural ecosystems [8–16].

Pathogens, as well as symbionts, secrete proteins that modulate interactions with their host plants. Such proteins are termed effectors and fulfill their function in the apoplastic space between invading fungal hyphae and plant cells or are taken up into the plant cytoplasm. Effectors promote fungal colonization by suppressing plant immune responses, shielding invading hyphae or altering host cell physiology and metabolism in favor of the fungus [6, 7, 17–23]. Genes encoding effector proteins in pathogenic fungi are thought to coevolve antagonistically with their plant targets, and this may lead to the emergence of new alleles in effector genes [24].

Low complexity regions in protein sequences are very abundant in eukaryotes and are characterized by the high enrichment in one or a few amino acids [25, 26]. Their emergence is thought to be linked to mitotic replication slippage or meiotic recombination [26, 27]. Two different scenarios are discussed to explain the abundance of low complexity regions. One hypothesis proposes that low complexity regions are merely neutral spacers between protein folds [28], or that low complexity regions are excised from the mature protein sequence and have therefore no structural or functional role [29]. Moreover, the high diversification of low complexity regions between species suggests that such regions evolve neutrally [30]. An opposite scenario proposes that the presence of low complexity regions is adaptive [31, 32], as they, for example, increase mRNA stability [33] and underlie protein-protein interactions [31]. Furthermore, the prevalence of low complexity regions in antigenic loci supported the idea that low complexity regions contribute to antigen diversification [34]. Low complexity regions are an important source of phenotypic variation [35], and this innovation can form the basis of adaptations [36]. In sum, these characteristics of low complexity regions suggest that they also contribute to the evolution of fungal effector proteins.

Several examples of effectors containing low complexity regions have been functionally characterized in plant colonizing fungi that belong to different taxonomic groups and that employ different lifestyles [37, 38]. Therefore, I sought to investigate if the evolution of low complexity regions varies between species with different lifestyles. To this end, I used publicly available genomic data of 121 plant colonizing fungi representing six different lifestyles from symbiontic to necrotrophic and wood degrading species. I compared typical features of low complexity regions like the fraction of a protein sequence that spans a low complexity region, amino acid composition, and relative positions of low complexity regions in a protein sequence. The presented results reveal differences between cytoplasmic and secreted proteins, ancestral and recent protein sequences, and between different fungal lifestyles.

Results

First, I predicted low complexity regions in all protein sequences of the 121 investigated species (supplementary table 1) and determined the fraction of proteins with low complexity regions. This yielded one value for each species and protein type (Fig. 1). I then compared this fraction between cytoplasmic proteins, secreted non-effector proteins, and effector proteins. It turned out that the fraction of protein sequences with low complexity regions was lowest in secreted non-effector proteins in 119 species (Fig. 1). Exceptions to this general trend were the two obligate biotroph species Blumeria graminis f.sp. tritici (short name 'Blugrt') and Erysiphe necator (short name 'Erynec'), both belonging to the Ascomycota. I found this pattern in only 29 to 71 species in 10,000 random permutations (see Methods). Moreover, the fraction of proteins with low complexity regions was in 118 species higher in cytoplasmic proteins compared to secreted proteins (effectors and non-effectors) than in cytoplasmic proteins (Fig. 1). The three exceptions from this general trend were Taphrina deformans (a facultative biotrophic pathogen belonging to the Ascomycota; abbreviation 'Tapdef') as well as the two Basidiomycete symbionts Tulasnella calospora (abbreviation 'Tulcal'), and Piriformospora indica (abbreviation 'Pirind'). This trend occurred only in 14 to 52 species in 10,000 random permutations. In conclusion, these observations indicate that the fraction of proteins with low complexity regions does not evolve by chance. However, evolutionary mechanisms that could explain systematic differences in the presence of low complexity regions between cytoplasmic and secreted proteins remain to be identified. For example, the high fraction of cytoplasmic proteins with low complexity regions could suggest that low complexity regions are functionally important; hence, their presence could be advantageous and selected. Alternatively, the occurrence of low complexity regions could be neutral in cytoplasmic proteins, and therefore low complexity regions accumulate in cytoplasmic proteins. Likewise, it remains to be elucidated if the presence of low complexity regions in secreted proteins is generally disadvantageous, which would then explain the low fraction of secreted proteins with low complexity regions. In particular, the evolutionary and molecular mechanisms that underlie differences between secreted non-effector proteins and effector proteins remain to be elucidated. In summary, the fraction of proteins with low complexity regions differed between cytoplasmic and secreted proteins, but the observed trend was largely consistent between different lifestyles and phyla (Fig. 1A to Fig. 1F).

Next, I calculated the fraction of each protein sequence that spans a low complexity region, thereby providing one value for all proteins in each species (supplementary table 2). I found that the median of these fractions was highest in effector proteins for 95 species (Fig. 2). This number ranged from 21 to 58 species in 10,000 random permutations, again indicating that this pattern does not evolve neutrally. Moreover, the median fraction of protein sequences spanning a low complexity region did not evolve by chance in the investigated protein categories (Table 1). Intriguingly, I found the highest median values in the group of effector proteins (Table 1). Together with the analysis of the protein fraction with low complexity regions, this finding indicates that low complexity regions are less common in effector proteins, but on average longer when they occur.

Table 1
Minimum and maximum median fractions of protein sequences that span a low complexity region as obtained from 10,000 permutations

	Observed	minimum	maximum	observed median
protein category	median value	median value[1]	median value[1]	value by chance[2]
effector	0.05367232	0.03999907	0.04338914	no
secreted non-effector	0.03812317	0.03915904	0.04451039	no
cytoplasmic	0.04198473	0.04112554	0.04178556	no
1) Reported are the minimum and maximum median values that are obtained from 10,000 random permutations				
2) An observed value is considered to evolve by chance if it lies between the minimum and maximum median values obtained from 10,000 random permutations				

Previous studies reported that low complexity regions differ in their amino acid composition, that is, certain amino acids were found to be overrepresented in low complexity regions [39–41]. My analysis revealed over-representation of certain amino acids as well; however, no lifestyle-specific or phylum-specific enrichments could be identified (supplementary table 3).

To investigate further a putative role of low complexity regions in the emergence of novel effector alleles, I inferred homologous relationships between all 73,484 effector sequences (supplementary table 2). In two independent analyses, I used the natural effector protein sequences or sequences where I replaced low complexity regions with 'X' as unknown amino acid, because low complexity regions can complicate the search for homology [42]. For both analyses, I reconstructed families of homologous sequences with OrthoFinder [43] (supplementary table 2, supplementary table 4, and supplementary table 5). Next, I aimed to identify all families of homologous effector proteins that contain at least one member from each species. This set of proteins represents likely ancestral sequences, as they are conserved in all species; however, no family of homologous proteins contained members from all species (Table 2). Therefore, I used those families of homologous proteins that covered the largest number of species as a proxy for truly ancestral sequences (Table 2). As a complementary approach, I identified all groups of homologous proteins containing only effector proteins from one species. Since these sequences are species-specific, they likely emerged only recently. I found that low complexity regions span a higher fraction of protein sequences in recent proteins compared to ancestral proteins (Fig. 3A; P-value < 2.2×10^{-16}, Wilcoxon Rank-Sum test). I obtained similar results when I used natural protein sequences (that is, low complexity regions are not masked) to infer homologous relationships (supplementary Fig. 1A; P-value < 2.2×10^{-16}, Wilcoxon Rank-Sum test). Information about families of homologous effector proteins based on natural

and masked protein sequences are summarized in supplementary table 2, supplementary table 4, and supplementary table 5. This finding is in line with a previous study showing that the fraction of a protein sequence that spans a low complexity region is higher in younger protein sequences when comparing mammalian proteins with other vertebrate and non-vertebrate sequences [41], suggesting that this observation reflects a general trend in eukaryotic proteins.

Table 2
Groups of homologous proteins and number of their members for ancestral and species-specific proteins as identified by OrthoFinder with native and masked protein sequences

	ancestral proteins	species-specific proteins
native effector protein sequences	one group (OG0000001) with 870 proteins conserved in 119 species	9026 groups with 9369 species-specific proteins
masked effector protein sequences	one group (OG0000001) with 864 proteins conserved in 119 species	19645 groups with 19645 species-specific proteins

To identify lifestyle-specific differences in the fraction of a protein sequence that spans a low complexity region, I identified all families of homologous effector proteins that contain at least one member in a species from each of the six lifestyles. This approach highlighted 156 families (with 10 to 998 members) when I used the results obtained from OrthoFinder with masked sequences as input (supplementary table 2 and supplementary table 4). Next, I calculated the average fraction of a protein sequence that span a low complexity region between all effector proteins of one family, yielding one value per family and lifestyle (supplementary table 6). I then used this data set as input for a principal component analysis and found that the first and second principal components explain about 69.1% of the observed data (Fig. 3B). Interestingly, data from proteins of obligate biotrophic and necrotrophic fungi showed the largest difference in the principal components. Moreover, wood degrading and hemibiotrophic fungi showed similar results to necrotrophic fungi, although their contribution to the principal components was smaller (Fig. 3B). Furthermore, data from symbiontic fungi were similar to those of obligate biotrophs, which might reflect their strong dependence on host plants for survival. I obtained similar results when I used natural protein sequences for the detection of homology (supplementary table 7 and supplementary Fig. 1B). To gain more fine-grained insights in the contribution of fungal lifestyles on the fraction of protein sequences that span a low complexity region, phylogenetic information need to be taken into account [44]. However, obtaining accurate alignments and phylogenetic trees is challenging in this data set, because the used effector protein sequences represent hundreds of million years of evolution [45].

To investigate potential layered effects between protein type, lifestyle, and phylogenetic relationships (phylum), I fitted a general linear model to the data of all proteins, regardless of their type (supplementary table 2). Specifically, I used the formula “fraction of protein sequences that span a low complexity region” ~ protein type * lifestyle * phylum. I then used the results to rank the models with different fixed-term effects according to the Bayesian information criterion, and I found that all three parameters together explain best the observed data (Table 3). In summary, the fraction of low complexity regions in a protein sequence is higher in younger protein sequences, indicating that low complexity regions contribute to the formation of novel alleles. Moreover, the results obtained from a principal component analysis and a generalized linear model suggest that lifestyle contributes to the evolution of the fraction of effector protein sequences that span low complexity regions.

Table 3

Bayesian information criteria of generalized linear models that fit the fractions of protein sequences spanning low complexity regions based on different strata

intercept	ls[1]	phy[2]	pt[3]	ls:phy	ls:pt	phy:pt	ls:phy:pt	df[4]	likelihood	BIC[5]	delta	weight
0.021787527	+	+	+	+	+			25	-2390553.51	4780751.949	0	0.993225959
0.021544955	+	+	+	+				15	-2390477.508	4780741.974	9.975720634	0.006774041
0.021784575	+	+	+	+	+	+		29	-2390557.614	4780703.345	48.60455593	2.77E-11
0.021538011	+	+	+	+		+		19	-2390479.323	4780688.793	63.15680041	1.92E-14
0.021687621	+	+	+	+	+	+	+	39	-2390614.684	4780675.456	76.49366543	2.44E-17
0.021166705	+	+		+				13	-2389468.048	4778751.459	2000.490346	0
0.025110773	+	+	+		+			21	-2389472.866	4778647.473	2104.47681	0
0.024853619	+	+	+					11	-2389392.821	4778629.411	2122.538928	0
0.025113029	+	+	+		+	+		25	-2389477.269	4778599.466	2152.48305	0
0.024848407	+	+	+			+		15	-2389394.459	4778575.875	2176.074258	0
0.025809602	+		+		+			19	-2389040.414	4777810.974	2940.975137	0
0.025550158	+		+					9	-2388957.341	4777786.856	2965.093633	0
0.024464376	+	+						9	-2388392.656	4776657.485	4094.464077	0
0.0204476		+	+					6	-2388228.708	4776372.198	4379.751247	0
0.020441439		+	+			+		10	-2388230.374	4776318.719	4433.230397	0
0.025188246	+							7	-2387955.191	4775810.962	4940.987682	0
0.020632524			+					4	-2387877.125	4775697.439	5054.509974	0
0.020037031		+						4	-2387216.409	4774376.006	6375.943801	0
0.020261673								2	-2386870.007	4773711.608	7040.341502	0

1) ls, lifestyle

2) phy, phylum

3) pt, protein type

4) df, degrees of freedom

5) BIC, value of the Bayesian Information Criterion

A previous study indicated that low complexity regions could play a position-dependent role and proteins where low complexity regions tended to localize towards the termini of a protein had a larger number of interaction partners [46]. To investigate if low complexity regions show different localization patterns in my set of fungal proteins, I determined the relative position of low complexity regions in all types of proteins, that is, cytoplasmic proteins, secreted non-effector proteins, and effector proteins. Figure 4 shows the result for each low complexity region in each protein and species. In 115 species, the median relative position of low complexity regions in cytoplasmic proteins was located closer to the N-terminus than the median relative position of low complexity regions in secreted proteins (effectors and non-effectors). Exceptions to this general finding were Taphrina deformans (an Ascomycete facultative biotroph, short 'Tapdef'), Ustilago maydis (a Basidiomycete facultative biotroph, short 'Ustmay'), Zymoseptoria tritici (an Ascomycete hemibiotroph, short 'Zymtri'), Rhizoctonia solani (a Basidiomycete necrotroph, short 'Rhisol'), Tuber aestivum var. urcinatum (an Ascomycete symbiont, short 'Tubaes'), and Wolfiporia cocos (a Basidiomycete wood degrading fungus, short 'Wolcoc'). This suggests that the position of low complexity regions evolves in general differently between cytoplasmic and secreted proteins, and this conclusion is corroborated by results from 10,000 random permutations, where cytoplasmic proteins were located closest to the N-terminus in only 14 to 51 species. Following the results reported by Coletta and colleagues [46], this would indicate that cytoplasmic proteins with low complexity regions have more interaction partners than secreted proteins with low complexity regions. In 52 species, the median relative position of low complexity regions in secreted non-effectors was closer to the N-terminus than in effectors, and in 69 species, the opposite trend was observed. This is consistent with randomized samples, where low complexity regions were closer located to the N-terminus in secreted non-effectors compared to effectors in 39 to 83 species, suggesting that the relative localization of low complexity regions is similar between different types of secreted proteins (effectors and non-effectors). To investigate further if the observed median values of relative positions evolved by chance, I randomly assigned each protein to one

protein type (cytoplasmic, secreted non-effector, and effector). I found that the median relative position in the different protein type does not evolve by chance (Table 4).

Table 4
Minimum and maximum median relative positions of low complexity regions as obtained from 10,000 permutations

	Observed	minimum	maximum	observed median
protein category	median value	median value[1]	median value[1]	value by chance[2]
effector	0.6216692	0.4660705	0.4987578	no
secreted non-effector	0.6446508	0.03915904	0.04451039	no
cytoplasmic	0.4762675	0.02892562	0.07200726	no
1) Reported are the minimum and maximum median values that are obtained from 10,000 random permutations				
2) An observed value is considered to evolve by chance if it lies between the minimum and maximum median values obtained from 10,000 random permutations				

I used the before described data set of homologous effector protein families to investigate if there is a difference in the relative positions of low complexity regions between anciently and recently emerged protein sequences (Table 2, supplementary table 2, supplementary table 4). I observed that the relative position is closer to the N-terminus in ancient proteins (Fig. 5A; P-value = 0.01875, Wilcoxon Rank-Sum test). I observed a similar trend when I used natural protein sequences to infer homologous relationships (supplementary Fig. 2A; P-value = 0.02432, Wilcoxon Rank-Sum test). If we assume that the relative position of a low complexity region is indicative of the number of interaction partners, this result suggests that effector proteins with low complexity regions evolve a larger number of interaction partners over time.

To analyze potential lifestyle differences in relative positions of low complexity regions, I calculated a mean value of all homologous proteins belonging to one lifestyle (supplementary table 8). This yielded 46 families of homologous protein sequences with 24 to 998 members. The smaller number of homologous effector protein families compared to the analysis of the protein sequences that span a low complexity region (supplementary table 6) originates from the need to exclude proteins that do not contain low complexity regions, because I cannot determine relative positions in such cases. A principal component analysis based on these data showed that the first two principal components explain around 72.2% of the observed data (Fig. 5B). Again, I observed a strong contribution from effector proteins of obligate biotrophic species to the observed data. Moreover, hemibiotrophic, necrotrophic, and facultative biotrophic species showed similar contributions, which may reflect that the lifestyle of those species covers also saprotrophic feeding strategies. I found similar trends when I used data based on natural effector protein sequences (supplementary table 9 and supplementary Fig. 2B). Again, I sought to detect layered effects of the categories protein type, lifestyle, and phylum, and I used a general linearized model to highlight the contributions of these factors as described above. I found again that all three parameters together explain best the observed relative position of low complexity regions (Table 5). In summary, I conclude that the relative position of low complexity regions differs in ancestral and recent protein sequences. In addition, the results obtained from a principal component analysis and a generalized linear model suggest that lifestyle contributes to the evolution of relative positions of low complexity regions in effector proteins.

Table 5

Bayesian information criteria of generalized linear models that fit the relative position of low complexity regions based on different strata

intercept	ls[1]	phy[2]	pt[3]	ls:phy	ls:pt	phy:pt	ls:phy:pt	df[4]	likelihood	BIC[5]	delta	weight
0.485809	+	+	+	+				15	-248734.2587	497674.6588	0	1
0.4857375	+	+	+	+		+		19	-248729.5896	497720.2917	45.63289662	1.23E-10
0.4858139	+	+	+	+	+			25	-248725.2206	497794.0102	119.3514804	1.21E-26
0.485567	+	+	+	+	+	+		29	-248716.2173	497830.9745	156.3157266	1.14E-34
0.4854967	+	+	+	+	+	+	+	38	-248704.2129	497930.6505	255.9917515	2.58E-56
0.4735661	+	+	+					11	-248903.1913	497957.553	282.8942366	3.72E-62
0.4734886	+	+	+			+		15	-248898.7062	498003.5538	328.8950795	3.81E-72
0.4705451	+		+					9	-248941.2639	498006.2127	331.5539289	1.01E-72
0.4735189	+	+	+		+			21	-248893.6163	498075.8306	401.1718379	7.70E-88
0.4732644	+	+	+		+	+		25	-248884.5653	498112.6996	438.0408635	7.60E-96
0.4704797	+		+		+			19	-248931.9806	498125.0735	450.4147729	1.56E-98
0.4843521		+	+					6	-249320.1771	498722.8108	1048.15203	2.49E-228
0.484943			+					4	-249339.2366	498733.4442	1058.785433	1.22E-230
0.4842792		+	+			+		10	-249315.7992	498769.026	1094.367215	2.30E-238
0.4890257	+	+		+				13	-250223.0786	500624.813	2950.154198	0
0.4766281	+	+						9	-250395.2361	500914.1571	3239.498331	0
0.4732229	+							7	-250436.2246	500968.6486	3293.989831	0
0.4879425		+						4	-250808.1222	501671.2154	3996.556604	0
0.4881567								2	-250822.7226	501672.9308	3998.272028	0
1) ls, lifestyle												
2) phy, phylum												
3) pt, protein type												
4) df, degrees of freedom												
5) BIC, value of the Bayesian Information Criterion												

Discussion

Several studies showed that effector proteins containing low complexity regions contribute to virulence in fungi that represent diverse taxonomic groups and different strategies of plant colonization. Therefore, I sought to investigate if differences in lifestyle could contribute to the evolution of low complexity regions. Previous studies suggested that fungal lifestyle is connected to the secretome composition because secreted effector proteins aid in plant colonization [7]. For example, a comparative study comprising fungi with different plant colonization strategies showed that necrotrophic and hemibiotrophic species possess a larger repertoire of plant cell wall degrading enzymes compared to biotrophic and symbiontic fungi [7]. Additionally, secreted non-effector proteins could contribute to the evolution of different lifestyles. For instance, secreted serine proteases are involved in the determination of fungal lifestyles [47]. Based on these findings, I hypothesized that effector proteins with low complexity regions could also play a role in the lifestyle evolution of plant colonizing fungi.

To investigate this idea, I made use of comparative genomics. This approach has several potential shortcomings that may influence the results presented here. First, this strategy depends critically on the availability of high-quality genome assemblies and annotations. For example, re-sequencing of the Verticillium dahliae genome with single-molecule real-time sequencing together with optical mapping considerably improved the genome assembly [48]. Recent advances in the assembly of fungal genome sequences were also reported for the plant pathogens Botrytis cinerea [49] and Ramularia collo-cygni [50], and it is conceivable that future comparative genomics studies will benefit from further improvements in genome assemblies and annotations.

Second, I did not consider unconventionally secreted (effector) proteins in my analysis, although there are several reports of unconventionally secreted effector proteins that play important roles in virulence [51–53]. Such proteins could be identified in silico with SecretomeP [54, 55]. However, this software was designed for mammalian and bacterial sequences, and no method is available for the screening of proteins originating from non-mammalian eukaryotes [56]. For example, applications to plant proteins yielded only unreliable results [57]. Therefore, it is questionable whether SecretomeP would be well suited for the identification of potential fungal effector proteins. Thus, I did not consider unconventionally secreted proteins in the present analysis, although this may come with the cost of looking at incomplete sets of secreted (effector) proteins.

Third, my predictions of effector proteins did not take into account information about effector functions. Some effectors containing low complexity regions may only be functional after posttranslational modifications like cleavage of the protein. A case in point is the effector protein Rsp3 of Ustilago maydis [58]. Moreover, effectors were found to localize and function in different plant cellular compartments [59]. Effectors were also shown to be expressed in a colonization-stage dependent manner, that is, some effectors are already expressed when the fungus grows on the plant surface, whereas others are expressed only during late infection stages [60–64]. Such data are not yet available for a larger number of pathosystems, which makes it difficult to integrate them in current comparative studies.

Fourth, the investigated species do not only differ in their mode of colonization but also in their host range, that is, the number of host plants they can infect. This number can range from one to several hundred species [4], and it was shown that codon usage is one genomic factor that contributes to the evolution of host ranges [65]. It is challenging to include data of the host range in the present analysis because such information focuses often on economically important plants [66]. Moreover, the assignment of lifestyle categories may not be adequate for all species. For instance, it is conceivable that necrotrophic fungi show a biotrophic colonization strategy, at least during early infection stages, because they also need to colonize living plant tissue first.

Fifth, while I considered only genomic information of one strain, it has been shown that patterns of plant colonization can vary for different strains of the same species, likely due to their highly variable effector repertoire. Examples comprise the wheat pathogens Zymoseptoria tritici [67] and Blumeria graminis f. sp. tritici [68], and the broad host range pathogen Verticillium dahlia [69]. Moreover, different strains can encode different alleles of the same effector gene; this was, for example, reported for the low complexity regions containing Ustilago maydis effector Rsp3 [58].

Several reports suggest that effector proteins containing low complexity regions can play important roles in virulence [37, 70], and results presented in the present work suggest that different fungal lifestyles contribute to the evolution of low complexity regions in effector proteins.

Conclusions

A comparative genomics study with 121 plant colonizing fungi representing six different lifestyles showed that protein type (cytoplasmic or secreted), protein age (ancestral proteins conserved in most species or recent species-specific proteins), and lifestyle of different fungal species contribute to the evolution of low complexity regions in effector proteins. Future work is required to elucidate the evolutionary and molecular mechanisms that explain the observed differences between protein types (especially secreted non-effectors and secreted effectors), and between different lifestyles.

Methods

To study the evolution of low complexity regions in proteins of plant colonizing fungi, I employed publicly available genome data from 21 necrotrophic, 34 hemibiotrophic, 11 obligate biotrophic, 19 facultative biotrophic, and 24 symbiotic fungi. Furthermore, I included genomic data from 12 wood degrading species as a contrasting set, because wood degrading fungi obtain nutrients also from plant material, but do not establish an interaction with the living plant [71]. Among all investigated species, 39 are Basidiomycota, 81 are Ascomycota, and one belongs to Glomeromycota (supplementary table 1). All species considered in this study, lifestyle information, taxonomic classifications, and sources of protein-coding sequences are listed in supplementary table 1. All protein-coding sequences were initially filtered according to two criteria. If a sequence length was not a multiple of three (that is, the sequence contained invalid codons) or if a sequence contained a non-terminal stop codon (that is, the sequence represents a potential pseudogene), it was not considered for further analysis. The total number of annotated genes and the number of genes that passed the two filtering steps in each species are listed in supplementary table 1. All valid protein-coding sequences were translated to amino acid sequences and assigned as cytoplasmic protein, secreted (but non-effector) protein or secreted effector protein. To define the total set of predicted secreted proteins, SignalP 4.1 [72] was used to identify N-terminal secretion signal peptides, and TMHMM 2.0c [73], as well as Phobius 1.01 [74], were employed to identify transmembrane domains. C-terminal endoplasmatic reticulum retention signals (ERRS) were predicted with ps_scan 1.88 [75] using the prosite pattern PS00014 (ER_TARGET). This prosite pattern considers the consensus sequence [KRHQSA]-[DENQ]-E-L for the prediction of an ERRS. A protein was considered as secreted if (i) a secretion signal peptide could be found, (ii) no transmembrane domain was identified downstream of the signal peptide, and (iii) no ERRS was found. It is conceivable that this set of secreted proteins does not only contain putative effector proteins, but also “housekeeping” proteins that are, for example, required for fungal cell wall synthesis or modification. Therefore, I employed three prediction programs to identify genuine effector proteins, namely Localizer 1.0.3 [76], ApoplastP 1.0 [77], and EffectorP 2.0 [78]. I considered a predicted secreted protein as a putative effector if one of the three software showed a positive prediction. The numbers of cytoplasmic proteins, secreted non-effector proteins and effector proteins in each species are listed in supplementary table 1, and the assigned classification for each protein is shown in supplementary table 2.

I inferred homologous relationships between protein sequences with two approaches using OrthoFinder 2.2.6 [43]. In the first approach, I employed the native protein sequences, and in a second approach, I masked all low complexity regions with 'X' (that is, unknown amino acids) to rule out the possibility that low complexity regions affect the search for homologs. Default settings were used in both cases. Searches for homologous protein sequences were performed only for effector proteins, because they are involved in establishing interactions with host plants. Supplementary table 2 lists all proteins and provides information about the group of homologs to which each analyzed protein belongs in the two analyses. Supplementary table 4 and supplementary table 5 show the number of effector proteins from each species in each group of homologues that were identified by using masked or native protein sequences.

I used three programs to scan protein sequences for the presence of low complexity regions, namely SEG [79], DisEMBL [80], and fLPS [81]. All programs were run with default settings. As a conservative approach for assigning low complexity regions, I considered only protein regions that were identified by all three software as low complexity region. Relative positions of low complexity regions in protein sequences were calculated by dividing the midpoint position of a low complexity region by the protein length.

Amino acid enrichments or depletions in low complexity regions compared to non-low complexity regions were determined as follows. For each protein, the frequency of each amino acid was determined in low complexity regions and non-low complexity regions. This analysis was done separately for each species and each protein category (that is, cytoplasmic proteins, secreted non-effector proteins, and effector proteins). Significant differences in amino acid frequencies were identified with the Wilcoxon rank-sum test, followed by Bonferroni correction to account for multiple testing. Specifically, I multiplied each P-value obtained with the Wilcoxon rank-sum test with 20 (the number of tested amino acids), and considered differences as significant if the P-value was smaller than 5% after this correction.

To assess if observed results can be explained by chance alone, I performed random permutation. Specifically, the total set of proteins comprises x_E effector proteins, x_s secreted non-effector proteins, and x_C cytoplasmic proteins. From the total set of proteins, I randomly assigned x_E proteins as effectors, x_s proteins as secreted non-effectors, and x_C proteins as cytoplasmic. This assignment was done without replacement (that is, the total number of proteins did not change, and each protein is assigned to only one category) and repeated 10,000 times. For each repetition, I noted the median fraction of a protein sequence that is spanned by a low complexity region and the relative position of a low complexity region. Moreover, I recorded (i) in how many cases the fraction of proteins with low complexity region was lowest in secreted non-effector proteins (Fig. 1), (ii) the number of cases where the fraction of proteins with low complexity regions was lower in secreted proteins (effectors and non-effectors) than in cytoplasmic proteins (Fig. 1), (iii) the number of cases where the median fraction of a protein sequence that spans a low complexity region was highest in effectors (Fig. 2), (iv) in how many cases the median relative position of a low complexity region was closest to the N-terminus in cytoplasmic proteins (Fig. 4), and (v) the number of cases where the relative position of low complexity regions was closer located to the N-terminus in secreted non-effectors compared to effectors (Fig. 4).

To examine if the three characteristics protein type, lifestyle, and phylum could have a layered effect on the observed values (fraction of a protein sequence that spans a low complexity region and relative position of low complexity regions), I fitted a generalized linear model to the data. The fractions of protein sequences that span a low complexity region were $\log_{10}$ transformed because of their skewed distribution. The resulting model was then used as input for the dredge function in the MuMIn package [82] of R (https://www.r-project.org). This function generates a set of models with combinations of fixed effect terms. I used the Bayesian information criterion (option BIC) to rank results.

Declarations

Ethics approval and consent to participate – not applicable.

Consent for publication – not applicable.

Availability of data and materials – this study is based on publicly available genomic data. The sources of all data are listed in supplementary table 1. Data that were generated in the present study are available in supplementary table 2 to supplementary table 9. All supplementary tables are available at FigShare as follows: supplementary table 1: 10.6084/m9.figshare.11920194; supplementary table 2: 10.6084/m9.figshare.11919048; supplementary table 3: 10.6084/m9.figshare.11920200; supplementary table 4: 10.6084/m9.figshare.11920209; supplementary table 5: 10.6084/m9.figshare.11920224; supplementary table 6: 10.6084/m9.figshare.11920230; supplementary table 7: 10.6084/m9.figshare.11920239; supplementary table 8: 10.6084/m9.figshare.11920326; supplementary table 9: 10.6084/m9.figshare.11920329. Figures, tables and supplementary figures are included in this manuscript.

Competing interests – the authors declare that they have no competing interests.

Funding – this project did not receive specific funding.

Authors' contributions – G. S. conceived the study, designed and performed analyses, interpreted results, and wrote the manuscript.

Acknowledgements – I am grateful to Pouria Dasmeh for very fruitful discussions and helpful advice on data analyses, to Felix Moerman for his help with model ranking, and to Eugenio Azpeitia and Libera Lo Presti for their constructive comments on an early version of this manuscript.

References

1. Remy W, Taylort TN, Hass H, Kerp H. Four hundred-million-year-old vesicular arbuscular mycorrhizae (Endomycorrhiae/symbiosis/fossil ft /mut). Proceedings of National Academy of Sciences, USA. 1994;91(December):11841–3.
2. Gehrig H, Schüssler A, Kluge M. Geosiphon pyriforme, a fungus forming endocytobiosis with Nostoc (cyanobacteria), is an ancestral member of the Glomales: evidence by SSU rRNA analysis. J Mol Evol. 1996;43:71–81.
3. Martin FM, Uroz S, Barker DG. Ancestral alliances: Plant mutualistic symbioses with fungi and bacteria. Science. 2017;356(6340):article eaad4501.
4. Dean R, Van Kan J, Pretorius Z, Hammond-Kosack K, Di Pietro A, Spanu P, Rudd J, Dickman M, Kahmann R, Ellis J, Foster G. The Top 10 fungal pathogens in molecular plant pathology. Molecular Plant Pathology. 2012;13(4):414–30.
5. Lowe RGT, Howlett BJ. Indifferent, affectionate, or deceitful: Lifestyles and secretomes of fungi. PLoS Pathogens. 2012;8(3):e1002515–e1002515.
6. van der Heijden MGA, Martin FM, Selosse MA, Sanders IR. Mycorrhizal ecology and evolution: The past, the present, and the future. New Phytologist. 2015;205(4):1406–23.
7. Lo Presti L, Lanver D, Schweizer G, Tanaka S, Liang L, Tollot M, Zuccaro A, Reissmann S, Kahmann R. Fungal Effectors and Plant Susceptibility. Annual Review of Plant Biology. 2015;66(1):513–45.
8. Parniske M. Arbuscular mycorrhiza: The mother of plant root endosymbioses. Nature Reviews Microbiology. 2008;6(10):763–75.
9. Fisher M, Henk D, Briggs C, Brownstein J, Madoff L, McCraw S, Gurr S. Emerging fungal threats to animal, plant and ecosystem health. Nature. 2012;484(7393):186–94.
10. Bagchi R, Gallery RE, Gripenberg S, Gurr SJ, Narayan L, Addis CE, Freckleton RP, Lewis OT. Pathogens and insect herbivores drive rainforest plant diversity and composition. Nature. 2014;506(7486):85–8.
11. Behie SW, Bidochka MJ. Nutrient transfer in plant-fungal symbioses. Trends in Plant Science. 2014;19(11):734–40.
12. Bever JD, Mangan SA, Alexander HM. Maintenance of Plant Species Diversity by Pathogens. Annual Review of Ecology, Evolution, and Systematics. 2015;46(1):305–25.
13. Berruti A, Lumini E, Balestrini R, Bianciotto V. Arbuscular mycorrhizal fungi as natural biofertilizers: Let's benefit from past successes. Frontiers in Microbiology. 2016;6:article1559.
14. Verzeaux J, Hirel B, Dubois F, Lea PJ, Tétu T. Agricultural practices to improve nitrogen use efficiency through the use of arbuscular mycorrhizae: Basic and agronomic aspects. Plant Science. 2017;264:48–56.
15. Choi J, Summers W, Paszkowski U. Mechanisms Underlying Establishment of Arbuscular Mycorrhizal Symbioses. Annual Review of Phytopathology. 2018;56(1):135–60.
16. Savary S, Willocquet L, Pethybridge SJ, Esker P, McRoberts N, Nelson A. The global burden of pathogens and pests on major food crops. Nature Ecology and Evolution. 2019;3(3):430–9.
17. Stergiopoulos I, de Wit PJGM. Fungal Effector Proteins. Annual Review of Phytopathology. 2009;47(1):233–63.
18. Giraldo MC, Valent B. Filamentous plant pathogen effectors in action. Nature Reviews Microbiology. 2013;11(11):800–14.
19. Ökmen B, Doehlemann G. Inside plant: Biotrophic strategies to modulate host immunity and metabolism. Current Opinion in Plant Biology. 2014;20:19–25.
20. Rövenich H, Boshoven JC, Thomma BPHJ. Filamentous pathogen effector functions: Of pathogens, hosts and microbiomes. Current Opinion in Plant Biology. 2014;20:96–103.
21. Zuccaro A, Lahrmann U, Langen G. Broad compatibility in fungal root symbioses. Current Opinion in Plant Biology. 2014;20:135–45.
22. Plett JM, Martin F. Reconsidering mutualistic plant-fungal interactions through the lens of effector biology. Current Opinion in Plant Biology. 2015;26:45–50.
23. Toruño TY, Stergiopoulos I, Coaker G. Plant-Pathogen Effectors: Cellular Probes Interfering with Plant Defenses in Spatial and Temporal Manners. Annual Review of Phytopathology. 2016;54:419–41.
24. Brown JKM, Tellier A. Plant-Parasite Coevolution: Bridging the Gap between Genetics and Ecology. Annual Review of Phytopathology. 2011;49(1):345–67.
25. Golding GB. Simple sequence is abundant in eukaryotic proteins. Protein Science. 1999;8(6):1358–61.
26. Marcotte EM, Pellegrini M, Yeates TO, Eisenberg D. A census of protein repeats. Journal of Molecular Biology. 1999;293(1):151–60.
27. Levinson G, Gutman GA. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Molecular Biology and Evolution. 1987;4(3):203–21.
28. Huntley M, Golding GB. Evolution of simple sequence in proteins. Journal of Molecular Evolution. 2000;51(2):131–40.
29. Pizzi E, Frontali C. Low-complexity regions in Plasmodium falciparum proteins. Genome Research. 2001;11(2):218–29.
30. Ekman D, Light S, Björklund ÅK, Elofsson A. What properties characterize the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae? Genome Biology. 2006;7(6):article R45.

31. Karlin S, Brocchieri L, Bergman A, Mrázek J, Gentles AJ. Amino acid runs in eukaryotic proteomes and disease associations. Proceedings of the National Academy of Sciences of the United States of America. 2002;99(1):333–8.
32. Clarke JL, Sodeinde O, Mason PJ. A unique insertion in plasmodium berghei glucose-6-phosphate dehydrogenase-6-phosphogluconolactonase: Evolutionary and functional studies. Molecular and Biochemical Parasitology. 2003;127(1):1–8.
33. Xue HY, Forsdyke DR. Low-complexity segments in Plasmodium falciparum proteins are primarily nucleic acid level adaptations. Molecular and Biochemical Parasitology. 2003;128:21–32.
34. Verstrepen KJ, Jansen A, Lewitter F, Fink GR. Intragenic tandem repeats generate functional variability. Nature Genetics. 2005;37:986–90.
35. Fondon III JW, Garner HR. Molecular origins of rapid and continuous morphological evolution. Proceedings of the National Academy of Sciences of the United States of America. 2004;101(52):18058–63.
36. Kashi Y, King DG. Simple sequence repeats as advantageous mutators in evolution. Trends in Genetics. 2006;22(5):253–9.
37. Mesarich CH, Bowen JK, Hamiaux C, Templeton MD. Repeat-containing protein effectors of plant-associated organisms. Frontiers in Plant Science. 2015;6:article 872.
38. Ma LS, Pellegrin C, Kahmann R. Repeat-containing effectors of filamentous pathogens and symbionts. Current Opinion in Microbiology. 2018;46:123–30.
39. DePristo MA, Zilversmit MM, Hartl DL. On the abundance, amino acid composition, and evolutionary dynamics of low-complexity regions in proteins. Gene. 2006;378(1–2):19–30.
40. Radó-Trilla N, Albà M. Dissecting the role of low-complexity regions in the evolution of vertebrate proteins. BMC Evolutionary Biology. 2012;12:article 155.
41. Toll-Riera M, Radó-Trilla N, Martys F, Albá MM. Role of low-complexity sequences in the formation of novel protein coding sequences. Molecular Biology and Evolution. 2012;29(3):883–6.
42. Wootton JC. Sequences with 'unusual' amino acid compositions. Current Opinion in Structural Biology. 1994;4(3):413–21.
43. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biology. 2015;16(1):article157.
44. Felsenstein J. Phylogenies and the Comparative Method. The American Naturalist. 1985;125(1):1–15.
45. Löytynoja A. Alignment Methods: Strategies, Challenges, Benchmarking, and Comparative Overview. In: Anisimova M, editor. Evolutionary Genomics: Statistical and Computational Methods [Internet]. Totowa, NJ: Humana Press; 2012. p. 203–35. Available from: https://doi.org/10.1007/978-1-61779-582-4_7
46. Coletta A, Pinney JW, Solís DYW, Marsh J, Pettifer SR, Attwood TK. Low-complexity regions within protein sequences have position-dependent roles. BMC Systems Biology. 2010;4:article 43.
47. Muszewska A, Stepniewska-Dziubinska MM, Steczkiewicz K, Pawlowska J, Dziedzic A, Ginalski K. Fungal lifestyle reflected in serine protease repertoire. Scientific Reports. 2017;7:article 9147.
48. Faino L, Seidl MF, Datema E, van den Berg GCM, Janssen A, Wittenberg AHJ, Thomma BPHJ. Single-Molecule Real-Time Sequencing Combined with Optical. mBio. 2015;6(4):e00936-15.
49. Van Kan JAL, Stassen JHM, Mosbach A, Van Der Lee TAJ, Faino L, Farmer AD, Papasotiriou DG, Zhou S, Seidl MF, Cottam E, Edel D, Hahn M, Schwartz DC, Dietrich RA, Widdison S, Scalliet G. A gapless genome sequence of the fungus Botrytis cinerea. Molecular Plant Pathology. 2017;18(1):75–89.
50. Stam R, Münsterkötter M, Pophaly SD, Fokkens L, Sghyer H, Güldener U, Hückelhoven R, Hess M. A new reference genome shows the one-speed genome structure of the barley pathogen ramularia collo-cygni. Genome Biology and Evolution. 2018;10(12):3243–9.
51. Giraldo MC, Dagdas YF, Gupta YK, Mentlak TA, Yi M, Martinez-Rocha AL, Saitoh H, Terauchi R, Talbot NJ, Valent B. Two distinct secretion systems facilitate tissue invasion by the rice blast fungus Magnaporthe oryzae. Nature Communications. 2013;4:article 1996.
52. Liu T, Song T, Zhang X, Yuan H, Su L, Li W, Xu J, Liu S, Chen L, Chen T, Zhang M, Gu L, Zhang B, Dou D. Unconventionally secreted effectors of two filamentous pathogens target plant salicylate biosynthesis. Nature Communications. 2014;5:article 4686.
53. Krombach S, Reissmann S, Kreibich S, Bochen F, Kahmann R. Virulence function of the Ustilago maydis sterol carrier protein 2. New Phytologist. 2018;220(2):553–66.
54. Bendtsen JD, Jensen LJ, Blom N, Von Heijne G, Brunak S. Feature-based prediction of non-classical and leaderless protein secretion. Protein Engineering, Design and Selection. 2004;17(4):349–56.
55. Bendtsen JD, Kiemer L, Fausbøll A, Brunak S. Non-classical protein secretion in bacteria. BMC Microbiology. 2005;5:article 58.
56. Nielsen H, Petsalaki EI, Zhao L, Stühler K. Predicting eukaryotic protein secretion without signals. Biochimica et Biophysica Acta - Proteins and Proteomics. 2019;1867(12):article 140174.
57. Lonsdale A, Davis MJ, Doblin MS, Bacic A. Better than nothing? Limitations of the prediction tool secretomeP in the search for leaderless secretory proteins (LSPs) in plants. Frontiers in Plant Science. 2016;7:article 1451.
58. Ma LS, Wang L, Trippel C, Mendoza-Mendoza A, Ullmann S, Moretti M, Carsten A, Kahnt J, Reissmann S, Zechmann B, Bange G, Kahmann R. The Ustilago maydis repetitive effector Rsp3 blocks the antifungal activity of mannose-binding maize proteins. Nature Communications. 2018;9:article

1711.
59. Khan M, Seto D, Subramaniam R, Desveaux D. Oh, the places they'll go! A survey of phytopathogen effectors and their host targets. Plant Journal. 2018;93(4):651–63.
60. Skibbe DS, Doehlemann G, Fernandes J, Walbot V. Maize Tumors Caused by Ustilago maydis Require Organ-Specific Genes in Host and Pathogen. Science. 2010;328(5974):89–92.
61. Hacquard S, Kracher B, Maekawa T, Vernaldi S, Schulze-Lefert P, Van Themaat EVL. Mosaic genome structure of the barley powdery mildew pathogen and conservation of transcriptional programs in divergent hosts. Proceedings of the National Academy of Sciences of the United States of America. 2013;110(24):2219–28.
62. Dobon A, Bunting DCE, Cabrera-Quio LE, Uauy C, Saunders DGO. The host-pathogen interaction between wheat and yellow rust induces temporally coordinated waves of gene expression. BMC Genomics. 2016;17(1):380–380.
63. Zeng FS, Menardo F, Xue MF, Zhang XJ, Gong SJ, Yang LJ, Shi WQ, Yu DZ. Transcriptome analyses shed new insights into primary metabolism and regulation of Blumeria graminis f. Sp. tritici during conidiation. Frontiers in Plant Science. 2017;8(June):1–17.
64. Lanver D, Müller AN, Happel P, Schweizer G, Haas FB, Franitza M, Pellegrin C, Reissmann S, Altmüller J, Rensing SA, Kahmann R. The Biotrophic Development of Ustilago maydis Studied by RNA-Seq Analysis. The Plant Cell. 2018;30(2):300–23.
65. Badet T, Peyraud R, Mbengue M, Navaud O, Derbyshire M, Oliver RP, Barbacci A, Raffaele S. Codon optimization underpins generalist parasitism in fungi. eLife. 2017;6:e22472.
66. Farr DF, Rossman AY, Palm ME, McCray EB. USDA Fungus-Host Database [Internet]. Available from: https://nt.ars-grin.gov/fungaldatabases/fungushost/fungushost.cfm
67. Haueisen J, Möller M, Eschenbrenner CJ, Grandaubert J, Seybold H, Adamiak H, Stukenbrock EH. Highly flexible infection programs in a specialized wheat pathogen. Ecology and Evolution. 2019;9(1):275–94.
68. Müller MC, Praz CR, Sotiropoulos AG, Menardo F, Kunz L, Schudel S, Oberhänsli S, Poretti M, Wehrli A, Bourras S, Keller B, Wicker T. A chromosome-scale genome assembly reveals a highly dynamic effector repertoire of wheat powdery mildew. New Phytologist. 2019;221(4):2176–89.
69. Gibriel HAY, Li J, Zhu L, Seidl MF, Thomma BPHJ. Verticillium dahliae strains that infect the same host plant display highly divergent effector catalogs. bioRxiv [Internet]. 2019; Available from: https://www.biorxiv.org/content/10.1101/528729v1
70. Ma LS, Pellegrin C, Kahmann R. Repeat-containing effectors of filamentous pathogens and symbionts. Current Opinion in Microbiology. 2018;46:123–30.
71. Ohm RA, Riley R, Salamov A, Min B, Choi IG, Grigoriev IV. Genomics of wood-degrading fungi. Fungal Genetics and Biology. 2014;72:82–90.
72. Petersen TN, Brunak S, Von Heijne G, Nielsen H. SignalP 4.0: Discriminating signal peptides from transmembrane regions. Nature Methods. 2011;8(10):785–6.
73. Krogh A, Larsson B, Von Heijne G, Sonnhammer ELL. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. Journal of Molecular Biology. 2001;305(3):567–80.
74. Käll L, Krogh A, Sonnhammer ELL. A combined transmembrane topology and signal peptide prediction method. Journal of Molecular Biology. 2004;338(5):1027–36.
75. de Castro E, Sigrist CJA, Gattiker A, Bulliard V, Langendijk-Genevaux PS, Gasteiger E, Bairoch A, Hulo N. ScanProsite: Detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Research. 2006;34(WEB. SERV. ISS.):362–5.
76. Sperschneider J, Catanzariti AM, Deboer K, Petre B, Gardiner DM, Singh KB, Dodds PN, Taylor JM. LOCALIZER: Subcellular localization prediction of both plant and effector proteins in the plant cell. Scientific Reports. 2017;7:article 44598.
77. Sperschneider J, Dodds PN, Singh KB, Taylor JM. ApoplastP: prediction of effectors and plant proteins in the apoplast using machine learning. New Phytologist. 2018;217(4):1764–78.
78. Sperschneider J, Dodds PN, Gardiner DM, Singh KB, Taylor JM. Improved prediction of fungal effector proteins from secretomes with EffectorP 2.0. Molecular Plant Pathology. 2018;19(9):2094–110.
79. Wootton JC, Federhen S. [33] Analysis of Compositionally Biased Regions in Sequence Databases. Methods in Enzymology. 1996;266:554–71.
80. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. Protein disorder prediction: Implications for structural proteomics. Structure. 2003;11(11):1453–9.
81. Harrison PM. fLPS: Fast discovery of compositional biases for the protein universe. BMC Bioinformatics. 2017;18(1):1–9.
82. Bartoń K. MuMIn package [Internet]. 2019. Available from: https://cran.r-project.org/web/packages/MuMIn/

Figures

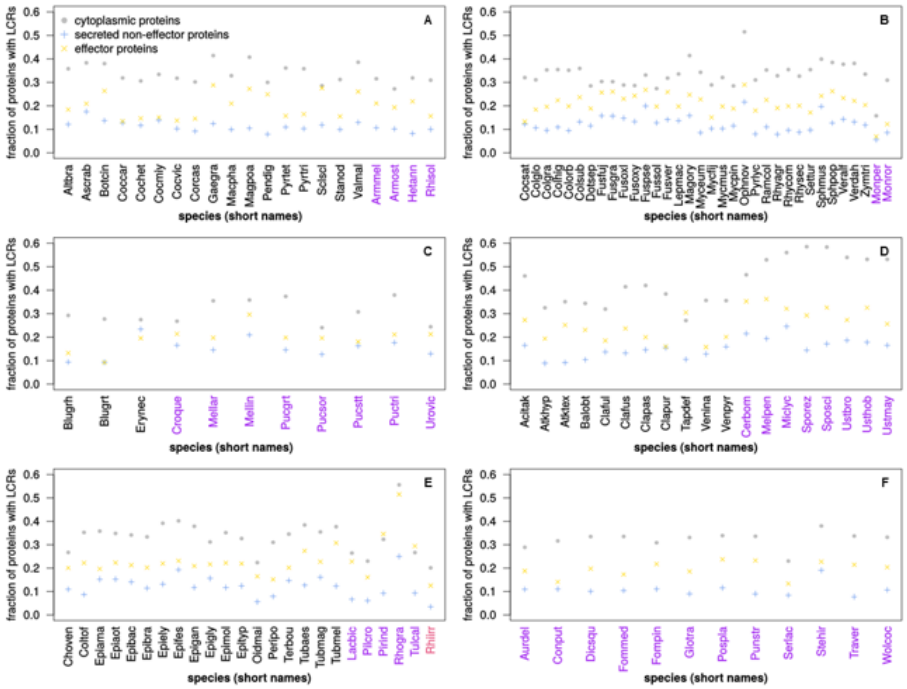

Figure 1

Fraction of proteins that contain low complexity regions. In each species, the fraction of cytoplasmic, secreted non-effectors and effectors with at least one low complexity region was determined (vertical axis). The data for each protein type are presented as illustrated by the legend in panel A (grey circle, cytoplasmic proteins; blue cross, secreted non-effector proteins; yellow diagonal cross, effector proteins). Each panel shows the results of species from one lifestyle, namely necrotroph (A), hemibiotroph (B), obligate biotroph (C), facultative biotroph (D), symbiont (E), or wood degrading (F). Within each panel, species are sorted by their phylum (Ascomycota, black; Basidiomycota, purple; Glomeromycota, pink). Within each phylum, species are sorted alphabetically, thereby grouping species from one genus together. The indicated short names of each species (horizontal axis) can be linked to full species names with information in supplementary table 1.

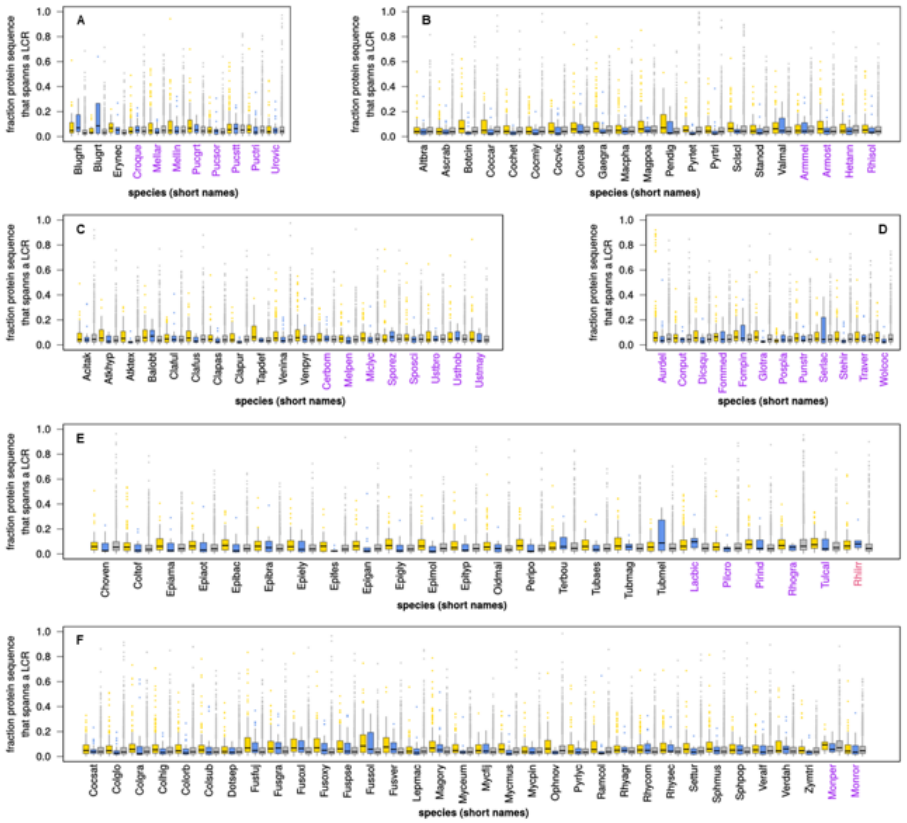

Figure 2

Fraction of protein sequences that spans a low complexity region. The fraction of a protein sequence that contains low complexity regions was calculated for each protein in each species (vertical axis). Data are represented separately for each protein type as indicated by different colors (grey, cytoplasmic proteins; blue, secreted non-effector proteins; yellow, effector proteins). Data are shown in the form of a boxplot. The thick black line in each box indicates the median value. The lower and upper box limit indicates the values of the first and third quartile, respectively. Whiskers show data within the 1.5-fold interquartile range and dots show outliers beyond the 1.5-fold interquartile range. Each panel shows the results of species from one lifestyle, namely obligate biotroph (A), necrotroph (B), facultative biotroph (C), wood degrading (D), symbiont (E), or hemibiotroph (F). Within each panel, species are sorted by their phylum (Ascomycota, black; Basidiomycota, purple; Glomeromycota, pink). Within each phylum, species are sorted alphabetically, thereby grouping species from one genus together. The indicated short names of each species (horizontal axis) can be linked to full species names with information in supplementary table 1.

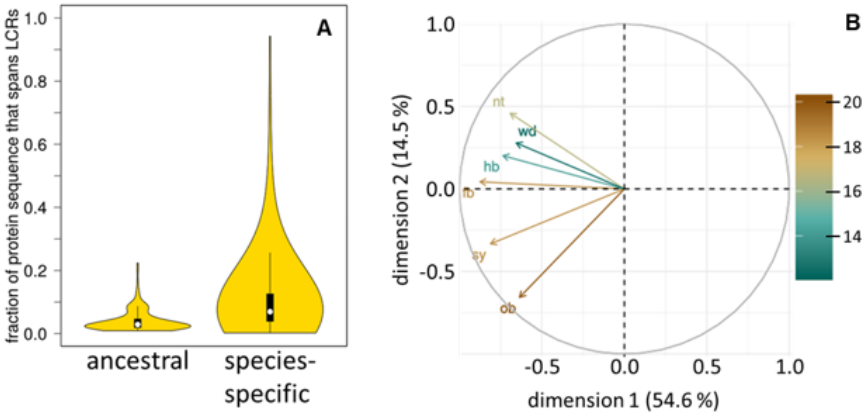

Figure 3

Fraction of secreted effector protein sequences that span a low complexity region. (A) Effector proteins are classified as ancestral or species-specific (horizontal axis) and their sequence fraction spanning a low complexity region is shown on the vertical axis. OrthoFinder was used to reconstruct groups of homologous proteins by using masked protein sequences (low complexity regions are replaced with 'X' as unknown amino acid). The data distributions are shown in the form of a violin plot. In addition to the boxplot shown in black, this plot indicates the density of data at each value. (B) Principal component analysis of effector protein sequences that span a low complexity region. All proteins from one family of homologous sequences were sorted in six groups according to the lifestyle of each species. Then, a mean value for each group of proteins was calculated. The horizontal and vertical axis show the first and second dimension of the obtained principal component analysis, and their contribution to the observed variation in the data is indicated in brackets. Each arrow illustrates the eigenvalue of each principal component according to lifestyle (nt, necrotroph; hb, hemibiotroph; fb, facultative biotroph; ob, obligate biotroph; sy, symbiont; wd, wood degrading). The color of each arrow shows its relative contribution (in percent) to the first two dimensions of the principal component analysis as indicated by the legend.

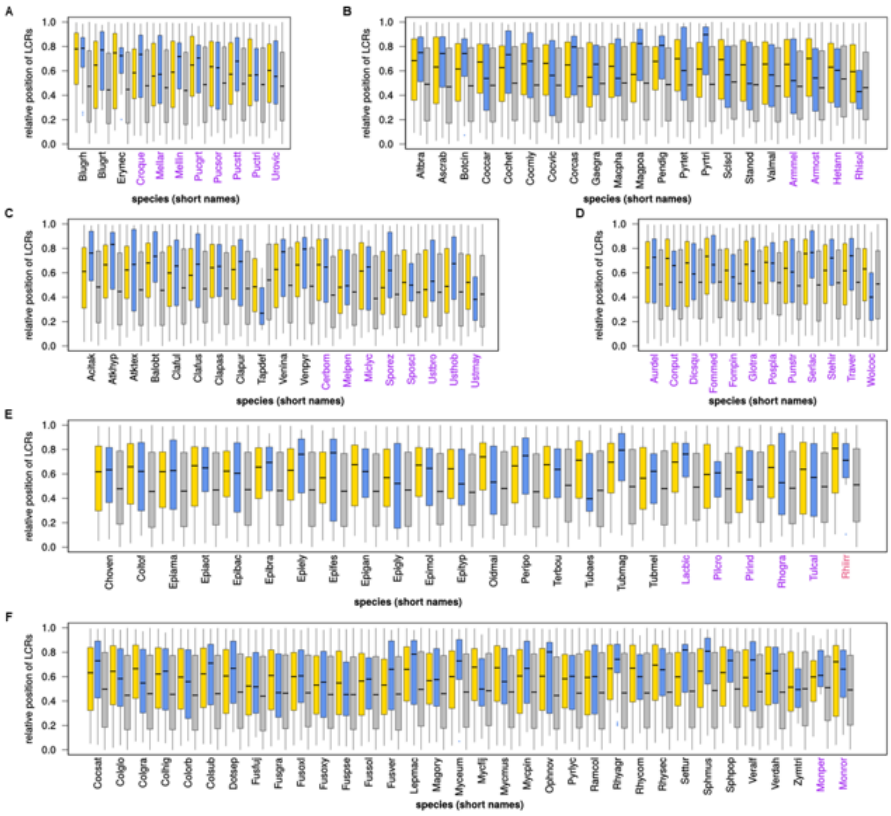

Figure 4

Relative position of low complexity regions. The relative position of each low complexity region in each protein was determined by dividing the midpoint of a low complexity region by protein length (vertical axis). Values close to 0 indicate a localization towards the N-terminus, whereas values close to 1 indicate a C-terminal localization. Data are represented separately for each protein type as indicated by different colors (grey, cytoplasmic proteins; blue, secreted non-effector proteins; yellow, effector proteins). Data are shown in form of a boxplot. The thick black line in each box indicates the median value. The lower and upper box limit indicates the values of the first and third quartile, respectively. Whiskers show data within the 1.5-fold interquartile range and dots show outliers beyond the 1.5-fold interquartile range. Each panel shows the results of species from one lifestyle, namely obligate biotroph (A), necrotroph (B), facultative biotroph (C), wood degrading (D), symbiont (E), or hemibiotroph (F). Within each panel, species are sorted by their phylum (Ascomycota, black; Basidiomycota, purple; Glomeromycota, pink). Within each phylum, species are sorted alphabetically, thereby grouping species from one genus together. The indicated short names of each species (horizontal axis) can be linked to full species names with information in supplementary table 1.

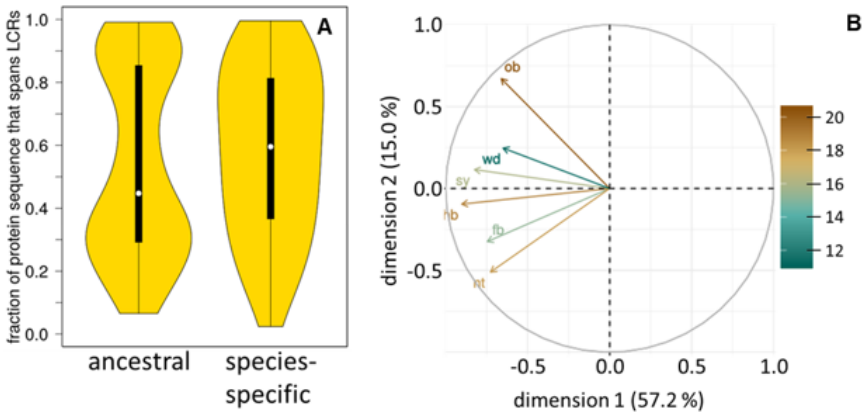

Figure 5

Relative positions of low complexity regions in effector proteins. (A) Effector proteins are classified as ancestral or species-specific (horizontal axis) and their relative position is shown on the vertical axis. Values close to 0 indicate a N-terminal localization and values close to 1 indicate a C-terminal localization. OrthoFinder was used to reconstruct groups of homologous proteins by using masked protein sequences (low complexity regions are

replaced with ‘X’ as unknown amino acid). The data distributions are shown in the form of a violin plot. In addition to the boxplot shown in black, this plot indicates the density of data at each value. (B) Principal component analysis of relative positions of low complexity regions in effector proteins. All proteins from one family of homologous sequences were sorted in six groups according to the lifestyle of each species. Then, a mean value for each group of proteins was calculated. The horizontal and vertical axis show the first and second dimension of the obtained principal component analysis, and their contribution to the observed variation in the data is indicated in brackets. Each arrow illustrates the eigenvalue of each principal component according to lifestyle (nt, necrotroph; hb, hemibiotroph; fb, facultative biotroph; ob, obligate biotroph; sy, symbiont; wd, wood degrading). The color of each arrow shows its relative contribution (in percent) to the first two dimensions of the principal component analysis as indicated by the legend.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementaryfigures.docx