MBE Advance Access published May 4, 2011

Revisiting the Yeast PPR Proteins – Application of an Iterative Hidden Markov Model Algorithm Reveals New Members of the Rapidly Evolving Family

Research Article

Kamil A. Lipinski[1,2], Olga Puchta[1,5], Vineeth Surandranath[3], Marek Kudla[1,4] and Pawel Golik[1,5]

[1]Institute of Genetics and Biotechnology, Faculty of Biology, University of Warsaw, Pawinskiego 5A, 02-106 Warsaw, Poland;

[2]Present address: Institute of Cell Biology, Swiss Federal Institute of Technology (ETH) Zurich, Schafmattstrasse 18, 8093 Zurich, Switzerland;

[3]Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstrasse 108, 01307 Dresden, Germany;

[4]Present address: Watson School of Biological Sciences, One Bungtown Road, Cold Spring Harbor, 11724 New York, USA;

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

Corresponding author: Pawel Golik, Institute of Genetics and Biotechnology, Faculty of Biology, University of Warsaw, Pawinskiego 5A, 02-106 Warsaw, Poland; tel. +48 (22) 592 3234; fax +48 (22) 658 4176; pgolik@igib.uw.edu.pl

Key words: PPR proteins, pentatricopeptide repeats, mitochondria, yeast, hidden Markov model, RNA-binding

Running head: Yeast PPR family

Abstract

Pentatricopeptide repeat (PPR) proteins form the largest known RNA-binding protein family, and are found in all eukaryotes, being particularly abundant in higher plants. PPR proteins localize mostly to mitochondria and chloroplasts, and many were shown to modulate organellar genome expression on the posttranscriptional level. While the genomes of land plants encode hundreds of PPR proteins, only a few have been identified in Fungi and Metazoa. As the current PPR motif profiles are built mainly on the basis of the predominant plant sequences, they are unlikely to be optimal for detecting fungal and animal members of the family, and many putative PPR proteins in these genomes may remain undetected. In order to verify this hypothesis we designed a HMM-based bioinformatic tool called SCIPHER (Supervised Clustering-based Iterative Phylogenetic Hidden Markov Model algorithm for the Evaluation of tandem Repeat motif families) using sequence data from orthologous clusters from available yeast genomes. This approach allowed us to assign twelve new proteins in *S. cerevisiae* to the PPR family. Similarly, in other yeast species we obtained a five-fold increase in the detection of PPR motifs, compared to the previous tools. All the newly identified *S. cerevisiae* PPR proteins localize in the mitochondrion and are a part of the RNA processing interaction network. Furthermore, the yeast PPR proteins seem to undergo an accelerated divergent evolution. Analysis of single and double amino acid substitutions in the Dmr1 protein of *S. cerevisiae* suggests that cooperative interactions between motifs and pseudoreversion could be the force driving this rapid evolution.

Introduction

The whole-genome sequencing project of the model land plant *Arabidopsis thaliana* revealed the existence of an unexpectedly large protein family with no ascribed functions for most of the hundreds of its members (Small and Peeters 2000; Lurin et al. 2004). The characteristic feature of all proteins grouped into this family is the presence of tandem repeats of a degenerated 35-amino-acid motif, usually containing the entire domain architecture (Small and Peeters 2000). The name pentatricopeptide repeat family (PPR) was given due to an obvious similarity to a previously known tetratricopeptide repeat protein family (TPR), bearing tandem repeats of a similar 34-residue motif. Based on the similarity of TPR and PPR motifs, it was postulated that a single PPR motif is comprised of a pair of anti-parallel α-helices (A and B) forming in tandem arrays a superhelix that encloses a central groove – a putative ligand binding site (Small and Peeters 2000). While in TPR motifs the residues projecting into the groove vary considerably, the side chains lining the central groove in PPR motifs are clearly hydrophilic and form a positively charged bottom. Thus, unlike the TPR proteins – which provide a platform for protein-protein interactions – it was proposed that a multitude of PPR motifs in a single protein could constitute a novel sequence-specific RNA-binding domain (Small and Peeters 2000; Lurin et al. 2004). In several cases direct binding of particular PPR proteins to their RNA targets has been demonstrated *in vivo* or *in vitro* in plants (reviewed in (Schmitz-Linneweber and Small 2008). In one recent study of the maize PPR10 protein, one PPR motif was postulated to recognize a single nucleotide in the RNA substrate (Prikryl et al. 2011). Specific RNA binding was also shown for the *S. cerevisiae* Dmr1p (Puchta et al. 2010) and human PTCD3 (Davies et al. 2009), which interact with the small subunit ribosomal RNA in mitochondria.

Pentratricopeptide repeat proteins are present universally and exclusively in eukaryotic genomes, with an extreme abundance in terrestrial plants. While the genomes of *Arabidopsis*

thaliana and *Oryza sativa* encode about four to five hundred PPR proteins (O'Toole et al. 2008), only 3 and 7 PPR proteins were known to be encoded by yeast and human genomes, respectively (Lurin et al. 2004; Lightowlers and Chrzanowska-Lightowlers 2008; Davies et al. 2009). Early predictions of subcellular localization suggested that the great majority of PPR proteins are targeted to organelles (Lurin et al. 2004). After nearly a decade of extensive genetic and biochemical studies it is now clear that PPR proteins mainly localize to chloroplasts and/or mitochondria, where they attach to diverse RNA species with target sequences located in UTRs, introns or intergenic spacers. At the molecular level, PPR proteins perform a variety of essential functions in many post-transcriptional steps of organellar genome expression in plants, animals and Fungi (Delannoy et al. 2007). These include regulation of processing, maturation and translation of organellar RNAs by direct or indirect involvement in mechanisms like splicing, RNA editing, modulation of transcript stability and translational control (Saha, Prasad, Srinivasan 2007; Schmitz-Linneweber and Small 2008). Therefore, PPR proteins – forming the largest known paralogous protein family in plants – seem to play a pivotal role in the expression of organellar genomes as either housekeeping factors or potential mediators in the nuclear control over the mitochondrial and chloroplast genomes in virtually all Eukaryotes.

Given the abundance of PPR proteins encoded by the plant genomes, a dearth of these proteins in other clades may seem questionable. Also, while a typical plant pentatricopeptide repeat protein consists of a multitude of tandem PPR motif repeats, only two and three motifs can be detected using the current tools in the *S. cerevisiae* Dmr1p and Aep3 sequences, respectively. PPR (and TPR) motifs are identified using profiles based upon known sequences (Karpenahalli, Lupas, Soding 2007). As 88% of all known PPR proteins, containing 94% of all PPR motifs (as identified by the Pfam database (Finn et al. 2010)) are encoded by genomes of *Viridiplantae*, this could introduce a significant bias. If the motif

signature is significantly different between various eukaryotic lineages, the plant model of a PPR motif will fail to detect many non-plant members of this family and the total number of PPR proteins in fungi or animals could be profoundly underestimated.

In this study we readdressed the question of PPR proteins encoded by non-plant eukaryotic genomes. We designed a sensitive and reliable method to search for pentatricopeptide repeat proteins (and other divergent repetitive proteins) using an approach combining iterative HMM profile refinement and expert assesment of results, applied it to the PPR protein family in 14 yeast species for which a complete genome sequence was available, and detected nearly two hundred new putative members.

As the PPR proteins prove to be among the most divergent in orthologous comparisons between different yeast species, we also decided to address one possible cause of this rapid divergent evolution experimentally, using the well-characterized *S. cerevisiae* Dmr1 protein as a model. Our results indicate the possible role of intragenic genetic interactions among different PPR motifs in accelerating the evolution of the pentatricopeptide-repeat proteins.

Materials and Methods

Yeast genomic data

Genome sequences of 14 yeast species were used in this study: *Saccharomyces cerevisiae, Saccharomyces castellii* (Cliften et al. 2003), *Candida glabrata* (Koszul et al. 2003), *Vanderwaltozyma polyspora* (Scannell et al. 2007), *Kluyveromyces lactis* (Zivanovic et al. 2005), *Kluyvromyces waltii* (Kellis, Birren, Lander 2004), *Ashbya gossypii* (Dietrich et al. 2004), *Pichia guilliermondii, Candida albicans* (Jones et al. 2004), *Pichia stipitis* (Jeffries et al. 2007), *Debaryomyces hansenii* (Dujon et al. 2004), *Clavispora lusitaniae, Yarrowia lipolytica* (Dujon et al. 2004) and *Schizosaccharomyces pombe* (Wood et al. 2002). The non-redundant database of protein sequences contained a set of 67943 translated open reading

frames downloaded via SRS (EBI) from UniProtKB/TrEMBL databases. Genome sequences of *S. castellii, K. waltii* and *C. lusitaniae* are not annotated, thus open reading frames of particular homologous sequences were derived using tBLASTn (NCBI). In total, sequences of 41 translated ORFs from these species were gradually appended to the created protein database after each iteration of the SCIPHER search, when new orthologous group were defined.

Building orthologous groups

Homologous sequences for particular proteins were found using BLASTp (NCBI) with default parameters. Clustering of proteins into orthologous groups was manual according to several general rules:

- Assessment of a direct orthology was based on the reciprocal BLAST hypothesis: sequence X from species A is considered as a direct ortholog of sequence Y from species B if a BLASTp search with X as a query hits Y in the first place and BLASTp search with Y as a query hits X in the first place.
- Direct orthology relationships were propagated according to the following rule: if A is a direct ortholog of B and B is a direct ortholog of C then A is a direct ortholog of C (and all sequences belong to the same orthologous group).
- Sequences were considered as in-paralogs if they fulfilled the following criterion: Sequence X and sequence Y from species A are in-paralogous if a direct ortholog of sequence X from species B is the same as a direct ortholog of sequence Y.

Initial definition of yeast PPR motif signature

The PPR model of the first iteration was constructed from an alignment of 194 motifs derived from direct orthologs of Pet309p (102 motifs), Dmr1p (66 motifs) and Aep3p (26 motifs) from the analyzed genomes (34 sequences). Motifs were detected using TPRpred (Karpenahalli, Lupas, Soding 2007). A hidden Markov model was constructed using

hmmbuild program (Eddy 1998) from the EMBOSS package (Rice, Longden, Bleasby 2000). The search was performed with hmmsearch program with default parameters (combined e-value threshold <10).

True positivity criteria (TPC)

When the results of HMM searches for PPR motifs using the SCIPHER procedure were analyzed, it was expected that previously unknown PPR proteins would be detected. A sequence was considered as a new PPR protein if the search results fulfilled three predefined true positivity criteria:

- At least two statistically significant motifs must to be found within the sequence (combined e-value < 10).
- At least two statistically significant motifs must be found in at least one other direct ortholog of this sequence (any other sequence in the orthologous group in the defined phylogenetic cluster).
- Motif hits must not overlap with other previously annotated tandem repeat motifs (e.g. TPR, HAT). However, if statistical significance of the PPR hits is higher than the significance of other tandem repeat motif hits in the same region of the protein (as evaluated in proper control search), this condition may be omitted.

Control searches

Control models of TPR and HAT motifs were constructed using two sets of true positive TPR (2822 sequences) and HAT (957 sequences) hits derived from the Pfam database (Finn et al. 2006). Hidden Markov models were constructed as described previously and used in control searches of the same protein database as SCIPHER searches for PPR proteins and with the same search parameters.

Comparing rates of sequence divergence between proteins

The dataset of TPR proteins contained 30 proteins being best hits in a control search for yeast TPR proteins using the list of Pfam true positive TPR motifs. The dataset of mitoribosomal proteins contained 37 and 12 proteins constituting small and large subunit of the yeast mitochondrial ribosomes, respectively. The dataset of RRM (RBD) sequences consisted of 22 proteins containing at least two RRM motifs per protein found in PROSITE (Sigrist et al. 2010). The dataset of PUF (Pumilio) sequences consisted of 6 proteins containing 5-8 PUF motifs each found in PROSITE. All sequences were derived from the UniprotKB/SwissProt database. Orthologous sequences were obtained by a series of BLASTp searches (NCBI). Sequence similarities between orthologous pairs were calculated as the percentage of identities in pairwise alignments. Data for PPR proteins were obtained by calculating sequence similarities between orthologs of Dmr1p, Pet309p, Aep3p, Rpm2p, Rmd9p, Cbp1p, Msc6p, Pet111p, Yer077cp, Atp22p, Aep2p, Sov1p and Aep1p. Whole-genome interspecies distributions of sequence conservation between pairs of orthologs were obtained from (Dujon et al. 2004).

Yeast strains, media and plasmids

Standard yeast media and basic genetic methods were as described previously (Dujardin et al. 1980; Burke, Dawson, Stearns 2000). Yeasts were transformed using either the rapid or high-efficiency LiAc/SS-DNA/PEG protocol (Gietz and Woods 2002). The *dmr1Δ* strain DPPR1 (*MAT* α*, ade2; trp1; ura3; leu2; his3; dmr1::kanMX4;* [rho$^{-/0}$]) (Puchta et al. 2010) was crossed to the isogenic wild-type strain W303/A/520 (*MAT* a*, ade2; trp1; ura3; leu2; his3;* [rho^{+}]) (Rogowska et al. 2006) to create the DDPPR1 diploid, which was subsequently transformed with the pDMR1-2μ plasmid (Puchta et al. 2010) containing the entire *DMR1* genomic region cloned in YEplac195 (2μ, *URA3*) vector. This strain was then sporulated and

haploid spores carrying the *dmr1::kanMX4* deletion were obtained yielding the HS-DPPR strain (*MAT* α*, ade2; trp1; ura3; leu2; his3; dmr1::kanMX4*; [rho^{+}]; [pDMR1-2μ]).

Error-prone PCR and construction of the mutated library

Random mutagenesis of the *DMR1* gene was performed by an error-prone PCR reaction using Diversify PCR Random Mutagenesis Kit (Clonetech, Cat.# 630703) with buffer condition 3, as described in the user's manual and DMR1_LBam (5'-ATGGATCCTTCCTTGTCGCACATTATCTTACT) and DMR1_RPst (5'-ATCTGCAGTACCTACTATATCGACCACTACGGG) primers described previously (Puchta et al. 2010). PCR products were digested with *Bam*HI and *Pst*I and cloned into the YCplac111 (2μ, *LEU2*) vector (Gietz and Sugino 1988). The obtained plasmid library was amplified and transformed into the HS-DPPR strain. Counterselection to induce the loss of the *URA3* pDMR1-2μ vector carrying the wild-type *DMR1* gene was then performed by passaging transformed clones on media containing uracil (20 μg/ml) and 5-fluoroorotic acid (5-FOA, 1 mg/ml) twice.

***In vivo* labeling of mitochondrial translation products**

The procedure was carried out as described previously (Funes and Herrmann 2007). Briefly, yeast strains were grown overnight at 30 °C in complete synthetic media without methionine, supplemented with 2% galactose. An amount of cells equivalent to 1 OD_{600} unit was taken from the o/n culture and incubated for 30 minutes at 30 °C in complete synthetic media without methionine, supplemented with 2% galactose, cycloheximide (0.375 mg/ml) and [^{35}S]-methionine (50 μCi/ml). Labeling was stopped by chasing with unlabeled methionine and cells were chilled on ice and lyzed. Proteins were precipitated with trichloroacetic acid (TCA) and separated on a 15% SDS-PAGE gel. Gels were stained with Coomassie Blue and dried, and the labeled products of mitochondrial translation were visualized by autoradiography. The total amount of protein in each sample was estimated by

densitometry of the Coomassie Blue staining and the obtained values were used to normalize the results of autoradiography.

Results

Application of current methods used to identify PPR motifs (TPRpred, (Karpenahalli, Lupas, Soding 2007)) to *S. cerevisiae* proteins identifies only three such sequences: Pet309p, Aep3p and Dmr1p, which was the subject of our recent study (Puchta et al. 2010). Only two PPR motifs are detected in the entire 864 amino acid sequence of Dmr1p. Similarly, only 3 motifs are identified in Aep3p. This appears questionable considering the expected architecture of all PPR proteins, where tandem arrays of motifs form a complex superhelical tertiary structure constituting a ligand-binding site. Also, dot-plot analysis of the Dmr1p sequence (not shown) reveals the presence of self-homology indicative of a repetitive character, and multiple alpha-helical regions covering nearly the entire sequence of the protein are predicted by psi-pred (Jones 1999; Buchan et al. 2010). Additional PPR motifs were identified using TPRpred in sequences of Dmr1p orthologs from other fungi, including *Ashbya gossypii* and *Kluyveromyces lactis,* in regions showing significant sequence similarity to the *S. cerevisiae* sequence.

These preliminary results strongly suggested that the Dmr1p sequence contained additional PPR motifs not detected by TPRpred, which uses a profile that is strongly biased towards plant motifs (94% of all known PPR motifs come from the 88% of the PPR proteins that are encoded by plant genomes). We have therefore decided to build profiles that could be used to specifically detect PPR motifs in yeast genomes. To this end we designed a new method, termed SCIPHER (Supervised Clustering-based Iterative Phylogenetic Hidden Markov Model algorithm for the Evaluation of tandem Repeat motif families), using sequence data from orthologous clusters from available yeast genomes

The SCIPHER search

The SCIPHER algorithm combines the iterative application of hidden Markov models (HMM) with a simple phylogenetic analysis aimed at clustering proteins into orthologous groups (see: Materials and Methods). The initial set of PPR proteins used to define the first generation model consisted of all direct orthologs of known *S. cerevisiae* PPR proteins (Pet309p, Dmr1p and Aep3p) from 14 analyzed *Saccharomycotina* and *Taphrinomycotina* species (34 sequences total). The model was constructed on the basis of 194 'canonical' PPR motifs derived from these sequences with the use of TPRpred, currently considered to be the most accurate bioinformatic tool for the tandem repeat motif detection (Karpenahalli, Lupas, Soding 2007).

The analysis flowchart is shown in Figure 1. For each known PPR protein in each of the analyzed organisms a BLASTp (or tBLASTn) search was used to find direct orthologs in other species. As PPR proteins are coorthologous, accurate and rigorous distinction of direct orthology from out-paralogy is required. Currently, several automated tools for clustering of sequences into orthologous groups (OGs) are available (Remm, Storm, Sonnhammer 2001; Alexeyenko et al. 2006; Huerta-Cepas et al. 2008), but their accuracy is dependent on sequence similarity, thus these are prone to errors in distinguishing out-paralogs from direct orthologs in highly divergent multiple gene families. In this study the homology relationships between sequences were elucidated manually (see Materials and Methods). After finding and clustering the orthologous sequences each protein sequence in each orthologous group was subjected to search for canonical PPR motifs using TPRpred (Karpenahalli, Lupas, Soding 2007). All motifs obtained this way were then used to create a hidden Markov model, which was subsequently used to search for PPR motifs in a protein database representing proteomes of all species is the analyzed dataset. New motifs found in already analyzed protein sequences were appended to the previously constructed model, while new protein sequences

in which motifs were detected were used to construct new orthologous groups, as described previously, if they matched predefined true positivity criteria (TPC, see Materials and Methods). About two thirds of sequences identified from the third iteration onwards were rejected as not fulfilling TPC, about 50% of those were TPR or HAT proteins. This shows that the PPR motif in yeasts is very divergent and closely related to TPR and HAT motifs, so that manual curation is needed to reliably tell them apart.

The algorithm was iterated until saturation (no new true positives with subsequent iterations), which was reached at the fifth iteration, resulting in the detection of 953 motifs in 150 protein sequences (Figure 2). Sequences of all the detected PPR motifs are available in Supplementary Data file S1.

According to the data generated, the 14 analyzed yeast genomes encode from 9 to 15 Pentatricopeptide Repeat Proteins each, representing at least 19 different orthologous groups (Supplementary Figure S1). In *Saccharomyces cerevisiae*, in addition to previously known PPR proteins (Pet309p, Dmr1p and Aep3p), in which new motifs were identified, the SCIPHER search unveiled the existence of 11 additional family members. These are Rpm2p, Rmd9p, Rmd9Lp, Cbp1p, Yer077cp, Atp22p, Msc6p, Pet111p, Aep2, Sov1p, and Aep1p (Figure 3).

In the case of *S. cerevisiae* we performed another extrapolation step, identifying motifs which, while failing to give significant hits with HMM profiles directly, share significant sequence similarity with regions identified as PPR motifs in orthologous sequences from other yeast genomes. These could represent very divergent, possibly degenerate PPR motifs. They were identified in 13 of the 14 identified PPR *S. cerevisiae* proteins (light gray boxes in Figure 3), additionally four such divergent PPR motifs were found in the sequence of Rpo41p, bringing the total number of putative *S. cerevisiae* PPR family members to 15. With the exception of Rpo41 (containing the RNA polymerase domain), no motifs other than PPR

were found in the identified proteins, which is in agreement with previous findings in non-plant members of this family.

PPR proteins in yeasts are involved in the expression of the mitochondrial genome

Rpm2p encodes a protein component of mitochondrial RNase P – a widely known enzyme responsible for the removal of a 5'-leader from mitochondrial tRNAs, and thus for the maturation of all mitochondrial polycistronic transcripts that contain tRNAs (Hollingsworth and Martin 1986). Additionally, Rpm2p is also localized in the cytoplasmic mRNA processing bodies and in the nucleus, where it acts as a transcriptional activator (Stribinskis et al. 2005; Stribinskis and Ramos 2007). The number of motifs detected in Rpm2p orthologs ranges from 3 in *C. albicans* to 10 in *D. hansenii*, covering the whole domain architecture of the protein (Figure 3). Therefore, in the RNase P complex, PPR motifs present in Rpm2p may serve as an innate scaffold for binding of the 9S RNA and/or might form a platform for the recognition of target RNA species. The above is consistent with the observation that classical (P subfamily) PPR proteins do not exhibit any catalytical activities but rather perform structural functions, bridging the interactions between proteins and RNAs (Delannoy et al. 2007).

Rmd9p, which is also localized in mitochondria, is an extrinsic membrane protein facing the matrix side of the mitochondrial inner membrane (Nouet et al. 2007). Rmd9p interacts physically with mitoribosomes (Williams et al. 2007) and it was proposed to play a role in the processing and stability of mitochondrial mRNAs, most probably by delivering mtRNAs to the translation machinery (Nouet et al. 2007). Yeast species that had undergone whole-genome duplication (post-WGD species: *S. cerevisiae, S. castellii, V. polyspora* and *K. glabrata*) retained a paralog of Rmd1p (named Rmd9Lp), however its molecular function still remains unknown.

Cbp1p, another newly identified putative yeast PPR protein, is involved in the processing pathway of *COB* mRNA (Dieckmann, Pape, Tzagoloff 1982; Dieckmann, Koerner, Tzagoloff 1984). This protein interacts with the 5'-UTR of *COB* mRNA and has a role in its stability and translation (Islas-Osuna et al. 2002). It was shown that at the inner mitochondrial membrane Cbp1p was associated with Pet127p – a putative 5'-3' exoribonuclease (Krause et al. 2004). In *S. cerevisiae* Cbp1p 5 PPR motifs were found, while the highest number of motifs was detected in the *V. polyspora* ortholog (8 motifs).

The product of *YER077C* ORF (Yer077cp) was suggested to be a member of PPR protein family in *S. cerevisiae* by previous bioinformatic studies (Lurin et al. 2004). This protein still remains uncharacterized, although it is known to localize to mitochondria, and null mutants exhibit a decrease in plasma membrane electron transport. The presence of PPR motifs in orthologs of Yer077cp allows to speculate that this protein may perform a function in the expression of the mitochondrial genome, possibly in the post-transcriptional processing of mitochondrial RNAs.

Atp22p is a mitochondrial inner membrane protein required for the assembly of the F0 sector of ATP synthase. It has been reported that levels of *ATP9* and *ATP6/8* transcripts in atp22 mutants remain unchanged, while the functionality of ATPase is defective (Helfenbein et al. 2003). Atp22p was found to act as a translational activator for the ATPase subunit-6 mRNA, with genetic data suggesting its interaction with the 5' UTR of the transcript (Zeng, Hourset, Tzagoloff 2007). The presence of PPR motifs in Atp22p also suggests that it should exhibit an RNA-binding activity. The number of motifs detected by SCIPHER varies from 0 in *P. stipitis* and *S. castellii* to 6 in *V. polyspora*.

Msc6p was previously identified in a whole-genome screen for mutants displaying unequal sister chromatid exchange during meiosis (Thompson and Stahl 1999). The exact molecular function of Msc6p remains unknown, although it is also localized in mitochondria

and null mutants exhibit a decreased respiratory growth rate. Direct orthologs of this protein are present only in yeast species from *S. cerevisiae* and *K. lactis* clades (Supplementary Figure S1). The maximum number of motifs was found in an ortholog from *V. polyspora*, where 6 motifs were detected.

Pet111p is a well-known regulatory protein localized in the inner mitochondrial membrane where it plays the role of a translational activator specific for *COX2* mRNA (Poutre and Fox 1987; Mulero and Fox 1993; Bonnefoy, Bsat, Fox 2001). At the same time Pet111p negatively regulates translation of *COX3* mRNA, in opposition to Pet122p, Pet494p and Pet54p (Naithani et al. 2003). 6 PPR motifs were found in the sequence of Pet111p from *S. cerevisiae* during the SCIPHER analysis.

Aep2p (Atp13p) was suggested to be involved in translation and stability of *ATP9* mRNA (Finnegan et al. 1991). This protein probably binds to the 5'-UTR of this transcript, as the phenotype of isolated thermosensitive mutants may be suppressed by a base substitution in this region of the target mRNA (Ellis et al. 1999). Species-specific interactions between Aep2p and *ATP9* mRNA were found to cause nucleo-mitochondrial incompatibility between *S. cerevisiae* and the closely related species *S. bayanus* (Lee et al. 2008). Only a few PPR motifs were detected in orthologs of Aep2p in the analyzed yeast species. This includes 4 motifs found in the *S. cerevisiae* protein as well as 3 motifs in *V. polyspora*, *K. waltii, K. lactis* and *C. albicans* and 2 in *P. stipitis*. YQ24 (SPCC11E10.04) from *S. pombe* also seems to be a direct ortholog of Aep2p, and 4 PPR motifs were detected in its sequence.

Sov1p is a protein of unknown function that also localizes to mitochondria and the phenotypes of its mutants indicate an involvement in the expression or maintenance of the mitochondrial genome (Merz and Westermann 2009). Direct orthologs of Sov1p are found in *S. cerevisiae* and *K. lactis* clades as well as in the *Y. lipolytica* genome. The number of motifs detected varies from 2 in *Y. lipolytica* to 7 in *A. gossypii*.

Aep3p, a previously known PPR protein, is thought to be involved in the expression of *ATP6/8* mRNA. Aep1p (Nca1p) – similarly to Aep2p – is implicated in the expression of the mitochondrial transcript encoding subunit 9 (Atp9p) of F1-F0 ATP synthase (Payne et al. 1993; Ziaja, Michaelis, Lisowsky 1993). The phenotype of the previously isolated thermosensitive mutant indicate that Aep1p presumably plays a role is the translation of the *ATP9* mRNA, as the levels of the cognate transcript are unchanged upon a temperature shift (Payne, Schweizer, Lukins 1991). Orthologs of Aep1p are present exclusively in *S. cerevisiae* and *K. lactis* clades and contain up to 4 PPR motifs.

Finally, the SCIPHER search identified PPR motifs located in N-termini of several orthologs of Rpo41p – the catalytic subunit of the mitochondrial RNA polymerase (Greenleaf, Kelly, Lehman 1986). It is widely accepted that eukaryotic mtRNA polymerases are related to T-odd bacteriophage RNA polymerases, although many contain an additional N-terminal extension (ATD) crucial for the coupling of transcription and translation in the mitochondrial expression system (for review see: (Shadel 2004). This auxiliary domain is speculated to interact with the newly synthesized RNA species and a post-transcriptional factor Nam1p, therefore spatially and functionally bridging the RNA polymerase complex with RNA processing machinery and mitoribosomes at the inner mitochondrial membrane (Rodeheffer et al. 2001). The detection of PPR motifs in yeast ATDs is in line with previous observations of such motifs in the human mitochondrial RNA polymerase POLRMT (Shadel 2004).

In addition, several other PPR proteins were found to be encoded by genomes of species from the *C. albicans* clade. These proteins fall into three separate orthologous groups. Proteins from one group are related to *Zea mays* PPR protein CRP1 (Fisk, Walker, Barkan 1999), while others exhibit similarity to *KIN11* from *Arabidopsis thaliana* – an SNF1-related protein kinase involved in signal transduction (Bhalerao et al. 1999). The third group has not

been named as proteins constituting this cluster do not resemble any protein characterized up to date. Also, an ortholog of human PPR protein *PTCD1* with 4 PPR motifs was found in the genome of *Y. lipolytica*. Human *PTCD1* is a mitochondrial matrix protein which associates with leucine tRNAs and precursor RNAs that contain leucine tRNAs (Rackham et al. 2009), while its homolog in *Neurospora crassa* was shown to be involved in the assembly of respiratory complex I (NADH:ubiquinone oxidoreductase) (Lightowlers and Chrzanowska-Lightowlers 2008). As *Y. lipolytica* retained mitochondrially encoded components of respiratory complex I (Matsuoka et al. 1994), this PPR protein might also play a role in their expression, although no data to support this hypothesis are currently available.

The remaining detected PPR proteins were not assigned into any orthologous group, as many appear unique or do not exhibit significant similarity with other known proteins.

This brief analysis indicates that both previously known and newly identified PPR proteins of *S. cerevisiae* localize to the mitochondrion, in agreement with whole-proteome localization studies (Huh et al. 2003) and single-gene analyses, and most of them are implicated in the expression of mitochondrial genome, which is expected for this family.

Yeast PPR proteins undergo rapid evolution

As the initial assessment of alignments suggested that PPR proteins show significant divergence even in closely related species, we compared pairwise sequence identities between direct orthologs in selected pairs of species to the values observed for other protein families. A metagenomic whole-genome dataset describing conservation of orthologous gene pairs for comparisons between *S. cerevisiae* and *C. glabrata*, *S. cerevisiae* and *K. lactis*, and *S. cerevisiae* and *D. hansenii* is available (Dujon et al. 2004). We plotted the mean identity values for PPR proteins and control data sets obtained for TPR proteins (sharing a similar repetitive structure), mitochondrial ribosomal proteins (sharing an involvement in mitochondrial gene expression) and RRM and Pumilio (PUF) proteins (containing RNA

binding motifs) in relation to the distribution of relative frequencies of pairs of orthologs exhibiting a particular level of sequence identity (derived from (Dujon et al. 2004)) (Figure 4).

The TPR and ribosomal proteins, as well as the control modular RNA binding proteins (RRM and PUF) exhibit divergence close to the typical (median) value obtained for the entire dataset of 1100 pairwise alignments (47.8-62.0% identity for *S. cerevisiae – C. glabrata* and *S. cerevisiae – K. lactis* pairs; 39.9-42.4% identity for *S. cerevisiae – D. hansenii* pairs).

PPR proteins, on the other hand, were clearly subjected to accelerated divergent evolution, as the average sequence identity between PPR orthologs is about two-fold less than between orthologs from control data sets (31.7% and 28.7% for *S. cerevisiae – C. glabrata* and *S. cerevisiae – K. lactis*, respectively, and 23.7% for *S. cerevisiae – D. hansenii*). In the plot showing the distribution of percentage identity between orthologous pairs (Fig. 4) they are clearly among the most divergent and fastest evolving proteins. As TPR and ribosomal proteins evolve with much lower rates, close to the average, the observed accelerated divergent evolution of the PPR protein family cannot be explained neither by their structural nor functional properties alone.

Pseudoreversion can contribute to the accelerated evolution of PPR proteins

Several factors can contribute to the accelerated rate of sequence divergence observed for PPR proteins. As the organization and sequence of the mitochondrial genomes is very divergent among different yeast lineages (Tian et al. 1991; Groth, Petersen, Piskur 2000; Nosek et al. 2006), variation in the mitochondrial RNA transcripts (particularly in untranslated regions) could drive fast changes in their binding partners, including the PPR proteins.

Another interesting possibility is related to the presumed mechanism of substrate binding by the PPR proteins. Repetitive character of these sequences and their substrates suggests a

cooperative mode of binding, demonstrated experimentally for the plant chloroplast protein HCF152 (Nakamura et al. 2003). One could speculate that proteins with such binding mechanism could tolerate hypomorphic mutations in a single motif better, as the remaining motifs would still provide sufficient activity. Also, if different repeats within a PPR protein sequence contribute to its RNA-binding activity in a cooperative manner, then mutations in one motif could drive compensatory mutations in the other motifs if the activity is to be retained.

In order to test this possibility we performed a random mutagenesis study on the *S. cerevisiae* Dmr1p, which is an RNA-binding protein from the pentatricopeptide repeat family, required for the stability of the mitochondrial 15S ribosomal RNA (Puchta et al. 2010). As a deletion of *DMR1* results in destabilization of the mitochondrial genome and thus irreversible respiratory deficiency (Puchta et al. 2010), the wild-type allele of *DMR1* must be present in the host strain before mutant alleles are introduced. This can be achieved by applying the plasmid shuffling technique. The wild-type *DMR1* gene cloned in a *URA3* vector (pDMR1-2μ) was introduced into a diploid strain heterozygous for the *dmr1Δ* deletion and carrying the functional rho^+ mitochondrial genome (DDPPR1 strain, all genotypes are given in Materials and Methods). After sporulation a haploid strain (HS-DPPR) was obtained, which carried the *dmr1Δ* deletion on the chromosome and the wild-type *DMR1* gene on the *URA3* plasmid vector. Mitochondrial respiratory function in this strain is thus entirely dependent on the plasmid-borne copy of the gene.

The host strain HS-DPPR was transformed with a library of random mutant alleles of the *DMR1* gene obtained by error-prone PCR, cloned in a *LEU2* vector (see Materials and Methods). The wild-type *DMR1* gene was subsequently removed by counterselection on 5-FOA, leaving only the mutagenized library *DMR1* alleles. About 500 clones were screened for hypomorphic (partial loss-of-function) mutations in *DMR1* by testing for respiratory

growth on YP-glycerol plates at permissive (30 °C) and restrictive (36.5 °C) temperatures. Three independent clones exhibiting slower growth at normal temperature and no growth at the restrictive temperature on respiratory medium were identified and analyzed by sequencing.

One of the hypomorphic mutants, named *dmr1-1,* was found to carry a single nonconservative amino acid substitution (D785V) in one of the PPR motifs identified by SCIPHER analysis. This residue appears to be highly conserved, all the members of *Saccharomycotina* that retain homology in this region have aspartate in this position, and a conservative substitution to glutamate is found in *S. pombe* (*Taphrinomycotina*) (Figure 5A).

The remaining two clones shared identical amino acid sequence of the *DMR1* gene and were further treated as one mutant, named *dmr1-2*. It was found to carry the same D785V substitution as *dmr1-1* plus an additional substitution T351A. This latter residue is not conserved itself, but is located within a conserved PPR motif (Fig. 5A).

The effect of the single D785V substitution (*dmr1-1*) and double T351A, D785V substitution (*dmr1-2*) on the function of Dmr1p was assessed by analyzing growth on respiratory media at permissive and restrictive temperature (Figure 5B) and the rate of mitochondrial translation (Figure 5C), measured by ^{35}S labeling *in vivo* in the presence of cycloheximide (Materials and Methods). The host strain HS-DPPR carrying the wild-type *DMR1* gene was used as a positive control, while the *dmr1Δ* DPPR1 strain was used as a negative control.

Both mutant strains show no respiratory growth at the restrictive temperature. At the permissive temperature the respiratory growth of the single *dmr1-1* mutant is visibly impaired, while that of the double mutant *dmr1-2* is closer to the positive control (Fig. 5B). The second T351A mutation has, therefore, partially alleviated the strong hypomorphic phenotype of the D785V substitution. Similar conclusions can be drawn from the analysis of

mitochondrial translation, carried out at the permissive temperature (Fig. 5C) using galactose as a carbon source. While the single D785V substitution in the *dmr1-1* mutant has a visible strong effect on mitochondrial protein synthesis (about two-fold reduction), the presence of the second T351A mutation in *dmr1-2* restores nearly wild-type translation product levels. The effect of *dmr1-1* and *dmr1-2* alleles on the mitochondrial translation, was, however, much less evident when either glycerol or glucose were used as carbon sources (not shown).

These results indicate that the phenotype of a mutation in one PPR motif can be partially alleviated by second-site mutations affecting another motif (pseudoreversion). Such intragenic genetic interactions can be expected if different motifs contribute to the same function (RNA-binding) in a cooperative fashion. These genetic interactions could also drive rapid divergence of protein sequences in evolution – when one of the motifs is changed by mutation, mutations in other motifs may be necessary to retain optimal function. This mechanism could therefore explain, at least in part, the observed divergence of PPR protein orthologs in yeasts revealed by our analyses.

Discussion

Previous bioinformatic whole-genome searches for members of PPR protein family uncovered only three *bona fide* PPR proteins in yeast *S. cerevisiae* (Small and Peeters 2000; Lurin et al. 2004; Puchta et al. 2010). However, as the motif signatures used were based mainly on plant sequences, they turned out to be not optimal for detection of non-plant PPR proteins. The presence of additional non-canonical motifs in some known non-plant proteins had already been reported e.g. in case of the PPR protein family from *Trypanosoma brucei*. The authors suggested that in virtually all cases sequence alignments as well as secondary structure predictions clearly implied that structurally similar regions adjoin known PPR motifs within particular protein sequences (Pusnik et al. 2007).

Iterative HMM models were proposed previously for structure prediction (Jones 1999; Karplus 2009) and, recently, homology searches (Johnson, Eddy, Portugaly 2010). Direct application of these tools to the PPR proteins is, however, prone to a significant number of false-positive hits, due to the repetitive nature of these proteins and general similarity to other motifs, such as TPR and HAT. Our approach was unique in providing extensive filtering to ensure that the identified proteins are indeed *bona fide* members of the PPR family.

The SCIPHER algorithm, developed in this work, allowed us to find many more PPR proteins in *S. cerevisiae*, as well as in the other 13 yeast species analyzed in this study. The phylogenetic clustering-based approach also provided extensive data about the organization and evolution of the family in *Saccharomycotina*. The metagenomic approach used in the SCIPHER algorithm takes into account natural evolutionary divergence of the motif signature, thus making the *in silico* search more reliable. Secondly, it simultaneously characterizes a set of proteins in multiple species, thus enabling comparative phylogenetic analysis, which is required for the complete bioinformatic characterization of the analyzed protein family. Our results clearly demonstrate that the predominantly plant-derived motifs are suboptimal in other eukaryotic lineages, which illustrates a significant divergence of the PPR motif and presents a significant analytical challenge. Manual curation according to the arbitrarily assumed (but based on the current knowledge about the PPR family structure and function) True Positivity Criteria remains a critical step in the analysis. The established algorithm is suitable for developing a model aimed at detection of members of any protein family bearing repeats of a motif of choice, e.g. HAT, TPR, HEAT or Armadillo repeats. The notable limitation is, however, that the evolutionary distance between the analyzed species in the chosen clade must be long enough to assure significant sequence ambiguity and short enough to allow a definite assignment of homology interrelationships between protein sequences. For example, a similar SCIPHER search for PPR proteins in vertebrates failed to

yield a conclusive outcome due to insufficient sequence dissimilarity between pairs of orthologs, which resulted in excessively rigorous HMM models.

In this work we found that the *S. cerevisiae* genome encodes at least 15 members of the PPR family, all of which localize to the mitochondrion. Moreover, data collected about their molecular functions or phenotypes of their mutants allow speculating that all these factors are involved in the expression of the mitochondrial genome. On the basis of literature data, integrated from high-throughput physical and genetic interaction studies (data from the BIOGRID database (Stark et al. 2006)), as well as single gene analyses, it is clear that the network of yeast PPR proteins forms a subnetwork of the yeast interactome, implicated in mitochondrial RNA processing (Supplementary Figure S2). The presence of PPR motifs within poorly characterized proteins like Yer077cp, Msc6p or Sov1p suggests that they should exhibit an RNA-binding activity. This result can guide further genetic and biochemical studies needed to verify this hypothesis and eventually identify their RNA targets.

In yeast, expression and maintenance of a small mitochondrial genome is regulated by a large set of nuclear encoded factors that localize in the mitochondrion and modulate replication, transcription, post-transcriptional processing and translation (reviewed in (Contamine and Picard 2000) and (Lipinski, Kaniak-Golik, Golik 2010). PPR proteins in yeast, like their plant counterparts, seem to play essential nonredundant roles, as most of them are indispensable for respiration and mtDNA stability.

The rate of amino acid substitution is widely used to measure functional constraints on gene products (Nei 2005). Comparison of sequence similarities between pairs of orthologs, derived from various yeast species, clearly indicates that the yeast PPR protein family is subjected to accelerated divergent evolution. Genes encoding these proteins fall into a relatively small group of yeast genes evolving so quickly that BLAST analysis encounters

problems with assessing homology relationships (Wolfe 2004). Previously, it was reported that functional categories of genes with increased evolutionary rates in post-WGD (whole genome duplication) species are enriched in factors involved in mitochondrial translation and their physical interaction partners (Jiang et al. 2008). The relaxed functional role of the mitochondrion in post-WGD species cannot, however, fully explain the phenomenon observed in the case of the PPR protein family, as the increased rate of evolution of its members is apparent also in pre-WGD species. Moreover, sequence similarities between orthologs of mitochondrial ribosomal proteins are higher than between orthologs of PPR proteins. On the other hand, accelerated rate of sequence divergence in this protein family seems not to be a feature simply resulting from structural properties, as structurally similar TPR proteins evolve much more slowly.

The observed accelerated rate of sequence divergence might be explained by coevolution of a protein and its RNA target. The increased evolutionary rate of PPR proteins could be driven by divergence of, for example, UTRs of particular mitochondrial RNAs. Incompatibility of mitochondrial genomes and nuclear-encoded factors involved in their expression was shown to cause hybrid sterility and reproductive isolation in *Saccharomyces* yeasts (Herbert et al. 1992; Lee et al. 2008; Chou et al. 2010) and proposed as a general mechanism driving speciation (Chou and Leu 2010). Interestingly, one of the protein factors involved in cytonuclear incompatibility in two species belonging to the *Saccharomyces sensu stricto* complex (*S. cerevisiae* and *S. bayanus*) is a PPR protein – Aep2 (Lee et al. 2008). *S. bayanus* Aep2p fails to facilitate translation of the *S. cerevisiae ATP9* mRNA and multiple mutations are critical for this functional diversification. The hypothesis of a nucleo-organellar genetic buffering has been already proposed to explain the rapid expansion of PPR proteins in terrestrial plants and the functional role of this family in general (Schmitz-Linneweber and Small 2008).

Analysis of single and double mutant alleles of a *S. cerevisiae* PPR protein Dmr1 suggests another mechanism that can contribute to the rapid evolution of this protein family. Intragenic genetic interactions between amino acid substitutions located in different PPR motifs of this protein suggest that cooperative functional interactions between motifs could contribute to a high mutation rate by enforcing concerted changes in several motifs. In the case of *S. bayanus* and *S. cerevisiae* Aep2 sequences, discussed above, multiple concerted critical mutations were found to be involved in the functional change and their separation resulted in nonfunctional proteins (Lee et al. 2008). Both the coevolution of PPR proteins and their RNA targets and intragenic genetic interactions due to the cooperative binding mechanism can therefore act together in shaping the rapid divergent evolution of this protein family.

In summary, by applying an iterative hidden Markov model and homolog-clustering approach we designed a sensitive motif-detection method and applied it to the analysis of the Pentatricopeptide Repeat (PPR) protein family in 14 yeast species. Our data indicate that the PPR family is much more abundant in these fungi than previously reported, varying from 9 to 15 per genome. All newly identified members of this family in *S. cerevisiae* localize to mitochondria and seem to play a role in post-transcriptional steps of gene regulation, presumably as RNA-binding factors. Moreover, all display a highly accelerated rate of sequence divergence, which may reflect an evolutionary interdependence of nuclear encoded RNA-binding proteins and their organellar RNA targets, as well as a cooperative interaction between motifs.

Supplementary Material

Supplementary Figure S1 (pdf). Distribution of the PPR proteins discovered using SCIPHER analysis in genomes of 15 yeast species from *Saccharomycotina* and *Taphrinomycotina* (Ascomycota).

Supplementary Figure S2 (pdf). PPR proteins in the physical and genetic interactome of yeast mitochondria (data from the BIOGRID database (Stark et al. 2006)). Putative PPR proteins are in red boxes. Protein nodes (circles) are colored by localization as in the legend, RNA nodes are shown as rectangles, PPR proteins are in red. Green lines indicate genetic interactions, black lines indicate physical intreractions, red lines – both genetic and physical interactions. A Cytoscape (Shannon et al. 2003) source file for this network is provided in Supplementary data file S3.

Supplementary Data S1 (txt file in FASTA format). PPR motifs identified in the last (5^{th}) iteration of the SCIPHER search in genomes of 15 yeast species. The header line for each motif begins with a number indicating its position in the sequence.

Supplementary data S2 (xls file). Coordinates and sequences of PPR motifs identified in *S. cerevisiae* proteins by SCIPHER profiles and by extrapolation from orthologous sequences.

Supplementary data S3 (cys file). A Cytoscape (Shannon et al. 2003) file describing the interaction network shown in Supplementary Figure S2.

Acknowledgements

This work was supported by grant N N303 550639 from the Ministry of Science and Higher Education of Poland and by the Operational Programme Innovative Economy POIG.02.02.00-14-024/08-00. We are grateful to prof. Ewa Bartnik for critical reading of the manuscript.

Literature Cited

Alexeyenko, A, I Tamas, G Liu, EL Sonnhammer. 2006. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. Bioinformatics 22:e9-15.

Bhalerao, RP, K Salchert, L Bako, L Okresz, L Szabados, T Muranaka, Y Machida, J Schell, C Koncz. 1999. Regulatory interaction of PRL1 WD protein with Arabidopsis SNF1-

like protein kinases. Proceedings of the National Academy of Sciences of the United States of America 96:5322-5327.

Bonnefoy, N, N Bsat, TD Fox. 2001. Mitochondrial translation of Saccharomyces cerevisiae COX2 mRNA is controlled by the nucleotide sequence specifying the pre-Cox2p leader peptide. Mol Cell Biol 21:2359-2372.

Buchan, DW, SM Ward, AE Lobley, TC Nugent, K Bryson, DT Jones. 2010. Protein annotation and modelling servers at University College London. Nucleic Acids Res 38 Suppl:W563-568.

Burke, D, D Dawson, T Stearns. 2000. Methods in Yeast Genetics. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Chou, JY, YS Hung, KH Lin, HY Lee, JY Leu. 2010. Multiple molecular mechanisms cause reproductive isolation between three yeast species. PLoS Biol 8:e1000432.

Chou, JY, JY Leu. 2010. Speciation through cytonuclear incompatibility: insights from yeast and implications for higher eukaryotes. Bioessays 32:401-411.

Cliften, P, P Sudarsanam, A Desikan, L Fulton, B Fulton, J Majors, R Waterston, BA Cohen, M Johnston. 2003. Finding functional features in Saccharomyces genomes by phylogenetic footprinting. Science 301:71-76.

Contamine, V, M Picard. 2000. Maintenance and integrity of the mitochondrial genome: a plethora of nuclear genes in the budding yeast. Microbiol Mol Biol Rev 64:281-315.

Davies, SM, O Rackham, AM Shearwood, KL Hamilton, R Narsai, J Whelan, A Filipovska. 2009. Pentatricopeptide repeat domain protein 3 associates with the mitochondrial small ribosomal subunit and regulates translation. FEBS Lett 583:1853-1858.

Delannoy, E, WA Stanley, CS Bond, ID Small. 2007. Pentatricopeptide repeat (PPR) proteins as sequence-specificity factors in post-transcriptional processes in organelles. Biochem Soc Trans 35:1643-1647.

Dieckmann, CL, TJ Koerner, A Tzagoloff. 1984. Assembly of the mitochondrial membrane system. CBP1, a yeast nuclear gene involved in 5' end processing of cytochrome b pre-mRNA. J Biol Chem 259:4722-4731.

Dieckmann, CL, LK Pape, A Tzagoloff. 1982. Identification and cloning of a yeast nuclear gene (CBP1) involved in expression of mitochondrial cytochrome b. Proceedings of the National Academy of Sciences of the United States of America 79:1805-1809.

Dietrich, FS, S Voegeli, S Brachat, et al. 2004. The Ashbya gossypii genome as a tool for mapping the ancient Saccharomyces cerevisiae genome. Science 304:304-307.

Dujardin, G, P Pajot, O Groudinsky, PP Slonimski. 1980. Long range control circuits within mitochondria and between nucleus and mitochondria. I. Methodology and phenomenology of suppressors. Mol Gen Genet 179:469-482.

Dujon, B, D Sherman, G Fischer, et al. 2004. Genome evolution in yeasts. Nature 430:35-44.

Eddy, SR. 1998. Profile hidden Markov models. Bioinformatics 14:755-763.

Ellis, TP, HB Lukins, P Nagley, BE Corner. 1999. Suppression of a nuclear aep2 mutation in Saccharomyces cerevisiae by a base substitution in the 5'-untranslated region of the mitochondrial oli1 gene encoding subunit 9 of ATP synthase. Genetics 151:1353-1363.

Finn, RD, J Mistry, B Schuster-Bockler, et al. 2006. Pfam: clans, web tools and services. Nucleic Acids Res 34:D247-251.

Finn, RD, J Mistry, J Tate, et al. 2010. The Pfam protein families database. Nucleic Acids Res 38:D211-222.

Finnegan, PM, MJ Payne, E Keramidaris, HB Lukins. 1991. Characterization of a yeast nuclear gene, AEP2, required for accumulation of mitochondrial mRNA encoding subunit 9 of the ATP synthase. Curr Genet 20:53-61.

Fisk, DG, MB Walker, A Barkan. 1999. Molecular cloning of the maize gene crp1 reveals similarity between regulators of mitochondrial and chloroplast gene expression. Embo J 18:2621-2630.

Funes, S, JM Herrmann. 2007. Analysis of mitochondrial protein synthesis in yeast. Methods Mol Biol 372:255-263.

Gietz, RD, A Sugino. 1988. New yeast-Escherichia coli shuttle vectors constructed with in vitro mutagenized yeast genes lacking six-base pair restriction sites. Gene 74:527-534.

Gietz, RD, RA Woods. 2002. Transformation of yeast by lithium acetate/single-stranded carrier DNA/polyethylene glycol method. Methods Enzymol 350:87-96.

Greenleaf, AL, JL Kelly, IR Lehman. 1986. Yeast RPO41 gene product is required for transcription and maintenance of the mitochondrial genome. Proceedings of the National Academy of Sciences of the United States of America 83:3391-3394.

Groth, C, RF Petersen, J Piskur. 2000. Diversity in organization and the origin of gene orders in the mitochondrial DNA molecules of the genus Saccharomyces. Mol Biol Evol 17:1833-1841.

Helfenbein, KG, TP Ellis, CL Dieckmann, A Tzagoloff. 2003. ATP22, a nuclear gene required for expression of the F0 sector of mitochondrial ATPase in Saccharomyces cerevisiae. J Biol Chem 278:19751-19756.

Herbert, CJ, C Macadre, AM Becam, J Lazowska, PP Slonimski. 1992. The MRS1 gene of S. douglasii: co-evolution of mitochondrial introns and specific splicing proteins encoded by nuclear genes. Gene Expr 2:203-214.

Hollingsworth, MJ, NC Martin. 1986. RNase P activity in the mitochondria of Saccharomyces cerevisiae depends on both mitochondrion and nucleus-encoded components. Mol Cell Biol 6:1058-1064.

Huerta-Cepas, J, A Bueno, J Dopazo, T Gabaldon. 2008. PhylomeDB: a database for genome-wide collections of gene phylogenies. Nucleic Acids Res 36:D491-496.

Huh, WK, JV Falvo, LC Gerke, AS Carroll, RW Howson, JS Weissman, EK O'Shea. 2003. Global analysis of protein localization in budding yeast. Nature 425:686-691.

Islas-Osuna, MA, TP Ellis, LL Marnell, TM Mittelmeier, CL Dieckmann. 2002. Cbp1 is required for translation of the mitochondrial cytochrome b mRNA of Saccharomyces cerevisiae. J Biol Chem 277:37987-37990.

Jeffries, TW, IV Grigoriev, J Grimwood, et al. 2007. Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast Pichia stipitis. Nat Biotechnol 25:319-326.

Jiang, H, W Guan, D Pinney, W Wang, Z Gu. 2008. Relaxation of yeast mitochondrial functions after whole-genome duplication. Genome research 18:1466-1471.

Johnson, LS, SR Eddy, E Portugaly. 2010. Hidden Markov model speed heuristic and iterative HMM search procedure. BMC bioinformatics 11:431.

Jones, DT. 1999. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292:195-202.

Jones, T, NA Federspiel, H Chibana, et al. 2004. The diploid genome sequence of Candida albicans. Proceedings of the National Academy of Sciences of the United States of America 101:7329-7334.

Karpenahalli, MR, AN Lupas, J Soding. 2007. TPRpred: a tool for prediction of TPR-, PPR- and SEL1-like repeats from protein sequences. BMC bioinformatics 8:2.

Karplus, K. 2009. SAM-T08, HMM-based protein structure prediction. Nucleic acids research 37:W492-497.

Kellis, M, BW Birren, ES Lander. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature 428:617-624.

Koszul, R, A Malpertuy, L Frangeul, C Bouchier, P Wincker, A Thierry, S Duthoy, S Ferris, C Hennequin, B Dujon. 2003. The complete mitochondrial genome sequence of the pathogenic yeast Candida (Torulopsis) glabrata. FEBS Lett 534:39-48.

Krause, K, R Lopes de Souza, DG Roberts, CL Dieckmann. 2004. The mitochondrial message-specific mRNA protectors Cbp1 and Pet309 are associated in a high-molecular weight complex. Mol Biol Cell 15:2674-2683.

Lee, HY, JY Chou, L Cheong, NH Chang, SY Yang, JY Leu. 2008. Incompatibility of nuclear and mitochondrial genomes causes hybrid sterility between two yeast species. Cell 135:1065-1073.

Lightowlers, RN, ZM Chrzanowska-Lightowlers. 2008. PPR (pentatricopeptide repeat) proteins in mammals: important aids to mitochondrial gene expression. Biochem J 416:e5-6.

Lipinski, KA, A Kaniak-Golik, P Golik. 2010. Maintenance and expression of the S. cerevisiae mitochondrial genome-From genetics to evolution and systems biology. Biochim Biophys Acta 1797:1086-1098.

Lurin, C, C Andres, S Aubourg, et al. 2004. Genome-wide analysis of Arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. Plant Cell 16:2089-2103.

Matsuoka, M, M Matsubara, J Inoue, M Kakehi, T Imanaka. 1994. Organization and transcription of the mitochondrial ATP synthase genes in the yeast Yarrowia lipolytica. Curr Genet 26:382-389.

Merz, S, B Westermann. 2009. Genome-wide deletion mutant analysis reveals genes required for respiratory growth, mitochondrial genome maintenance and mitochondrial protein synthesis in Saccharomyces cerevisiae. Genome Biol 10:R95.

Mulero, JJ, TD Fox. 1993. PET111 acts in the 5'-leader of the Saccharomyces cerevisiae mitochondrial COX2 mRNA to promote its translation. Genetics 133:509-516.

Naithani, S, SA Saracco, CA Butler, TD Fox. 2003. Interactions among COX1, COX2, and COX3 mRNA-specific translational activator proteins on the inner surface of the mitochondrial inner membrane of Saccharomyces cerevisiae. Mol Biol Cell 14:324-333.

Nakamura, T, K Meierhoff, P Westhoff, G Schuster. 2003. RNA-binding properties of HCF152, an Arabidopsis PPR protein involved in the processing of chloroplast RNA. Eur J Biochem 270:4070-4081.

Nei, M. 2005. Selectionism and neutralism in molecular evolution. Mol Biol Evol 22:2318-2342.

Nosek, J, L Tomaska, M Bolotin-Fukuhara, I Miyakawa. 2006. Mitochondrial chromosome structure: an insight from analysis of complete yeast genomes. FEMS Yeast Res 6:356-370.

Nouet, C, M Bourens, O Hlavacek, S Marsy, C Lemaire, G Dujardin. 2007. Rmd9p controls the processing/stability of mitochondrial mRNAs and its overexpression compensates for a partial deficiency of oxa1p in Saccharomyces cerevisiae. Genetics 175:1105-1115.

O'Toole, N, M Hattori, C Andres, K Iida, C Lurin, C Schmitz-Linneweber, M Sugita, I Small. 2008. On the expansion of the pentatricopeptide repeat gene family in plants. Mol Biol Evol 25:1120-1128.

Payne, MJ, PM Finnegan, PM Smooker, HB Lukins. 1993. Characterization of a second nuclear gene, AEP1, required for expression of the mitochondrial OLI1 gene in Saccharomyces cerevisiae. Curr Genet 24:126-135.

Payne, MJ, E Schweizer, HB Lukins. 1991. Properties of two nuclear pet mutants affecting expression of the mitochondrial oli1 gene of Saccharomyces cerevisiae. Curr Genet 19:343-351.

Poutre, CG, TD Fox. 1987. PET111, a Saccharomyces cerevisiae nuclear gene required for translation of the mitochondrial mRNA encoding cytochrome c oxidase subunit II. Genetics 115:637-647.

Prikryl, J, M Rojas, G Schuster, A Barkan. 2011. Mechanism of RNA stabilization and translational activation by a pentatricopeptide repeat protein. Proceedings of the National Academy of Sciences of the United States of America 108:415-420.

Puchta, O, M Lubas, KA Lipinski, J Piatkowski, M Malecki, P Golik. 2010. DMR1 (CCM1/YGR150C) of Saccharomyces cerevisiae encodes an RNA-binding protein from the pentatricopeptide repeat family required for the maintenance of the mitochondrial 15S ribosomal RNA. Genetics 184:959-973.

Pusnik, M, I Small, LK Read, T Fabbro, A Schneider. 2007. Pentatricopeptide repeat proteins in Trypanosoma brucei function in mitochondrial ribosomes. Mol Cell Biol 27:6876-6888.

Rackham, O, SM Davies, AM Shearwood, KL Hamilton, J Whelan, A Filipovska. 2009. Pentatricopeptide repeat domain protein 1 lowers the levels of mitochondrial leucine tRNAs in cells. Nucleic Acids Res 37:5859-5867.

Remm, M, CE Storm, EL Sonnhammer. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol 314:1041-1052.

Rice, P, I Longden, A Bleasby. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16:276-277.

Rodeheffer, MS, BE Boone, AC Bryan, GS Shadel. 2001. Nam1p, a protein involved in RNA processing and translation, is coupled to transcription through an interaction with yeast mitochondrial RNA polymerase. J Biol Chem 276:8616-8622.

Rogowska, AT, O Puchta, AM Czarnecka, A Kaniak, PP Stepien, P Golik. 2006. Balance between transcription and RNA degradation is vital for Saccharomyces cerevisiae mitochondria: reduced transcription rescues the phenotype of deficient RNA degradation. Mol Biol Cell 17:1184-1193.

Saha, D, AM Prasad, R Srinivasan. 2007. Pentatricopeptide repeat proteins and their emerging roles in plants. Plant Physiol Biochem 45:521-534.

Scannell, DR, AC Frank, GC Conant, KP Byrne, M Woolfit, KH Wolfe. 2007. Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. Proceedings of the National Academy of Sciences of the United States of America 104:8397-8402.

Schmitz-Linneweber, C, I Small. 2008. Pentatricopeptide repeat proteins: a socket set for organelle gene expression. Trends Plant Sci 13:663-670.

Shadel, GS. 2004. Coupling the mitochondrial transcription machinery to human disease. Trends Genet 20:513-519.

Shannon, P, A Markiel, O Ozier, NS Baliga, JT Wang, D Ramage, N Amin, B Schwikowski, T Ideker. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research 13:2498-2504.

Sigrist, CJ, L Cerutti, E de Castro, PS Langendijk-Genevaux, V Bulliard, A Bairoch, N Hulo. 2010. PROSITE, a protein domain database for functional characterization and annotation. Nucleic acids research 38:D161-166.

Small, ID, N Peeters. 2000. The PPR motif - a TPR-related motif prevalent in plant organellar proteins. Trends Biochem Sci 25:46-47.

Stark, C, BJ Breitkreutz, T Reguly, L Boucher, A Breitkreutz, M Tyers. 2006. BioGRID: a general repository for interaction datasets. Nucleic Acids Res 34:D535-539.

Stribinskis, V, HC Heyman, SR Ellis, MC Steffen, NC Martin. 2005. Rpm2p, a component of yeast mitochondrial RNase P, acts as a transcriptional activator in the nucleus. Mol Cell Biol 25:6546-6558.

Stribinskis, V, KS Ramos. 2007. Rpm2p, a protein subunit of mitochondrial RNase P, physically and genetically interacts with cytoplasmic processing bodies. Nucleic Acids Res 35:1301-1311.

Thompson, DA, FW Stahl. 1999. Genetic control of recombination partner preference in yeast meiosis. Isolation and characterization of mutants elevated for meiotic unequal sister-chromatid recombination. Genetics 153:621-641.

Tian, GL, C Macadre, A Kruszewska, et al. 1991. Incipient mitochondrial evolution in yeasts. I. The physical map and gene order of Saccharomyces douglasii mitochondrial DNA discloses a translocation of a segment of 15,000 base-pairs and the presence of new introns in comparison with Saccharomyces cerevisiae. J Mol Biol 218:735-746.

Williams, EH, CA Butler, N Bonnefoy, TD Fox. 2007. Translation initiation in Saccharomyces cerevisiae mitochondria: functional interactions among mitochondrial ribosomal protein Rsm28p, initiation factor 2, methionyl-tRNA-formyltransferase and novel protein Rmd9p. Genetics 175:1117-1126.

Wolfe, K. 2004. Evolutionary genomics: yeasts accelerate beyond BLAST. Curr Biol 14:R392-394.

Wood, VR GwilliamMA Rajandream, et al. 2002. The genome sequence of Schizosaccharomyces pombe. Nature 415:871-880.

Zeng, X, A Hourset, A Tzagoloff. 2007. The Saccharomyces cerevisiae ATP22 gene codes for the mitochondrial ATPase subunit 6-specific translation factor. Genetics 175:55-63.

Ziaja, K, G Michaelis, T Lisowsky. 1993. Nuclear control of the messenger RNA expression for mitochondrial ATPase subunit 9 in a new yeast mutant. J Mol Biol 229:909-916.

Zivanovic, Y, P Wincker, B Vacherie, M Bolotin-Fukuhara, H Fukuhara. 2005. Complete nucleotide sequence of the mitochondrial DNA from Kluyveromyces lactis. FEMS Yeast Res 5:315-322.

Figure legends

Figure 1. The SCIPHER algorithm flow chart. The search starts by defining a target cluster of organisms, obtaining a current motif signature, and retrieving protein sequences already known to belong to the analyzed protein family. For each sequence BLAST analysis is performed in order to obtain direct orthologs and group proteins into orthologous groups (reciprocal BLAST criterion). The current motif signature is then used to locate motifs in each sequence and each orthologous group (OG). Next, all the found motifs are aligned and used to create a new motif signature (a HMM profile). New hits in already analyzed sequences are added directly to previous multiple sequence alignments (MSAs) (right). New sequences are analyzed on the basis of predefined true positivity criteria (TPC, see Materials and Methods for details) (left). If they fulfill TPC, new sequences are appended, orthologous groups are formed and analyzed as previously. The search is iterated until saturation.

Figure 2. Evolution of the complexity of the yeast PPR SCIPHER model. The chart presents the increase in the number of PPR protein sequences (dark) and individual PPR motifs (light) with each iteration of the SCIPHER method. In this study, the model reached its saturation at the fifth iteration, when no new motifs or sequences were found.

Figure 3. PPR proteins in *Saccharomyces cerevisiae* detected with the use of the SCIPHER algorithm. Systematic ORF names are given in parentheses. Dark gray boxes represent PPR motifs detected directly by the HMM profiles, while the light gray boxes are regions sharing significant sequence similarity with PPR motifs detected in orthologous sequences from other yeast species analyzed in this work. The empty box represents the DNA-dependent RNA-polymerase domain of Rpo41p (inferred from the PFAM database). Coordinates and sequences of PPR motifs shown in the figure are provided in a Supplementary Data S2 file.

Figure 4. Accelerated divergent evolution of the PPR protein family in yeast. Conservation of the PPR protein family in yeast genomes was estimated by calculating the mean sequence

identity between pairs of direct orthologs of 14 PPR proteins (Dmr1p, Pet309p, Aep3p, Rpm2p, Rmd9p, Cbp1p, Msc6p, Pet111p, Yer077cp, Atp22p, Aep2p, Sov1p, Aep1p) in *S. cerevisiae* vs. *K. lactis* (A) *S. cerevisiae* vs. *C. glabrata* (B) and *S. cerevisiae* vs. *D. hansenii* (C) comparisons. The obtained mean percentages of identities were compared with control data sets derived from the analysis of 30 pairs of orthologs of TPR proteins (structural similarity), proteins constituting small (12 proteins) and large (37 proteins) subunits of mitochondrial ribosomes (functional relatedness). Other modular RNA binding proteins – RRM family (22 proteins) and Pumilio (PUF) family (6 proteins) were also included as controls. In order to relate the observed evolution rates to the genomic context, these mean percentage identities were shown in relation to the plots of distributions of percentage identity between pairs of homologous proteins from the four yeast species (data from (Dujon et al. 2004)), where the X-axis represents sequence similarity between pairs of orthologs (in 5% intervals) while the Y-axis shows relative frequency of pairs exhibiting that similarity). Mean ortholog identities for PPR proteins, TPR proteins, mitochondrial ribosomal proteins (MRP), RRM proteins and Pumilio (PUF) are indicated relative to the distributions. In (C) the values for MRP, RRM and PUF are within 2% of each other and are indicated using a single arrow.

Figure 5. A second-site intragenic mutation partially alleviates the phenotype of a point hypomorphic allele of *S. cerevisiae DMR1* encoding a PPR protein. (A) Amino acid substitutions in a single (*dmr1-1*) and double (*dmr1-2*) mutant of *DMR1*. Gray boxes in the top diagram denote PPR motifs detected by SCIPHER analysis. Conservation of mutated residues in orthologs from 15 yeast species is indicated in the alignment. (B) A single substitution in *dmr1-1* imparts a slow-growth phenotype on respiratory medium (glycerol) at a permissive temperature, while the growth of a double mutant (*dmr1-2*) is closer to that observed with a wild-type *DMR1* allele. Neither *dmr1-1* nor *dmr1-2* grow on respiratory

medium at the restrictive temperature (36.5 °C). The *dmr1Δ* nullomorph results in a complete loss of respiratory growth at both temperatures. (C) Mitochondrial protein synthesis revealed by *in vivo* ^{35}S incorporation with cytosolic translation inhibited by cycloheximide and SDS-PAGE of labeled proteins. Incorporation in the single *dmr1-1* mutant is about 50% of that observed in the wild-type *DMR1* strain, while that of the double *dmr1-2* mutant is close to wild-type. Cells were grown at permissive temperature using galactose as the carbon source.

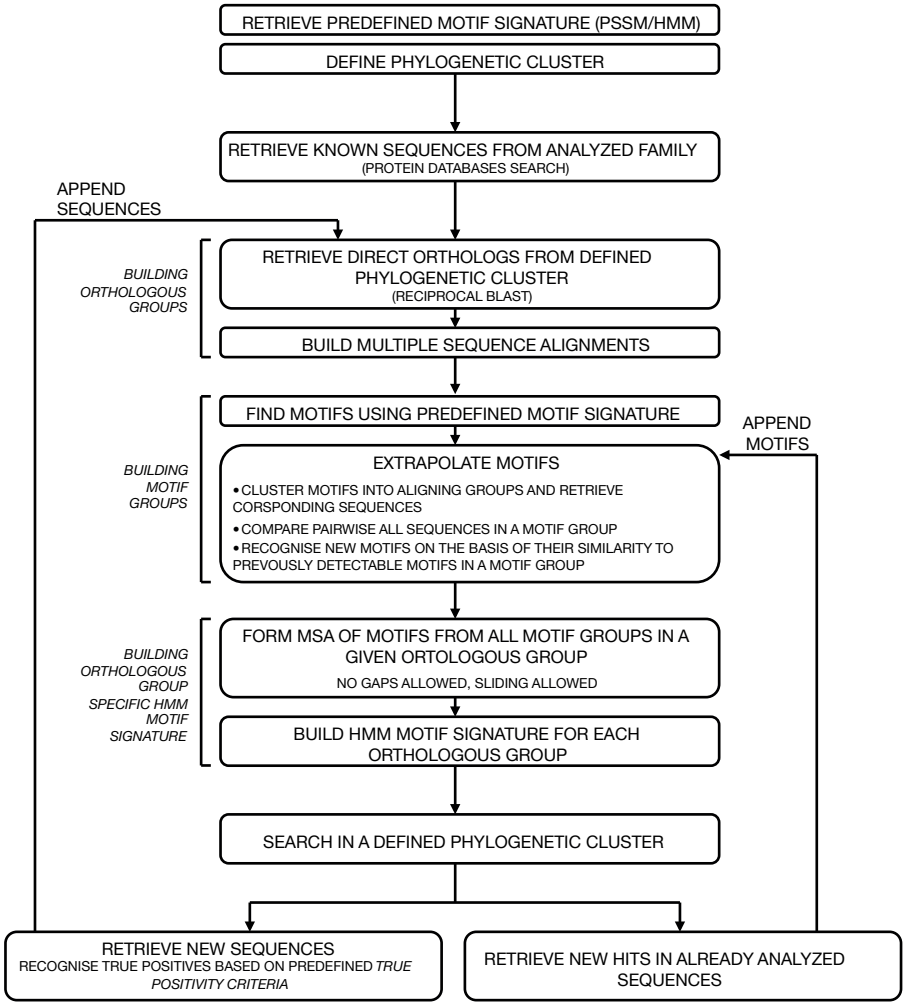

RETRIEVE PREDEFINED MOTIF SIGNATURE (PSSM/HMM)
DEFINE PHYLOGENETIC CLUSTER
RETRIEVE KNOWN SEQUENCES FROM ANALYZED FAMILY
(PROTEIN DATABASES SEARCH)
APPEND SEQUENCES
BUILDING ORTHOLOGOUS GROUPS
RETRIEVE DIRECT ORTHOLOGS FROM DEFINED PHYLOGENETIC CLUSTER
(RECIPROCAL BLAST)
BUILD MULTIPLE SEQUENCE ALIGNMENTS
FIND MOTIFS USING PREDEFINED MOTIF SIGNATURE
APPEND MOTIFS
BUILDING MOTIF GROUPS
EXTRAPOLATE MOTIFS
• CLUSTER MOTIFS INTO ALIGNING GROUPS AND RETRIEVE CORSPONDING SEQUENCES
• COMPARE PAIRWISE ALL SEQUENCES IN A MOTIF GROUP
• RECOGNISE NEW MOTIFS ON THE BASIS OF THEIR SIMILARITY TO PREVOUSLY DETECTABLE MOTIFS IN A MOTIF GROUP
BUILDING ORTHOLOGOUS GROUP SPECIFIC HMM MOTIF SIGNATURE
FORM MSA OF MOTIFS FROM ALL MOTIF GROUPS IN A GIVEN ORTOLOGOUS GROUP
NO GAPS ALLOWED, SLIDING ALLOWED
BUILD HMM MOTIF SIGNATURE FOR EACH ORTHOLOGOUS GROUP
SEARCH IN A DEFINED PHYLOGENETIC CLUSTER
RETRIEVE NEW SEQUENCES
RECOGNISE TRUE POSITIVES BASED ON PREDEFINED TRUE POSITIVITY CRITERIA
RETRIEVE NEW HITS IN ALREADY ANALYZED SEQUENCES

Figure 1

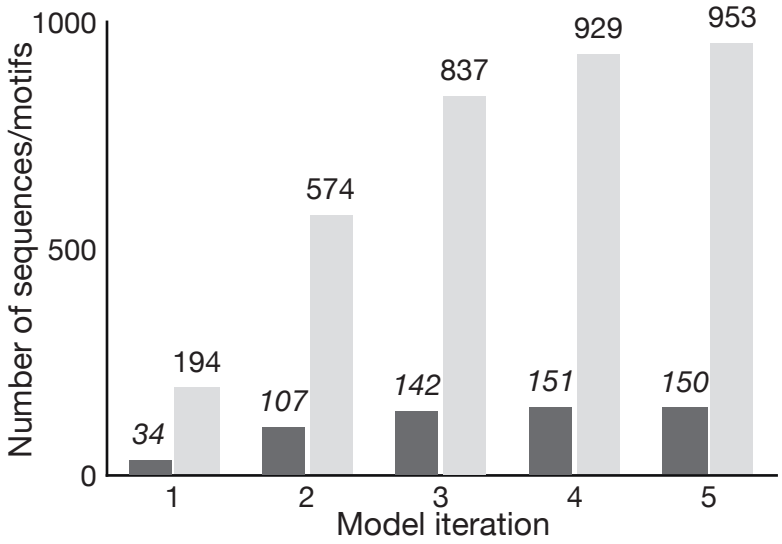

1000
500
0
Number of sequences/motifs
34
194
107
574
142
837
151
929
150
953
1
2
3
4
5
Model iteration

Figure 2

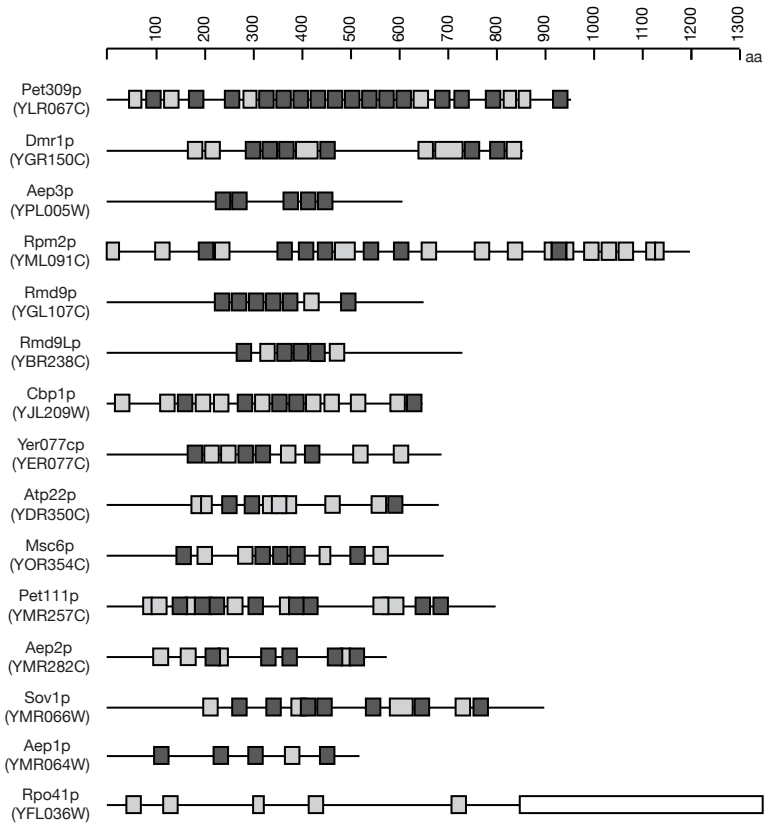

100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
aa
Pet309p
(YLR067C)
Dmr1p
(YGR150C)
Aep3p
(YPL005W)
Rpm2p
(YML091C)
Rmd9p
(YGL107C)
Rmd9Lp
(YBR238C)
Cbp1p
(YJL209W)
Yer077cp
(YER077C)
Atp22p
(YDR350C)
Msc6p
(YOR354C)
Pet111p
(YMR257C)
Aep2p
(YMR282C)
Sov1p
(YMR066W)
Aep1p
(YMR064W)
Rpo41p
(YFL036W)

Figure 3

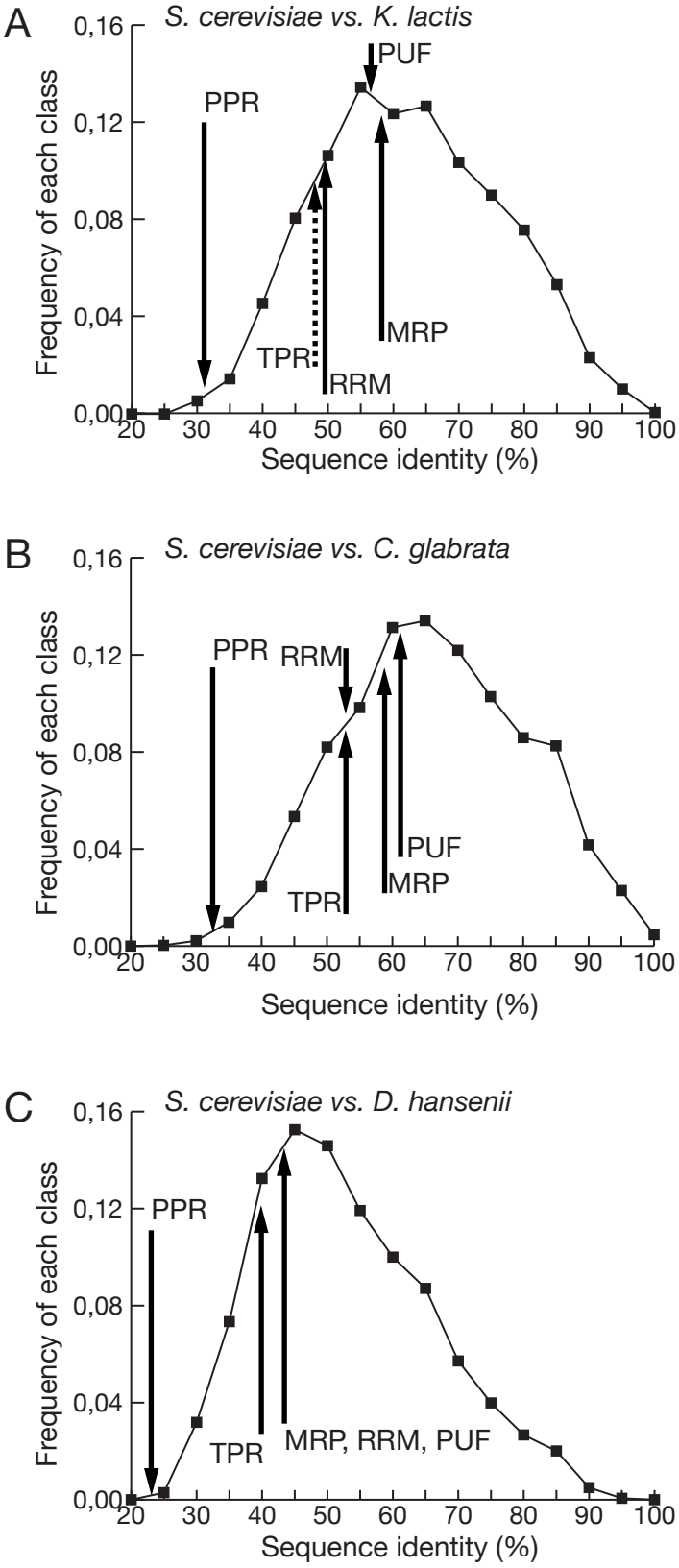
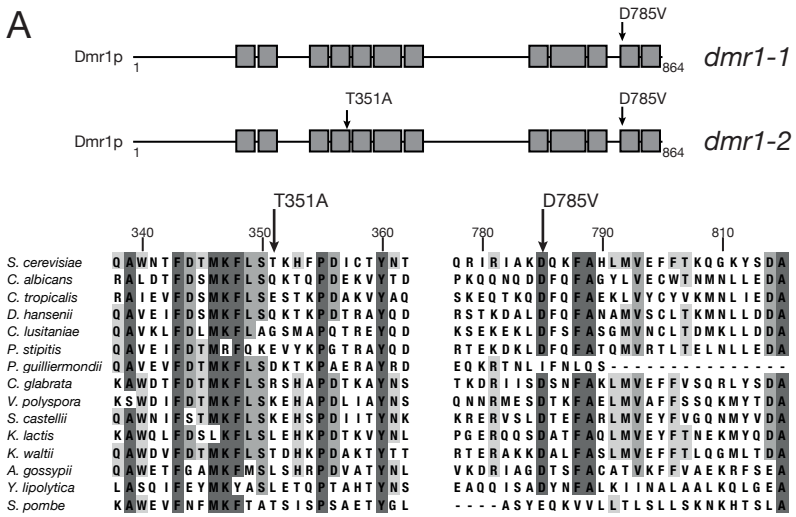
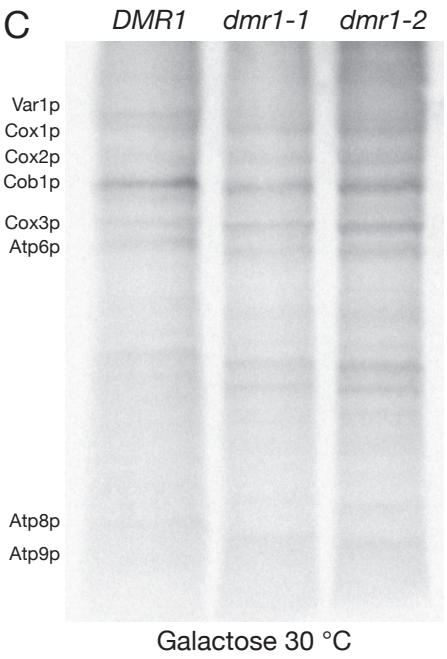

A
S. cerevisiae vs. K. lactis
PUF
PPR
MRP
TPR
RRM
Frequency of each class
Sequence identity (%)
B
S. cerevisiae vs. C. glabrata
PPR
RRM
PUF
MRP
TPR
Frequency of each class
Sequence identity (%)
C
S. cerevisiae vs. D. hansenii
PPR
TPR
MRP, RRM, PUF
Frequency of each class
Sequence identity (%)

Figure 4


A
D785V
Dmr1p
1
864
dmr1-1
T351A
D785V
Dmr1p
1
864
dmr1-2
T351A
D785V
340
350
360
780
790
810
S. cerevisiae
C. albicans
C. tropicalis
D. hansenii
C. lusitaniae
P. stipitis
P. guilliermondii
C. glabrata
V. polyspora
S. castellii
K. lactis
K. waltii
A. gossypii
Y. lipolytica
S. pombe


C
DMR1
dmr1-1
dmr1-2
Var1p
Cox1p
Cox2p
Cob1p
Cox3p
Atp6p
Atp8p
Atp9p
Galactose 30 °C

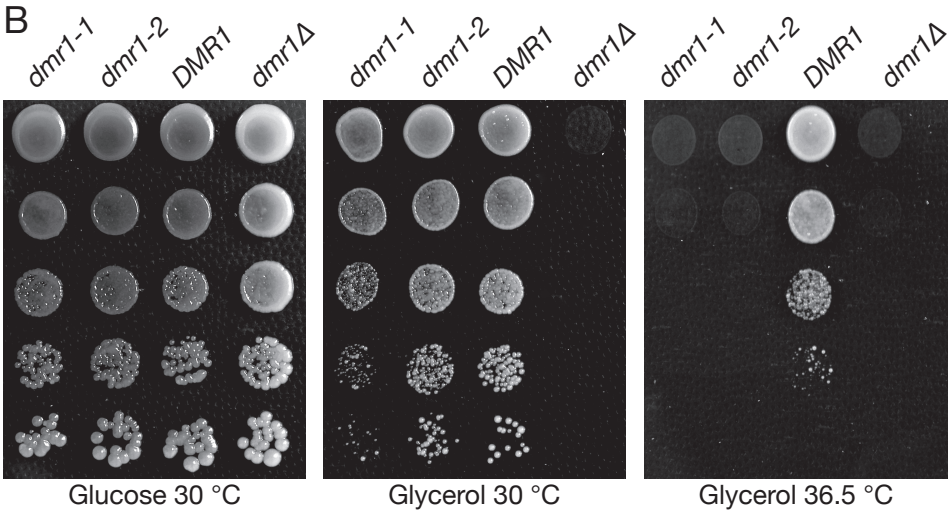

B
dmr1-1
dmr1-2
DMR1
dmr1Δ
Glucose 30 °C
Glycerol 30 °C
Glycerol 36.5 °C

Figure 5