Comparative and Functional Characterization of Intragenic Tandem Repeats in 10 *Aspergillus* Genomes

John G. Gibbons and Antonis Rokas

Department of Biological Sciences, Vanderbilt University, Nashville

Intragenic tandem repeats (ITRs) are consecutive repeats of three or more nucleotides found in coding regions. ITRs are the underlying cause of several human genetic diseases and have been associated with phenotypic variation, including pathogenesis, in several clades of the tree of life. We have examined the evolution and functional role of ITRs in 10 genomes spanning the fungal genus *Aspergillus*, a clade of relevance to medicine, agriculture, and industry. We identified several hundred ITRs in each of the species examined. ITR content varied extensively between species, with an average 79% of ITRs unique to a given species. For the fraction of conserved ITR regions, sequence comparisons within species and between close relatives revealed that they were highly variable. ITR-containing proteins were evolutionarily less conserved, compositionally distinct, and overrepresented for domains associated with cell-surface localization and function relative to the rest of the proteome. Furthermore, ITRs were preferentially found in proteins involved in transcription, cellular communication, and cell-type differentiation but were underrepresented in proteins involved in metabolism and energy. Importantly, although ITRs were evolutionarily labile, their functional associations appeared. To be remarkably conserved across eukaryotes. Fungal ITRs likely participate in a variety of developmental processes and cell-surface-associated functions, suggesting that their contribution to fungal lifestyle and evolution may be more general than previously assumed.

Introduction

A recurrent theme in eukaryotic genomes is the presence of repetitive DNA (Toth et al. 2000; Katti et al. 2001; Li et al. 2004; Hancock and Simon 2005; Karaoglu et al. 2005; Thomas 2005; Kashi and King 2006), notably highlighted by the human genome which is composed of approximately 35% repetitive elements (Venter et al. 2001). One such class of repetitive DNA is intragenic tandem repeats (ITRs) that comprise of three or more nucleotides repeated in tandem, found in protein-coding regions. ITRs include short sequence repeats, such as microsatellites (Ellegren 2004; Li et al. 2004), as well as longer repeats that span tens to hundreds of nucleotides, such as the ~100-nt repeat identified in the *FLO1*, *FLO5*, and *FLO9* genes in *Saccharomyces cerevisiae* (Verstrepen et al. 2005).

ITR sequences are typically mutationally unstable and prone to local expansion and contraction either via unequal recombination events or slip-strand mispairing (Levinson and Gutman 1987; Schlotterer and Tautz 1994; Bichara et al. 2006). This inherent mutational instability frequently results in elevated mutation rates relative to the rest of the genome, especially for short ITRs (Dieringer and Schlotterer 2003; Kashi and King 2006; Moxon et al. 2006). For example, in humans, microsatellite mutation rates are as frequent as 10^{-3} to 10^{-4} per locus per generation (Weber and Wong 1993) compared with a rate of 10^{-8} per generation for single-nucleotide substitutions (Drake et al. 1998). Not surprisingly, variation in ITRs is associated with several human hereditary disorders (Sherman et al. 1985; Sutherland and Richards 1995; Pearson et al. 2005; Mirkin 2007). For example, Huntington's disease, an inherited autosomal dominant neurodegenerative disorder, is caused by an expansion of a CAG repeat in exon 1 of the *IT-15* gene, which results in a glutamine expansion in the protein product (Schilling et al. 1995). However, variation in ITRs has also been associated with natural variation, as, for example, in circadian clock adjustment in the fruit fly (Sawyer et al. 1997) and skeletal morphology in domesticated dog breeds (Fondon and Garner 2004).

Genotypic (Balajee et al. 2007; Levdansky et al. 2007) and phenotypic (Verstrepen et al. 2004, 2005; Fidalgo et al. 2006; Michael et al. 2007) variation associated with ITRs has also been observed in fungi. For example, ITR variation in the FLO1 protein of *S. cerevisiae* is positively associated with an increase in cell–cell adhesion (Verstrepen et al. 2005), whereas ITR variation in the FLO11 protein of the same species contributes to the formation of self-supporting biofilm (Fidalgo et al. 2006). ITRs have also been identified in members of the agglutinin-like sequence (ALS) protein family that are thought to play a similar role in mediating adhesion to other cells and substrates in *Candida albicans* and *Candida. glabrata* (Verstrepen et al. 2004; Oh et al. 2005). These findings have led to the hypothesis that fungal ITRs may be implicated in the generation of variation in cell-surface proteins, molecules with active roles in the colonization of host tissue, and evasion of its immune system (Jordan et al. 2003; Verstrepen et al. 2004; Levdansky et al. 2007).

We were particularly interested in studying the evolution and function of ITRs in the filamentous ascomycete *Aspergillus*, a genus with a large societal impact, both beneficial and detrimental. For example, the species *Aspergillus oryzae* and *Aspergillus niger* are commercially exploited for a variety of industrial purposes (Machida et al. 2005; Pel et al. 2007). In contrast, *Aspergillus flavus* is a producer of the carcinogenic compound aflatoxin as well as an agricultural pathogen (corn, cotton, and peanuts) that causes annual losses totaling hundreds of millions of dollars (Yu et al. 2004, 2005). *Aspergillus fumigatus* and *Aspergillus terreus* are potentially lethal opportunistic pathogens and the leading causes of invasive pulmonary aspergillosis (Patterson et al. 2000; Yu et al. 2004). It is perhaps a testament to the relevance of this genus to human affairs that 10 draft genomes from eight species are already available (Galagan et al. 2005; Machida et al. 2005; Nierman et al. 2006; Pel et al. 2007; Fedorova et al. 2008).

Key words: fungi, *Aspergillus*, intragenic tandem repeats, polymorphism, comparative functional genomics.

E-mail: antonis.rokas@vanderbilt.edu.

Mol. Biol. Evol. 26(3):591–602. 2009
doi:10.1093/molbev/msn277
Advance Access publication December 4, 2008

To understand the comparative biology of ITRs in this important fungal genus, we first identified and calculated the frequency and distribution of ITRs across the genomes of eight *Aspergillus* species, and evaluated their relative placement in proteins. We next assessed the evolutionary conservation of ITRs and ITR-containing proteins by comparing the relative proportion of orthologous ITRs and ITR-containing proteins in the entire, background, and ITR proteomes. We then examined ITR variation levels by analyzing the observed differences within and between closely related species. To gain an insight into the functional biology of *Aspergillus* ITR-containing proteins, we first determined whether ITR-containing proteins were compositionally distinct from the background proteome. We next examined ITR-containing proteins for the presence of a variety of cell-surface-associated protein domains. Finally, we evaluated whether ITR-containing proteins were associated with specific functional categories.

Materials and Methods

Genome Sequences

The coding sequences analyzed in this study were downloaded from the *Aspergillus* comparative site at the Broad Institute (http://www.broad.mit.edu/annotation/genome/aspergillus_group/MultiHome.html). They are also available in public databases under the accession numbers *A. flavus* NRRL 3357 (Genbank: AAIH01000000); *A. oryzae* RIB 40 (DDJB: AP007150–AP007177); *A. terreus* NIH 2624 (Genbank: AAJN01000000); *A. niger* CBS 513.88 (EMBL: AM270980–AM270998); *A. niger* ATCC1015 (http://genome.jgi-psf.org/Aspni1/Aspni1.home.html); *Neosartorya fischeri* NRRL 181 (Genbank: AAKE03000000) (*N. fischeri* is the taxonomic name assigned to the sexual generation of *Aspergillus fischerianus*); *A. fumigatus* CEA10 (Genbank: ABDB01000000); *A. fumigatus* Af293 (Genbank: AAHF01000000); *Aspergillus clavatus* NRRL 1 (Genbank: AAKD00000000); and *Aspergillus nidulans* FGSC A4 (Genbank: AACD00000000).

Genomic Identification of ITRs

The EMBOSS ETANDEM software was used to identify short (3–39 nt) and long (40–500 nt) ITRs in each of the 10 analyzed transcriptomes (Rice et al. 2000). ITRs with a consensus sequence conservation $\geq$85% and an absolute sequence length of at least 24 nt (i.e., eight copies of a trinucleotide repeat; six copies of a tetranucleotide repeat; two copies of any large repeat unit) were considered significant (supplementary data file 1, Supplementary Material online). The 24-nt cutoff is the minimum identification criterion for trinucleotide repeats in our ITR detection software (ETANDEM), which allowed as inclusive a characterization of ITRs from the genus *Aspergillus* as possible.

Conservation of ITR-Containing Genes

Orthologs were identified using the reciprocal-best-Blast-hit approach for each pairwise species comparison with a cutoff E-value of 1E−06 (Rokas et al. 2007; Moreno-Hagelsieb and Latimer 2008). Average conservation was calculated by dividing the total number of orthologs shared by a species pair by the total number of proteins of the species with the smaller proteome. We used this method to calculate the conservation of the entire proteome, background proteome and ITR-containing proteome. ITR conservation was calculated by identifying the number of shared ITRs between orthologs of each species pair and dividing it by the total number of ITRs in the orthologs.

ITR Variation within and between Species

Investigation of ITR variation was examined in two intraspecific and two interspecific comparisons: *A. fumigatus* strain Af293 versus CEA10, *A. niger* strain CBS 518.33 versus ATCC1015, *A. flavus* versus *A. oryzae* and *A. fumigatus* Af293 versus *N. fischeri*. In each case, the orthologs of all ITR-containing genes were compared. ITR orthologs were categorized as 1) Monomorphic, 2) Variable, 3) Ortholog no ITR, or 4) No ortholog. Because the current annotation of *A. niger* strain CBS 513.88 contains many instances of multiple stop codons per gene, only genes with a single-stop codon were used.

The effect of repeat unit copy number and longest pure tract on ITR variation rates was examined via Student's *t*-test (Sokal and Rohlf 1995). Data from the two interspecific and intraspecific comparisons, respectively, were pooled as the distributions did not significantly deviate from each other. The average copy number and average pure tract length of monomorphic and variable ITRs were independently assessed in trinucleotide and hexanucleotide repeats.

Amino Acid Composition

For each species, the absolute and relative frequency of each amino acid was calculated in the background and ITR-containing proteomes. The relative proportions of each amino acid were analyzed via Fisher's exact test (Sokal and Rohlf 1995). To limit overall experimentwise error rates due to multiple comparisons, we used a Bonferroni corrected P value = 0.0003125.

Hydropathy Index

For each species, an average hydropathy score was calculated for each protein by assigning each amino acid with a numerical value based on the Kyte and Doolittle Hydropathy Index, and dividing the sum by the total number of amino acids in the protein (Kyte and Doolittle 1982). For each species, the mean hydropathy scores of background and ITR proteomes were compared via Student's *t*-test (Sokal and Rohlf 1995). To limit experimentwise error rates due to multiple comparisons, we used a Bonferroni corrected P value = 0.00625.

The Relative Position of ITRs within Proteins

To test the hypothesis that ITRs were randomly distributed throughout the protein, we adapted the methods of Huntley et al. (Huntley and Clark 2007). Briefly, each

Table 1
General Characteristics and ITR Summary of the *Aspergilli*

	Aspergillus flavus	*Aspergillus oryzae*	*Aspergillus terreus*	*Aspergillus niger*	*Neosartorya fischeri*	*Aspergillus fumigatus*	*Aspergillus clavatus*	*Aspergillus nidulans*
Strain	NRRL 3357	RIB 40	NIH 2624	CBS 513.88	NRRL 181	Af 293	NRRL 1	FGSC A4
Number of genes	12,587	12,063	10,406	13,912	10,403	9,887	9,120	10,665
Total ITRs	235	204	172	345	222	222	313	215
ITR-containing genes	210	182	154	317	194	207	278	200
Genes containing multiple ITRs	20	17	14	22	24	15	28	12
Short ITRs	161	136	123	277	155	187	262	160
Long ITRs	74	68	49	68	67	35	51	55

protein was separated into three equal-sized segments, the N-terminal, midsegment and C-terminal (Huntley and Clark 2007). For each species, we calculated the midpoints of all ITRs and identified the protein segments the midpoints were located in. We generated the expected frequencies of ITR position by using the following equations: midsegment = $(L/3)/(L - l)$ and N-terminal and C-terminal = $((L/3) - (l/2))/(L - l)$, where L is the protein length and l is the total ITR length (Huntley and Clark 2007). For each species, we then averaged the probabilities of each ITR and multiplied this by the total number of ITRs. G-tests were used to assess whether observed frequencies deviated from expected (Sokal and Rohlf 1995).

Functional Annotation and Classification

The SignalP version 3.0 software was used to predict signal peptides using the hidden Markov model constructed with eukaryotic proteins (Bendtsen et al. 2004). Glycosylphosphatidylinisotol (GPI) anchors were predicted using the big-Pi predictor software (Eisenhaber et al. 2004). Transmembrane helices were predicted using the TMHMM version 2.0 software, with hits considered significant when more than 18 amino acids in the transmembrane helix were predicted (Krogh et al. 2001). The proportions of each predicted motif in each proteome were assessed via Fisher's exact test (Sokal and Rohlf 1995). To limit experimentwise error rates due to multiple comparisons, we used a Bonferroni corrected P value = 0.00625 for each motif. All statistical analyses were performed using the JMP software, version 5.0.1a (Frenkel and Blumenthal 2002). Putative functional domains were identified using the Pfam annotation of the eight *Aspergillus* proteomes (Finn et al. 2006), which were downloaded from the *Aspergillus* comparative site at the Broad Institute.

To test the hypothesis that ITR-containing proteins differed in specific functions, the major FunCat (Ruepp et al. 2004) categories for *A. oryzae*, *A. terreus*, *A. fumigatus*, and *A. nidulans* were retrieved from the MIPS PEDANT database (http://pedant.gsf.de/). For each category, the proportion of ITR-containing proteins and background proteins was assessed via Fisher's exact test (Sokal and Rohlf 1995).

In a whole genome expression profiling microarray analysis, Nierman et al. (2006) identified 458 genes in *A. fumigatus* that were differentially expressed at 30, 37, and 48 °C. The latter two temperatures represent ones that the species experiences in the human body and compost, respectively, and both are presumed to be more stressful than the 30 °C one. To test whether the ITR-containing genes of *A. fumigatus* were over or underrepresented in the gene set that is differentially expressed under temperature stress, we examined the proportion of ITR-containing genes relative to the background ones in the gene set via Fisher's exact test (Sokal and Rohlf 1995).

Results

Identification and Distribution of ITRs across *Aspergillus*

We identified short (3–39-nt) and long (>39-nt) ITRs in each of the 10 transcriptomes using the ETANDEM software (Rice et al. 2000). The total number of ITRs ranged from 172 to 345 per species (table 1, fig. 1, supplementary data file 1, Supplementary Material online). The number of ITR-containing genes ranged from 154 to 317 (table 1). Several ITR-containing genes are well characterized and are known to play key roles in fungal lifestyle and pathogenicity (fig. 2). In many instances, individual genes harbored multiple ITRs (average = 1.11 ITRs per gene). In all species analyzed, short ITRs were far more abundant than long ITRs (table 1, fig. 1). ITR abundance was not associated with genome size ($r^2 = 0.15$; $F = 1.02$; $n = 8$; $P = 0.35$).

We found that approximately 95% (1,835 of 1,928) of ITR repeat units had lengths divisible by three, a result likely reflecting selection toward repeat units that do not alter the reading frame (Metzgar et al. 2000). The remaining 5% (93 of 1,928) of ITRs consisted of repeat units that were not divisible by three (supplementary table 1, Supplementary Material online). Thirty-five such ITRs were identified as tetranucleotides, although 16 of them could be further collapsed to dinucleotides (supplementary table 1, Supplementary Material online). The remaining 58 ITRs exhibited repeat unit lengths between 5 and 104 nt and occurred less frequently.

Approximately 40% (35 of 93) of the ITRs, whose repeat units were not divisible by three, had an absolute sequence length divisible by three. Interestingly, only one of these 35 ITRs was variable (found in the *A. fumigatus* gene *afu5g06790*), with both repeat unit copy number variants remaining in frame. The remaining 60% (58 of 93) of the ITRs did not have an absolute sequence length divisible by three and was not variable. Because our study was restricted to coding regions without internal stop codons, we

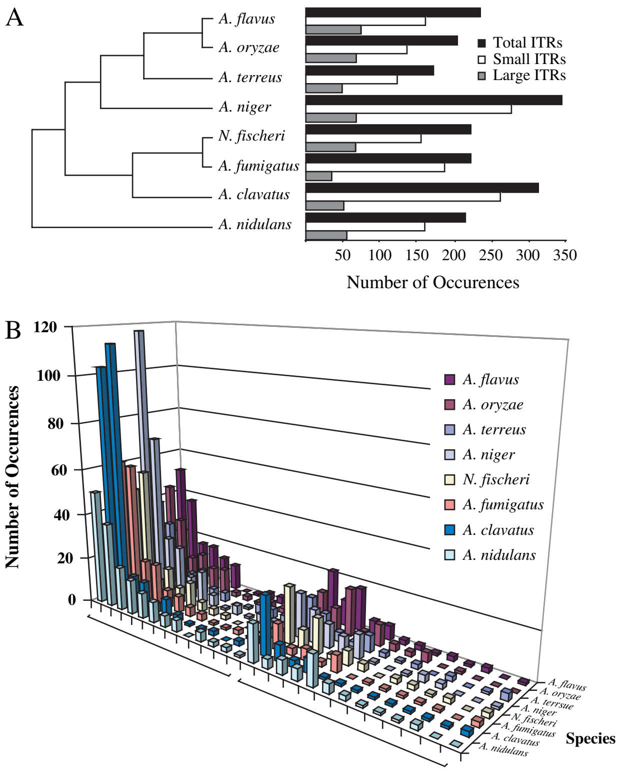

FIG. 1.—Distribution of ITRs across the *Aspergilli*. (*A*) Phylogenetic relationships (Rokas et al. 2007) and distribution of ITRs across the analyzed *Aspergillus* species. For each species, black bars represent total ITRs identified, white bars represent short ITRs (repeat unit 3–39 nt), and gray bars represent long ITRs (repeat unit ≥40 nt). (*B*) Distribution of short and long ITRs across the *Aspergilli*. ITRs were binned according to repeat unit size (*X* axis). The *Y* axis represents the total number of occurrences. Each species is indicated by a different color and label on the *Z* axis.

currently do not know whether any of these proteins exhibits phase variation (van der Woude and Baumler 2004; Moxon et al. 2006).

ITRs with short repeat unit sizes were more abundant than ITRs with long repeat unit sizes (fig. 1). The mean short ITR repeat unit copy number (average = 9.08) was significantly larger than the mean long ITR repeat unit copy number (average = 3.20) (t = −17.214; df = 1926; P = 7.0E−E−70). The largest repeat unit copy numbers identified were two 97-repeat trinucleotides, found in the non-orthologous ITR-containing genes of *A. niger* (*angc_134081920*) and *A. oryzae* (*ao090010000583*).

Identifying the Relative Position of ITRs within Proteins

The tendency of ITRs to occur toward the end of a protein has been frequently observed in eukaryotes (Alba and Guigo 2004; Siwach et al. 2006; Huntley and Clark 2007). To test whether this was the case in the *Aspergilli*, we first identified the relative midpoint location of each ITR in its respective protein and divided each protein into three equally sized regions (N-terminal, midsegment, and C-terminal), following a previously developed protocol (Huntley and Clark 2007). We then compared the observed frequencies with those expected by chance for each species.

We found that the distribution of ITRs across protein lengths was random for seven of the eight species. *Aspergillus niger* was the only species to significantly deviate from the null distribution (P = 0.0015). Similar results were obtained when short and long ITRs were independently tested for each species. However, when we pooled data from all species, we did find a significant deviation from the null distribution (df = 2, g = 25.657, P = 2.7E−06). We further observed a clear, but nonsignificant, trend of ITR overrepresentation in the C-terminal portion of ITR-containing proteins across all species other than *A. oryzae* and *Aspergillus clavatus* (supplementary data file 1, Supplementary Material online).

Conservation of ITRs and ITR-Containing Proteins

To evaluate the evolutionary conservation of ITR-containing proteins relative to the entire and background proteomes, we identified all orthologs for each pairwise species comparison (table 2*A*). We found that the ITR proteome was less conserved than the entire and background proteomes, with the latter two being essentially identical. Even more noticeable were the low levels of ITR conservation between species pairs, with an average 21% of ITRs found in a given species present in another one. For example, comparison of the sister species *A. flavus* and *A. oryzae* revealed that 84% of background proteins shared an ortholog, compared with only 75% of ITR-containing proteins (table 2*A*). Within these ITR-containing proteins, only 56% of ITRs were conserved (monomorphic or variable) (table 2*B*).

ITR Variation within and between Species

ITRs are typically highly variable, within and among species, and it is for this reason that they are frequently used in biomedical and population-based applications (Selkoe and Toonen 2006; Balajee et al. 2007). To assess ITR variation within and between *Aspergillus* species, we performed two intraspecific (*A. fumigatus*: strain Af293 vs. strain CEA10 and *A. niger*: strain CBS 513.88 vs. strain ATCC 1015) and two interspecific (*A. flavus* vs. *A. oryzae* and *A. fumigatus* vs. *N. fischeri*) comparisons. ITRs were categorized as either Monomorphic (no difference in repeat unit copy number), V (difference of at least one repeat unit), Ortholog no ITR (the ITR was absent in the ortholog), or No ortholog.

In the two intraspecific comparisons, as well as in the interspecific *A. flavus* versus *A. oryzae* comparison, approximately 25% of ITRs were in the Variable category, whereas in the *A. fumigatus* versus *N. fischeri* comparison, approximately 40% of ITRs were categorized as Variable (fig. 3*A*). The most abundant variations in repeat unit number were single repeat unit differences (~32%); however, repeat unit differences as large as 36 and 41 were identified (fig. 3*B*). The main difference between the intraspecific and interspecific comparisons was that ~32% of ITRs in the *A. fumigatus* versus *N. fischeri* case belonged to the Ortholog no ITR category, compared with 13% in the other between species comparison of *A. flavus* and *A. oryzae* and

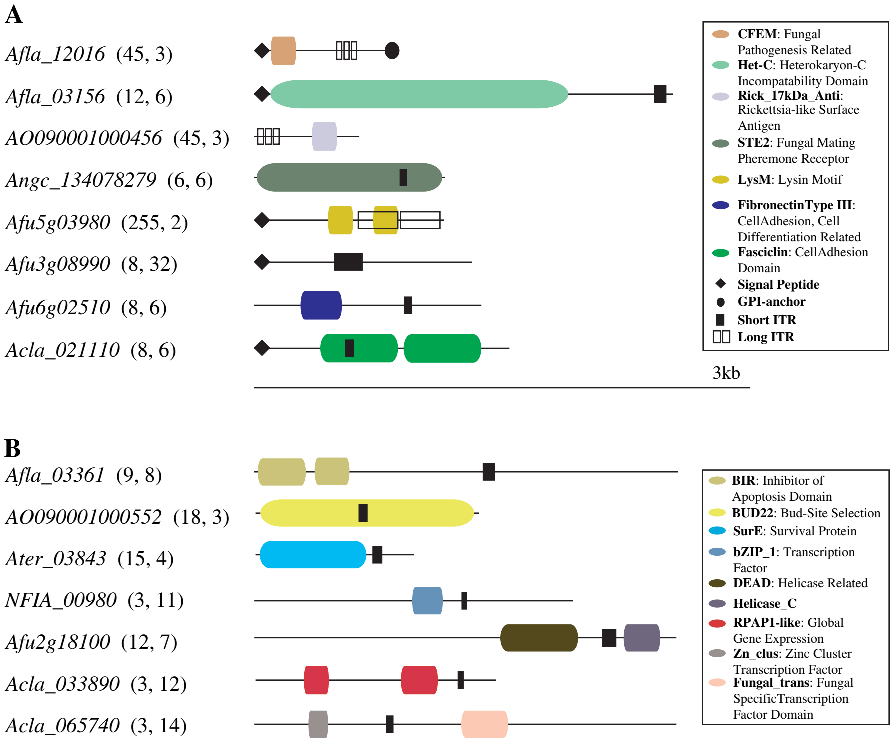
A
Afla_12016 (45, 3)
Afla_03156 (12, 6)
AO090001000456 (45, 3)
Angc_134078279 (6, 6)
Afu5g03980 (255, 2)
Afu3g08990 (8, 32)
Afu6g02510 (8, 6)
Acla_021110 (8, 6)
CFEM: Fungal Pathogenesis Related
Het-C: Heterokaryon-C Incompatability Domain
Rick_17kDa_Anti: Rickettsia-like Surface Antigen
STE2: Fungal Mating Pheremone Receptor
LysM: Lysin Motif
FibronectinType III: CellAdhesion, Cell Differentiation Related
Fasciclin: CellAdhesion Domain
Signal Peptide
GPI-anchor
Short ITR
Long ITR
3kb
B
Afla_03361 (9, 8)
AO090001000552 (18, 3)
Ater_03843 (15, 4)
NFIA_00980 (3, 11)
Afu2g18100 (12, 7)
Acla_033890 (3, 12)
Acla_065740 (3, 14)
BIR: Inhibitor of Apoptosis Domain
BUD22: Bud-Site Selection
SurE: Survival Protein
bZIP_1: Transcription Factor
DEAD: Helicase Related
Helicase_C
RPAP1-like: Global Gene Expression
Zn_clus: Zinc Cluster Transcription Factor
Fungal_trans: Fungal SpecificTranscription Factor Domain

FIG. 2.—Representative *Aspergillus* ITR-containing proteins. (*A*) Cell-surface-associated ITR-containing proteins. (*B*) ITR-containing proteins associated with transcriptional regulation. The numbers in parentheses after each gene identifier are the repeat unit size and copy number, respectively. Short repeats (repeat unit 3–39 nt) are represented by a single solid black rectangle. Long repeats (repeat unit ≥40 nt) are represented by a hollow rectangle for each repeat unit. Pfam functional domains are represented by colored ellipses. Note that this set of protein examples includes several well-characterized genes. For example, *Afla_03156* is a member of the HET gene family, whereas *Angc_134078279* is a STE2 (mating pheromone receptor) homolog.

8% and 20% in the two within species comparisons (*A. fumigatus* and *A. niger*, respectively) (fig. 3*A*). Additionally, only 9% of ITRs in the *A. fumigatus* versus *N. fischeri* case were in the Monomorphic category, compared with 35% in the other between species comparison, and 52% and 62% in the within species (*A. niger* and *A. fumigatus*, respectively) comparisons.

Interestingly, the distribution of ITRs in the Monomorphic and Ortholog no ITR categories from the interspecific comparison between *A. flavus* and *A. oryzae* was much more similar to the within-species cases than to the interspecific comparison between *A. fumigatus* and *N. fischeri* (fig. 3*A* and *B*). In agreement with several lines of genetic, molecular, and genomic data, these observed patterns also suggest that, despite their distinct species status, *A. oryzae* is a domesticated ecotype of *A. flavus* (Kurtzman et al. 1986; Geiser et al. 1998; Kumeda and Asao 2001; Montiel et al. 2003; Rokas et al. 2007).

Previous research has indicated a positive association between repeat unit copy number and sequence instability, thereby increasing polymorphism rates (Brohede and Ellegren 1999; Lai et al. 2003; Shinde et al. 2003). Therefore, we examined the relationship between repeat unit copy number and levels of trinucleotide and hexanucleotide ITR variation, using data from the two interspecific cases. We found no significant difference in average monomorphic and variable repeat unit copy number in both trinucleotide and hexanucleotide repeats ($t = -0.899$, $N = 142$, df = 141, $P =$ 0.37 and $t = 0.241$, $N = 101$, df = 100, $P = 0.81$, respectively). However, previous work has suggested that the longest uninterrupted tract of repeat units, or "pure tract," may actually be a more accurate predictor of polymorphism (Lai and Sun 2003; Butland et al. 2007; Anmarkrud et al. 2008). To test this hypothesis, we compared the average longest pure tract sequence between monomorphic and variable ITRs in the two intraspecific cases. We found that size-variable ITRs had significantly longer average pure tracts in trinucleotide repeat units (monomorphic = 4.59; variable = 10.43) but not in hexanucleotide repeat units (monomorphic = 2.80; variable = 3.71) compared with monomorphic ITRs ($t = 4.22$, $N = 94$, df = 93, $P = 5.62\text{E}{-}05$ and $t = 1.67$, $N = 56$, df = 55, $P = 0.10$, respectively).

Amino Acid Composition of ITR-Containing Proteins

To test whether ITR-containing proteins were compositionally distinct from background proteins, we compared amino acid frequencies and mean hydropathy, a measure of a protein's interaction with water, across proteomes. In both analyses, an underlying difference between ITR-containing proteins and background proteins was apparent. Of the 20 amino acids analyzed, only histidine showed no significant difference between the ITR and background proteomes in all eight species (table 3, supplementary table 2, Supplementary Material online). Twelve of the 20y amino acids

Table 2
Evolutionary Conservation of ITRs and ITR-Containing Proteins

	Aspergillus flavus	*Aspergillus oryzae*	*Aspergillus terreus*	*Aspergillus niger*	*Neosartorya fischeri*	*Aspergillus fumigatus*	*Aspergillus clavatus*	*Aspergillus nidulans*	Proteome
A. Average entire, background and ITR-containing proteome conservation									
A. flavus	X								Entire Background ITR-containing
A. oryzae	84%	X							E
	84%								B
	75%								I
A. terreus	73%	69%							E
	73%	70%							B
	67%	60%	X						I
A. niger	63%	62%	69%						E
	63%	63%	70%						B
	58%	55%	60%	X					I
N. fischeri	75%	71%	71%	72%					E
	75%	61%	71%	72%					B
	72%	61%	62%	64%	X				I
A. fumigatus	74%	72%	71%	72%	86%	X			E
	74%	72%	72%	72%	86%				B
	65%	58%	59%	65%	83%				I
A. clavatus	79%	76%	76%	78%	86%	82%	X		E
	80%	77%	77%	78%	87%	83%			B
	57%	54%	62%	76%	73%	70%			I
A. nidulans	70%	67%	68%	68%	71%	72%	77%		E
	70%	67%	69%	68%	71%	72%	78%		B
	60%	57%	63%	58%	64%	62%	61%	X	I
	Proteome “Entire” or “E” includes all proteins.								
	Proteome “Background” or “B” includes proteins that do not contain an ITR.								
	Proteome “ITR-containing” or ITR includes proteins that contain at least one ITR.								
B. Average ITR conservation									
A. flavus	x								
A. oryzae	56%	x							
A. terreus	14%	20%	x						
A. niger	16%	22%	23%	X					
N. fischeri	19%	18%	20%	16%	x				
A. fumigatus	19%	20%	25%	16%	47%	x			
A. clavatus	15%	17%	19%	17%	20%	22%	x		
A. nidulans	12%	14%	20%	14%	13%	14%	15%	x	

(cysteine, glutamine, glutamic acid, isoleucine, leucine, methionine, phenylalanine, proline, serine, threonine, tryptophan, and tyrosine) were significantly differentially distributed across proteomes in all eight species (table 3, supplementary table 2, Supplementary Material online). Similarly, in each of the eight species, average hydropathy of the ITR and background proteomes differed significantly, with the ITR proteome always being less hydrophobic than the background proteome (table 4).

Functional Characterization of ITR-Containing Proteins

Previous studies have suggested that ITRs may play an important role in fungal pathogenesis by generating structural diversity in cell-surface associated proteins (Verstrepen et al. 2004, 2005; Levdansky et al. 2007). To test this hypothesis, we bioinformatically identified signal peptides, indicators of secreted proteins, transmembrane helices, hallmarks of transmembrane proteins, and GPI anchors, molecules attached to some cell-surface proteins, in the background and ITR proteomes for each of the eight species (supplementary data file 1, Supplementary Material online). Signal peptides and GPI anchors were significantly overrepresented in the ITR proteomes, whereas transmembrane helices showed no significant difference across proteomes (table 5). The overrepresentation of signal peptides and GPI anchors in ITR proteomes provides further support that ITRs may play an active role in cell-surface-associated proteins (Hamada et al. 1999; Levdansky et al. 2007).

We further investigated the functional role of ITRs by examining the occurrences of background and ITR-containing proteins in 4 of the 10 genomes (*A. oryzae*, *A. terreus*, *A. fumigatus*, and *A. nidulans*), according to the FunCat annotation scheme (Ruepp et al. 2004). Given the distributional similarities within species (fig. 4) and the small percentage of ITR-containing genes, data from the four species were pooled in order to have sufficient data points for reliable statistical analysis. ITR-containing genes were significantly underrepresented in the Metabolism ($P = 3e - 5$) and Energy ($P = 0.0015$) categories (fig. 4). In contrast, ITRs were significantly overrepresented in the Transcription ($P = 0.0007$), Cellular communication/Signal transduction mechanism ($P = 0.0073$) and Cell-type differentiation ($P = 0.0007$) categories (fig. 4).

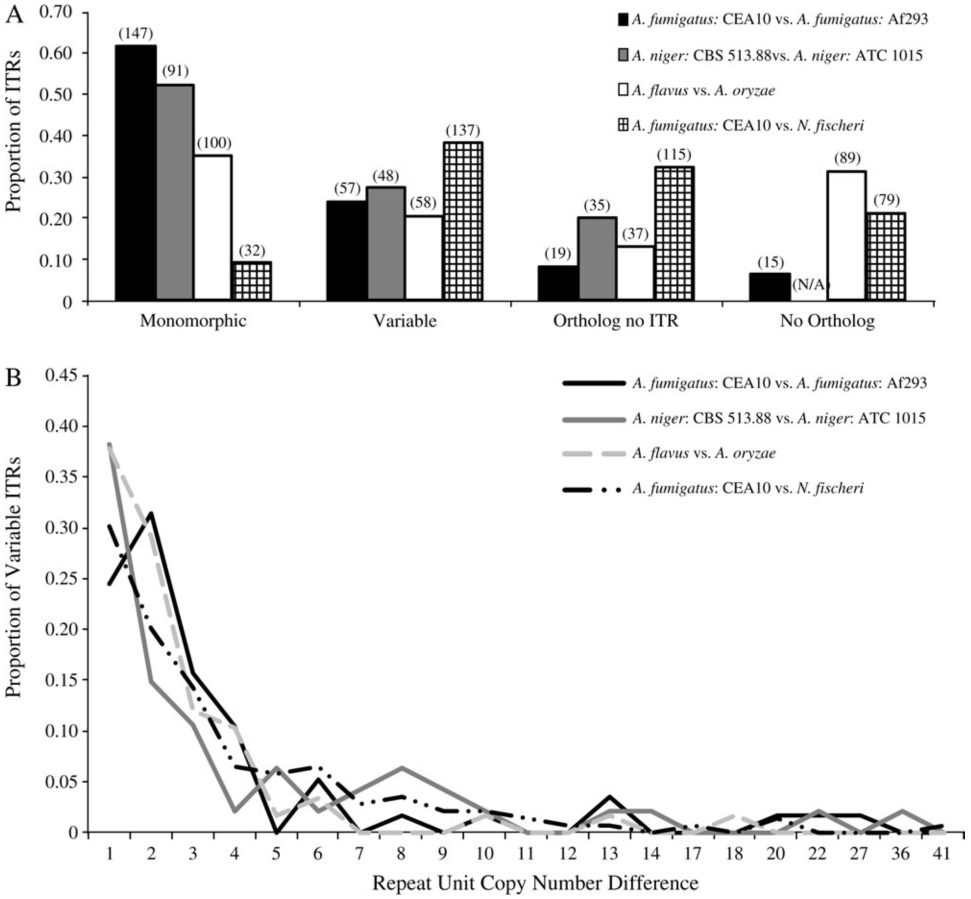

FIG. 3.—ITR variation within and between species: (*A*) Categorical proportions of ITR-containing genes in two within-species and two between-species comparisons. Black bars represent *Aspergillus fumigatus* strain CEA10 versus *fumigatus* strain Af293, dark gray bars represent *Aspergillus niger* strain 513.88 versus *A. niger* strain ATCC1015, white bars represent *Aspergillus flavus* versus *Aspergillus oryzae*, and the white mesh bars represent *Aspergillus fumigatus* strain CEA10 versus *Neosartorya fischeri*. Numbers in parentheses above bars correspond to the total number of occurrences. The *Y* axis is the proportion of total ITRs. ITRs were grouped into four categories indicated on the *X* axis. Monomorphic ITRS showed identical repeat unit copy number, whereas Variable ITRs showed a difference of at least one repeat unit copy number. ITRs in the Ortholog no ITR category consisted of orthologs in which no ITR was present only in one of the two taxa compared. Due to the poor annotation quality of the *A. niger* strain ATCC1015, only orthologs with identified ITRs were used in this analysis. ITR-containing genes in one taxon that did not have an identifiable ortholog in the other taxon were placed in the No ortholog category. (*B*) ITR repeat unit copy number variation. The *X* axis is the difference in repeat unit copy number in variable ITRs. The *Y* axis represents the proportion of total variable ITRs. The solid black line represents the *A. fumigatus* strain CEA10 versus *A. fumigatus* strain Af293 comparison, the dashed dark gray line represents the *A. niger* strain 513.88 versus *A. niger* strain ATCC1015 comparison, the dashed light gray line represents the *A. flavus* versus *Aspergillus oryzae* comparison, and the dashed black line represents the *A. fumigatus* strain CEA10 versus *N. fischeri* comparison. Note that proportion values in the *A. niger* intraspecific comparison are slightly inflated due to the lack of data for the No ortholog category.

Interestingly, studies in *Escherichia coli* and *S. cerevisiae* have identified an overrepresentation of microsatellites in stress response genes (Rocha et al. 2002; Bowen et al. 2005). However, there was no association between ITR-containing proteins and the Biogenesis of cell-wall components ($P = 0.66$) and Stress response ($P = 0.53$) categories. Furthermore, examination of the presence of ITR-containing genes in a previously identified *A. fumigatus* gene set that showed differential expression under temperature-induced stress (Nierman et al. 2006) also did not reveal any association ($P = 0.35$).

Discussion

ITRs are frequently associated with genetic disease and pathogenesis (Sherman et al. 1985; Sutherland and Richards 1995; Fondon and Garner 2004; Pearson et al. 2005; Verstrepen et al. 2005; Mirkin 2007), but also with adaptation to changing environments and phenotypic evolution (Verstrepen et al. 2004, 2005; Oh et al. 2005; Fidalgo et al. 2006; Michael et al. 2007). Nearly 2,000 ITRs are distributed throughout the genomes of the eight *Aspergillus* examined in this study. These ITR regions are highly variable, and the proteins containing them are less conserved and compositionally distinct relative to the rest of the proteome. ITR-containing proteins also appear to be functionally distinct. They are more likely to contain signal peptides and GPI anchors, motifs strongly suggestive of functional involvement on or around the cell surface. Furthermore, ITR-containing proteins are preferentially associated with certain cellular processes (transcription, cellular communication, and cell-type differentiation) and dissociated from

Table 3
Amino Acid Composition of ITR-Containing Proteome

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	Y	W	V
Aspergillus flavus	×	×	×	+	—	+	+	×	×	—	—	+	—	—	+	+	+	—	—	—
Aspergillus oryzae	×	×	×	×	—	+	+	+	×	—	—	+	—	—	+	+	+	—	—	—
Aspergillus terreus	+	×	×	+	—	+	+	×	×	—	—	—	—	—	+	+	+	—	—	—
Aspergillus niger	×	×	×	×	—	+	+	×	×	—	—	×	—	—	+	+	+	—	—	—
Neosartorya fischeri	+	×	—	+	—	+	+	×	×	—	—	×	—	—	+	+	+	—	—	—
Aspergillus fumigatus	×	×	×	×	—	+	+	×	×	—	—	×	—	—	+	+	+	—	—	—
Aspergillus clavatus	+	—	—	×	—	+	+	+	×	—	—	—	—	—	+	+	+	—	—	×
Aspergillus nidulans	+	×	—	×	—	+	+	×	×	—	—	×	—	—	+	+	+	—	—	—

+ Amino acids that are overrepresented in the ITR-containing proteomes (Bonferroni corrected $P = 0.0003125$).

— Amino acids that are underrepresented in the ITR-containing proteomes (Bonferroni corrected $P = 0.0003125$).

× Amino acids that do not differ in composition between ITR-containing and background proteomes.

others (metabolism and energy). These results bear on our understanding of the comparative and functional biology of eukaryotic ITRs, as well as the realization of the fungal lifestyle.

The availability of several comparative studies allows us to identify several general features of ITR-containing proteins. ITR abundance in eukaryotic proteomes is not correlated with genome size (Karaoglu et al. 2005; Huntley and Clark 2007), and ITR content varies extensively between species (figs. 1–3) (Huntley and Clark 2007). ITR regions and proteins are, on average, more hydrophilic than the rest (table 4) (Katti et al. 2001; Kim et al. 2008). Although it has been hypothesized that hydrophilic tandem repeat peptides in regions linking protein domains may produce more tolerated structural formations (Katti et al. 2000), a general explanation explaining the hydrophilic nature of ITR regions is still lacking. Finally, ITR regions are consistently highly variable, both within species as well as between close relatives (fig. 3) (Jordan et al. 2003; Bowen et al. 2005; O'Dushlaine et al. 2005; Levdansky et al. 2007; Kim et al. 2008). This variation, coupled with the increasing abundance of multiple genomes from a variety of clades, raise the possibility to efficiently identify and develop a suite of cladewide microsatellite and minisatellite markers that assess variation in taxa separated by hundreds of million years of evolution.

Perhaps more surprisingly, ITR-containing proteins across eukaryotes also share a number of functional features. Most strikingly, ITRs are consistently overrepresented in proteins associated with transcriptional, developmental, and signaling processes (fig. 4) (Katti et al. 2000; Young et al. 2000; Alba and Guigo 2004; O'Dushlaine et al. 2005; Huntley and Clark 2007), whereas they are underrepresented in proteins participating in metabolic and housekeeping processes (fig. 4) (Young et al. 2000; Huntley and Clark 2007). This conservation of functional association is observed across organisms separated by large evolutionary distances and persists despite differences across studies in the identification and functional classification of tandem repeats. This enrichment of tandem repeats has been attributed to their general involvement in modulating protein–protein interactions (Hancock and Simon 2005), where slight variations in tandem repeat regions can potentially generate slight variations in the structure of the protein–protein interaction network (King et al. 1997). Furthermore, the presence and functional role of ITRs in proteins participating in key processes, such as transcription, may also be the explanation as to why human repeat-based disorders are so common and devastating (Gatchel and Zoghbi 2005).

Studies in *S. cerevisiae* and *C. albicans* have shown that ITR variation can modulate the adhesiveness of several cell-surface proteins (Bowen et al. 2005; Oh et al. 2005; Verstrepen et al. 2005), a key trait for understanding fungal pathogenesis and virulence (Oh et al. 2005; Verstrepen et al. 2005). Several *Aspergillus* species are also capable of colonizing human tissue and causing potentially fatal infections (Patterson et al. 2000; Iversen et al. 2007). We too found that *Aspergillus* ITR-containing proteins were significantly enriched for cell-surface-associated motifs (fig. 2, table 5) (Levdansky et al. 2007), although no association between ITRs and cell-surface proteins could be established

Table 4
Hydropathy of ITR-Containing Proteome

	Aspergillus flavus		*Aspergillus oryzae*		*Aspergillus terreus*		*Aspergillus niger*		*Neosartorya fischeri*		*Aspergillus fumigatus*		*Aspergillus clavatus*		*Aspergillus nidulans*	
Proteome	Back	ITR	Back	ITR	Back	ITR	Back	ITR	Back	ITR	Back	ITR	Back	ITR	Back	ITR
Number of proteins	12,377	210	11,881	182	10,252	154	13,595	317	10,209	194	9,680	207	8,842	278	10,465	200
Average hydropathy	−0.3	−0.6	−0.3	−0.6	−0.3	−0.6	−0.6	−0.8	−0.3	−0.6	−0.3	−0.5	−0.3	−0.8	−0.3	−0.8
P value	*4.62E−20		*8.18E−20		*5.72E−15		*1.81E−10		*4.62E−15		*2.61E−05		*3.15E−05		*7.21E−36	

Proteome "Back" = proteins which do not possess an ITR.

Proteome "ITR" = proteins containing one or more ITRs.

*Significant at Bonferroni corrected *P* value = 0.00625.

Table 5
Protein Motif Comparison of ITR-Containing Genes and Background Genes

		Signal Peptide				Transmembrane Helix				GPI Anchor			
Species	Proteome	−	+	Prop	*P* Value	−	+	Prop	*P* Value	−	+	Prop	*P* Value
Aspergillus flavus	Back	10,413	1,964	0.189		11,162	1,214	0.109	1.000	12,360	16	0.001	
	ITR	155	55	0.355	**9E−05	191	20	0.105		206	5	0.024	**2E−05
Aspergillus oryzae	Back	10,817	1,064	0.098		10,693	1,188	0.111	0.170	11,864	17	0.001	
	ITR	149	33	0.221	**1E−03	158	24	0.152		179	3	0.017	**1E−v03
Aspergillus terreus	Back	8,629	1,623	0.188		9,205	1,047	0.114	0.504	10,230	22	0.002	
	ITR	121	33	0.273	0.0747	136	18	0.132		150	4	0.027	**5E−04
Aspergillus niger	Back	11,635	1,959	0.168		12,351	1,243	0.101	0.375	13,564	30	0.002	
	ITR	257	61	0.237	*0.0195	294	24	0.082		315	3	0.010	*0.0389
Neosartorya fischeri	Back	8,650	1,560	0.180		9,234	976	0.106	1.000	10,185	25	0.002	
	ITR	147	46	0.313	**2E−03	175	18	0.103		189	4	0.021	**2E−03
Aspergillus fumigatus	Back	8,303	1,377	0.166		8,729	951	0.109	0.195	9,672	8	0.001	
	ITR	160	47	0.294	**1E−03	181	26	0.144		197	10	0.051	**5E−13
Aspergillus clavatus	Back	7,569	1,273	0.168		7,983	859	0.108	0.258	8,819	23	0.003	
	ITR	207	71	0.343	**2E−06	245	33	0.135		270	8	0.030	**3E−06
Aspergillus nidulans	Back	8,894	1,571	0.177		9,450	1,015	0.107	0.717	10,437	28	0.003	
	ITR	159	41	0.258	*0.0362	179	21	0.117		199	1	0.005	0.432

Proteome Back = genes that do not contain the presence of an ITR.
Proteome ITR = genes containing one or more ITR.
"−" = absence of a predicted protein motif.
"+" = presence of predicted protein motif.
Prop = proportion of total proteome set.
*Statistically significant at *P* value = 0.05.
**Statistically significant at Bonferroni *P* value = 0.00625.

on the basis of the FunCat analysis (fig. 4). Importantly, the two main adhesion families in *Saccharomyces* and *Candida* (FLO and ALS, respectively) are not found in *Aspergillus* species (data not shown) (Levdansky et al. 2007). Although the composition of the *Aspergillus* cell surface is not well understood, it does include a small number of as yet uncharacterized ITR-containing proteins (Levdansky et al. 2007).

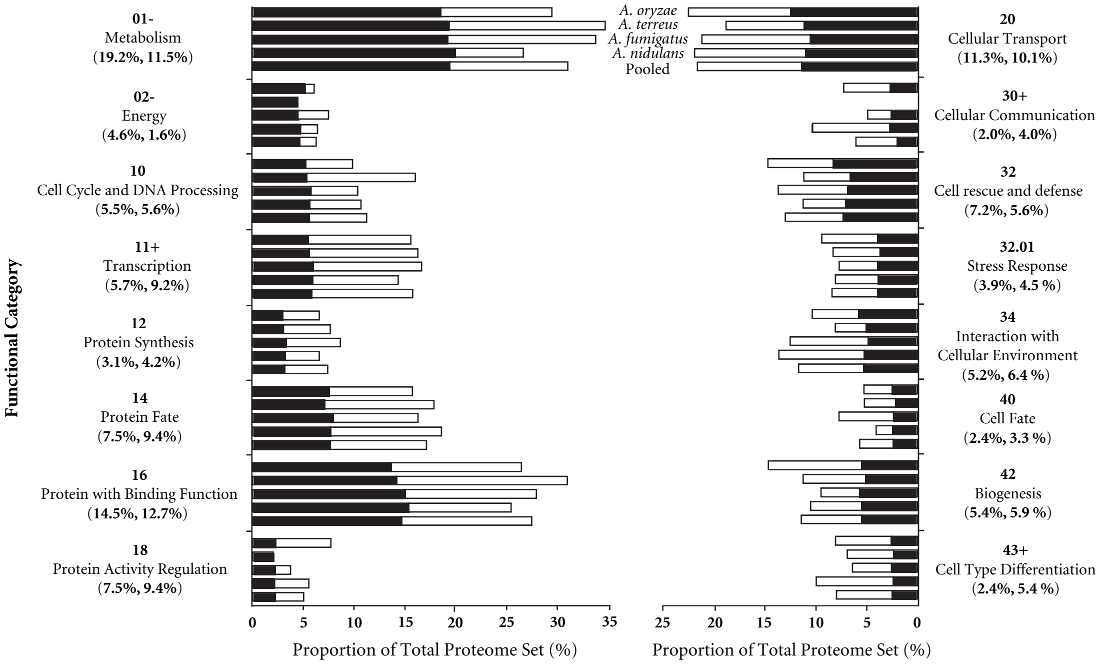

FIG. 4.—Functional classification of ITR and background proteomes according to the FunCat scheme. The ITR-containing proteome and background proteome proportions of 16 FunCat categories for *Aspergillus oryzae*, *Aspergillus terreus*, *Aspergillus fumigatus*, and *Aspergillus nidulans* and the pooled data of all species are reported. The FunCat category number is shown above each FunCat Category. A minus sign (−) next to the FunCat number represents a statistically significant underrepresentation of ITR-containing proteins, whereas a plus sign (+) represents a statistically significant overrepresentation of ITR-containing proteins. The percentages displayed under each FunCat category are the pooled percentages of the background and ITR proteomes. For each FunCat category and species set, the first black bar is the proportion of background proteins belonging to the category, whereas the white bar is the proportion of ITR-containing proteins belonging to the same category.

The relationship between ITRs and variation at the fungal cell surface notwithstanding, the enrichment of several functional processes with ITR-containing proteins suggests that their role in fungi might be more diverse than previously thought (Verstrepen et al. 2005; Levdansky et al. 2007). Experimental (Michael et al. 2007; Paoletti et al. 2007) and bioinformatic evidence (this study) both suggest that ITR variability may be key in a wide variety of physiological and developmentally important processes. For example, ITR variation in the *Neurospora crassa* protein WC-1 likely plays an important role in regulating the organism's circadian clock behavior in response to environmental cues (Michael et al. 2007). In *Aspergillus* (and other filamentous fungi), an important role for ITR variation may be in the control of self/non-self-recognition during somatic cell fusion (Paoletti et al. 2007). We noted that several ITR-containing proteins in the *Aspergillus* possessed Pfam domains characteristic of heterokaryon incompatibility (HET) proteins. The HET protein family is thought to control self-/non-self-recognition across filamentous fungi (Espagne et al. 2002; Paoletti et al. 2007), with repeat variation in these proteins playing a key role in establishing recognition specificity (Paoletti et al. 2007).

Conclusion

Several comparative and functional characteristics of ITR regions appear to be remarkably conserved across eukaryotes, although the ITR regions themselves are very poorly conserved even between very close relatives. The presence of ITRs in a functionally diverse collection of proteins involved in transcriptional regulation and cell-surface activities (fig. 2) suggests that ITRs may be key contributors to the astounding diversity of lifestyles exhibited by fungi.

Supplementary Material

Supplementary tables 1 and 2 and Supplementary data file 1 are available at *Molecular Biology and Evolution*http://www.mbe.oxfordjournals.org/). Supplementary data file 1 is an Excel (.xls) spreadsheet containing all the identified ITRs from the transcriptomes of the eight *Aspergillus* species. For each ITR, the gene name, ETANDEM threshold score, ITR location, ITR repeat unit size, ITR repeat unit copy number, percent consensus sequence conservation, and ITR consensus sequence are reported. For each ITR-containing protein, the predicted signal peptides, transmembrane helices, GPI anchors, PFAM domains, and genuswide orthologs are reported.

Acknowledgments

We thank Melanie Huntley for information on her published experimental protocol on the placement of tandem repeats and Stefanie Butland for information on the methodology of calculating longest pure repeat tracts. This work was conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University. J.G.G. is funded by the Graduate Program in Biological Sciences at Vanderbilt University. Research in A.R.'s lab is supported by the Searle Scholars Program and Vanderbilt University.

Literature Cited

Alba MM, Guigo R. 2004. Comparative analysis of amino acid repeats in rodents and humans. Genome Res. 14: 549–554.

Anmarkrud JA, Kleven O, Bachmann L, Lifjeld JT. 2008. Microsatellite evolution: mutations, sequence variation, and homoplasy in the hypervariable avian microsatellite locus HrU10. BMC Evol Biol. 8.

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. Eukaryot Cell. 6:1392–1399.

Bendtsen JD, Nielsen H, von Heijne G, Brunak S. 2004. Improved prediction of signal peptides: SignalP 3.0. J Mol Biol. 340:783–795.

Bichara M, Wagner J, Lambert IB. 2006. Mechanisms of tandem repeat instability in bacteria. Mutat Res. 598:144–163.

Bowen S, Roberts C, Wheals AE. 2005. Patterns of polymorphism and divergence in stress-related yeast proteins. Yeast. 22:659–668.

Brohede J, Ellegren H. 1999. Microsatellite evolution: polarity of substitutions within repeats and neutrality of flanking sequences. Proc R Soc Lond B – Biol Sci. 266:825–833.

Butland SL, Devon RS, Huang Y, Mead CL, Meynert AM, Neal SJ, Lee SS, Wilkinson A, Yang GS, Yuen MMS. (13 co-authors). 2007. CAG-encoded polyglutamine length polymorphism in the human genome. BMC Genomics. 8:126.

Dieringer D, Schlotterer C. 2003. Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species. Genome Res. 13: 2242–2251.

Drake JW, Charlesworth B, Charlesworth D, Crow JF. 1998. Rates of spontaneous mutation. Genetics. 148:1667–1686.

Eisenhaber B, Schneider G, Wildpaner M, Eisenhaber F. 2004. A sensitive predictor for potential GPI lipid modification sites in fungal protein sequences and its application to genome-wide studies for *Aspergillus nidulans*, *Candida albicans*, *Neurospora crassa*, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. J Mol Biol. 337:243–253.

Ellegren H. 2004. Microsatellites: simple sequences with complex evolution. Nat Rev Genet. 5:435–445.

Espagne E, Balhadere 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.

Fedorova ND, Khaldi N, Joardar VS, Maiti R, Amedeo P, Anderson MJ, Crabtree J, Silva JC, Badger JH, Albarraq A. (38 co-authors). 2008. Genomic islands in the pathogenic filamentous fungus *Aspergillus fumigatus*. PLoS Genet. 4:e1000046.

Fidalgo M, Barrales RR, Ibeas JI, Jimenez J. 2006. Adaptive evolution by mutations in the FLO11 gene. ProcNatl Acad Sci USA. 103:11228–11233.

Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R. (13 co-authors). 2006. Pfam: clans, web tools and services. Nucleic Acids Res. 34:D247–D251.

Fondon JW, Garner HR. 2004. Molecular origins of rapid and continuous morphological evolution. Proc Natl Acad Sci USA. 101:18058–18063.

Frenkel S, Blumenthal EZ. 2002. Jmp in, Ver 4. Jama-Journal of the American Medical Association. 287:1733–1734.

Galagan JE, Calvo SE, Cuomo C, Ma LJ, Wortman JR, Batzoglou S, Lee SI, Basturkmen M, Spevak CC, Clutterbuck J. (50 co-authors). 2005. Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*. Nature. 438:1105–1115.

Gatchel JR, Zoghbi HY. 2005. Diseases of unstable repeat expansion: mechanisms and common principles. Nat Rev Genet. 6:743–755.

Geiser DM, Pitt JI, Taylor JW. 1998. Cryptic speciation and recombination in the aflatoxin-producing fungus *Aspergillus flavus*. Proc Natl Acad Sci USA. 95:388–393.

Hamada K, Terashima H, Arisawa M, Yabuki N, Kitada K. 1999. Amino acid residues in the omega-minus region participate in cellular localization of yeast glycosylphosphatidylinositol-attached proteins. J Bacteriol. 181:3886–3889.

Hancock JM, Simon M. 2005. Simple sequence repeats in proteins and their significance for network evolution. Gene. 345: 113–118.

Huntley MA, Clark AG. 2007. Evolutionary analysis of amino acid repeats across the genomes of 12 *Drosophila* species. Mol Biol Evol. 24:2598–2609.

Iversen M, Burton CM, Vand S, Skovfoged L, Carlsen J, Milman N, Andersen CB, Rasmussen M, Tvede M. 2007. Aspergillus infection in lung transplant patients: incidence and prognosis. Eur J Clin Microbiol Infect Dis. 26:879–886.

Jordan P, Snyder LAS, Saunders NJ. 2003. Diversity in coding tandem repeats in related *Neisseria* spp. BMC Microbiol. 3:23.

Karaoglu H, Lee CM, Meyer W. 2005. Survey of simple sequence repeats in completed fungal genomes. Mol Biol Evol. 22:639–649.

Kashi Y, King DG. 2006. Simple sequence repeats as advantageous mutators in evolution. Trends Genet. 22:253–259.

Katti MV, Ranjekar PK, Gupta VS. 2001. Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol Biol Evol. 18:1161–1167.

Katti MV, Sami-Subbu R, Ranjekar PK, Gupta VS. 2000. Amino acid repeat patterns in protein sequences: their diversity and structural–functional implications. Protein Sci. 9:1203–1209.

Kim TS, Booth JG, Gauch HG, Sun Q, Park J, Lee YH, Lee K. 2008. Simple sequence repeats in *Neurospora crassa*: distribution, polymorphism and evolutionary inference. BMC Genomics. 9:31.

King DG, Soller M, Kashi Y. 1997. Evolutionary tuning knobs. Endeavour. 21:36–40.

Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 305:567–580.

Kumeda Y, Asao T. 2001. Heteroduplex panel analysis, a novel method for genetic identification of *Aspergillus* section *Flavi* strains. Appl Environ Microbiol. 67:4084–4090.

Kurtzman CP, Smiley MJ, Robnett CJ, Wicklow DT. 1986. DNA relatedness among wild and domesticated species in the *Aspergillus–Flavus* froup. Mycologia. 78:955–959.

Kyte J, Doolittle RF. 1982. A simple method for displaying the hydropathic character of a protein. J Mol Biol. 157:105–132.

Lai YL, Shinde D, Arnheim N, Sun FZ. 2003. The mutation process of microsatellites during the polymerase chain reaction. J Comput Biol. 10:143–155.

Lai YL, Sun FZ. 2003. The relationship between microsatellite slippage mutation rate and the number of repeat units. Mol Biol Evol. 20:2123–2131.

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.

Levinson G, Gutman GA. 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol. 4:203–221.

Li YC, Korol AB, Fahima T, Nevo E. 2004. Microsatellites within genes: structure, function, and evolution. Mol Biol Evol. 21:991–1007.

Machida M, Asai K, Sano M, Tanaka T, Kumagai T, Terai G, Kusumoto K, Arima T, Akita O, Kashiwagi Y. (63 co-authors). 2005. Genome sequencing and analysis of *Aspergillus oryzae*. Nature. 438:1157–1161.

Metzgar D, Bytof J, Wills C. 2000. Selection against frameshift mutations limits microsatellite expansion in coding DNA. Genome Res. 10:72–80.

Michael TP, Park S, Kim TS, Booth J, Byer A, Sun Q, Chory J, Lee K. 2007. Simple sequence repeats provide a substrate for phenotypic variation in the *Neurospora crassa* circadian clock. PLoS ONE. 2:e795.

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

Montiel D, Dickinson MJ, Lee HA, Dyer PS, Jeenes DJ, Roberts IN, James S, Fuller LJ, Matsuchima K, Archer DB. 2003. Genetic differentiation of the *Aspergillus* section *Flavi* complex using AFLP fingerprints. Mycol Res. 107:1427–1434.

Moreno-Hagelsieb G, Latimer K. 2008. Choosing BLAST options for better detection of orthologs as reciprocal best hits. Bioinformatics. 24:319–324.

Moxon R, Bayliss C, Hood D. 2006. Bacterial contingency loci: the role of simple sequence DNA repeats in bacterial adaptation. Annu Rev Genet. 40:307–333.

Nierman WC, Pain A, Anderson MJ, Wortman JR, Kim HS, Arroyo J, Berriman M, Abe K, Archer DB, Bermejo C. (97 co-authors). 2006. Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*. Nature. 439:1151–1156.

O'Dushlaine CT, Edwards RJ, Park SD, Shields DC. 2005. Tandem repeat copy-number variation in protein-coding regions of human genes. Genome Biol. 6:R69.

Oh SH, Cheng G, Nuessen JA, Jajko R, Yeater KM, Zhao XM, 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. Microbiol – SGM. 151:673–681.

Paoletti M, Saupe SJ, Clave C. 2007. Genesis of a fungal non-self recognition repertoire. PLoS ONE. 2:e283.

Patterson TF, Kirkpatrick WR, White M, Hiemenz JW, Wingard JR, Dupont B, Rinaldi MG, Stevens DA, Graybill JR. 2000. Invasive aspergillosis. Disease spectrum, treatment practices, and outcomes. I3 *Aspergillus* Study Group. Medicine. 79:250–260.

Pearson CE, Edamura KN, Cleary JD. 2005. Repeat instability: mechanisms of dynamic mutations. Nat Rev Genet. 6:729–742.

Pel HJ, de Winde JH, Archer DB, Dyer PS, Hofmann G, Schaap PJ, Turner G, de Vries RP, Albang R, Albermann K. (69 co-authors). 2007. Genome sequencing and analysis of the versatile cell factory *Aspergillus niger* CBS 513.88. Nat Biotechnol. 25:221–231.

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

Rocha EPC, Matic I, Taddei F. 2002. Over-representation of repeats in stress response genes: a strategy to increase versatility under stressful conditions? Nucleic Acids Res. 30:1886–1894.

Rokas A, Payne G, Fedorova ND, Baker SE, Machida M, Yu J, Georgianna DR, Dean RA, Bhatnagar D, Cleveland TE. (15 co-authors). 2007. What can comparative genomics tell us about species concepts in the genus *Aspergillus*? Stud Mycol. 59:11–17.

Ruepp A, Zollner A, Maier D, Albermann K, Hani J, Mokrejs M, Tetko I, Guldener U, Mannhaupt G, Munsterkotter M. (10 co-authors). 2004. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Res. 32:5539–5545.

Sawyer LA, Hennessy JM, Peixoto AA, Rosato E, Parkinson H, Costa R, Kyriacou CP. 1997. Natural variation in a *Drosophila* clock gene and temperature compensation. Science. 278: 2117–2120.

Schilling G, Sharp AH, Loev SJ, Wagster MV, Li SH, Stine OC, Ross CA. 1995. Expression of the Huntingtons-disease (It15) protein product in Hd patients. Hum Mol Genet. 4:1365–1371.

Schlotterer C, Tautz D. 1994. Chromosomal homogeneity of *Drosophila* ribosomal DNA arrays suggests intrachromosomal exchanges drive concerted evolution. Curr Biol. 4: 777–783.

Selkoe KA, Toonen RJ. 2006. Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. Ecol Lett. 9:615–629.

Sherman SL, Jacobs PA, Morton NE. 1985. Further segregation analysis of the fragile X-syndrome with special reference to transmitting males – reply. Hum Genet. 71:183–183.

Shinde D, Lai YL, Sun FZ, Arnheim N. 2003. Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)(n) and (A/T)(n) microsatellites. Nucleic Acids Res. 31:974–980.

Siwach P, Pophaly SD, Ganesh S. 2006. Genomic and evolutionary insights into genes encoding proteins with single amino acid repeats. Mol Biol Evol. 23:1357–1369.

Sokal RR, Rohlf FJ. 1995. Biometry: the principles and practice of statistics in biological research. New York: W.H. Freeman.

Sutherland GR, Richards RI. 1995. Simple tandem DNA repeats and human genetic-disease. Proc Natl Acad Sci USA. 92:3636–3641.

Thomas EE. 2005. Short, local duplications in eukaryotic genomes. Curr Opin Genet Dev. 15:640–644.

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

van der Woude MW, Baumler AJ. 2004. Phase and antigenic variation in bacteria. Clin Microbiol Rev. 17:581–611.

Venter JCMD, Adams EW, Myers PW, Li RJ, Mural GG, Sutton HO, Smith M, Yandell CA, Evans RA, Holt JD. (263 co-authors). 2001. The sequence of the human genome. Science. 291:1304–1351.

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

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

Weber JL, Wong C. 1993. Mutation of human short tandem repeats. Hum Mol Genet. 2:1123–1128.

Young ET, Sloan JS, Van Riper K. 2000. Trinucleotide repeats are clustered in regulatory genes in *Saccharomyces cerevisiae*. Genetics. 154:1053–1068.

Yu J, Cleveland TE, Nierman WC, Bennett JW. 2005. *Aspergillus flavus* genomics: gateway to human and animal health, food safety, and crop resistance to diseases. Rev Iberoam Micol. 22:194–202.

Yu J, Whitelaw CA, Nierman WC, Bhatnagar D, Cleveland TE. 2004. *Aspergillus flavus* expressed sequence tags for identification of genes with putative roles in aflatoxin contamination of crops. FEMS Microbiol Lett. 237:333–340.

Laura Katz, Associate Editor

Accepted November 30, 2008