BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

Modeling the evolution dynamics of exon-intron structure with a general random fragmentation process

Liya Wang[1*] and Lincoln D Stein[1,2,3*]

Abstract

Background: Most eukaryotic genes are interrupted by spliceosomal introns. The evolution of exon-intron structure remains mysterious despite rapid advance in genome sequencing technique. In this work, a novel approach is taken based on the assumptions that the evolution of exon-intron structure is a stochastic process, and that the characteristics of this process can be understood by examining its historical outcome, the present-day size distribution of internal translated exons (exon). Through the combination of simulation and modeling the size distribution of exons in different species, we propose a general random fragmentation process (GRFP) to characterize the evolution dynamics of exon-intron structure. This model accurately predicts the probability that an exon will be split by a new intron and the distribution of novel insertions along the length of the exon.

Results: As the first observation from this model, we show that the chance for an exon to obtain an intron is proportional to its size to the 3rd power. We also show that such size dependence is nearly constant across gene, with the exception of the exons adjacent to the 5′ UTR. As the second conclusion from the model, we show that intron insertion loci follow a normal distribution with a mean of 0.5 (center of the exon) and a standard deviation of 0.11. Finally, we show that intron insertions within a gene are independent of each other for vertebrates, but are more negatively correlated for non-vertebrate. We use simulation to demonstrate that the negative correlation might result from significant intron loss during evolution, which could be explained by selection against multi-intron genes in these organisms.

Conclusions: The GRFP model suggests that intron gain is dynamic with a higher chance for longer exons; introns are inserted into exons randomly with the highest probability at the center of the exon. GRFP estimates that there are 78 introns in every 10 kb coding sequences for vertebrate genomes, agreeing with empirical observations. GRFP also estimates that there are significant intron losses in the evolution of non-vertebrate genomes, with extreme cases of around 57% intron loss in *Drosophila melanogaster*, 28% in *Caenorhabditis elegans*, and 24% in *Oryza sativa*.

Keywords: Evolution of exon-intron structure, General random fragmentation process, Simulation

Background

Most eukaryotic genes contain spliceosomal introns, which are removed from mRNA after transcription by the RNA splicing apparatus. The biological origins of introns are uncertain. Since the discovery of introns, there has been significant debate as to whether introns in modern-day organisms were inherited from a common, ancient ancestor, the intron-early hypothesis [1-3], or whether they appeared in genes more recently in the evolutionary process, the intron-late hypothesis [4,5], or indeed whether they result from a mixed model [6,7]. The mixed model suggests that most introns were gained very early in the evolution of eukaryotic genes, followed by intron loss/gain during the course of eukaryotic diversification. The details of such processes, however, remain elusive.

* Correspondence: wangli@cshl.edu; lincolnstein@gmail.com
[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA
[2]Ontario Institute for Cancer Research, 101 College St. Suite 800, Toronto, ON M5G0A3, Canada
Full list of author information is available at the end of the article

One way to understand the process is to examine the size distribution of internal translated exons, referring to exons that are fully translated and referred to as *itexon* in [8]. However, to avoid introducing an unfamiliar term, we will simply refer to itexons as "exon" within this communication unless specified. Although the distribution of exons is just a snapshot of the present day world, fitting it to a model based on well characterized mathematical functions may provide insights into the evolution of exon-intron structure. In a previous study, Gudlaugsdottir et al. [9] have suggested that the size distribution of exons can be approximated with a combination of Weibull [10] and exponential distributions. The Weibull distribution is a particular case of the generalized extreme value distribution. It is widely used in survival analysis, describing the size of particles generated by grinding, milling and crushing operations, etc. The exponential distribution can be used to describe the length of intervals between uniformly distributed points. Therefore, GJudlaugsdottir et al. hypothesized that the exponential distribution is the outcome of random insertion of introns (intron-late). However, they then related the Weibull distribution to the intron-early theory without providing a stochastic model that explains the observed distribution.

In later work, Ryabov and Gribskov [11] showed that a combination of two lognormal distributions gives the best fit quality to the size distribution of exons. The lognormal distribution could result from a random Kolmogoroff fractioning process [12], which assumes that the chance of fragmentation is independent of exon size. Inserting an intron into an exon is equivalent to fragmentation (splitting) of the exon. Therefore, they hypothesized that the process of intron insertion is independent of exon size.

On the other hand, Tenchov and Yanev [13] demonstrated that the Weibull distribution could result from a uniform random fragmentation process. Here, "uniform" means that the chance of fragmentation is linearly proportional to the size of the particle (or exon). Under certain conditions, the resulted Weibull distribution is indistinguishable from lognormal distribution. Hence they concluded that the model of random fragmentation could not be inferred based on the basis of fit quality. Therefore, the hypothesis that intron insertion is independent of exon size is debatable.

One assumption made in the exon size based approaches [9,11] is that introns are inserted into exon randomly. The notion of random insertion of intron has also been proposed based on the analysis of intron distribution in ancient paralogs [14]. Others have argued that there exist certain favored sites for intron insertion - the so-called proto-splice sites [4,15,16]. Unfortunately, none of the size-based approaches provide evidence to support the assumption of random intron insertion.

In this work, we aim to revisit these competing hypotheses by addressing the following open questions: Do longer exons have an increased chance of gaining a new intron? For intron gain events, will the intron be inserted into exon randomly or at some proto-splice sites? Is there an intron gain/loss bias? Are intron insertion events independent of each other? Is there a common mechanism to explain intron gain/loss in different species? In order to answer these and other related questions, we propose a General Random Fragmentation Process (GRFP) to characterize the evolution dynamics of exon-intron structures. The parameters of GRFP are determined by combining simulation and analysis of real genomic data.

Methods

GRFP model

The model of GRFP is motivated by generalizing both Kolmogoroff fractioning process and the uniform random fragmentation process. In GRFP, the probability for an exon to split (gaining an intron) is assumed to be exponentially proportional to the length of the k-th exon (L_k) as L_k^a. Under such a generalization, the Kolmogoroff fractioning process, in which insertion events are independent of exon length, is a particular case of GRFP with $\alpha = 0$, while the uniform random fragmentation process, in which insertions are linearly proportional to exon length, is another special case with $\alpha = 1$. The generalization not only allows GRFP to model either Kolmogoroff or uniform fragmentation process but also allows it to model the fragmentation process of exons (intron gain) with varying α. In the results section, we will use the empirical size distribution of exons to determine the value of α. GRFP also assumes that introns insert into exons randomly and independently, and these assumptions are confirmed by the analysis of real genome data.

The model of GRFP, illustrated in Figure 1, is summarized below:

1. Given a set of n exons, the opportunity for k-th exon to acquire a new intron is proportional to its size to the α-th power:

$$P_s(K) = L_k^a / \sum_{k=1}^{n} L_k^a \quad (1)$$

2. Within k-th exon, the new intron insertion loci follow a normal distribution:

$$P_I(x) \sim L_k * N(\mu_I, \sigma_I), x \in (0, \quad \mathrm{L_k}) \quad (2)$$

3. Intron gains are independent of each other.

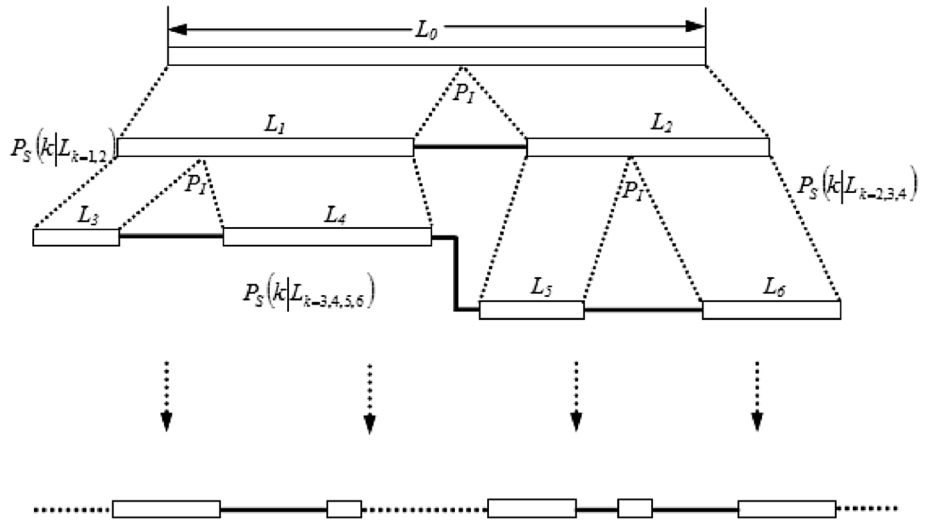

Figure 1 Demonstration of GRFP on splitting a long exon with initial size L_0. The probability of picking which exon to split is proportional to the length of the exon, $P_S(k) \sim L_k^{\alpha}$. The probability of picking an inserting point ($x \epsilon (0, L_k)$) for exon k follows a normal distribution, $P_I(x) \sim L_k * N(\mu_I, \sigma_I)$.

Where $P_S(K)$ denotes the probability for k-th exon to obtain an intron; $P_I(x)$, probability of inserting an intron after x-th position within k-th exon; L_k, length of the exon k; α, dependency value; μ_I and σ_I, mean and standard deviation of the distribution of insertion loci. The model of GRFP has three unknown parameters to be determined, α, μ_I and σ_I.

Simulation testing

We start each simulation with a long exon. The diagram in Figure 1 shows how the splitting of a long exon during evolution is simulated. The diagram shows intron gains in the following order. After inserting the first intron, $P_S(K|L_{k=1,2})$ denotes the picking probability between L_1 and L_2; Assuming L_1 is selected and split by a new intron; the next exon to be split will be chosen from L_3, L_4 and L_2 with probability $P_S(K|L_{k=2,3,4})$; Assuming L_2 is selected, and so on.

For simulations, sequence of pseudorandom number is obtained using the *Mersenne Twister* algorithm [17] implemented in standard MATLAB 7.12. In this study, all of the simulations start with one single exon. This simplifies the simulation, but it does not imply that the evolution of eukaryotic genes always starts with one long exon. Under such simplification, two more parameters are the initial size of starting exon (L_0) and the number of splitting (m). For a given species, we can estimate them from the annotated gene sets.

We evaluate the properties of GRFP using three simulation experiments. In each, we simulate a set of ordered fragments and quantify their statistical characters given different parameters. The three sets of quantifications listed below are used to justify the three assumptions of GRFP respectively for both simulated fragments and real exons.

1. Mean and standard deviation of the size distribution by fitting it with lognormal distribution (equation (3)) or Weibull distribution (equation (4)):

$$dN = \left(N/\left(\sigma_E\sqrt{2\pi}\right)e^{-\left((1nE-\mu_E)/\sqrt{2\sigma_E}\right)^2}\right)d\ \ln E \tag{3}$$

$$dN = \left(Nkz^{k-I}e^{-zk}/\lambda\right)d\ \ln E \tag{4}$$

Where z=E/λ, E is exon length, dN the number of exons in a bin (bin size is 0.1 unless specified), N the amplitudes of the peak; k shape parameter, and λ scale parameter of the Weibull distribution; μ_E the mean position, and σ_E the standard deviation of the lognormal distribution. These and subsequent fittings in this study are performed using the nonlinear Trust-Region-Reflective curve-fitting algorithm [18,19] implemented in MATLAB 7.12. Simulation demonstrates that σ_E is primarily determined by the choice of α (in equation (1)).

2. Mean and standard deviation of the insertion ratio defined below:

$$x_i = L_i/(L_i + L_{i+1}) \tag{5}$$

Where L_i and L_{i+1} are the length of two adjacent fragments (exons). This is an indirect estimation of insertion loci (x in equation (2)). Both simulation and real genome data indicate that x follows a normal distribution (with a standard deviation σ_x).

3. Correlation between x_i and x_j defined by equation (6):

$$\rho(i,j) = \frac{\sigma^2_{x_i+x_j} - \sigma^2_{x_i} - \sigma^2_{x_j}}{2\sigma_{x_i}\sigma_{x_j}} \quad (6)$$

Where σ_x is estimated from fitting the histograms of ratio x with a normal distribution. In theory, $x_i + x_j$ is still normally distributed, and the mean value is the sum of the means. However, the variances are not additive if x_i and x_j are correlated. We can estimate the relationship between x_i and x_j with equation (6).

In the first experiment, we examine the relationship between GRFP parameters and the size distribution of the simulated fragments. With fixed μ_I, σ_I, initial size of starting exon (L_0), and the number of splitting (m), one long exon is fragmented with different choices of α. μ_E and σ_E are estimated through fitting a lognormal distribution to the size distribution of the resulted fragments. The correlation between μ_E, σ_E and α is examined. Then, with fixed μ_I, σ_I, and α, the relationship between μ_E, σ_E and initial size of starting exon (L_0), the number of splitting (m) is examined through similar simulations.

In the second experiment, we examine the relationship between real σ_I (in equation (2)) and estimated σ_x (from equation (5)). By fragmenting a long exon, we construct a binary tree to track the splitting process. We classify the adjacent fragments pair (the order is maintained during fragmentation) into four groups based on whether they have the same parent nodes, or if not same parents, comparing their depths. The size distribution of each group and the mixture (equation (5)) is examined. With fixed μ_I, L_0, m, and α, the correlation between σ_I and σ_x is examined by simulations with different choices of σ_I. Then, by coupling with empirical observations, we use Expectation-Maximization (EM) iteration to determine the value of α and σ_I.

In the third experiment, we examined the effects of intron loss on the statistical characters of resulted fragments. By introducing various percentages of intron loss after intron gain, we evaluate how σ_E, σ_x, and $\rho(i,j)$ of the resulted fragments are changed.

Empirical data analysis

In this study, we obtained the cDNA sequences of 14 species (*Homo sapiens* (GRCh37.p8), *Mus musculus* (GRCm38), *Rattus norvegicus* (RGSC3.4), *Danio rerio* (Zv9), *Caenorhabditis elegans* (WBcel215), *Drosophila melanogaster* (BDGP5), *Bos taurus* (UMD3.1), *Pan troglodytes* (CHIMP2.1.4), *Gallus gallus* (WASHUC2), *Sus scrofa* (Sscrofa10.2), *Arabidopsis thaliana* (TAIR10), *Oryza sativa* (MSU6), *Sorghum bicolor* (Sorbi1), *Zea mays* (AGPv2)) from Ensembl and plant Ensembl database [20]. To ensure the quality of the data, we only use the cDNA sequences of protein coding genes with both RefSeq mRNA ID and the known status of both gene and transcript. To examine the size distribution of exons, we extracted the genomic positions of the exons from cDNA sequences to compute exon sizes. We also extracted the genomic positions of the 5′ and 3′ UTRs and used them to identify internal translated exons.

For testing the first assumption of GRFP, we fitted both Weibull and normal distribution to the size distribution of vertebrate exons (logarithm scale). We also grouped exons by positions for testing position bias of intron gain/loss. For the second assumption, we fitted a normal distribution to x (equation (5)). For the third assumption, we computed $\rho(i,j)$ for exon pairs at i-th and j-th position in all protein coding genes. Finally, we examined the differences in the parameters fitted to vertebrate and non-vertebrate species.

Results

Empirical data analysis

Statistical counts of empirical data

Statistical counts of the extracted data are shown in Table 1. The transcript with the longest CDS (Coding DNA Sequence) for each gene is used for counting the number of protein coding genes, number of splitting, and total CDS length. In these counts, a coding gene is excluded if it does not contain any internal translated exons. The total CDS length is the summation of the length of all exons. The number of splitting events is the total number of exons minus one (for reversion of splitting, a long exon can be reconstructed by connecting all exons together). The last two columns of Table 1 are estimated through GRFP simulations that will be discussed later.

In this study, we ignored non-internal translated exons considering the rate of indels (a type of mutations affecting exon size distribution) is significantly lower in the coding region than the non-coding region [21]. It is true that introns can be inserted anywhere, including non-coding exons or even another intron, but their size distribution is confounded by the appearance of more frequent indels.

Size distribution of exons

Figure 2 shows the histograms of exons for eight vertebrate species. Both Weibull (solid line) and normal (dashed line) functions provide a reasonable fit to the histograms of exons, with the fitted parameters shown in Table 2. Notably, in Table 2, the fitted parameters are almost identical across species (e. g., μ_E and σ_E), which might indicate that these vertebrate genomes have undergone a similar stochastic process on the exon-intron structure during evolution. For the six non-vertebrate species, a mixture of two normal functions (dashed line) fits the histograms well (Additional file 1: Figure S1). This

Table 1 Statistical counts of coding genes, splitting (number of exons minus one) and total CDS length (b.p.)

	Number of coding genes	Total CDS length (10^7)	Number of splitting (m, 10^5)	Estimated splitting (m_e, 10^5)	$\frac{m-m_e}{m_e}$
H. sapiens	17275	2.443	1.827	1.901	- 3.9%
M. musculus	16319	2.276	1.705	1.768	- 2.7%
R. norvegicus	17354	2.193	1.722	1.703	1.0%
D. rerio	15068	1.932	1.462	1.501	- 2.7%
G. gallus	5416	0.655	0.537	0.509	5.4%
P. troglodytes	12508	1.694	1.295	1.316	- 1.6%
B. taurus	11948	1.413	1.142	1.099	4.0%
S. scrofa	5498	0.524	0.433	0.408	6.1%
C. elegans	17684	1.833	1.024	1.425	- 28.4%
D. melangaster	8063	1.141	0.383	0.886	- 57.1%
A. thaliana	16547	1.501	1.083	1.167	- 7.2%
O. sativa	23566	2.255	1.329	1.749	- 24.0%
S. bicolor	17769	1.445	1.041	1.123	- 7.3%
Z. mays	15887	1.320	0.987	1.025	- 3.7%

Annotation data for each species is extracted from Ensembl database. Protein coding genes are counted only if they contain at least one internal translated exon. Total CDS length is the summation of all internal translated exon length in these genes. Number of splitting is estimated by the number of internal translated exons minus one. Estimated splitting is determined from GRFP simulation.

might suggest that the evolution of non-vertebrate exon-intron structure has undergone other processes.

In order to assess whether the size distribution of vertebrate exons is position-dependent, we grouped their exons from all protein coding genes according to their positions relative to 5′ UTRs/3′ UTRs. For the five well annotated vertebrates, the standard deviations (σ_E) of the fitted normal functions at each position (e.g. Additional file 1: Figure S2 for *H. sapiens*) are shown in the left panel of Figure 3. The right panel shows corresponding α values calculated using equation (9). The mean values of these distributions are almost constant at all positions (results not shown). Figure 3 shows that σ_E is almost constant for exons across gene body, with exceptions of the first three exons right after 5′ UTR (see solid line), where it increases markedly. For exons next to the 3′ UTR (in dashed line), no similar trend is observed.

These observations suggest that the size distribution of vertebrate exons could be properly fit with either Weibull or normal distribution. The Weibull distribution gives a better fit to both left and right tails (e.g., Additional file 1: Figure S2) because the distribution is skewed to the left. For numerical simulations, we will show that similar size distribution of fragments will be generated based on GRFP model.

Distribution of insertion ratio

For every gene of the selected species, we calculated the insertion ratio x (equation (5)) for each adjacent exon pairs L_i and L_{i+1}. Figure 4 shows the histograms of the ratios for the 14 species. The histograms are fitted well with a normal distribution, and the fitted parameters are shown in Table 3. The fitted parameters are almost identical across vertebrate species with μ_x= 0.5 and σ_x= 0.13. The insertion ratio for non-vertebrates fits a normal distribution reasonably well but with much larger σ_x.

Another interesting observation in Figure 4 is the sharp spike at 0.5, which suggests that there are excessive adjacent exons pairs with the same length. This is consistent with the observation of tandem exon duplication [22]. Because the spikes are located right on the center, mathematically such deviation has little effect on the fitting of the histogram.

The normal function fitted in Figure 4 describes where introns get inserted into an exon. To assess whether it is consistent or against the hypothesis of proto-spliced sites, we calculate the position distribution of four possible proto-splice sites (tested in [23,24]) within human coding sequences, and the results are shown in Additional file 1: Figure S3. The position for each of the four sites (G|G, AG|G, AG|GT and (C/A)AG(A/G)) is calculated by dividing the distance between the intron starting site and start codon by the length of the coding sequence. All coordinates are extracted from Ensembl annotation of *H. sapiens*. The symbol "|" stands for the intron position and "/" indicates wo alternative states of one nucleotide site. Additional file 1: Figure S3 shows that these proto-spliced sites are distributed nearly uniformly within CDS. If introns strongly prefer to be inserted into these sites, the insertion ratio should follow a uniform distribution instead of normal distribution as we observed. Therefore, the analysis here does not favor the proto-splice site hypothesis.

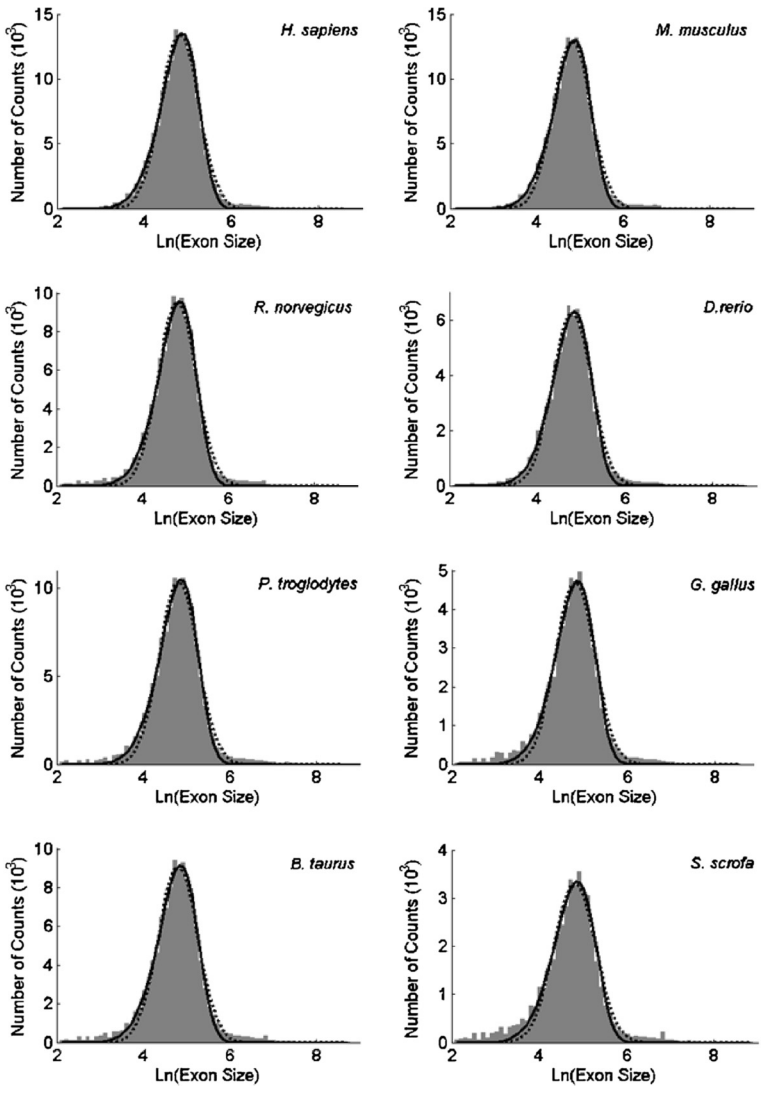

Figure 2 Size distributions of vertebrate exons fitting with normal distribution. The histograms of exons are fitted with a Weibull function (solid line) and normal function (dashed line).

Correlation between insertion ratios

The correlations calculated using equation (6) are shown in Figure 5. For better visualization purpose, we set the correlation between x and itself to zero (the diagonal blocks). The color bar on the right indicates the correlation value between insertion ratio and ratios with i or j distance to it. Figure 5 shows that $\rho(i, j)$ is significantly negative if $|i - j| = 1$. While $\rho(i, j)$ is close to zero for vertebrates if $|i - j| > 1$. The negative correlation between adjacent insertion ratio is not surprising since the calculation uses the same exon length as dividend (equation (5)) but with opposite signs. For example, considering three exons in order with length p, L-p, and q, the insertion ratios are $x_1 = p/(p+L-p) = p/L$ and $x_2 = (L-p)/(L-p+q) \approx 1-p/L$ if $p \approx q$; Both x_1 and x_2 are proportional to p but with different signs.

Table 2 Fitted parameters for size distributions of observed vertebrate exons

	Weibull		Normal	
	λ	κ	μ_E	σ_E
H. sapiens	2.81	6.98	4.81	0.432
M. musculus	2.81	7.01	4.82	0.431
R. norvegicus	2.80	6.89	4.81	0.437
D. rerio	2.79	6.87	4.79	0.437
G. gallus	2.80	6.80	4.81	0.442
P. troglodytes	2.81	6.87	4.81	0.440
B. taurus	2.80	6.69	4.80	0.449
S. scrofa	2.82	6.44	4.81	0.472

The size distribution of exon (logarithmic scale) for each vertebrate species is shown in Figure 2. Each distribution is fitted with equation (3) and (4) separately. The normal function fitting is shown as dashed line. The Weibull function fitting is shown as a solid line.

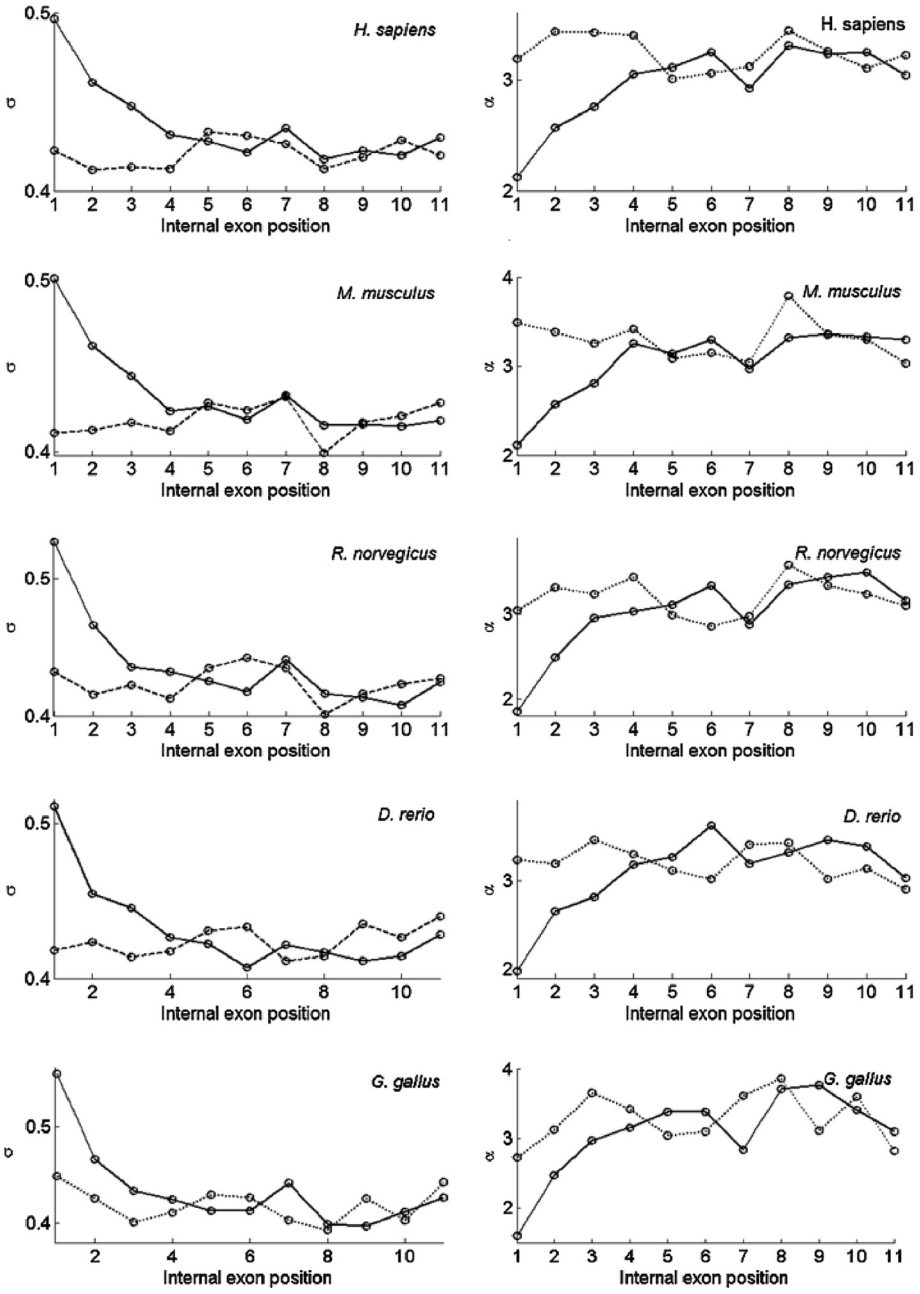

Figure 3 Fitted standard deviation (σ_E) and dependency (α) for internal exons with positions relative to 5′ UTRs (solid line) or 3′ UTRs (dashed line). The dependency value α is calculated using equation (9).

The key observation in Figure 5 is that nonadjacent insertion ratios are nearly uncorrelated for vertebrate genomes. However, the correlation between intron insertion ratios has quite different patterns for non-vertebrates. Their insertion ratios are more negatively correlated with each other than those for vertebrates, especially for *C. elegans*, *D. melanogaster*, and *O. sativa*.

In summary, analysis of empirical data reveals three significant differences between vertebrate and non-vertebrate genomes. First, a mixture of two normal functions gives a better fit to the size distribution of non-vertebrate exons, instead of one normal function for that of vertebrate exons; Second, the insertion ratio of non-vertebrates also follows a normal distribution but with larger standard deviation than that of vertebrates; Third, the insertion ratios of non-vertebrates are more negatively correlated than that of vertebrates.

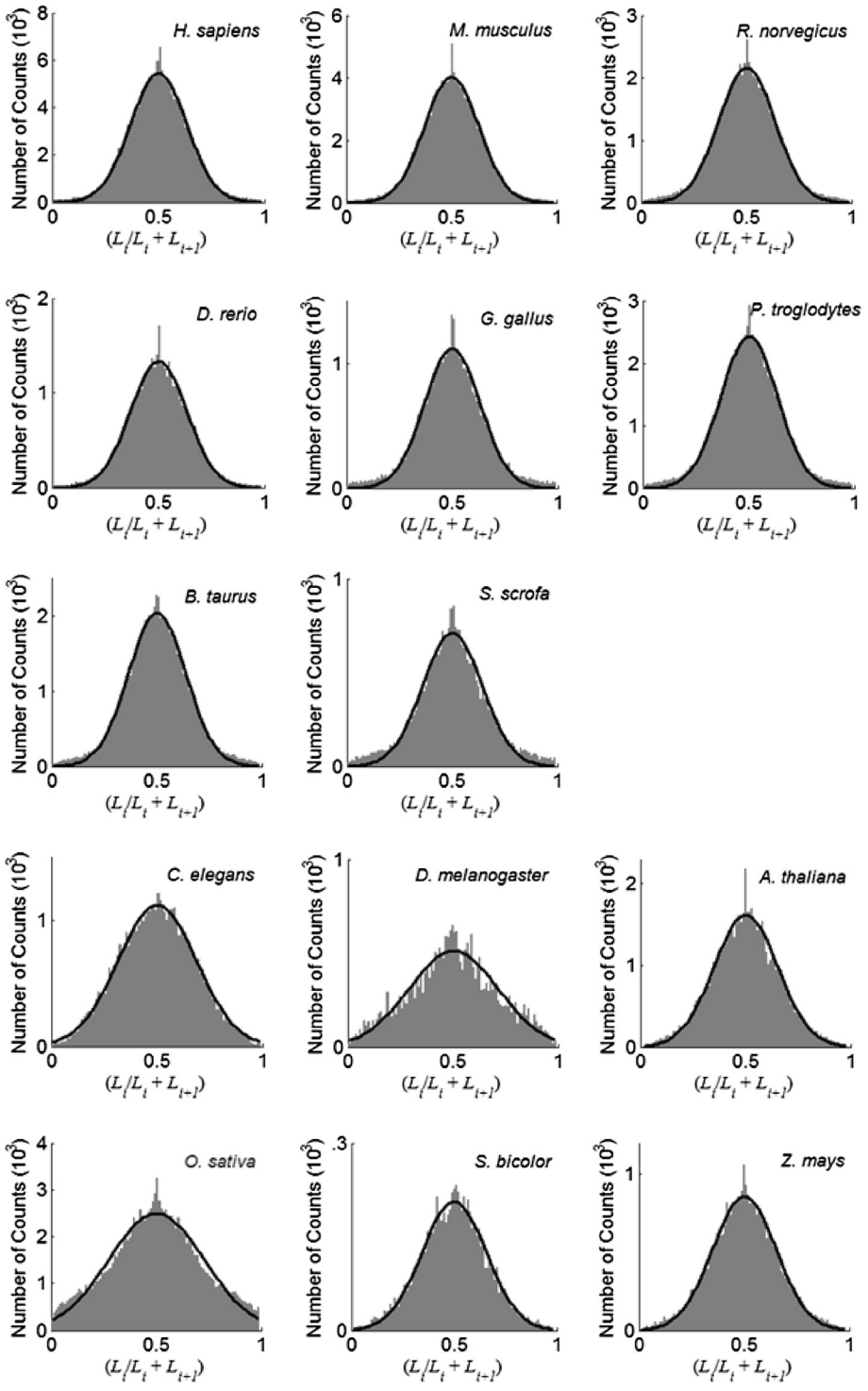

Figure 4 Genome wide distribution of $L_i/(L_i + L_{i+1})$. The histograms are drawn with bin size of 0.01, and fitted with a Normal function.

Table 3 Fitted parameters for distribution of insertion ratios from empirical data

	μ_x	σ_x
H. sapiens	0.501	0.132
M. musculus	0.501	0.132
R. norvegicus	0.501	0.135
D. rerio	0.501	0.132
G. gallus	0.501	0.132
P. troglodytes	0.502	0.134
B. taurus	0.502	0.136
S. scrofa	0.502	0.142
C. elegans	0.499	0.185
D. melangaster	0.502	0.215
A. thaliana	0.501	0.152
O. sativa	0.502	0.226
S. bicolor	0.501	0.157
Z. mays	0.501	0.152

The distribution of insertion ratios (equation (5)) for each species is shown in Figure 4. Each distribution is fitted with a normal distribution (equation (3)).

Simulation testing

Default values for L_0, m, σ_I, and μ_I

As mentioned before, we start each simulation with a long exon. Using the counts for *H. sapiens* (Table 1), we set the initial exon size and number of splitting to following values for all simulations unless specified:

$$L_0 = 2.4 \times 10^7, m = 1.8 \times 10^5 \quad (7)$$

For the remaining unknown parameters of GRFP, α, σ_I, and μ_I, we chose to examine α first with following values for σ_I and μ_I:

$$\sigma_I = 0.11, \mu_I = 0.5 \quad (8)$$

These values are determined through an EM iteration process that will be discussed in the simulation testing section. The EM iteration uses observed values of σ_x and μ_x for vertebrates (Table 3). The simulation described below shows that σ_x overestimates but is linearly proportional to σ_I, while μ_x approximates μ_I extremely well.

Relationship between α, L_0, m and σ_E, μ_E

Using the values of L_0, m, σ_I and μ_I in equations (7) and (8), we performed three GRFP simulations with α values of 0.3, 1 and 3. The size distributions of the GRFP fragments are shown in Additional file 1: Figure S4. Both the Weibull and normal functions were used for fitting to the logarithm size distribution. Fitted parameters are shown in Table 4. Numerically the Weibull function is unstable for fitting the histogram when α is close to zero. Therefore, we picked α = 0.3 to mimic a random Kolmogoroff fractioning process. α = 1 would correspond to a uniform random fragmentation process, and α = 3 will generate size distribution similar to that of real exons for the vertebrate genomes (in both shape and standard deviation). We found that although the Weibull function fits the tails of the distributions better, the fitted parameters for Weibull are extremely sensitive to the minimum size of the GRFP fragments. Therefore, in this study we use σ_E of the fitted normal function to characterize the peak width of the size distribution. It is worthwhile reemphasizing that both empirical and simulated distributions are skewed to the left; thus both tails of the peak are better fitted by the Weibull distribution.

These simulations show that σ_E (or width of the peak) decreases as α increases. To quantify how σ_E is dependent on α, we performed three GRFP simulations for each α value range from 0 to 4. The size distribution of GRFP was fitted to a normal function, and the fitted σ_E and μ_E values (mean ± 3 standard deviations) were plotted against α in Figure 6A. For α values ranging between 2 and 4, we found that the relationship between σ_E and α can be fitted with the following equation:

$$\sigma_E = 0.54/a^{0.56} + 0.14 \quad (9)$$

From equation (9), we estimate that $\alpha \approx 3$ gives the observed $\sigma_E \approx 0.43$ (Table 2). This suggests the chance of intron gain is proportional to the exon length to the 3rd power, which disagrees with the independency hypothesis of earlier work [11].

Similarly, we performed a series of GRFP simulations with different choices of L_0 and m, and the results are shown in Figure 6C-F. Figure 6C and 6E show that the estimated σ_E is independent of both L_0 and m, while Figure 6D and 6F demonstrates that the mean value (μ_E) of the resulting size distribution is dependent on both L_0 and m. From Figure 6F, we can estimate m via the GRFP simulation given σ_E and L_0, using the intersection between the dashed line (μ_E of *H. sapiens*) and the solid curve. Given that μ_E is approximately 4.81 across vertebrate genomes, we used GRFP simulation to estimate the number of splitting (m_e) for each species (Table 1). The percentage of intron loss is estimated by comparing m_e with m. For other species, the corresponding L_0 in Table 1 is used to estimate m_e. m_e and the percentages of intron loss are calculated in the same way. Note that here we use the same μ_E value of 4.8 for invertebrates although the size distributions of their exons (Additional file 1: Figure S1) are quite different (we hypothesize that they are resulting from intron loss). These estimations show that there are approximately 24% intron loss in *O. sativa*, 28% intron loss in *C. elegans*, and 57% intron

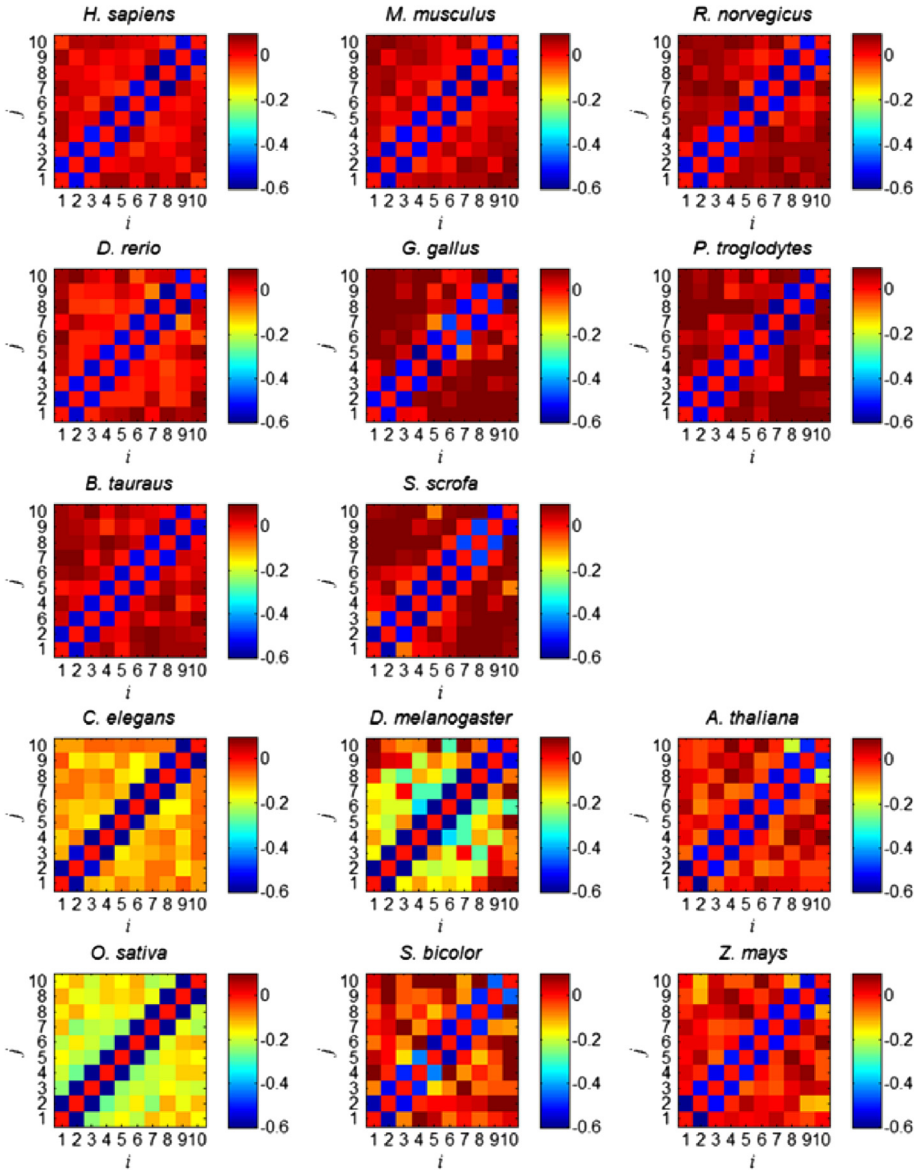

Figure 5 Correlation of insertion ratios for different species. The correlation between *x* and itself is set to zero (the diagonal blocks). The color bar on the right indicates the correlation value between insertion ratio and ratios with *i* or *j* distance to it.

loss in *D. melanogaster*, relative to what is predicted by GRFP model based on CDS length.

$$m = 0.0078L_0 + 84 \tag{10}$$

In Figure 7, we plot m_e (estimated number of splitting, open circle) against CDS length and fit it with a linear function (equation (10)). The observed number of splitting (closed circle) events is also plotted for comparison. The first observation from Figure 7 is that, under the GRFP model, number of splitting is linearly proportional to CDS length. On average, 78 splitting events will occur in an exon with length of 10000 bps (or around 8 events per 1000 bps). The second observation is that the observed number of splitting agrees well with the estimation from GRFP model, with the exception of non-vertebrates, especially *O. sativa*, *C. elegans*, and *D. melanogaster*.

Parameterizing GRFP via EM iteration

In the previous simulation studies, with the assumption of known σ_I, we have shown that σ_E is dependent on α but not on L_0 and m, which suggests that the value of α can be estimated from σ_E. However, simulations show that σ_E is also dependent on σ_I. To derive the values of α and σ_I simultaneously without assuming knowing any one of them, here we determine their values through

Table 4 Fitted parameters for size distribution of simulated exons

	Weibull		Normal	
	λ	κ	μ_E	σ_E
$\alpha = 0.3$	4.58	3.95	4.24	1.216
$\alpha = 1$	2.88	4.42	4.72	0.688
$\alpha = 3$	2.93	7.25	4.85	0.430

The size distribution of simulated exon (logarithmic scale) for each choice of dependency value (α) is shown in Additional file 1: Figure S4. Each distribution is fitted with equation (3) and (4) separately. The normal function fitting is shown as dashed line. The Weibull function fitting is shown as a solid line.

EM iterations, by combining simulations with empirically observed σ_I (Table 2), μ_x, and σ_x (Table 3) for vertebrate genomes.

Before performing EM iteration, we need to quantify how σ_I is related to σ_x. We performed a series of GRFP simulations with σ_I ranging from 0.06 to 0.18 and $\alpha = 3$. For each simulation, we calculate the insertion ratio (equation (5)) from the resulted fragments, and estimated σ_x from fitting a normal function to the histogram of insertion ratios. σ_x is plotted against given σ_I in Additional file 1: Figure S5A. The plot shows that there is a linear relationship between the two. From this relationship, we estimate that real σ_I is closer to 0.11 than the 0.13 estimated from Figure 4. To see the over estimation of σ_I, we show the simulation process and results in Additional file 1: Figure S6 and Additional file 1: Figure S7. By classifying adjacent exon pairs into four different groups, we show that the mixture of these four groups still follows a normal distribution but with larger σ_x.

For EM iteration, we use σ_I to estimate α, then use estimated α to re-estimate σ_I. The iteration start with $\sigma_I = \sigma_x = 0.13$ (Table 3).

1. Given σ_I, determine the relationship between α and σ using simulation (Figure 6A)
2. With observed σ_E (Table 2) and the estimated relationship (equation (9)), estimate α
3. With α, determine the relationship between σ_x and σ_I (Additional file 1: Figure S5A)
4. With the relationship and $\sigma_x = 0.13$, estimate σ_I
5. Return to step (1), iterate until convergence

The results of the EM iteration are shown in Additional file 1: Figure S5B and Additional file 1: Figure S5C. At the end, σ_I converges to 0.11; α converges to approximately 3. Again, $\alpha \approx 3$ suggests that, during evolution, longer exon

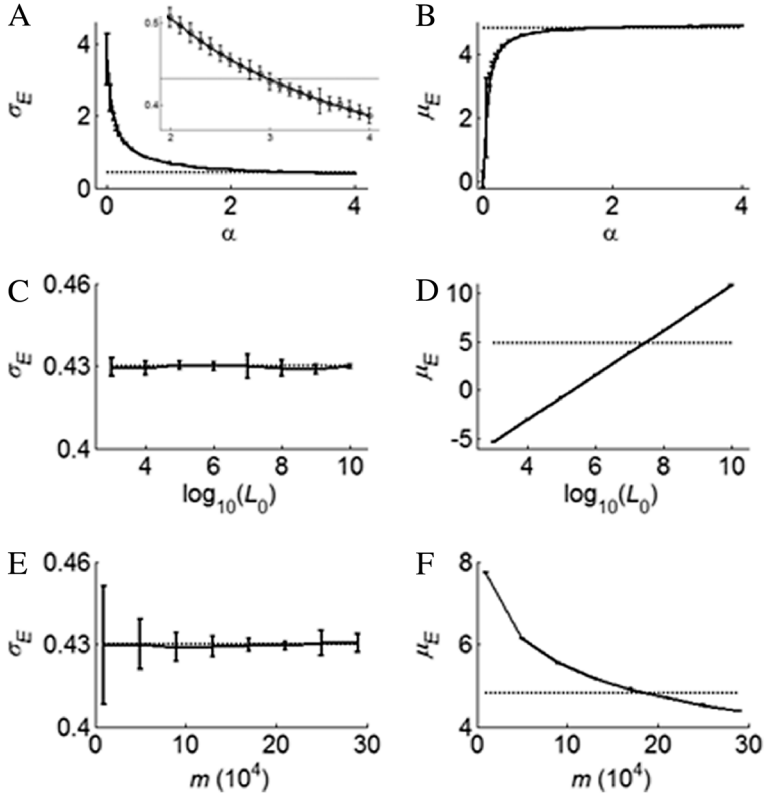

Figure 6 Relationship between GRFP parameters and σ_E, μ_E. (**A**) Plot of σ_E and (**B**) μ_E as a function of α; (**C**) Plot of σ_E and (**D**) μ_E as a function of L_0); (**E**) Plot of σ_E and (**F**) μ_E as a function of m. 3 simulations are performed for each test and plus/minus three standard deviations are shown in vertical bars. The dashed lines show where σ_E and μ_E of *H. sapiens* are (Table 2).

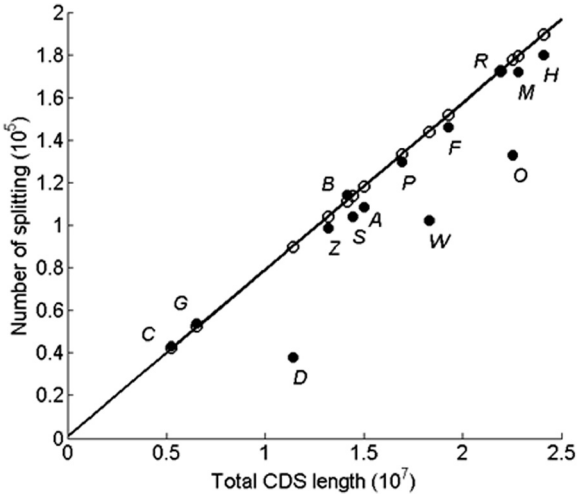

Figure 7 Plot of the number of splittings as a function of total CDS length. Estimated splitting (open circle) is from GRFP simulation with different CDS lengths (Table 1) and fits with a linear function (solid line) shown in equation (10). The observed splitting (closed circle) is also plot. Species are *Arabidopsis thaliana* (***A***), *Bos taurus* (***B***), *Sus scrofa* (***C***), *Drosophila melanogaster* (***D***), *Danio rerio* (***F***), *Gallus gallus* (***G***) *Homo sapiens* (***H***), *Mus musculus* (***M***), *Oryza sativa* (***O***), *Pan troglodytes* (***P***), *Rattus norvegicus* (***R***), *Sorghum bicolor* (***S***), *Caenorhabditis elegans* (***W***), and *Zea mays* (***Z***).

has much higher chance to gain an intron than the shorter one, with a probability proportional to its size to the 3rd power.

Intron losses accounting for increasing σ_I, σ_E and more negative $\rho(i, j)$

In Table 3 and Figure 4, we show that σ_x of non-vertebrates is significantly larger than those of vertebrate genomes. In Figure 5, we also observed more negative correlation between insertion ratios for non-vertebrates. Additional file 1: Figure S1 also shows that the size distributions of non-vertebrate exons are different from those vertebrates (Figure 2). In Table 1, we have estimated that there is a significant amount of intron losses in non-vertebrate genomes. Next, numerical simulations indicate that these three differences could result from excessive intron loss during the evolution of non-vertebrate genomes.

With simulated GRFP fragments, we gradually introduce 5-50% of "intron loss" by randomly reconnecting adjacent fragment pairs. The size distributions of the resulted fragments are fitted with a normal function, (Additional file 1: Figure S8) and the fitted σ_E is plotted against percentage of intron loss in Additional file 1: Figure S9A. The insertion ratio between each adjacent fragment pairs is calculated using equation (5). Their histograms are fitted with a normal function (Additional file 1: Figure S10) with the fitted σ_x plot against intron loss in Additional file 1: Figure S9B. In Additional file 1: Figure S9C, we calculate the correlation of insertion ratios between i and i+4 sites using equation (6) for each of the intron loss simulations and plot them against intron loss. Results in Additional file 1: Figure S9 suggest that intron loss might account for increasing σ_x (Figure 4 and Table 3), and more negative correlation between non-adjacent insertion ratios (Figure 5).

Additional file 1: Figure S8 also shows that the size distribution of exons no longer can be fit properly to a normal function. As the percentage of intron loss increases, a second peak is appearing, as the size distribution of exons for non-vertebrates in Additional file 1: Figure S1.

Discussion and conclusion

In this study, we analyze the size distribution of exons for 14 species, including eight vertebrates and six non-vertebrates. Our approach overcomes the limits of using orthologous genes, thus allowing us to infer evolutionary processes affecting the exon-intron structure across widely divergent species. The use of size distributions is more reliable than alignment based approaches if considering the accumulation of repeating intron gain/loss. Based on the size distribution of exons, we propose GRFP to characterize the evolution of eukaryotic genes. The solid agreements between GRFP simulations and observations on genomic data provide several key findings on the evolution of exon-intron gene structures.

Chance of intron gain is proportional to exon size to the 3rd power

GRFP reveals that longer exons have a higher chance to gain an intron during evolution, and reveals the novel finding that the chance of intron gain is proportional to the exon length to the third power. This finding was derived after investigating real genome data, comparing with numerical simulations, and excluding various effects on GRFP through EM iterations. This finding might explain why long exons are rare in modern organisms. E.g., statistical study has shown that only 3.5% of the primate exons are longer than 300 nt [8,9,25].

The "third power" is derived from σ_E (or width) of the exon size distributions. The model of GRFP indicates that σ_E will remain constant given the same dependency value α. Thus, the nearly identical σ_E (0.43) in Table 2 suggests the existence of a common dependency value ($\alpha \approx 3$) across vertebrate species. However, the 5′ deviations of α value might indicate that intron is less preferred there. For non-vertebrate species, α cannot be directly estimated since their exons follow a mixture of two lognormal distributions (instead of one) possibly due to excessive intron loss. For estimation of intron loss in Table 1, we simply assume that $\alpha \approx 3$ holds for non-vertebrates.

Why is the probability of intron gain proportional to the exon length to the third power? Given that the third power is usually related to volume, it might be possible that exon occupies a volume proportional to its length to the third power due to dynamic movement, and the chance of an intron attacking it is proportional to this volume. Further investigation will be needed to support this hypothesis.

No evidence for site-specific bias of intron insertion

We derive this finding from indirectly estimating the position distribution of intron insertion loci. We demonstrate that the insertion loci follow a normal distribution, peaking around the center of the exon with a standard deviation (σ_I) of 0.11. This observation does not support the proto-splice site hypothesis. If there were proto-splice sites in the exon, the insertion loci would follow the position distribution of these sites, which will most likely be a uniform distribution (Additional file 1: Figure S3).

In Figure 5, we also demonstrate that, for vertebrate genomes, intron gains are independent of each other. This observation is also one of the core assumptions of GRFP simulation. It holds on vertebrate genomes and non-vertebrate genomes if the effect of intron losses is considered. Another simplification of GRFP is that the effect of exon duplications is ignored. As mentioned earlier, the sharp spikes in Figure 4 are related to tandem exon duplication [22]. Such effect is not considered since overall their contribution is not significant in estimating either insertion loci or the chance of intron gain. This is illustrated in Additional file 1: Figure S7 and Additional file 1: Figure S10, where no such spikes are observed.

The assumption behind the estimation of insertion ratio (equation (5)) is that the order of the exons within each gene is maintained during evolution. In the cases of tandem exon duplication, exon shuffling, or intron loss, the order is just locally disrupted. Simulation also shows that the estimated insertion ratio is a mixture of four different groups of adjacent fragment pairs (Additional file 1: Figure S7, Additional file 1: Figure S8), but σ_x is linearly related to σ_I (Additional file 1: Figure S5A).

Suggesting 5′ intron gain/loss bias

By grouping exons by positions within a gene, we demonstrate that exons next to the 5′ UTR have bigger standard deviation (σ_E) than other exons. One may argue that the deviation near the 5′ UTR is caused by the fact that on average exons are longer for genes contain fewer exons. If this is the case, similar trend near the 3′ UTR should have been observed. From equation (9), bigger σ_E indicates smaller GRFP dependency value (α). The dropping of α values for exons adjacent to the 5′ UTR implies that introns are not favored there; In the GRFP model (Equation (1)), a smaller dependency value indicates a lower chance in acquiring introns during evolution. Alternatively, it might be explained as intron loss bias, that is, introns right after 5′ UTR has a tendency to lose than other introns. Certainly such comparison is limited to introns in the coding region and debatable due to the unclear stochastic process of intron loss.

Excessive intron losses accounting for deviations from GRFP

In this study, we show that exons of non-vertebrates are different from those of vertebrates in three aspects. First, the size distribution of their exons fit a mixture of two normal distributions (Additional file 1: Figure S1) instead of one for vertebrates (Figure 2). Second, their insertion ratios have much larger standard deviations (σ_x) as shown in Table 3. Third, their non-adjacent insertion ratios are more negatively correlated as shown in Figure 5.

The estimations in Table 1 (also Figure 7) suggest that there are excessive intron losses in non-vertebrate genomes. Based on this, we performed simulations of intron loss after GRFP fragmentation. Additional file 1: Figure S9 demonstrates that, with increasing intron losses, σ_E increases, σ_I increases, and $\rho(i, j)$ decreases from zero, consistent with the observation on empirical data analysis here. Therefore, GRFP model holds on non-vertebrate genomes when the effect of intron loss is considered. Comparative approaches also show that frequent intron loss has been inferred during the evolution of Nematode [26,27] and Drosophila genomes [28], though they cannot provide a genome wide estimation of percentage of intron losses.

Here, the excessive intron loss hypothesis in non-vertebrate genomes is interpreted as breaking the equilibrium between intron gain and intron loss. Although GRFP model is built on modeling intron gain events, it does not assume that introns in vertebrate genomes are never lost. Instead, we interpret the straight line in Figure 7 as a "dynamic equilibrium line for vertebrates", where the genome reaches a state of stability for its intron count. In Additional file 1: Figure S11, we observed the similar linear relationship between intron counts and CDS length by grouping genes by chromosomes for *H. sapiens*. This further proves that the "equilibrium" is reached in each chromosome and the linear relationship is independent of lineage (since human chromosomes are not related by a simple lineage relationship). If there are many intron losses during evolution, subsequent gains must have also occurred to bring the genome back to the equilibrium line. Therefore, the statistical measurements of the exons can remain the same across examined vertebrate genomes. For the non-vertebrate genomes that fall below the line, equilibrium is shifted to intron loss relative to vertebrates. The shifting

(variation of intron density) might be related to the differences in the generation time of each species [29].

The size distribution of exons (Additional file 1: Figure S1), the insertion ratios (Table 3) and the correlation map (Figure 5) suggest that *A. thaliana*, *S. bicolor*, and *Z. mays* underwent intron loss during evolution though not as significant as *O. sativa*. The estimated intron losses are around 7% for *A. thaliana* and *S. bicolor*, and 4% for *Z. mays* (Table 1). However, such percentage of intron loss is not significant enough to justify the size distribution of exons for these three plant species, a mixture of two Gaussians instead of one as shown in Additional file 1: Figure S1. Another possible explanation is that plants have undergone significant genome duplications and the rate of indels is higher for the duplicate genes [30]. The reason is that one copy of the duplicate genes is freed from the selection pressure.

Weakness and strength of GRFP

In this work, we propose the GRFP model to capture the dynamic processes describing the evolution of exon-intron structures. For vertebrate genomes, the model fits well with the well annotated genome data, including exon size distribution, distribution of insertion loci, total CDS length, number of introns, independency among intron gains, and 5′ intron gain bias. For non-vertebrate genomes, simulations show that the deviations from the vertebrate genome can be explained by excessive intron loss. The GRFP model implies that the evolution of gene structure is purely random, from picking which exon to split (gains intron) to picking intron insertion loci on the selected exon. The solid agreements between GRFP simulations and real genome data confirm that GRFP model provides one possible explanation on the exon-intron structure evolution.

It is well known that a modern genome is a collection of introns that have accreted (and been deleted) over at least a billion years. Here, by considering the whole process as a black box, we reproduce the output of this box (the current day genomes) with numerical simulations. The size distribution of exons serves as the key component in building GRFP model because of two reasons. First, the dominant factor that can shape such distribution is intron gain/loss (fragmentation). Second, the most prominent cofounding factor on exon sizes - the rate of indels in them during evolution is low. Certainly, the mechanism of intron gain is complicated considering differences across lineage, differences in rates of insertion across sites, the age of introns, the possibility of indels to maximize fit to epigenomic structures that can occur following intron gain, alternative splicing, the different models of intron gain, and so on. The process of exon fragmentation (or intron gain) might be as straightforward as the model of GRFP describes. By focusing on internal translated exons only, we have demonstrated outstanding agreements between empirical observations and GRFP simulations.

It is crucial to note that GRFP does not make any assumptions on the rate of intron gain/loss. Recent studies [6,7,31-36] have suggested that the early eukaryotes experienced a rapid burst of intron gain. In later evolution, intron loss dominates the landscape, with occasional bursts of intron gains. This does not contradict the results presented here since GRFP makes no assumptions about the rate of intron gain or loss during evolution, but instead estimates the chance for an exon to gain an intron somewhere along its length and predicts the distribution of those insertion events.

One may argue that the agreement between the GRFP model and well annotated genome structure could be fortuitous. While we cannot rule out that other models might reproduce the exons of modern genomes, the predictive power of GRFP is striking, and we believe that it is a promising approach to understanding the evolution of exon-intron structures, and an excellent starting point for new models for revealing the hidden stochastic processes of evolution.

Unanswered questions and future studies

GRFP model provides explicit rules on the exon-intron structure evolution. However, it does not address the origin of introns, the mechanism of intron insertion, and the rate of intron gain/loss. In other words, GRFP addresses where introns are inserted (which exon and where in the exon), but not when and how introns are inserted. Future research will focus on extending GRFP to model the evolution of noncoding exon, and developing GRFP-based methods for comparative genomics studies.

Additional files

Additional file 1: Figure S1. has the size distribution of non-vertebrate exons. **Figure S2** has the size distributions of *H. sapiens* exons grouped by position, supporting the plot in Figure 3. **Figure S3** shows that the distribution of proto-splice sites within *H. sapiens* coding sequences is uniform. **Figure S4** shows the size distribution of simulated exons with different dependency values. **Figure S5** shows the linear relationship between expected and observed standard deviation of insertion ratios. **Figure S6** illustrates four different groups of insertion ratios. **Figure S7** shows the distribution of insertion ratio for each of the four groups and their mixture. **Figure S8** shows the distribution of fragment size after a certain percentage of intron losses, supporting **Figure S9A**. **Figure S10** shows the distribution of insertion ratios after a certain percentage of intron loss, supporting **Figure S9B**. **Figure S11** shows the linear relationship between the number of splitting and total CDS length for each *H. sapiens* chromosome.

Abbreviations

GRFP: General Random Fragmentation Process; UTR: Untranslated region; CDS: Coding DNA Sequence; EM: Expectation Maximization.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions
LW developed the model, performed the data analysis and designed the simulation experiment. LW and LDS wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements
We thank the National Science Foundation (DBI-0735191) and National Institute of Health (P41-HG02223) for funding aspects of this work.

Author details
[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA. [2]Ontario Institute for Cancer Research, 101 College St. Suite 800, Toronto, ON M5G0A3, Canada. [3]Department of Molecular Genetics, University of Toronto, 1 King's College Circle, Toronto, ON M5S 1A8, Canada.

Received: 20 November 2012 Accepted: 22 February 2013
Published: 28 February 2013

References

1. Gilbert W: **The exon theory of genes.** *Cold Spring Harb Symp Quant Biol* 1987, **52:**901–905.
2. Gilbert W, Glynias M: **On the ancient nature of introns.** *Gene* 1993, **135**(1–2):137–144.
3. Penny D, Hoeppner MP, Poole AM, Jeffares DC: **An overview of the introns-first theory.** *J Mol Evol* 2009, **69**(5):527–540.
4. Stoltzfus A, Spencer DF, Zuker M, Logsdon JM Jr, Doolittle WF: **Testing the exon theory of genes: the evidence from protein structure.** *Science* 1994, **265**(5169):202–207.
5. Logsdon JM Jr, Tyshenko MG, Dixon C, DJ J, Walker VK, Palmer JD: **Seven newly discovered intron positions in the triose-phosphate isomerase gene: evidence for the introns-late theory.** *Proc Natl Acad Sci USA* 1995, **92**(18):8507–8511.
6. Rogozin IB, Carmel L, Csuros M, Koonin EV: **Origin and evolution of spliceosomal introns.** *Biol Direct* 2012, **7:**11.
7. Csuros M, Rogozin IB, Koonin EV: **A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes.** *PLoS Comput Biol* 2011, **7**(9):e1002150.
8. Zhang MQ: **Statistical features of human exons and their flanking regions.** *Hum Mol Genet* 1998, **7**(5):919–932.
9. Gudlaugsdottir S, Boswell DR, Wood GR, Ma J: **Exon size distribution and the origin of introns.** *Genetica* 2007, **131**(3):299–306.
10. Weibull GW: **Citation Classic - a Statistical Distribution Function of Wide Applicability.** *Curr Cont/Eng Technol Appl Sci* 1981, **10:**18–18.
11. Ryabov Y, Gribskov M: **Spontaneous symmetry breaking in genome evolution.** *Nucleic Acids Res* 2008, **36**(8):2756–2763.
12. Kolmogoroff AN: **Concerning the logarithmic normal distribution principle of dimensions of particles during dispersal.** *Cr Acad Sci Urss* 1941, **31:**99–101.
13. Tenchov BG, Yanev TK: **Weibull Distribution of Particle Sizes Obtained by Uniform Random Fragmentation.** *J Colloid Interface Sci* 1986, **111**(1):1–7.
14. Cho G, Doolittle RF: **Intron distribution in ancient paralogs supports random insertion and not random loss.** *J Mol Evol* 1997, **44**(6):573–584.
15. Rogozin IB, Sverdlov AV, Babenko VN, Koonin EV: **Analysis of evolution of exon-intron structure of eukaryotic genes.** *Brief Bioinform* 2005, **6**(2):118–134.
16. Logsdon JM Jr, Palmer JD: **Origin of introns--early or late.** *Nature* 1994, **369**(6481):526. author reply 527–528.
17. Matsumoto M, Nishimura T: **Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator.** *ACM Trans Model Comput Simul* 1998, **8**(1):3–30.
18. Coleman TF, Li YY: **An interior trust region approach for nonlinear minimization subject to bounds.** *Siam J Optim* 1996, **6**(2):418–445.
19. Thomas FC, Li YY: **On the Convergence of Interior-Reflective Newton Methods for Nonlinear Minimization Subject to Bounds.** *Math Program* 1994, **67**(2):189–224.
20. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, *et al:* **Ensembl 2012.** *Nucleic Acids Res* 2012, **40**(Database issue):D84–90.
21. Chen JQ, Wu Y, Yang H, Bergelson J, Kreitman M, Tian D: **Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria.** *Mol Biol Evol* 2009, **26**(7):1523–1531.
22. Letunic I, Copley RR, Bork P: **Common exon duplication in animals and its role in alternative splicing.** *Hum Mol Genet* 2002, **11**(13):1561–1567.
23. Long MY, De Souza SJ, Rosenberg C, Gilbert WE: **Relationship between "proto-splice sites" and intron phases: Evidence from dicodon analysis.** *Proc Natl Acad Sci USA* 1998, **95**(1):219–223.
24. Long MY, Rosenberg C: **Testing the "proto-splice sites" model of intron origin: Evidence from analysis of intron phase correlations.** *Mol Biol Evol* 2000, **17**(12):1789–1796.
25. Berget SM: **Exon recognition in vertebrate splicing.** *J Biol Chem* 1995, **270**(6):2411–2414.
26. Cho S, Jin SW, Cohen A, Ellis RE: **A phylogeny of caenorhabditis reveals frequent loss of introns during nematode evolution.** *Genome Res* 2004, **14**(7):1207–1220.
27. Kiontke K, Gavin NP, Raynes Y, Roehrig C, Piano F, Fitch DH: **Caenorhabditis phylogeny predicts convergence of hermaphroditism and extensive intron loss.** *Proc Natl Acad Sci U S A* 2004, **101**(24):9003–9008.
28. Coulombe-Huntington J, Majewski J: **Intron loss and gain in Drosophila.** *Mol Biol Evol* 2007, **24**(12):2842–2850.
29. Jeffares DC, Mourier T, Penny D: **The biology of intron gain and loss.** *Trends Genet* 2006, **22**(1):16–22.
30. Xu G, Guo C, Shan H, Kong H: **Divergence of duplicate genes in exon-intron structure.** *Proc Natl Acad Sci USA* 2012, **109**(4):1187–1192.
31. Yenerall P, Zhou L: **Identifying the mechanisms of intron gain: progress and trends.** *Biol Direct* 2012, **7:**29.
32. Koonin EV, Csuros M, Rogozin IB: **Whence genes in pieces: reconstruction of the exon-intron gene structures of the last eukaryotic common ancestor and other ancestral eukaryotes.** *Wiley Interdiscip Rev RNA* 2013, **4**(1):93–105.
33. Roy SW, Gilbert W: **The evolution of spliceosomal introns: patterns, puzzles and progress.** *Nat Rev Genet* 2006, **7**(3):211–221.
34. Rodriguez-Trelles F, Tarrio R, Ayala FJ: **Origins and evolution of spliceosomal introns.** *Annu Rev Genet* 2006, **40:**47–76.
35. Koonin EV: **The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate.** *Biol Direct* 2006, **1:**1–22.
36. Whitney KD, Garland T: **Did Genetic Drift Drive Increases in Genome Complexity.** *PLoS Genet* 2010, **6**(8):1–6.

doi:10.1186/1471-2148-13-57
Cite this article as: Wang and Stein: Modeling the evolution dynamics of exon-intron structure with a general random fragmentation process. *BMC Evolutionary Biology* 2013 **13**:57.