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; Karaoğlu 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. 2005; Thomas 2005; Kashi and King 2006), notably high

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

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

doi:10.1093/molbev/msn277
Advance Access publication December 4, 2008

© The Author 2008. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved.
For permissions, please e-mail: journals.permissions@oxfordjournals.org

Downloaded from https://academic.oup.com/mbe/article-abstract/26/3/591/976508/Comparative-and-Functional-Characterization-of by guest on 15 September 2017
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: AAIIH01000000); 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/Aspn1/Aspn1_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: ABD01000000); A. fumigatus Af293 (Genbank: AAFP01000000); Aspergillus clavatus NRRL 1 (Genbank: AAD00000000); 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 ≥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 intraspecific 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
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 = \(\frac{L}{3}\) and N-terminal and C-terminal = \(\frac{(L/3) - (2/3)}{L - 1}\), 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). Glycosylphosphatidylinositol (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

Table 1

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

We identified short (3–39-nt) and long (>39-nt) ITRs in each of the 10 transciptomes using the ETANDEM software (Rice et al. 2000). The total number of ITRs ranged from 172 to 345 per species (table 1, fig. 1). The number of ITR-containing genes ranged from 154 to 317 (table 1). Several ITR-containing genes were 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 tetraneucleotides, although 18 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
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 2A). 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 2A). Within these ITR-containing proteins, only 56% of ITRs were conserved (monomorphic or variable) (table 2B).

I TR 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), Variable (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. 3A). 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. 3B). 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 (the ITR was absent in the ortholog), or No ortholog.

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.
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 *Ange_134078279* is a STE2 (mating pheromone receptor) homolog.

8% and 20% in the two within species comparisons (*A. fumigatus* and *A. niger*, respectively) (fig. 3A). 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. 3A 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 ($t = 4.22$, $N = 94$, $df = 93$, $P = 5.62E-05$ 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 hydrophathy, 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 20 amino acids...
(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 (Vestrepen et al. 2004, 2005; Levndansky 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; Levndansky 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 ($P = 3 \times 10^{-5}$) 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).

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Evolutionary Conservation of ITRs and ITR-Containing Proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. flavus</td>
<td>Aspergillus oryzae</td>
</tr>
<tr>
<td>A. flavus</td>
<td>X</td>
</tr>
<tr>
<td>A. oryzae</td>
<td>84%</td>
</tr>
<tr>
<td>A. terreus</td>
<td>73%</td>
</tr>
<tr>
<td>A. niger</td>
<td>63%</td>
</tr>
<tr>
<td>N. fischeri</td>
<td>75%</td>
</tr>
<tr>
<td>A. fumigatus</td>
<td>74%</td>
</tr>
<tr>
<td>A. clavatus</td>
<td>79%</td>
</tr>
<tr>
<td>A. nidulans</td>
<td>70%</td>
</tr>
<tr>
<td>60%</td>
<td>57%</td>
</tr>
</tbody>
</table>

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 |
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
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 genera 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 3
Amino Acid Composition of ITR-Containing Proteome

| 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.

Table 4
Hydropathy of ITR-Containing Proteome

<table>
<thead>
<tr>
<th>Proteome</th>
<th>Aspergillus flavus</th>
<th>Aspergillus oryzae</th>
<th>Aspergillus terreus</th>
<th>Aspergillus niger</th>
<th>Neosartorya fischeri</th>
<th>Aspergillus fumigatus</th>
<th>Aspergillus clavatus</th>
<th>Aspergillus nidulans</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proteome</td>
<td>Back ITR</td>
<td>Back ITR</td>
<td>Back ITR</td>
<td>Back ITR</td>
<td>Back ITR</td>
<td>Back ITR</td>
<td>Back ITR</td>
<td>Back ITR</td>
</tr>
<tr>
<td>Number of proteins</td>
<td>12,377</td>
<td>210</td>
<td>11,881</td>
<td>182</td>
<td>10,252</td>
<td>154</td>
<td>13,595</td>
<td>317</td>
</tr>
<tr>
<td>Average hydrophathy</td>
<td>−0.3</td>
<td>−0.6</td>
<td>−0.3</td>
<td>−0.6</td>
<td>−0.6</td>
<td>−0.8</td>
<td>−0.3</td>
<td>−0.6</td>
</tr>
</tbody>
</table>

Proteome “Back” = proteins which do not possess an ITR.
Proteome “ITR” = proteins containing one or more ITRs.
*Significant at Bonferroni corrected P value = 0.000625.
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).

### Table 5

**Protein Motif Comparison of ITR-Containing Genes and Background Genes**

<table>
<thead>
<tr>
<th>Species</th>
<th>ITR</th>
<th>Prop</th>
<th>Prop</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Aspergillus flavus</em></td>
<td></td>
<td>0.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Aspergillus oryzae</em></td>
<td></td>
<td>0.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Aspergillus niger</em></td>
<td></td>
<td>0.13</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Aspergillus fumigatus</em></td>
<td></td>
<td>0.07</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Neosartorya fischeri</em></td>
<td></td>
<td>0.10</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Aspergillus terreus</em></td>
<td></td>
<td>0.10</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Aspergillus oryzae</em></td>
<td></td>
<td>0.10</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Aspergillus terreus</em></td>
<td></td>
<td>0.10</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Proteome Back = genes that do not contain the presence of an ITR.*

*Proteome ITR = genes containing one or more ITR.*

*"-0" = absence of a predicted protein motif.*

*"_0" = 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.*

**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 proteome proportions of 16 FunCat categories 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/](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**


Laura Katz, Associate Editor

Accepted November 30, 2008