OPEN ACCESS
Journal of Bioinformatics and Computational Biology
Vol. 19, No. 6 (2021) 2140013 (17 pages)

DOI: 10.1142/S0219720021400138

World Scientific
www.worldscientific.com

Evidence for exon shuffling is sensitive to model choice

Xiaoyue Cui[*,§], Maureen Stolzer[†,¶] and Dannie Durand[‡,‖]

[*]*Department of Computational Biology*
Carnegie Mellon University, 4400 Fifth Avenue
Pittsburgh, PA 15213, USA

[†]*Department of Biological Sciences*
Carnegie Mellon University, 4400 Fifth Avenue
Pittsburgh, PA 15213, USA

[‡]*Departments of Biological Sciences and Computational Biology*
Carnegie Mellon University, 4400 Fifth Avenue
Pittsburgh, PA 15213, USA
[§]*xiaoyuec@andrew.cmu.edu*
[¶]*mstolzer@andrew.cmu.edu*
[‖]*durand@cmu.edu*

Received 5 September 2021
Accepted 24 September 2021
Published 19 November 2021

The exon shuffling theory posits that intronic recombination creates new domain combinations, facilitating the evolution of novel protein function. This theory predicts that introns will be preferentially situated near domain boundaries. Many studies have sought evidence for exon shuffling by testing the correspondence between introns and domain boundaries against chance intron positioning. Here, we present an empirical investigation of how the choice of null model influences significance. Although genome-wide studies have used a uniform null model, exclusively, more realistic null models have been proposed for single gene studies. We extended these models for genome-wide analyses and applied them to 21 metazoan and fungal genomes. Our results show that compared with the other two models, the uniform model does not recapitulate genuine exon lengths, dramatically underestimates the probability of chance agreement, and overestimates the significance of intron-domain correspondence by as much as 100 orders of magnitude. Model choice had much greater impact on the assessment of exon shuffling in fungal genomes than in metazoa, leading to different evolutionary conclusions in seven of the 16 fungal genomes tested. Genome-wide studies that use this overly permissive null model may exaggerate the importance of exon shuffling as a general mechanism of multidomain evolution.

Keywords: Exon shuffling; null model; intron.

1. Introduction

Soon after the discovery of introns, Gilbert[1] hypothesized that exon-intron gene organization could facilitate the evolution of gene function, through a process he called *exon shuffling*. If exons encode specific functions, then new combinations of these functional units can arise through recombination within introns. Blake[2] subsequently observed that a new architecture arising from reassortment of exons must encode a foldable protein to be advantageous. This is most likely to occur if the exons themselves correspond to structural units. These observations, taken together, predict (1) that sequences flanked by introns encode "integrally folded protein units"[3] and (2) that introns will be situated near the boundaries of those units.

The notion of what constitutes an "integrally folded protein unit" evolved over time. Exons are typically too short to encode entire folds. That exons might encode smaller, structurally compact regions or elements of secondary structure was considered, but no compelling relationship between exons and a quantum of protein structure emerged.[4] Moreover, intron gains and losses can obscure an one-to-one relationship between ancestral exons and units of protein structure. With this in mind, the formation of novel protein architectures through intronic recombination was reframed in terms of larger structural modules, so-called *domains*, sequences that encode an entire fold and may be encoded by more than one exon.

Over the intervening 40 years, studies have probed the role of exon shuffling in the origins of ancient genes (the "exon theory of genes"; see Ref. 5 for a detailed review) and the evolution of modular protein architectures,[6–10] especially during emergence of metazoan multicellularity.[11] Due to the limited availability of sequence and structural data, early studies focused on the spatial relationship between intron and structural units one gene at a time.[5]

Sequencing of eukaryotic genomes provided a much larger sample of genes with intron/exon structure for such studies. Moreover, the rapid growth of sequence data enabled prediction of domains from multiple sequence alignments, relaxing the need for structural data. In the first genome-scale study, Liu and Grigoriev[7] tested the second prediction, that introns will be preferentially situated near the boundaries of domains, in nine metazoan genomes against a null model of uniformly distributed intron positions. They reported "a striking correlation" between introns and domain boundaries, concluding that "exon shuffling was extensive throughout evolution of eukaryotes and contributed significantly to the complexity of their proteomes". They subsequently examined the evolutionary role of domains flanked by introns at both ends, but did not carry out genome-scale statistical tests.[8] Fifteen years later, Smithers *et al.*[10] applied a similar approach to a larger and more broadly taxonomically distributed set of eukaryotic genomes, and concluded that "domain shuffling … is indisputably found widely across the eukaryotic tree".

These conclusions depend crucially on the assessment of significance of intron-domain boundary agreement, which in turn depends on the use of realistic null models. In early studies,[12] three null models were developed that preserve features of

the genuine data to different extents. Only one, the uniform model, was used in later genome-scale studies.[7,10] The appropriateness of a uniform model of random intron positioning and its influence on the conclusions of the study were not examined.

Here, we investigate how the choice of null model affects the assessment of the exon shuffling hypothesis empirically. We specify test statistics for both predictions, that domains flanked by introns at both ends will be overrepresented and that introns will be preferentially situated at domain boundaries. We extend null models from studies of single genes[12] for use in genome-wide analyses and use them to assess the significance of both test statistics in five metazoan and 16 fungal genomes.

Our empirical results show that these null models vary substantially in their propensity for Type I errors in genome scale studies and the extent to which they preserve the properties of gene architecture. In particular, the widely-used uniform model does not recapitulate exon length distributions, even approximately, and results in highly exaggerated significance estimates. The impact on metazoan genomes is minimal; exon shuffling statistics are significant under all three models. However, statistical tests in fungal genomes are highly sensitive to the choice of model. Moreover, even when highly significant, the effect size in fungal genomes is extremely small. Only 3% of domains, on average, coincide with an intron at both ends. Our results are consistent with conclusions of prior studies that exon shuffling contributed to metazoan, but not fungal genome evolution.[11] Importantly, this work demonstrates the importance of selecting null models that preserve the features of genuine data: more permissive null models may overestimate the significance, leading to incorrect biological conclusions.

2. Models for Testing the Exon Shuffling Hypothesis

Agreement between "the exon-structure of the genes and the domain-organization of proteins"[13] is a source of evidence for exon shuffling. Here, we consider two test statistics that capture different aspects of this correlation, expressed in terms of the relative positions of introns and domain boundaries. For each of these test statistics, we use three different null models to assess the deviation from chance agreement between exon and domain organization.

Our analyses use the following general procedure for all six combinations of test statistic and null model, with one exception discussed below. Let g be a gene in genome $\mathcal{G}$ of length $l(g)$ codons with $K(g)$ exons and $D(g)$ domains, and let T_g be a gene-specific test statistic that quantifies the agreement between introns and domain boundaries in g. We define a genome-wide test statistic $T_{\mathcal{G}} = \sum_{g \in \mathcal{G}} T_g$ to assess this agreement across the genome as a whole.

The expected value of $T_{\mathcal{G}}$ is estimated by repeatedly generating ensembles of randomized intron positions and calculating the genome-wide test statistic for each ensemble. This procedure is repeated for M iterations, resulting in M estimates of the genome-wide test statistic. In this study, $M = 10^8$.

This calculation is carried out on a per-gene basis at each iteration. For each gene $g \in \mathcal{G}$, $K(g) - 1$ random intron positions are generated according to the null model. The value of the gene-specific test statistic for the ith ensemble, $T_g^{(i)}$, is obtained by comparing the simulated intron positions with the true domain boundaries. From these, the ith genome-wide test statistic is calculated: $T_{\mathcal{G}}^{(i)} = \sum_{g \in \mathcal{G}} T_g^{(i)}$.

The expected value of the genome-wide test statistic, $E[T_{\mathcal{G}}]$, is the mean of the M genome-wide test statistics, $\{T_{\mathcal{G}}^{(1)}, \ldots, T_{\mathcal{G}}^{(M)}\}$, generated by this procedure. We then use a χ^2 goodness-of-fit test with one degree of freedom to assess whether the observed coincidence between domain boundaries and introns differs significantly from the coincidence expected under a null model of intron positioning.

Test statistics. We consider two test statistics (Fig. 1) corresponding to two properties that are predicted to facilitate the formation of novel protein architectures by intronic recombination.

The intron test statistic: Recombination in introns located outside sequences that encode protein modules is less likely to disrupt the structural integrity of the protein. According to the exon shuffling theory, the presence of introns separating sequences that encode domains is advantageous for acquisition of novel domain architectures and therefore, there should be an over-representation of domain-flanking introns, that is, introns in domain boundary boxes, defined in what follows.

This property is represented by the number of introns that agree with domain boundaries (T_I). Following Refs. 7, 8 and 10, we define a domain boundary box to be w contiguous amino acids straddling the end of a domain. Two values of w were considered[7,8]: For $w = 20$, the box extends 10 amino acids on either side of the domain boundary. For $w = 6$, the box consists of 5 amino acids outside and 1 amino acid inside the domain boundary. Then, $T_{I,g}$ and $T_{I,\mathcal{G}}$ are defined to be the number of introns located in domain boxes in gene g and genome-wide, respectively.

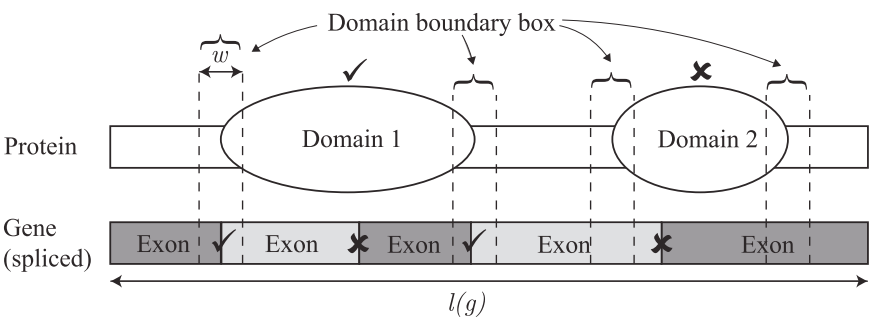

Fig. 1. Calculation of the test statistics T_I and T_D for a hypothetical gene with $K(g) = 5$ exons encoding a protein with $D(g) = 2$ domains. Two introns (exon boundaries) labeled with check-marks fall into a boundary box of domain 1. The other two introns, labeled with cross-marks, do not agree with any domain boundary box in the gene. So, in this example, $T_{I,g} = 2$. Both of the first domains boxes agree with some intron; this is not true for domain 2. Thus, $T_{D,g} = 1$ in this example.

The domain test statistic: Protein modules that are encoded by an integral number of exons are more likely to fold correctly following intronic recombination. That is, if intron-mediated recombination of sequences encoding structural or functional modules plays an important role in the formation of novel protein architectures, then sequences that encode those modules should be flanked by introns at both ends. This property is quantified by the number of domains that contain at least one intron in each domain boundary box (T_D).

2.1. *Null models of intron positions*

Three approaches have been introduced to model the chance intron positions in a single gene: uniform, sampling from the empirical exon length distribution, and permutating exon order.[4] Here, we extend those models for use in genome-wide tests.

The uniform null model. Intron positions in gene g are simulated by sampling $K(g) - 1$ integers uniformly at random from the interval $[0, l(g)]$. This model preserves the number of introns per gene, but not characteristic exon lengths.

For the uniform model, the expected value of the intron test statistic ($T_{I,g}$) can also be estimated from known quantities without resorting to simulation. The probability of an intron falling into a domain boundary box in g is $w(g)$, the fraction of the $l(g)$ residues in g that are within any boundary box in g. The expected number of introns that agree with some domain boundary in g is

$$E[T_{I,g}] = (K(g) - 1)w(g).$$

This expression was used to assess significance in two previous studies.[8,10] Liu and Grigoriev used an analogous expression to calculate a genome-wise intron test statistic directly, without the intermediate step of comparing intron positions with domain boundaries on a per-gene basis.[7] That approach preserves the number of introns in the genome as a whole, but not the number of introns per gene.

The permutation null model. Intron positions in gene g are simulated by permuting the order of the exons in g. All $K(g)!$ permutations are assigned with equal probability. This model preserves both the number of introns and the gene-specific exon lengths. It does not, however, preserve length distributions associated with exon order. The number of permutations grows super-exponentially with the number of exons, with potential complications at both ends of the scale. For some genes, the total number of permutations will be smaller than M, the number of ensembles to be generated, necessitating sampling with replacement. For other genes, the number of permutations will be so large as to require subsampling.

To address these issues, for genes with nine or fewer exons, we first enumerate all permutations and calculate the associated values of T_g. These precalculated values are then sampled with replacement to obtain M gene-specific test statistics. To reduce the computational overhead for genes with more than nine exons, the number of permutations generated is capped at 9!. The associated precalculated test statistics

are then sampled with replacement to obtain M values of T_g. Note that this process still generates M distinct genome-wide ensembles because the same permutation will be combined with different permutations of other genes in each replicate.

The empirical exon length null model. Intron positions in gene g are simulated by sampling $K(g)$ exons of lengths $\{l_1, \ldots, l_{K(g)}\}$ from the genome-wide empirical exon length distribution. Some early studies sampled lengths from a lognormal distribution with mean and standard deviation calculated from the empirical data.[4] With the genome-scale data sets now available, exon lengths can be modeled directly by the genome-wide empirical distribution. This model preserves the number of introns and approximately preserves the genome-wide distribution of exon lengths. Like the permutation model, it does not account for differences in exon lengths at different ordinal positions in the gene.

To ensure that the simulated gene length agrees with the actual gene length, the lengths are scaled by a factor of $l(g)/\hat{l}$, where $\hat{l} = \sum_{k=1}^{K(g)} l_k$. The resulting distribution of scaled exon lengths will deviate from the empirical distribution from which the lengths were originally sampled. This deviation can be mitigated by repeatedly sampling sets of $K(g)$ exons until their combined length is close to the actual gene length, but at a considerable increase in running time.

To balance these needs, we introduce a procedure where the tradeoff between accuracy and performance is controlled by three adjustable parameters, θ, τ and M'. For each $g \in \mathcal{G}$, sets of $K(g)$ lengths are sampled repeatedly from the empirical exon length distribution until either $|\hat{l} - l(g)| < \theta$ or the number of tries reaches τ. Upon termination, the sample is scaled and added to the set of ensembles for g. We determined empirically that $\tau = 20$ and $\theta = 100$ represent a reasonable tradeoff.

We further limit the computational costs by sampling fewer than M ensembles per gene and apply a memoization strategy similar to that used for the permutation model. For each gene, $M' < M$ ensembles of randomized intron positions are sampled in advance and $T_{I,g}$ and $T_{D,g}$ are calculated for each ensemble. In this study, the number of ensembles is limited to $M' = 10{,}000$. Next, M different genome ensembles are generated by sampling with replacement from the M' precomputed test statistics for each gene. The choice of $M' = 10{,}000$ improves speed without unduly compromising the statistical power: for all genes with seven or fewer exons, the empirical model offers more statistical power than the permutation model, since $10{,}000 > 7!$. This is especially relevant for intron-poor organisms. In almost all fungal species examined here, more than 80% genes examined have seven exons or fewer.

3. Results

In order to determine how model choice influences conclusions about the exon shuffling hypothesis, we examined the evidence empirically using both the intron-based and the domain-based test statistics and all three null models discussed in the previous section. The test statistics were calculated with two different domain box sizes used in prior studies[7,8]: $w = 20$ and 6.

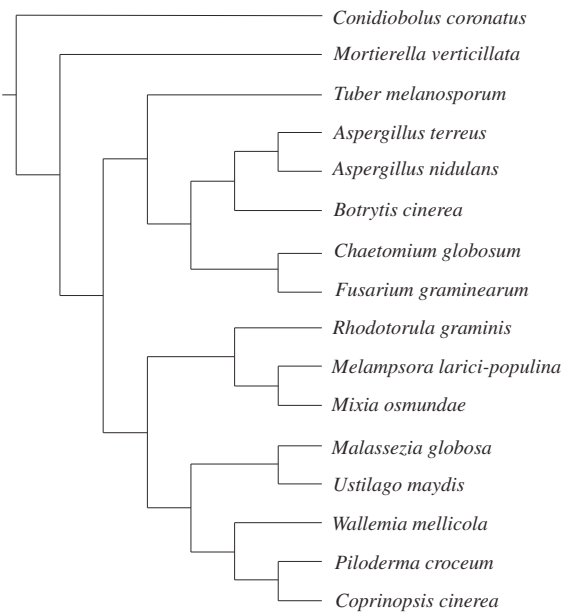

Fig. 2. Tree of the 16 fungal species in this study, adapted from Spatafora *et al.*[17]

We analyzed the five of the nine metazoan genomes originally analyzed by Ref. 7, as well as 16 fungal genomes (Fig. 2). Fungal genomes were selected for this study because they possess a broad range of intron sizes and frequencies,[14,15] ranging from 0.4 to 4.7 introns per gene in our dataset. Gene model coordinates were downloaded from NCBI, Ensembl, and JGI (Table S1). Domain predictions were extracted from the SUPERFAMILY 2 database.[16] Following Refs. 7 and 8, genes with at least one intron and at least one annotated domain that does not coincide with the first or last w amino acids of the protein, were considered.

3.1. *The uniform model generates unrealistic exon lengths*

Testing the exon shuffling hypothesis requires models of chance intron positions that are consistent with intron-exon structure in actual genes. To assess the suitability of the null models used here, we compared the exon length distributions generated by the uniform and empirical models to the genuine exon length distribution (Fig. 3). (Exon lengths in ensembles simulated by permutation are the same as the genuine data and were not included in this comparison.)

Visual inspection shows that while neither model preserves the genuine exon length distribution (Fig. 3(a)), the deviation is much greater for the uniform model. A quantitative comparison using the Kolmogorov–Smirnov (KS) distance, a measure of the difference between two cumulative distribution functions with range [0,1], indicates that the uniform model provides a much poorer fit than the empirical model for all 21 genomes analyzed (Fig. 3(b)).

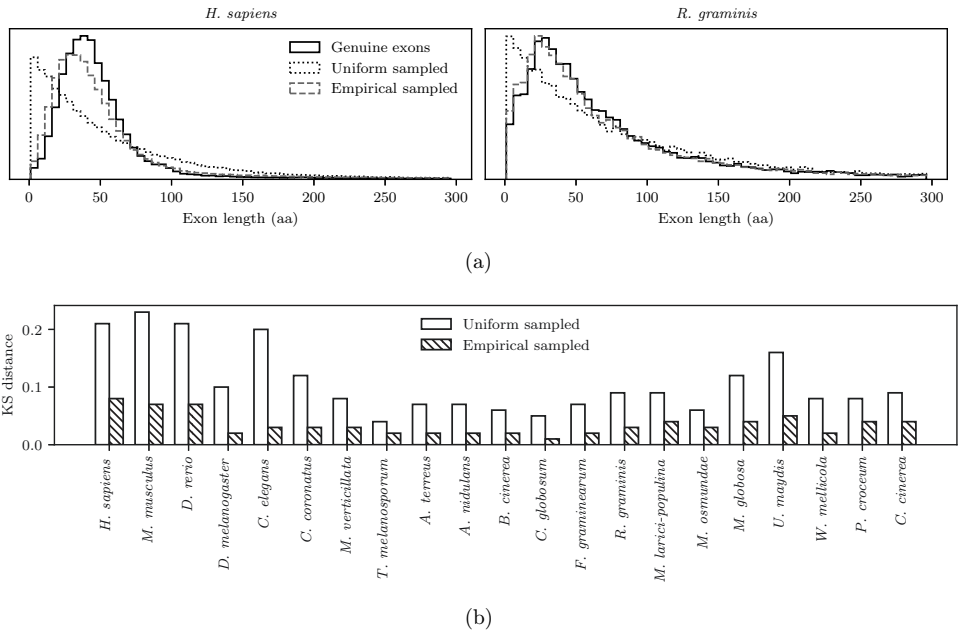

Fig. 3. (a) Histograms of exon lengths from the genuine data and two simulation models for a representative metazoan and a representative fungal genome. (b) KS distances between the genuine exon length distribution and the exon length distributions implied by the uniform model (white bars) and the empirical model (hatched bars), respectively.

Figure 3 provides an assessment of the agreement between the genome-wide distributions of simulated and genuine exon lengths, but does not address the possibility that exon lengths might differ in different classes of genes or domains. A violation of the assumption that the genome-wide distribution is an appropriate model for all genes could be particularly problematic if mobile or promiscuous domains are encoded by exons with a different characteristic length distribution.

However, visual comparison of exon sizes across proteins with varying numbers of domains (Fig. S1) suggests that they are not dramatically different and this is not a huge factor in this data set. Another possibility not addressed is that exons lengths might differ at different positions in the exon–intron structure. Indeed, empirical evidence suggests that exon length distributions vary with ordinal position in the gene and that this effect varies across taxonomic lineages.[18] None of the models used in this study account for a possible interaction between exon position and exon length.

3.2. *Estimates of chance agreement are highly sensitive to model choice*

We next asked how the choice of null model influences the significance of genome-wide intron-domain boundary agreement.

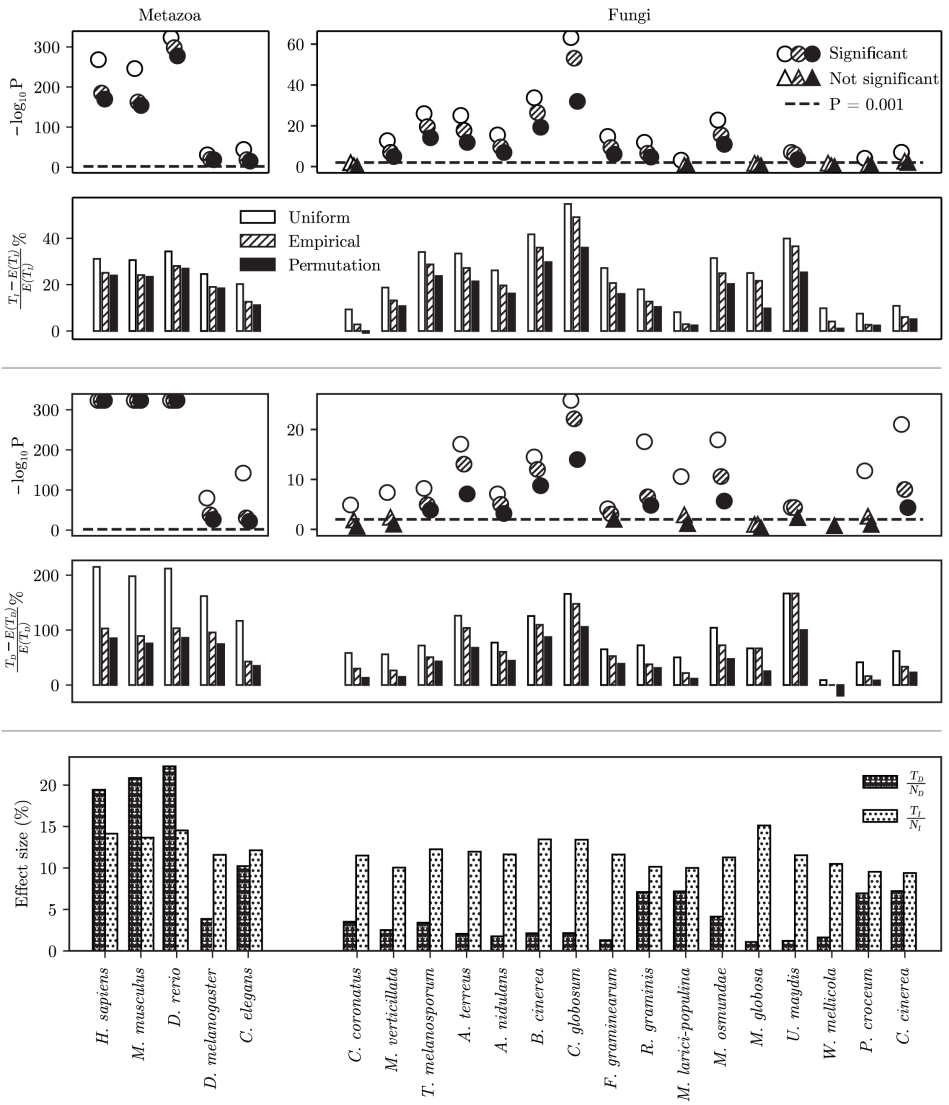

Fig. 4. χ^2 Goodness-of-Fit ($-\log_{10}(p)$) and relative difference ($(T-E)/E$) for the intron-based (top) and domain-based (middle) test statistics ($w = 20$), with the effect sizes (bottom). Dashed lines indicate significance threshold ($\alpha = 0.001$). Circles and triangles indicate p-values that are significant and not significant, respectively, at the α level. N_D and N_I are the total number of domains and introns in $\mathcal{G}$, respectively. The smallest p-value obtainable with the R function `pchisq()` is 5E-324.

The numerical range of p-values obtained with the three models is astounding. The uniform null model consistently yields the smallest expected agreement, the most stringent p-values, and the greatest significance. Compared with the most conservative estimate (obtained with the permutation model), p-values inferred with the uniform model can differ by ten to ~ 100 orders of magnitude (Fig. 4).

For the five metazoan genomes, the p-values are so extreme that the choice of null model is of minor importance; for both test statistics and especially T_D, intron-domain boundary agreement in these genomes is highly significant under all null models. Most fungal genomes, however, have much less stringent p-values. Intron-domain boundary agreement is not significant under all null models in three fungal genomes with respect to T_I and two for T_D. Importantly, for T_I, at a significance threshold of 0.001, three species have significant agreement under the uniform model, but are not significant with the other null models. This is also the case for four fungal genomes when T_D is considered. The choice of null model leads to different conclusions for three genomes for the smaller box size ($w = 6$, Fig. S2) as well. Thus, for fungal genomes, the choice of null model not only leads to different numerical values, but potentially can result in different biological conclusions.

3.3. *Few domains are flanked by introns*

Statistical significance provides a measure of the frequency of intron-domain boundary agreement relative to chance expectations, but does not tell us how important this agreement is to the evolution of novel protein architectures. As an assessment of the size of this effect, the fraction of introns that are associated with a domain boundary is consistently modest, ranging from 9% to 15% across the 21 genomes studied (Fig. 4).

We further asked what fraction of domains coincide with an intron at both ends. In metazoa, on average, 21% of all domain instances are flanked by introns in the vertebrates and 7% in the invertebrates. In contrast, in the fungal genomes tested, on average, only $\sim 3\%$ of all domain instances are flanked by introns at both ends.

4. Discussion

The exon shuffling hypothesis, later recast in terms of domain shuffling, makes two predictions: (1) sequences that fold independently will be preferentially encoded by an integral number of exons (i.e. will be flanked by introns at both ends) and (2) introns will tend to be located outside of the sequences that encode these modules. One commonly used strategy for testing this hypothesis is to assess the frequency of introns positioned near structural boundaries against chance models of intron positioning. Although several null models of intron positioning have been proposed for the analysis of individual genes,[4,12,19,20] genome-scale studies have only considered uniformly distributed introns, leaving the choice of null model out of the discussion. In addition, genome-scale studies that have used this approach have focused exclusively on the second prediction.[7,10] The importance of domains flanked by introns to multidomain evolution has been discussed,[8] but not tested statistically.

In this study, we evaluated the performance of three null intron position models empirically in five metazoan and 16 fungal genomes. We considered both predictions of the exon shuffling hypothesis: In addition to the widely used intron test statistic,

we introduced a new statistic representing domains that are flanked by introns at both ends. The models were compared with respect to their propensity to reject the null hypothesis, the extent to which they preserve gene and genome properties, and ease of computation.

Preservation of gene and genome features: Generally speaking, statistical hypothesis testing can be compromised when the null model preserves too few aspects of the genuine data. Our results show that the exon length distributions generated by uniformly distributed intron positions deviate greatly from genuine exon length distributions, with an unrealistic excess of very short exons (Fig. 3). The genomic exon length distribution is preserved exactly by the permutation model and approximately by the empirical model, although some distortion is introduced by the length adjustment required to keep the gene length constant. We did not probe the accuracy of the models in reproducing exon length distributions in different classes of genes or domains, although we do observe that, in this data set, exon length distributions do not vary greatly with the number of domains encoded. Exon length distributions are known to vary with exon position,[18] a phenomenon not accounted for by any of the models tested here. The importance of models that capture exon length variation on a finer scale warrants further investigation.

Propensity to reject the null hypothesis: The models differ in how well they recapitulate genuine exon lengths. As might be expected, they also differ in their assessment of significance. Indeed, the impact of model choice on p-values is dramatic. Despite this enormous variation (up to 100 orders of magnitude in our data), the variation in p-values may have little impact on exon shuffling tests in metazoa. In all metazoan genomes tested, intron-domain boundary agreement is significant with both test statistics and with all three models. In contrast, in fungal genomes, model choice has a real impact. For almost half (7/16) of the fungal genomes studied, using different null models leads to different conclusions with at least one of the two test statistics used (Fig. 4).

Ease of computation: The models used in this study were originally designed for per-gene statistical tests. Extending these models for genome scale simulation required developing heuristics to mitigate the computational burden associated with genome-scale sampling without unduly compromising the properties of the model. Our permutation-based randomization procedure accounts for the wide variation in the number of possible permutations across genes, depending on intron count. The empirical model required a randomization strategy that satisfies the constraint that sampled exons must agree with the gene length, but also preserves typical exon lengths. With the uniform model, simulation is required to estimate $E[T_D]$, the expected number of domains flanked by introns. Exceptionally, the expected number of introns associated with domain boundaries ($E[T_I]$) can be calculated analytically, allowing for rapid determination of significance using a χ^2 goodness of fit test. This may be why genome-scale statistical tests have used the intron statistic with the uniform model exclusively. Despite this computationally compelling advantage, our

empirical results suggest that, outside metazoa, the use of the uniform model can compromise the integrity of the analysis.

Possible confounding factors: The results presented here could be influenced by several factors that we did not consider in this study. Paralogous genes with similar domain content and intron–exon structure could distort the signal through a "double counting" effect; see, for example, Refs. 6 and 21. We did not correct for duplicated genes, consistent with the studies that inspired this work,[7,10] which allowed for comparisons with the results of those studies. Correction for paralogy should be carried out for any definitive study of exon shuffling as an evolutionary mechanism. Another potential source of error arises from misannotation of gene models and domains,[22] especially in the more recently sequenced fungal genomes. Finally, some pairs of genomes in this study are too closely related to provide independent assessments of the coincidence of features in gene and protein architectures; a phylogenetic correction is needed to discount results from closely related species.

Other gene and protein properties: This study is focused on the accuracy of null models for testing the coincidence between the architectural features of gene and protein sequences. Other types of biological information can contribute to an understanding of the role of exon structure in promoting the emergence of new proteins. For example, a comprehensive test of the exon shuffling hypothesis should also consider intron phase and exon symmetry.[9,13,23–26] In another example, Smithers *et al.*[10] examined protein age and the presence of disordered regions, providing contextual information about when and how new protein coding genes emerge.

Impact of model choice on biological interpretation: Large sample sizes can lead to highly significant associations even when the number of such associations is quite small. This is the case in our analysis: in fungal genomes, the percentage of domains that are flanked by introns is tiny (3% on average). The number of introns found at a domain boundary is larger, although still modest, and not markedly different in metazoa and fungi. This observation suggests several hypotheses. Exon shuffling may have played an important role in metazoan, but not fungal, evolution, as has previously been suggested.[11] Alternatively, this could indicate that while intronic recombination contributes to the evolution of novel domain architectures, domains are not the "integrally folded protein units" that are shuffled by this process or that shuffling is more resilient to imprecise boundaries than originally hypothesized.[2] A variant on this explanation is that flanking introns do contribute to domain mobility, but only in a small number of domain families.

Another possibility is that the weak association between introns and domain boundaries arises for reasons unrelated to exon shuffling. Other forces acting on gene and/or protein architecture may drive the juxtaposition of introns and domain boundaries. For example, intron positions may be constrained by the requirements of the splicing machinery.[27] Similarly, the foldability requirement may constrain domain lengths and, by extension, the locations of domain boundaries. Constraints such as these could result in exon and domain length distributions that are under-

dispersed, which in turn could increase the chance probability of introns in close proximity to domain boundaries.

In summary: We observe that the uniform null model widely-used for testing the exon-shuffling hypotheses results in a highly skewed estimate of the exon length distribution. This, in turn, leads to exaggerated assessments of statistical significance. Modeling intron positions by sampling from the empirical distribution or permuting exon order results in a much more realistic distribution of intron positions.

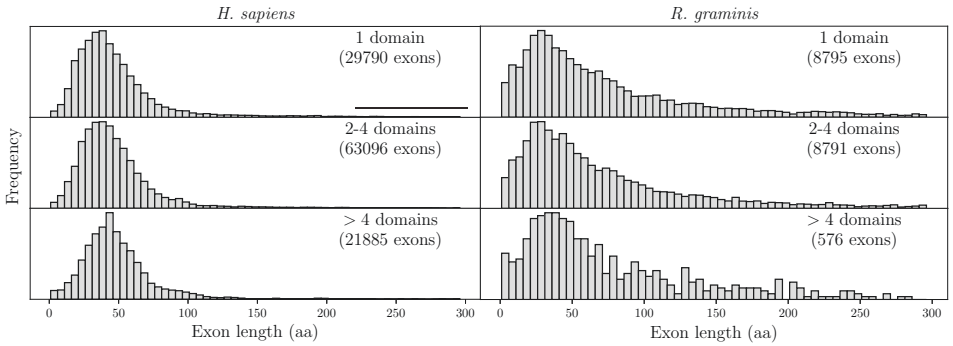

Fig. S1. Distribution of exon lengths in proteins with one domain, two to four domains, and more than four domains. (left) Distribution in human, representative species for metazoa. (right) Distribution in *R. graminis*, representative species for fungi.

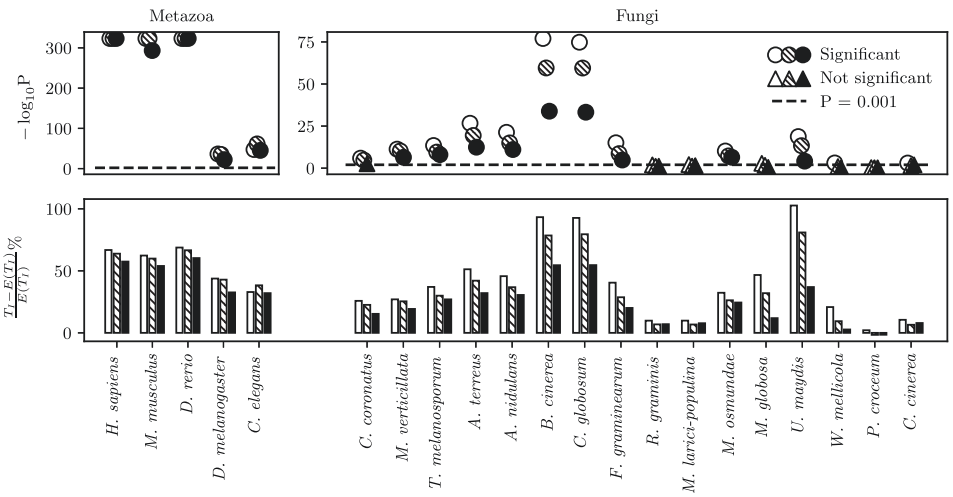

Fig. S2. p-values (top) and relative differences (bottom) of intron-domain boundary agreement with respect to T_I, with window size of $w = 6$ under the uniform per gene model, empirical exon length distribution model, and permutation model. Dashed line indicates threshold for significance; circles and triangles indicate p-values that are significant and not significant, respectively. The minimal p-values obtainable with the `pchisq()` function in R is 5E-324.

Table S1. Genomes used in this study. All metazoan genome annotations were downloaded from Ensembl release 69.

Genome	GenBank accession	Exon/gene	Domain/gene	GC%
Conidiobolus coronatus NRRL28638[28]	GCA_001566745.1	4.7	0.78	31.9
Mortierella verticillata	GCA_000739165.1	5.2	0.90	48.9
Tuber melanosporum Mel28[29]	GCA_000151645.1	5.4	0.81	44.9
Aspergillus terreus NIH2624	GCA_000149615.1	4.0	0.95	52.4
Aspergillus nidulans FGSC A4[30]	GCA_000149205.2	4.0	0.92	50.2
Botrytis cinerea B05.10[31]	GCA_000143535.4	3.8	0.53	42.4
Chaetomium globosum CBS 148.51[32]	GCA_000143365.1	4.0	0.80	55.4
Fusarium graminearum PH-1[33]	GCA_000240135.3	3.6	0.77	48.3
Rhodotorula graminis WP1[34]	GCA_001329695.1	7.9	0.85	67.5
Melampsora larici-populina 98AG31[35]	GCA_000204055.1	7.5	0.41	41.3
Mixia osmundae IAM 14324[36]	GCA_000708205.1	5.6	0.83	55.5
Malassezia globosa CBS 7966[37]	GCA_000181695.1	2.8	0.97	52.0
Ustilago maydis 521[38]	GCA_000328475.2	3.2	0.91	54.0
Wallemia mellicola CBS 633.66[39]	GCA_000263375.1	4.5	0.92	40
Piloderma croceum F 1598[40]	GCA_000827315.1	8.1	0.51	46.6
Coprinopsis cinerea okayama7#130[41]	GCA_000182895.1	8.0	0.66	51.6

Application of these models is more costly, computationally, but we suggest that their use is essential to obtain accurate statistical tests outside of fungi.

Further, in probing effect size, we observe that the number of domains flanked by introns is modest in metazoa and vanishingly small in fungi. Taken together, these observations lead us to question whether exon shuffling is really widespread across the eukaryotic tree.[10] Additional investigations of exon shuffling in eukaryotic lineages outside of metazoa are an exciting area for future work.

Acknowledgments

We thank Arlin Stoltzfus for providing details on how the original gene-specific tests were implemented and four anonymous reviewers for biological insights and technical suggestions. This work was supported in part by NSF Grants DBI-1838344 and DBI-1759943. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

1. Gilbert W, Why genes in pieces? *Nature* **271**(5645):501–501, 1978.
2. Blake CC, Do genes-in-pieces imply proteins-in-pieces? *Nature* **273**(5660):267–267, 1978.
3. Blake C, Exons and the evolution of proteins, *Int Rev Cytol* **93**:149–185, 1985.
4. Stoltzfus A, Spencer DF, Zuker M, Logsdon JM, Doolittle WF, Testing the exon theory of genes: The evidence from protein structure, *Science* **265**(5169):202–207, 1994.
5. Roy SW, Recent evidence for the exon theory of genes, *Genetica* **118**(2):251–266, 2003.

6. Vibranovski MD, Sakabe NJ, De Oliveira RS, de Souza SJ, Signs of ancient and modern exon-shuffling are correlated to the distribution of ancient and modern domains along proteins, *J Mol Evol* **61**(3):341–350, 2005.
7. Liu M, Grigoriev A, Protein domains correlate strongly with exons in multiple eukaryotic genomes-evidence of exon shuffling? *Trends Genet* **20**(9):399–403, 2004.
8. Liu M, Walch H, Wu S, Grigoriev A, Significant expansion of exon-bordering protein domains during animal proteome evolution, *Nucleic Acids Res* **33**(1):95–105, 2005.
9. França GS, Cancherini DV, de Souza SJ, Evolutionary history of exon shuffling, *Genetica* **140**(4):249–257, 2012.
10. Smithers B, Oates M, Gough J, why genes in pieces? revisited, *Nucl Acids Res* **47**(10):4970–4973, 2019.
11. Patthy L, Exon shuffling played a decisive role in the evolution of the genetic toolkit for the multicellular body plan of metazoa, *Genes* **12**(3):382, 2021.
12. Stoltzfus A, Spencer DF, Doolittle WF, Methods for evaluating exon–protein correspondences, *Bioinformatics* **11**(5):509–515, 1995.
13. Patthy L, Genome evolution and the evolution of exon-shufflinga review, *Gene* **238**(1):103–114, 1999.
14. Croll D, McDonald BA, Intron gains and losses in the evolution of *Fusarium* and Cryptococcus fungi, *Genome Biol Evol* **4**(11):1148–1161, 2012.
15. Stajich JE, Dietrich FS, Roy SW, Comparative genomic analysis of fungal genomes reveals intron-rich ancestors, *Genome Biol* **8**(10):1–13, 2007.
16. Pandurangan AP, Stahlhacke J, Oates ME, Smithers B, Gough J, The SUPERFAMILY 2.0 database: A significant proteome update and a new webserver, *Nucl Acids Res* **47**(D1):D490–D494, 2019.
17. Spatafora JW *et al.*, A phylum-level phylogenetic classification of zygomycete fungi based on genome-scale data, *Mycologia* **108**(5):1028–1046, 2016.
18. Zhu L, Zhang Y, Zhang W, Yang S, Chen JQ, Tian D, Patterns of exon-intron architecture variation of genes in eukaryotic genomes, *BMC Genom* **10**:47, 2009.
19. Gilbert W, Glynias M, On the ancient nature of introns, *Gene* **135**(1–2):137–144, 1993.
20. Weber K, Kabsch W, Intron positions in actin genes seem unrelated to the secondary structure of the protein, *EMBO J* **13**(6):1280–1286, 1994.
21. Cancherini DV, França GS, de Souza SJ, The role of exon shuffling in shaping protein-protein interaction networks, *BMC Genom* **11**(5):1–13, 2010.
22. Nagy A, Patthy L, Fixpred: A resource for correction of erroneous protein sequences, *Database* **2014**, 2014.
23. De Souza SJ *et al.*, Toward a resolution of the introns early/late debate: Only phase zero introns are correlated with the structure of ancient proteins, *P Natl Acad Sci USA* **95**(9):5094–5099, 1998.
24. Kaessmann H, Zöllner S, Nekrutenko A, Li WH, Signatures of domain shuffling in the human genome, *Genome Res* **12**(11):1642–1650, 2002.
25. Long M, Rosenberg C, Gilbert W, Intron phase correlations and the evolution of the intron/exon structure of genes, *P Natl Acad Sci USA* **92**(26):12495–12499, 1995.
26. Patthy L, Exon shuffling and other ways of module exchange, *Matrix Biol* **15**(5):301–310, 1996.
27. De Kee DW, Gopalan V, Stoltzfus A, A sequence-based model accounts largely for the relationship of intron positions to protein structural features, *Mol Biol Evol* **24**(10):2158–2168, 2007.
28. Chang Y *et al.*, Phylogenomic analyses indicate that early fungi evolved digesting cell walls of algal ancestors of land plants, *Genome Biol Evol* **7**(6):1590–1601, 2015.

29. Martin F *et al.*, Périgord black truffle genome uncovers evolutionary origins and mechanisms of symbiosis, *Nature* **464**(7291):1033–1038, 2010.
30. Galagan JE *et al.*, Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*, *Nature* **438**(7071):1105–1115, 2005.
31. Amselem J *et al.*, Genomic analysis of the necrotrophic fungal pathogens *Sclerotinia sclerotiorum* and *Botrytis Cinerea*, *PLoS Genet* **7**(8):e1002230, 2011.
32. Cuomo CA, Untereiner WA, Ma LJ, Grabherr M, Birren BW, Draft genome sequence of the cellulolytic fungus *Chaetomium globosum*, *Genome Announc* **3**(1):e00021–15, 2015.
33. Ma LJ *et al.*, Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*, *Nature* **464**(7287):367–373, 2010.
34. Firrincieli A *et al.*, Genome sequence of the plant growth promoting endophytic yeast *Rhodotorula graminis* WP1, *Front Microbiol* **6**:978, 2015.
35. Duplessis S *et al.*, Obligate biotrophy features unraveled by the genomic analysis of rust fungi, *P Natl Acad Sci USA* **108**(22):9166–9171, 2011.
36. Toome M *et al.*, Genome sequencing provides insight into the reproductive biology, nutritional mode and ploidy of the fern pathogen *Mixia osmundae*, *New Phytol* **202**(2):554–564, 2014.
37. Xu J *et al.*, Dandruff-associated malassezia genomes reveal convergent and divergent virulence traits shared with plant and human fungal pathogens, *P Natl Acad Sci USA* **104**(47):18730–18735, 2007.
38. Kämper J *et al.*, Insights from the genome of the biotrophic fungal plant pathogen *Ustilago maydis*, *Nature* **444**(7115):97–101, 2006.
39. Padamsee M *et al.*, The genome of the xerotolerant mold *Wallemia sebi* reveals adaptations to osmotic stress and suggests cryptic sexual reproduction, *Fungal Genet Biol* **49**(3):217–226, 2012.
40. Kohler A *et al.*, Convergent losses of decay mechanisms and rapid turnover of symbiosis genes in mycorrhizal mutualists, *Natu Genet* **47**(4):410–415, 2015.
41. Stajich JE *et al.*, Insights into evolution of multicellular fungi from the assembled chromosomes of the mushroom *Coprinopsis cinerea* (*Coprinus cinereus*), *P Natl Acad Sci USA* **107**(26):11889–11894, 2010.

Xiaoyue Cui has a BS in Biological Sciences from Tsinghua University and a MS in Computational Biology from Carnegie Mellon University. She is currently a Ph.D. student in Computational Biology at Carnegie Mellon University. She uses statistical methods to study the evolution of multidomain proteins.

Maureen Stolzer received her Ph.D. degree in Biological Sciences from Carnegie Mellon University, USA, in 2011. She is currently a Research Scientist in the Department of Biological Sciences at Carnegie Mellon University.

Stolzer leads the Notung software development team and is the author of Notung-DM software, which reconciles trees on three levels of biological organization: domains, genes and species. Her research focuses on the intertwining evolution of entities at multiple levels of biological organization and the evolution of multi-domain proteins.

Dannie Durand has a BS in Physics from MIT and a Ph.D. in Computer Science from Columbia University. Following an NSF Nato postdoctoral fellowship at INRIA in Rennes, France, she worked as a Member of Technical Staff at Bellcore in Morristown, NJ. She began to work on biological problems in 1995, with Warren Ewens at the University of Pennsylvania and Lee Silver at Princeton, under the aegis of a Sloan Fellowship.

In 2000, Durand joined the Biological Sciences faculty at Carnegie Mellon University in Pittsburgh, where she is also a member of the Center for Evolutionary Biology and Medicine. She was awarded a David and Lucile Packard Foundation Fellowship in 2001. Durand's research focuses on the emergence of novel genes via gene duplication, horizontal transfer and domain shuffling. Durand and her team develop and distribute the NOTUNG phylogenetic reconciliation software package. NOTUNG is used worldwide to investigate the evolution of gene families in the context of species evolution. Durand is a member of the Quest for Orthologs Consortium, an international community standards organization.