British Mycological Society promoting fungal science

journal homepage: www.elsevier.com/locate/fbr

Fungal Biology Reviews

Review

Coding fungal tandem repeats as generators of fungal diversity

Emma LEVDANSKY, Haim SHARON, Nir OSHEROV[*]

Department of Human Microbiology, Sackler School of Medicine, Tel-Aviv University, Ramat-Aviv, 69978 Tel-Aviv, Israel

ARTICLE INFO

Article history:
Received 10 April 2008
Received in revised form
21 July 2008
Accepted 12 August 2008
Published online ■

Keywords:
Coding tandem repeats
Fungal diversity
Microsatellites
Pathogenesis
Simple sequence repeats

ABSTRACT

Coding tandem repeats are adjacent sequences that are directly repeated. The repeated units can be identical or partially degenerate. They are completely contained within a coding sequence and are composed of repeated units in which copy number does not disrupt the reading frame. They have been observed in viruses, prokaryotes and eukaryotes. The benefits offered by repeats include the modular construction of new proteins and introduction of rapidly evolving protein sequences which allow faster adaptation to new environments. Here we review the subject of tandem repeats and their relevance in fungi. Emphasis is given to repeat-containing fungal cell wall proteins and their role in generating diversity, adaptation to the environment, immunogenicity, adhesion, and pathogenesis. We describe in detail the recent studies analyzing coding tandem repeats in the model yeast *Saccharomyces cerevisiae* and the important human pathogens *Candida albicans* and *Aspergillus fumigatus*. Numerous unanswered questions are highlighted, providing a rich hunting ground for future research.

1. Tandem repeats: an overview

Tandem repeats (TRs, or simply ‘repeats’) are adjacent DNA sequences of 2–200 nucleotides in length that are directly repeated, the repeated units of which may be identical or partially degenerate (Pâques *et al.*, 2001; Strand *et al.*, 1993). TRs are also known as microsatellites, or simple sequence repeats (SSRs) when they are shorter than 10 nucleotides and as minisatellites when they are 10–200 nucleotides long. Repeats were described in the Archaea, Bacteria and Eucaryota kingdoms as well as in viruses (Bart-Delabesse *et al.*, 2001; Metzgar *et al.*, 2001; Trivedi, 2006). Most repeats are in non-coding regions, but some are found in coding sequences or pseudogenes (Verstrepen *et al.*, 2004).

Repeats are caused by replication slippage, genetic recombination during mitosis or meiosis and double strand break repair. Repeat variability is an outcome of three main genetic mechanisms: (i) by DNA strand slippage during replication. This occurs at the repetitive sequences when the new strand mispairs with the template strand. Backward slippage leads to insertional mutations whereas forward slippage to deletions (Kunkel, 1993), (ii) by genetic recombination following unequal crossing-over between the repeats on homologous chromosomes during meiosis and in mitotically dividing cells, resulting in the addition of repeats to one allele and a reduction to the other (Pearson *et al.*, 2005), (iii) by double strand break repair in which repair of the break is mediated by sequence information from a sister or homologous chromosome, leading to

[*] Corresponding author. Tel.: +972 3 640 9599; fax: +972 3 640 9160.
E-mail address: nosherov@post.tau.ac.il (N. Osherov).

1749-4613/$ – see front matter © 2008 The British Mycological Society. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.fbr.2008.08.001

changes in the number of repeats or the generation of chimeric genes (Richard *et al.*, 1999).

Tandem repeats can be roughly classified in two categories – non-coding and coding.

Non-coding repeats can subtly affect gene expression. Eukaryotic non-coding repeats can be found in 5′ untranslated regions (UTR), in introns and in 3′ UTRs. In mammals repeat variations in 5′-UTRs can regulate gene expression by affecting transcription and translation (Kenneson *et al.*, 2001). Repeat expansions and variations in 3′-UTRs can cause transcription slippage and produce expanded mRNA that can disrupt splicing and, possibly, disrupt other cellular functions (Mankodi *et al.*, 2002). Mammalian intronic repeats can affect gene transcription, mRNA splicing, or export to the cytoplasm (Davis *et al.*, 1997; Meloni *et al.*, 1998; Sirand-Pugnet *et al.*, 1995). Surprisingly, little is known about the effects of non-coding repeats in fungi. In general, fungal non-coding repeats appear to be distributed randomly throughout the genome, and there are relatively few of them compared to the number of coding repeats. Unlike coding repeats, they are not necessarily composed of nucleotide triplets and they are generally shorter and less skewed towards a high GC content (Fabre *et al.*, 2002; Richard and Dujon, 2006; and our unpublished work). However, the role of non-coding fungal repeats in modulating gene expression and RNA stability in pathogenic fungi remains to be determined.

Coding repeats generate variability in all living organisms. Coding repeats are located in-frame within the coding sequence of the gene, and are transcribed into mRNA and translated into a protein product. Coding repeat expansions and/or contractions can lead to a gain or loss of gene function *via* frameshift mutations or expanded toxic mRNA (Garcia-Lopez *et al.*, 2008). They can also lead to more subtle phenotypic changes by altering the number of in-frame coding repeats among different isolates leading to expansion or contraction of amino-acid blocks (Li *et al.*, 2004).

Coding repeats have been observed in viruses, archea, prokaryotes and eukaryotes. There is very little overlap between the repeat-rich genes in each of the three primary kingdoms (Marcotte *et al.*, 1999; Björklund *et al.*, 2006). On average eukaryotes have significantly higher incidences of coding repeats than prokaryotes and viruses, perhaps providing them with an extra source of variability to compensate for their low generation rate.

In viruses, comparative genomic studies of attenuated and virulent strains of *Gallid herpesvirus* 2 (GaHV-2) have identified differences in the number of repeats in the *UL36* and *UL47* genes that are correlated to virulence (Spatz and Silva, 2007). Glycoprotein I (gI) of *herpes simplex* virus type 1 (HSV-1) also contains a repeat region including the amino-acids serine and threonine, residues that can undergo *O*-glycosylation (Norberg *et al.*, 2007). This may lead to protease resistance (Byrd and Bresalier, 2004) and to variable structural rigidity of the extended region creating phenotypic alterations among different viral isolates.

Coding repeats are important in generating variability in several prokaryotic pathogens. By altering the morphology of cell-surface immunogenic antigens and adhesins, they enable these pathogens to evade the immune system thereby enhancing pathogenicity. Notable bacterial examples include *Streptococcal* alphaC, *emm* and *PspA* (Gravekamp *et al.*, 1998; Podbielski *et al.*, 1994; Waltman *et al.*, 1990), *Staphylococcus aureus* MSCRAMM genes (Patti *et al.*, 1994), *Neisseria meningitidis PilQ* and *DcaC* (Jordan *et al.*, 2003), and *Mycoplasma hyorhinis vlp* (Citti *et al.*, 1997).

Numerous repeats also exist in the ORFs of higher eukaryotes including *Drosophila melanogaster*, *Caenorhabditis elegans*, plants, mammals and humans (Katti *et al.*, 2001; Kantety *et al.*, 2002; Li *et al.*, 2004; Morgante *et al.*, 2002; Toth *et al.*, 2000).

Both coding and non-coding repeat expansions have been implicated in human disease. Expansions of simple DNA repeats are implicated in nearly 30 human hereditary disorders (Mirkin, 2007; Pearson *et al.*, 2005). Expandable repeats can be located in various regions of their resident genes: first, the coding regions, as occurs in numerous diseases mediated by polyglutamine or polyalanine runs in proteins; second, the 5′ untranslated regions (5′-UTRs), as in the case of fragile X syndrome, fragile X mental retardation associated with the *FRAXE* site, fragile X tremor and ataxia syndrome, and spinocerebellar ataxia 12; third, 3′-UTRs, as is observed for myotonic dystrophy 1, spinocerebellar ataxia 8 and Huntington's-disease-like 2; fourth, introns, as in the case of myotonic dystrophy 2, Friedreich's ataxia and spinocerebellar ataxia 10; and fifth, promoter regions, as occurs in progressive myoclonic epilepsy 1.

2. Coding fungal tandem repeats: an overview

Identification of coding fungal repeats. Several algorithms are available to detect tandem repeats in a nucleotide sequence, including ETANDEM (Rice *et al.*, 2000), mREPS (Kolpakov *et al.*, 2003), and Tandem Repeat Finder (TRF) (Benson, 1999). These linear programs calculate a repeat score based on the length of each repeat, the conservation of sequence between the repeats, and the number of repeat units. A recent non-linear model, SERV, produces a numerical VAR score that can predict the probability that a repeat sequence will vary in the number of repeats among different strains. A VAR score larger than 1 suggests a high probability that the repeats within a particular gene will vary among different strains or isolates of a particular species (Legendre *et al.*, 2007). In this review we used the SERV model analysis of the fungal coding repeats in *Aspergillus fumigatus*, *Saccharomyces cerevisiae* and *Candida albicans* (available at http://hulsweb1.cgr.harvard.edu/TandemRepeat/). This general non-linear model outperforms the models described above and is capable of predicting repeat variability for all types of tandem repeats (microsatellites and minisatellites) in a wide range of organisms spanning the major kingdoms of life (Legendre *et al.*, 2007). The tables we generated contain the VAR score and TRF score for each of the most repeat-rich genes in each category.

Significant coding repeats were identified in all three fungal species in approximately 1 % of all genes. It is probably safe to assume that repeat-containing genes will be found throughout the fungal kingdom (Karaoglu *et al.*, 2005).

Coding repeats were studied in detail in *S. cerevisiae* (Richard and Dujon, 2006; Verstrepen *et al.*, 2005), in the ALS adhesins from *C. albicans* (reviewed in Hoyer *et al.*, 2007) and in *A. fumigatus* (Levdansky *et al.*, 2007). We will first discuss the generalizations that can be deduced from these studies and then look at the specific findings for each species.

Repeats are found in all classes of fungal proteins. Genomic analysis reveals that coding repeats are found in ORFs that can be classified into three groups based on functional motifs: (i) proteins destined for transport to the plasma membrane and/or cell wall and containing a signal peptide sequence and a glycosylphosphatidylinositol (GPI)-anchor motif (Table 1, genes annotated with superscript b) or PIR (proteins with internal repeats) motifs (Table 1, genes annotated with superscript c), (ii) proteins containing a signal peptide sequence only that are destined primarily for secretion (Table 2), and (iii) proteins lacking these motifs, being located inside the cell (Table 3).

The first group of genes encoding repeat-rich cell wall or plasma membrane proteins will be the focus of this review because of their potential ability to mediate interactions between the organism and its surroundings. The second group, which encodes proteins with a potential to be secreted, has not been studied in detail and contains primarily uncharacterized genes (Table 2). They are a diverse group of genes, with little overlap among the three species. Interestingly, the *S. cerevisiae* genome contains relatively few genes in this category. Of the few that have been characterized (Table 2, underlined) several potentially interesting findings emerge: (i) the *MFalpha* gene encodes the secreted alpha factor mating pheromone of *S. cerevisiae* and *C. albicans* and contains three repeats. The protein is cleaved by a Kex2 protease into 3 repeat-containing fragments, each one a pheromone peptide in itself (Fuller *et al*., 1988; Panwar *et al*., 2003). This mechanism can help to amplify and modify the mating signal. (ii) Ankyrin and WD40 domain repeats are found in the two most repeat-rich genes in *A. fumigatus* (*Afu1g01020* and *Afu7g08500*) (Table 2). These repeats are typically found in proteins involved in signal transduction, pre-mRNA processing and cytoskeleton assembly (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml). They form a rigid repeat structure that is involved in protein–protein interactions, suggesting that the putative secreted proteins encoded by *Afu1g01020* and *Afu7g08500* form protein complexes. The third group, which encodes repeat-rich intracellular proteins, also contains primarily uncharacterized genes (Table 3). Perhaps not unexpectedly, the Ubiquitin gene (Ub), encoded as a linear repeat of individual Ub molecules, is found in all 3 species of fungi. There are, however, several specific findings for each species: in *A. fumigatus*, the WD40 domain encoding genes *Afu7g01700*, *Afu7g07030* and *Afu7g079030* are highly homologous to the *Podospora anserina hetD* and *hetE* genes, involved in vegetative incompatibility. In *P. anserina*, both genes require a minimal number of 11 WD40 repeats to be active in incompatibility (Espagne *et al*., 2002). In *S. cerevisiae* there is enrichment of genes encoding nuclear proteins and in particular, helicases. The four helicases identified contain highly similar repeats, and all are similar to helicases that are encoded within subtelomeric Y′ elements and are involved in telomerase-independent telomere maintenance (Yamada *et al*., 1998). In *C. albicans*, four of the nine genes identified encode for genes involved in stress responses (*ASR1*, *ASR2*, *DDR48* and *PNG2*) although the function of these genes has not been elucidated.

Repeats are more commonly found in fungal cell wall proteins (CWPs) than in other classes of proteins. There is a substantial enrichment of putative cell-surface proteins which contain internal repeats (Fig 1). For example, in *A. fumigatus* 4 of the 100 most repeat-rich genes (4 %) encode size-variable GPI-anchored CWPs, whereas this class of gene constitutes only 0.8 % of the number of genes in the genome, a 5-fold enrichment (Table 4 and Levdansky *et al*., 2007). Similarly, an unexpectedly large fraction (12.5 %) of *S. cerevisiae* CWPs contains tandem repeats (Table 4 and Verstrepen *et al*., 2004, 2005). Interestingly, in *C. albicans*, a commensal pathogen, the total number of repeat-rich CWPs is substantially larger than that found in *S. cerevisiae* or *A. fumigatus* (Table 4). This is probably because *C. albicans* has undergone a large increase in the number of genes encoding repeat-rich cell-surface adhesins, enabling it to adapt to life in the human host.

The number of repeats in many repeat-rich fungal CWPs varies among isolates, generating diversity. There is abundant experimental evidence demonstrating that the number of repeats in many of the repeat-rich fungal CWPs varies among isolates of the same species of fungus (see genes designated with superscript d in Table 1 and b in Tables 2 and 3). These CWPs include most of the ALS adhesin genes in *C. albicans* (reviewed in Hoyer, 2001; Hoyer *et al*., 2007), all four genes encoding GPI-anchored proteins in *A. fumigatus* (Levdansky *et al*., 2007) and most of the agglutinins and CWPs in *S. cerevisiae* (Verstrepen and Klis, 2006). This variability is proposed to generate diversity within a population of cells, for example endowing sub-populations with differing adhesive abilities. Under changing external conditions, such as changes in the adhesive properties of the substrate or host, there is a greater probability that some of these sub-populations will be able to adapt and thrive.

Many repeat-containing fungal CWPs are involved in adhesion. Many of the fungal adhesins contain tandem repeats, including the *S. cerevisiae FLO* genes that mediate adhesion of yeast cells in suspension to form large aggregates or ‘flocs’, and the *C. albicans ALS*, *EAP1* and *HWP1* adhesins that mediate adhesion to the host (Table 1). The function of several of these genes (*FLO1*, *ALS1*, *ALS3* and *ALS5*) is affected by the number of repeats they contain. Adhesion increases with additional repeats until an optimum number of repeats are reached (Loza *et al*., 2004; Oh *et al*., 2005; Rauceo *et al*., 2006; Verstrepen *et al*., 2005). The reason for this is not entirely clear. Adhesion apparently resides in the N-terminal binding domain of these proteins, whereas the repeats are found within their central regions and are probably not directly involved in adhesion. The repeats often encode Ser/Thr amino-acid residue (see Table 1) that are heavily mannosylated (Verstrepen and Klis, 2006). The mannosylated repeat region has been proposed to either (a) form an elongated stalk to present the binding domain at the cell wall surface (Hoyer *et al*., 2007; Loza *et al*., 2004) or (b) form covalent bonds to the cell wall polysaccharides, tightly anchoring and stabilizing the adhesin within the cell wall. This may enhance adhesion by securely presenting the N-terminal ligand-binding domain towards the substrate (Sheppard *et al*., 2004) or (c) alter the spatial structure of the binding domain thereby increasing its affinity to the substrate (Rauceo *et al*., 2006).

Table 1 – Top ranking repeat-rich putative CWPs[a]

Fungal species/ gene number	Annotation	TRF score	VAR score	Repeat consensus sequence
A. fumigatus				
AFU3G08990[b,d]	Cell-surface protein	738	1.55	QPSVPG
AFU2G05150[b,d]	Cell wall galactomannoprotein MP2/allergen	651	0.53	ETSTPCETTTTTTT
AFU4G09600[b,d]	GPI-anchored protein, putative	568	−0.62	RGFHKRGGGDTTVIGGPSGDDGGNSAEVEFESTYESSVKDYYKDDHSVDIENHVIHPPPVFHPPPV
AFU6G14090[b,d]	CFEM domain protein	196	1.24	GS
S. cerevisiae				
FLO1[b,d]	Flocculation protein FLO1	2690	−1.82	TTTEPWNGTFTSTSTEMTTVTGTNGLPTDETIIVIRTPTTATTA
FLO9[b,d]	Flocculation protein FLO9	2481	−1.82	TAITTTQPWNDTFTSTSTEMTTVTGTNGLPTDETIIVIRTPTTA
FLO5[b,d]	Flocculation protein FLO5	1619	−1.82	TEPWTGTFTSTSTEMTTITGTNGQPTDETVIVIRTPTSEGLITTT
FLO10[b,d]	Flocculation protein FLO10	1548	−1.83	TSSFSSSSEVCTECTETESTSTSTPYVTSSSSSSSEVCTECTETESTSYVTPYVSSSTAAAN
HKR1[d]	Mucin, osmosensor	1518	4.44	SAPVAVSSTYTSS
MUC1/FLO11[d]	Mucin-like, flocculation	1231	−1.83	SSTTESSSAPVPTPSSSTTESSSAPVTSSTTESSSAPVPTPSTSSNITSSAPVPTP
DAN4	Cell wall protein	976	2.27	TSTTSTTSTTPTTSTTST
FIT1[d]	Cell wall protein, involved in iron retention	821	−1.83	ETSVAAETSVAEPSTSAQGTSADEGSGSSITTTITATKNGHVYTKTVTQDATFVWTGEGERAPASTVATV
TIR4[c,d]	Cell wall mannoprotein of the Srp1p/Tip1p family	615	1.77	SSSVAPSSSEVV
HPF1[b]	Mannoprotein, glucosidase	555	2.64	SQVSDTPVSYTTSSSS
YNL190W[b]	Cell wall protein	555	2.54	THKYGKFNKTSKSKTPNHTG
SED1[b,d]	Stress-induced CWP	544	−1.83	SGSSVSGSTSTTESGSSASSSSSATESGSSASGSSSATESGSSVSGSSTATESGSSSAT
EGT2[b,d]	Cell wall endoglucanase	533	−1.83	TTEYTVVTEYTTYCPEPTTFTTNGKTYTVTEPTTLTITDCPCTIEKPTTTS
MSB2[d]	Mucin, osmosensor	477	−1.59	ESVVAGYSTTVGAAQYAQHTSLVPVSTIKGSKTSLSTE
PIR1[c,d]	Protein PIR1 (covalently linked cell wall protein)	413	0.32	AAVSQIGDGQIQATTKTTA
HSP150 (PIR2)[d]	Heat-shock protein (covalently linked cell wall protein)	409	1.67	AAVSQIGDGQVQATTKTTA
AGA1[b,d]	A-agglutinin mating attachment subunit	399	0.93	TSPSST
WSC3	Cell wall integrity sensor	271	1028	TSST
TIR3[c]	Cell wall protein	242	−0.15	SSAA
TIR2[c]	Cell wall protein	219	0.21	SSAVASSSEASSTETTSSAVASSSEA
MTL1	Mid2 p like cell wall sensor	149	1.04	SSSS
C. albicans				
ALS2[b,d]	ALS family adhesin	4852	−1.67	NPTVTTTEYWSQSYATTTTVTGPPGGTDTVIIREPP
ALS4[b,d]	ALS family adhesion	3948	−1.55	NPTVTTTEYWSQSYATTTTVTAPPGGTDTVIIREPP
ALS9[b,d]	ALS family adhesin	2847	−1.41	NPTVTTTEFWSESFASTTTITNPPDGTNSVIVKEPH
ALS1[b,d]	ALS family adhesin	1696	−1.83	NHTVTTTEYWSQSYATTTTVTAPPGGTDTVIIREPP
PGA55[b]	Putative CWP, unknown function	1543	−1.83	SSSSEV
ALS3[b,d]	ALS family adhesin	1492	−1.83	NPTVTTTEYWSQSYTTTTTVIAPPGGTDSVIIREPP
CSA1[b]	Heme-binding cell-surface CFEM domain protein	1096	−1.83	SINGFADRIYDQLPECAKPCMFQNTGVTPCPYWDTGCLCIMPTFAGAIGSCIAEKCKGQDVVSATSLGTSICSVAGVWDPYWMVPANVQSSLSAAATAVASSSEQPVETSSEPAGSSQSVESSQPAETSSSEPAETSSSEPAETSSETSSEQPASSEPAETSSEESSTITSAPSTPEDNPYTIYPSVAKTASINGFADRIYDQLPECAKPCMFQNTGVTPCPYWDTGCLCIMPTFAGAIGSCIAEKCKGQDVVSATSLGTSICSVAGVWDPYWMVPANVQSS
EAP1[b]	Cell wall adhesin	908	−1.8	TPAAPGTPVESQPVIPGTETTPAAPGTPVESQPATTPVAPGTE
HYR3[b]	Putative CWP, unknown function	794	−1.74	TSEYTTTWTTTNSDGSVSTESGIVSQSGSSFTTITTFAPDA
PGA18[b]	Putative CWP, unknown function	717	−1.83	SSSATTPGTSSVESTPGSSSATTPGSSTIESTSGSSSATTPGSSSATTPG

ALS5[b]	ALS family adhesin	531	−1.38	NPTVTTTEFWSESYATTETITNYPEGTDSVIVREPH
ALS6[b]	ALS family adhesin	508	−1.43	NPTVTTTEFWSESFATTTVTNGPEGTDSVIVREPH
PGA25[b]	Putative CWP, unknown function	503	−1.83	VGWIVGISVSQSVSSSSSSEVADFVGRTVIDPDPVGMIVAV
PGA62[b]	Putative CWP, unknown function	463	−1.83	TTVVTITSCEENKCHETEVTTGVTTVTEGDTTYTTYCPLPTTEAPAPATSTDVS
PGA54[b]	Putative CWP, unknown function	422	−1.83	EDNETITSTILQYVTVTSSDTTYVSATNTLTTTLTTKPTQAITPKKKKT
PIR1[c]	Structural glucan-linked CWP	421	−1.51	TVQPVAQISDGQIQHQTVKASATPVQQIGDGQIQHQ
IFF5[b]	Putative CWP, unknown function	417	−1.82	YIPTIIHSSDIQTQFISTWTATNSDGSVVTESGVVSQSGTSLTTI
RBR3[b]	Putative CWP, unknown function	404	−1.82	YHIEYFCSNYLSGAVETEFTSTWVVVTILMDQCLRIRYCRSVGYI
PGA23[b]	Putative CWP, unknown function	372	1.23	GAADTATSGAAGAAKLLPQVP
HWP1[b]	Adhesin	372	−1.12	QEPCDYPQQQP
YWP1[b]	Adhesin	345	−1.83	TYCPLTSYETVESTKVITILACDENKCQETTAEATPTEATTVVEGVVTEY
ALS7[b,d]	ALS family adhesin	340	−1.39	NPTVTTTKFWSESFATTETITNGPQGTDSVIIKEPH
PGA58[b]	Putative CWP, unknown function	337	−1.83	PQPPQLLQLPQLLQLAPSASAPAPAPPASPAALAPAPSAPAPAPEQPEQPA
RBT1[b]	Virulence-associated CWP	324	−1.7	TTPESSAPESSVPESSAPE
IFF6[b]	Putative CWP, unknown function	300	0.59	DSSTDSNTGATESSTATDTNTDAT
IFF4[b]	Adhesin	211	0.08	TPSESSLLVKQTSKNHHILMKCF
RBR1[b]	CWP essential for filamentous growth	181	−0.31	SAASAAKSGA
HYR1[b]	Putative CWP, unknown function	168	0.67	GSNNGSG
CHT2[b]	Putative chitinase	165	−0.49	QSATTTSAAVT
IFF8[b]	Putative CWP, unknown function	162	0.86	NNN
HWP2[b]	Putative CWP, unknown function	159	−0.42	STTPIISSA
PGA57[b]	Putative CWP, unknown function	130	−0.68	GHSSGGGHSSS
PGA39[b]	Putative CWP, unknown function	118	−0.04	TTDSA
PGA42[b]	Putative CWP, unknown function	116	0.18	TEYSSF
PGA37[b]	Putative CWP, unknown function	112	−1.03	SSSGSRGGSRGG
PGA60[b]	Putative CWP, unknown function	110	−0.69	SNESLTTT

a Genes with a TRF score > 100 are characterized as top ranking.

b Gene encoding a putative GPI-anchored CWP.

c Gene encoding a putative PIR-CWP.

d Gene containing repeats that vary in number among strains.

Table 2 – Top ranking repeat-rich putative secreted proteins[a]

Fungal species/ gene number	Annotation	TRF score	VAR score	Repeat consensus sequence
A. fumigatus				
AFU1G01020	NACHT and Ankyrin domain protein	1876	0.39	KLLIDKGADVNVRDNDGWTPLSRASDEGHEEVAKLLIDKGADVNVRDNDGWTPLSRALLSGHEE
AFU7G08500	NACHT and WD40 domain protein	821	−0.12	SVAFSPDGQRIVSGSDDNTIKLWDAQTGSELQSLQGHSDSVH
AFU5G03760	Class III chitinase ChiA1	766	0.3	VASSTPVVPGTSASSSPVSSSSAVASSTPVVPGTSASSSPVSSSSAVASSTPVVPGTSTSPSTPAIPGTSASSSPVSSSS
AFU1G04130	FG-GAP repeat protein, putative	675	−0.1	HQDPQHRHRPVEVHVASGASNYQTRIQEVGTTFYPEDNGVWQMIDFNRDGMLDLV
AFU1G05670	Conserved hypothetical protein	579	−0.23	TPELFKQICTLLNNGNNLLTADFVKEVNGLIGNANTLLTADFVKETRALIEAVAPML
AFU3G07400	Conserved hypothetical protein	575	−0.08	DPVCHKNSDCGPGVGYCYHGICLADPPKLTSRDDPICHKNSDCGPGVGYCYHGICVADPPKDPRERR
AFU3G13110	Extracellular serine–threonine rich protein	562	−0.09	TTTVVTYETVTTCPVTETISTSGTVTTSTYSTVSTVTLTSTATICTACEASTTPAPSAAPVTTAPAPEDM
AFU6G10930	Extracellular protein, putative	304	−0.34	VQPSVIISSQPAVRYKPQSSSQATAQLGYQPESQTTP
AFU8G00630	Conserved hypothetical protein	194	−0.7	SKAPASTTSKASASTTSKGSVST
AFU3G07870	Extracellular serine-rich protein, putative	170	2.0	SSSSSSSSSSSSSSSSSSSSSSSSSSSS
S. cerevisiae				
YIL169C	Putative protein	859	−1.83	FSKSYTTATVTHCDDNGCNTKTVTSEAPEATTTTVSPKTYTTATVTQCDDNGCSTKTVTSEAPEETSATT
YPL283W-A	Hypothetical protein	558	−1.38	VDTGSGSSTSPDVGAGSGSSISAGVGTCSGSRTSP
MNN4[b]	Positive regulator of mannosylphosphate transferase Mnn6p	449	0.96	EKKKKEE
MF (ALPHA)1[c]	Mating factor alpha-1	411	2.25	AEAWHWLQLKPGQPMYKREAD
YOR053W	Hypothetical protein	174	1.1	RR
SCW11[b]	Putative glucosidase	168	0.84	TSS
PRY2	Pathogen related protein	135	−0.41	PTTTAS
C. albicans				
orf19.1725	Hypothetical protein	886	−1.83	PGGSVVTVTVTESTVETITGPGFSTTVTLTPGTNVITSPTGPATEPTGPSTKPTG
orf19.206	Hypothetical protein	739	−1.83	GSSDDANTSSTDDSTDEISQTTTDSSSTATGIDDGDDENNDMKEYPQCFNKQDDQPKREHCCFDDNDRVLYPKPC
orf19.750	Hypothetical protein	467	−1.80	SVEESKRLDADVAAQLAVTF
RBR3	Putative CWP, no GPI anchor	404	−1.82	SSSSKSSSTTP
orf19.7606	Hypothetical protein	418	−1.75	AAAPSDPISQVIGLVSNILEGGFSTSGALLHNLIG
orf19.7167	Hypothetical protein	401	0.50	QSSSELSPESLSESLSESLSVPFHVI
MSB2	Uncharacterized protein	275	0.27	PTTSEAPDTPTTSEAPN
MFALPHA	Alpha factor mating pheromone	246	0.43	RDANAEAGFRLTNLVILNLV
orf19.4330	Hypothetical protein	225	−0.64	SLFLVHDLLVLSLMFLFLFSCSFCFSCS

a Top 10 genes with the highest TRF score were selected for each category.
b Gene containing repeats that vary in number among strains.
c Underlined, previously characterized gene.

Table 3 – Top ranking repeat-rich putative cellular proteins[a]

Fungal species/top 10 TR-rich proteins	Annotation	TRF score	VAR score	Repeat consensus sequence
A. fumigatus				
AFU7G07100	NACHT and WD repeat vegetative incompatibility domain protein	2153	0.64	QVLKGHENSVNAVAFSPDGQTVASASDDKTIRLWDAASGAEK
AFU2G17000	PT repeat family protein	2110	−8.42	AEPA
AFU7G08290[b]	Vegetative incompatibility WD repeat protein, putative	1496	0.45	QLLASGSDDKTIKLWDPTTGALKHTLEGHSDSIRSVAFSQDGQFLASGSHDKTIKLWDPTTGNLKHTLEGHSDWVRSVAFWKD
AFU7G07030	Vegetative incompatibility WD repeat protein, putative	1309	0.85	GHSDWVRSVAFSQNSQLLASGSDDKTIKLWDPTTGALKHTLEGHSDSIRSVAFSQDGQLLASGSDDETIKLWDPTTSALKQTLEGHSDSILTVAFSQDGQLLASGSHDKTIKLWDPTTGTLKHTLE
AFU7G08310[b]	Conserved hypothetical protein	1063	0.43	HTSSPPGDPLPRTSTGEGSDVSEPIRMDISESSDSEDLEPQPGVHTSSPPREPSPRTSIGEGSDVSEPATIDISESSDSRDPEPQPGA
Ubi4[b]	Polyubiquitin UbiD/Ubi4, putative	825	0.08	VKTLTGKTITLEVESSDTIDNVKSKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRLRGGCKS
AFU7G08240	Hypothetical protein	821	−0.08	FNPQPYLTYTPAPRPPDMSDPTQFGITRDLPFQQLHMTSASSDTQSDQSQMNITFD
AFU6G09340[b]	Hypothetical protein	729	1.45	SVSAL
AFU6G09360[b]	Proline–glycine rich protein, putative	559	−0.26	GVDAPYGVRTPRGTEATCGPRHP
AFUA7G07060[b]	Hypothetical protein	544	−0.37	HTSSPPREPSPRTSTGEGSDVSEPIRMDISESSDSEDPGPQPGA
S. cerevisiae				
NUM1[b]	Nuclear migration protein	3123	−1.83	ELEKKLEQPSLEYLVEHAKATDHHLLSDSAYEDLVKCKENPDMEFLKEKSAKLGHTVVSNEAYS
UBI4	Ubiquitin	1593	−1.83	ANFVKTLTGKTITLEVESSDTIDNVKSKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRLRGG
NSP1	Nucleoporin	1157	−1.83	FGAKSDENKASATSKPAFSFGAKPEEKKDDNSSKPAFSFGAKSNEDKQDGTAKPAFSFGAKPAEKNNNETSKPAFSFGAKSDEKKDGDASKPAFS
YJL225C	Putative ATP-dependent helicase	807	2.48	STNSSTNATTTE
YIL177C	Putative ATP-dependent helicase	807	2.487	STNSSTNATTTE
YMR317W	Hypothetical protein	720	1.98	SSVSSEAPSSTS
YEL077C	Hypothetical protein	718	−1.377	STNSSTNATTTASTNVRTSATTTASTNSNTSATTTE
YPR204W	DNA helicase	668	1.98	TNSSTSATTTE
YLL067C	Helicase-like	661	1.98	TNSSTSATTTE
C. albicans				
UBI4	Ubiquitin precursor	1237	−1.83	MQIFVKTLTGKTITLEVESSDTIDNVKSKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRLRGG
orf19.7239	Hypothetical protein	780	−1.83	AQPVSDNQDTLKTTVLPKEEPHHPSLAGEPGIVIPKEKDALSAFEKVEDADAKALNKNVTEVGTANA
orf19.267	Hypothetical protein	617	−1.83	PVKMSTASSASIVNSNVANESGSDGYIDIDIKAAGLAFVPVKTGVLQL
orf19.2296	Mucin-like hypothetical	515	−1.83	AGTGAGLAAGSSAHSHAAEQEPTHKSQLDPELKKDLYSQGYTKGKSSHSSGPSST
orf19.5401	Hypothetical protein	488	1.43	STSVVTPATNQESTTDTSSDNNV
ASR2	HSP-like gene regulated by cAMP and by osmotic stress	439	−1.41	AVDDVGIVLKDIKKGAEA
ASR1	HSP-like gene regulated by cAMP and by osmotic stress	437	−1.83	THGTTGYGSWRTGSHGASGAHDSTGYGSSQTGSHGTAGYGSSQTGTH
DDR48	Immunogenic stress-associated protein	331	−0.003	DSYGSSNTDSYGSSNRRGNDSYGSSN
PNG2	Caspofungin and azole induced gene	303	0.43	PHEPPHEPPHEP

[a] Top 10 genes with the highest TRF score were selected for each category.

[b] Gene containing repeats that vary in number among strains.

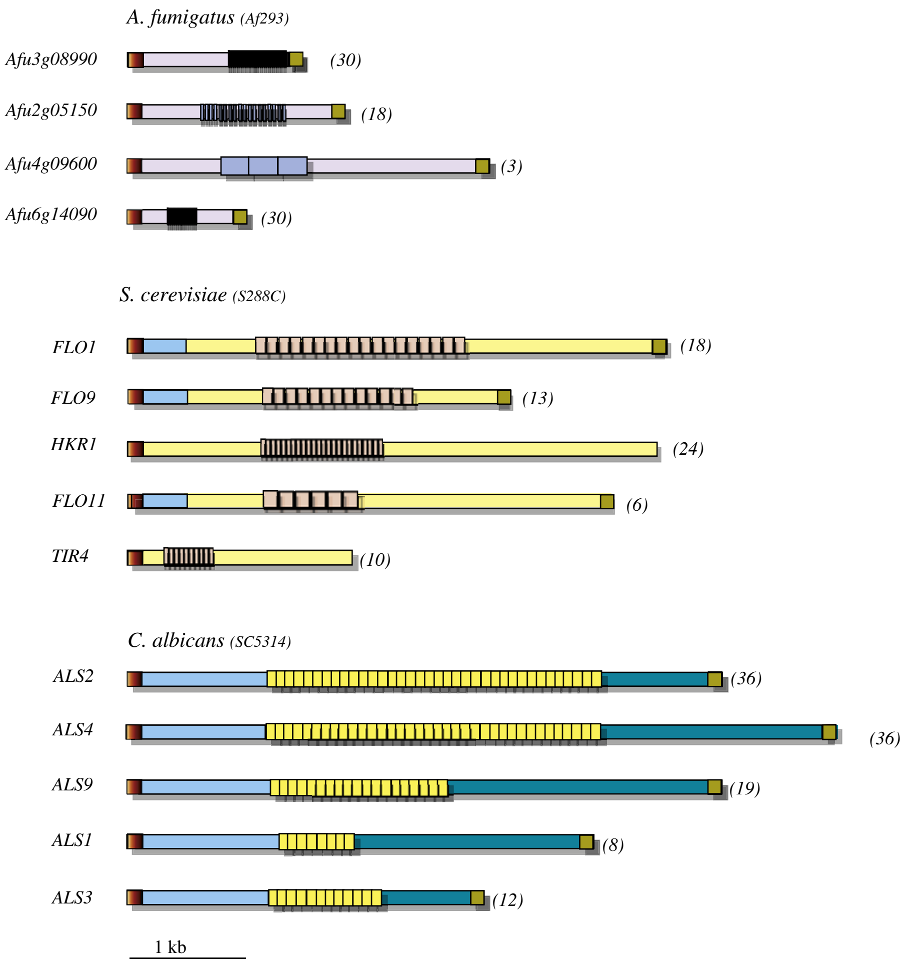

Fig. 1 – Schematic representation of top-scoring repeat-rich CWPs in *A. fumigatus* (strain Af293), *S. cerevisiae* (strain S288C) and *C. albicans* (strain SC5314). Scoring was performed using the SERV model (Legendre et al., 2007). All the genes depicted in the figure exhibit isolate-specific size variability. Key: red squares = leader sequence; light blue squares = ligand-binding domain; tan squares = GPI anchor motif. Note: leader sequences and GPI anchor motifs are not drawn to scale. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3. Coding repeats in *S. cerevisiae* CWPs

There are four main groups of repeat-containing CWPs in *S. cerevisiae*: (i) the flocculins encoded by *FLO1*, 5, 9, 10 and 11, (ii) the Pir family proteins that stabilize the cell wall (*PIR1*, 2), (iii) the Dan/Tir family of mannoproteins involved in adaptation to anaerobic conditions (*TIR2*, 3, 4 and *DAN4*) (Sertil et al., 2007) and (iv) the mucin-like HOG-pathway osmosensors Hkr1p and Msb2p (Table 1). Deletion of the *HKR1* and *MSB2* repeat domain leads to constitutive activation of the HOG pathway, suggesting that the repeats have an inhibitory role (Tatebayashi et al., 2007). Deletion of the repeat region of *PIR4*, that is closely related to *PIR1* and *PIR2*, results in the loss of binding of Pir4p to β-1,3 glucan, suggesting that the

Table 4 – Repeat-containing genes encoding putative CWPs are enriched in fungal genomes

Fungal species	# Genes with TRF > 100	# Genes encoding putative CWPs[a] (TRF > 100)	% Genes encoding putative CWPs (TRF > 100)	% of CWPs in genome[b]	Fold enrichment
A. fumigatus	100	4	4	0.8	5
C. albicans	233	36	15	1.7	8.8
S. cerevisiae	167	21	12.5	1	12.5

a CWPs include GPI-anchored and Pir proteins, and proteins lacking these motifs but experimentally shown to localize to the cell wall.
b Calculated by dividing the total number of putative CWPs in each organism, by the total number of genes in its genome.

repeats are directly involved in cross-linking the protein to the cell wall (Castillo *et al.*, 2003).

The functional role of repeats has been studied most extensively in the flocculins, and they will be discussed in more detail below.

Flocculins: a primer. Flocculins are GPI-anchored CWPs containing an N-terminal lectin-like substrate-binding domain followed by a conserved repeat element. *FLO1*, *9*, 5, and *10* are closely related, encoding proteins which bind mannose sugar residues on neighboring cells, promoting cell–cell adhesion to form multi-cellular clumps that sediment out of solution. This ability is used in the brewing industry to separate the yeast after fermentation is complete (Verstrepen and Klis, 2006). *FLO11* is more similar to *AGA1* and mucins and mediates hydrophobicity-based adhesion to abiotic surfaces. Flo11p is involved in diploid filamentation and haploid invasive growth. By binding the cells tightly to the agar, it enables them to resist washes and to tunnel into the substrate (Guo *et al.*, 2000).

Flocculins generate functional diversity through changes in the number of repeats and by epigenetic control of expression. To understand the effect of varying repeat number on flocculin function, Verstrepen *et al.* (2005) generated an isogenic series of *FLO1* mutant strains containing different numbers of repeats, and measured their ability to flocculate and to adhere to plastic. The results showed that there was a linear correlation between the number of repeats and the extent of adhesion: as the Flo1 protein became longer (carrying more repeats), the adhesion properties gradually became stronger. All *FLO* genes naturally vary in repeat number within a population of cells, suggesting that similar mechanisms may be generally applicable to the entire family.

The fact that *S. cerevisiae* contains numerous highly similar flocculin-encoding genes presents another advantage: the *FLO* repeats provide ideal sites for recombination and the generation of novel chimeric genes. This process can quickly generate diversity (Verstrepen and Klis, 2006). Also, *FLO1*, *FLO5* and *FLO9* genes have adjacent, truncated, non-functional copies, which are annotated as pseudogenes in the SWISSPROT/SGD/MIPS databases. These pseudogenes may provide a reservoir of sequences that could become incorporated into the adjacent functional *FLO* genes by recombination through the tandem repeats (Harrison *et al.*, 2002).

Another mechanism that generates diversity, at least for *FLO11*, is epigenetic switching of gene expression. Under strong inducing conditions, not all the cells continuously express Flo11p. This switching of *FLO11* between 'on' and 'off' states is due to reversible epigenetic repression by chromatin-binding proteins (Halme *et al.*, 2004). Those cells within the population that express Flo11p form a filament, whereas those that do not, continue to divide as single-celled yeast. This switching means that even a strain with a single *FLO11* gene has cells with two different cell surfaces: those that have Flo11p in their cell walls and those that do not.

Another possible reservoir of cell–cell variation is provided by the subtelomeric localization of *FLO1*, *5* and *10*, that, at least in laboratory yeast strains, silences their expression. However, the silent genes can be activated by mutations that occur at high frequency to the *IRA1* or *IRA2* genes, encoding Ras GTPase-activating proteins. In *IRA* null mutants, the *FLO10* gene is expressed and confers hyperfilamentation and hyperadhesion (Halme *et al.*, 2004).

4. Coding repeats in *C. albicans* CWPs

There are three main groups of characterized repeat-containing CWPs in *C. albicans* based on sequence homology: the ALS (agglutinin like sequence) family of adhesins (*ALS1-7*, *ALS9*), the *EAP1*/*HWP1* adhesins and *RBT1*, and the *PIR1* family protein that stabilizes the cell wall (Table 1) (De Groot *et al.*, 2003; Ruiz-Herrera *et al.*, 2006). Research towards understanding the role of repeats in these proteins has focused almost exclusively on Hwp1p and the ALS adhesins, and they will be highlighted in the proceeding section.

The N-terminal repeats in Hwp1p undergo covalent cross-linking to host cells. The 10-amino-acid long N-terminal repeat in the Hwp1p adhesin is rich in proline (P) and glutamine (Q) residues (Table 1). It undergoes transglutamination by endogenous host transglutaminases (TGases) to form covalent bonds between the Hwp1p glutamines to lysine residues on the cell surface of human buccal epithelial cells (BECs). The Hwp1p repeat is an extraordinary case of molecular mimicry: a similar 8-amino-acid repeat is found in mammalian small proline-rich (SPR) proteins that form a protective TGase-induced cross-linked barrier on human buccal and gingival tissues (Staab *et al.*, 2004). In essence, *C. albicans* hijacks this system by mimicking the sequence of the SPRs and inducing the endogenous TGases to stably cross-link it to the host surface. Deletion of *HWP1* in *C. albicans* reduces the stable adhesion of hyphae to BECs, and results in reduced virulence in a mouse model for systemic candidiasis, suggesting that Hwp1p-dependent adhesion may also occur in internal body tissues (Staab *et al.*, 1999).

The ALS family of adhesins: a brief overview. The ALS adhesins are a family of 8 genes (*ALS1-7*, *ALS9*) related to the *S. cerevisiae* alpha-agglutinins involved in mating (reviewed in Hoyer, 2001; Hoyer *et al.*, 2007). They are GPI-anchored CWPs containing an N-terminal adhesin domain followed by a conserved repeat element of 108 bp and a 3′ domain, both rich in Ser–Thr residues and heavily glycosylated. The current working model for the Als proteins is that the heavily glycosylated repeats and 3′ regions assume an elongated conformation that presents the N-terminal adhesin domain at the cell wall surface. Their primary role is to enable *C. albicans* cells to adhere to the host and in the formation of a biofilm (Hoyer *et al.*, 2007).

ALS genes generate functional diversity through changes in the number of repeats. There is widespread variability in the number of ALS repeats among isolates of *C. albicans*. For example, in a study of over 100 bloodstream isolates of *C. albicans*, the number of repeats in *ALS1* varied from 4 to 37 and the most common allele had 16 copies (Lott *et al.*, 1999). Similar variability has also been detected in *ALS3* and *ALS7* (Oh *et al.*, 2005; Zhang *et al.*, 2003). In contrast, there was less variation in the number of tandem repeat copies in *ALS5* and *ALS6* with a mean of nearly 5 copies for *ALS5* and nearly 4 copies for *ALS6* (Zhao *et al.*, 2007).

The evidence suggests that the number of repeats in the ALS genes correlates to *C. albicans* adhesion. Deletion of 15 of the 20 tandem repeats of *ALS1* and expression of the truncated gene in non-adherent *S. cerevisiae* cells reduced adherence by 50 %, whereas deletion of all the repeats abolished adherence completely (Loza *et al.*, 2004). Oh *et al.* (2005) engineered isogenic *C. albicans* strains to express a single functional copy of *ALS3* with either 9 or 12 repeats. Proteins with 12 repeats contributed more to *C. albicans* adhesion to endothelial or epithelial cells than did those with 9 copies. Rauceo et al. (2006) prepared *S. cerevisiae* strains expressing Als5p with 0–6 repeats. Adhesion to FN-coated beads and aggregation was positively correlated to the number of tandem repeats. Similar results were also shown for the *Candida glabrata EPA1* (epithelial adherence) gene encoding a flocculin-like adhesion (Frieman *et al.*, 2002).

Little is known about the contribution of adhesins to *C. albicans* virulence *in vivo*. Deletion of *ALS1* leads to reduced virulence in two murine models of disseminated candidiasis and oropharyngeal candidiasis (Fu *et al.*, 2002; Kamai *et al.*, 2002). However, there is currently no evidence directly linking the number of ALS repeats to altered virulence in animal models for candidiasis.

5. Coding repeats in *A. fumigatus* CWPs

The number of *A. fumigatus* CWPs containing high-scoring repeats is relatively small compared to that of *C. albicans* and *S. cerevisiae*, and they show no significant homology to any of the genes found in yeast. This may be a result of the large evolutionary distance between the yeast and the filamentous fungi, as they are estimated to have diverged 300–400 million years ago (Dujon, 2006). Ten of the highest scoring repeat-containing putative GPI-anchored putative CWP-encoding genes in *A. fumigatus* were analyzed for variability of the repetitive sequence among both clinical and environmental isolates (Levdansky *et al.*, 2007). In all, only the four highest scoring repeat-containing ORFs showed size variability of the repetitive region at both the DNA and RNA levels (*Afu4g09600*, *Afu2g05150/MP-2*, *Afu6g14090* and *Afu3g08990*) (Table 1) (Levdansky *et al.*, 2007). All four genes are conserved among the filamentous fungi and have no yeast homologs. They do not contain an N-terminal substrate-binding domain similar to that found in the *S. cerevisiae* flocculins or *C. albicans* adhesins.

Afu4g09600 encodes a hypothetical protein with 2–3 large (66 amino-acid) repeats. *Afu2g05150/AfMP-2* encodes an immunogenic protein (Afmp2p) of unknown function belonging to the antigenic mannoprotein superfamily (Chong *et al.*, 2004). It contains a variably sized Ser/Thr-rich repeat region (amino-acid residues 239–368) composed of a 13-amino-acid repeat. AfMP2p is found in the cell wall and culture medium of *A. fumigatus*. Patients with aspergilloma and invasive aspergillosis develop a specific antibody response against this protein, although it was not shown if the response is specifically directed to the repeat domain (Chong *et al.*, 2004). *Afu6g14090* has an N-terminal CFEM domain (amino-acid residues 18–85) adjacent to the variable-size Ser-rich repeat region (amino-acid residues 140–219). CFEM is a -fungus-specific eight-cysteine-containing domain. Some CFEM-containing proteins, such as the Pth11p receptor from *Magnaporthe grisea* and the Rbt5p plasma membrane-anchored heme-binding protein in *C. albicans*, are proposed to participate in fungal pathogenesis (Kulkarni *et al.*, 2003). *Afu3g08990* encodes a hypothetical protein conserved specifically in the aspergilli. It contains a variable 6-amino-acid Ser/Pro-rich repeat showing significant homology to repeats found in the immunoglobulin A-binding beta antigen of *Streptococcus agalactiae* and to the extended rod domain of mammalian type XXI collagen. Three of the four genes (*Afu2g05150*, *Afu3g08990*, and *Afu6g14090*) were deleted, but only *Afu3g08990* deletion resulted in a clear mutant phenotype. *Afu3g08990* deletion leads to rapid conidial germination and reduced adherence to extracellular matrix suggestive of an alteration in cell wall characteristics (Levdansky *et al.*, 2007). Mutant conidia exhibit an abnormal cell wall morphology and increased sensitivity to zymolase and mechanical agitation (our unpublished results). Deletion of *Afu3g08990* does not affect virulence in a murine model for disseminated aspergillosis. *Afu3g08990* protein is localized to the cell walls of dormant and germinating conidia and has been proposed to act like cement, strengthening and increasing the elasticity of the cell wall. Interestingly, the repeat region of *Afu3g08990* was recently used to subtype 55 outbreak isolates of *A. fumigatus*. The method was able to identify ‘‘clonal’’ and genotypically distinct *A. fumigatus* isolates, and could therefore be used in hospital settings to indicate the source of the fungal infection and the route of transmission in a rapid and accessible manner (Balajee *et al.*, 2007).

6. Conclusions

The field of coding fungal tandem repeats is now ripe with potential. The tools needed to identify and analyze coding repeat-containing genes in many species of fungi are now available. Yet, as can be determined from this review, we know little about most of these genes. What is the role of repeats in cellular and secreted fungal proteins? Does repeat number affect their function? For those genes that have been studied, much remains unclear. For example, what is the precise role of repeats in the *S. cerevisiae* and *Candida* adhesins? Are they important for virulence? Do they interact with the host immune system? Can it be experimentally proven that repeat variability confers selective advantages in pathogenesis?

Answering these questions in the near future should benefit both basic molecular biology and our understanding of fungal pathogenic strategies.

Acknowledgements

We thank Kevin Verstrepen and Yana Shadkchan for their comments and suggestions. Research on coding repeats in *A. fumigatus* in the Osherov lab is supported by the Israel Science Foundation (ISF) and International Cancer Research Foundation (ICRF).

REFERENCES

Balajee SA, Tay ST, Lasker BA, Hurst SF, Rooney AP, 2007. Characterization of a novel gene for strain typing reveals substructuring of *Aspergillus fumigatus* across North America. *Eukaryotic Cell* **6**: 1392–1399.

Bart-Delabesse E, Sarfati J, Debeaupuis JP, van Leeuwen W, van Belkum A, Bretagne S, Latge JP, 2001. Comparison of restriction fragment length polymorphism, microsatellite length polymorphism, and random amplification of polymorphic DNA analyses for fingerprinting *Aspergillus fumigatus* isolates. *Journal of Clinical Microbiology* **39**: 2683–2686.

Benson G, 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**: 573–580.

Björklund ÅK, Ekman D, Elofsson A, 2006. Expansion of protein domain repeats. *PLoS Computational Biology* **2**: e114.

Byrd JC, Bresalier RS, 2004. Mucins and mucin binding proteins in colorectal cancer. *Cancer Metastasis Review* **23**: 77–99.

Castillo L, Martinez AI, Garcerá A, Elorza MV, Valentín E, Sentandreu R, 2003. Functional analysis of the cysteine residues and the repetitive sequence of *Saccharomyces cerevisiae* Pir4/Cis3: the repetitive sequence is needed for binding to the cell wall beta-1,3-glucan. *Yeast* **20**: 973–983.

Chong KT, Woo PC, Lau SK, Huang Y, Yuen KY, 2004. *AFMP2* encodes a novel immunogenic protein of the antigenic mannoprotein superfamily in *Aspergillus fumigatus*. *Journal of Clinical Microbiology* **42**: 2287–2291.

Citti C, Kim MF, Wise KS, 1997. Elongated versions of *Vlp* surface lipoproteins protect *Mycoplasma hyorhinis* escape variants from growth-inhibiting host antibodies. *Infection and Immunity* **65**: 1773–1785.

Davis BM, McCurrach ME, Taneja KL, Singer RH, Housman DE, 1997. Expansion of a CUG trinucleotide repeat in the 39 untranslated region of myotonic dystrophy protein kinase transcripts results in nuclear retention of transcripts. *Proceedings of the National Academy of Sciences of the United States of America* **94**: 7388–7393.

De Groot PW, Hellingwerf KJ, Klis FM, 2003. Genome-wide identification of fungal GPI proteins. *Yeast* **20**: 781–796.

Dujon B, 2006. Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution. *Trends in Genetics* **7**: 375–387.

Espagne E, Balhadère P, Penin ML, Barreau C, Turcq B, 2002. *HET-E* and *HET-D* belong to a new subfamily of WD40 proteins involved in vegetative incompatibility specificity in the fungus *Podospora anserina*. *Genetics* **161**: 71–81.

Fabre E, Dujon B, Richard GF, 2002. Transcription and nuclear transport of CAG/CTG trinucleotide repeats in yeast. *Nucleic Acids Research* **30**: 3540–3547.

Frieman MB, McCaffery JM, Cormack BP, 2002. Modular domain structure in the *Candida glabrata* adhesin Epa1p, a beta1,6 glucan-cross-linked cell-wall protein. *Molecular Microbiology* **46**: 479–492.

Fu Y, Ibrahim AS, Sheppard DC, Chen YC, French SW, Cutler JE, Filler SG, Edwards JE, 2002. *Candida albicans* Als1p: an adhesin that is a downstream effector of the EFG1 filamentation pathway. *Molecular Microbiology* **44**: 61–72.

Fuller RS, Sterne RE, Thorner J, 1988. Enzymes required for yeast prohormone processing. *Annual Review of Physiology* **50**: 345–362.

Garcia-Lopez A, Monferrer L, Garcia-Alcover I, Vicente-Crespo M, Alvarez-Abril MC, Artero RD, 2008. Genetic and Chemical Modifiers of a CUG Toxicity Model in Drosophila. *PLoS ONE* **13**: e1595 (online).

Gravekamp C, Rosner B, Madoff LC, 1998. Deletion of repeats in the alpha C protein enhances the pathogenicity of group B streptococci in immune mice. *Infection and Immunity* **66**: 4347–4354.

Guo B, Styles CA, Feng Q, Fink GR, 2000. A *Saccharomyces* gene family involved in invasive growth, cell–cell adhesion and mating. *Proceedings of the National Academy of Sciences of the United States of America* **97**: 12158–12163.

Halme A, Bumgarner S, Styles CA, Fink GR, 2004. Genetic and epigenetic regulation of the *FLO* gene family generates cell-surface variation in yeast. *Cell* **116**: 405–415.

Harrison P, Kumar A, Lan N, Echols N, Snyder M, Gerstein M, 2002. A small reservoir of disabled ORFs in the yeast genome and its implications for the dynamics of proteome evolution. *Journal of Molecular Biology* **316**: 409–419.

Hoyer LL, 2001. The ALS gene family of *Candida albicans*. *Trends in Microbiology* **9**: 176–180.

Hoyer LL, Green CB, Oh SH, Zhao X, 2007. Discovering the secrets of the *Candida albicans* agglutinin-like sequence (ALS) gene family – a sticky pursuit. *Medical Mycology* **20**: 1–15.

Jordan P, Snyder LA, Saunders NJ, 2003. Diversity in coding tandem repeats in related *Neisseria* spp. *BioMed Central Microbiology* **3**: 23.

Kamai Y, Kubota M, Kamai Y, Hosokawa T, Fukuoka T, Filler SG, 2002. Contribution of *Candida albicans* ALS1 to the pathogenesis of experimental oropharyngeal candidiasis. *Infection and Immunity* **9**: 5256–5258.

Kantety RV, La Rota M, Matthews DE, Sorrells ME, 2002. Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. *Plant Molecular Biology* **48**: 501–510.

Karaoglu H, Lee CM, Meyer W, 2005. Survey of simple sequence repeats in completed fungal genomes. *Molecular Biology and Evolution* **22**: 639–649.

Katti MV, Ranjekar PK, Gupta VS, 2001. Differential distribution of simple sequence repeats in eukaryotic genome sequences. *Molecular Biology and Evolution* **18**: 1161–1167.

Kenneson A, Zhang F, Hagedorn CH, Warren ST, 2001. Reduced FMRP and increased FMR1 transcription is proportionally associated with CGG repeat number in intermediate-length and premutation carriers. *Human Molecular Genetics* **10**: 1449–1454.

Kolpakov R, Bana G, Kucherov G, 2003. mreps: Efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Research* **31**: 3672–3678.

Kulkarni RD, Kelkar HS, Dean RA, 2003. An eight-cysteine containing CFEM domain unique to a group of fungal membrane proteins. *Trends in Biochemical Science* **28**: 118–121.

Kunkel TA, 1993. Slippery DNA and diseases. *Nature* **365**: 207–208.

Legendre M, Pochet N, Pak T, Verstrepen KJ, 2007. Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome Research* **17**: 1787–1796.

Levdansky E, Romano J, Shadkchan Y, Sharon H, Verstrepen KJ, Fink GR, Osherov N, 2007. Coding tandem repeats generate diversity in *Aspergillus fumigatus* genes. *Eukaryotic Cell* **6**: 1380–1391.

Li YC, Korol AB, Fahima T, Nevo E, 2004. Microsatellites within genes: structure, function and evolution. *Molecular Biology and Evolution* **21**: 991–1007.

Lott TJ, Holloway BP, Logan DA, Fundyga R, Arnold J, 1999. Towards understanding the evolution of the human commensal yeast *Candida albicans*. *Microbiology* **145**: 1137–1143.

Loza L, Fu Y, Ibrahim AS, Sheppard DC, Filler SG, Edwards Jr JE, 2004. Functional analysis of the *Candida albicans ALS1* gene product. *Yeast* **2**: 473–482.

Mankodi A, Takahashi MP, Jiang H, Beck CL, Bowers WJ, Moxley RT, Cannon SC, Thornton CA, 2002. Expanded CUG repeats trigger aberrant splicing of ClC-1 chloride channel pre-mRNA and hyperexcitability of skeletal muscle in myotonic dystrophy. *Molecular Cell* **10**: 35–44.

Marcotte EM, Pellegrini M, Yeates TO, Eisenberg D, 1999. A census of protein repeats. *Journal of Molecular Biology* **293**: 151–160.

Meloni R, Albanese V, Ravassard P, Treilhou F, Mallet J, 1998. A tetranucleotide polymorphic microsatellite, located in the first intron of the tyrosine hydroxylase gene, acts as a transcription regulatory element in vitro. *Human Molecular Genetics* **7**: 423–428.

Metzgar D, Thomas E, Davis C, Field D, Wills C, 2001. The microsatellites of *Escherichia coli*: rapidly evolving repetitive DNAs in a non-pathogenic prokaryote. *Molecular Microbiology* **39**: 183–190.

Mirkin SM, 2007. Expandable DNA repeats and human disease. *Nature* **447**: 932–940.

Please cite this article in press as: Levdansky E et al., Coding fungal tandem repeats as generators of fungal diversity, Fungal Biology Reviews (2008), doi:10.1016/j.fbr.2008.08.001

Morgante M, Hanafey M, Powell W, 2002. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nature Genetics* **30**: 194–200.

Norberg P, Olofsson S, Tarp MA, Clausen H, Bergström T, Liljeqvist JA, 2007. Glycoprotein I of herpes simplex virus type 1 contains a unique polymorphic tandem-repeated mucin region. *Journal of General Virology* **88**: 1683–1688.

Oh SH, Cheng G, Nuessen JA, Jajko R, Yeater KM, Zhao X, Pujol C, Soll DR, Hoyer LL, 2005. Functional specificity of *Candida albicans* Als3p proteins and clade specificity of *ALS3* alleles discriminated by the number of copies of the tandem repeat sequence in the central domain. *Microbiology* **151**: 673–681.

Panwar SL, Legrand M, Dignard D, Whiteway M, Magee PT, 2003. *MFalpha1,* the gene encoding the alpha mating pheromone of *Candida albicans*. *Eukaryotic Cell* **2**: 1350–1360.

Pâques F, Richard GF, Haber JE, 2001. Expansions and contractions in 36-bp minisatellites by gene conversion in yeast. *Genetics* **158**: 155–166.

Patti JM, Allen BL, McGavin MJ, Hook M, 1994. MSCRAMM mediated adherence of microorganisms to host tissues. *Annual Review of Microbiology* **48**: 585–617.

Pearson CE, Nichol Edamura K, Cleary JD, 2005. Repeat instability: mechanisms of dynamic mutations. *Nature Reviews Genetics* **6**: 729–742.

Podbielski A, Krebs B, Kaufhold A, 1994. Genetic variability of the *emm*-related gene of the large *vir* regulon of group A streptococci: potential intra- and intergenomic recombination events. *Molecular & General Genetics* **243**: 691–698.

Rauceo JM, De Armond R, Otoo H, Kahn PC, Klotz SA, Gaur NK, Lipke PN, 2006. Threonine-rich repeats increase fibronectin binding in the *Candida albicans* adhesin Als5p. *Eukaryotic Cell* **5**: 1664–1673.

Rice P, Longden I, Bleasby A, 2000. EMBOSS: the European molecular biology open software suite. *Trends in Genetics* **16**: 276–277.

Richard GF, Dujon B, Haber JE, 1999. Double-strand break repair can lead to high frequencies of deletions within short CAG/CTG trinucleotide repeats. *Molecular & General Genetics* **261**: 871–882.

Richard GF, Dujon B, 2006. Molecular evolution of minisatellites in hemiascomycetous yeasts. *Molecular Biology and Evolution* **23**: 189–202.

Ruiz-Herrera J, Elorza MV, Valentín E, Sentandreu R, 2006. Molecular organization of the cell wall of *Candida albicans* and its relation to pathogenicity. *FEMS Yeast Research* **6**: 14–29.

Sertil O, Vemula A, Salmon SL, Morse RH, Lowry CV, 2007. Direct role for the Rpd3 complex in transcriptional induction of the anaerobic DAN/TIR genes in yeast. *Molecular Cell Biology* **27**: 2037–2047.

Sheppard DC, Yeaman MR, Welch WH, Phan QT, Fu Y, Ibrahim AS, Filler SG, Zhang M, Waring AJ, Edwards JE, 2004. Functional and structural diversity in the Als protein family of *Candida albicans*. *Journal of Biological Chemistry* **1279**: 30480–30489.

Sirand-Pugnet P, Durosay P, Brody E, Marie J, 1995. An intronic (A/U)GGG repeat enhances the splicing of an alternative intron of the chickrn b-tropomyosin pre-mRNA. *Nucleic Acids Research* **23**: 3501–3507.

Spatz SJ, Silva RF, 2007. Sequence determination of variable regions within the genomes of gallid herpesvirus-2 pathotypes. *Archives of Virology* **152**: 1665–1678.

Staab JF, Bradway SD, Fidel PL, Sundstrom P, 1999. Adhesive and mammalian transglutaminase substrate properties of *Candida albicans* Hwp1. *Science* **283**: 1535–1538.

Staab JF, Bahn YS, Tai CH, Cook PF, Sundstrom P, 2004. Expression of transglutaminase substrate activity on *Candida albicans* germ tubes through a coiled, disulfide-bonded N-terminal domain of Hwp1 requires C-terminal glycosylphosphatidylinositol modification. *Journal of Biological Chemistry* **279**: 40737–40747.

Strand M, Prolla TA, Liskay RM, Petes TD, 1993. Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. *Nature* **365**: 274–276.

Tatebayashi K, Tanaka K, Yang HY, Yamamoto K, Matsushita Y, Tomida T, Imai M, Saito H, 2007. Transmembrane mucins Hkr1 and Msb2 are putative osmosensors in the SHO1 branch of yeast HOG pathway. *EMBO Journal* **26**: 3521–3533.

Toth G, Gáspári Z, Jurka J, 2000. Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Research* **10**: 967–981.

Trivedi S, 2006. Comparison of simple sequence repeats in 19 Archaea. *Genetic and Molecular Research* **5**: 741–772.

Verstrepen KJ, Reynolds TB, Fink GR, 2004. Origins of variation in the fungal cell surface. *Nature Reviews Microbiology* **2**: 533–540.

Verstrepen KJ, Jansen A, Lewitter F, Fink GR, 2005. Intragenic tandem repeats generate functional variability. *Nature Genetics* **37**: 986–990.

Verstrepen KJ, Klis FM, 2006. Flocculation, adhesion and biofilm formation in yeasts. *Molecular Microbiology* **60**: 5–15.

Waltman WD, McDaniel LS, Gray BM, Briles DE, 1990. Variation in the molecular weight of *PspA* (pneumococcal surface protein A) among *Streptococcus pneumoniae*. *Microbial Pathogenesis* **8**: 61–69.

Yamada M, Hayatsu N, Matsuura A, Ishikawa F, 1998. Y′-Help1, a DNA helicase encoded by the yeast subtelomeric Y′ element, is induced in survivors defective for telomerase. *Journal of Biological Chemistry* **273**: 33360–33366.

Zhao X, Oh SH, Jajko R, Diekema DJ, Pfaller MA, Pujol C, Soll DR, Hoyer LL, 2007. Analysis of *ALS5* and *ALS6* allelic variability in a geographically diverse collection of *Candida albicans* isolates. *Fungal Genetics and Biology* **44**: 1298–1309.

Zhang N, Harrex AL, Holland BR, Fenton LE, Cannon RD, Schmid J, 2003. Sixty alleles of the *ALS7* open reading frame in *Candida albicans*: *ALS7* is a hypermutable contingency locus. *Genome Research* **13**: 2005–2017.

Please cite this article in press as: Levdansky E et al., Coding fungal tandem repeats as generators of fungal diversity, Fungal Biology Reviews (2008), doi:10.1016/j.fbr.2008.08.001