OPEN BIOLOGY

royalsocietypublishing.org/journal/rsob

Research

Cite this article: Loewenthal G, Wygoda E, Nagar N, Glick L, Mayrose I, Pupko T. 2022 The evolutionary dynamics that retain long neutral genomic sequences in face of indel deletion bias: a model and its application to human introns. *Open Biol.* **12**: 220223. https://doi.org/10.1098/rsob.220223

Received: 25 July 2022
Accepted: 9 November 2022

Subject Area:
genomics, bioinformatics, genetics, systems biology

Keywords:
indel, intron, deletion bias, c-value paradox, genome evolution, border-induced selection

Author for correspondence:
Tal Pupko
e-mail: talp@tauex.tau.ac.il

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.6328016.

THE ROYAL SOCIETY PUBLISHING

The evolutionary dynamics that retain long neutral genomic sequences in face of indel deletion bias: a model and its application to human introns

Gil Loewenthal[1], Elya Wygoda[1], Natan Nagar[1], Lior Glick[2], Itay Mayrose[2] and Tal Pupko[1]

[1]The Shmunis School of Biomedicine and Cancer Research and [2]School of Plant Sciences and Food Security, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978, Israel

TP, 0000-0001-9463-2575

Insertions and deletions (indels) of short DNA segments are common evolutionary events. Numerous studies showed that deletions occur more often than insertions in both prokaryotes and eukaryotes. It raises the question why neutral sequences are not eradicated from the genome. We suggest that this is due to a phenomenon we term *border-induced selection*. Accordingly, a neutral sequence is bordered between conserved regions. Deletions occurring near the borders occasionally protrude to the conserved region and are thereby subject to strong purifying selection. Thus, for short neutral sequences, an insertion bias is expected. Here, we develop a set of increasingly complex models of indel dynamics that incorporate border-induced selection. Furthermore, we show that short conserved sequences within the neutrally evolving sequence help explain: (i) the presence of very long sequences; (ii) the high variance of sequence lengths; and (iii) the possible emergence of multimodality in sequence length distributions. Finally, we fitted our models to the human intron length distribution, as introns are thought to be mostly neutral and bordered by conserved exons. We show that when accounting for the occurrence of short conserved sequences within introns, we reproduce the main features, including the presence of long introns and the multimodality of intron distribution.

1. Introduction

Insertions and deletions (indels) of short DNA segments are common molecular evolutionary events [1], whose effect expands to macro-evolutionary processes, such as the divergence among species [2–5]. By analysing homologous genomic sequences across various prokaryotic and eukaryotic taxa, it was repeatedly shown that deletions are more common than insertions [6–15], a phenomenon termed 'deletion bias'. The deletion bias raises a question: why genomes and non-coding regions such as introns do not shrink over the course of evolution? Intriguingly, the opposite has supposedly happened, as eukaryotes have larger genomes [16], longer proteins [17] and much larger intergenic regions [18] compared to prokaryotes. Petrov [19] suggested that the genome size is determined by two competing forces: short indels that reduce the genome size and large insertions (e.g. segmental duplications and the addition of transposable elements) that increase it. While this description may partially explain the overall genome size, it does not explain the length distribution of neutral sequences, such as introns, and the presence of short introns over a long evolutionary time.

He *et al*. [20] developed a deterministic model that describes how the length of a neutral sequence evolves given insertion and deletion rates, and assuming

that only indels of length one are allowed. The sequence length grows exponentially when there is an insertion bias, as one would expect. However, the sequence length grows linearly when the insertion and deletion rates are equal, which is quite counterintuitive. The authors explain it by noting that insertions emerge in between nucleotides, and thus, given a sequence of length N, there are $N+1$ possible positions for insertions and only N possible positions for deletions. Under the setting of deletion bias, He *et al*. [20] suggested that neutral sequences will be eliminated. However, the more elaborated statistical model TKF91 [21] that similarly allowed for indels of size one only, demonstrated that under very weak deletion bias, neutral sequences will be maintained.

Indel dynamics may partially explain the distribution of intron lengths within and among organisms, and the length difference between introns of closely related species is correlated to indels [12,22]. Introns are non-coding sequences that are mostly neutral [23], but reside between exons, which are usually highly conserved [24]. The distribution of the intron lengths is highly dispersed and thus it is usually plotted on a log scale. On such a scale it is often multimodal [25]. For example, the distribution of human intron lengths is bimodal and ranges from 30 to 1 160 411 base-pairs [26]. Introns length distributions of various organisms were fitted statistically with a Frechet mixture model and demonstrated that in almost all eukaryotes, the log intron length distribution is composed of multiple distinct components. This phenomenon was hypothesized to stem from the presence of alternative splicing mechanisms [25]. Other studies classified introns according to their lengths and suggested that different classes are characterized by different splicing signals [27] or the presence of conserved elements [28].

In this work, we develop a general statistical framework for indel dynamics and derive a set of models with increasing complexity that depict the length distribution of neutral sequences. We start with a simple model allowing indels of length one only and reproduce TKF91 result stating that under a very weak deletion bias, arbitrarily large sequences are likely to appear. We extend this model by allowing indels of various lengths and show that this allows the occurrence of neutral sequences even when the deletion bias is substantial. This is due to selection against deletions that encompass conserved regions at the neutral sequence borders, a phenomenon we term *border-induced selection*. Moreover, we suggest a model that includes small-conserved elements embedded within the neutral sequence. The presence of these elements may significantly increase the neutral sequence length as they multiply the intensity of border-induced selection. Finally, we test how well our indel models explain the empirical intron length distribution in human. We show that the quantitative fit of the models improves with model complexity. Moreover, our framework provides an explanation for the multimodality observed in the distribution of intron lengths.

2. Results

2.1. General model of length evolution

Our general goal is to understand how the length of neutral sequences evolves through generations. We start by describing a simple stochastic process for sequence length evolution. As we are only interested in length variation, substitutions are ignored, i.e. we implicitly assume that indel evolutionary dynamics is context independent, that is, the probability of indel events and their type does not vary as a result of substitutions. Further, we assume that the length can vary only due to indel events, and thus we ignore the possible contribution of rare events such as segmental duplications. In general, the variation of sequence length through generation can be described as follows:

$$L_n = L_{n-1} + \Delta L_{n-1}, \tag{2.1}$$

where L_n is a random variable denoting the length of the sequence in generation n, and ΔL_{n-1} is a random variable that quantifies the sum of insertion and deletion lengths in the transition from generation $n-1$ to generation n. Different assumptions regarding the indel dynamics would change the distributions of ΔL_{n-1} and thus the stationary distribution of L_n. In the models proposed below, we focus on neutral segments that are bordered between highly conserved segments. We demonstrate the applicability of our models to introns, which we approximate as neutrally evolving sequences.

2.2. Human intron length distribution—empirical dataset for model validation

Below, increasingly complex models were tested for their fit to the human intron length distribution, as a representative of a large and well-curated empirical dataset. The human length distribution is characterized by the following features: mean intron length of approximately 7000 base-pairs (bp), standard deviation (s.d.) of approximately 20 000 bp and a range that spans over five orders of magnitude: the minimal intron size is 30 bp, and the maximal is 1 160 411 bp [26]. Furthermore, the distribution of the logarithm of the length is bimodal [25], with the main and minor modes at 2100 and 100 bp, respectively (figure 1).

2.3. Model with indels of size one

We start with a simple model (M1) that allows only insertions and deletions of size one and a uniform distribution of indel events along the sequence. We also assume that the sequence in question is placed between two conserved sequences that cannot be deleted. Therefore, even if the sequence length goes to zero in a certain generation, it can revive. This is analogous to the immortal link of the TKF91 model [21]. Under this model, the distribution of ΔL_{n-1} is

$$\Delta L_{n-1} = \begin{cases} 1 & p_i(L_{n-1}+1) \\ 0 & \text{otherwise} \\ -1 & p_d L_{n-1} \end{cases}. \tag{2.2}$$

In this model, p_i and p_d are the probabilities of an insertion and deletion event, per character per generation, respectively. In each generation, the length can vary by no more than a single character. We also assume that events are extremely rare, and thus, both $p_i(L_{n-1}+1)$ and $p_d L_{n-1}$ are much smaller than 1.0, even for sequences longer than a million characters [29]. Since insertions occur between characters, there is an additional place for insertions compared to deletions, i.e., deletions can only occur upstream to each character while an insertion can also occur downstream to the last

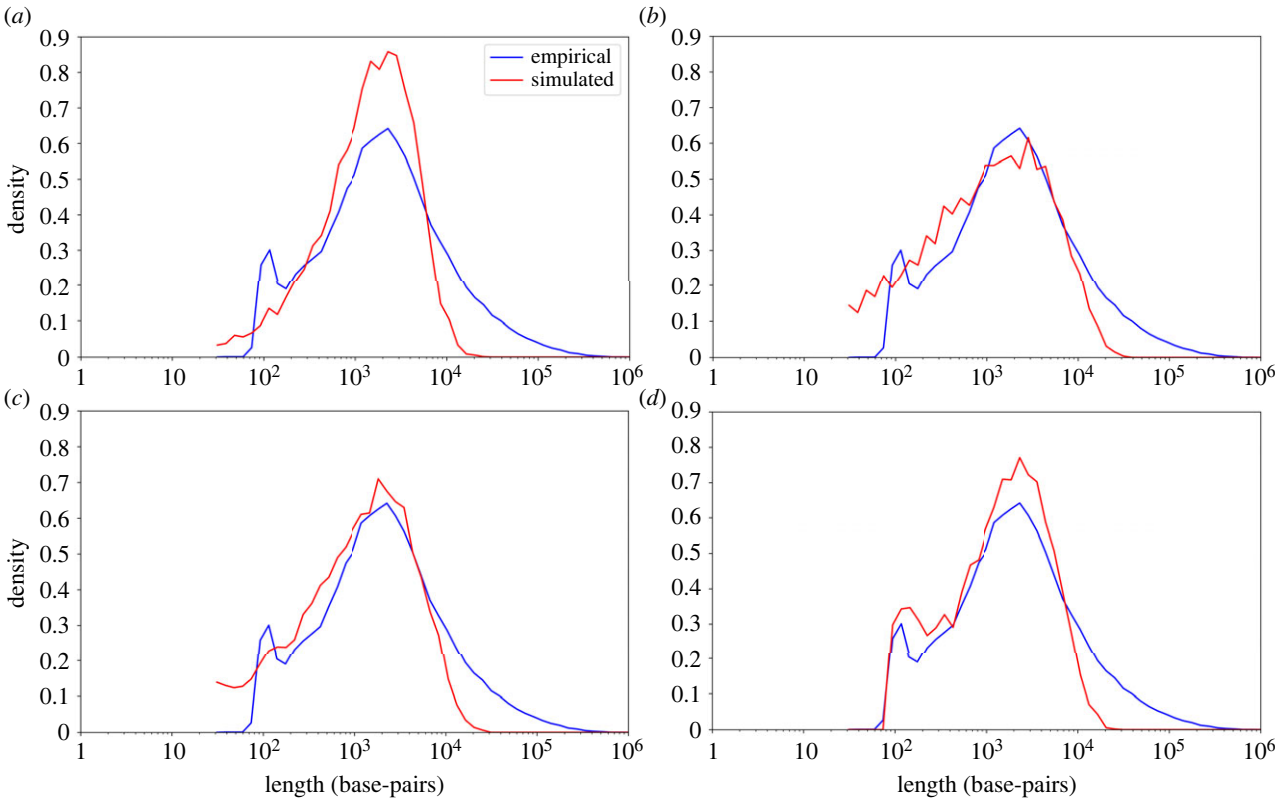

Figure 1. Empirical and simulated distributions of intron lengths in human. In each panel, the blue line shows a length distribution derived from the human intron empirical data. The red line is the distribution obtained using simulations with one of the models M1–M4. In all models, the length distribution was derived from 10,000 simulations. (*a*) The simulations are derived from M1, with the following parameters: $r = 0.9995$, $p_i = 0.9995 \cdot 10^{-7}$, $p_d = 10^{-7}$. The MSE is 0.64; (*b*) the simulations are derived from M2 with the following parameters: $r = 0.9975$, $p_i = 0.9975 \cdot 10^{-7}$, $p_d = 10^{-7}$, $\mu_i = 17$, $\mu_d = 5$. The MSE is 0.34; (*c*) the simulations are derived from M3 with the following parameters: $r = 0.983$, $p_i = 2.68 \cdot 10^{-8}$, $p_d = 10^{-7}$, $\mu_i = 16.5$, $\mu_d = 4.5$. The MSE is 0.31; (*d*) The simulations are derived from M4 that relies on the output of M3 model. The M3 parameters used here are $r = 0.9776$, $p_i = 2.65 \cdot 10^{-8}$,$p_d = 10^{-7}$, $\mu_i = 16.5$, and $\mu_d = 4.5$. The M4 model parameters are $l_e = 88$, $l_i = 35$, and $p_c = 0.69$. The MSE is 0.29.

character. For example, if an intron is of length three bases, insertions can occur at four possible locations, while deletions can occur at only three locations (i.e. upstream of each base).

Given the stochastic process described above, taking expectations from both sides of equation (2.1) yields:

$$\overline{L_n} = \overline{L_{n-1}} + p_i(\overline{L_{n-1}} + 1) - p_d\overline{L_{n-1}}. \tag{2.3}$$

Equation (2.3) coincides with the model of He *et al*. [20]. The solution for equation (2.3) for the case in which $p_i = p_d$ is a linear growth, where L_0 is the expectation of the sequence length at the beginning of the process:

$$\overline{L_n} = L_0 + np_i. \tag{2.4}$$

When $p_i \neq p_d$ the solution is

$$\overline{L_n} = (L_0 - L_\infty)(1 + p_i - p_d)^n + L_\infty, \tag{2.5}$$

where $L_\infty \equiv p_i/(p_d - p_i)$. This notation is used as the length converges to L_∞ when $p_i < p_d$ and not to p_i as reported by He *et al*. [20]. If $p_i > p_d$, the exponential term grows to infinity. Figure 2 demonstrates the behaviour of the solution of equation (2.3) for the three regimes: $p_i > p_d$, $p_i = p_d$, and $p_i < p_d$.

We will focus on the deletion bias regime (i.e. $p_i < p_d$), as it was repeatedly reported that deletions are more common than insertions. The steady-state length, L_∞, depends solely on the ratio between the insertion and deletion probabilities,

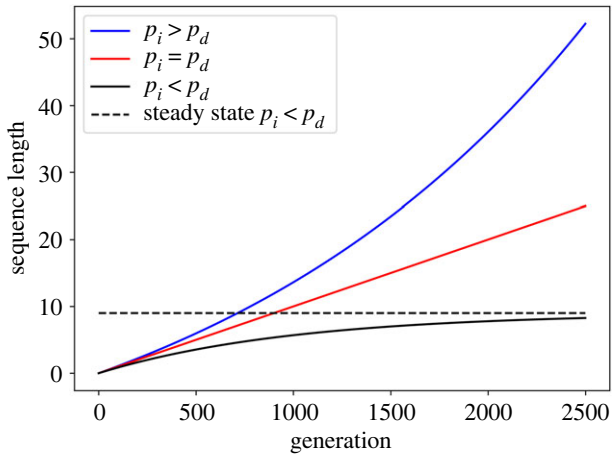

Figure 2. The expectation of the model with single-character indels (M1). Three types of solution are possible: exponential growth when $p_i > p_d$, linear growth when $p_i = p_d$, exponential decay to a steady-state when $p_i < p_d$. The parameters used to generate the graphs were $L_0 = 0$, $p_d = 0.01$, $p_i =$ (0.0105 or 0.01 or 0.009).

$r \equiv p_i/p_d$ (multiplying the value of p_i and p_d by a fixed factor has no effect on the stationary distribution—it only affects the time till convergence; see Appendix A):

$$L_\infty = \frac{r}{1-r}. \tag{2.6}$$

Equation (2.6) shows that L_∞ can be arbitrarily long by selecting the appropriate r-value. For example, when $r = 0.9, 0.99, 0.999$ then $L_\infty \approx 9, 99, 999$, respectively. Under this model, when the sequence is shorter than L_∞, it has an insertion bias even though $p_i < p_d$. For example, when the sequence length is one, there are two possible insertions and a single possible deletion, thus if $p_i > p_d/2$ the sequence will have an insertion bias. Of note, equation (2.6) is the same as reported in the TKF91 model [21].

The above model clearly does not fit the human empirical intron data. Shown in figure 1 is the distribution of the intron lengths of the human genome (see Materials and Methods). The mode of this distribution is 2100 bp. Simulations with the above model allow obtaining estimate of the stationary distribution for each value of r. We searched for the value of r that provides the best fit in terms of mean squared error (MSE) between the empirical and simulated length distributions of human introns (see Materials and Methods). The optimal value of r was 0.9995 with an MSE of 0.64. Although for this value of r the main mode of the empirical distribution matches the mode of the simulated stationary distribution, the two distributions vary greatly with respect to their shapes (figure 1*a*). Specifically, while the empirical distribution has a heavy right tail, these long introns are missing from the stationary distribution generated by the model. In addition, the empirical distribution has a second mode near 100 bp, which is missing from the simulated distribution. Given this discrepancy, we now turn to a more complex model that relaxes the oversimplified assumption that all indels are of size one.

2.4. Model with indels of fixed arbitrary size

We generalize the above-described model by adding parameters μ_i and μ_d that are the insertion and deletion lengths, respectively. Of note, these lengths are considered constant (below, we relaxed this assumption by allowing a distribution of indel sizes). Under M2, the distribution of L_{n-1} is

$$\Delta L_{n-1} = \begin{cases} \mu_i & p_i(L_{n-1}+1) \\ 0 & \text{otherwise} \\ -\mu_d & p_d p_{\text{valid}} L_{n-1} \end{cases}. \tag{2.7}$$

We note that because deletions are no longer restricted to have a length of one, some deletions may extend from the neutral sequence to its conserved flanking regions, entailing substantial fitness reduction, and thus such deletions are rejected. This is reflected in the extra factor p_{valid} in equation (2.7). Given the value of L_{n-1} and μ_d, p_{valid} can be computed by

$$p_{\text{valid}} = \max\left(\frac{L_{n-1} - \mu_d + 1}{L_{n-1}}, 0\right). \tag{2.8}$$

Of note, if the proposed deletion length is larger than the current sequence length, equation (2.8) will assign a probability of zero to p_{valid}, suggesting that neutral sequence segments are immune to deletions larger than their size. Under this scenario, there is a bias for insertions in neutral sequences that are very short (see also [30]).

Taking the expectation of both sides of equation (2.1), accounting for the distribution of ΔL_{n-1} as in equations (2.7–2.8) yields

$$\overline{L_n} = \overline{L_{n-1}} + p_i\mu_i(\overline{L_{n-1}} + 1) - p_d\mu_d \begin{cases} (\overline{L_{n-1}} - \mu_d + 1), & \text{if } \mu_d \le \overline{L_{n-1}} \\ 0, & \text{if } \mu_d > \overline{L_{n-1}} \end{cases}. \tag{2.9}$$

The first two terms on the right-hand side resemble equation (2.3), except that the second term is multiplied by the insertion length μ_i. The third term indicates that no deletions are allowed when $\mu_d > \overline{L_{n-1}}$. As expected, when we choose $\mu_i = \mu_d = 1$, equation (2.9) reduces to equation (2.3). Of note, equation (2.9) is effectively a three-parameter difference equation. Let r be the ratio between the expectation of the insertion length and the expectation of the deletion length: $r \equiv (p_i\mu_i)/(p_d\mu_d)$. Note that r, as defined for M1 (equation 2.6), is a special case of the r in M2, when $\mu_i = \mu_d = 1$. Using these definitions, we can rewrite equation (2.9) with three parameters r, p_d, and μ_d:

$$\overline{L_n} = \overline{L_{n-1}} + rp_d\mu_d(\overline{L_{n-1}} + 1) - p_d\mu_d \begin{cases} (\overline{L_{n-1}} - \mu_d + 1), & \text{if } \mu_d \le \overline{L_{n-1}} \\ 0, & \text{if } \mu_d > \overline{L_{n-1}} \end{cases}. \tag{2.10}$$

When the steady-state length, L_∞, is substantially larger than μ_d, we can approximately ignore the $\mu_d > \overline{L_{n-1}}$ condition and solve the following equation:

$$\overline{L_n} = \overline{L_{n-1}} + rp_d\mu_d(\overline{L_{n-1}} + 1) - p_d\mu_d(\overline{L_{n-1}} - \mu_d + 1). \tag{2.11}$$

The solution of equation (2.11) resembles the solution of equation (2.3):

$$\overline{L_n} = (L_0 - L_\infty)(1 + (r-1)p_d\mu_d)^n + L_\infty. \tag{2.12}$$

The steady-state sequence length, L_∞ is

$$L_\infty = \frac{r + \mu_d - 1}{1 - r}. \tag{2.13}$$

It is interesting to compare the properties of M2 and M1. First, as expected, if μ_d and μ_i are set to be 1, M2 reduces to M1. Second, when r is close to 1, L_∞ under M2 is roughly μ_d fold larger than L_∞ under M1. Third, when r is close to 0, there are substantially more deletions than insertions, and thus the $\mu_d > \overline{L_{n-1}}$ condition of equation (2.10) may not be negligible. Under such a high-deletion regime, many deletions are rejected and L_∞ should be larger than the value predicted by equation (2.13). Hence, in this case, the value in equation (2.13) can be considered as a lower bound for the steady-state length.

To fit this model to the human empirical intron data, we assume that the mean insertion and deletion lengths are 16.5 and 4.5 bp, respectively, as reported by Matthee *et al*. [31] for introns in mammals. We use 17 and 5 bp, as this model supports only integers. We scanned the r parameter, using the same procedure we applied for M1, and the r-value that yielded the optimal fit was 0.9975 with an MSE of 0.34, which is a substantial improvement over the MSE obtained for M1. The inferred r-value is slightly lower than that obtained using the M1 model. The increased fit and the fact that the shape of the M1 and M2 distributions are different emphasize the importance of the conserved regions at the edges of the neutrally evolving sequence, i.e. boundary-induced selection. Figure 1*b* shows that despite the increased fit as measured by the MSE value, substantial discrepancies remain between the simulated and the empirical distributions.

2.5. Model with indels of varying sizes

We generalize the above model by relaxing the assumption that the insertions and deletions are of constant length. Thus, in M3 we assume that the length of each indel is drawn from a specified distribution. Let $f(\cdot)$ and $g(\cdot)$ be the length distributions of insertions and deletions, respectively. The distribution of L_{n-1} under this model is

$$\Delta L_{n-1} = \begin{cases} i & p_i f(i)(L_{n-1}+1) \quad 1 \le i \le m_{\text{zip}} \\ 0 & \text{otherwise} \\ -d & p_d g(d) p_{\text{valid}}(d) L_{n-1} \quad 1 \le d \le m_{\text{zip}} \end{cases}. \tag{2.14}$$

The probability that a deletion is valid depends on the deletion length, as described in equation (2.8). The larger the deletion length d is, the smaller $p_{\text{valid}}(d)$ is. In equation (2.14), $p_{\text{valid}}(d)$ is a function of L_{n-1}, which complicates the analytic computation of the expectation of ΔL_{n-1}.

In this work, we assume a truncated Zipfian distribution for both insertions and deletions, as in Loewenthal *et al.* [10]:

$$f(k|a,m_{\text{zip}}) = \frac{k^{-a}}{\sum_{i=1}^{m_{\text{zip}}} i^{-a}}. \tag{2.15}$$

This distribution has two parameters a and m_{zip}, which control the shape of the distribution and the maximally allowed indel length, respectively. We assume that both insertion and deletion lengths are Zipfian distributed, but we allow different a parameters for insertions and deletions. Unless otherwise stated, m_{zip} is set to 150 throughout this work. We denote by μ_i and μ_d the expectations of the truncated Zipfian distribution for insertions and deletions, respectively.

In Appendix B, we show using simulations that the mean value of this distribution, L_∞, for M3 is about 4.5 fold higher compared to model M2 when using the same set of parameter values. The higher mean in M3 compared to M2 stems from the higher probabilities that proposed long deletions are rejected.

The mean of the stationary distribution in M3 (i.e. L_∞) is often larger than the mean deletion length, even under a very strong deletion bias regime. For example, when $r = 0.25$, $\mu_d = 15$, and $\mu_i = 5$, the mean length is 23.2 bp (figure 3). This can be explained by the fact that when the segment length is shorter than the mean deletion length, most deletions are rejected, and thus, effectively, a strong bias for insertions exists.

Applying this model to the human intron length distribution, we found that the best fit is obtained with $r = 0.983$, yielding an MSE of 0.31. In this computation we applied the mean indel lengths as reported in Matthee *et al.* [31] (i.e. $\mu_d = 4.5$ and $\mu_i = 16.5$; figure 1*c*). The modes of the empirical and simulated distributions are similar. However, major discrepancies between the shapes of the two distributions exist: the means and s.d. of the distributions are (6793; 21 860) and (1684; 2354) bp for the empirical and simulated distributions, respectively. There is also an additional peak in the empirical distribution (bimodality) that is absent in the simulated distribution.

2.6. Conserved segments

We propose a toy statistical model, M4, to qualitatively demonstrate that conserved segments embedded within the neutrally evolving sequence may explain the gap between the theoretical and empirical distributions. Accordingly, in M4 we assume that in each neutral sequence there is some probability, p_c, that it includes a single conserved sequence of length l_i within it. Let l_e denote the total length of the conserved sequences in the edges of an intron (i.e. the 5′ and the 3′ splice sites). Of note, that the length of the conserved sequence on each edge may be different, and l_e represents their sum. To simulate this model, for each neutral sequence, a Bernoulli trial is executed with probability p_c to decide if there is a conserved sequence within the intron sequence (in M4, we only allow a single internal conserved sequence). If there is no conserved sequence, then the length of the intron is the sum of l_e and the length of a single stochastic simulation under M3. If a conserved sequence is introduced, the length of the intron is the sum of l_e and l_i, and the lengths of two stochastic simulations under M3. For simplicity, we assume that the parameters l_e, l_i, and p_c are the same for all the simulated introns. The optimized simulated distribution is shown in figure 1*d*. In contrast with the fit in model M3, now the simulated distribution is bimodal similarly to the empirical distribution. The fit between the two distributions slightly improved (the MSE decreased from 0.31 to 0.29), the value of r decreased to 0.9776, and the mean and the s.d. of the sequence length both slightly increased, reaching (2149; 2365) bp. The fitted parameters under M4 are $l_e = 88$, $l_i = 35$, and $p_c = 0.69$. The high p_c value suggests that most of the introns have a conserved internal segment. Thus, conserved segments may explain the low peak in the intron length distribution, it widened the length distribution and resulted in a lower value of r.

Here and in previous works [25], the intron length values are transformed using the log function prior to their visualization as a distribution. This is justified, as the intron lengths are spread over five orders of magnitude. This results in a bimodal distribution. Note that when the same empirical distribution is plotted without log scaling, the bimodality disappears (figure 4*a*), meaning that this is an artefact of the contraction made by the log scale. Our analyses suggest that the inclusion of conserved segments (both within and in the border of introns) leads to the appearance of bimodality in the log scale and to longer introns. The bimodality due to the introduction of an internal conserved segment is more intuitive, as M4 splits the introns to two groups: introns with

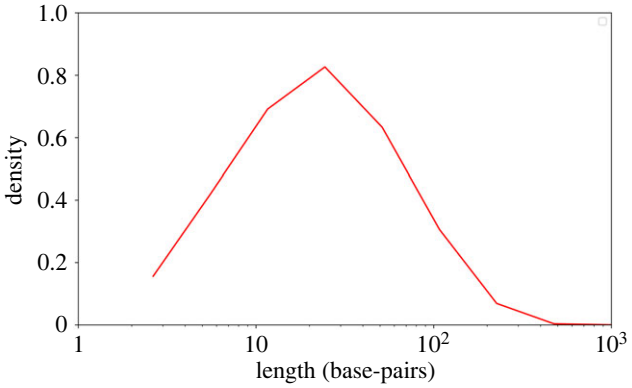

Figure 3. Long neutral sequences are probable in M3 under a high-deletion bias regime. The red line is the distribution obtained using M3 under a strong deletion bias regime ($r = 0.25$). The length distribution was derived from 10 000 simulations with the following parameters: $p_i = 7.5 \cdot 10^{-9}$, $p_d = 10^{-8}$, $\mu_i = 5$, $\mu_d = 15$.

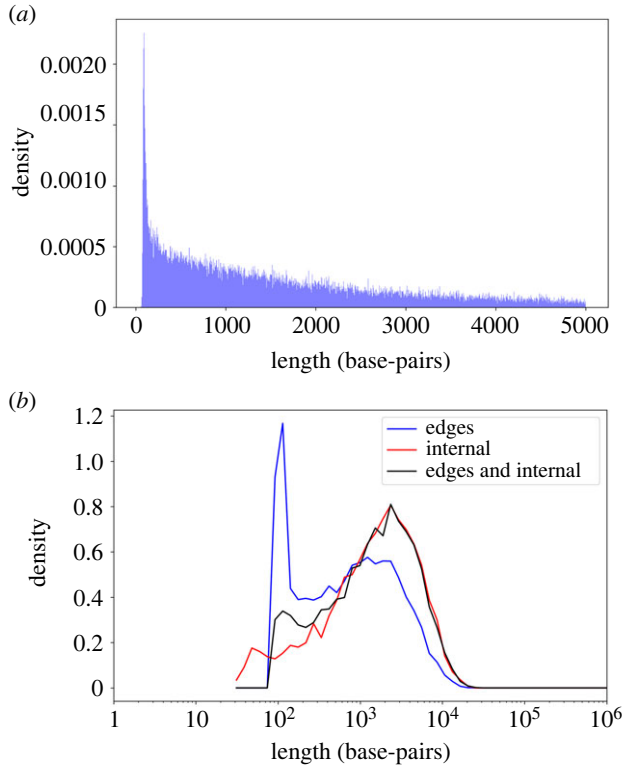

Figure 4. Empirical length distribution in linear scale and M4 parameter. (*a*) The human intron length distribution of introns shorter than 5000 bp in a linear scale. There is a single mode, as opposed to the logarithmic scale. Thus, the two modes are an artefact of the logarithmic transformation. (*b*) Three M4 simulations were run with the same M3 parameters: $r = 0.9776$, $p_i = 2.65 \cdot 10^{-8}$, $p_d = 10^{-7}$, $\mu_i = 16.5$, and $\mu_d = 4.5$. In blue, we use the M4 parameters: $l_e = 88$, $l_i = 0$, and $p_c = 0$. Thus, we only add a constant to the M3 simulation, reflecting the conserved intronic splice sites at the edges of the intron. It can be seen that bimodality emerged. The mean length is 1295 bp. In red, we use the M4 parameters: $l_e = 0$, $l_i = 35$, and $p_c = 0.69$. These parameters dictate the presence of introns with an internal conserved segment, but there are no conserved segments at the edges of the introns. The mean value increased to 2067 bp, reflecting how internal conserved segments may significantly increase the intron lengths. In black, running the simulations with conserved segments both within and at the edges of the intron. The M4 parameters were $l_e = 88$, $l_i = 35$, and $p_c = 0.69$. The mean length is 2158 bp. Of note, this distribution was generated with the same parameters as in figure 1*d* and the small differences reflect stochastic variations.

an internal conserved segment and introns without one. The edge conserved segments, which are mathematically equivalent to adding a constant to the distribution, can also create a mode in log scale, but not on a linear scale, because of the contraction of the log scale. Specifically, it is the introduction of conserved elements in the borders of introns that mostly explains the bimodality and the presence of internal conserved elements that lead to the generation of longer introns (figure 4*b*).

3. Discussion

In this work, we presented several increasingly complex models for the length distribution of neutrally evolving sequences. Critical to our models is the assumption that neutrally evolving segments are placed between highly conserved sequences. We focused on the deletion bias regime, which was shown in a large number of studies to be prevalent across all domains of life [6–15]. It was previously suggested that this deletion bias leads to shrinkage of genomes over evolutionary times [19]. Here, we showed that the placement of conserved flanking sequences can lead to the emergence of long sequences, even in a high-deletion bias regime. The counterintuitive result that long neutrally evolving sequences can emerge even under a strong deletion bias is due to the rejection of deletions that invade the highly conserved borders of the neutral sequences. We hence propose the term border-induced selection for this phenomenon.

To test the fit of our models to empirical genomic data, we studied the length distribution of human introns, which are thought to evolve mostly neutrally [23] and are in between exons, which are generally highly conserved [24]. Using the M3 model, we reconstructed the main mode of the empirical distribution. However, M3 does not reproduce a secondary lower peak of the distribution and does not explain the extremely high variance of the lengths of introns, which in this case spans over five orders of magnitude. Yet, the M3 model does not account for conserved segments within and at the edges of introns. Examples for such conserved segments are the 3′ and 5′ splice sites, as well as intron splicing enhancers and silencers [32–34]. We modelled the presence of conserved segments within introns using M4, and it resulted in both the emergence of a second peak and a slight increase in the variance. We note that M4 only allows a single intermediate conserved segment, and we expect that a more elaborate model that allows multiple conserved internal segments will better explain the presence of very long introns.

As is often the case with models, many assumptions are clear oversimplifications of biological reality. First, the output of the stochastic model depends on the length distribution of indels. As in previous work, we assumed that this distribution follows a truncated Zipfian distribution [10] with a cutoff of 150 characters. In our work, we did not study if this is the best-fitting distribution, and it is possible that other distributions may provide better fit to the data. Our model also assumes perfect neutrality of the sequence of interest, which is likely to be an oversimplified assumption for species with a large effective population size. The model also assumes a perfect conservation of the bordering conserved elements. This is also true for the conserved regions within introns. The effect of relaxing these assumptions needs to be further studied. Of note, M4 is not a genuine stochastic model with specified parameters controlling the probability of emergence and loss of conserved regions. However, we expect that a more complex model, which addresses these limitations, will not change the main result of our model, namely, that neutral sequences are not purged under a deletion bias regime. Moreover, mobile genetic elements, microsatellites, and genome rearrangement events are all ignored in our study. Clearly, these factors should be integrated when moving towards complex models that aim to capture the main forces dictating genome dynamics evolution. Finally, throughout this work, we assumed that the empirical length distributions are in equilibrium, and we thus compare them to the stationary distributions of our models. It may be the case that this assumption too is an oversimplification of reality. While in this work we focused on presenting the theory behind our

models, and demonstrated them only on the evolution of human intron lengths, our models provide a framework to study length distribution of introns of other organisms, as well as other neutrally evolving sequences such as prokaryotic spacers [35]. For verification of these models in other organisms, further studies regarding indel rates, length distributions, and distribution of internal conserved segments are required.

Our analyses show that as we move to increasingly more complex models, the insertion-to-deletion rate ratio, r, gets further away from the value of one. Equation (2.13) indicates that as r gets closer to one, small perturbations of r lead to sharp changes in intron length distribution. Since in our models r is closer to one, the mode is unstable. For example, changing the value of r between values such as, say, 0.9995 and 0.9998, would generate distributions with very different means: from 2000 to 5000 bp in M1, respectively. Indeed, the decrease in r as we move to more advanced models, reaching an r of 0.98 in M4 lends an additional level of justification for these advanced models. We anticipate that incorporating multiple conserved elements will further lead to more stable models.

Previous studies provide indirect support for our proposed models. First, Pozzoli *et al*. [28] compared mouse and human introns, and showed that the deletion rate is higher for long introns, in line with our models, because deletions in short introns are often rejected, while in long introns, there is little to none border-induced selection. Pozzoli *et al*. also examined introns of similar length and found that the number of conserved sequences is negatively correlated to deletion rate, again in line with the existence of border-induced selection. Moreover, the authors also showed that almost all introns longer than 10 000 bp harbour conserved sequences, emphasizing the important role conserved segments play in generating the heavy tail of the intron length distribution. Second, Yang *et al*. [36] have recently shown that within a genome, the intron size is correlated to the alternative splicing level and prevalence. Sironi *et al*. [37] showed a correlation between the logarithm of intron length and the number of conserved sequences within the intron. These observations can fit a general model, in which tight regulation of splicing is associated with conserved intronic regulatory elements, which, as we showed, lead to long introns. Third, it was shown that first introns are much longer, typically about double, than other introns, which may be partially explained by the observation that functional motifs are more frequent in first introns [38]. This observation further supports the M4 model, in which the presence of conserved segments leads to longer introns. Thus, both the conserved edges of the neutral sequence and the conserved elements within it contribute to the prevalence of long neutral sequences. We note that in M4, the introns are no longer truly neutral, as they are embedded with conserved segments within them.

Our model provides a plausible explanation for the extremely large variance in intron lengths within a species. However, it does not directly explain differences in distributions among species. One trivial explanation is that the model parameters themselves evolve. Thus, different species have different insertion-to-deletion rate ratios and, possibly, different propensity for the emergence of conserved regions within introns. These factors may be relevant not only to the distribution of intron lengths, but rather for the entire genome size. Indeed, eukaryotes generally have a lower deletion bias than prokaryotes [9], which may partially explain the higher eukaryotes genome sizes and their higher variation [39]. It was previously shown that the total indel rate is negatively correlated to the effective population size [29]. It was also shown that the effective population size times the mutation rate is correlated to the mean length of introns [40]. A dependence between the insertion-to-deletion ratio and the effective population size, if exists, may help explain this relationship: smaller population size leads to an increased r, which in turn leads to longer introns.

4. Materials and methods

4.1. Intron length distribution

We downloaded the canonical genome of human from the UCSC Genome Browser database [41]. The canonical genome introns annotation is based on the longest coding sequence isoform for each gene. The complete distribution of intron lengths in the canonical human genome is provided at https://github.com/elyawy/Luigi (last accessed 29 June 2022).

4.2. Simulations and optimization of model parameters

The simulations of M1, M2, and M3 are based on the Gillespie algorithm [42]. We used discrete generations, and thus waiting times were geometrically distributed. The number of generations needed to reach stationarity is dictated by the transient part of equation (2.12), i.e. $(1 + (r - 1)p_d\mu_d)^n$. We simulated until this factor was below 10^{-6}.

Model parameters for M1–M3 were optimized using a grid search over the r parameter in the range [0.37, 0.9999]. The optimal r parameter had the lowest MSE between the simulated and empirical length distribution (in logarithmic scale). For M4, we heuristically searched for the values of r, l_e, l_i, and p_c that best fit the empirical distribution according to the MSE criterion. This was done using the module optimize of Python SciPy package [43] using the 'trf' option, which is based on the trust region algorithm described in Gould *et al*. [44].

4.3. Source code and implementation details

The source code and documentation of the C++ (models M1–M3) and Python (model M4) implementations of the stochastic simulations are available at https://github.com/elyawy/Luigi (last accessed 29 June 2022).

Data accessibility. The data are provided in the electronic supplementary material [45].

Authors' contributions. G.L.: conceptualization, data curation, formal analysis, investigation, methodology, resources, software, validation, visualization, writing—original draft and writing—review and editing; E.W.: data curation, methodology, software, validation and writing—original draft; N.N.: methodology, resources, software and writing—original draft; L.G.: methodology, resources and writing—original draft; I.M.: formal analysis, methodology, supervision, validation, writing—original draft and writing—review and editing; T.P.: conceptualization, formal analysis, funding acquisition, investigation, methodology, project administration, software, supervision, writing—original draft and writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Table 1. Multiplying the value of p_i and p_d by a fixed factor has no effect on the stationary distribution. Each row in the table represents a set of 10 000 simulations under M1 with the parameters p_i and p_d specified in the first two columns. Each set of simulations were run until convergence to the stationary distribution. In each row, the parameters p_i and p_d were multiplied by 0.1 compared to the row above, so the r is fixed to 0.99 in all rows. The empirical means and s.d. of the various sets are similar. We also performed a two-sided Kolmogorov–Smirnov (KS) test between the first simulation set and the other sets. The null hypothesis, namely that the distributions are the same, cannot be rejected.

p_i	p_d	r	empirical mean	empirical s.d.	KS. Stat.	KS. p-value
9.90×10^{-07}	1.00×10^{-06}	0.99	98.50	99.86	—	—
9.90×10^{-08}	1.00×10^{-07}	0.99	99.03	98.84	0.01	0.34
9.90×10^{-09}	1.00×10^{-08}	0.99	98.23	99.22	0.01	0.63
9.90×10^{-10}	1.00×10^{-09}	0.99	100.24	102.71	0.01	0.58

Funding. This study was funded by Israel Science Foundation (ISF) grant no. 2818/21 to T.P.

Acknowledgements. G.L., E.W., N.N. and L.G. were supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel Aviv University. We thank Jotun Hein and Alon Itzkovitch for fruitful discussion.

Appendix A

See table 1.

Appendix B

We computed L_∞ for M3 using simulations with various values for r, μ_d, and μ_i. Here, μ_i and μ_d denote the expectations of the truncated Zipfian distribution for insertions and deletions, respectively. Our simulations suggest that a strong linear correlation exists between the simulated value of L_∞ and the L_∞ values calculated based on equation (2.13). Let κ_{zipf} be the slope of the regression line (assuming an intercept of zero). Thus,

$$L_\infty \cong \kappa_{\text{zipf}} \frac{r + \mu_d - 1}{1 - r}. \tag{B.1}$$

Note that in this case, the value of r depends on both μ_d and μ_i. We also note that the truncated Zipfian distribution depends on two parameters: a and m_{zip}. The values of μ_d and μ_i can be computed given these two parameters. In equation (B.1), κ_{zipf} may vary depending on m_{zip}. The correlation coefficient between the estimated L_∞ based on equation (2.13) and L_∞ estimated using simulations was found to be higher than 0.97 for all tested m_{zip} values (electronic supplementary material, figure S1).

The fact that the slope of the regression line is higher than one in all cases (electronic supplementary material, figure S1) suggests that introducing variation to the indel lengths pushes the distribution of sequence lengths to higher values, including increasing the average length L_∞ by a factor $\kappa_{\text{zipf}} > 1$. We hypothesize that the reason for the shift in sequence lengths is due to a reduction of the expectation of the mean length of accepted proposed deletions, which stems from the variation of indel lengths. Indeed, the expectation of the length of accepted proposed deletions for an arbitrary deletion length distribution $g(d)$ for a sequence of length L, $E[a.d]$, is given by $E[a.d] = \sum_i i g(d) p_{\text{valid}}$. This expectation has a compact form when the maximal deletion length is lower than L: $E[a.d] = \mu_d - (1/L)(\mu_d(\mu_d - 1) + V_d)$ where μ_d and V_d are the mean length and variance of $g(d)$, respectively. As expected, when $L \to \infty$, only the first term contributes so $E[a.d] = \mu_d$, but for a finite L, the negative second term reduces $E[a.d]$, and thus the reduction is higher when variation in the allowed deletion length is introduced.

References

1. Cartwright RA. 2009 Problems and solutions for estimating indel rates and length distributions. *Mol. Biol. Evol.* **26**, 473–480. (doi:10.1093/molbev/msn275)
2. Anzai T *et al.* 2003 Comparative sequencing of human and chimpanzee MHC class I regions unveils insertions/deletions as the major path to genomic divergence. *Proc. Natl Acad. Sci. USA* **100**, 7708–7713. (doi:10.1073/pnas.1230533100)
3. Britten RJ. 2002 Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels. *Proc. Natl Acad. Sci. USA* **99**, 13633–13635. (doi:10.1073/pnas.172510699)
4. Britten RJ, Rowen L, Williams J, Cameron RA. 2003 Majority of divergence between closely related DNA samples is due to indels. *Proc. Natl Acad. Sci. USA* **100**, 4661–4665. (doi:10.1073/pnas.0330964100)
5. Wetterbom A, Sevov M, Cavelier L, Bergström TF. 2006 Comparative genomic analysis of human and chimpanzee indicates a key role for indels in primate evolution. *J. Mol. Evol.* **63**, 682–690. (doi:10.1007/s00239-006-0045-7)
6. Fitch WM. 1973 Aspects of molecular evolution. *Annu. Rev. Genet.* **7**, 343–380. (doi:10.1146/annurev.ge.07.120173.002015)
7. Graur D, Shuali Y, Li WH. 1989 Deletions in processed pseudogenes accumulate faster in rodents than in humans. *J. Mol. Evol.* **28**, 279–285. (doi:10.1007/BF02103423)
8. De Jong WW, Rydén L. 1981 Causes of more frequent deletions than insertions in mutations and protein evolution. *Nature* **290**, 157–159. (doi:10.1038/290157a0)
9. Kuo CH, Ochman H. 2009 Deletional bias across the three domains of life. *Genome Biol. Evol.* **1**, 145–152. (doi:10.1093/gbe/evp016)
10. Loewenthal G *et al.* 2021 A Probabilistic model for indel evolution: differentiating insertions from deletions. *Mol. Biol. Evol.* **38**, 5769–5781. (doi:10.1093/molbev/msab266)
11. Mira A, Ochman H, Moran NA. 2001 Deletional bias and the evolution of bacterial genomes. *Trends Genet.* **17**, 589–596. (doi:10.1016/S0168-9525(01)02447-7)
12. Ogata H, Fujibuchi W, Kanehisa M. 1996 The size differences among mammalian introns are due to the accumulation of small deletions. *FEBS*

Lett. **390**, 99–103. (doi:10.1016/0014-5793(96)00636-9)

13. Van Passel MWJ, Smillie CS, Ochman H. 2007 Gene decay in archaea. *Archaea* **2**, 137–143. (doi:10.1155/2007/165723)
14. Fan Y, Wang W, Ma G, Liang L, Shi Q, Tao S. 2008 Patterns of insertion and deletion in mammalian genomes. *Curr. Genomics* **8**, 370–378. (doi:10.2174/138920207783406479)
15. Zhang Z, Gerstein M. 2003 Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acids Res.* **31**, 5338–5348. (doi:10.1093/nar/gkg745)
16. Cavalier-Smith T. 1982 Skeletal DNA and the evolution of genome size. *Annu. Rev. Biophys. Bioeng.* **11**, 273–302. (doi:10.1146/annurev.bb.11.060182.001421)
17. Brocchieri L, Karlin S. 2005 Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Res.* **33**, 3390–3400. (doi:10.1093/nar/gki615)
18. Ahnert SE, Fink TMA, Zinovyev A. 2008 How much non-coding DNA do eukaryotes require? *J. Theor. Biol.* **252**, 587–592. (doi:10.1016/j.jtbi.2008.02.005)
19. Petrov DA. 2002 Mutational equilibrium model of genome size evolution. *Theor. Popul. Biol.* **61**, 531–544. (doi:10.1006/tpbi.2002.1605)
20. He Y, Tian S, Tian P. 2019 Fundamental asymmetry of insertions and deletions in genomes size evolution. *J. Theor. Biol.* **482**, 109983. (doi:10.1016/j.jtbi.2019.08.014)
21. Thorne JL, Kishino H, Felsenstein J. 1991 An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.* **33**, 114–124. (doi:10.1007/BF02193625)
22. Moriyama EN, Petrov DA, Hartl DL. 1998 Genome size and intron size in *Drosophila*. *Mol. Biol. Evol.* **15**, 770–773. (doi:10.1093/oxfordjournals.molbev.a025980)
23. Resch AM, Carmel L, Mariño-Ramírez L, Ogurtsov AY, Shabalina SA, Rogozin IB, Koonin EV. 2007 Widespread positive selection in synonymous sites of mammalian genes. *Mol. Biol. Evol.* **24**, 1821–1831. (doi:10.1093/molbev/msm100)
24. Siepel A, Haussler D. 2004 *Computational identification of evolutionarily conserved exons*. In *Proc. of the Annu. Int. Conf. on Computational Molecular Biology, RECOMB*, pp. 177–186. New York, NY: Association for Computing Machinery.
25. Gotoh O. 2018 Modeling one thousand intron length distributions with fitild. *Bioinformatics* **34**, 3258–3264. (doi:10.1093/bioinformatics/bty353)
26. Piovesan A, Caracausi M, Ricci M, Strippoli P, Vitale L, Pelleri MC. 2015 Identification of minimal eukaryotic introns through GeneBase, a user-friendly tool for parsing the NCBI Gene databank. *DNA Res.* **22**, 495–503. (doi:10.1093/dnares/dsv028)
27. Mount SM, Burks C, Herts G, Stormo GD, White O, Fields C. 1992 Splicing signals in *Drosophila*: intron size, information content, and consensus sequences. *Nucleic Acids Res.* **20**, 4255–4262. (doi:10.1093/nar/20.16.4255)
28. Pozzoli U, Menozzi G, Comi GP, Cagliani R, Bresolin N, Sironi M. 2007 Intron size in mammals: complexity comes to terms with economy. *Trends Genet.* **23**, 20–24. (doi:10.1016/j.tig.2006.10.003)
29. Sung W, Ackerman MS, Dillon MM, Platt TG, Fuqua C, Cooper VS, Lynch M. 2016 Evolution of the insertion-deletion mutation rate across the tree of life. *G3* **6**, 2583–2591.
30. Ptak SE, Petrov DA. 2002 How intron splicing affects the deletion and insertion profile in *Drosophila melanogaster*. *Genetics* **162**, 1233–1244. (doi:10.1093/genetics/162.3.1233)
31. Matthee CA, Eick G, Willows-Munro S, Montgelard C, Pardini AT, Robinson TJ. 2007 Indel evolution of mammalian introns and the utility of non-coding nuclear markers in eutherian phylogenetics. *Mol. Phylogenet. Evol.* **42**, 827–837. (doi:10.1016/j.ympev.2006.10.002)
32. Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, Blencowe BJ, Frey BJ. 2010 Deciphering the splicing code. *Nature* **465**, 53–59. (doi:10.1038/nature09000)
33. Cooper DN. 2010 Functional intronic polymorphisms: buried treasure awaiting discovery within our genes. *Hum. Genomics* **4**, 284–288. (doi:10.1186/1479-7364-4-5-284)
34. Lin CL, Taggart AJ, Fairbrother WG. 2016 RNA structure in splicing: an evolutionary perspective. *RNA Biol.* **13**, 766–771. (doi:10.1080/15476286.2016.1208893)
35. Rédei GP. 2008 *Encyclopedia of genetics, genomics, proteomics and informatics*. New York City, NY: Springer.
36. Yang P, Wang D, Kang L. 2021 Alternative splicing level related to intron size and organism complexity. *BMC Genomics* **22**, 853. (doi:10.1186/s12864-021-08172-2)
37. Sironi M, Menozzi G, Comi GP, Bresolin N, Cagliani R, Pozzoli U. 2005 Fixation of conserved sequences shapes human intron size and influences transposon-insertion dynamics. *Trends Genet.* **21**, 484–488. (doi:10.1016/j.tig.2005.06.009)
38. Bradnam KR, Korf I. 2008 Longer first introns are a general property of eukaryotic gene structure. *PLoS ONE* **3**, e3093. (doi:10.1371/journal.pone.0003093)
39. Bohlin J, Pettersson JHO. 2019 Evolution of genomic base composition: from single cell microbes to multicellular animals. *Comput. Struct. Biotechnol. J.* **17**, 362–370. (doi:10.1016/j.csbj.2019.03.001)
40. Lynch M, Conery JS. 2003 The origins of genome complexity. *Science* **302**, 1401–1404. (doi:10.1126/science.1089370)
41. Rosenbloom KR *et al.* 2015 The UCSC genome browser database: 2015 update. *Nucleic Acids Res.* **43**, D670–D681. (doi:10.1093/nar/gku1177)
42. Gillespie DT. 1977 Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **81**, 2340–2361. (doi:10.1021/j100540a008)
43. Virtanen P *et al.* 2020 SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272. (doi:10.1038/s41592-019-0686-2)
44. Gould NIM, Lucidi S, Roma M, Toint PL. 1999 Solving the trust-region subproblem using the Lanczos method. *SIAM J. Optim.* **9**, 504–525. (doi:10.1137/S1052623497322735)
45. Loewenthal G, Wygoda E, Nagar N, Glick L, Mayrose I, Pupko T. 2022 The evolutionary dynamics that retain long neutral genomic sequences in face of indel deletion bias: a model and its application to human introns. Figshare. (doi:10.6084/m9.figshare.c.6328016)