Horizontal gene transfer in 44 early diverging fungi favors short, metabolic, extracellular proteins from associated bacteria

Michał Ciach*[1,3] ,Julia Pawłowska[2], Anna Muszewska*[1]

[1] Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Pawinskiego 5A, 02-106 Warsaw, Poland

[2] Department of Molecular Phylogenetics and Evolution, Faculty of Biology, Biological and Chemical Research Centre, University of Warsaw, ul. Żwirki i Wigury 101, 02-089 Warsaw, Poland;

Running title: HGT in early diverging fungal lineages

Keywords: fungi, HGT, genome architecture, fungal ecology, early diverging fungi, associated bacteria

Abstract

Numerous studies have been devoted to individual cases of horizontally acquired genes in fungi. It has been shown that such genes expand their metabolic capabilities and contribute to their adaptations as parasites or symbionts. Some studies have provided a more extensive characterization of the horizontal gene transfer (HGT) in *Dikarya*. However, in the early diverging fungi (EDF), the overall influence of HGT on the ecological adaptation and evolution is largely unknown. In order to fill this gap, we have designed a computational pipeline to obtain a sample of over 600 phylogenetic trees with evidence for recent to moderately old HGT across multiple EDF genomes ranging from *Chytridiomycota* and *Blastocladiomycota* to *Mucoromycota*. Our pipeline is designed to obtain a small sample of reliable HGT events with a possibly minimal number of false detections that distort the overall statistical patterns. We show that transfer rates differ greatly between closely related species and strains, but the ancestrally aquatic fungi are generally more likely to acquire foreign genetic material than terrestrial ones. A close ecological relationship with another organism is a predisposing condition, but does not always result in an extensive gene exchange, with some fungal lineages showing a preference for HGT from loosely associated soil bacteria.

Importance

Although it is now recognized that horizontal gene exchange is a factor influencing the adaptation and evolution of eukaryotic organisms, the so far described cases in early diverging fungi (EDF) are fragmentary, and a large-scale comprehensive study is lacking. We have designed a methodology to obtain a reliable, statistical sample of inter-kingdom xenologs across the tree of life of EDF to give a preliminary characterization of their general properties and patterns. We study how different fungal lineages vary in terms of the number of xenologs, what are their ecological associations, and the molecular properties of proteins encoded by the acquired genes. Our results help to better understand to what extent and in

what way the incorporation of foreign genetic material shaped the present biodiversity of fungi.

Introduction

Horizontal gene transfer (HGT) is a process of exchange of genetic material between distantly related species. It has been extensively studied in prokaryotes and proved to be an important mechanism that increases the genomic diversity, contributes to niche adaptation, antibiotic resistance, new metabolic pathways and numerous other traits. In eukaryotes, the significance and frequency of HGT remain mostly unclear, and the mechanisms of gene transfer beyond *Agrobacterium*-mediated gene transformation are still a matter of debate[1]. It has been shown, however, that HGT does occur both within eukaryotes (including animals) and between eukaryotes and prokaryotes [2].

Fungi represent very diverse life strategies, making them a unique model to study the connections between HGT and ecological adaptations. Genomes of fungi are protected against the incorporation of foreign DNA at multiple levels: starting from the use of alternative genetic codes (e.g. in CTG *Candida albicans* and *Pachysolen tannophilus* [3]), through foreign DNA elimination mechanisms RNAi and RIP [4], which are ubiquitous in fungi [5], to the physical separation of chromatin from the rest of the cell by the nuclear envelope. On the other hand, although the genomes of fungi are protected against the incorporation of foreign DNA at multiple levels [3–5], the osmotrophic lifestyle [6] and hyphal fusion [7,8] makes the intake of foreign DNA more common. The complex interplay of factors facilitating and preventing the incorporation of foreign DNA places fungi somewhere between bacteria and animals in terms of the rate of HGT and contributes to large differences in the numbers of HGT between different fungal lineages.

Several horizontally transferred genes have documented adaptive roles for fungal fitness. A fungal pathogen of insects, *Metarhizium robertsii*, has acquired sterol carrier and peptidase genes that facilitate the penetration of the victim's body [9,10]. Multiple transfers of toxins in plant-infecting fungi have improved their ability to infect new hosts [11,12]. Bacteria have been described as a source of extensive HGT in plant-associated *Colletotrichum* sp. [13], *Verticillium dahliae* [14] and animal skin pathogens *Malassezia* [15]. Spectacular examples of transfers of clusters of functionally-linked genes [16,17], genes facilitating metabolism expansions [18] or even conditionally dispensable chromosomes in pathogenic *Fusarium* species [19] have influenced the perception of genomic evolution in fungi. They showed that HGT can provide a faster way to gain the ability to colonize new hosts and degrade new compounds than vertical inheritance.

Existing studies of fungal HGT have mostly focused on the Dikarya fungi [20,21], and the abundance of HGT in early diverging fungi (EDF, non-Dikarya fungi) is relatively understudied. However, it is already known that it does occur and may have a profound impact on the recipient organism. Recent reports show that EDF experienced several documented waves of transfers, like the acquisition of secondary metabolite clusters in *Basidiobolus* [22], *Mortierella* [23] and *Rhizophagus* [24], or a number of enzymes transferred to *Neocallimastigomycota* involved in the metabolism of nucleotides [25] and carbohydrates [26]. A more complex scenario was reported for *Rhizophagus irregularis* that acquired genes both from the host plant and its endosymbiotic bacterium [27]. Catalase

genes of bacterial origin, which presumably increase ROS resistance, were found in the amphibian pathogen *Batrachochytrium dendrobatidis* [28]. However, despite numerous individual examples, no comprehensive characterization of HGT in EDF has been presented so far.

In this work, we study the patterns of HGT in 44 EDFs from a taxonomic, ecological, and molecular perspective. We have designed a custom phylogenetic-based pipeline to obtain a reliable sample of xenologs, focusing on limiting the number of false positive results that distort the observed signal while retaining a sufficient sample size to infer statistical trends. The pipeline is designed after a comparison of multiple existing approaches and a theoretical and experimental identification of potential pitfalls that can result in false detections. This allows us to make a comparative study of HGT in diverse EDF lineages.

Methods

Dataset

Proteomes for 44 early diverging fungi were downloaded from NCBI in April 2018 [29] all assemblies analysed in this study are listed in **Supplementary Table S1**. Fungal ecological tagging was derived from Fungal Traits [30]. The taxonomical annotation was derived from the NCBI taxonomy database, with manual fine-tuning of taxonomic names and ranks when needed according to [31]. Genomic features were obtained from GFF files. A species tree was inferred from whole proteomes with OMA2 based on OMA groups of orthologs [32].

HGT identification

Low-complexity regions in all proteins were masked using ncbi-seg [33] (window length 12, trigger complexity 1.8, extension complexity 2.0). Proteins with less than 30 non-masked amino acids, more than 2500 characters (including masked amino-acids), or more than 50% masking were discarded as prone to produce false-positive hits. Homologous families within the 55 proteomes were detected using blastp suite [29] with an e-value threshold 1e-05. Proteins with homologs in a single fungal family were retained for further analysis.

Homologous sequences for the retained proteins were found in the Non-Redundant NCBI database (last updated December 01, 2020) using blastp suite with an e-value threshold 1e-04 and maximum 500 target sequences. Detected homologs were masked using ncbi-seg (default settings). Excessively short, long, or masked (70% coverage threshold) proteins were discarded as before. Only homologs from bacteria, archaea, or eukarya were retained, in order to discard viral or artificial sequences. Taxonomic lineage of each homolog was inspected using the NCBI Taxonomy database (last updated May 17, 2021). Homologs with ’unclassified’, ‘uncultured’ or ‘environmental’ keywords anywhere in their lineage, as well as non-fungal homologs with the ‘*incertae sedis’* keyword, were discarded as prone to produce false-positive hits due to uncertain taxonomy (note that some of the fungi analyzed in this paper belong to *incertae sedis* families; their homologs were retained despite the uncertain taxonomy).

Remaining fungal sequences and their homologs were clustered using psi-cd-hit in a stepwise clustering procedure. First all homologs were clustered with cd-hit 90% identity threshold (-c 0.9 -n 5 -aL 0.4), then cd-hit 60% identity threshold (-c 0.6 -n 4 -aL 0.4) and

finally psi-cd-hit with 30% sequence identity threshold (-c 0.3 -aL 0.5 -ce 1e-5) [34]. Clusters without proteins from the 55 target proteomes, clusters without at least two non-fungal species, and clusters with proteins from less than 4 different species (in which case the species tree has an uninformative topology) were discarded.

Remaining clusters were aligned using mafft 3.7 [35] with local iterative mode (localpair, maxiterate 1000). Alignments were trimmed using GBlocks allowing for half of gap characters in each column (-b5=h) in order to remove poorly aligned regions which may disturb the phylogenetic tree topology [36]. In each alignment, sequences with more than 50% gaps after trimming were discarded. Alignments with less than 20 columns, less than 4 sequences, or more than 70% of masked characters were discarded. Maximum likelihood (ML) trees were inferred using IqTree [37] with automated model selection.

Trees in which fungi did not form a single clade were discarded. Corresponding species trees were downloaded from a local copy of the NCBI Taxonomy database (last updated August 8, 2021). The location of the fungal clade was compared in both trees using an in-house developed algorithm (described in the **Supplementary Methods**). Briefly, sets of species in neighbouring clades of the fungal clade were inspected in the gene and the species tree. The fungal clade was considered to be displaced if the sets differed between trees.

Phylogenetic trees with a single fungal clade and at least two non-fungal species were classified into three separate groups of putative HGT events. First, if the branch joining the fungal clade with the rest of the tree had zero length (indicating an identical protein), the homologous family was considered as a possible contamination. Second, if the fungal clade had the same location in the gene and the species tree, it was considered as a possible, but weakly supported, case of HGT. Finally, if the fungal clade was displaced, it was considered as a strongly supported case of HGT.

Xenolog protein products were mapped against PFAM [38] using pfamscan.pl. Additionally, proteins were scanned with SignalP for signal peptides [39] and cellular localization with WolfPsort II [40]. Statistical and phylogenetic analyses were performed with the R statistical package (https://www.R-project.org/) and the Python 3 programming language using the Jupyter Notebook [41] and numpy [42], scipy [43] and ete3 [44] libraries. Plots were created using the matplotlib [45] library.

Results

We have detected 619 gene trees containing a strong indication of a horizontal gene transfer (Fig.1., see the Methods section and the **Supplementary Methods** for details). The protein clusters corresponding to those trees jointly contained 1193 fungal proteins, out of which 972 belonged to the 44 EDF analyzed in this study. We have additionally identified 712 clusters with partial evidence for HGT, including 184 clusters with a likely contamination. Due to incomplete evidence of HGT, these clusters are not analyzed in this paper, but their list is available in the **Supplementary Table S1**.

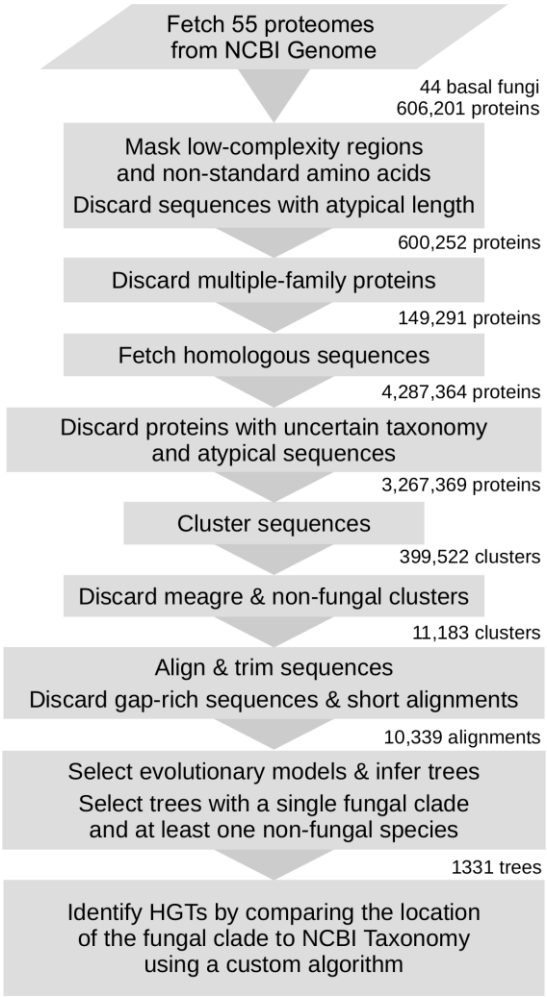


Fig. 1. Summary workflow of all of the steps in the HGT identification pipeline.

Over half of the trees contained a single fungal protein (**Fig. 2**). The largest xenologous family contained 27 proteins (including fungi from outside of this study). However, the distribution of family sizes appears to follow a power-law distribution which is characterized by infrequent occurrences of extremely large values. This suggests that large xenologous families are likely to exist, but are not present in our dataset due to our restrictive criteria.

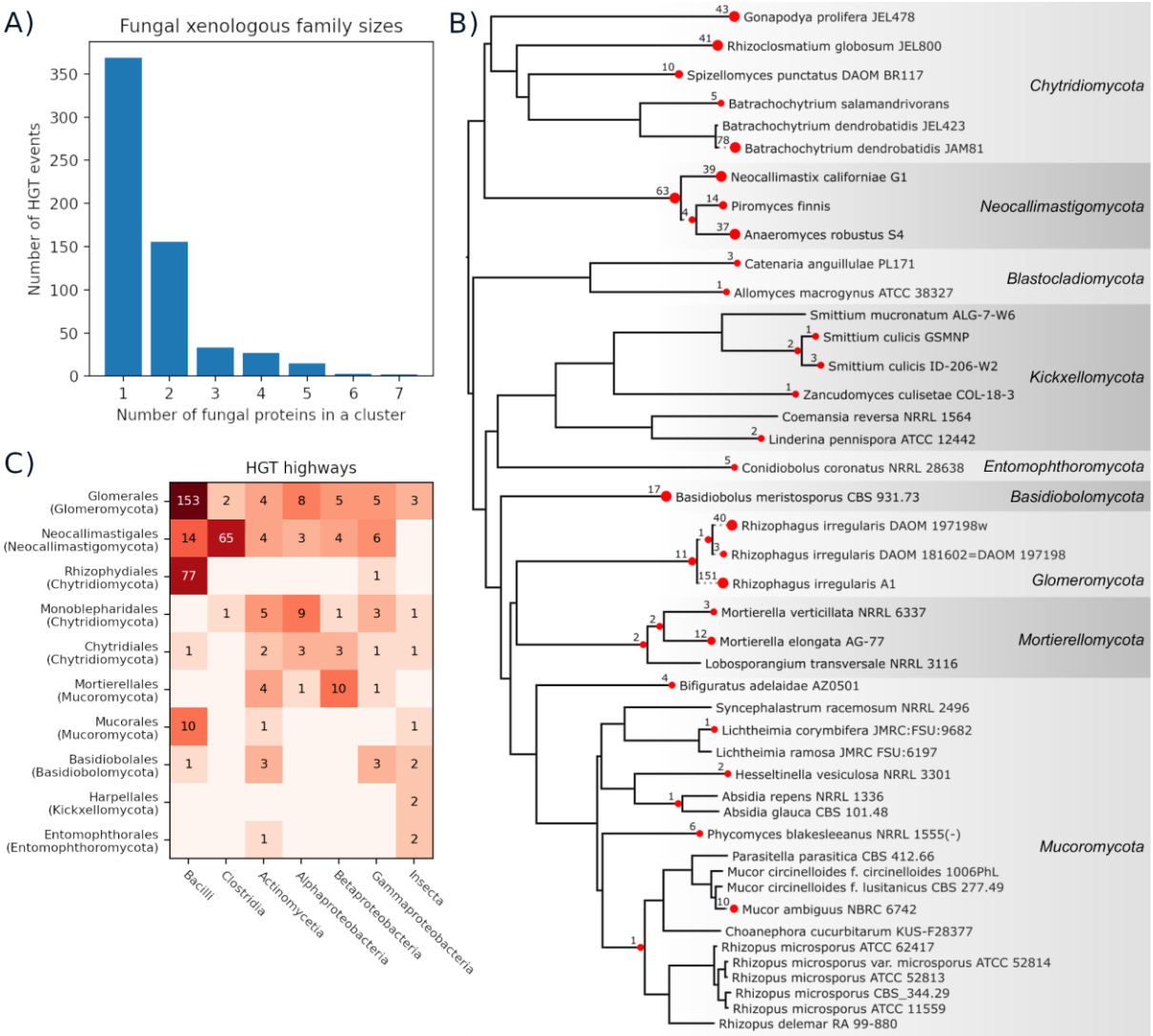


Fig. 2. The HGT patterns in 44 early diverging fungi. A) The number of fungal xenologs in a xenologous family. B) The OMA2 phylogenetic tree of fungi annotated with the numbers of detected HGTs. C) The numbers of HGT into fungi per acceptor-donor pair. Taxonomic names based on NCBI Taxonomy with manual fine-tuning according to Voigt et al. [31].

The number of xenologs detected in each fungus is depicted in **Supplementary Fig. S1**. Symbiotic species of the gut-dwelling *Neocallimastigomycota* and arbuscular mycorrhizal *Glomeromycota* contained the most xenologs in their proteomes, followed by the parasitic *Chytridiomycota*. The number of xenologs was highly divergent across fungal species, with 10 fungi from those three groups accounting for 87% of all detected xenologs. As shown in the next sections, those xenologs are of both direct origin (when the given strain was the recipient of a transfer) and the ancestral one (when the gene was transferred to an ancestor of a strain), albeit with different proportions---mostly ancestral for *Neocallimastix californiae* G1, and mostly direct for *Rhizophagus irregularis* A1.

Proteome size does not influence the xenolog count. Initially, we have observed a statistically significant correlation between the number of xenologs and proteome size (ρ=0.43, p=0.003). However, this apparent correlation was caused only by two outlying fungi

with. After discarding *R. irregularis* A1 and *N. californiae G1* from the dataset, the correlation was no longer significant (p=0.20). We conclude that there is no biologically significant correlation between the proteome size and the number of xenologs.

Transfer rates are highly divergent across the fungal tree of life. We have generated an OMA2 phylogenomic tree of the 44 EDFs and annotated its nodes with the numbers of HGTs they received, shown in **Fig. 2**. Note that the results are only an approximation of the true transfer acceptors, as the transfers can be more ancestral than the data suggests.

We have discovered a highly uneven distribution of HGTs across the fungi. There is a clear pattern dividing ancestrally aquatic taxa grouped in *Blastocladiomycota*, *Neocallimastigomycota* (living in the digestive tract of terrestrial herbivores) and *Chytridiomycota* with higher HGT rates compared to terrestrial fungi. The mainly saprotrophic *Mucoromycotina*, despite highest sampling in our dataset, shows the lowest number of HGT events per genome, just as the animal-related fungi *Entomophtoromycota* and *Kickxellomycota*.

Strains of the same species have profoundly different numbers of xenologs. *R. irregularis* A1 had the highest number of xenologous families, with 151 protein clusters containing 167 sequences, most of which were received from a bacterium similar to *Lysinibacillus massiliensis* (Firmicutes, Bacilli). Related bacteria were reported from *R. irregularis* mycorrhizosphere [46]. On the other hand, the strain *R. irregularis* DAOM 197198w was an acceptor of 40 gene families, containing 54 genes from different species of Proteobacteria, while *R. irregularis* DAOM 181602 was a direct acceptor in only three HGT events which, after duplications in this strain, resulted in 10 xenologs (not including ancestral transfers). Genomic diversity was already documented between *R. irregularis* isolates pointing at an ample spectrum of trait heterogeneity within one species [47].
In a similar manner, *Batrachochytrium dendrobatidis* JAM81 was a recipient of 78 transferred gene families resulting in 156 xenologs, mostly from different *Oceanobacillus* species, while we did not detect any transferred gene in *B. dendrobatidis* JEL423.

Transfer rates reflect physical proximity. To identify major HGT pathways, for each cluster we have matched the taxonomic order of its acceptor with the taxonomic class of the donor (**Fig. 2**, **Supplementary Fig. S2**). We have not included 42 clusters in which the donor had a higher taxonomic rank.

Among the identified pathways, the representatives of *Bacilli* and *Burkholderiaceae*, bacterial groups often associated with fungal hyphae, stand out as the main source of xenologs to distinct fungal recipients. In *Mortierellomycota* species, where *Burkholderia*-related endobacteria (BRE) are common, HGT from endosymbionts dominates. However, in *Glomeromycota* and *Chytridiomycota* fungi, which do not have associated BRE, xenologs from endohyphal bacteria were outnumbered by xenologs from loosely associated *Bacilli*. Other soil inhabiting bacteria, such as *Actinomycetes* (*Actinomycetia*), *Hyphomicrobiales* (*Alphaproteobateria*) and *Bacteroidetes* (*Fibrobacteres*, *Chlorobi*, and *Bacteroidetes* group) also contributed genes to at least four different fungal orders. Finally, we did not detect any HGT from mycoplasma endosymbionts (MRE) in *Glomeromycota*, possibly due to the highly reduced genome of these bacteria which offers less possibilities for transfer.

Our results confirm a major HGT event in the ancestor of *Neocallimastigomycota* that might have driven its evolution towards a gut-dwelling fungus [26]. This group of fungi is the only one that has obtained numerous transfers from the anaerobic *Clostridia*, which purportedly enabled them to adapt to the gut environment [26]. Moreover, *Neocallimastigomycota* obtained genes from a myriad of diverse bacteria groups, including *Fibrobacteria* (*Fibrobacteres*, *Chlorobi*, and *Bacteroidetes* group), *Negativicutes* (*Firmicutes*), and *Spirochaetia* (*Spirochaetes*) which may have contributed to their adaptive toolkit, a result that agrees with previous reports [48].

Transfers from Eukarya are less common than from Bacteria. Bacteria, with 490 xenologous families, were by far the most represented donors, followed by 128 families from Eukarya and a single transfer from Archaea (a *Gonapodya prolifera* molybdopterin binding oxidoreductase KXS19169.1). We note, however, that our method may be biased towards interkingdom transfers, as discussed at the end of this article. We have identified 19 Viridiplantae and 63 Metazoan families. Those numbers confirm an uncommon cross-kingdom transfer of genes between eukaryotes [49–51]. Interestingly, insects contributed uncharacterized genes to diverse fungal lineages, including both parasitic *Entomophthorales* and mycorrhizal *Glomeromycota*.

Xenolog sequences evolve according to ordinary substitution models. For 66% of the xenologous protein families, iqTree selected the LG model as the best-fitting one, followed by WAG, JTT and VT models (**Supplementary Fig. S3**), which jointly account for 89% of clusters. All of the four most common models are general models of amino acid substitutions suited for nuclear sequences. The less frequent models were nearly equally divided into chloroplast (3%), viral (3%), mitochondrial (2.7%) and nuclear (2.3%). This result suggests that xenolog evolution follows the same substitution patterns as most protein coding genes.

Long and repetitive sequences are rarely transferred. We have found that 49% of xenologs do not contain any low-complexity regions, compared to 33% for the proteome background. Furthermore, the distribution of the proportions of protein sequences covered by low-complexity regions has a lighter right tail for xenologs (**Fig. 3**). The results suggest that proteins with no low-complexity regions are transferred preferably, while extensive low-complexity regions tend to inhibit transfers.

A transferred fungal protein was, on average, 120 amino acids shorter than an average fungal protein. The distribution of sequence length of xenologs had a lighter right tail than the one for the background (**Fig. 3**). The two longest xenologs we have found had 1548 and 1465 amino acids (visible as a sudden increase in the right part of the qq-plot in Fig. 3), followed by 1146 amino acids and less.

Comparison of the average sequence lengths and low-complexity proportions of fungal xenologs and the donor group showed no visible trend (**Supplementary Fig. S4**). Therefore, the differences between fungal xenologs and the proteome background described above seem to influence the likelihood of an HGT instead of being caused by it.

Xenologs are secreted more often than typical fungal proteins, but determining their precise subcellular location poses difficulties. TargetP 2.0 detected a signal peptide in

14.6% of identified xenologs, compared to 8.1% for the background. This stands in agreement with the osmotrophic lifestyle of fungi and the massive fungal secretome. Notably, we have detected only 5 xenologs with mitochondrial transfer peptides, accounting for 0.5% of xenologs, compared to 3% for the whole proteomes, consistent with the bacterial origin of most xenologs. We did not observe significant differences in numbers of transmembrane helices, predicted by TMHMM, between xenologs and background proteins.
The subcellular localization predicted by WolfPsort was consistent for 87% of clusters, compared to 97% for TargetP (**Fig. 3**, **Supplementary Fig. S5**). Compared to the proteome background, we have noted a lower proportion of xenologs localized in the nucleus and the mitochondria, while a higher proportion localizes in the cytoplasm, the endoplasmic reticulum, and the extracellular space. In contrast to TargetP, WolfPSort predicts 148 proteins to localize in the mitochondria.

Interestingly, WolfPSort predicts the nucleus as the most common location of xenologs. This unexpected phenomenon could be caused by a bias of the method. To verify this hypothesis, we have generated a decoy dataset by randomly permuting each sequence from the 44 proteomes. We have found that the distribution of locations is very similar for the proteome background and the decoy dataset, with the nucleus as the most common location, and the decoy proteins attain similar prediction scores (**Supplementary Fig. 4**). This is likely caused by the fact that WolfPSort is based on the *k*-Nearest Neighbours classification algorithm, which, when faced with a new, randomly selected data point, tends to reflect the distribution of the underlying training dataset.
However, the location of transposon-derived proteins predicted by WolfPSort is, as expected, consistently nuclear. This suggests that the software classifies nuclear proteins correctly, but is biased towards nuclear/mitochondrial localization for proteins that are difficult to classify. Therefore, the actual proportions of cytosol and extracellular xenologs may be even higher.

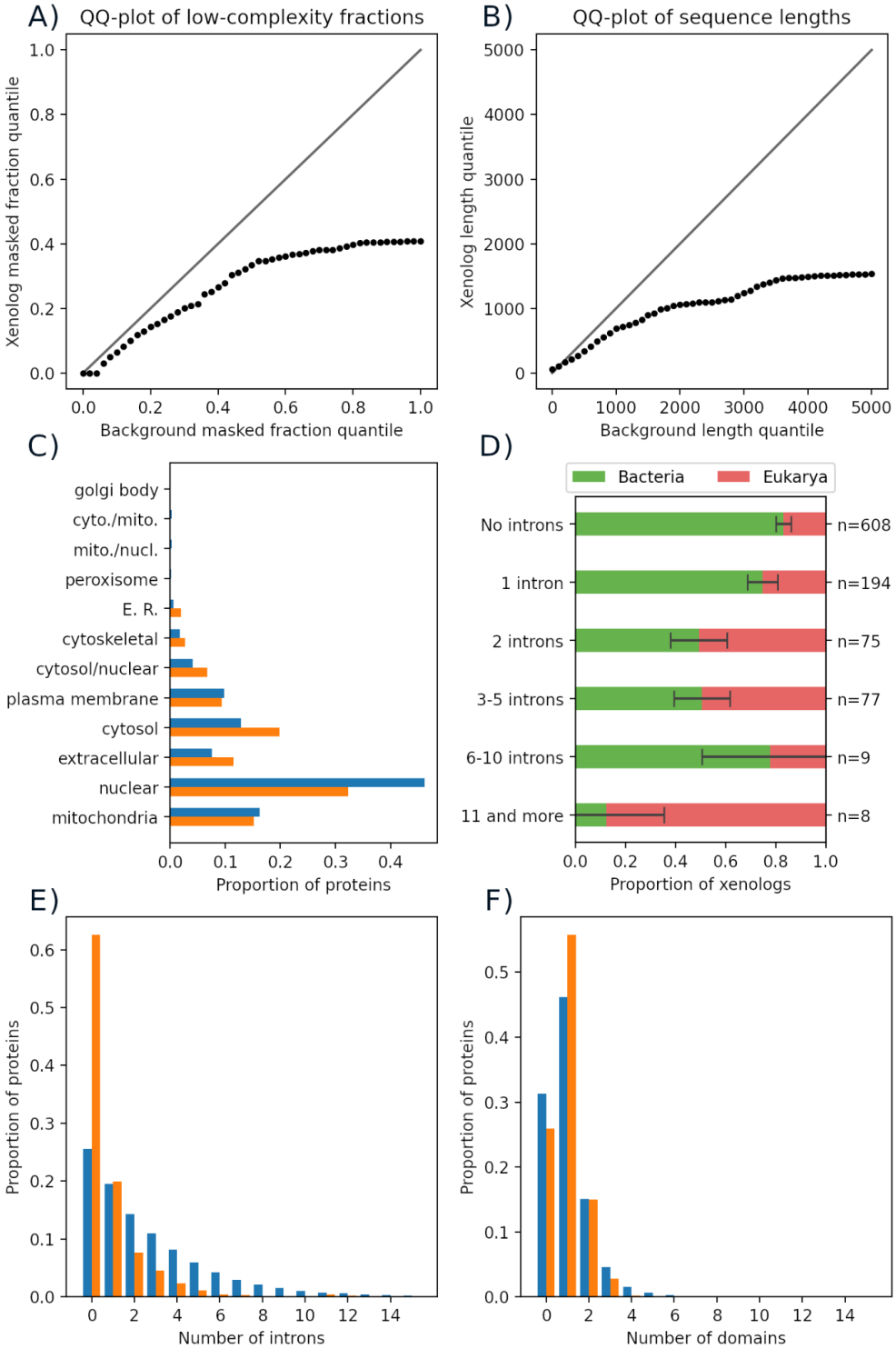


Fig. 3: Molecular characteristics of xenologs (orange) and background proteins (blue). A, B) Quantile plots comparing the sequence low-complexity fractions and lengths between the xenologs and the proteome background. C) Subcellular localizations. D) The dependence of the number of introns of a xenolog on the kingdom of the donor. E) The number of introns in xenologs and background proteins. F) Number of PFAM domains in a xenolog and a background protein.

Xenologs typically contain few introns. Approximately 63% of xenologs did not contain any intron, and 90% contained at most two, compared to 25% and 59% for the background,

respectively (**Fig. 3**). This result is in line with recommendations on HGT verification measures [13]. However, we have observed notable exceptions, especially in the case of *Gonapodya prolifera* and *Rhizoclosmatium globosum* (*Chytridiomycota*) which are particularly prone to have 4 and more introns in their xenologs. These proteins come from both unicellular eukaryotic and prokaryotic donors. We have detected 8 xenologs, mostly of eukaryotic origin, with 11 or more introns (**Fig. 3**), including two *G. prolifera* xenologs: a phosphorylase with twelve introns transferred from green algae (KXS22044.1) and a copper dependent monooxygenase with eleven introns transferred from molluscs (KXS15796.1). In both cases, the donor homologs also contained multiple introns. Although there are chytrid species referred to as *intron-rich chytrids* [52], neither *G. prolifera* nor *R. globosum* have been described as such in their genomic papers [53,54].

It has been shown that genes transferred from bacteria can acquire introns after transfer [55]. This seems to be the case in two *R. globosum* xenologs, a *Parvularcula mediterranea* glutaminase with seven introns (ORY47769.1) and a deltaproteobacterium *Labilithrix luteola* magnesium transporter with twelve introns (ORY15655.1).

Xenologs acquire domains post-transfer. Compared to the proteome background, fungal sequences obtained *via* HGT are more likely to map to one or more PFAM domains (**Fig. 3**). We have detected 542 xenologous sequences, accounting for 56% of all detected xenologs, that map to a single domain, and 178 that map to at least two. However, we have found that not all domains of fungal xenologs originate from the gene transfer event. In order to inspect whether xenologs obtain new domains after transfer, in each cluster we have looked for domains present in at least one fungal sequence and which had no clan-level siblings in any non-fungal sequence in the cluster, referred to as *novel* domains. We have detected 102 different novel domains in 99 fungal proteins spread among 59 clusters, accounting for approximately 10% of xenologs and clusters respectively, indicating that addition of a domain after HGT is widespread. The most commonly added domain was the carbohydrate binding module CBM_10 (PF02013), a part of the dockerin cellulosome typical of anaerobic fungi, acquired post-transfer in 33 fungal proteins, followed by a bacterial/plant ATPase associated domain SecA_PP_bind (PF01043) in 7 proteins and a typically eukaryotic Presenilin domain (PF01080), involved in calcium homeostasis in humans and extensively studied in the context of Alzheimer's disease, acquired after transfer in 6 proteins. The latter domain was lost in *Dikarya* but occurs in animals, plants and eukaryotic microorganisms, including the EDF.

On the other hand, we have identified *forgotten domains*, defined as domains absent in fungi, but present (or having a clan-level sibling) in at least 20% of sequences in both clades neighbouring with the fungal one in the corresponding gene tree. Such domains can be assumed to be ancestral for the whole protein cluster, but not inherited in a transfer. We have identified 103 xenologs with "forgotten domains" in 56 protein clusters. An example of a "forgotten domain" is the PKC52377.1 protein of *R. irregularis* A1, which is homologous to C-terminal fragments of *Lentibacillus* and *Staphylococcus* proteins. The bacterial homologs contain two domains, a phage regulatory domain Phage_pRha and a phage antirepressor domain ANT. Only the latter is present in the *Rhizophagus* homolog, indicating that the fragment of the gene coding for the N-terminal ANT domain has been lost during or after transfer.

Both results indicate that caution should be taken when analyzing domains of transferred proteins to elucidate how HGT has shaped the evolution of organisms. A domain found in a xenolog can be acquired after a HGT, rather than being obtained in the process, and, since HGT can lead to acquisition of gene fragments, an acquired gene can inherit only selected functions of the original gene. In total, we have found 194 proteins that had a novel or a forgotten domain (or, in a case of 8 xenologs, both). This observation indicates that approximately 20% of homologs can perform different functions in the recipient of the gene than in the donor.

Metabolic, DNA binding, and transposon domains constitute the majority of xenologs. To analyze HGT-derived protein domains of fungal xenologs, for each cluster we have selected domains which were present in at least a half of fungal proteins and in at least 20% of proteins in one of its neighbouring clades.

We have found numerous transfers of genes coding abhydrolases, acetyltransferases and esterases. Notably, many carbohydrate active enzymes, such as glycohydrolases and glycosyltransferases, were transferred to *Neocallimastigomycota*. On the other hand, peptidases were underrepresented in the dataset (peptidase xenologs n=14) compared to the carbohydrate active enzymes (n=37), but were transferred to diverse recipients, including metallopeptidases transferred from *Thiofilum flexile* to *Basidiobolus meristosporus* (ORY01793.1) and from *Micromonosporaceae* to *G. prolifera* (KXS08983.1)

Xenologs with DNA binding modules enriched the genomes of *Mucorales*, *Rhizophagus* spp. and *Basidiobolus.* We have also identified several phage and bacterial proteins without clear function in fungi (eg. spore germination protein GerD, PF17898 and Phage tail lysozyme, PF18013). Domains of unknown function (DUFs/UPFs) constitute a numerous functional class with 70 diverse Pfam families found in 102 xenologs.

In line with the reports on ubiquitous horizontal transposon transfers (HTTs), we have detected transfers of transposases to *Smittium culicis* (OMJ25253.1) from Oomycota and from *Lysinibacillus massiliensis* to *R. irregularis* (PKC52615.1). We have also observed transfers of homing endonucleases with a GIY-YIG domain from protists and bacteria to *R. irregularis* (PKC55335.1). The inteins found in *R. irregularis* seem to be distantly related to remaining fungal inteins, pointing to the mobile character of inteins in general.

We have detected several toxins, genes involved in resistance, secretion system elements, classes of genes regarded as prone to HGT. Putative betalactamase resistance genes were transferred to *G. prolifera, B. dendrobatidis* and *Spizellomyces punctatus* (eg. KXS21489.1), MFS_1 and ABC transporters to *Neocallimastigomycota* (eg. ORY49205.1). *Basidiobolus meristosporus* gained a potential resistance kinase from Dictyostelids (ORY05370.1). Several *Neocallimastigomycota* representatives gained ricin type sugar binding modules (eg. ORX83988.1). Similar proteins were recently documented in *Mucor* acting as toxins (mucoricin) [56]. Interestingly, a Zonular occludens toxin (Zot) was transferred from a *Mycobacterium* or *Bacillus* bacterium to *Rhizophagus irregularis* (EXX79716.1) and an insect-specific toxin with Toxin_30 domain (PF08117) has been transferred from *Pristhesancus plagipennis* to *Condiobolus coronatus* (KXN66526.1).

Mortierellomycota acquired part of a T3SS system (OAQ25560.1) from a *Betaproteobacteria*-related bacterium (MBM3372935.1). Those systems were shown to play a role in the interaction between fungus *Rhizopus microsporus* and their endohyphal bacteria *Mycetohabitans rhizoxinica (*syn. *Burkholderia rhizoxynica)* [57].

Discussion

Our aim in this work was to provide a general, statistical characterization of the properties of HGT across diverse EDF. To design a pipeline appropriate for the context of this work, we have compared multiple existing approaches with respect to their applicability to a high-throughput screening for xenologs. We have then analyzed their key steps both theoretically and experimentally, by analyzing the results of each step and checking how often it produces false discoveries. We have concluded that multiple factors may inflate the false positive rate when HGT detection is done in a high-throughput setting, including:

- Relying on sequence similarity instead of phylogenetic trees,
- Excessively long or short sequences,
- Viral and environmental homologs,
- Choice of the clustering algorithm,
- Low-quality alignments and wrong choice of the trimming software,
- Methods relying on rooted gene trees,
- Using reconciliation algorithms with improper event weights,
- Complex gene trees with multiple fungal clades.

How and why each of the above steps influences the false positive rate is described and discussed in detail in the **Supplementary Methods**. In order to avoid them, we have developed a custom-made algorithm for detection of gene trees with an unexpected location of the fungal clade, which allows us to detect HGT events and identify gene donors in unrooted trees. Due to the application of multiple filtering criteria and avoidance of risky steps in the pipeline, our data should provide a reliable source of fungal xenologs which can further be used to benchmark new HGT identification methods (**Supplementary Dataset**).

Our method has provided us with a dataset of xenologs that allows us to draw general conclusions about the nature of the moderately recent HGT into EDF. The receptiveness for foreign genes varies greatly between fungal lineages, with bacteria being the most common source of xenologs. Some lineages of EDF have associated bacteria, either endohyphal or acquired from the environment, which have been pointed as likely donors of xenologs [58]. There are several documented cases where the prokaryotic symbionts are co-evolving with the fungal host and transmitted vertically [59]. Such close endosymbiotic relations pose a natural opportunity for a large-scale gene exchange, with the prominent example of a massive gene transfer from the mitochondrion to the nucleus of eukaryotic organisms [60]. However, the majority of detected transfers originated from the soil-dwelling bacteria from the *Bacilli* group, which are likely to be more loosely associated with the fungi. Nearly all fungal lineages that experienced massive transfers have acquired some genes from this group, with a notable exception of *Mortierellomycota* fungi, for which the majority of xenologs did originate from the likely endohyphal *Betaproteobacteria* (BRE).

Our results confirm the reports on a massive enzyme encoding gene acquisition by *Neocallimastigomycota [26]* which has likely played a role in their transition to gut-dwelling fungi. We also complement the reports of an acquisition of catalases by *Batrachochytrium*

dendrobatidis [28]. The *B. dendrobatidis* catalase reported in the original study (EGF75958.1, identical to XP_006683415.1) was not included in our dataset because it contains distantly related homologs in numerous fungi. On the other hand, we have detected transfers of an ABC transporter, a betalactamase, and a bacteriocin resistance cysteine peptidase (C39), all from species of *Bacilli*. Together with the reports on the catalase transfer, those results show an enrichment of diverse aspects of *B. dendrobatidis* metabolism through HGT.

In agreement with previous reports, we have found that fungi belonging to ancestrally aquatic groups experienced more HGT events than terrestrial lineages. This can be explained by the easier exchange of genetic material in the aquatic ecosystem, the ubiquity of unicellular aquatic life forms, and the possibility of forming biofilms with bacteria [61–63]. Another possible explanation is that terrestrial fungi may have lost some of the HGT predisposing traits or gained defensive mechanisms preventing foreign gene incorporation.

Despite the impact of the aquatic environment, the relationship with the host and the necessity to adapt to a new environment seem to have a higher influence on the rate of HGT. Within fungi with at least 10 xenologs, most have a documented long-term association with another organism. On the other hand, many mycorrhizal fungi did not experience extensive transfers, in some cases even none at all. These observations suggest that, while symbiosis or parasitism does not always result in gene exchange, it is a predisposing condition.

We have observed clear patterns in the properties of protein products of transferred genes. Xenologs are characterized by relatively short length and short or no low-complexity regions. Regions with skewed amino acid abundance constitute 20% of eukaryotic proteins [64] and are involved in diverse molecular functions usually linked to regulation [65] and protein-protein interactions [66]. The depletion of LCRs and short sequence preference can be related to the so-called *complexity hypothesis*, which states that xenologs tend to be far from central hubs of interaction networks [67], and instead participate in secondary metabolism, drug resistance etc. In agreement with the current notions, the xenologs in our dataset tend to contain few introns. We have, however, detected notable exceptions to this rule that may result from the acquisition of intron-rich genes or intronization of xenologs in chytrids.

We have found no correlation between the number of xenologs and proteome size. This implies that HGT seems to have no influence on the proteome sizes of EDF. Moreover, large proteomes do not predispose to HGT and the overall amount of genetic material does not seem to be a limiting factor for incorporation of transferred genes. The number of xenologs in a proteome does not seem to be subject to evolutionary selection even in the case of parasites with a reduced genome.

Acknowledgements

This work was supported by National Science Centre grants (#2017/25/B/NZ2/01880 to AM, #2019/33/B/ST6/00737 to Paweł Górecki supporting M.C., #2017/25/B/NZ8/00473 to J.P.)

Author Contributions

M.C and A.M. designed the study, M.C. implemented software, M.C. and A.M. prepared the dataset and performed genome analyses, M.C. performed statistical analyses, and M.C., J.P. and A.M. interpreted the data and wrote the manuscript.

References

1. Naranjo-Ortiz MA, Gabaldón T. Fungal evolution: cellular, genomic and metabolic complexity. Biol Rev Camb Philos Soc. 2020;95: 1198–1232.

2. Gabaldón T. Patterns and impacts of nonvertical evolution in eukaryotes: a paradigm shift. Ann N Y Acad Sci. 2020;1476: 78–92.

3. Krassowski T, Coughlan AY, Shen X-X, Zhou X, Kominek J, Opulente DA, et al. Evolutionary instability of CUG-Leu in the genetic code of budding yeasts. Nat Commun. 2018;9: 1887.

4. Clutterbuck AJ, John Clutterbuck A. Genomic evidence of repeat-induced point mutation (RIP) in filamentous ascomycetes. Fungal Genetics and Biology. 2011. pp. 306–326. doi:10.1016/j.fgb.2010.09.002

5. Torres-Martínez S, Ruiz-Vázquez RM. The RNAi Universe in Fungi: A Varied Landscape of Small RNAs and Biological Functions. Annu Rev Microbiol. 2017;71: 371–391.

6. Richards TA, Talbot NJ. Horizontal gene transfer in osmotrophs: playing with public goods. Nature Reviews Microbiology. 2013. pp. 720–727. doi:10.1038/nrmicro3108

7. Priest SJ, Yadav V, Heitman J. Advances in understanding the evolution of fungal genome architecture. F1000Res. 2020;9. doi:10.12688/f1000research.25424.1

8. Feurtey A, Stukenbrock EH. Interspecific Gene Exchange as a Driver of Adaptive Evolution in Fungi. Annu Rev Microbiol. 2018;72: 377–398.

9. Zhao H, Xu C, Lu H-L, Chen X, St. Leger RJ, Fang W. Host-to-Pathogen Gene Transfer Facilitated Infection of Insects by a Pathogenic Fungus. PLoS Pathogens. 2014. p. e1004009. doi:10.1371/journal.ppat.1004009

10. Zhang Q, Chen X, Xu C, Zhao H, Zhang X, Zeng G, et al. Horizontal gene transfer allowed the emergence of broad host range entomopathogens. Proc Natl Acad Sci U S A. 2019;116: 7982–7989.

11. Mehrabi R, Bahkali AH, Abd-Elsalam KA, Moslem M, Ben M'Barek S, Gohari AM, et al. Horizontal gene and chromosome transfer in plant pathogenic fungi affecting host range. FEMS Microbiology Reviews. 2011. pp. 542–554. doi:10.1111/j.1574-6976.2010.00263.x

12. Ambrose KV, Koppenhöfer AM, Belanger FC. Horizontal gene transfer of a bacterial insect toxin gene into the Epichloë fungal symbionts of grasses. Sci Rep. 2014;4: 5562.

13. Jaramillo VDA, Armijos Jaramillo VD, Sukno SA, Thon MR. Identification of horizontally transferred genes in the genus Colletotrichum reveals a steady tempo of bacterial to fungal gene transfer. BMC Genomics. 2015. doi:10.1186/1471-2164-16-2

14. Shi-Kunne X, van Kooten M, Depotter JRL, Thomma BPHJ, Seidl MF. The Genome of the Fungal Pathogen Verticillium dahliae Reveals Extensive Bacterial to Fungal Gene Transfer. Genome Biol Evol. 2019;11: 855–868.

15. Ianiri G, Coelho MA, Ruchti F, Sparber F, McMahon TJ, Fu C, et al. HGT in the human and skin commensalMalassezia: A bacterially derived flavohemoglobin is required for NO resistance and host interaction. Proceedings of the National Academy of Sciences. 2020. pp. 15884–15894. doi:10.1073/pnas.2003473117

16. Wisecaver JH, Rokas A. Fungal metabolic gene clustersâ€"caravans traveling across genomes and environments. Frontiers in Microbiology. 2015. doi:10.3389/fmicb.2015.00161

17. Wisecaver JH, Slot JC, Rokas A. The evolution of fungal metabolic pathways. PLoS Genet. 2014;10: e1004816.

18. Gao S, Gold SE, Wisecaver JH, Zhang Y, Guo L, Ma L-J, et al. Genome-wide analysis of Fusarium verticillioides reveals inter-kingdom contribution of horizontal gene transfer to the expansion of metabolism. Fungal Genet Biol. 2019;128: 60–73.

19. Ma L-J, van der Does HC, Borkovich KA, Coleman JJ, Daboussi M-J, Di Pietro A, et al. Comparative genomics reveals mobile pathogenicity chromosomes in Fusarium. Nature. 2010. pp. 367–373. doi:10.1038/nature08850

20. Richards TA, Leonard G, Soanes DM, Talbot NJ. Gene transfer into the fungi. Fungal Biology Reviews. 2011. pp. 98–110. doi:10.1016/j.fbr.2011.04.003

21. Marcet-Houben M, Gabaldón T. Acquisition of prokaryotic genes by fungal genomes. Trends Genet. 2010;26: 5–8.

22. Tabima JF, Trautman IA, Chang Y, Wang Y, Mondo S, Kuo A, et al. Phylogenomic Analyses of Non-Dikarya Fungi Supports Horizontal Gene Transfer Driving Diversification of Secondary Metabolism in the Amphibian Gastrointestinal Symbiont,. G3 . 2020;10: 3417–3433.

23. Wurlitzer JM, Stanišić A, Wasmuth I, Jungmann S, Fischer D, Kries H, et al. Bacterial-Like Nonribosomal Peptide Synthetases Produce Cyclopeptides in the Zygomycetous Fungus Mortierella alpina. Applied and Environmental Microbiology. 2021. doi:10.1128/aem.02051-20

24. Venice F, Desirò A, Silva G, Salvioli A, Bonfante P. The Mosaic Architecture of NRPS-PKS in the Arbuscular Mycorrhizal Fungus Gigaspora margarita Shows a Domain With Bacterial Signature. Frontiers in Microbiology. 2020. doi:10.3389/fmicb.2020.581313

25. Alexander WG, Wisecaver JH, Rokas A, Hittinger CT. Horizontally acquired genes in early-diverging pathogenic fungi enable the use of host nucleosides and nucleotides. Proc Natl Acad Sci U S A. 2016;113: 4116–4121.

26. Murphy CL, Youssef NH, Hanafy RA, Couger MB, Stajich JE, Wang Y, et al. Horizontal Gene Transfer as an Indispensable Driver for Evolution of Neocallimastigomycota into a Distinct Gut-Dwelling Fungal Lineage. Appl Environ Microbiol. 2019;85. doi:10.1128/AEM.00988-19

27. Li M, Zhao J, Tang N, Sun H, Huang J. Horizontal Gene Transfer From Bacteria and Plants to the Arbuscular Mycorrhizal Fungus Rhizophagus irregularis. Frontiers in Plant Science. 2018. doi:10.3389/fpls.2018.00701

28. Hansberg W, Salas-Lizana R, Domínguez L. Fungal catalases: function, phylogenetic origin and structure. Arch Biochem Biophys. 2012;525: 170–180.

29. NCBI Resource Coordinators. Database Resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2017;45: D12–D17.

30. Põlme S, Abarenkov K, Henrik Nilsson R, Lindahl BD, Clemmensen KE, Kauserud H, et al. FungalTraits: a user-friendly traits database of fungi and fungus-like stramenopiles. Fungal Diversity. 2020. pp. 1–16. doi:10.1007/s13225-020-00466-2

31. Voigt K, James TY, Kirk PM, Santiago ALCM de A, Waldman B, Griffith GW, et al. Early-diverging fungal phyla: taxonomy, species concept, ecology, distribution, anthropogenic impact, and novel phylogenetic proposals. Fungal Divers. 2021; 1–40.

32. Altenhoff AM, Levy J, Zarowiecki M, Tomiczek B, Warwick Vesztrocy A, Dalquen DA, et al. OMA standalone: orthology inference among public and custom genomes and transcriptomes. Genome Res. 2019;29: 1152–1163.

33. Wootton JC, Federhen S. Statistics of local complexity in amino acid sequences and sequence databases. Computers & Chemistry. 1993. pp. 149–163. doi:10.1016/0097-8485(93)85006-x

34. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28: 3150–3152.

35. Nakamura T, Yamada KD, Tomii K, Katoh K. Parallelization of MAFFT for large-scale multiple sequence alignments. Bioinformatics. 2018. pp. 2490–2492. doi:10.1093/bioinformatics/bty121

36. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000;17: 540–552.

37. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol. 2020;37: 1530–1534.

38. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic Acids Research. 2019. pp. D427–D432. doi:10.1093/nar/gky995

39. Armenteros JJA, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nature Biotechnology. 2019. pp. 420–423. doi:10.1038/s41587-019-0036-z

40. Horton P, Park K-J, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, et al. WoLF PSORT: protein localization predictor. Nucleic Acids Res. 2007;35: W585–7.

41. Beg M, Taka J, Kluyver T, Konovalov A, Ragan-Kelley M, Thiery NM, et al. Using Jupyter for Reproducible Scientific Workflows. Computing in Science & Engineering. 2021. pp. 36–46. doi:10.1109/mcse.2021.3052101

42. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature. 2020;585: 357–362.

43. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods. 2020. pp. 261–272. doi:10.1038/s41592-019-0686-2

44. Huerta-Cepas J, Serra F, Bork P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol Biol Evol. 2016;33: 1635–1638.

45. Bisong E. Matplotlib and Seaborn. Building Machine Learning and Deep Learning Models on Google Cloud Platform. 2019. pp. 151–165. doi:10.1007/978-1-4842-4470-8_12

46. Taktek S, Trépanier M, Servin PM, St-Arnaud M, Piché Y, Fortin J-A, et al. Trapping of phosphate solubilizing bacteria on hyphae of the arbuscular mycorrhizal fungus Rhizophagus irregularis DAOM 197198. Soil Biology and Biochemistry. 2015. pp. 1–9. doi:10.1016/j.soilbio.2015.07.016

47. Chen ECH, Morin E, Beaudet D, Noel J, Yildirir G, Ndikumana S, et al. High intraspecific genome diversity in the model arbuscular mycorrhizal symbiont Rhizophagus irregularis. New Phytol. 2018;220: 1161–1171.

48. Alexandre FS, Della Flora LV, Henrique IG, da Silva DC, Mercedes AP, Silva AC, et al. Arbuscular Mycorrhizal Fungi (Rhizophagus clarus) and Rhizobacteria (Bacillus subtilis) Can Improve the Clonal Propagation and Development of Teak for Commercial Plantings. Frontiers in Plant Science. 2021. doi:10.3389/fpls.2021.628769

49. Schönknecht G, Weber APM, Lercher MJ. Horizontal gene acquisitions by eukaryotes as drivers of adaptive evolution. Bioessays. 2014;36: 9–20.

50. Van Etten J, Bhattacharya D. Horizontal Gene Transfer in Eukaryotes: Not if, but How Much? Trends in Genetics. 2020. pp. 915–925. doi:10.1016/j.tig.2020.08.006

51. Richards TA, Soanes DM, Foster PG, Leonard G, Thornton CR, Talbot NJ. Phylogenomic analysis demonstrates a pattern of rare and ancient horizontal gene transfer between plants and fungi. Plant Cell. 2009;21: 1897–1911.

52. van de Vossenberg BTLH, Warris S, Nguyen HDT, van Gent-Pelzer MPE, Joly DL, van de Geest HC, et al. Comparative genomics of chytrid fungi reveal insights into the obligate biotrophic and pathogenic lifestyle of Synchytrium endobioticum. Sci Rep. 2019;9: 8672.

53. Chang Y, Wang S, Sekimoto S, Aerts AL, Choi C, Clum A, et al. Phylogenomic Analyses Indicate that Early Fungi Evolved Digesting Cell Walls of Algal Ancestors of Land Plants. Genome Biol Evol. 2015;7: 1590–1601.

54. Mondo SJ, Dannebaum RO, Kuo RC, Louie KB, Bewick AJ, LaButti K, et al. Widespread adenine N6-methylation of active genes in fungi. Nat Genet. 2017;49: 964–968.

55. Lage J-LD, Da Lage J-L, Binder M, Hua-Van A, Janeček Š, Casane D. Gene make-up: rapid and massive intron gains after horizontal transfer of a bacterial α-amylase gene to Basidiomycetes. BMC Evolutionary Biology. 2013. p. 40. doi:10.1186/1471-2148-13-40

56. Soliman SSM, Baldin C, Gu Y, Singh S, Gebremariam T, Swidergall M, et al. Mucoricin is a ricin-like toxin that is critical for the pathogenesis of mucormycosis. Nature Microbiology. 2021. pp. 313–326. doi:10.1038/s41564-020-00837-0

57. Carter ME, Carpenter SCD, Dubrow ZE, Sabol MR, Rinaldi FC, Lastovetsky OA, et al. A TAL effector-like protein of an endofungal bacterium increases the stress tolerance and alters the transcriptome of the host. Proceedings of the National Academy of Sciences. 2020. pp. 17122–17129. doi:10.1073/pnas.2003857117

58. Naito M, Morton JB, Pawlowska TE. Minimal genomes of mycoplasma-related endobacteria are plastic and contain host-derived genes for sustained life within Glomeromycota. Proc Natl Acad Sci U S A. 2015;112: 7791–7796.

59. Pawlowska TE, Gaspar ML, Lastovetsky OA, Mondo SJ, Real-Ramirez I, Shakya E, et al. Biology of Fungi and Their Bacterial Endosymbionts. Annu Rev Phytopathol. 2018;56: 289–309.

60. Anselmetti Y, El-Mabrouk N, Lafond M, Ouangraoua A. Gene tree and species tree reconciliation with endosymbiotic gene transfer. Bioinformatics. 2021. pp. i120–i132. doi:10.1093/bioinformatics/btab328

61. McDaniel LD, Young E, Delaney J, Ruhnau F, Ritchie KB, Paul JH. High frequency of horizontal gene transfer in the oceans. Science. 2010;330: 50.

62. Grüll MP, Mulligan ME, Lang AS. Small extracellular particles with big potential for horizontal gene transfer: membrane vesicles and gene transfer agents. FEMS Microbiology Letters. 2018. doi:10.1093/femsle/fny192

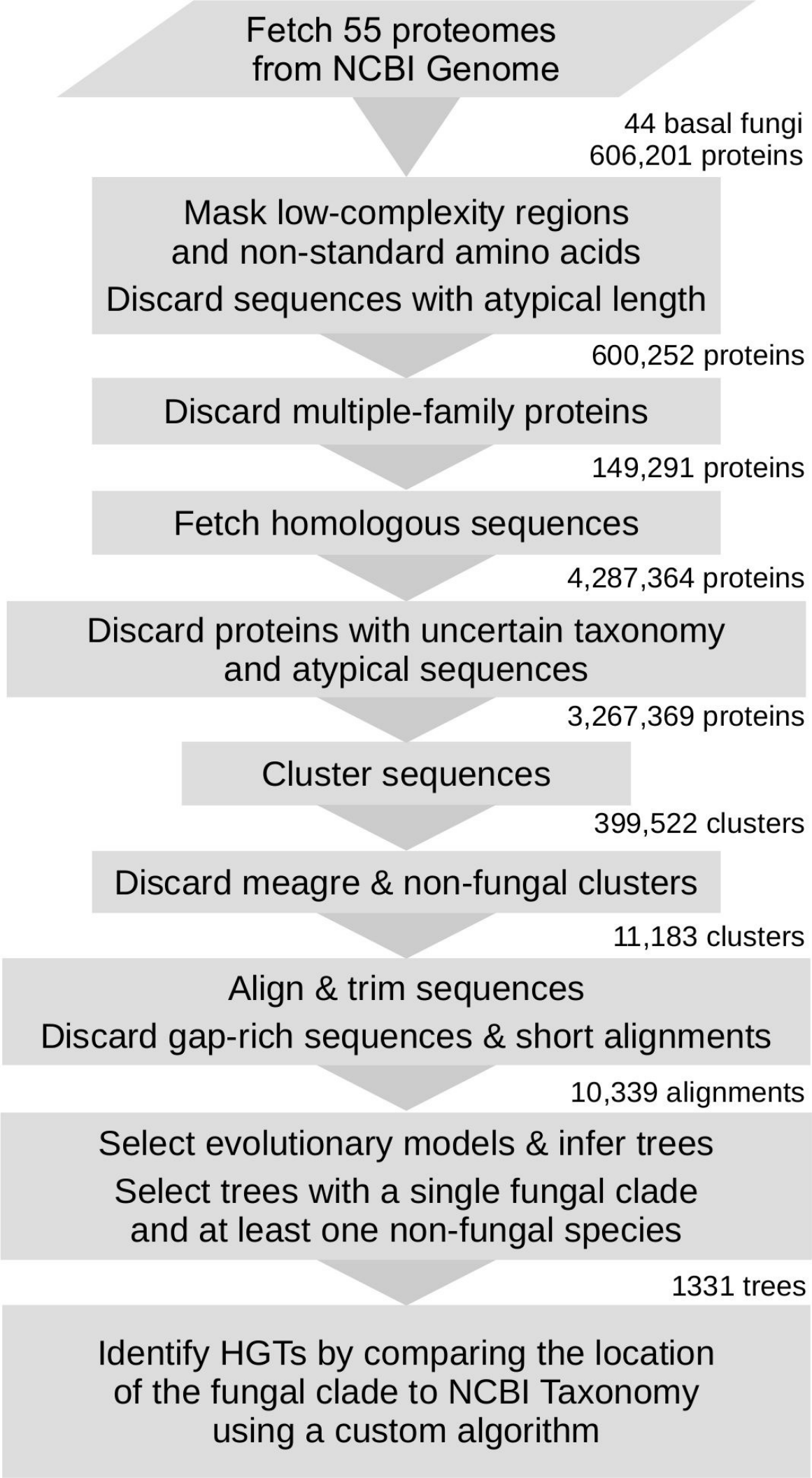
63. Abe K, Nomura N, Suzuki S. Biofilms: hot spots of horizontal gene transfer (HGT) in aquatic environments, with a focus on a new HGT mechanism. FEMS Microbiol Ecol. 2020;96. doi:10.1093/femsec/fiaa031

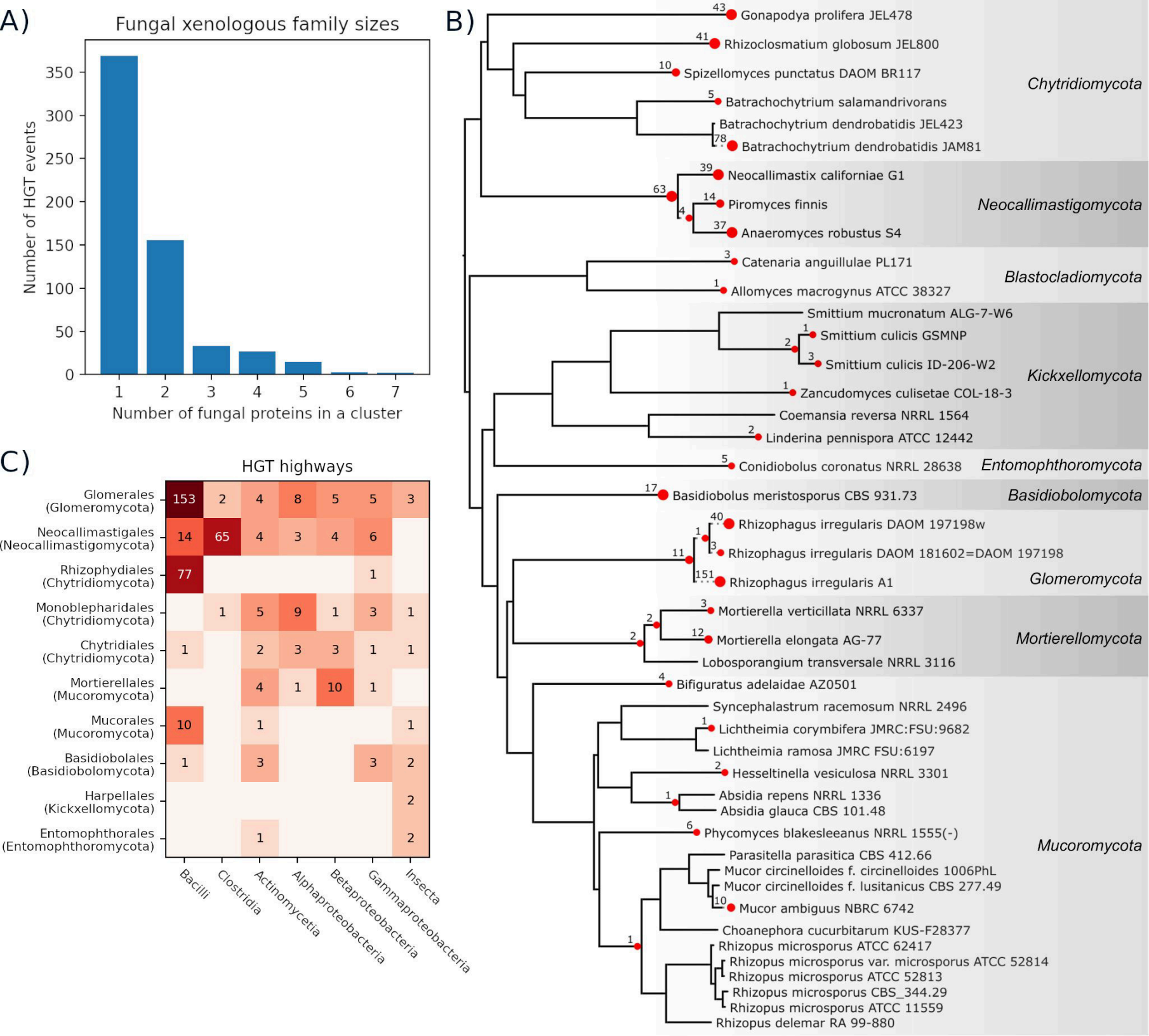
64. Mier P, Paladin L, Tamana S, Petrosian S, Hajdu-Soltész B, Urbanek A, et al. Disentangling the complexity of low complexity proteins. Brief Bioinform. 2020;21: 458–472.

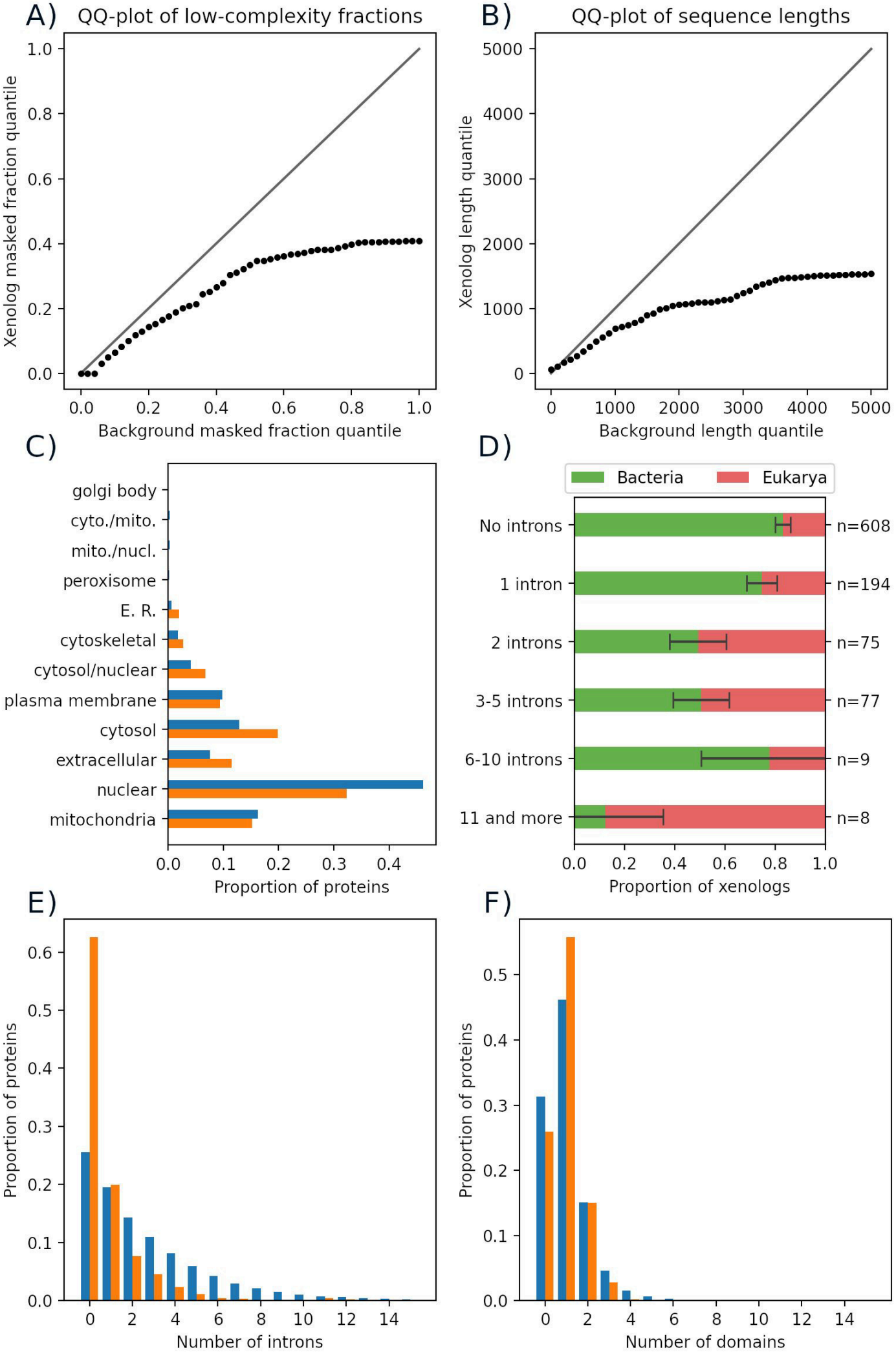
65. Wright PE, Dyson HJ. Intrinsically disordered proteins in cellular signalling and regulation. Nat Rev Mol Cell Biol. 2015;16: 18–29.

66. Coletta A, Pinney JW, Weiss Solís DY, Marsh J, Pettifer SR, Attwood TK. Low-complexity regions within protein sequences have position-dependent roles. BMC Systems Biology. 2010. doi:10.1186/1752-0509-4-43

67. Cohen O, Gophna U, Pupko T. The complexity hypothesis revisited: connectivity rather than function constitutes a barrier to horizontal gene transfer. Mol Biol Evol. 2011;28: 1481–1489.


Fetch 55 proteomes
from NCBI Genome
44 basal fungi
606,201 proteins
Mask low-complexity regions
and non-standard amino acids
Discard sequences with atypical length
600,252 proteins
Discard multiple-family proteins
149,291 proteins
Fetch homologous sequences
4,287,364 proteins
Discard proteins with uncertain taxonomy
and atypical sequences
3,267,369 proteins
Cluster sequences
399,522 clusters
Discard meagre & non-fungal clusters
11,183 clusters
Align & trim sequences
Discard gap-rich sequences & short alignments
10,339 alignments
Select evolutionary models & infer trees
Select trees with a single fungal clade
and at least one non-fungal species
1331 trees
Identify HGTs by comparing the location
of the fungal clade to NCBI Taxonomy
using a custom algorithm


A)
Fungal xenologous family sizes
Number of HGT events
0
50
100
150
200
250
300
350
1
2
3
4
5
6
7
Number of fungal proteins in a cluster
B)
Gonapodya prolifera JEL478
Rhizoclosmatium globosum JEL800
Spizellomyces punctatus DAOM BR117
Batrachochytrium salamandrivorans
Batrachochytrium dendrobatidis JEL423
Batrachochytrium dendrobatidis JAM81
Chytridiomycota
Neocallimastix californiae G1
Piromyces finnis
Anaeromyces robustus S4
Neocallimastigomycota
Catenaria anguillulae PL171
Allomyces macrogynus ATCC 38327
Blastocladiomycota
Smittium mucronatum ALG-7-W6
Smittium culicis GSMNP
Smittium culicis ID-206-W2
Zancudomyces culisetae COL-18-3
Coemansia reversa NRRL 1564
Linderina pennispora ATCC 12442
Kickxellomycota
Conidiobolus coronatus NRRL 28638
Entomophthoromycota
Basidiobolus meristosporus CBS 931.73
Basidiobolomycota
Rhizophagus irregularis DAOM 197198w
Rhizophagus irregularis DAOM 181602=DAOM 197198
Rhizophagus irregularis A1
Glomeromycota
Mortierella verticillata NRRL 6337
Mortierella elongata AG-77
Lobosporangium transversale NRRL 3116
Mortierellomycota
Bifiguratus adelaidae AZ0501
Syncephalastrum racemosum NRRL 2496
Lichtheimia corymbifera JMRC:FSU:9682
Lichtheimia ramosa JMRC FSU:6197
Hesseltinella vesiculosa NRRL 3301
Absidia repens NRRL 1336
Absidia glauca CBS 101.48
Phycomyces blakesleeanus NRRL 1555(-)
Parasitella parasitica CBS 412.66
Mucor circinelloides f. circinelloides 1006PhL
Mucor circinelloides f. lusitanicus CBS 277.49
Mucor ambiguus NBRC 6742
Choanephora cucurbitarum KUS-F28377
Rhizopus microsporus ATCC 62417
Rhizopus microsporus var. microsporus ATCC 52814
Rhizopus microsporus ATCC 52813
Rhizopus microsporus CBS_344.29
Rhizopus microsporus ATCC 11559
Rhizopus delemar RA 99-880
Mucoromycota
C)
HGT highways
Glomerales (Glomeromycota)
Neocallimastigales (Neocallimastigomycota)
Rhizophydiales (Chytridiomycota)
Monoblepharidales (Chytridiomycota)
Chytridiales (Chytridiomycota)
Mortierellales (Mucoromycota)
Mucorales (Mucoromycota)
Basidiobolales (Basidiobolomycota)
Harpellales (Kickxellomycota)
Entomophthorales (Entomophthoromycota)
Bacilli
Clostridia
Actinomycetia
Alphaproteobacteria
Betaproteobacteria
Gammaproteobacteria
Insecta


A)
QQ-plot of low-complexity fractions
Xenolog masked fraction quantile
Background masked fraction quantile
B)
QQ-plot of sequence lengths
Xenolog length quantile
Background length quantile
C)
golgi body
cyto./mito.
mito./nucl.
peroxisome
E. R.
cytoskeletal
cytosol/nuclear
plasma membrane
cytosol
extracellular
nuclear
mitochondria
Proportion of proteins
D)
Bacteria
Eukarya
No introns
1 intron
2 introns
3-5 introns
6-10 introns
11 and more
n=608
n=194
n=75
n=77
n=9
n=8
Proportion of xenologs
E)
Proportion of proteins
Number of introns
F)
Proportion of proteins
Number of domains