Patterns of Conservation and Diversification in the Fungal Polarization Network

Eveline T. Diepeveen[1], Thies Gehrmann[2,3], Valérie Pourquié[1,2], Thomas Abeel[2,4], and Liedewij Laan[1,*]

[1]Department of Bionanoscience, Faculty of Applied Sciences, Kavli Institute of NanoScience, Delft University of Technology, The Netherlands

[2]Delft Bioinformatics Lab, Faculty of Electrical Engineering, Mathematics and Computer Science, Intelligent Systems, Delft University of Technology, The Netherlands

[3]Department of Molecular Epidemiology, Leiden Computational Biology Center, Leiden University Medical Centre, The Netherlands

[4]Genome Sequencing and Analysis Program, Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts

*Corresponding author: E-mail: l.laan@tudelft.nl.

Accepted: June 18, 2018

Abstract

The combined actions of proteins in networks underlie all fundamental cellular functions. Deeper insights into the dynamics of network composition across species and their functional consequences are crucial to fully understand protein network evolution. Large-scale comparative studies with high phylogenetic resolution are now feasible through the recent rise in available genomic data sets of both model *and* nonmodel species. Here, we focus on the polarity network, which is universally essential for cell proliferation and studied in great detail in the model organism, *Saccharomyces cerevisiae*. We examine 42 proteins, directly related to cell polarization, across 298 fungal strains/species to determine the composition of the network and patterns of conservation and diversification. We observe strong protein conservation for a group of 23 core proteins: >95% of all examined strains/species possess at least 14 of these core proteins, albeit in varying compositions, and non of the individual core proteins is 100% conserved. We find high levels of variation in prevalence and sequence identity in the remaining 19 proteins, resulting in distinct lineage-specific compositions of the network in the majority of strains/species. We show that the observed diversification in network composition correlates with lineage, lifestyle, and genetic distance. Yeast, filamentous and basal unicellular fungi, form distinctive groups based on these analyses, with substantial differences to their polarization network. Our study shows that the fungal polarization network is highly dynamic, even between closely related species, and that functional conservation appears to be achieved by varying the specific components of the fungal polarization repertoire.

Key words: protein network evolution, cell polarity, protein network, evolution, fungi, adaptation.

Introduction

Fundamental cellular functions, such as respiration, biosynthesis, and homeostasis are crucial to a cell's existence. These complex functions are carried out by the combined action of proteins in protein networks with distinct cellular tasks (Pawson and Nash 2003; Papin et al. 2005). Through evolution of protein networks, by means of, for example, amino acid mutations, network expansion/reduction, and interaction effects, diversity in network composition, levels of protein conservation and divergence, and expression levels is generated (Schüler and Bornberg-Bauer 2010; Voordeckers et al. 2015) that can ultimately lead to the evolution of new functions (see Gladieux et al. 2014 for a list of reviews). Comparative genomics and/or interaction studies of protein networks, such as the citric-acid cycle (Huynen et al. 1999), mitotic checkpoint (Vleugel et al. 2012), and the mitogen-activated protein kinase pathway (Mody et al. 2009) illustrate such patterns. Most of these studies have examined protein network composition through cross kingdom comparisons, covering ~20 divergent species. Although such approaches are insightful for testing if proteins are commonly found in distantly related clades, patterns such as parallel evolution are difficult to disentangle because of the lack of phylogenetic resolution. Examining protein network evolution at phylogenetic dense levels is essential to gain deeper insights into the dynamics of protein networks.

Numerous factors have been presented that promote protein evolution (see Pál et al. 2006; Zhang and Yang 2015).

© The Author(s) 2018. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

To simplify, these factors can be divided into two broad categories: sources of genetic variation, those relating to regional genomic properties, such as variation in mutation or recombination rate; and selection on genetic variation, factors dependent on specific protein properties, such as the proportion and distribution of sites that are involved in a specific function, protein structure, expression level, and competition or adaptation (Pál et al. 2006). These factors often do not act independently, making it hard to identify the relative importance of each factor. In yeast, for instance, the functional importance of a protein influences the rate of protein evolution (Hirsh and Fraser 2001; Drummond et al. 2006; Wall et al. 2005), nonessential genes evolve on an average faster than essential genes (Wall et al. 2005), and loci with more protein–protein interactions evolve on an average slower (Jordan et al. 2003). Various studies have shown that expression rates have the most prominent effect on the rate of protein evolution (Drummond et al. 2006; Wall et al. 2005).

Proteins within the same network often differ substantially in characteristics at the network level. For instance, they can vary in the number and type of interactions, the position within the network (e.g., central vs. peripheral), and the overall number of incorporated proteins in the network. Protein networks can change compositions by losing proteins or including novel proteins through, for example, duplication events followed by neo- or subfunctionalization (Evlampiev and Isambert 2007, 2008). They can compensate for loss of proteins and new functions can evolve (Schüler and Bornberg-Bauer 2010). By inferring the evolutionary dynamics of protein networks, various patterns have emerged. Proteins with many interactions in the network ("hub" proteins) evolve slower than average (Fraser et al. 2002; Kim et al. 2006), especially when they use multiple binding interfaces (Kim et al. 2006). Interacting proteins evolve at similar rates (Fraser et al. 2002). An early comparative genomics study shows that the fundamental citric-acid cycle protein network, involved in energy release, is characterized by variation in protein composition (Huynen *et al.* 1999). Overall, protein networks are characterized by both conservation in topology and function, and substantial divergence in network constitution (Liang et al. 2006; Vleugel et al. 2012).

In this article, we ask how a fundamental protein network evolved with respect to its composition (as in which proteins make up the network), as well as how much the components themself evolve (as in levels of conservation and divergence). We focus on polarity establishment, a process essential for proliferation in basically all unicellular and multicellular organisms. To polarize, cells need to break the symmetrical distribution of cellular content and self-organize in a polarized way. The small GTPase, Cdc42, is a central key protein in this process (Johnson 1999; Etienne-Manneville 2004; Park and Bi 2007). The asymmetrical distribution of so-called polarization proteins, recruited by Cdc42, determines the site of local growth, or budding in the well-described case of the budding yeast *S. cerevisiae*, which is essential for proper cell division and mating.

Cdc42 is a highly conserved protein throughout eukaryotes at both the sequence and functional level (Johnson 1999; Martin 2015) and its activity is regulated through well-documented feedback mechanisms (Irazoqui et al. 2003; Wedlich-Soldner et al. 2003; Goryachev and Pokhilko 2008; Martin 2015). The proteins that directly interact with Cdc42 can be divided into five groups: the GTPase activating proteins (GAPs), that hydrolyze GTP to GDP and change Cdc42 to its inactive state; the guanine nucleotide exchange factors (GEFs), that catalyze the exchange of GDP for a new GTP molecule which activates Cdc42; the GDP dissociation inhibitors (GDIs) that extract Cdc42 from the membrane (Rdi1 is the only GDI in budding yeast; Richman et al. 2004); proteins involved in regulatory mechanisms, such as positive feedback (e.g., the scaffold protein Bem1; Butty et al. 2002); and a wide range of Cdc42 effector proteins which are activated by the active GTP bound state of Cdc42 (fig. 1*A*). Examples of Cdc42 effector proteins are the p21-associated kinases (PAK) Ste20, Cla4, and Skm1 (Johnson 1999), and the GTPase Interactive Components Gic1 and Gic2 (Brown et al. 1997). These proteins colocalize with Cdc42 during polarity establishment and form a protein complex by recruiting other proteins that are needed for actin and microtubule polarization (Brown et al. 1997; Johnson 1999; Drees et al. 2001). Besides functional studies of these proteins, data on potential promoting factors of protein network evolution, for example, the number of genetic and/or physical interactions and expression levels, is also available in budding yeast (see fig. 1*B*). This makes the polarization protein network ideal to test hypotheses on protein network composition, conservation, and divergence. We investigate this among the ecologically and genetically highly diverse clade: the Fungi (Galagan et al. 2005; Mueller and Schmit 2007; Ebersberger et al. 2012).

The eukaryote kingdom of fungi is estimated >760 Myr old (Lucking et al. 2009) and consists of up to 5.1 million estimated extant species (O'Brien et al. 2005). It includes an abundance of species with ecological, agricultural, medical, and scientific relevance. Lifestyles can be restricted to a unicellular lifestyle, either yeast-like or nonyeast as observed in the basal clade of Microsporidia, or multicellular (i.e., pseudohyphal and filamentous species), or can consist of different stages, switching between two or more lifestyles (i.e., di-, trimorphic species). The wealth of different ecologies together with the available genomic and phenotypic resources and tools, such as the Saccharomyces Genome Database (www.yeastgenome.org), make fungi an excellent tool for comparative studies. A vast increase of available fungal genomic data sets, especially fueled by initiatives such as the Fungal Genome Initiative (Rhind et al. 2011), the FungiDB (Stajich et al. 2012), the 1-K fungal genomes project (http://1000.fungalgenomes.org/home/; see also Sharma 2016) and the Joint Genome Institute's Fungal

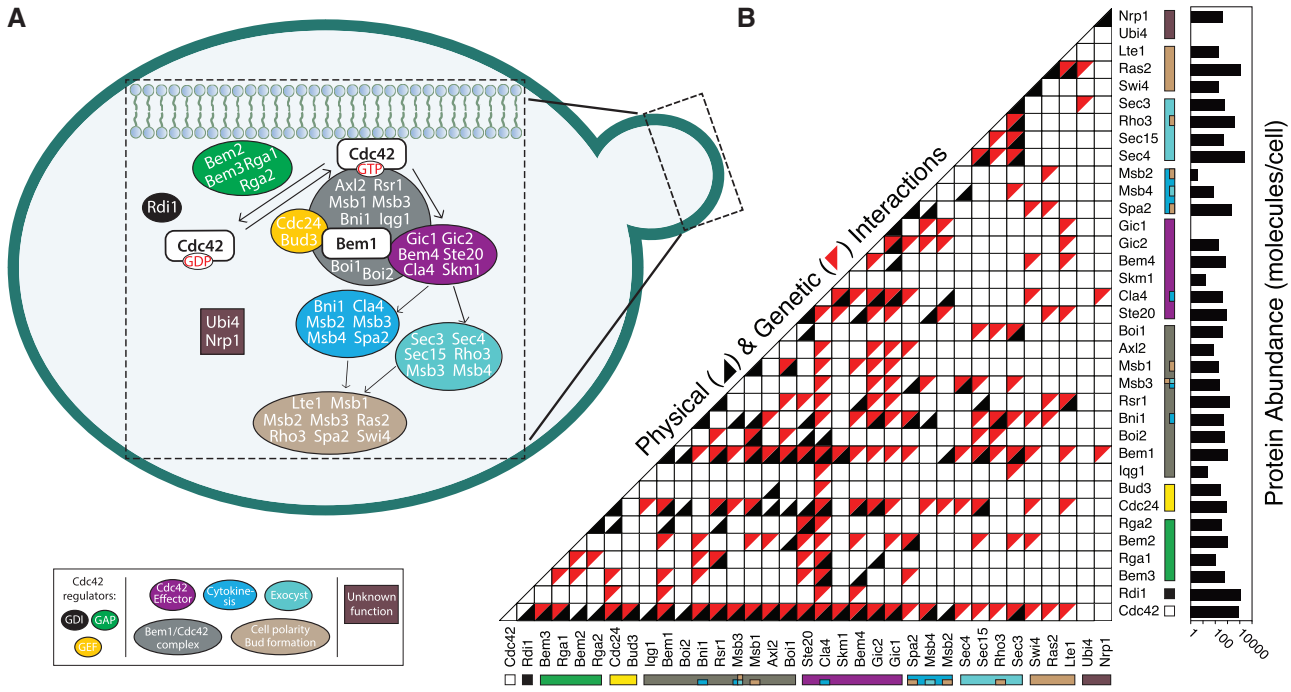

Fig. 1.—The central part of budding yeast's polarization protein network. (*A*) Polarity establishment and subsequent budding takes place at one location of the budding yeast cell membrane (cartoon). Insert: schematic overview of the 35 proteins selected from budding yeast and their functional groupings based on SGD (http://www.yeastgenome.org/) (Drees et al. 2001; Chang and Peter 2003; Madhani 2007; Martin and Arkowitz 2014). Cdc42 cycles between an active membrane-bound state (GTP) and an inactive cytosolic state (GDP). Depicted are the Cdc42 regulators and effectors, the Bem1/Cdc42 protein complex, and several downstream steps (color coded). Nrp1 has a presumed function in polarity establishment (see Laan et al. 2015), Ubi4 has a described epistatic interaction with Cdc42 (BioGRID; thebiogrid.org). (*B*) Matrix of the genetic (in red) and physical (in black) interactions between the 35 selected polarization proteins based on SGD protein data. Proteins are color coded with the functional groupings from the (*A*) panel. Protein abundance following Kulak et al. (2014) is displayed in the most right panel. Note that for Gic1 and Ubi4 no expression data were available.

Genome Project (Grigoriev et al. 2011), took place in the last years prior to this study and provides the desirable phylogenetic resolution.

Although the processes of cell polarity and morphogenesis have been studied extensively in *S. cerevisiae* (Chant 1999; Pruyne and Bretscher 2000; Drees et al. 2001; Chang and Peter 2003; Pruyne et al. 2004; Madhani 2007; Park and Bi 2007; Bi and Park 2012; Martin and Arkowitz 2014), it is unknown to what extent the network's topology is conserved across the fungal phylogeny. This is mainly because only a small number of divergent species has been examined, which are characterized by variation in both network composition and phenotypes (Diepeveen et al. 2017). Due to its fundamental function in cell proliferation, the polarization protein network is hypothesized to be a conserved system (Chang and Peter 2003; Pruyne et al. 2004) and several network members, such as Cdc42, Cdc24, and Sec15, are found to be essential in *S. cerevisiae* (Liu et al. 2015). Previously we showed that, under laboratory settings, the polarization network in *S. cerevisiae* is able to adapt to genetic perturbations to one of the core proteins: Bem1, which regulates Cdc42 (Laan et al. 2015). It is unknown to what extent this represents adaption under natural conditions. Thus, there is some information available on the conservation and the ability of a species to evolve adaptive diversity of a small number of individual polarization proteins. A larger screen to quantify the evolutionary conservation across large phylogenetic distances is currently lacking.

Here, we start to untangle patterns of protein network composition, conservation, and divergence with high phylogenetic resolution within a single, but phenotypically diverse kingdom. We aim to elucidate lineage-specific, independently recurrent and/or conserved patterns of protein network composition, and levels of protein conservation (i.e., both the presence and sequence conservation) and divergence of 42 polarization loci among 298 fungal species. We aim to elucidate factors underlying the observed patterns of this fundamental protein network.

Materials and Methods

Focal Polarization Protein List and Selected Strains and Species

We selected 42 polarization proteins (see table 1 and fig. 1) based on described physical or genetic interaction with the

Table 1
Input Polarization Proteins

Gene	*Saccharomyces cerevisiae*	*Schizosaccharomyces pombe*	*Neurospora crassa*	*Ustilago maydis*
Axl2	X	X	X	X
Bem1	X	X	X	X
Bem2	X			X
Bem3	X	X	X	X
Bem4	X			
Bni1	X		X	
Boi1	X	X		
Boi2	X		X	X
Bud3	X		X	
Cdc24	X	X	X	X
Cdc42	X	X	X	X
Cla4	X	X	X	X
Don1				X
For3		X		
Gic1	X			
Gic2	X			
Iqg1	X	X	X	X
Lte1	X		X	X
Msb1	X		X	
Msb2	X			
Msb3	X	X	X	
Msb4	X			
Nrp1	X	X	X	
Rac1				X
Ras2	X	X	X	X
Rdi1	X	X	X	X
Rga1	X	X	X	X
Rga2	X			
Rho3	X	X	X	X
Rsr1	X		X	
Scd1		X		
Sec15	X	X	X	X
Sec3	X		X	X
Sec4	X			
SepA			X	X
Dia				X
Skm1	X			
Spa2	X	X	X	X
Ste20	X	X	X	X
Swi4	X	X	X	X
Tea1	X	X	X	X
Ubi4	X	X	X	X

small GTPase, Cdc42, a key regulator of polarization (Etienne-Manneville 2004; Park and Bi 2007) and/or described functions of the protein in the polarization network on the Saccharomyces Genome Database (SGD; www.yeastgenome.org; June 2015; Cherry et al. 2012) and in Diepeveen et al. (2017) for orthologs and nonbudding yeast polarization proteins in *Schizosaccharomyces pombe*, *Ustilago maydis*, and *Neurospora crassa*. For each of these proteins we downloaded the amino acid sequences from the SGD for *S. cerevisiae* (August 2017), PomBase for *S. pombe* (August 2017; www.pombase.org/), Ensemble Fungi for *U. maydis* and *N. crassa* (August 2017; fungi.ensembl.org/index.html). We checked orthology by performing reciprocal BLAST searches. We used these sequences as reference or input query in the analyses described below.

Using the Joint Genome Institute (JGI) API, we downloaded the genome sequence and gene models for all published genomes on the JGI website (September 2017; genome.jgi.doe.gov/programs/fungi/index.jsf). We converted all GFF files to GFF3 files using a custom python script. This resulted in 298 strains/species (see supplementary file 1, Supplementary Material online) spanning the fungal kingdom from basal non-Dikarya to Dikarya Basidiomycota and Ascomycota. We provided the genome sequences and gene models to gffread (version 0.9.9; provided by the cufflinks package; Trapnell et al. 2012) to extract each species' proteome to use in the following steps.

Phylogenetic Tree Construction

In order to study the polarization protein network across the 298 strains/species, we constructed a phylogeny. We generated a multiple sequence alignment of 242 shared proteins as input for the phylogenetic analyses. Using the complete *S. cerevisiae* proteome as a starting point, we identified all reciprocal best protein BLAST hits (Altschul et al. 1997) for each *S. cerevisiae* protein in each of the 298 strains/species. We selected only those proteins that had a match in at least 95% of the 298 species. This resulted in a set of 242 proteins (including Cdc42). Each set of homologous proteins was aligned using Clustal Omega (version 1.2.4; Sievers et al. 2014). For species where the sequence was missing, we added gaps. The multiple alignments were concatenated into a single multiple alignment, representing a multiple alignment of 242 proteins (combined length of 183, 160 aa in *S. cerevisiae*). The total length of the full alignment was 613, 783 aa. The multiple alignment was given to FastTree2 (version 2.1.10; Price et al. 2010) with the JTT model of amino acid evolution (Jones et al. 1992) to produce a phylogenetic tree. The tree is rooted with the Microsporidia clade as outgroup. We visualized the tree in the interactive tree of life (iTOL; Letunic and Bork 2011) and edited it in FigTree v1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/). We checked the obtained support values and only reported them in the phylogeny, when they are < 0.9. In addition to this, we also used GBlocks (Castresana 2000; Talavera and Castresana 2007) on our alignment reducing the total length to 108.491 aa, and reran the phylogenetic analysis. This tree showed minor differences in various branches (within: Agaricomycotina, Pezizomycotina, Mucoromycota; between: Neocallimastigomycota, Chytridiomycota, Blastocladiomycota, and Zoopagomycota) and is available upon request.

Genome Quality

To determine if the quality of the genomic resources had an effect on the obtained results, we tested if there was a correlation between the total number of scaffolds, the N50 of scaffolds, L50 of scaffolds, and the total number of proteins we obtained per strain/species. We observed that the quality of the 298 selected strains/species' genomic resources was highly variable (supplementary file 1, Supplementary Material online). We observed great variation in the number of scaffolds, and the N50 and L50 of scaffolds. Genomic resources with short N50 may suffer from missing data such as missing exons and/or gaps (English et al. 2012), which could include, or result in, missing loci. A recent survey of >200 fungal genomes indicates that potentially only 40% of genomes reach the set cut-off for representative completeness (Cisse and Stajich 2016). We found weak but significant correlations between the total number of obtained proteins per species and the number of scaffolds (Spearman rho = −0.13, *P* value 0.03), scaffold L50 (R = −0.13, *P* value 0.03), and a stronger correlation for the scaffold N50 (R = 0.17, *P* value $< 1.0 \times 10^{-2}$; fig. 2). Thus, we find more polarization proteins for species with higher quality genomes (i.e., lower number of scaffolds, lower L50, longer N50). To test how strong this effect of potentially false negatives, or mismatches due to missing data is in our protein matrix, we also performed a multiple factor analysis (see below).

Identifying Orthologs (ggMatch)

Traditional tools to identify orthologous genes between genomes, such as orthoFinder (Emms and Kelly 2015) and orthoMCL (Li et al. 2003) require a vast amount of computational resources to deal with 298 genomes. The computational bottleneck is generally the expensive all-vs-all BLAST queries. Therefore, we implemented a method, Greedy-Gene-Match (ggMatch), which does not attempt to identify orthologs at a genome wide scale, but rather for a reduced set of queries. Typically, searching for orthologs from a single query limits the discovery of orthologs in distant species. Therefore, we approach this problem iteratively, by using results from a previous iteration to extend the query space for a following iteration. Briefly, ggMatch reduces the number of BLAST queries by extending the set of orthologs iteratively. Starting with an initial seed protein p from genome g, ggMatch identifies reciprocal best BLAST hits for protein p between genome g and all other genomes. In the next iteration, the newly discovered proteins are used in a similar fashion to extend the set, searching for further orthologs only in genomes for which no reciprocal best hits were found in previous iterations (see supplementary file 2, Supplementary Material online, for details). This results in a sparse construction of the all-vs-all BLAST matrix. This procedure can be compared with PSI-BLAST (Altschul et al. 1997), except that, rather than filtering based on an e-value threshold, we filter

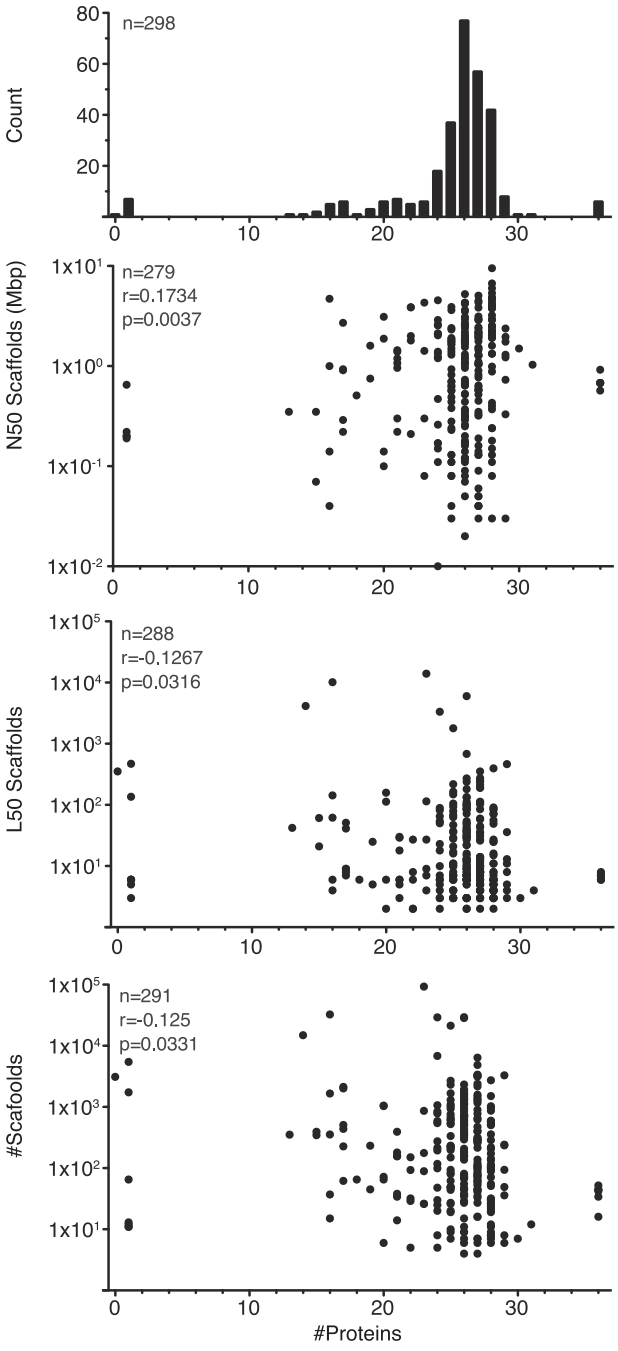

Fig. 2.—Correlation between genome quality and number of retrieved proteins. The top panel (Count) shows the distribution of strains/species for the number of retrieved proteins. The top center panel shows a statistically significant positive correlation between the N50 of scaffolds of the genome and the number of retrieved proteins. The bottom center plot shows a statistically significant negative correlation between L50 (scaffolds) and the number of proteins. The bottom panel shows statistically significant negative correlation between the number of scaffolds in the genome and the number of retrieved proteins.

based on hit reciprocity. Rather than BLAST, we use Diamond (version 0.9.14; Buchfink et al. 2015) to perform protein queries.

Once all iterations have been completed, we can validate our orthologous clusters using functional annotations. Using InterProScan (version 5.27-64.0; Zdobnov and Apweiler 2001), we predict functional annotations for each protein in the cluster. We filter all proteins from the final set S_n which do not have an overlapping annotation with annotations for the proteins in the original set S_0. The result is a set of proteins that a putative orthologs of the protein p_0 in species s_0.

The user can specify multiple queries, and multiple seed proteins in each seed set. In this case, we run one iteration for each query, before starting the next iteration. Each protein may only be assigned to a single query, and will be assigned to the first query it is matched to (hence, this method is greedy).

ggMatch is implemented in python using SnakeMake (Köster and Rahmann 2012). Code for ggMatch can be found on github at: https://github.com/thiesgehrmann/ggMatch. Code to download JGI genomes and convert GFF2 files to GFF3 files can be found on github at: https://github.com/thiesgehrmann/download_jgi.

Polarization Protein Conservation Matrix

As input to ggMatch, we provided the *S. cerevisiae* polarization network proteins, augmented with known orthologs from *S. pombe, N. crassa*, and *U. maydis*. In addition to this, we also added species-specific polarization proteins from these species that were not present in other species. These are summarized in table 1.

For the obtained orthologs we corrected the similarity scores (i.e., the number of positive-scoring matches) of the best hit to the query protein length of that given reference species, thereby obtaining the fraction of similarity per hit. For the species for which we did not obtain orthologs we assigned a similarity score of 0. We generated a matrix of similarity scores by combining all the obtained scores and organizing them according to the species order as observed in our constructed phylogeny for each protein. These steps were performed with in-house python scripts. The generated data matrix was displayed as gray-scale matrix in R version 3.1.2 (R Core Team 2014).

To determine if there was a group of proteins systematically present in a high number of species, we plotted the overall prevalence (%) across the 298 strains/species for each protein. We also plotted the overall prevalence of the two main fungal lineages Basidiomycota and Ascomycota separately and for the five non-Dikarya lineages: Neocallimastigomycota, Chytridiomycota, Blastocladiomycota, Zoopagomycota, and Mucoromycota (i.e., the Microsporidia were not included, due to their highly different pattern) together in order to test for lineage specific proteins. We used a 70% cut-off value as criteria for high prevalence and call the proteins with high prevalence in all 298 species together, and both Basidiomycota and Ascomycota individually, the conserved core of polarization proteins. We based this cut-off value on the observation that there is a gap in prevalence between 60% and 80% for the full data set of 298 strains/species, dividing proteins into two groups. We also plotted the difference in prevalence between the Ascomycota and the Basidiomycota, to depict proteins that are particularly prevalent in either group. To exclude any putative influence of genome quality, we constructed a reduced matrix of the 43 species with highest quality parameters (number of scaffolds < 40, number of contigs < 400, N50 > 1, 5 Mb).

Statistical Analyses

We tested for a potential correlation between our obtained pattern of orthologs (i.e., the total number of observed orthologs per strain/species) and genome quality for which we used three assembly statistics. We obtained the number of scaffolds, the associated N50 (i.e., length of the shortest scaffold in the group of the scaffolds [sorted by length from longest to smallest] that constitutes 50% of the bases in the assembly) and L50 (i.e., the number of scaffolds whose summed length constitutes 50% of the bases in the assembly) for each of the genomes of the 298 strains/species from the JGI genome portal (https://genome.jgi.doe.gov/programs/fungi/index.jsf). Data were tested for normality with D'Agostino and Pearson omnibus normality tests as implemented in GraphPad Prism version 5.0 for Mac OS X (GraphPad Software, La Jolla, CA, www.graphpad.com). Correlations were tested with a spearman's rank correlations as implemented in GraphPad Prism.

To test hypotheses about the cause(s) of differences observed in prevalence between the core and noncore proteins, we tested if these groups of proteins differed in number of genetic interactions, number of physical interactions and overall abundance of the proteins, as indicative for gene expression level. Data for the number of interactions of the 35 *S. cerevisiae* proteins with other members in this network were obtained from SGD. We gathered information about protein abundance (molecules/cells) in *S. cerevisiae* from Kulak *et al.* (2014). Data were tested for normality with D'Agostino and Pearson omnibus normality tests as implemented in GraphPad Prism. We also included the prevalence in the analyses. We performed, a Kruskal–Wallis test (for interactions) and a Mann Whitney tests (for prevalence and protein abundance) with GraphPad Prism.

Multiple Factor Analysis

Because we expected multiple continuous and categorical variables to potentially covary and correlate with the number of observed orthologs per species, we performed Multiple Factor Analysis (MFA) on a data set of the 298 strains/species. We included the following variables: proteins (i.e., the total numbers of proteins per species as observed in the full protein matrix), genome quality (i.e., the number of Scaffolds

and N50 of Scaffolds), Lineage (i.e., the main retrieved phylogenetic clades: Microsporidia, Blastocladiomycota, Chytridiomycota, Neocallimastigomycota, Zoopagomycota, Mucoromycota, Pucciniomycotina, Ustilaginomycotina, Agaricomycotina, Taphrinomycotina, Saccharomycotina, Pezizomycotina), genetic distances (in respect to the references *S. cerevisiae*, *S. pombe*, *N. crassa*, and *Sporisorium reilianum*; as reference for *U. maydis*), and lifestyle (i.e., unicellular, yeast, filamentous, dimorphic yeast-filamentous, dimorphic yeast-pseudohyphal, and trimorphic, i.e., species that have yeast, filamentous and pseudohyphal stages). We calculated the genetic distance between the examined strains/species and the references. We used the concatenated amino acid sequences of the 242 proteins from the phylogenetic analyses (613783 aa; see above) and calculated the genetic distance by using the JTT model of amino acid evolution in MEGA 7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger data sets (Kumar et al. 2016). We obtained the lifestyle information from the Fungal Databases of the CBS-KNAW Fungal Biodiversity Centre (http://www.cbs.knaw.nl/; last accessed November 2016), the JGI genome portal and literature (Bastidas and Heitman 2009; Nagy et al. 2014; Gauthier 2015). We performed the MFA with the FactoMineR R package version 1.33 (Lê et al. 2008) package in R version 3.3.2 (R Core Team 2014) under Rcmdr version 2.3-2 (Fox 2005, 2016; Fox and Bouchet-Valat 2017). We used three quantitative groups: genetic distance (i.e., four variables), genome quality (i.e., two variables) and proteins, and two qualitative groups; lifestyle and lineage. Continuous variables were scaled and standard settings were used. We first checked the eigenvalues for the first ten dimensions to determine the appropriate number of dimensions to consider. In particular we checked for a drop in decline in variance (i.e., broken stick method; Jackson 1993). Length and directions of continuous variables were plotted onto the first two dimensions and were visually checked. Partial axes for the first two dimensions were visually checked. The five groups were plotted onto the first two dimensions. We plotted individuals onto the first two dimensions and color-coded them according to lineage.

Results

The 298 Fungal Strains/Species Group into Major Phyla

In order to examine the protein network of fungal polarity establishment across species, we first estimated the phylogenetic relationship of our 298 focal species in order to display the protein matrix according to phylogeny. We inferred the phylogeny by means of the approximately maximum likelihood method on 242 homologous proteins (see Materials and Methods for details; fig. 3). We retrieve high support values for nearly all branches throughout the tree, which includes two vast monophyletic phyla: the club fungi and relatives (Basidiomycota); and the sac fungi (Ascomycota). Within the Basidiomycota, we found 100% support for the monophyletic subphyla Ustilaginomycotina, Pucciniomycotina, and Agaricomycotina, consistent with previous work (James et al. 2006; Wang et al. 2009). Within the Ascomycota, we found full support for the monophyly of the Taphrinomycotina, Saccharomycotina, and the Pezizomycotina, consistent with previous findings (James et al. 2006; Schoch et al. 2009; Wang et al. 2009). We observe paraphyly for the Chytridiomycota with a low support value for one of the two branches, and polyphyly for the Zoopagomycota. A discussion on relationships of deeper braches and clades is, however, beyond the scope of this work.

The Polarization Protein Network Is Dynamic and Has a Conserved Core

We constructed a protein matrix consisting of the 42 polarizations proteins and the 298 strains/species based on our iterative ggMatch approach (see Materials and Methods for details, also on the effect of genome quality). This approach resulted in a detailed protein matrix indicating the presence, level of divergence in respect to the references, and absence of the 42 polarization proteins in the examined strains/species (fig. 3). We screened the protein matrices for variation in protein prevalence (i.e., the overall number of species a protein is present in), levels of protein divergence and the combinations of proteins (i.e., the composition of polarization network per species). We observed variation in the polarization network at different levels and magnitudes. We observed great variation at the level of amino acid similarity between different proteins, across different strains/species and between different lineages (figs. 3 and 4). For instance, Cdc42 is present with high levels of similarity in nearly all species (except in the Microsporidia species). Cla4, Ste20, and Cdc24 are found throughout the phylogeny at high levels of prevalence as well, but their similarity scores vary greatly across species.

To examine the variability of the composition of the studied polarization protein repertoire across species, we assessed the overall number of different protein combinations and the total number of unique combinations across the reduced matrix and the full matrix of 298 strains/species. Please note 1) that we do not claim to present the complete polarity repertoire for each species, as we only focus on the 42 selected proteins, and 2) if an incomplete repertoire is unique in a species, the complete repertoire of proteins will necessarily also be unique. First of all, we did not observe significant differences in the number of proteins observed between the strains/species belonging to the Saccharomycotina and Pezizomycotina in both matrices (Mann Whitney test; P value > 0.05), indicating that genome quality does not influence these results. We observed substantial fractions for the total number of different protein combinations for the two matrices (i.e., 26/43 and 149/298).

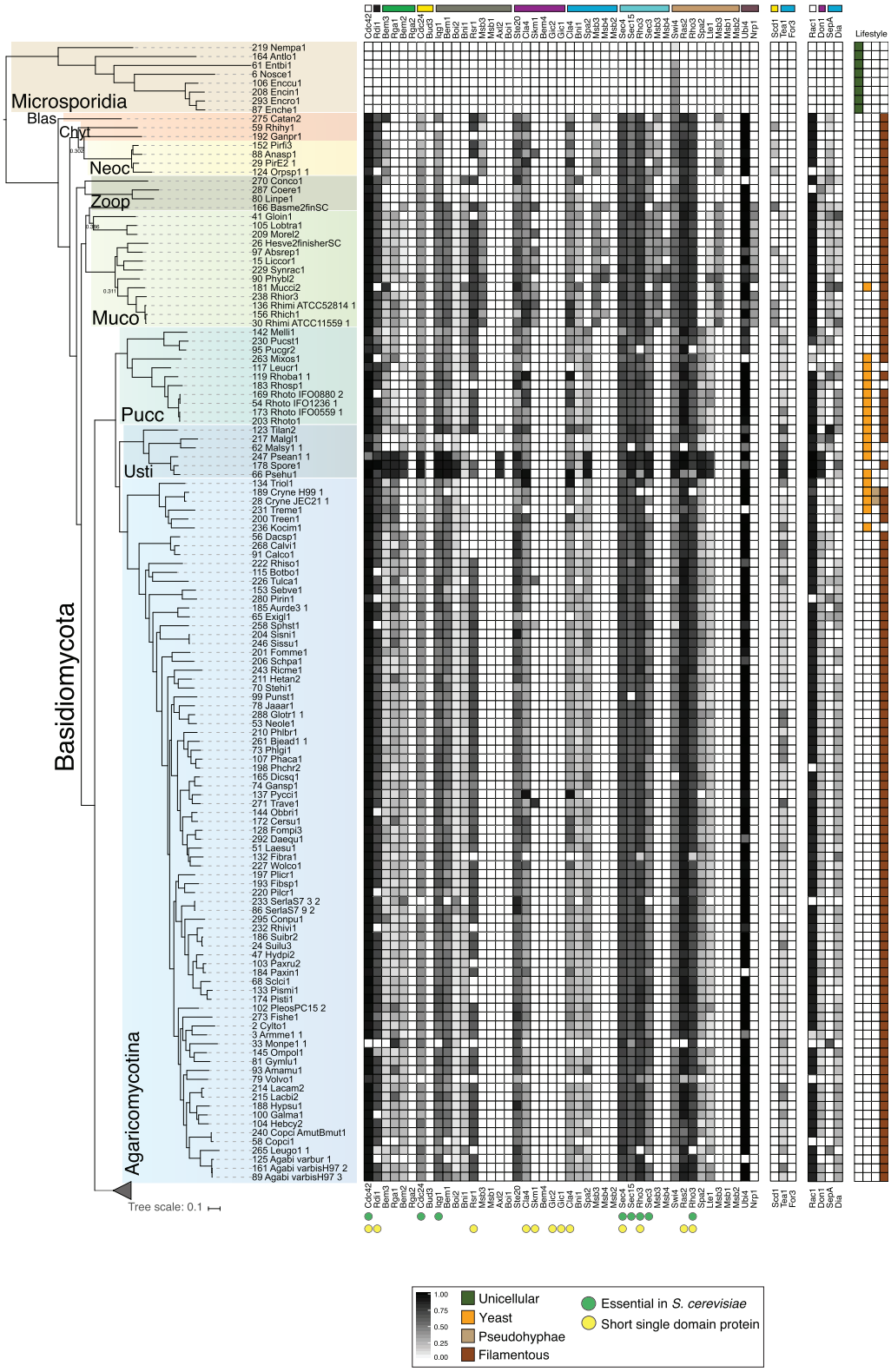

Fig. 3.—Phylogenetic relationships between the 298 fungal strains/species and the protein matrix for the 42 selected polarization proteins. The phylogeny is based on 242 protein sequences (613783 aa) and the approximately maximum likelihood method and the JTT model of amino acid evolution. Support values are almost exclusively >0.9, except when shown on the tree (11 instances). The tree includes the phyla: Microsporidia (in beige), Blastocladiomycota, and Chytridiomycota (in orange), Neocallimastigomycota (in yellow), Zoopagomycota (in dark green), Mucoromycota (in light green), Basidiomycota (in blue), and the Ascomycota (in purple). Subphyla are shades of the same phylum color.

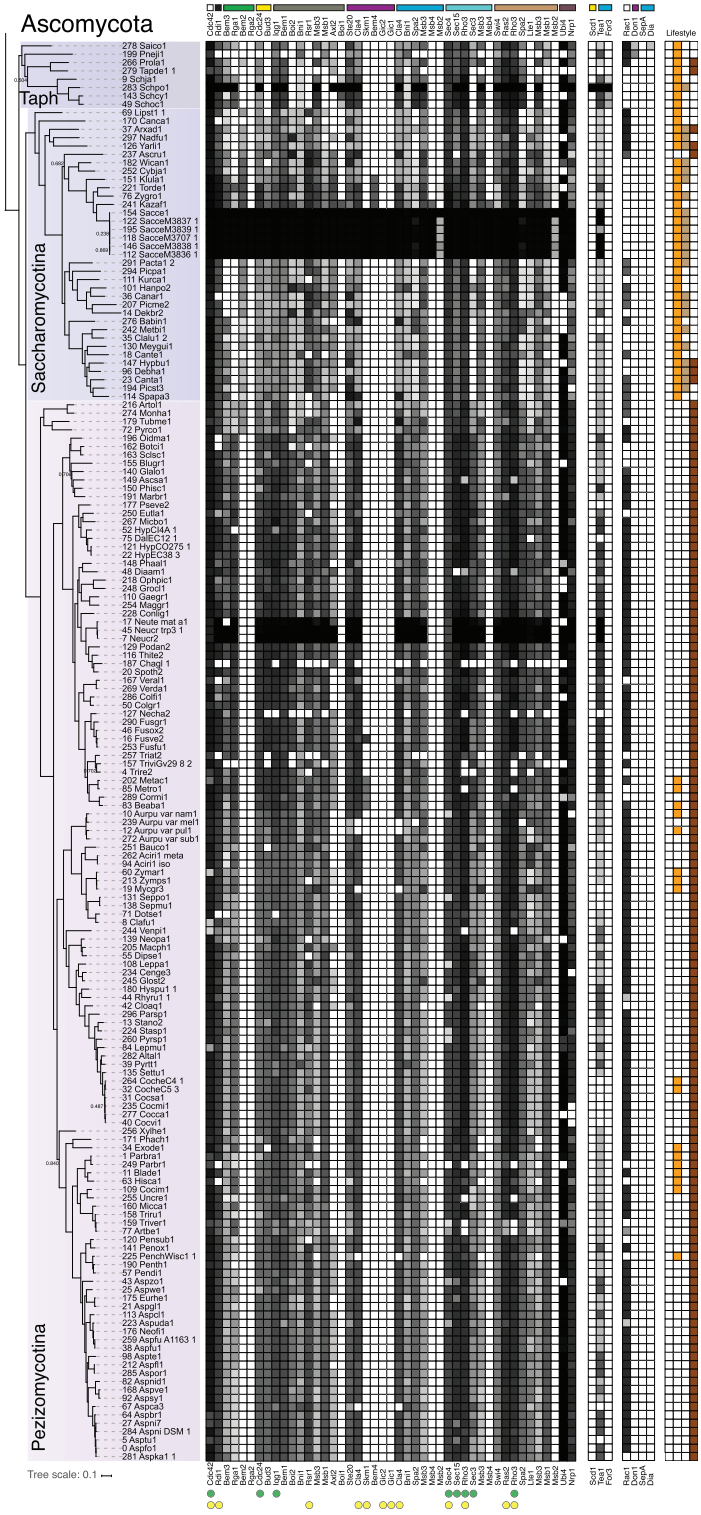

Fig. 3.—(Continued). Phylogenetic relationships follow known relationships at (sub)phylum level (Schüßler et al. 2001; James et al. 2006; Schoch et al. 2009; Wang et al. 2009; Nagahama et al. 2011; Shen et al. 2016), although we retrieved several nonmonophyletic clades (e.g., Chytridiomycota, Zoopagomycota). The protein matrix displays the similarity scores of the iterative ggMatch approach. White fields represent no match of the query proteins in the respective species; black field represent a match with 100% similarity score; gray fields represent a match with <100% similarity score. Proteins are ordered and color coded following figure 1*A*, followed by the three *Schizosaccharomyces pombe* proteins and four *Ustilago maydis* proteins. Essential proteins (in *Saccharomyces cerevisiae*) and short single domain proteins are labeled with green and yellow bullets at the bottom of the matrix. Life styles of the fungal species (yeast-like [orange], nonyeast-like unicellular [green], pseudohyphal [light brown], filamentous [dark brown]) are displayed in the last columns.

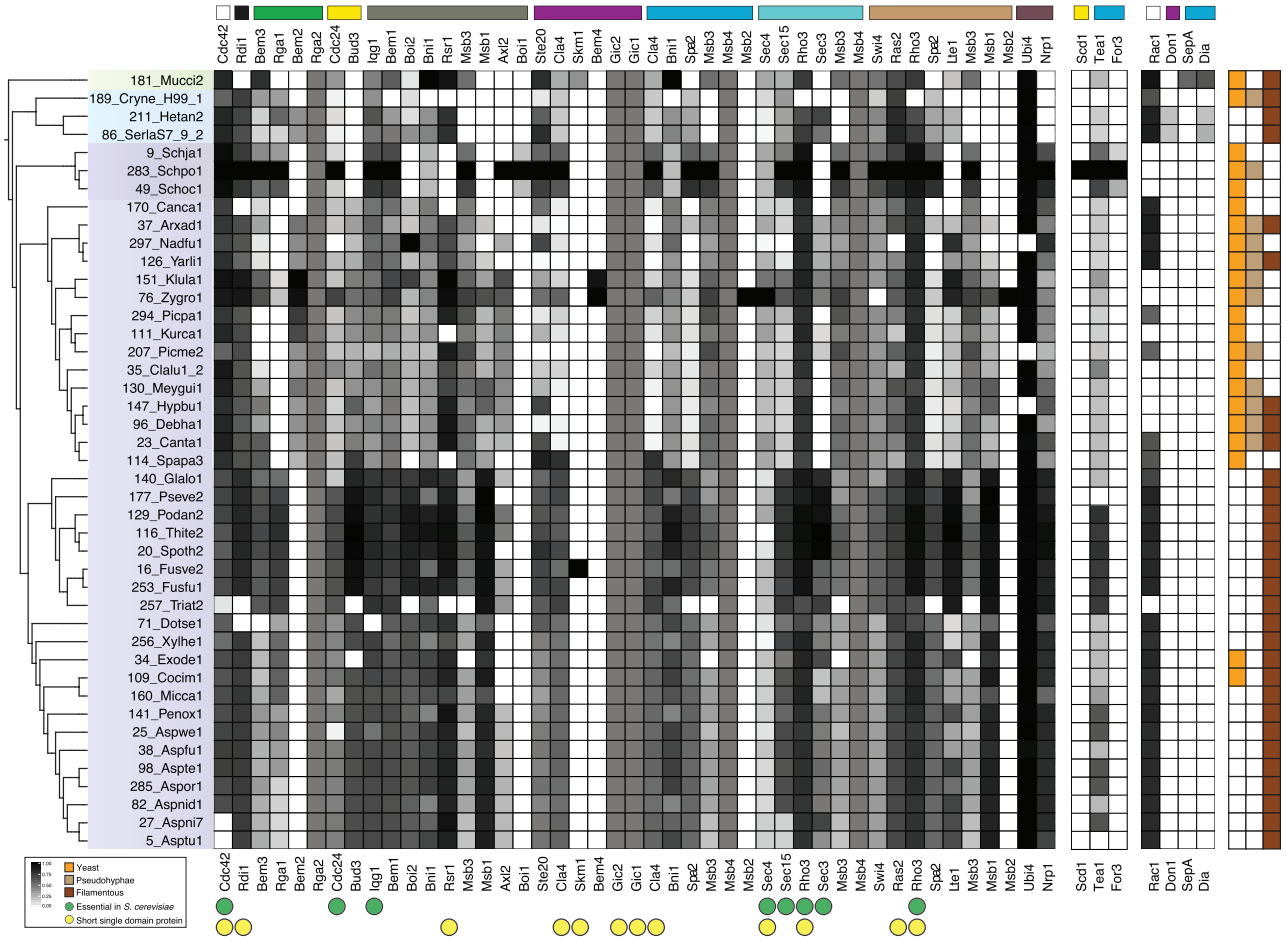

Fig. 4.—Protein matrix of the 43 species with highest genome quality. The matrix displays the similarity scores of the iterative ggMatch approach for species of the Mucoromycota (in green), Basidiomycota (in blue), and the Ascomycota (in purple). Proteins are ordered and color coded following figures 1*A* and 3. Essential proteins (in *Saccharomyces cerevisiae*) and short single domain proteins are labeled with green and yellow bullets at the bottom of the matrix. The life styles of the species are depicted in the far right column.

This indicates that the composition of repertoire is highly variable, with many different observed combinations of proteins. For both matrices the overall fraction of unique combinations (i.e., protein repertoire observed in a single species) was also very similar, 0.49 (reduced matrix) and 0.42 (full matrix). Interestingly, even essential genes in budding yeast, such as Cdc24, Iqg1, and Rho3, are repeatedly, independently lost in the fungal tree (fig. 3). We thus find that the majority of species are characterized by a unique set of polarization proteins not found in other species (see supplementary file 3, Supplementary Material online, for specific combinations). We also observed several specific combinations in multiple species. For instance, we observed the same pattern for seven out of eight Microsporidia species (Swi4). We observed most cases of repeated combinations in the species-rich and well-covered lineages Pezizomycotina and Agaricomycotina lineages (supplementary file 3, Supplementary Material online). These combinations include prevalent, but functionally diverse, proteins such as Rdi1, Bem1/3, Bni1, Boi2, Cla4, Ste20, Sec3/4/15, Iqg1, Lte1, Ras2, Rga1, Rho3, Rsr1, and Cdc24.

To examine the overall prevalence of each protein across the 298 strains/species in more detail, we screened the full matrix and plotted the prevalence (fig. 5). We observed 23 proteins that were present in $\geq$ 70% of all examined species (e.g., Iqg1), four proteins are more commonly found in the Basidiomycota (e.g., Bem2), and five proteins highly present in the Ascomycota (e.g., Msb1/3, Nrp1, and Bud3). We observed a perceived threshold at ~70% prevalence for proteins across all species examined that clearly divides the data set (fig. 5). We found 19 proteins that are present in <69% of the 298 strains/species, while the other 23 proteins are present in at least 79%. We used this 70% mark as cut-off value to determine conserved proteins. This cut-off value includes only proteins that are prevalent at >70% in both Ascomycota and Basidiomycota, individually, thereby excluding, for example, Ascomycota-specific proteins. We called these proteins the conserved core of polarization across fungi, although

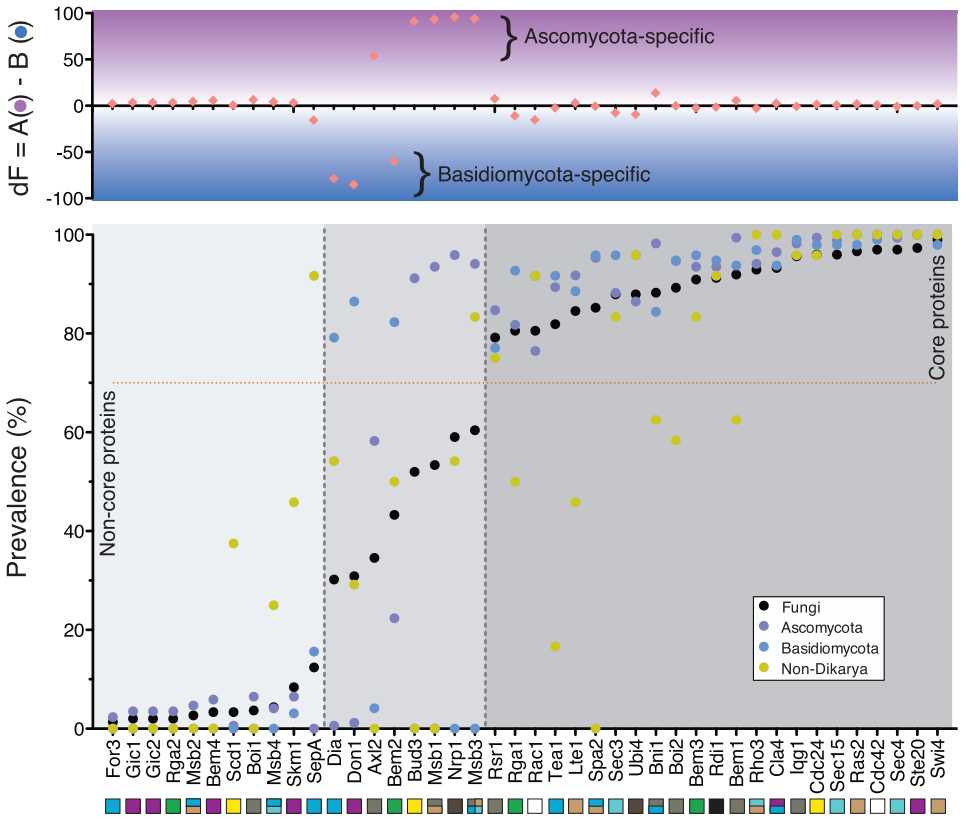

Fig. 5.—Polarization proteins prevalence. Prevalence of the 42 polarization proteins for all examined fungal species (black circles), the non-Dikarya species (excluding the Microsporidia; yellow circles), the Basidiomycota species (blue circles), and the Ascomycota species (purple circles). Proteins are ordered based on their overall prevalence in all examined strains/species. The 70% criterion is marked by a horizontal red dotted line. Shading in the bottom part reflects grouping of proteins with < 20% prevalence in the all Fungi group (light gray; left), proteins with prevalence 20% < 70% (gray; center), proteins with >70% prevalence in all examined groups (i.e., core proteins; dark gray; right). Difference in prevalence between the Ascomycota and Basidiomycota is presented in the top panel (pink diamonds).

none of the individual proteins is 100% conserved. We found this set of 23 core proteins in 126 out of 298 strains/species (supplementary file 3, Supplementary Material online) and >95% of all strains/species had a protein network consisting of 14 or more core proteins (fig. 6), albeit in different compositions (supplementary file 3, Supplementary Material online).

Core Proteins Have Higher Protein Abundance but Not More Interactions

As we observed a group of proteins at high prevalence across clades, we tested if there is a correlation between this conserved core of proteins and factors known to influence protein (network) evolution, such as number of protein–protein interactions and expression levels. We tested whether core proteins are conserved because they are either functionally important and/or because they are present in high quantity. For these factors, data for *S. cerevisiae* are available and therefore we excluded the two non-*S. cerevisiae* specific proteins, Rac1 and Tea1 from these analyses.

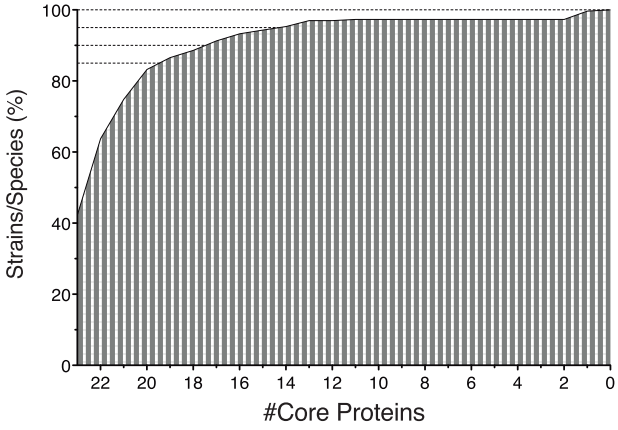

Fig. 6.—The number of core proteins and strains/species. Depicted is the percentage of strains/species and the number of core proteins. Dotted horizontal lines represent the 85%, 90%, 95%, and 100% of strains/species levels.

Core proteins had higher prevalence than noncore proteins (P value < 0.0001; fig. 7*A*). We found no significant difference in the number of either genetic or physical interactions

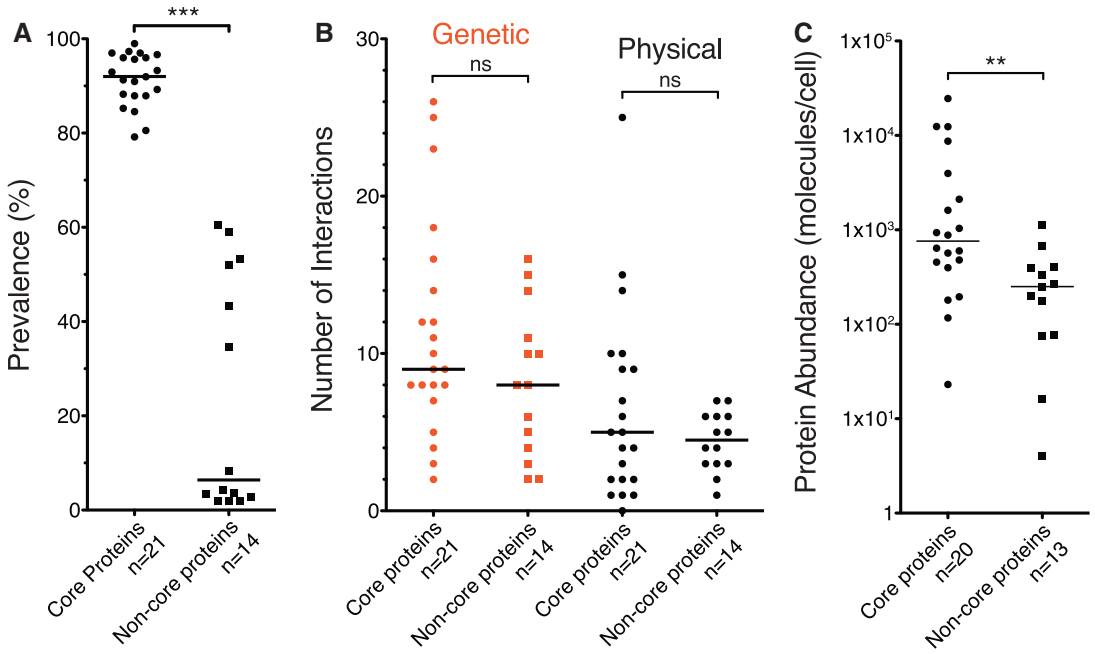

Fig. 7.—Comparison between budding yeast's core and noncore proteins. (*A*) Significant difference in the observed prevalence of the core and noncore proteins (*P* value < 0.0001). (*B*) Number of genetic interactions (in red) and physical interactions (in black) of the 35 examined polarization proteins. No difference was observed between the core and noncore proteins in the number of genetic or physical interactions. (*C*) Significant difference in protein abundance between the two groups. Core proteins have higher protein abundance (*P* value $= 0.005$). Note that data for Gic1 (noncore) and Ubi4 (core) were unavailable. Core proteins are depicted as circles, while noncore proteins are depicted as squares. Black lines depict medians.

(based on observations in *S. cerevisiae*) between the core proteins and the noncore proteins (fig. 7*B*). We did find a significant difference in protein abundance (as measured as molecules per cell in *S. cerevisiae*; Kulak et al. 2014) between the core proteins and noncore proteins (*P* value $= 0.005$; fig. 7*C*). Core proteins might thus be characterized by overall higher protein abundance than noncore proteins, based on *S. cerevisiae* data.

Lineage, Lifestyle, and Genetic Distance Covary with Protein Network Size

In order to test for correlations between factors that could influence protein network evolution and the observed patterns of differences in overall number of polarization proteins in the 298 strains/species, we performed a Multiple Factor Analysis (fig. 8). We considered the following factors: size of the studied protein repertoire (i.e., the total number of proteins we detected with iterative ggMatch approach per strain/species), lifestyle, lineage, genome quality (i.e., the number of scaffolds and the N50) and genetic distance to the four reference species used based on the 242 shared protein sequence alignment.

To determine the adequate number of dimensions to screen, we used the broken stick method (Jackson 1993). We found a drop in variance after the second dimension (supplementary file 4*A*, Supplementary Material online), therefore we only considered the first two dimensions. Dimension 1 is constructed based on four groups: lineage (contribution is 26.47%), lifestyle (24.71%), genetic distance (24.42%), and proteins (22.28%). Dimension 2 is based on lineage (40.25%), lifestyle (37.98%), and genetic distance (20.82%). Together these two dimensions account for 27.96% of the variance in the data. Dimension 1 explained 17.02%, dimension 2 10.93% (fig. 8*D*). We did not find a substantial contribution of genome quality, indicating that the number of scaffolds and/or N50 of scaffolds did not explain the variation in the number of proteins we observe and other examined factors. Supplementary file 4*B*, Supplementary Material online, shows that lifestyle and lineage vary closely together and that they further vary with the number of proteins and genetic distance. Supplementary files 4*C* and 4*D*, Supplementary Material online, indicate that number of proteins only correlates with dimension 1, while lineage, lifestyle, and genetic distance also correlate with dimensions 2.

We plotted the 298 strains/species onto the first two dimensions to visually examine if they cluster to specific patterns based on, for example, morphology or descent (fig. 8*A*). Overall, the individual species seem to cluster to lineages (as color coded [sub]phyla seem to form clouds), the number of proteins they have (the number of the proteins seems to decline horizontally, from left to right), and lifestyle (the top left corner represents the yeast-like fungi, the far right represents the unicellular nonyeast-like fungi, and the lower part

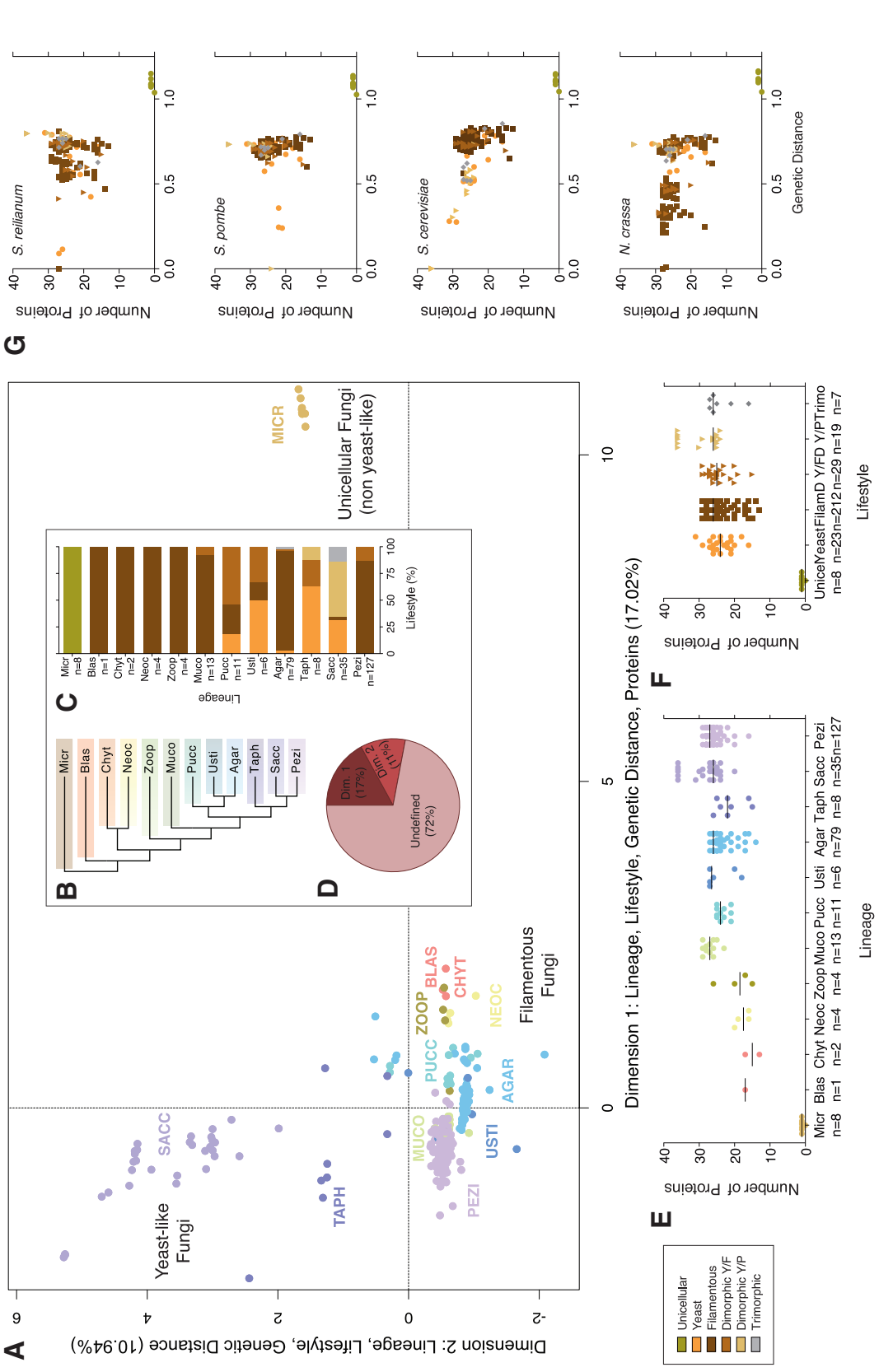

Fig. 8.—Multiple factor analysis and correlations. (*A*) Multiple factor analysis of the number of polarization proteins, lineage, lifestyle, genomic quality, and genetic distances. The 298 strains/species are plotted and color-coded according to their phylogenetic lineage as in figure 3. Dimension 1 explains 17.02% of the observed variation and the following four factors constitute to its construction (in order of importance): lineage, lifestyle, genetic distance, number of observed proteins. Dimension 2 explains 10.94% of the variation in the data and is based on the variables lineage, lifestyle, and genetic distance. Main areas occupied by specific lineages are labeled accordingly for clarity. A clear distinction can be made between yeast-like fungi (left top corner), filamentous fungi (lower part), and unicellular nonyeast like fungi (right). (*B*) Cartoon depicting the topology of the major clades. The length of branches do not represent observed branch lengths. See figure 3 for full phylogeny. (*C*) The distribution of lifestyles (in percentages) for the twelve different phylogenetic lineages. The number of strains/species per lineages is given. Lifestyles are color-coded as in legend at the bottom left of the figure. The 298 strains/species are classified as unicellular, yeast, filamentous, dimorphic (either yeast/filamentous or yeast/pseudohyphal), and trimorphic following figure 3. (*D*) Pie plot depicting the percentage of variation explained by the three main dimensions. The two dimensions account for 28% of the observed variation, leaving 72% undefined. (*E*) The number of observed proteins in twelve different phylogenetic lineages. Groups are color-coded per lineage as in figure 3. Medians are given as black lines. (*F*) The number observed proteins in the different lifestyles. Black lines represent medians. (*G*) The number of observed proteins plotted versus the genetic distance (in respect to *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Ustilago maydis*, *Sporisorium reilianum*). Strains/species are color-coded according to their lifestyle morphology.

represents the filamentous fungi) (fig. 8*A* and *C*). Interestingly, the mostly filamentous Pezizomycotina (Ascomycota) clustered together with the Basidiomycota individuals in the filamentous group. This observation is not in line with the phylogenetic relationships between these clades (see figs. 3 and 8*B*), but is likely caused by the shared lifestyle of the nonyeast Ascomycota and the Basidiomycota species. We observed further interesting patterns regarding lifestyle in the Ustilaginomycotina, Agaricomycotina, and Pucciniomycotina lineages. These lineages consist of species with a variety of lifestyles (fig. 8*C*). For all three lineages we observe that individuals with yeast-like morphologies cluster closer to the yeast-like cloud consisted of the Saccharomycotina and Taphrinomycotina than the other species of their lineage with filamentous morphologies. Within the Taphrinomycotina we observe the species with dimorphic Y/F morphologies closest to the filamentous cloud.

We also plotted lineage, lifestyle, and the four genetic distances in relation to the number of protein observed for each species (fig. 8*E*–*G*). We observe a relative high number of hits for the studied proteins in lineages such as the Mucoromycota, Saccharomycotina, and Pezizomycotina, a smaller number in the Neocallimastigomycota, Pucciniomycotina, and Taphrinomycotina, and a severely reduced number in the unicellular Microsporidia (fig. 8*E*). Filamentous species and dimorphic (Y/P) species tend to have a higher number of studied proteins than yeast species (fig. 8*F*). Lastly, the genetic distance versus protein plots hint to a decreasing number of polarization proteins with greater genetic distance to the references *S. cerevisiae*, and a larger number of proteins with increasing distance to *S. reilianum* (fig. 8*G*).

Discussion

Here, we assessed the composition, conservation, and divergence of the fungal polarization network at high phylogenetic resolution. We observed that the fungal polarization protein network is characterized by both strong protein conservation *and* variation in protein prevalence, sequence similarity, and network composition. Our results indicate that while certain proteins are nearly always needed, potentially for specific functions (i.e., functional conservation), the majority of functional steps seem to be fulfilled by a variable combination of proteins, indicating flexibility in the network composition. Below, we discuss these observations in context of protein network dynamics, functionality of the protein network, and potential causal factors of protein network evolution.

The Fungal Polarization Network Is Highly Variable

It is clear that protein network evolution has a variety of outcomes, such as network expansion/reductions, interaction effects, and protein divergence (Schüler and Bornberg-Bauer 2010; Voordeckers et al. 2015), brought forward by, for example, gen(om)e duplication, selection on protein function or structure and drift (Pál et al. 2006). We found that most proteins of the polarization network have high levels of divergence in amino acid sequence across fungi and that the specific buildup of the protein network per strains/species is highly variable. We find both variation at large phylogenetic distances, such as between subphyla, and between strains/species of the same clade. This indicates that, although the polarization network is involved in fundamental cellular functions across organisms, the network, that we know in *S. cerevisiae*, is not a conserved entity. Work based on the first available fungal genomes reveal remarkable levels of divergence (Galagan et al. 2005), with even <50% similarity in amino acid sequence in comparisons of Ascomycota species (Dean et al. 2005). Screening these genomes for networks reveal that especially regulatory pathways are recurrently characterized by substantial levels of variation, in that elements can be gained or lost over time (Tanay et al. 2005; Tuch et al. 2008; Habib et al. 2012; Muñoz et al. 2016). Our work provides further support for the eminent finding that proteomes and networks constantly change (Coulombe-Huntington and Xia 2017), not only in Ascomycota as previously shown but also in Basidiomycota and non-Dikarya lineages as Mucoromycota, Zoopagomycota, Neocallimastigomycota, and Chytridiomycota (fig. 3).

The substantial levels of variation that we observed in the polarization network could be caused by the remarkable differences in how fungal species polarize and grow (e.g., isotropic, [a]symmetric), among other factors (e.g., the genomic reductions of, e.g., yeasts, Nagy et al. 2014; Microsporidia, Vivarès et al. 2002; Miranda-Saavedra et al. 2007; Peyretaillade et al. 2011). In fact, we do find a clear clustering of yeast-like fungi, nonyeast like unicellular fungi and filamentous fungi in our MFA analysis. While budding yeast polarizes in a switch-like way, filamentous species are characterized by continuous hyphal growth and thus need a constant state of polarity. Differences at the protein levels between species with differences in polarization/growth mode have also been described. The Rho GTPase Rac1 has partly overlapping functions with Cdc42 in regulating polarization in a variety of filamentous species (Banuett et al. 2008), but not in *S. cerevisiae*.

To what extent does this high variability of the protein network affect functionality? As functional studies are not available for the majority of examined species, we made use of the functional classification of proteins of *S. cerevisiae* (see fig. 1). We found that 89% of examined strains/species have at least one protein present from all nine defined functional groups. This could imply that the overall functional pathway of polarity establishment, by means of regulation of a GTPase (Cdc42 and/or Rac1), might be similar across the fungal tree. Further functional exploration of protein networks in

nonmodel species is needed to determine the level of orthology of this network.

Variation in Polarization Network; from Stark Reductions to Lineage-Specific Additions

We found high levels of lineage-specific patterns, of which various patterns coincide with monophyletic clades. For instance, Axl2 is repeatedly lost in different Pezizomycotina clades (fig. 3). The protein matrix also showed very similar patterns for the thirteen Mucoromycota species, for example, with Rga1, Boi2, Msb3, and Nrp1 orthologs present in (nearly) all species, which is dissimilar from the other non-Dikarya clades. These species have a higher number of proteins, compared with the other non-Dikarya and Basidiomycota clades. This is possibly caused by the extensive genome duplications in Mucoromycota fungi (Corrochano et al. 2016).

We found that nearly all examined polarization proteins are absent in the Microsporidia (fig. 3), including most of the conserved core. The only protein that we observed is the Swi4 ortholog. Interestingly, we did not observe this pattern in the other non-Dikarya phyla. We believe that our observation is a true lineage-specific loss in the Microsporidia, as the majority of the polarization proteins (29 out of 42 proteins) are found in nonfungal eukaryotes, such as animals, amoeba and/or plants (see supplementary file 5, Supplementary Material online). The genomes of the parasitic Microsporidia are known to be highly condensed and lack other essential proteins, such as MAP kinases and proteins involved in stress response (Vivarès et al. 2002; Miranda-Saavedra et al. 2007; Peyretaillade et al. 2011). These species have very distinct ecologies, and it is hypothesized that this strong reduction in the proteome is an adaptation to their parasitic life style. It is currently not understood which proteins play a role in polarized cell growth in this genus.

In contrast to the strong reduction in the Microsporidia, we observed lineage-specific gain of polarization proteins in the budding yeast species Saccharomycetaceae. This lineage contains the main reference species *S. cerevisiae*, which automatically results in the full set of 35 *S. cerevisiae* proteins. Furthermore, search tools based on sequence similarity can be affected by the inability of detecting orthologs in nonreference species, for example, in cases of high sequence divergence, resulting in a reference-species bias. We observe that Bem4 is, for instance restricted to the Saccharomycetaceae clade, and Rga2, Gic1, and Gic2 to the Saccharomyces (fig. 3). Various causes can be involved. Genome-wide comparisons across the eukaryote tree have identified an increase in proteins domains in the lineage toward the Ascomycota (Zmasek and Godzik 2011). Furthermore, a whole genome duplication occurred in the *Saccharomyces* lineage after the divergence from the *Kluyveromyces* lineage, and has resulted in many duplicated genes (i.e., paralogs) and instances of accelerated evolution (Wolfe and Shields 1997; Kellis et al. 2004). Our results indicate that different processes have resulted in a myriad of lineage-specific patterns across the fungal tree.

The Conserved Core of Polarization; Functional versus Structural Conservation

We observe a group of 23 core proteins that are recurrently present in the vast majority of examined species (fig. 5). Interestingly, the 23 core proteins cover all functional groups from Cdc42 regulators and effectors to proteins involved in cytokinesis and exocytosis (fig. 1). Even though this group consists of the most prevalent polarization proteins, it does not represent the absolute minimal system needed for polarization. In fact, the majority of species does not have the full set of core proteins (i.e., the complete core is present in 126 out of the 298 strains/species), which can be seen as another indicator of high uniqueness of structural constitution (i.e., the presence of specific proteins across species) of the polarization network across fungi. Different strains/species might achieve functional conservation of the core functions of the network by having different combinations of core proteins. In fact, we observed 18 or more core proteins (i.e., 78%) in 264 strains/species (i.e., 88,6%). These results suggest that functional conservation of the polarization network is high, but that structural conservation, in the sense of network composition, of the individual proteins varies across the fungal strains/species.

Various protein characteristics have been elucidated that are (in part) responsible for protein network conservation, such as position within the network, whether the proteins are essential and the number of interactions (Giaever et al. 2002; Liu et al. 2015). We observed high proportions of essential proteins (7 out of 7) and short single domain proteins (7 out of 10) for the core proteins (fig. 3). Selection is thought to be strong on these classes of proteins, because of their crucial functions and long protein domains (Pál et al. 2006; Buljan and Bateman 2009). These functional characteristics are based on studies in *S. cerevisiae* and could be less relevant in other species. We did not find significant differences in the number of genetic and physical interactions between the conserved core proteins and the noncore proteins. Interestingly, the core proteins Cdc42, Bem1, Cdc24, and Cla4 have the most interactions with the other proteins. These proteins also take central parts in the polarization network, as key regulator (Cdc42; Johnson 1999; Etienne-Manneville 2004; Park and Bi 2007), scaffolding for protein complex (Cdc24 and Bem1; Butty et al. 2002), and signal transducing (Cla4; Johnson 1999). At the same time, we do find a low number of physical and genetic interactions for the core proteins, Rdi1, Rho3. Our results show that not only essential proteins and proteins with many interactions are among the conserved core proteins. We did find a significant effect of protein abundance on the

conservation of core proteins. Although this observation is only based on *S. cerevisiae* data, it does support the hypothesis that conserved proteins are generally more expressed in a cell, as discussed previously (Drummond et al. 2006; Wall et al. 2005).

Link to Causal Factors for Variation in the Polarization Protein Network

Here, we aim to uncover potential causal factors influencing the size of the protein network. Our Multiple Factor Analysis results show that the factors lifestyles, lineage, and genetic distance covary with the size of the studied protein network. These results indicate that the evolutionary background, adaptation to specific lifestyles (i.e., yeast-like, unicellular, and filamentous) and evolutionary time, and thus an indirect measure of genetic drift, of a given species influence their polarization network size. The examined factors do not explain all variation observed in the data, as 72% is undefined, indicative of missing causal factors. The discovery of, and interplay of, causal factors of adaptation and differentiation between species has gained much and long-term attention in the literature (Haldane 1927; Kimura 1967; Orr 2005; Futuyma 2009; Masel 2011). The long-lasting history of population genetics has shown that genetic variation, and thus sequence similarity and ultimately presence/absence of proteins, is caused by the interplay of mutation, natural selection, drift, and gene flow, with descent and thus the heritable characteristics as the starting conditions. It is clear that not all these potential causal factors are incorporated in our study, mainly due to the scale of our study and the unavailability of the particular data for our focal species. Furthermore, the number of proteins that we observed per species is in most cases an under estimation, due to undetected orthologs and lineage/species-specific proteins absent in our original protein list. Even though our analysis does not examine all factors that are likely to have played a role during the protein network evolution, we identified several factors that are, in part, responsible for the complex and highly dynamic polarization protein network evolution of fungi. Further expansion of experimental data sets and development of reliable large-scale comparative tools, should aid a better assessment of empirical data in light of the available theoretical models to study the full scope of real life protein network evolution.

Our study characterizes the fungal polarization protein network as highly dynamic across species, and we identify gene expression level, lineage, lifestyle, and genetic drift, as factors correlating with the observed patterns of conservation, variation, and adaptation. Our results provide further evidence that protein networks are often characterized by shared (ancient) conserved components as well as taxa-specific components that are variable between even closely related species. Our work sheds new light on the level and intensity of protein network evolution across broad and deep phylogenetic levels.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We want to thank the members of the Laan Lab for providing critical feedback during earlier phases of this project. Thanks to W.K.G. Daalman for advice on the statistical analyses. We would like to express our gratitude to the participants of the Gordon Research Conference on Cellular and Molecular Fungal Biology 2016 for constructive feedback on this project. Special thanks to Q.A. Justman, P.J. Boyton, and E.M. Hyland for providing valuable comments on earlier drafts of this article and to S.W.M. Pelders for assisting with writing python code. This work was supported by the Netherlands Organization for Scientific Research (NWO/OCW), as part of the Frontiers of Nanoscience program. The authors declare to have no competing interests.

Literature Cited

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25(17):3389–3402.

Banuett F, Quintanilla RH Jr, Reynaga-Peña CG. 2008. The machinery for cell polarity, cell morphogenesis, and the cytoskeleton in the Basidiomycete fungus *Ustilago maydis*—a survey of the genome sequence. Fungal Genet Biol. 45:S3–S14.

Bastidas RJ, Heitman J. 2009. Trimorphic stepping stones pave the way to fungal virulence. Proc Natl Acad Sci USA. 106(2):351–352.

Bi E, Park HO. 2012. Cell polarization and cytokinesis in budding yeast. Genetics 191(2):347–387.

Brown JL, Jaquenoud M, Gulli MP, Chant J, Peter M. 1997. Novel Cdc42-binding proteins Gic1 and Gic2 control cell polarity in yeast. Genes Dev. 11(22):2972–2982.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nat Methods 12(1):59–60.

Buljan M, Bateman A. 2009. The evolution of protein domain families. Biochem Soc Trans. 37(4):751–755.

Butty A-C. et al. 2002. A positive feedback loop stabilizes the guanine-nucleotide exchange factor Cdc24 at sites of polarization. EMBO J. 21(7):1565–1576.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 17(4):540–552.

Chang F, Peter M. 2003. Yeasts make their mark. Nat Cell Biol. 5(4):294–299.

Chant J. 1999. Cell polarity in yeast. Annu Rev Cell Dev Biol. 15(1):365–391.

Cherry JM, et al. 2012. Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Res. 40(D1):D700–D705.

Cisse OH, Stajich JE. 2016. FGMP: assessing fungal genome completeness and gene content. bioRxiv 049619.

Corrochano LM, et al. 2016. Expansion of signal transduction of pathways in fungi by extensive genome duplication. Curr Biol. 26(12):1577–1584.

Coulombe-Huntington J, Xia Y. 2017. Network centrality analysis in fungi reveals complex regulation of lost and gained genes. PLoS One 12(1):e0169459.

Dean RA, et al. 2005. The genome sequence of the rice blast fungus *Magnaporthe grisea*. Nature 434(7036):980–986.
Diepeveen ET, Iñigo de la Cruz L, Laan L. 2017. Evolutionary dynamics in the fungal polarization network, a mechanistic perspective. Biophys Rev. 9(4):375–387.
Drees BL, et al. 2001. A protein interaction map for cell polarity development. J Cell Biol. 154(3):549–571.
Drummond DA, Raval A, Wilke CO. 2006. A single determinant dominates the rate of yeast protein evolution. Mol Biol Evol. 23(2):327–337.
Ebersberger I, et al. 2012. A consistent phylogenetic backbone for the fungi. Mol Biol Evol. 29(5):1319–1334.
Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16(1):157.
English AC, et al. 2012. Mind the gap: upgrading genomes with pacific biosciences RS long-read sequencing technology. PLoS One 7(11):e47768.
Etienne-Manneville S. 2004. Cdc42 – the centre of polarity. J Cell Sci. 117(8):1291–1300.
Evlampiev K, Isambert H. 2007. Modeling protein network evolution under genome duplication and domain shuffling. BMC Syst Biol. 1(1):49.
Evlampiev K, Isambert H. 2008. Conservation and topology of protein interaction networks under duplication-divergence evolution. Proc Natl Acad Sci USA. 105(29):9863–9868.
Fox J. 2005. The R commander: a basic-statistics graphical user interface to R. J Stat Softw. 14(9):1–42.
Fox J. 2016. Using the R commander: a point-and-click interface for R. Boca Raton, Florida: Chapman and Hall/CRC.
Fox J, Bouchet-Valat M. 2017. Rcmdr: R commander. R package version 2.3-2. https://socialsciences.mcmaster.ca/jfox/Misc/Rcmdr/
Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. 2002. Evolutionary rate in the protein interaction network. Science 296(5568):750–752.
Futuyma DJ. 2009. Evolution. Sunderland (MA): Sinauer Associates. p. 279–301.
Galagan JE, Henn MR, Ma L-J, Cuomo CA, Birren B. 2005. Genomics of the fungal kingdom: insights into eukaryotic biology. Genome Res. 15(12):1620–1631.
Gauthier GM. 2015. Dimorphism in fungal pathogens of mammals, plants, and insects. PLoS Pathog. 11(2):e1004608.
Giaever G, et al. 2002. Functional profiling of the *Saccharomyces cerevisiae* genome. Nature 418(6896):387–391.
Gladieux P, et al. 2014. Fungal evolutionary genomics provides insight into the mechanisms of adaptive divergence in eukaryotes. Mol Ecol. 23(4):753–773.
Goryachev AB, Pokhilko AV. 2008. Dynamics of Cdc42 network embodies a Turing-type mechanism of yeast cell polarity. FEBS Lett. 582(10):1437–1443.
Grigoriev IV, et al. 2011. Fueling the future with fungal genomics. Mycology 2:192–209.
Habib N, Wapinski I, Margalit H, Regev A, Friedman N. 2012. A functional selection model explains evolutionary robustness despite plasticity in regulatory networks. Mol Syst Biol. 8:619.
Haldane JBS. 1927. A mathematical theory of natural and artificial selection, Part V: selection and Mutation. Math Proc Camb Philos Soc. 23(07):838–844.
Hirsh AE, Fraser HB. 2001. Protein dispensability and rate of evolution. Nature 411(6841):1046–1049.
Huynen MA, Dandekar T, Bork P. 1999. Variation and evolution of the citric-acid cycle: a genomic perspective. Trends Microbiol. 7(7):281–291.
Irazoqui JE, Gladfelter AS, Lew DJ. 2003. Scaffold-mediated symmetry breaking by Cdc42p. Nat Cell Biol. 5(12):1062–1070.
Jackson DA. 1993. Stopping rules in principal components analysis: a comparison of heuristic and statistical approaches. Ecology 74(8):2204–2214.
James TY, et al. 2006. Reconstructing the early evolution of fungi using a six-gene phylogeny. Nature 443(7113):818–822.
Johnson DI. 1999. Cdc42: an essential Rho-type GTPase controlling eukaryotic cell polarity. Microbiol Mol Biol Rev. 63(1):54–105.
Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci. 8(3):275–282.
Jordan IK, Wolf YI, Koonin EV. 2003. No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly. BMC Evol Biol. 3:1.
Kellis M, Birren BW, Lander ES. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. Nature 428(6983):617–624.
Kim PM, Lu LJ, Xia Y, Gerstein MB. 2006. Relating three-dimensional structures to protein networks provides evolutionary insights. Science 314(5807):1938–1941.
Kimura M. 1967. On the evolutionary adjustment of spontaneous mutation rates. Genet Res. 9(01):23–34.
Köster J, Rahmann S. 2012. Snakemake-a scalable bioinformatics workflow engine. Bioinformatics 28(19):2520–2522.
Kulak NA, Pichler G, Paron I, Nagaraj N, Mann M. 2014. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. Nat Methods 11(3): 319–324.
Kumar S, Stecher G, Tamura K. 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol. 33(7):1870–1874.
Laan L, Koschwanez JH, Murray AW. 2015. Evolutionary adaptation after crippling cell polarization follows reproducible trajectories. Elife 4: 10.7554/eLife.09638.
Lê S, Josse J, Husson F. 2008. FactoMineR: an R package for multivariate analysis. J Stat Soft 25(1):1–18.
Letunic I, Bork P. 2011. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Res. 39(suppl):W475–W478.
Li L, Stoeckert CJ Jr, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13(9):2178–2189.
Liang Z, Xu M, Teng M, Niu L. 2006. Comparison of protein interaction network reveals species conservation and divergence. BMC Bioinformatics 7(1):457.
Liu G, et al. 2015. Gene essentiality is a quantitative property linked to cellular evolvability. Cell 163(6):1388–1399.
Lucking R, Huhndorf S, Pfister DH, Plata ER, Lumbsch HT. 2009. Fungi evolved right on track. Mycologia 101(6):810–822.
Madhani H. 2007. From a to alpha: yeast as a model for cellular differentiation. Cold Spring Harbor, New York.
Martin SG. 2015. Spontaneous cell polarization: feedback control of Cdc42 GTPase breaks cellular symmetry. Bioessays 37(11): 1193–1201.
Martin SG, Arkowitz RA. 2014. Cell polarization in budding and fission yeasts. FEMS Microbiol Rev. 38(2):228–253.
Masel J. 2011. Genetic drift. Curr Biol. 21(20):R837–R838.
Miranda-Saavedra D, et al. 2007. The complement of protein kinases of the microsporidium *Encephalitozoon cuniculi* in relation to those of *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. BMC Genomics 8(1):309.
Mody A, Weiner J, Ramanathan S. 2009. Modularity of MAP kinases allows deformation of their signalling pathways. Nat Cell Biol. 11(4):484–491.

Mueller GM, Schmit JP. 2007. Fungal biodiversity: what do we know? What can we predict? Biodivers Conserv. 16(1):1–5.
Muñoz A, Santos Muñoz D, Zimin A, Yorke JA. 2016. Evolution of transcriptional networks in yeast: alternative teams of transcriptional factors for different species. BMC Genomics 17(S10):826.
Nagahama T, et al. 2011. Molecular evidence that deep-branching fungi are major fungal components in deep-sea methane cold-seep sediments. Environ Microbiol. 13(8):2359–2370.
Nagy LG, et al. 2014. Latent homology and convergent regulatory evolution underlies the repeated emergence of yeasts. Nat Commun. 5(1):4471.
O'Brien HE, Parrent JL, Jackson JA, Moncalvo JM, Vilgalys R. 2005. Fungal community analysis by large-scale sequencing of environmental samples. Appl Environ Microbiol. 71(9):5544–5550.
Orr HA. 2005. The genetic theory of adaptation: a brief history. Nat Rev Genet. 6(2):119–127.
Pál C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. Nat Rev Genet. 7(5):337–348.
Papin JA, Hunter T, Palsson BO, Subramaniam S. 2005. Reconstruction of cellular signalling networks and analysis of their properties. Nat Rev Mol Cell Biol. 6(2):99–111.
Park HO, Bi E. 2007. Central roles of small GTPases in the development of cell polarity in yeast and beyond. Microbiol Mol Biol Rev. 71(1):48–96.
Pawson T, Nash P. 2003. Assembly of cell regulatory systems through protein interaction domains. Science 300(5618):445–452.
Peyretaillade E, et al. 2011. Extreme reduction and compaction of microsporidian genomes. Res Microbiol. 162(6):598–606.
Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – approximately maximum-likelihood trees for large alignments. PLoS One 5(3):e9490.
Pruyne D, Bretscher A. 2000. Polarization of cell growth in yeast. I. Establishment and maintenance of polarity states. J Cell Sci. 113:365–375.
Pruyne D, Legesse-Miller A, Gao L, Dong Y, Bretscher A. 2004. Mechanisms of polarized growth and organelle segregation in yeast. Annu Rev Cell Dev Biol. 20(1):559–591.
R Core Team. 2014. R: A language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.
Rhind N, et al. 2011. Comparative functional genomics of the fission yeasts. Science 332(6032):930–936.
Richman TJ, et al. 2004. Analysis of cell-cycle specific localization of the Rdi1p RhoGDI and the structural determinants required for Cdc42p membrane localization and clustering at sites of polarized growth. Curr Genet. 45(6):339–349.
Schoch CL, et al. 2009. The Ascomycota tree of life: a phylum-wide phylogeny clarifies the origin and evolution of fundamental reproductive and ecological traits. Syst Biol. 58(2):224–239.
Schüler A, Bornberg-Bauer E. 2010. The evolution of protein interaction networks. Totowa (NJ): Humana Press.
Schüβler A, Schwarzott D, Walker C. 2001. A new fungal phylum, the Glomeromycota: phylogeny and evolution*. Mycol Res. 105(12):1413–1421.
Sharma KK. 2016. Fungal genome sequencing: basic biology to biotechnology. Crit Rev Biotechnol. 36(4):743–759.
Shen XX, et al. 2016. Reconstructing the backbone of the Saccharomycotina yeast phylogeny using genome-scale data. G3 6:3927–3939.
Sievers F, et al. 2014. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 7(1):539.
Stajich JE, et al. 2012. FungiDB: an integrated functional genomics database for fungi. Nucleic Acids Res. 40(D1):D675–D681.
Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol. 56(4):564–577.
Tanay A, Regev A, Shamir R. 2005. Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast. Proc Natl Acad Sci USA. 102(20):7203–7208.
Trapnell C, et al. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 7(3):562–578.
Tuch BB, Galgoczy DJ, Hernday AD, Li H, Johnson AD. 2008. The evolution of combinatorial gene regulation in fungi. PLoS Biol. 6(2): e38–364.
Vivarès CP, Gouy M, Thomarat F, Méténier G. 2002. Functional and evolutionary analysis of a eukaryotic parasitic genome. Curr Opin Microbiol. 5(5):499–505.
Vleugel M, Hoogendoorn E, Snel B, Kops GJPL. 2012. Evolution and function of the mitotic checkpoint. Dev Cell 23(2):239–250.
Voordeckers K, Pougach K, Verstrepen KJ. 2015. How do regulatory networks evolve and expand throughout evolution? Curr Opin Biotechnol. 34:180–188.
Wall DP, et al. 2005. Functional genomic analysis of the rates of protein evolution. Proc Natl Acad Sci USA. 102(15):5483–5488.
Wang H, Xu Z, Gao L, Hao B. 2009. A fungal phylogeny based on 82 complete genomes using the composition vector method. BMC Evol Biol. 9(1):195.
Wedlich-Soldner R, Altschuler S, Wu L, Li R. 2003. Spontaneous cell polarization through actomyosin-based delivery of the Cdc42 GTPase. Science 299(5610):1231–1235.
Wolfe KH, Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387(6634):708–713.
Zdobnov EM, Apweiler R. 2001. InterProScan–an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17(9):847–848.
Zhang J, Yang J-R. 2015. Determinants of the rate of protein sequence evolution. Nat Rev Genet. 16(7):409–420.
Zmasek CM, Godzik A. 2011. Strong functional patterns in the evolution of eukaryotic genomes revealed by the reconstruction of ancestral protein domain repertoires. Genome Biol. 12(1):R4.

Associate editor: Balazs Papp