# RESEARCH ARTICLE

# Population Genomics of Intron Splicing in 38 *Saccharomyces cerevisiae* Genome Sequences

*Daniel A. Skelly, James Ronald, Caitlin F. Connelly, and Joshua M. Akey*

Department of Genome Sciences, University of Washington

Introns are a ubiquitous feature of eukaryotic genomes, and the dynamics of intron evolution between species has been extensively studied. However, comparatively few analyses have focused on the evolutionary forces shaping patterns of intron variation within species. To better understand the population genetic characteristics of introns, we performed an extensive population genetics analysis on key intron splice sequences obtained from 38 strains of *Saccharomyces cerevisiae*. As expected, we found that purifying selection is the dominant force governing intron splice sequence evolution in yeast, formally confirming that intron-containing alleles are a mutational liability. In addition, through extensive coalescent simulations, we obtain quantitative estimates of the strength of purifying selection ($2N_e s \approx 19$) and use diffusion approximations to provide insights into the evolutionary dynamics and sojourn times of newly arising splice sequence mutations in natural yeast populations. In contrast to previous functional studies, evolutionary analyses comparing the prevalence of introns in essential and nonessential genes suggest that introns in nonribosomal protein genes are functionally important and tend to be actively maintained in natural populations of *S. cerevisiae*. Finally, we demonstrate that heritable variation in splicing efficiency is common in intron-containing genes with splice sequence polymorphisms. More generally, our study highlights the advantages of population genomics analyses for exploring the forces that have generated extant patterns of genome variation and for illuminating basic biological processes.

## Introduction

A distinguishing feature of eukaryotic genomes is the presence of intervening nucleotides that interrupt protein-coding sequences. The majority of these introns are removed by the spliceosome, an ancient molecular machine that was likely present in the most recent common ancestor of all living eukaryotes (Nixon et al. 2002; Simpson et al. 2002; Vanácová et al. 2005). The abundance of spliceosomal introns varies widely between taxa, from just a handful of introns in some protists (Morrison et al. 2007) to hundreds of thousands of introns in vertebrates and plants. Intron sizes, too, vary over several orders of magnitude between species (e.g., Russell et al. 1994; Lander et al. 2001). Despite these profound differences in intron characteristics between taxa, all introns are governed by the same evolutionary forces that regulate genetic elements in any genome (Lynch 2002).

The budding yeast *Saccharomyces cerevisiae* is an important model system for examining the evolution of eukaryotic genomes. The introns of *S. cerevisiae* are unusual among eukaryotes in several respects. Although introns are still being discovered and characterized in this well-annotated genome (Miura et al. 2006; Juneau et al. 2007; Zhang et al. 2007; Yassour et al. 2009), less than 10% of yeast genes contain introns. *Saccharomyces cerevisiae* introns are small (typically $<$600 bp; Spingola et al. 1999), and only a few yeast genes have been reported to undergo alternative splicing (Davis et al. 2000; Miura et al. 2006; Yassour et al. 2009). *Saccharomyces cerevisiae* introns are characterized by highly conserved 5′, 3′, and branch point sequences. The first yeast introns discovered possessed splicing sequences that fit a strict consensus motif (Langford et al. 1984; Teem et al. 1984; Woolford 1989). More recently, a limited number of introns with splice motifs that match a more relaxed consensus have been identified (Davis et al. 2000; Juneau et al. 2007; Zhang et al. 2007; Yassour et al. 2009), although the information content of short yeast intron splice sequences tends to exceed that present in the short introns of a typical multicellular eukaryote (Lim and Burge 2001). Molecular studies have revealed that all positions in the 5′, 3′, and branch point sequences of yeast introns are likely to play some role in determining splicing efficiency, with a few positions especially critical for pre-mRNA splicing (Jacquier et al. 1985; Fouser and Friesen 1986; Woolford 1989). In particular, the first and second positions of the intron, the adenosine in the penultimate position of the branch point, and the AG terminating the intron appear to be necessary to achieve any appreciable level of proper splicing (Newman et al. 1985; Fouser and Friesen 1986; Jacquier and Rosbash 1986; Vijayraghavan et al. 1986). In addition, interdependencies between bases, even outside the conserved splice sites, can render the effects of mutations at some positions unpredictable (e.g., Castanotto and Rossi 1998).

The functional significance of introns in *S. cerevisiae* is poorly understood. The ancestor of extant fungi was likely intron-rich, with an estimated density of roughly four introns per kilobase (Stajich et al. 2007). It has been hypothesized that introns are on their way out of the yeast genome, with intron loss mediated by homologous recombination of reverse-transcribed cDNAs (Fink 1987). An implication of this model is that yeast introns are largely genomic relics unlikely to have functional significance. In support of this hypothesis, there are many examples of introns that can be deleted from the genome without obvious phenotypic consequences, at least under standard laboratory conditions (Ng et al. 1985; Ho and Abelson 1988; Parenteau et al. 2008). In contrast, although most yeast introns have no known functional importance, it is clear that some encode functional elements, such as snoRNAs (Maxwell and Fournier 1995) or promoters (Thompson-Jäger and Domdey

Key words: yeast, intron evolution, purifying selection, ancestral selection graph.

E-mail: akeyj@u.washington.edu.

*Genome. Biol. Evol.* Vol. 2009:466–478.
doi:10.1093/gbe/evp046
Advance Access publication November 17, 2009

1990). Moreover, yeast introns appear to play a more general role in the regulation of gene expression and protein production (Juneau et al. 2006; Pleiss et al. 2007). Introns can be involved in splicing autoregulation (Li et al. 1996), transcriptional and translational enhancement (Furger et al. 2002; Juneau et al. 2006), and transcriptional response to environmental stresses (Pleiss et al. 2007). Notably, perturbation of subtle layers of transcript regulation or systems for responding to environmental stimuli may not be detectable under standard laboratory conditions but might be critical to organismal fitness in more challenging wild environments.

The evolutionary forces governing intron dynamics have been subject to considerable debate (Belshaw and Bensasson 2006). Evolutionary analyses of introns have surveyed a variety of phylogenetic depths, from kingdom (Fedorov et al. 2002; Rogozin et al. 2003; Stajich et al. 2007) to subphylum (Bon et al. 2003; Sharpton et al. 2008). These comparisons have revealed that intron gains and losses are common over long evolutionary timescales. Notably, however, there have been few studies examining the evolutionary forces shaping patterns of intron variation over shorter timescales (Llopart et al. 2002; Lynch 2002; Omilian et al. 2008). Population genetic analyses of intron polymorphism are a powerful approach for exploring the evolutionary trajectory of polymorphisms within introns and the importance of introns as genomic elements.

Here, we describe the first systematic population genomics analysis of intron splicing in yeast. Specifically, we analyzed patterns of polymorphism in key intron splice sequences in 38 strains of *S. cerevisiae* with fully sequenced genomes (Liti et al. 2008). As expected, polymorphisms are rare in sequences important for pre-mRNA splicing in *S. cerevisiae*, consistent with the elimination of deleterious mutations by purifying selection. We performed extensive simulations using the ancestral selection graph (Neuhauser and Krone 1997) to derive quantitative estimates of the strength of purifying selection acting upon these critical intron splice sequences. We compare these estimates to the strength of selection acting on nonsynonymous sites and apply diffusion approximations to explore the evolutionary dynamics of splice sequence polymorphisms. The strong purifying selection we observe acting on intron splice sequences formally confirms that intron-containing alleles are a mutational liability (Lynch 2002) and renews questions about why introns exist in the yeast genome. Additional analyses suggest that extant introns in yeast are not merely genomic relics but that introns tend to be actively maintained in natural populations of *S. cerevisiae*.

## Materials and Methods

### Sequence Data

We used complete haploid genome sequences for 35 *S. cerevisiae* strains sequenced and assembled as part of the *Saccharomyces* Genome Resequencing Project (Liti et al. 2008), along with the reference *S. cerevisiae* genome (October 2007 sequence; http://www.yeastgenome.org/) and two previously sequenced genomes, RM11-1a (http://www.broad.mit.edu/annotation/genome/saccharomyces_cerevisiae/) and YJM789 (Wei et al. 2007). We examined sequences annotated as spliceosomal introns in *Saccharomyces* Genome Database (SGD Project 2008), excluding introns in dubious genes and introns lacking evidence of splicing in previous experimental studies (Spingola et al. 1999; Davis et al. 2000). We also included spliceosomal introns deposited in the Yeast Intron Database (Spingola et al. 1999) and reported in the recent literature (Juneau et al. 2007; Zhang et al. 2007), for a total of 292 introns in 276 genes. The majority of the introns that we studied have experimental support (>75%), with nine introns being initially discovered using unbiased experimental techniques for identifying spliceosomal introns (Juneau et al. 2007; Zhang et al. 2007). The remaining introns were largely annotated by gene- and intron-finding programs based on the *S. cerevisiae* genome and characteristics of known experimentally verified introns (Spingola et al. 1999; SGD Project 2008). We did not include novel splice variants observed in recent large-scale studies of the yeast transcriptome (Miura et al. 2006; Yassour et al. 2009), as many of these observations lack additional experimental support and it is often unclear whether low-abundance sequences might reflect rare alternative splice variants or mis-splicing. After retrieval of the intron-containing gene sequences from the reference genome, we used MEGA Blast (Zhang et al. 2000) to identify homologous sequences in the remaining 37 yeast strains. We aligned the 38 sequences for each gene using MAFFT (Katoh and Toh 2008). We obtained maximum likelihood estimates of genome-wide levels of synonymous and nonsynonymous site divergence for all pairwise comparisons between strains using PAML (Yang 2007).

The strains we examined from the *Saccharomyces* Genome Resequencing Project include nucleotides that have been imputed by taking into account phylogenetic relationships between strains to correct likely sequencing errors and fill in missing data (Liti et al. 2008). We reanalyzed the data using only strains that had <3% imputed data and found qualitatively similar results (supplementary text, Supplementary Material online). As such, we used complete assemblies (including imputed data) for all further analyses, and we expect our conclusions to be robust to the presence of imputed nucleotides.

### Experimental Determination of Splicing Efficiency

We estimated intron splicing efficiency across seven introns in *S. cerevisiae* strains BY4716 (isogenic to S288C, the yeast reference genome strain), DBVPG1373, K11, UWOPS03-461-4, UWOPS83-787-3, YJM975, YS2, and YS4. We grew the strains to mid-log phase ($OD_{660}$ 0.8–1.0) in yeast extract peptone dextrose and extracted RNA by the acid phenol method (Schmitt et al. 1990). We made cDNA by random priming using the Superscript III First-Strand Synthesis kit (Invitrogen Corp). We used primers designed in Primer3 (Rozen and Skaletsky 2000) to amplify the products of genes YBL108C, YBR215W, YLR199C, YLR445W, YML025C, YNL004W, and YNL038W. We visualized the gene products on 2% agarose gels using ethidium bromide and used the program ImageQuant (Molecular Dynamics, Inc) to quantify the amount of spliced and unspliced gene product. We

quantified splicing efficiency as the ratio of the amount of spliced product to the sum of the amounts of spliced and unspliced product. We obtained four biological replicates per strain per intron and analyzed differences in intron splicing efficiency using the nonparametric Kruskal–Wallis rank sum test in R (R Development Core Team 2008).

## Sequence Analysis

*Saccharomyces cerevisiae* introns are characterized by highly conserved 5′, 3′, and branch point sequences (Langford et al. 1984; Teem et al. 1984; Woolford 1989). We examined six, three, and seven base pairs (bp), respectively, of these sequences, which constitute the positions corresponding to the most highly conserved residues in consensus splice sequences (Woolford 1989; Lopez and Séraphin 1999). To summarize nucleotide variation between strains, we used the formula $\hat{\pi} = \sum_{i=1}^{S} h_i w_i$, where $h_i$ is an unbiased estimate of nucleotide diversity for the $i$th segregating site (Tajima 1989), and $w_i$ is a weight for each site calculated by dividing the number of strains with nongap nucleotides at site $i$ by the total number of strains. By ignoring sequence gaps, we minimized the effect of alignment errors or incomplete sequences on our calculations. We plotted position-specific nucleotide diversities, and generated sequence logos, using the R software environment (Bembom 2007; R Development Core Team 2008).

## Estimating the Magnitude of Purifying Selection

To obtain quantitative estimates of the magnitude of selection acting on intron splice sequences, we conducted simulations using the ancestral selection graph (Neuhauser and Krone 1997). We simulated strains as sampled individuals from one common population (panmictic model) or from a model that included population structure (structure model). The complete demographic history of these strains is likely to be complex, but we sought to construct a simple structured model that recapitulated levels of synonymous site divergence observed between strains to gauge the robustness of our results to demographic uncertainty. We constructed a phylogenetic tree based on synonymous site divergences and observed a topology very similar to a tree based on genome-wide pairwise single nucleotide polymorphism differences (Liti et al. 2008). At a crude level, this tree can be subdivided into two divergent groups, loosely corresponding to strains involved in baking and wine production and those from Europe versus Asian, African, and several wild non-European strains. We estimated the time to most recent common ancestor (TMRCA) of these groups to be approximately $1.3N_e$ generations (roughly 11,500 years, adopting estimates of the mutation rate and generation time provided in Fay and Benavides 2005) using the method of Tang et al. (2002).

As the "European" group exhibits markedly lower synonymous site divergence than the "Asian/African" group (0.0058 vs. 0.0097), we chose to model a population bottleneck in the European group occurring after the estimated TMRCA of the two groups. A population bottleneck can be parameterized in terms of the increase in population homozygosity that results from a decrease in population size. In a randomly mating haploid population of finite size $N$, the inbreeding coefficient $F$ reflects the chance that two randomly drawn copies of a gene are identical by descent (Crow and Kimura 1970). Working with $H = 1 - F$, the probability of nonidentity of two gene copies after $t$ generations is $H_t = (1 - 1/N)H_{t-1} = (1 - 1/N)^t$ assuming the population is noninbred at generation 0. Therefore, the increase in homozygosity caused by a bottleneck where the population is held at size $N$ for $t$ generations is $F_t = 1 - (1 - 1/N)^t$. Using the approximation $\log(1 - x) \approx -x$ (for small $x$) leads to the formula $F = t/N$, although this approximation breaks down at about $F > 0.2$, and it becomes more accurate to parameterize severe bottlenecks using the formula $\log(1 - F) = -t/N$. We modeled a bottleneck that began 0.5 $N_e$ generations (roughly 5,000 years) ago and searched a coarse grid of $F = [0.1, 0.2, \ldots, 0.9]$ to determine that a relatively severe bottleneck was necessary to fit the observed data. We conducted a finer search across a range of bottlenecks where $F = [0.675, 0.7, 0.725, \ldots, 0.975, 0.999, 0.9999]$. We selected the best-fitting bottleneck by minimizing the sum of squared differences between the observed and simulated ratios of nucleotide diversity: 1) between-group to overall, 2) European group to Asian/African group, and 3) between-group to mean within-group, calculated at synonymous or simulated neutral sites. A severe bottleneck ($F = 0.875$) provided a very close fit to the observed data using these measures.

Our implementation of the ancestral selection graph applies to evolution in haploid populations or diploid populations in which selection acts additively. We simulated using a four-allele model, where each simulation included a neutral site and a linked selected site at which the scaled selection coefficient was $\sigma = 2N_e s$ for the selectively-favored allele and $\sigma = 0$ for the remaining three alleles. Direct comparison of selected and linked neutral sites ensures that our estimates reflect the effect of selection acting directly on intron splice sequences rather than hitchhiking or background selection. Mutation occurred at the neutral and selected loci with rate $\theta/2$ along each branch with $\theta = 2N_e\mu = 0.0095$ (estimated from the synonymous site substitution rate). We ran simulations for $2N_e s = 0.0, 0.2, 0.4, \ldots, 13.0$ and $2N_e s = 14.0, 15.0, 16.0, \ldots, 50.0$. To increase the speed of simulations for strong selection, for $2N_e s > 13$, we sampled the remaining lineages from the stationary distribution at time 40 $N_e$ generations in the past as suggested by Pritchard (2001). We verified the results for strong selection using a theoretical formula for the distribution of gene frequencies in a panmictic population under a two-allele model with reversible asymmetric mutation and one selectively favored allele (supplementary fig. 1; supplementary text, Supplementary Material online). A program implementing the ancestral selection graph with recombination, balancing selection, and flexible population demographic models is available upon request from the authors.

We simulated 1,000 replicates of 292 simulations for each selective coefficient. To assess the correspondence of simulations conducted for each selective class with the observed data, we calculated the reduction in nucleotide diversity at selected (intronic) sites, relative to neutral (synonymous) sites. To obtain confidence intervals (CIs) for our model selection statistic, we calculated nucleotide diversity

for each of the 1,000 replicates and obtained the interval containing 95% of the realized nucleotide diversities. The number of simulations (unlinked sites) per replicate, 292, was chosen to match the size of our intron data set, which consists of 292 introns.

### Diffusion Approximations for Evolutionary Dynamics of Splice Sequence Polymorphisms

We used diffusion approximations derived by Kimura and Ohta (1969) to explore the evolutionary dynamics of intron splice sequence polymorphisms. These formulas allow for the examination of the fixation probabilities and sojourn times of alleles subject to arbitrary selective advantage or disadvantage and present at arbitrary initial frequencies in the population. We calculated the mean sojourn time of an allele subject to selective disadvantage $s$ using the formula $u(p)\bar{t}_1(p) + [1 - u(p)]\bar{t}_0(p)$, where $u(p)$ is the probability of ultimate fixation of an allele present at initial frequency $p$. $\bar{t}_1(p)$ and $\bar{t}_0(p)$ are the average number of generations until fixation conditional on ultimate fixation of the allele and the average number of generations until loss conditional on ultimate loss of the allele, respectively. These formulas are (Kimura and Ohta 1969):

$$\bar{t}_1(p) = \int_p^1 \psi(\xi)u(\xi)\{1 - u(\xi)\}\mathrm{d}\xi + \frac{1 - u(p)}{u(p)}\int_0^p \psi(\xi)u^2(\xi)\mathrm{d}\xi$$

$$\bar{t}_0(p) = \frac{u(p)}{1 - u(p)}\int_p^1 \psi(\xi)\{1 - u(\xi)\}^2\mathrm{d}\xi + \int_0^p \psi(\xi)\{1 - u(\xi)\}u(\xi)\mathrm{d}\xi$$

Assuming no recurrent mutation, for selection against an allele with disadvantage $s$ in a haploid population of size $N_e$, $u(p)=\frac{1-\exp(2N_esp)}{1-\exp(2N_es)}$ and $\psi(x)=2N_e\frac{\int_0^1 \exp(2N_esx)\mathrm{d}x}{x(1-x)\exp(2N_esx)}$. Similarly, the probability that a mutant at frequency $p$ rises to at least frequency $p'$ is $u(p)=\frac{1-\exp(2N_esp)}{1-\exp(2N_esp')}$. The waiting time until the frequency of such events (which are rare under the conditions we discuss) is exponentially distributed with parameter $\lambda$ equal to the frequency of the event, with an expected value of $1/\lambda$.

To calculate the number of new intron splice sequence mutations per day, we used the following estimates: 1) wild yeast are likely to reproduce at a rate of approximately eight generations per day (Fay and Benavides 2005), 2) the mutation rate at synonymous sites is approximately $u = 1.8 \times 10^{-10}$ (Fay and Benavides 2005), and 3) the effective population size of yeast is roughly $N_e = 26$ million (calculated using $\theta=2N_eu=0.0095$ estimated from synonymous sites).

### Modeling Intron Presence/Absence using Genic Characteristics

We used logistic regression to model the presence/absence of introns using genic characteristics as linear predictors. Specifically, we considered 1) the classification of each gene as essential or nonessential under standard laboratory conditions (Giaever et al. 2002), 2) whether the gene encodes a ribosomal protein, 3) the codon adaptation index (Sharp and Li 1987) as an estimate of the relative expression level of each gene, 4) the genic GC content, and 5) $d_N/d_S$ for sequences from all 38 strains as a proxy for the rate of protein evolution, calculated using PAML (Yang 2007). We implemented the model using the glm function in R (R Development Core Team 2008). Starting with all single predictors and second-order interaction terms, we used the drop1 function (R Development Core Team 2008) to remove predictors that did not significantly improve the fit of the model. Our final model consisted of predictors 1–4 above as well as the interaction between the first and second predictors.

## Results

### Polymorphisms Are Rare in Key Splice Sequences

We compiled a list of 292 introns in 276 genes assembled from a variety of sources (SGD project 2008; Spingola et al. 1999; Juneau et al. 2007; Zhang et al. 2007), excluding introns in dubious genes and introns lacking evidence of splicing in previous experimental studies (Spingola et al. 1999; Davis et al. 2000). In the 38 yeast strains we examined, we were able to identify sequences corresponding to the majority (>50%) of each intron for 286 introns. For the remaining six introns, there were a total of 28 instances where a strain was missing over half the intron sequence. We do not believe these reflect true intron presence–absence polymorphisms. Rather, we attribute the missing bases to incomplete sequence coverage (24/28 instances exist in strains with $<1.5\times$ genome sequence coverage; in all 20 instances where the complete intron is missing, a portion of the coding sequence is missing as well; see Material and Methods and supplementary text, Supplementary Material online for a discussion of the use of sequence data that varies in coverage level). However, it remains a formal possibility that these six introns represent true polymorphic deletions of a large portion of the intron (Llopart et al. 2002).

We focused on six, seven, and three bp of the 5′, branch point, and 3′ splice sequences, respectively (fig. 1). Among the 38 strains we examined, we identified 21 polymorphisms within 23 introns in 20 genes (two polymorphisms occur in splice sequences that are shared among multiple splice variants; table 1). We found no polymorphisms in the remaining 269 introns in 256 genes. It is likely that many of the polymorphisms we identified are functionally neutral. For example, 8 of the 21 polymorphisms involve the alleles [C/T]AG at the 3′ splice site. Because 125 introns in our set use a CAG 3′ splice site and 154 use a TAG 3′ splice site, it is unlikely that a switch between the two has dramatic effects on splicing. However, as we demonstrate below, a subset of these polymorphisms do affect splicing efficiency.

We estimated position-specific nucleotide diversity for the conserved intron splice sequences (fig. 1). Residues previously identified as most critical for achieving splicing—the first two and last two positions of the intron and the

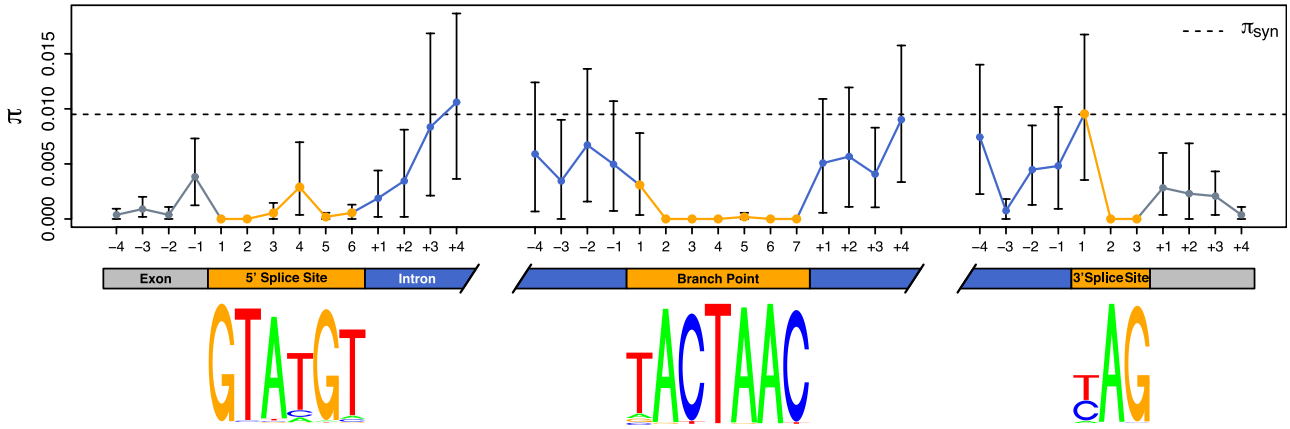


FIG. 1.—Position-specific nucleotide diversity across *Saccharomyces cerevisiae* splice site sequences (orange dots). A window of four bases on each side of the splice site sequences is shown (colored gray for exonic sequences and blue for intronic sequences). The dotted line shows mean nucleotide diversity across synonymous sites. Error bars show 95% CIs based on 1,000 resamplings. Sequence logos beneath splice site sequences were generated using our set of 292 introns, with the height of each position scaled according to information content.

adenosine in the penultimate position of the branch point (Newman et al. 1985; Fouser and Friesen 1986; Jacquier and Rosbash 1986; Vijayraghavan et al. 1986)—showed complete invariance, with no polymorphisms present in any strain for any intron. Interestingly, several positions located in the branch point sequence (2, 3, 4, 5, and 7) and the fifth position in the 5′ splice site also showed either extremely low levels of polymorphism or complete invariance (fig. 1). Experimental studies of the effects of mutations at these positions have produced mixed results (Langford et al. 1984; Jacquier et al. 1985; Parker and Guthrie 1985; Fouser and Friesen 1986; Vijayraghavan et al. 1986); the low levels of variation we observed suggest that levels of functional constraint at these sites are of a similar order of magnitude to previously identified sites critical to splicing.

## Natural Variation in Intron Splicing Efficiency

To better understand natural variation in intron splicing, we surveyed intron splicing efficiency across seven introns in eight strains. We examined introns that had at least one polymorphic splice sequence nucleotide and focused on eight strains chosen to ensure that we captured variation present at both alleles for all seven introns (fig. 2). We estimated splicing efficiency by quantifying the amounts of spliced and unspliced gene product on electrophoretic gels,

**Table 1**
**Summary of Polymorphisms Identified in Yeast Intron Splice Sequences**

| ORF | Gene name[a] | Location | Reference sequence | Alternate sequence | *Saccharomyces paradoxus* sequence[b] |
|---|---|---|---|---|---|
| YBL018C | *POP8* | 5′ splice site | GTA**T**GT | GTA**C**GT | GTACGT |
| YDR367W | | 5′ splice site | GTA**TG**T | GTT**GA**T | — |
| YGL033W | *HOP2* | 5′ splice site | GT**T**AAG | GT**C**AAG | GTAAAG |
| YKL186C | *MTR2* | 5′ splice site | GTATG**T** | GTATG**A** | ACATGA |
| YLR445W | | 5′ splice site | GTA**A**GT | GTA**G**GT | GTAAGT |
| YML025C | *YML6* | 5′ splice site | GTA**C**GT | GTA**T**GT | GTACGT |
| YNL246W | *VPS75* | 5′ splice site | GTA**T**GT | GTA**A**GT | GTAAGT |
| YBR215W | *HPC2* | Branch point | **G**ATTAAC | **C**ATTAAC | TACTAAC |
| YCL002C | | Branch point | **G**ACTAAC | **A**ACTAAC | GACTAAC |
| YKL150W | *MCR1* | Branch point | **T**ACTAAC | **A**ACTAAC | TACTAAC |
| YLR199C | *PBA1* | Branch point | **G**ACTAAC | **A**ACTAAC | GACTAAC |
| YLR316C | *TAD3* | Branch point | **A**ACTAAC | **G**ACTAAC | AACTAAC |
| YNL004W | *HRB1* | Branch point | TACT**A**AT | TACT**G**AT | TACTAAT |
| YBR084C-A | *RPL19A* | 3′ splice site | **C**AG | **T**AG | CAG |
| YBR089C-A | *NHP6B* | 3′ splice site | **T**AG | **C**AG | TAG |
| YKL006C-A | *SFT1* | 3′ splice site | **C**AG | **T**AG | CAG |
| YKL186C | *MTR2* | 3′ splice site | **C**AG | **T**AG | — |
| YNL038W | *GPI15* | 3′ splice site | **C**AG | **T**AG | CAG |
| YNL312W | *RFA2* | 3′ splice site | **C**AG | **T**AG | TAG |
| YOR182C | *RPS30B* | 3′ splice site | **T**AG | **C**AG | CAG |
| YOR234C | *RPL33B* | 3′ splice site | **T**AG | **C**AG | TAG |

[a] Blank gene names indicate uncharacterized open reading frames (ORFs).
[b] Dash indicates that the *S. paradoxus* allele at this position could not be confidently identified.

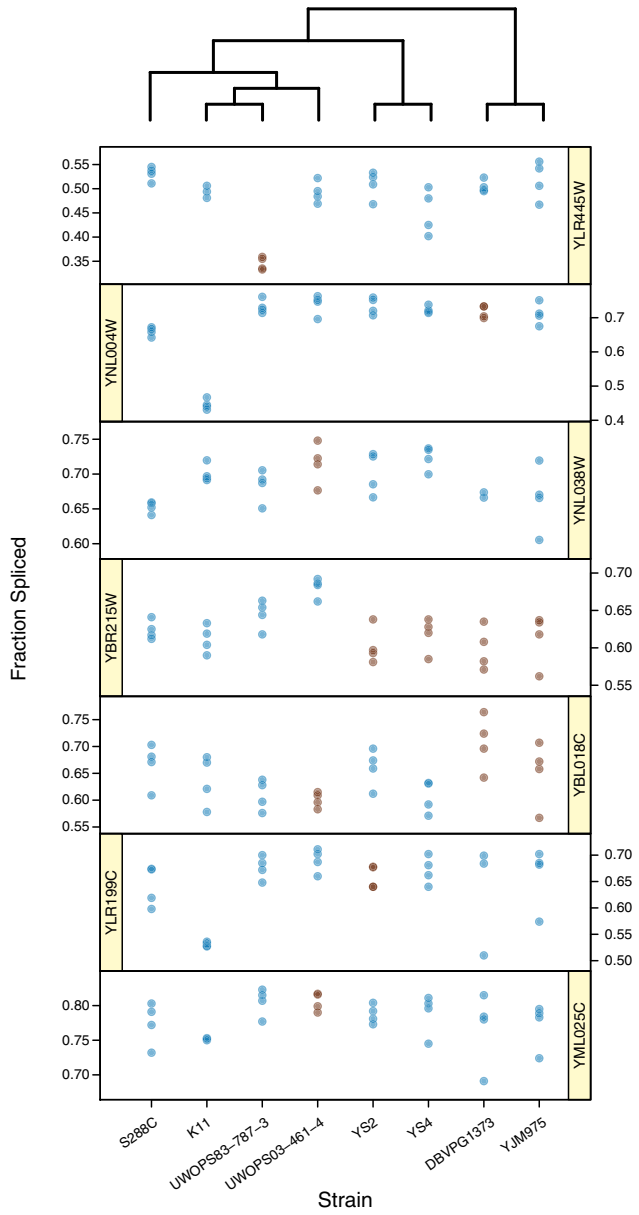


FIG. 2.—Splicing efficiency for seven introns measured across the eight strains shown at bottom. Panels indicate the fraction of spliced gene product measured for four biological replicates of each strain for the gene indicated along the axis. Blue dots indicate strains with the reference (strain S288C) splice sequence allele, and dark red dots indicate strains with the alternative allele. The top four panels show introns with significant variation in splicing efficiency among strains (Kruskal–Wallis rank sum test, $P < 0.05$). Results for introns in genes YBL018C and YLR199C were marginally significant ($P = 0.052$ and $P = 0.060$, respectively), and YML025C did not show significant variation in splicing efficiency among strains ($P = 0.13$). Note that scaling of the vertical axis differs between introns. The phylogeny above the figure depicts the approximate genealogical relationship between strains.

using four biological replicates per strain. Interestingly, we observed significant variation among strains in splicing efficiency for the majority of introns (4/7 introns; Kruskal–Wallis rank sum test, $P < 0.05$; an additional two introns are marginally significant; fig. 2). In one case (YLR445W, a protein of unknown function), a 5′ splice site polymorphism from the common GTAAGT (used in 13 other introns and present at the orthologous 5′ splice site in *Saccharomyces paradoxus*, the closest known relative of *S. cerevisiae*) to the sequence GTAGGT (not found as a 5′ splice site in other introns) resulted in dramatically lower splicing efficiency in strain UWOPS83-787-3 (fig. 2, top panel). Thus, heritable variation in splicing efficiency is common, and polymorphisms in splice sequences can contribute to this variation.

In the remaining six introns, the polymorphisms we observed in splice sequences did not clearly coincide with increases or decreases in experimentally measured splicing efficiencies (fig. 2), suggesting that the genetic basis of splicing efficiency is complex. In addition to conserved splicing sequences, these results suggest that *trans*-acting factors and other *cis*-acting factors (such as additional sequence motifs or spurious splice sequences) also contribute to the efficiency of the splicing reaction (e.g., Couto et al. 1987; Kivens and Siliciano 1996; Castanotto and Rossi 1998; Spingola and Ares 2000). It is not entirely surprising that many of the splice sequence polymorphisms we examined did not track with our measured splicing efficiencies. Specifically, our test set of introns included three polymorphisms between alternate 5′ or 3′ splice site sequences that are common across the global set of introns, suggesting that they should not dramatically affect splicing. Moreover, polymorphisms present in the population may persist precisely because their functional effect on splicing is minimal. Finally, splicing efficiency in these strains was only measured under standard laboratory conditions, and some polymorphisms may have environment-specific effects on splicing.

## Quantitative Estimates of the Strength of Selection Acting on Intron Splice Sequences

The significantly reduced levels of diversity within critical splicing sequences (fig. 1) suggest that most newly arisen mutations at these sites are deleterious and removed by purifying selection. To obtain quantitative estimates of the strength of purifying selection acting on intron splice sequences, we used the ancestral selection graph (Neuhauser and Krone 1997). The ancestral selection graph describes a genealogical process that extends the coalescent by properly taking into account the effect of natural selection (Neuhauser and Krone 1997). We simulated strains as sampled individuals from one common population (panmictic model) or from a model that included population structure, with two subpopulations that split at some time in the past, one of which subsequently experienced a bottleneck (structure model). Obviously, both models are simplifications of the real demographic history of these 38 strains. However, the simple model of population structure we used recapitulates major patterns of synonymous site divergence within and between subpopulations (see Materials and Methods). In addition, it is useful to examine varying models to gauge the robustness of our results to demographic uncertainty (Akey et al. 2004).

We evaluated the fit of the panmictic and structure models to the observed data using the ratio of nucleotide diversity at selected (intronic) sites to diversity at neutral (synonymous) sites. Synonymous sites are subject to weak

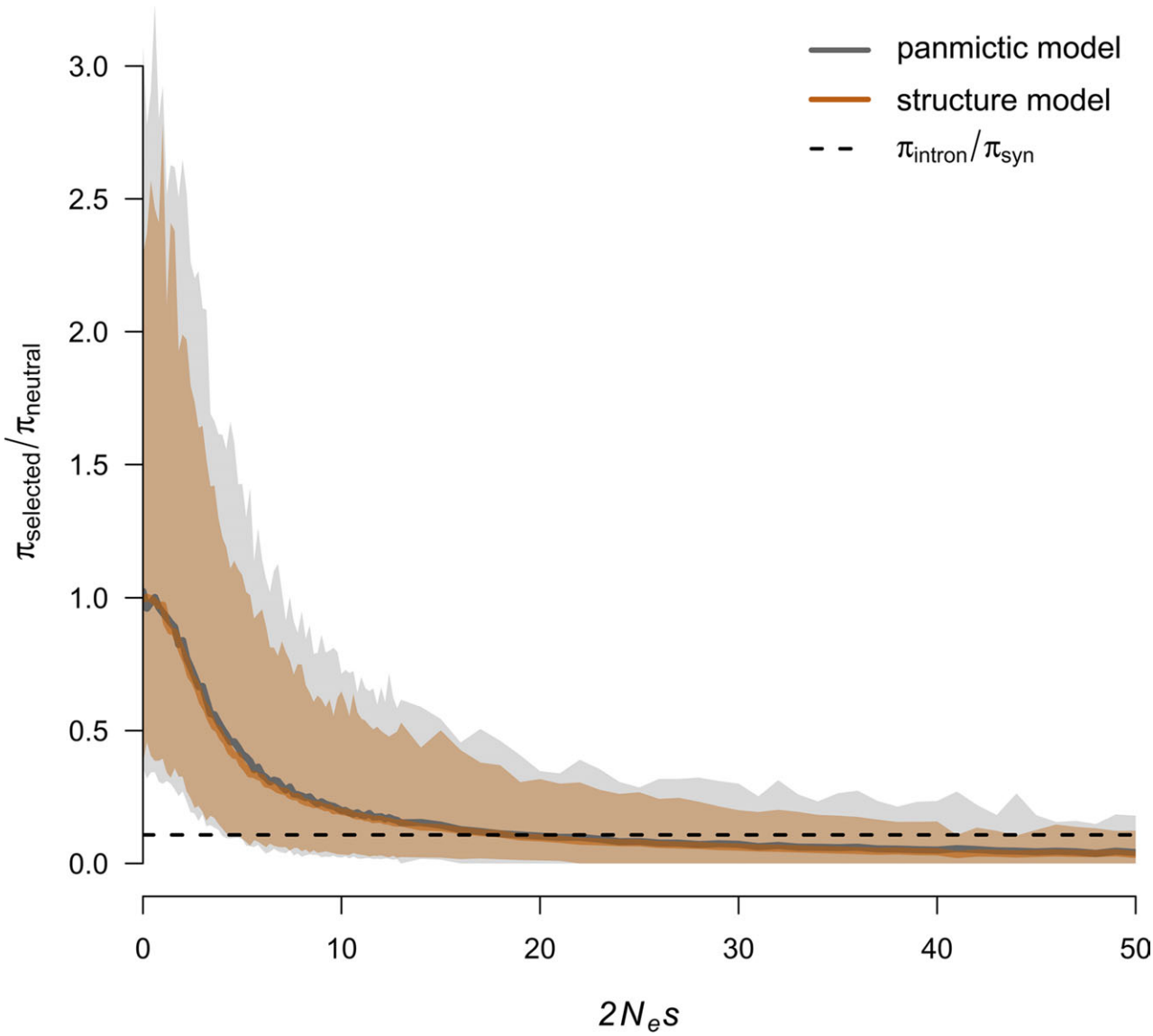


FIG. 3.—Reduction in diversity at simulated selected site, relative to linked neutral site, as a function of the strength of selection against new mutations at the site. Solid lines show the mean of this ratio, across selection coefficients, for the two demographic models studied. Lighter shading gives 95% CIs based on 1,000 simulated replicates sized to match the intron data set. The dashed line shows the reduction in diversity observed in the intron splice sequence data set, relative to synonymous sites.

selective constraint in yeast (Akashi 2001), which suggests that normalization using synonymous sites will lead to slight underestimates of the magnitude of selection against splice sequence polymorphisms; nevertheless, this bias will not be present for relative comparisons between intronic and nonsynonymous sites that are both normalized using synonymous sites (see below). The value of this summary statistic was broadly similar across selection coefficients for our two demographic models, with slightly lower values for the structure model (fig. 3). We first estimated the strength of purifying selection acting on intron splice sequences as a class, then considered the strength of selection acting on each intronic site. Given that the panmictic and structure models give very similar results across the range of selective classes we examined (fig. 3), we provide estimates of the strength of selection based on the results of simulations using the panmictic model. The best fit to the observed intron splice sequence data occurred at $2N_es \approx 19$ (fig. 3); 95% CIs based on simulations suggest that the minimum strength of purifying selection is $2N_es > 4$. Our simulation scheme ensures that these estimates reflect selection acting directly on intron splice sequences rather than hitchhiking or background selection (see Materials and Methods).

Next we investigated the strength of selection acting on individual splice sequence nucleotides. This analysis is complicated by the lack of variation at individual sites for sites subject to very strong selection. To this end, we focused on estimating the lower limits to selection at each site. Lower limits were obtained using the CIs associated with our simulations (fig. 2), which properly account for the stochastic properties of coalescent genealogies as well as sampling variation present in the relatively small intron

**Table 2**
**Position-Specific Estimates of the Minimum Strength of Selection on Intron Splice Sequences (Specified in Terms of the Scaled Selection Coefficient, $2N_es$)**

| | 5′ splice site | | | | | | Branch point | | | | | | | 3′ splice site | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Consensus | G | T | A | T | G | T | T | A | C | T | A | A | C | C/T | A | G |
| Position | 1 | 2 | 3 | 4 | 5 | 6 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 1 | 2 | 3 |
| Estimate | 13.0 | 13.0 | 6.0 | 1.6 | 11.0 | 6.0 | 0.8 | 13.0 | 13.0 | 13.0 | 11.0 | 13.0 | 13.0 | 0.0 | 13.0 | 13.0 |

data set. We observed considerable heterogeneity in the magnitude of selection across yeast intron splice sequence nucleotides (table 2). One class of sites, as mentioned above, showed either very low levels of polymorphism or complete invariance. For these sites (first and last two positions in the intron, branch point positions 2–7, and fifth position in 5′ splice site), we estimate the scaled selection coefficient to be at least $2N_es \approx 11$ (table 2). The lower limit to the strength of selection was approximately one order of magnitude weaker at the fourth position in the 5′ splice site and first position of the branch point (table 2). Notably, levels of variation were indistinguishable from neutrality only at the first position in the 3′ splice site.

Finally, we estimated the strength of selection on nonsynonymous sites using the same demographic models described above. A minority of nonsynonymous mutations may have been driven to high frequency by positive selection (which would downwardly bias our estimates), although positive selection acting on intron splice sequence mutations is also conceivable. We evaluated the fit of models with varying selection intensities using the same metric as above, the ratio of the nucleotide diversity at selected (nonsynonymous) versus putatively neutral (synonymous) sites. Our estimate of the magnitude of purifying selection acting on the average nonsynonymous site is $2N_es = 10.6$. Although this estimate is subject to uncertainty, these results suggest that the strength of purifying selection acting on intron splice sequences as a class is nearly double that acting on an average nonsynonymous site.

## Evolutionary Dynamics of Intron Splice Sequence Polymorphisms

An advantage of obtaining quantitative estimates of the strength of selection acting on intron splice sequences (fig. 3 and table 2) is that these estimates can be used to better understand the evolutionary dynamics of existing genetic variation and newly arising mutations. We used diffusion approximations derived by Kimura and Ohta (1969) to explore the fixation probabilities and sojourn times of alleles as a function of the estimated selection coefficients across intron splice sequence positions (fig. 4). Above, we estimate the average strength of selection against splice sequence mutations as a whole to be $2N_es \approx 19$. With purifying selection of this magnitude, it is approximately 9 million times more likely that a newly arising mutation at a neutral site will eventually rise to fixation than a newly arising mutation at a selectively constrained site. Nevertheless, the mean sojourn times of newly arising neutral and deleterious mutations are quite similar (roughly 20% longer for neutral than for deleterious mutations) because a large fraction of both classes of mutations are lost soon after arising (fig. 4, middle panel). Interestingly, newly arising strongly deleterious ($2N_es \approx 19$) mutations that are destined for ultimate fixation arrive there nearly three times faster on average than new neutral mutations (fig. 4, left panel). This somewhat counterintuitive result arises from the fact that low and moderate frequency mutations that are strongly deleterious are overwhelmingly likely to be lost. Thus, the only new strongly deleterious mutations that rise to ultimate fixation are those exceptionally rare mutants that rapidly and continually increase stochastically in frequency faster than the mutant alleles can be purged from the population by purifying selection.

Heterogeneous selective pressures across intron splice sequence nucleotides result in significantly different predicted evolutionary trajectories for polymorphisms that arise at different positions within splice sequences. Using estimates of the mutation rate, reproductive capacity, and effective population size of yeast (see Materials and Methods), we estimate that an average of 11 new mutations arise each day at any particular intronic splice sequence position in the global yeast population, scattered among the 292 introns we study. The fate of these mutations varies widely, depending on the magnitude of selection against new mutations at the position where they occur. Even for selectively neutral mutations, the probability of ultimate fixation is only about one in 25 million for the large global yeast population. For the splice sequence positions where purifying selection is detectable but weak (table 2), substitutions occur at about 60% the rate at neutral sites, whereas for the class of sites that shows about an order of magnitude stronger selection, the substitution rate is only roughly 0.02% than that at neutral sites. The differences are less extreme when considering the probability that a newly arising mutation becomes common in the population because the fate of new mutants is determined largely by drift while they remain rare. For example, the average waiting time until a new mutation at a particular splice sequence nucleotide (within 1 of the 292 introns we study) attains 10% frequency is roughly 650 years for a neutral mutation, 690 years for a weakly deleterious mutation ($2N_es = 1$), and 1,120 years for a strongly deleterious mutation ($2N_es = 10$) (see fig. 4, right panel, for waiting times scaled by $N_e$). Thus, although strongly deleterious new mutants are ultimately fixed at exceedingly low rates, their behavior at relatively low frequencies does not differ greatly from neutral variants. In some cases, changes in environmental conditions or compensatory evolution could lead initially deleterious mutations managing to drift to moderate frequency to become selectively favored and rise to ultimate fixation.

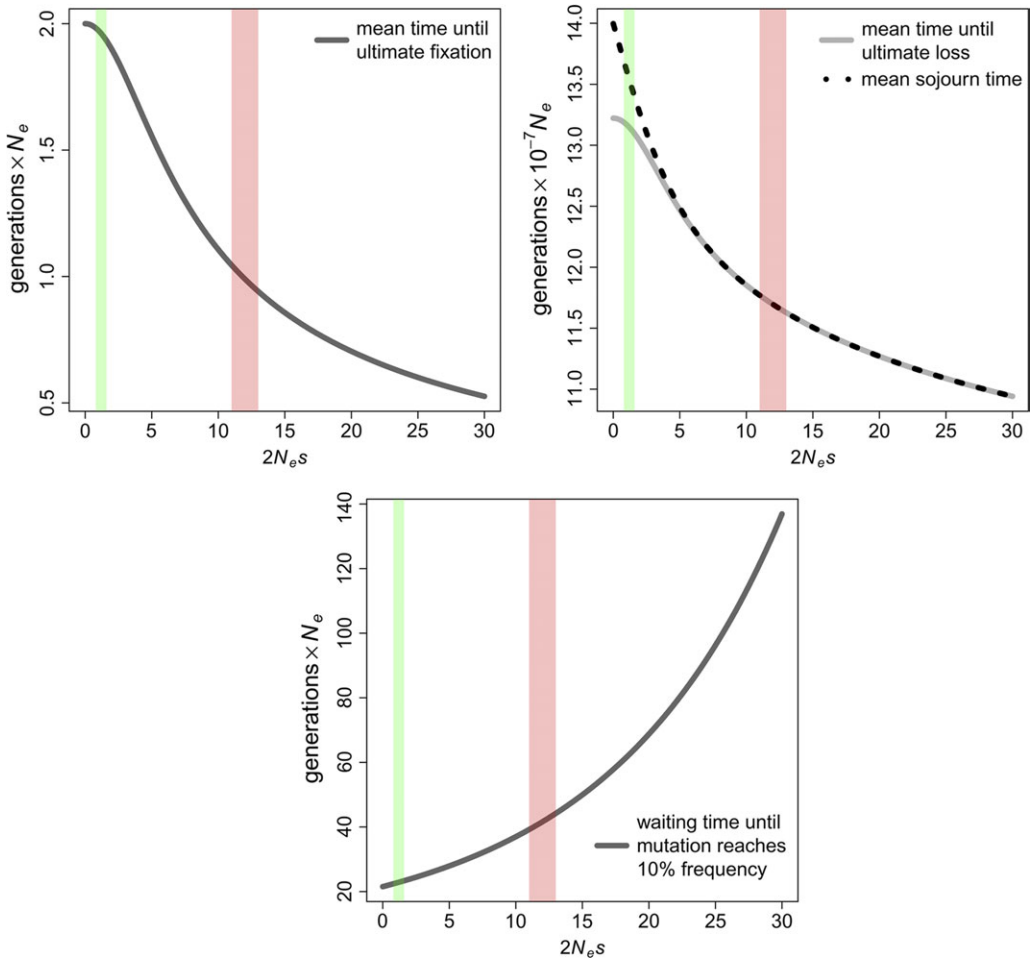


FIG. 4.—Evolutionary dynamics of newly arising mutations. The three panels depict the fates of newly arising mutations as a function of the selection coefficient against new mutations. Times shown on the *y* axis are scaled in terms of the effective population size as shown. Transparent boxes depict lower limits to the magnitude of selection at weakly constrained splice nucleotides (green) and strongly constrained splice nucleotides (red), as discussed in the text and table 2. Left panel shows the mean time for newly arising mutations to achieve fixation conditional upon ultimate fixation. Middle panel shows the mean time for newly arising mutations to be lost conditional upon ultimate loss (solid line) and the mean sojourn time of newly arising mutations (dotted line). The lines converge as $2N_es$ increases because ultimate loss becomes increasingly more probable as selection against new mutants increases. Right panel shows the expected waiting time for a newly arising mutation to reach 10% frequency.

## Why Are Extant Yeast Introns Retained?

Although our analyses clearly demonstrate that purifying selection acts on intron splice sequences, it is less clear what specifically is deleterious about perturbing intron splicing in yeast. One possibility is that most introns are on their way out of the yeast genome (Fink 1987) and are not functionally important. Under this scenario, the sole consequence of perturbing splicing arises when mutations in critical intronic splice sequences disrupt the splicing process, leading to the accumulation of splicing intermediates or improperly spliced transcripts (e.g., Parker and Guthrie 1985; Fouser and Friesen 1986). These defective precursor molecules either result in proteins likely to have impaired function or are eliminated by nonsense-mediated mRNA decay and alternative degradation pathways (Danin-Kreiselman et al. 2003; Hilleren and Parker 2003; Sayani et al. 2008), which would effectively knock out a gene.

An alternative possibility is that yeast introns, in general, are functionally important and play a role in gene regulation. In yeast, the modulation of splicing efficiency can effect rapid response to environmental challenges (Pleiss et al. 2007) and contribute to important biological processes, such as the initiation of meiosis (Engebrecht et al. 1991). Moreover, the interactions between the spliceosome and proteins responsible for transcription, capping, polyadenylation, RNA export, and nonsense-mediated mRNA decay (Maniatis and Reed 2002) support an integral role for introns in affecting transcriptional and translational yield (Juneau et al. 2006). Under this scenario, perturbing splicing could have the same severe effects on the splicing reaction itself as described above but could also have deleterious consequences for normal gene regulation.

To examine the functional importance of yeast introns, we compared the prevalence of introns within genes classified as essential or nonessential under standard laboratory conditions (Giaever et al. 2002). The rationale of this analysis is that because mutations in critical splice sequences disrupt proper splicing and often lead to loss of gene function, introns create a sizeable target for mutation to null alleles (Lynch 2002). Our above estimates of the strong

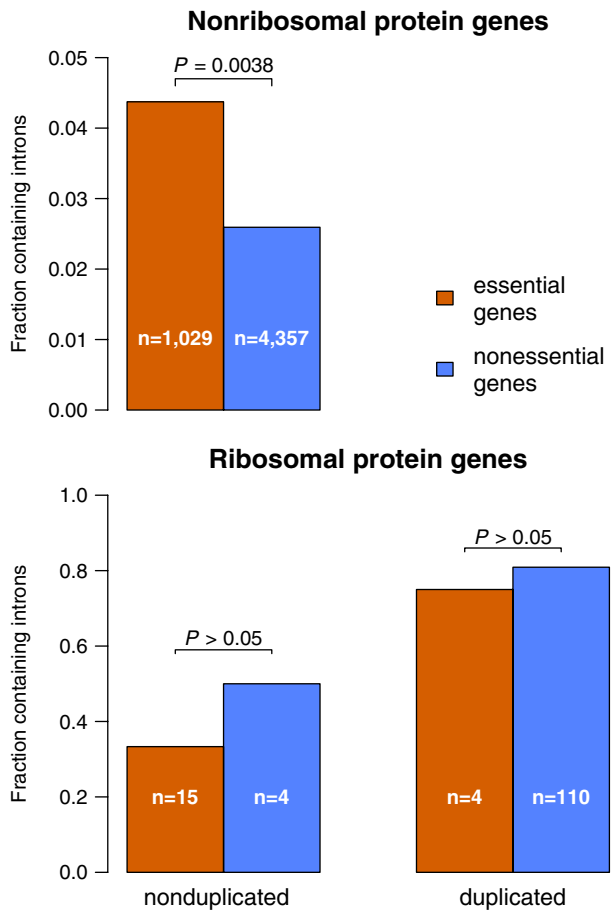


FIG. 5.—Proportion of genes containing introns, divided according to whether the genes are classified as essential (Giaever et al. 2002). Top panel depicts the fraction of intron-containing genes among nonribosomal protein genes. Bottom panel depicts the fraction of intron-containing genes among ribosomal protein genes, divided into ribosomal protein genes that are duplicated and those that are not. Note the difference in vertical scaling between top and bottom panels, as introns are much more common in ribosomal protein genes. *P* values show results from Fisher's exact test of the null hypothesis that the proportion of genes containing introns does not differ between each pair of orange and blue bars. Sample sizes for each category are noted in white text within each bar. Genes that were not classified as essential or nonessential by Giaever et al. (2002) are omitted from this figure.

purifying selection acting on intron splice sequence mutations formally justify this premise by confirming that intron-containing alleles are a mutational liability, since newly arising splice sequence mutations tend to be strongly deleterious. As such, functionless introns that reside passively within genes will tend to be lost (on an evolutionary timescale) more rapidly from essential genes than from nonessential genes. Conversely, functionally important introns will tend to be preserved in all genes, perhaps even more so in essential genes in cases where the function of the intron involves regulation of the gene in which it resides. In the results described below, we analyzed ribosomal and nonribosomal intron-containing genes separately because ribosomal genes exhibit several features (such as high mean levels of expression and larger mean intron sizes) that distinguish them from nonribosomal protein genes (Ares et al. 1999; Spingola et al. 1999).

Interestingly, introns are significantly overrepresented in essential genes that code for nonribosomal proteins (fig. 5). The prevalence of introns in this class of genes suggests that introns tend to have important functions that preserve their presence within nonribosomal protein genes. Among ribosomal protein genes, there is no significant difference in the proportion of essential versus nonessential genes containing introns, within each of the duplicated and nonduplicated subclasses of ribosomal protein genes (fig. 5). These data do not support the hypothesis that introns in ribosomal protein genes are functionless genomic relics, since essential ribosomal protein genes are not significantly less likely to harbor introns. Even so, the lack of a clearer pattern may result from a combination of small sample sizes as well as the fact that many ribosomal protein genes classified as nonessential are nevertheless likely to impose severe growth defects in homozygous mutant form (Giaever et al. 2002).

One proposed mechanism for intron loss in yeast involves homologous recombination of reverse-transcribed cDNAs (Fink 1987). Although other processes may also contribute to intron loss (e.g., simple genomic deletion; Lynch and Richardson 2002), the loss of intronic sequence by RNA-mediated recombination has been experimentally demonstrated in *S. cerevisiae* (Derr et al. 1991). This mechanism predicts that highly expressed genes should lose their introns more rapidly than those expressed at lower levels. To assess whether the distribution of introns in essential and nonessential genes (fig. 5) might be driven by this neutral process rather than reflecting the preservation of functionally important introns, we used logistic regression to model the presence/absence of introns using genic characteristics as linear predictors (see Materials and Methods). We found that a gene's essentiality classification remained a significant predictor of intron presence ($P < 0.01$) even after accounting for transcript abundance (using codon bias as a surrogate measure of expression level), suggesting that our observation of an overrepresentation of introns in essential nonribosomal protein genes does reflect the functional importance of these introns.

## Discussion

The molecular details of intron splicing have been studied in considerable detail. Experimental studies have been extraordinarily valuable for unraveling the molecular basis of the splicing process and for determining the molecular consequences of specific mutations in splice sequences (e.g., Jacquier et al. 1985; Fouser and Friesen 1986; Vijayraghavan et al. 1986). Nevertheless, such approaches are inherently limited by the difficulty in accurately recapitulating the conditions experienced by wild yeast populations, as well as the incomplete detection of every phenotype that affects fitness. Our analyses complement such studies by taking advantage of the characteristic signature imparted on DNA sequence variation by natural selection. Our results are consistent with previous studies that have uniformly identified the first and last two bases of the intron and the sixth branch point position as critical for proper splicing (Newman et al. 1985; Fouser and Friesen

1986; Jacquier and Rosbash 1986; Vijayraghavan et al. 1986). For several other positions where the consensus is less clear (the fifth position of the 5′ splice site as well as branch point positions 2, 3, 4, 5, and 7; Langford et al. 1984; Jacquier et al. 1985; Parker and Guthrie 1985; Fouser and Friesen 1986; Vijayraghavan et al. 1986), we suggest that the level of selective constraint is comparable, and the positions are similarly integral to the splicing process in wild environments.

We estimate the strength of purifying selection acting on yeast intron splice sequences as a class to be nearly double that governing the evolution of an average nonsynonymous polymorphism or a *cis*-acting polymorphism influencing gene expression (Ronald and Akey 2007). Thus, most newly arisen mutations in intron splice sequences are deleterious and are eliminated by purifying selection, although some mildly deleterious alleles may attain appreciable frequencies. This high level of selective constraint partially reflects the specific sequences we examined, which constitute the most critical residues for intron splicing. In contrast, collections of nonsynonymous polymorphisms or *cis*-regulatory polymorphisms are likely to contain many positions at which mutations are functionally neutral, leading to an underestimate of the strength of selection at functionally critical sites. For example, it has been estimated that 36% of nonsynonymous single-nucleotide polymorphisms are deleterious in *S. cerevisiae* (Doniger et al. 2008). Similarly, estimates of the strength of selection acting on extant *cis*-regulatory polymorphisms would be diluted by the presence of neutral polymorphisms in promoters and 3′ untranslated regions (Ronald and Akey 2007). Some of the constraint we detect may also reflect the functional importance of regulated splicing (see below). Might the strength of purifying selection acting on yeast intron splice sequences vary systematically between genes? We examined a variety of genic features (gene ontology terms, GC content, $d_N/d_S$, and codon bias) separately in ribosomal protein encoding and nonribosomal protein-encoding genes with and without polymorphic intron splice sequences (see supplementary text, Supplementary Material online). We observed no detectable heterogeneity between genes with or without polymorphic intron splice sequences, suggesting that levels of functional constraint for intron splicing are broadly similar within ribosomal protein-encoding genes and within nonribosomal protein-encoding genes.

In this paper, we present evidence suggesting that introns tend to be actively maintained in *S. cerevisiae*. It is important to note that there are several possible mechanisms through which intronic sequences might contribute to organismal function. First, the modulation of splicing efficiency contributes to important biological processes, such as the initiation of meiosis (Engebrecht et al. 1991) and can be an important mechanism for gene regulation (Pleiss et al. 2007). Selective constraint attributable to this function would be reflected in the strong purifying selection we observe acting on key splice sequences and would contribute to a greater retention of introns in essential genes (fig. 5). Second, the close association between the spliceosome and transcriptional machinery (Maniatis and Reed 2002) points to a general role for introns in affecting transcriptional and translational yield of the genes in which they reside (Juneau et al. 2006). Because this intronic feature is independent of specific splice sequence nucleotides, it is not reflected in purifying selection on splice sequences, although it is likely to contribute to the retention of introns in essential genes. Finally, intronic bases not directly involved in the splicing reaction could encode functional elements such as promoters (Thompson-Jäger and Domdey 1990), snoRNAs (Maxwell and Fournier 1995), or binding sites for regulatory proteins. When the functional importance of such intron-encoded elements is unrelated to the importance of the gene in which the intron resides, this form of constraint is not detectable by our analyses.

Fink's (1987) proposal that intron loss in yeast occurs largely through homologous recombination of reverse-transcribed cDNAs predicts the 5′ bias in intron location observed in the yeast genome. It might be expected that this bias would be absent in essential genes, where we argue that introns tend to be preserved due to their functional importance (fig. 5). In fact, we observed a strong 5′ bias in location for introns in both essential and nonessential genes (supplementary fig. 2, Supplementary Material online). For both ribosomal and nonribosomal protein genes, there is no significant difference in the distribution of intron locations between essential and nonessential genes (Wilcoxon–Mann–Whitney test; $P > 0.05$). However, this does not contradict our assertion that introns in essential nonribosomal protein genes have been preserved due to functional importance. First, the sample sizes for the nonparametric test above are small, and the power to detect a subtle difference in locational bias is likely to be low. Second, an evolutionary process of intron gain occurring uniformly throughout genes and intron loss preferentially occurring at the 3′ end of genes (as predicted by Fink's model) predicts a steady-state distribution of introns that is 5′ biased. Thus, unless insertion of a new intron into a gene is immediately adaptive, most introns that filter through the sieve of natural selection will be positioned near the 5′ end of the gene.

The population genomics analyses we present allowed us to characterize the forces governing the evolutionary trajectory of polymorphisms in key splice sequences in *S. cerevisiae*. Our quantitative analyses demonstrate that these sequences are subject to strong functional constraint. We propose that introns are not merely genomic relics on their way out of the yeast genome; patterns of intron prevalence in essential and nonessential genes suggest that, at least in nonribosomal protein genes, introns appear to be actively maintained for their functional importance. Our relatively high estimate of the magnitude of purifying selection governing the evolution of splice sequences reflects the need for intronic bases to be properly removed from transcripts but is also likely to arise from the functional importance of regulated splicing. Ultimately, disentangling the possible contributions of yeast introns to organismal function will require detailed studies of splicing in different environments and at different points in the cell cycle. Obtaining a better understanding of the dynamics and functional importance of introns in yeast may inform our understanding of the prevalence of introns in more complex eukaryotic genomes.

## Supplementary Material

Supplementary text and figures 1–2 are available at *Genome Biology and Evolution* online (http://www.oxfordjournals.org/our_journals/gbe/).

## Acknowledgments

We wish to thank the researchers involved in the *Saccharomyces* Genome Resequencing Project for openly releasing and documenting their data and making their prepublication manuscript freely available. We thank Beth Dumont for helpful comments on the manuscript. This work was supported by an NIH training grant to the University of Washington (D.A.S.) and a Sloan fellowship in Computational Biology (J.M.A.).

## Literature Cited

Akashi H. 2001. Gene expression and molecular evolution. Curr Opin Genet Dev. 11:660–666.

Akey JM, et al. 2004. Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS Biol. 2:e286.

Ares M Jr., Grate L, Pauling MH. 1999. A handful of intron-containing genes produces the lion's share of yeast mRNA. RNA. 5:1138–1139.

Belshaw R, Bensasson D. 2006. The rise and falls of introns. Heredity. 96:208–213.

Bembom O. 2007. Sequence logos for DNA sequence alignments. R package version 1.6.0.

Bon E, et al. 2003. Molecular evolution of eukaryotic genomes: hemiascomycetous yeast spliceosomal introns. Nucleic Acids Res. 31:1121–1135.

Castanotto D, Rossi JJ. 1998. Cooperative interaction of branch signals in the actin intron of Saccharomyces cerevisiae. Nucleic Acids Res. 26:4137–4145.

Couto JR, Tamm J, Parker R, Guthrie C. 1987. A trans-acting suppressor restores splicing of a yeast intron with a branch point mutation. Genes Dev. 1:445–455.

Crow JF, Kimura M. 1970. An introduction to population genetics theory. New York: Harper and Row.

Danin-Kreiselman M, Lee CY, Chanfreau G. 2003. RNAse III-mediated degradation of unspliced pre-mRNAs and lariat introns. Mol Cell. 11:1279–1289.

Davis CA, Grate L, Spingola M, Ares M Jr. 2000. Test of intron predictions reveals novel splice sites, alternatively spliced mRNAs and new introns in meiotically regulated genes of yeast. Nucleic Acids Res. 28:1700–1706.

Derr LK, Strathern JN, Garfinkel DJ. 1991. RNA-mediated recombination in Saccharomyces cerevisiae. Cell. 67:355–364.

Doniger SW, et al. 2008. A catalog of neutral and deleterious polymorphism in yeast. PLoS Genet. 4:e1000183.

Engebrecht JA, Voelkel-Meiman K, Roeder GS. 1991. Meiosis-specific RNA splicing in yeast. Cell. 66:1257–1268.

Fay JC, Benavides JA. 2005. Evidence for domesticated and wild populations of Saccharomyces cerevisiae. PLoS Genet. 1:e5.

Fedorov A, Merican AF, Gilbert W. 2002. Large-scale comparison of intron positions among animal, plant, and fungal genes. Proc Natl Acad Sci USA. 99:16128–16133.

Fink GR. 1987. Pseudogenes in yeast? Cell. 49:5–6.

Fouser LA, Friesen JD. 1986. Mutations in a yeast intron demonstrate the importance of specific conserved nucleotides for the two stages of nuclear mRNA splicing. Cell. 45:81–93.

Furger A, O'Sullivan JM, Binnie A, Lee BA, Proudfoot NJ. 2002. Promoter proximal splice sites enhance transcription. Genes Dev. 16:2792–2799.

Giaever G, et al. 2002. Functional profiling of the Saccharomyces cerevisiae genome. Nature. 418:387–391.

Hilleren PJ, Parker R. 2003. Cytoplasmic degradation of splice-defective pre-mRNAs and intermediates. Mol Cell. 12:1453–1465.

Ho CK, Abelson J. 1988. Testing for intron function in the essential Saccharomyces cerevisiae tRNA(SerUCG) gene. J Mol Biol. 202:667–672.

Jacquier A, Rodriguez JR, Rosbash M. 1985. A quantitative analysis of the effects of 5′ junction and TACTAAC box mutants and mutant combinations on yeast mRNA splicing. Cell. 43:423–430.

Jacquier A, Rosbash M. 1986. RNA splicing and intron turnover are greatly diminished by a mutant yeast branch point. Proc Natl Acad Sci USA. 83:5835–5839.

Juneau K, Miranda M, Hillenmeyer ME, Nislow C, Davis RW. 2006. Introns regulate RNA and protein abundance in yeast. Genetics. 174:511–518.

Juneau K, Palm C, Miranda M, Davis RW. 2007. High-density yeast-tiling array reveals previously undiscovered introns and extensive regulation of meiotic splicing. Proc Natl Acad Sci USA. 104:1522–1527.

Katoh K, Toh H. 2008. Recent developments in the MAFFT multiple sequence alignment program. Brief Bioinform. 9:286–298.

Kimura M, Ohta T. 1969. The average number of generations until fixation of a mutant gene in a finite population. Genetics. 61:763–771.

Kivens W, Siliciano PG. 1996. RNA sequences upstream of the 3′ splice site repress splicing of mutant yeast ACT1 introns. RNA. 2:492–505.

Lander ES, et al. 2001. Initial sequencing and analysis of the human genome. Nature. 409:860–921.

Langford CJ, Klinz FJ, Donath C, Gallwitz D. 1984. Point mutations identify the conserved, intron-contained TACTAAC box as an essential splicing signal sequence in yeast. Cell. 36:645–653.

Li B, Vilardell J, Warner JR. 1996. An RNA structure involved in feedback regulation of splicing and of translation is critical for biological fitness. Proc Natl Acad Sci USA. 93: 1596–1600.

Lim LP, Burge CB. 2001. A computational analysis of sequence features involved in recognition of short introns. Proc Natl Acad Sci USA. 98:11193–11198.

Liti G, et al. 2008. Population genomics of domestic and wild yeasts. Nature. doi:10.1038/nature07743

Llopart A, Comeron JM, Brunet FG, Lachaise D, Long M. 2002. Intron presence-absence polymorphism in Drosophila driven by positive Darwinian selection. Proc Natl Acad Sci USA. 99:8121–8126.

Lopez PJ, Séraphin B. 1999. Genomic-scale quantitative analysis of yeast pre-mRNA splicing: implications for splice-site recognition. RNA. 5:1135–1137.

Lynch M. 2002. Intron evolution as a population-genetic process. Proc Natl Acad Sci USA. 99:6118–6123.

Lynch M, Richardson AO. 2002. The evolution of spliceosomal introns. Curr Opin Genet Dev. 12:701–710.

Maniatis T, Reed R. 2002. An extensive network of coupling among gene expression machines. Nature. 416:499–506.

Maxwell ES, Fournier MJ. 1995. The small nucleolar RNAs. Annu Rev Biochem. 64:897–934.

Miura F, et al. 2006. A large-scale full-length cDNA analysis to explore the budding yeast transcriptomes. Proc Natl Acad Sci USA. 103:17846–17851.

Morrison HG, et al. 2007. Genomic minimalism in the early diverging intestinal parasite Giardia lamblia. Science. 317: 1921–1926.
Neuhauser C, Krone SM. 1997. The genealogy of samples in models with selection. Genetics. 145:519–534.
Newman AJ, Lin RJ, Cheng SC, Abelson J. 1985. Molecular consequences of specific intron mutations on yeast mRNA splicing in vivo and in vitro. Cell. 42:335–344.
Ng R, Domdey H, Larson G, Rossi JJ, Abelson J. 1985. A test for intron function in the yeast actin gene. Nature. 314:183–184.
Nixon JEJ, et al. 2002. A spliceosomal intron in Giardia lamblia. Proc Natl Acad Sci USA. 106:3701–3705.
Omilian AR, Scofield DG, Lynch M. 2008. Intron presence-absence polymorphisms in Daphnia. Mol Biol Evol. 25: 2129–2139.
Parenteau J, et al. 2008. Deletion of many yeast introns reveals a minority of genes that require splicing for function. Mol Biol Cell. 19:1932–1941.
Parker R, Guthrie C. 1985. A point mutation in the conserved hexanucleotide at a yeast 5′ splice junction uncouples recognition, cleavage, and ligation. Cell. 41:107–118.
Pleiss JA, Whitworth GB, Bergkessel M, Guthrie C. 2007. Rapid, transcript-specific changes in splicing in response to environmental stress. Mol Cell. 27:928–937.
Pritchard JK. 2001. Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet. 69:124–137.
R Development Core Team. 2008. R: a language and environment for statistical computing [Internet]. Vienna (Austria): R Foundation for Statistical Computing.ISBN 3-900051-07-0. [cited 2008 December 10]. Available from: http://www.R-project.org.
Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV. 2003. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. Curr Biol. 13:1512–1517.
Ronald J, Akey JM. 2007. The evolution of gene expression QTL in Saccharomyces cerevisiae. PLoS ONE. 2:e678.
Rozen S, Skaletsky HJ. 2000. Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S, editors. Bioinformatics methods and protocols: methods in molecular biology. Totowa (NJ): Humana Press. p. 365–386.
Russell CB, Fraga D, Hinrichsen RD. 1994. Extremely short 20-33 nucleotide introns are the standard length in Paramecium tetraurelia. Nucleic Acids Res. 22:1221–1225.
Sayani S, Janis M, Lee CY, Toesca I, Chanfreau GF. 2008. Widespread impact of nonsense-mediated mRNA decay on the yeast intronome. Mol Cell. 31:360–370.
Schmitt ME, Brown TA, Trumpower BL. 1990. A rapid and simple method for preparation of RNA from Saccharomyces cerevisiae. Nucleic Acids Res. 25:3091–3092.
SGD Project. 2008. Saccharomyces genome database [Internet]. [cited 2008 Oct 15]. Available from: http://yeastgenome.org
Sharp PM, Li WH. 1987. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15:1281–1295.
Sharpton T, Neafsey D, Galagan J, Taylor J. 2008. Mechanisms of intron gain and loss in Cryptococcus. Genome Biol. 9:R24.
Simpson AG, MacQuarrie EK, Roger AJ. 2002. Eukaryotic evolution: early origin of canonical introns. Nature. 419:270.
Spingola M, Ares M Jr. 2000. A yeast intronic splicing enhancer and Nam8p are required for Mer1p-activated splicing. Mol Cell. 6:329–338.
Spingola M, Grate L, Haussler D, Ares M Jr. 1999. Genome-wide bioinformatic and molecular analysis of introns in Saccharomyces cerevisiae. RNA. 5:221–234.
Stajich JE, Dietrich FS, Roy SW. 2007. Comparative genomic analysis of fungal genomes reveals intron-rich ancestors. Genome Biol. 8:R223.
Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 123: 585–595.
Tang H, Siegmund DO, Shen P, Oefner P, Feldman MW. 2002. Frequentist estimation of coalescence times from nucleotide sequence data using a tree-based partition. Genetics. 161: 447–459.
Teem JL, et al. 1984. A comparison of yeast ribosomal protein gene DNA sequences. Nucleic Acids Res. 12:8295–8312.
Thompson-Jäger S, Domdey H. 1990. The intron of the yeast actin gene contains the promoter for an antisense RNA. Curr Genet. 17:269–273.
Vanácová S, Yan W, Carlton JM, Johnson PJ. 2005. Spliceosomal introns in the deep-branching eukaryote Trichomas vaginalis. Proc Natl Acad Sci USA. 102:4430–4435.
Vijayraghavan U, et al. 1986. Mutations in conserved intron sequences affect multiple steps in the yeast splicing pathway, particularly assembly of the spliceosome. EMBO J. 5: 1683–1695.
Woolford JL. 1989. Nuclear pre-mRNA splicing in yeast. Yeast. 5:439–457.
Wei W, et al. 2007. Genome sequencing and comparative analysis of Saccharomyces cerevisiae strain YJM789. Proc Natl Acad Sci USA. 104:12825–12830.
Yang Z. 2007. PAML 4: a program package for phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24: 1586–1591.
Yassour M, et al. 2009. Ab initio construction of a eukaryotic transcriptomes by massively parallel mRNA sequencing. Proc Natl Acad Sci USA. doi:10.1073/pnas.0812841106
Zhang Z, Hesselberth JR, Fields S. 2007. Genome-wide identification of spliced introns using a tiling microarray. Genome Res. 17:503–509.
Zhang Z, Schwartz S, Wagner L, Miller W. 2000. A greedy algorithm for aligning DNA sequences. J Comput Biol. 7: 203–214.

Chung-I Wu, Associate Editor

Accepted November 6, 2009