Complex Selection on Intron Size in Cryptococcus neoformans

Stephanie S. Hughes, Cedric O. Buckley, and Daniel E. Neafsey
Microbial Analysis Group, Broad Institute of MIT and Harvard, Cambridge

We conducted a genome-wide analysis of the roles of mutation and selection in sculpting intron size in the fungal pathogen Cryptococcus neoformans. We find that deletion rate is positively associated with intron length and that insertion rate exhibits a weak negative association with intron length. These patterns suggest that long introns as well as extremely short introns in this unusually intron-rich fungal genome are in mutation–selection disequilibrium and that the proportion of constrained functional sequence in introns does not scale linearly with size. We find that untranslated region introns are longer than coding-region introns and that first introns are substantially longer than subsequent introns, suggesting heterogeneous distribution of constrained functional sequence and/or selective pressures on intron size within genes. In contrast to Drosophila, we find a positive correlation between \( d_{\text{SI}} \) and first intron or last intron length and a negative correlation between \( d_{\text{SI}} \) and internal intron length. This contrasting pattern may indicate that terminal introns and internal introns are differentially subject to hypothesized selection pressures modulating intron size and provides further evidence of widespread selective constraints on noncoding sequences.

Introduction

Introns are ubiquitous features of eukaryotic genomes. Despite their broad distribution, the forces affecting their evolution are only just now becoming clear. Although introns are subject to less evolutionary constraint than coding sequence, they do not evolve neutrally at the sequence level (Waterston et al. 2002; Hare and Palumbi 2003). Introns have been found to be subject to selective constraint at the level of primary sequence for many reasons, including maintenance of splicing efficiency or alternative splicing (Sorek and Ast 2003; Voelker and Berglund 2007) and conservation of transcription factor binding sites (Majewski and Ott 2002).

Though intron size varies widely within and across genomes of different strains, there is evidence to suggest that intron size is not a neutral feature. First, intron size distributions are not symmetrically distributed. Most eukaryotes exhibit a highly skewed intron size distribution, with many introns at what is presumably the minimum length necessary for efficient splicing and a very long tail of introns exhibiting greater length (Mount et al. 1992; Lim and Burge 2001; Yu et al. 2002). In humans, nematodes, and Drosophila, introns in highly expressed genes are shorter than introns in lowly expressed genes (Castillo-Davis et al. 2002; Urrutia and Hurst 2003; Marais et al. 2005), suggesting that selection may be acting to minimize the energetic costs associated with transcription of nontranslated sequence and/or increase the rate of transcription. In addition, there is an inverse relationship between recombination rate and intron size in Drosophila melanogaster (Carvalho and Clark 1999; Comeron and Kreitman 2000) and in Caenorhabditis elegans (Prachumwat et al. 2004). This suggests that regions of high recombination may be preventing the fixation of insertions that augment intron length beyond the minimum size required for splicing and/or that weak selection may preferentially be fixing insertions in low-recombination regions to reduce Hill–Robertson effects and increase the genetic distance between selected loci in flanking exons (Hill and Robertson 1966). Finally, Pragreaves (2006) has found recent evidence in D. melanogaster that small insertions exhibit an elevated probability of fixation relative to other insertions or deletions (indels) and that the indel fixation probabilities may vary according to chromosome.

Eukaryotic genomes generally exhibit a mutation pressure biased toward deletions over mutations (Petrov and Hartl 1997; Petrov 2002), suggesting that most introns in recombining genomic regions should drift to a minimum size required for efficient splicing unless they contain regulatory sites or other functional sequences. Ptak and Petrov (2002) have observed that introns in D. melanogaster exhibit a lower rate of deletion relative to neutrally evolving, “dead-on-arrival” retroelement sequences, suggesting that introns do indeed harbor functional sequences subject to selective constraint.

In the manuscript, we investigate whether introns that are longer than modal size are at selection–mutation equilibrium (presumably because they contain a greater proportion of constrained functional sequences) or alternatively if long introns represent temporary departures from selection–mutation equilibrium and are destined to be trimmed of superfluous nucleotides by deletion-biased mutation pressure. We addressed this question by analyzing the distribution of recent insertion and deletion mutations in 4 recently sequenced genomes from the Cryptococcus neoformans species group. Cryptococcus neoformans is a pathogenic, basidiomycetous yeast exhibiting high intron density relative to other fungi, with an average of 5.3 introns/gene (vs. 1–2 introns/gene for most sequenced ascomycetes; Lofus et al. 2005). In addition to high intron density, C. neoformans offers several other key traits for this analysis, including 4 independent evolutionary lineages of sufficiently close relatedness to permit alignment of noncoding sequences and inference of indels (average \( d_{\text{SI}} = 0.37; \) Neafsey and Galagan 2007) and an intron size distribution that clusters extremely tightly around the modal size of 52 bp (fig. 1).

We find that C. neoformans introns that are shorter than the modal size exhibit an insertion bias and introns

1 Present address: Department of Biology, Jackson State University.

Key words: introns, size, natural selection, Cryptococcus, deletion.

E-mail: neafsey@broad.mit.edu.

doi:10.1093/molbev/msn220
Advance Access publication January 2, 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
longer than the modal size exhibit a deletion bias. Assuming a similar indel mutation profile across all introns, this finding implies that selection is differentially enhancing the fixation probability of insertions in short introns and deletions in long introns and that extremely long or short introns are in mutation–selection disequilibrium. This suggests that long introns do not necessarily contain greater amounts of functional information per capita relative to introns closer to the modal size. We further report a reduced rate of deletion in introns flanked by coding exons relative to introns flanked by noncoding exons in the untranslated region (UTR) or intergenic sequence, perhaps indicating a higher density of constrained functional sequence in coding-region introns relative to other noncoding sequence classes. We find a difference in the evolutionary profile of first-ranked and last-ranked coding-region introns and internal coding-region introns, suggesting different selective forces sculpt intron length variation within genes according to intron position. These findings further demonstrate the pervasive impact of selection on noncoding sequence and better elucidate the relative roles of neutral mutation and selection in defining fundamental genomic features such as intron length distributions.

Materials and Methods

Genome assemblies of C. neoformans strains JEC21, H99, WM276, and R265 were acquired and aligned as previously described (Neafsey and Galagan 2007). We will refer to these genomes as strains in the manuscript, though pairwise synonymous divergence among the strains is generally concordant with species-level divergence. For the purposes of our analysis, therefore, we considered differences observed among strains to be fixed differences rather than polymorphisms. Introns and intergenic regions in the alignment were identified according to the annotation produced by the Institute for Genome Research for strain JEC21, which was based on a library of ~23,000 full-length cDNA-paired sequencing reads (Loftus et al. 2005). All analyses of mutations were conducted using custom Perl scripts. Only introns exhibiting canonical splice sites (GT/AG) in all 4 strains were analyzed, resulting in the elimination of 3,511 coding-region introns and 349 UTR introns. In total, we analyzed 24,702 coding-region introns, 669 UTR introns, and 867 intergenic regions. Indels where sequence was present in 1 strain and absent in 3 were assumed to be insertions, and indels where sequence was present in 3 strains and absent in 1 were assumed to be deletions. Our assumption that the minor allelic state is the derived state may be subject to error when parallel insertions or deletions occur or when the minor allelic state occurs in the strain closest to the root and is in fact ancestral. We expect parallel indel mutations of the same size and position to be quite rare, however, and phylogenetic analyses with more distantly related species of fungi indicate that these Cryptococcus strains are most likely rooted on the internal branch of their clade rather than any of the terminal branches (results not shown). Indels that were present in 2 strains and absent in 2 strains were not analyzed due to inability to infer their ancestral state. Only those indels occurring in alignable regions were retained for analysis, so as to not bias the size spectrum or ratio of observed deletions and insertions. We required >70% alignment similarity and no gaps in the flanking ± 5 bp to consider a region alignable. This filter removed approximately 17% (2,256) of candidate deletions and 15% (1,131) of candidate insertions. Three large insertions exhibiting sequence similarity to transposons were not included in rate analyses because they were extreme outliers with respect to insertion size.

Rates of insertion and deletion were scaled by the number of terminal-branch base pair mutations in the alignments. Counts of terminal-branch mutations were performed using the Baseml program in the PAML v3.14 software package (Yang 1997). An HKY85 substitution model was used, and the following topology was assumed: (JEC21, H99, (WM276, R265)). Rates of insertion were calculated by dividing the total number of inserted base pairs by the number of terminal-branch substitutions. Rates of deletion were calculated in a similar way but with a correction factor to account for the reduced probability or opportunity for fixing large deletions in small introns relative to large introns (Blumenstiel et al. 2002). This “window size” correction for deletion rate ($D$) was implemented as follows:

$$D = \frac{MD}{\sum_{j=1}^{m} \sum_{i=1}^{n} \alpha_{ij} t_i}$$

where $n_{ij}$ is the number of deletions of size $j$ in fragment $i$, $\alpha_{ij}$ is the number of sites available for deletions of size $j$ in fragment $i$ (length $- j + 1$), MD is the maximum observed deletion size, and $t_i$ is the number of terminal substitutions per base pair in fragment $i$ as estimated in PAML.

An additional filter was used to exclude aligned introns and intergenic regions containing misannotated coding sequence. The greater degeneracy of the third position of codons relative to the first 2 positions yields a 3mer-based periodicity to the pattern of divergence between orthologous coding sequences that is not observed between

**Fig. 1.**—Intron size distribution in Cryptococcus neoformans, Drosophila melanogaster, and Homo sapiens. Distributions are based on 29,146 C. neoformans introns, 61,031 D. melanogaster introns, and 239,661 H. sapiens introns. Numbers on horizontal axis indicate midpoint of each 10 bp interval. Homo sapiens and D. melanogaster intron data were drawn, respectively, from the March 2006 and April 2006 assemblies available at the UCSC genome browser database (Karolchik et al. 2003).
Results and Discussion

Intron Size Disequilibrium

We report a strong relationship between indel rates and coding-region intron size in *C. neoformans* (fig. 2). When introns are binned into quintiles by length, insertion rate is significantly greater than deletion rate only for the shortest set of introns (10–49 bp; *P* < 10⁻⁴). Insertion rate and deletion rate are not significantly different for introns in the second length quintile, which contains the mode (mode = 52 bp; bin = 50–52 bp; *P* = 0.110). For all quintiles containing longer introns, deletion rate significantly exceeds insertion rate (*P* < 10⁻⁴). Deletion rate becomes progressively greater with intron size and is significantly different between all consecutive quintile bins (*P* < 10⁻⁴). Insertion rate does not significantly differ between any consecutive quintile bins, but the bin containing the shortest introns (10–49 bp) exhibits an insertion rate significantly greater than introns in the 2 bins, the bins containing the longest introns (bins = 57–63 bp, 64–971 bp; *P* < 10⁻⁴). Assuming that the neutral profile of insertions and deletions is similar across all introns, these patterns imply the role of natural selection in differentially affecting the fixation probability of insertions and deletions according to intron size. The much greater contrast in deletion rate among intron size categories relative to insertion rate indicates that selective modulation of intron size is primarily achieved through regulation of the fixation of deletions. Intron size is not significantly different among the 4 strains (Kruskal–Wallis test, *H* = 3.2, *P* = 0.366), eliminating the possibility that these results could be explained by drift in intron size in one or more strains.

The increased rate of deletion in longer introns could be due to a greater opportunity for fixing large deletions in such introns. Deletions that straddle the intron/exon boundaries or otherwise disrupt proper splicing would presumably be selectively deleterious, and large deletions might be more likely to disrupt splicing in short introns. However, we consider this explanation for the deletion rate variation to be unlikely, given that the size spectrum of deletions is heavily biased toward very short deletions (95% ≤ 5 bp), and our method for calculating deletion rate takes into account variation in mutational opportunity as a function of intron size. The size spectra of insertions and deletions observed in coding-region introns, UTR introns, and intergenic regions are presented in figure 3.

Counts of insertions, deletions, and singletons (mutations present in only one strain) for each noncoding sequence class are presented in table 1. In total, we analyzed 24,703 coding-region introns, 669 UTR introns, and 867 intergenic regions. Overall rates of insertion and deletion in the different sequence classes are presented in figure 4. We observe significant variation in both insertion rate (Kruskal–Wallis test, *H* = 115.3, *P* < 10⁻⁴) and deletion rate (Kruskal–Wallis test, *H* = 109.7, *P* < 10⁻⁴) among the noncoding sequence classes. Insertion rates are similar in UTR introns (0.039 bp/substitution [95% CI: 0.030–0.048]) and intergenic sequence (0.039 bp/substitution [95% CI: 0.0235–0.044]) but significantly lower (bootstrapping: *P* < 10⁻⁴) in coding-region introns (0.032 bp/substitution [95% CI: 0.023–0.044]) and UTR introns (0.032 bp/substitution [95% CI: 0.023–0.044]) but significantly lower (bootstrapping: *P* < 10⁻⁴) in coding-region introns (0.032 bp/substitution [95% CI: 0.023–0.044]) and UTR introns (0.032 bp/substitution [95% CI: 0.023–0.044]) but significantly lower (bootstrapping: *P* < 10⁻⁴) in coding-region introns (0.032 bp/substitution [95% CI: 0.023–0.044]) and UTR introns (0.032 bp/substitution [95% CI: 0.023–0.044]) but significantly lower (bootstrapping: *P* < 10⁻⁴) in coding-region introns (0.032 bp/substitution [95% CI: 0.023–0.044]) and UTR introns (0.032 bp/substitution [95% CI: 0.023–0.044]). Deletion rates are also similar in intergenic sequence (0.075 bp/substitution [95% CI: 0.068–0.083]) and UTR introns (0.084 bp/substitution [95% CI: 0.073–0.096]) but significantly lower (bootstrapping: *P* < 10⁻⁴) in coding-region introns (0.055 bp/substitution [95% CI: 0.039–0.071]) and UTR introns (0.055 bp/substitution [95% CI: 0.039–0.071]). These observations may indicate a higher density of constrained functional sequence in coding introns relative to the other noncoding sequence classes.

Our findings suggest that introns in *Cryptococcus* that significantly deviate in size from the mode are largely in mutation–selection disequilibrium and that long introns therefore do not necessarily exhibit a greater density of functional sequence than short introns. On the contrary, there is a weak but significant positive correlation between
Intron Size Variation within Genes

In addition to global, genome-wide forces affecting intron size, selection is known to affect intron size variation among introns within genes and may influence the genome-wide intron length distribution. As has been observed before in other organisms (Duret 2001; Marais et al. 2005; Hong et al. 2006), we find that UTR introns (average size = 108 bp, standard deviation [SD] = 102 bp) are significantly longer than coding-region introns (average size = 68 bp, SD = 39 bp) in C. neoformans (Mann–Whitney U test, \( U = 6,031,095, P < 10^{-5} \)). This suggests that patterns of intron evolution and functionality in intron-rich fungal lineages may be comparable with those of other eukaryotic lineages that have retained much of their ancestral intron complement (Roy and Gilbert 2005). In Drosophila
and vertebrates, first introns are significantly longer than subsequent introns within genes (Duret 2001; Marais et al. 2005). We observe the same pattern in Cryptococcus, where first coding-region introns (average size = 75.9 bp, SD = 61.3 bp) are significantly longer than subsequent coding introns (average size = 64.4 bp, SD = 42.0 bp; Mann–Whitney U test, $U = 49,127,030$, $P < 10^{-4}$).  Though intron length may vary with rank, the positive association between length and deletion rate we report in figure 2 holds true independently of rank. We find that longer introns (80th length percentile or above) exhibiting ranks 1, 2, 3, 4, and 5 all exhibit a higher deletion rate than shorter introns (20th length percentile or below) of the same rank ($P < 10^{-4}$).

Marais et al. (2005) detect a significant negative correlation ($R_S = -0.20, P < 10^{-4}$) between $d_S$ and intron size for first introns in Drosophila but not for subsequent introns (table 2). Given that regulatory elements are more common in first introns than subsequent introns in mammals and Drosophila (Duret 2001; Majewski and Ott 2002; Keightley and Gaffney 2003; Chamary and Hurst 2004), Marais et al. (2005) attribute this negative correlation between first intron length and $d_S$ to a higher density of regulatory elements in the first introns of slowly evolving genes relative to quickly evolving genes.

We observe a quite different relationship between intron size and $d_S$ in Cryptococcus (table 2). First intron length in Cryptococcus exhibits a weak but significant positive correlation with $d_S$ (Spearman’s $R_S = 0.0753$, $P < 10^{-4}$), as well as $d_S/d_S$ (Spearman’s $R_S = 0.0379$, $P = 0.005$), suggesting that slowly evolving genes in this lineage may be less regulated than their counterparts in Drosophila, or may utilize first introns for regulation to a lesser degree. Surprisingly, however, we observe a weak but significant negative correlation between cumulative subsequent intron length and $d_S$ (Spearman’s $R_S = -0.024$, $P = 0.020$) or $d_S/d_S$ (Spearman’s $R_S = -0.0357$, $P < 10^{-4}$) in Cryptococcus. The incongruence of these associations indicates that first introns and subsequent introns are subject to different selective pressures with regard to length. A summary of correlates of intron length in Cryptococcus and Drosophila is presented in table 2.

### Testing Hypotheses of Selection on Intron Size

Interpretation of the nature of these selective pressures on intron length within genes is not straightforward. The evolutionary rate metrics $d_S$ and $d_S/d_S$ are affected by many different factors (Plotkin and Fraser 2007). In addition, there is controversy as to whether most amino acid changes are adaptive and driven by positive selection (Sawyer et al. 2007) or slightly deleterious and fixed because of insufficiently strong purifying selection (Shapiro et al. 2007).

None of the existing hypotheses of selection on intron size can fully accommodate the observations in table 2. If most of the amino acid changes in Cryptococcus were fixed by positive selection, the negative correlation between subsequent intron length and $d_S$ could be explained by the cost of expression hypothesis (Carvalho and Clark 1999). Under this scenario, the hypothesis predicts that genes that are best able to fix adaptive amino acid substitutions ($= high d_S$) are also best able to fix deletions or prevent fixation of insertions, thereby maximizing expression efficiency and/or rate. This hypothesis neither easily accommodates the positive correlation we observed between first intron length and $d_S$ nor very compatible with the observation that the predominant manifestation of selection in Cryptococcus introns is a reduction in the deletion rate, as opposed to a reduction in insertion rate or acceleration of deletion rate (figs. 2 and 4).

Alternatively, if variation in protein evolutionary rate is mainly driven by the fixation of slightly deleterious mutations, the selective interference hypothesis predicts that longer introns would actually be adaptive in low-recombination regions. Under this hypothesis, longer introns could more effectively decrease Hill–Robertson interference between selected sites in adjacent exons than short introns, reducing the rate of fixation of mildly deleterious replacement mutations and decrementing the $d_S$
and \(d_5/d_3\) metrics. This could potentially explain the negative correlation we observe between \(d_5\) and cumulative subsequent intron length. The interference hypothesis may also explain the incongruence in the \(d_5\) versus length relationship between first introns and subsequent introns, as first introns are less subject to interference selection as a result of their being flanked by at most a single 3' exon, whereas most subsequent introns on average will be flanked by a greater amount of coding sequence and presumably more selected sites (Comeron and Kreitman 2000).

If this interference selection explanation applies, it would predict that the length of last introns would similarly be positively correlated with the rate of protein evolution, as last introns are flanked by at most a single 3' exon. In fact, we find a weak but positive correlation between last intron length and \(d_5\) (Spearman's \(R_S = 0.051, P = 0.0008\)) as well as last intron length and \(d_5/d_3\) (Spearman's \(R_S = 0.038, P = 0.0088\)), similar to the finding for first introns. This suggests that part of the observed disparity in evolutionary profile between first introns and subsequent introns (Marais et al. 2005) may not be driven by a distinct regulatory role played by first introns but instead by the differential capacity of first introns to reduce Hill–Robertson interference relative to internal introns.

Alternatively, it is possible that some unknown selection pressure, perhaps related to transcript folding or splicing efficiency, may regulate intron length within genes as a function of distance from transcript ends. Further analysis of intron size variation within genes across a wider sampling of eukaryotic genomes may help to elucidate any position-dependent selective pressures modulating intron size variation within genes.

## Conclusions

We report the first genome-wide analysis of intronic insertion and deletion profiles in a fungus. In *Cryptococcus*, most introns are very close to the modal size, but natural selection is gradually lengthening short introns and gradually shortening long introns. This suggests that most introns that depart from the modal size are nonoptimal in terms of organismal fitness and may have resulted from chance fixation of a large insertion or deletion mutation sometime in the past. Heterogeneous selective pressures govern intron size variation within genes. Within genes, first introns and subsequent introns differ in length and exhibit opposing correlations with the evolutionary rate of the protein encoded by the parent gene, suggesting that selection to reduce the cost of gene expression, selection to reduce Hill–Robertson interference, or possibly some unknown selective pressure may differentially explain intron size variation according to the position of introns within genes.

## Acknowledgments

S.S.H. was supported by a National Human Genome Research Institute grant to the Broad Institute for summer undergraduate research and the National Institutes of Health-funded RISE program for undergraduate research at Jackson State University. D.E.N. was supported at the Broad Institute by National Institute of Allergy and Infectious Diseases funding. We are grateful to Bruce Birren and Angela Brunache for helping to foster this collaboration. We thank Scott Roy, Justin Blumenstiel, and 2 anonymous reviewers for helpful comments and discussion about the manuscript.

## Literature Cited


## Table 2

### Summary of Correlates with Intron Length in *Cryptococcus* and *Drosophila*

<table>
<thead>
<tr>
<th>Variable 1</th>
<th>Variable 2</th>
<th><em>Cryptococcus</em> Correlation</th>
<th><em>Drosophila</em> Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coding intron length</td>
<td>Intron divergence</td>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td>First coding intron length (d_5) or (d_5/d_3)</td>
<td>Positive</td>
<td>Negative</td>
<td></td>
</tr>
<tr>
<td>Subsequent intron length (d_5) or (d_5/d_3)</td>
<td>Negative</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>Last coding intron length (d_5) or (d_5/d_3)</td>
<td>Positive</td>
<td>Unknown</td>
<td></td>
</tr>
</tbody>
</table>

\(a\) Data from (Marais et al. 2005).


Aoife McLysaght, Associate Editor

Accepted October 2, 2007