Splice sites seldom slide — Supplementary Material

Steven Sêton Bocco Miklós Csűrös[*]

March 3, 2016

Contents

List of Tables

List of Figures

[*] csurosm@gmail.com

A Methods

A.1 Formal framework for gene structure annotations and alignments

Consider a set of orthologous genes from organisms indexed by $1, \ldots, n$. Gene i is a $3\ell_i$-length coding sequence along genome i, defined by the *coding positions* $x_1^{(i)}, \ldots, x_{3\ell_i}^{(i)}$. For the sake of simplicity, we assume without loss of generality that all genes are transcribed on the forward ("Watson" (Cartwright and Graur, 2011)) strand: $x_1^{(i)} < x_2^{(i)} < \cdots < x_{3\ell_i}^{(i)}$; $x_1^{(i)}$ is the first position of the start codon, and $x_{3\ell_i}^{(i)}$ precedes the terminal codon. The transcribed sequence $t_{1\cdots 3\ell}^{(i)}$ is formed by the nucleotides $t_j^{(i)} \in \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$ at positions $x_j^{(i)}$.

A *nucleotide alignment* partitions the set of genomic positions $\{(i, j) \colon i = 1, \ldots, n; j = 1, \ldots, 3\ell_i\}$ into equivalence relations; each equivalence relation is a homology statement between nucleotide positions of different sequences, written as $(i, j) \overset{\mathsf{nuc}}{\sim} (i', j')$. Every equivalence class includes at most one position for each i, and the statements are *pairwise collinear* (Ghiurcuta and Moret, 2013): for every (i, j_1) and (i, j_2) with $j_1 < j_2$, there exist no position pairs $(i', j_1'), (i', j_2')$ in any other gene i' with $j_1' > j_2'$ that are simultaneously homologous $(i, j_1) \overset{\mathsf{nuc}}{\sim} (i', j_1')$ and $(i, j_2) \overset{\mathsf{nuc}}{\sim} (i, j_2')$.

A *codon index* is formed by phased triplets $\langle i, k \rangle = [(i, 3k-2), (i, 3k-1), (i, 3k)]$, and a *codon alignment* imposes triplet homologies: for any i, i', k, k', $(i, 3k-2) \overset{\mathsf{nuc}}{\sim} (i; 3k'-2)$, $(i, 3k-1) \overset{\mathsf{nuc}}{\sim} (i', 3k'-1)$ and $(i, 3k) \overset{\mathsf{nuc}}{\sim} (i', 3k')$ hold either simultaneously at once, denoted as $\langle i, k \rangle \overset{\mathsf{codon}}{\sim} \langle i', k' \rangle$, or not at all. The corresponding codon-level equivalence classes are the alignment columns $j = 1, \ldots, m$, indexed in increasing order of genomic coordinates of the underlying genes; every column contains at most one codon $\langle i, k \rangle$ for every sequence i. Accordingly, the alignment is described as an $n \times m$ matrix $\mathbf{A} = (a_{i,j})$, where each entry $a_{i,j} \colon i = 1, \ldots n; j = 1, \ldots, m$ is either a gap character, or a codon index $\langle i, k \rangle$. By collinearity, $a_{i,j} \colon j = 1, \ldots, m$ contain all $\langle i, k \rangle \colon k = 1, \ldots, \ell_i$, in consecutive order of k.

A.2 Joint representation of exon junctions and codon-level alignment

The structure of gene i is described by the intron gaps between consecutive genomic positions $x_j^{(i)}$; *exons* are formed by the maximal segments $x_{j+\delta}^{(i)} =$

$\delta + x_j^{(i)}: \delta = 0, 1, 2, \ldots,$. A length-$\lambda$ *intron* at $x_{j+1}^{(i)} = x_j^{(i)} + 1 + \lambda$ is either in *phase 1* if $j = 3k - 2$, or in *phase 2* if $j = 3k - 1$, or in *phase 0* if $j = 3k$ for some k. Phase 1 and phase 2 introns belong to a unique column j in the alignment $\mathbf{A}$, where $a_{i,j} = \langle i, k \rangle$. Phase 0 introns may fall into a gap, when $\langle i, k \rangle$ and $\langle i, k+1 \rangle$ are in non-adjacent alignment columns.

Consider the codon homologies in genes i, i'. For any codon $\langle i, k \rangle$, either exists a unique homologous codon $\langle i, k \rangle \overset{\mathsf{codon}}{\sim} \langle i', k' \rangle$, or there is no homology in i'. A *match segment* is a maximal segment of successive codon homologies: $\langle i, k + \delta \rangle \overset{\mathsf{codon}}{\sim} \langle i', k' + \delta \rangle$ for all $0 \leq \delta < d$ in case of a segment of length d starting with the (k, k') codon pair. A codon $\langle i, k \rangle$ is *deleted* in gene i' if $\langle i, k \rangle \overset{\mathsf{codon}}{\not\sim} \langle i', k' \rangle$ for any k'. As for the exon-intron structure of gene i, we are interested in an *intron's gap distance*, i.e., in the number of codon matches (upstream or downstream) next to an exon junction. For a phase 1 or 2 intron, which interrupts codon $\langle i, k \rangle$ that has no homolog in gene i', both gap distances are 0. Otherwise, if codon $\langle i, k \rangle$ is within a match segment covering codons $\langle i, k - a \rangle, \ldots, \langle i, k + b \rangle$, then the upstream gap distance is $(a + 1)$ and the downstream gap distance is $(b + 1)$.

In order to calculate the statistical significance of bias in gap distance distributions, we collect information on the segment lengths for a pair of genomes i, i'. Let $n_{\mathsf{M}}(d)$ denote the number of match segments with length d, across all the aligned orthologous genes, and let D denote the total number of deleted codons in genes of i. Define

$$\begin{aligned} N_0 &= D & L_0 &= D + \sum_{j>0} j \cdot n_{\mathsf{M}}(j) & \\ N_d &= \sum_{j \geq d} n_{\mathsf{M}}(j) & L_d &= L_{d-1} - N_{d-1} & \{d > 0\} \end{aligned}$$

Finally, let $p_d = N_d / L_d$. Now, p_d is the probability that a phase 1 or phase 2 intron, placed uniformly along all genes, falls exactly d codons away from a gap, provided it is at least as far away as d. If r_d denotes the number of introns with distance d from the nearest downstream gap, then the null hypothesis of uniform intron placement implies that r_d is distributed as binomial random variable with parameters $R_d = \sum_{j \geq d} r_d$ and p_d. Accordingly, P-values for the gap distances are computed as $P_d = \sum_{j \geq r_d} \binom{R_d}{j} p_d^j (1 - p_d)^{R_d - j}$.

A.3 Intron contexts

Intron contexts are built from conserved codons in pairwise alignments. For an intron in gene i, conservation with respect to all other genes i' is considered. Conservation is measured by log-odds scores (Durbin *et al.*, 1998) for matches $\langle i, k\rangle \overset{\mathsf{codon}}{\sim} \langle i', k'\rangle$, calculated specifically for the genome pair i, i'. Taking all the aligned codons in genes of i, i', we compute the number of columns where every non-terminal codon pair (a, b) appears, denoted as $c_{a\to b}$, as well as the number of times every non-terminal codon b is used in genes of i', denoted as c_b. Then the score for matching codon a with b is

$$\mathsf{score}(a \to b) = \left\lceil \log_\alpha \frac{c_{a\to b}/\sum_{b'} c_{a\to b'}}{c_b/\sum_{b'} c_{b'}} \right\rceil,$$

where $\lceil\cdot\rceil$ means rounding up to the nearest integer, and $\alpha = \sqrt[1000]{10}$ (i.e., conservation is measured in *millibans*). The score for a run of matches $\langle i, k+\delta\rangle \overset{\mathsf{codon}}{\sim} \langle i', k'+\delta\rangle : 0 \le \delta < d$ is the sum of match scores

$$\sum_{\delta=0}^{d-1} \mathsf{score}\Big(t^{(i)}_{3(k+\delta)-2..3(k+\delta)}, t^{(i')}_{3(k'+\delta)-2..3(k'+\delta)}\Big).$$

An *anchor* is defined as a run of matches within the same match segment, scoring above a predefined threshold τ, without intervening introns in any of the two sequences. (Note that there might be no upstream or downstream anchor.) In our analysis, threshold τ is chosen as the expected score of four matches:

$$\tau = 4 \times \frac{\sum_{a,b} c_{a,b} \cdot \mathsf{score}(a \to b)}{\sum_{a,b} c_{a,b}}.$$

We find the nearest upstream and downstream anchors from an exon junction by adapting Kadane's linear-time algorithm for finding a maximum-scoring segment (Bentley, 1984). An anchor that includes the match $\langle i, k\rangle \overset{\mathsf{codon}}{\sim} \langle i', k'\rangle$ lies on the *diagonal* with offset $\Delta_{i\to i'} = x^{(i')}_{3k'} - x^{(i)}_{3k}$.

Intron contexts are formed by overlapping regions bracketed by upstream and downstream anchors. An intron context provides a *consistent* coordinate system, if its upstream and downstream anchors are consistent among themselves: for all triples of genes i, i', i'' for which the upstream (or downstream) anchors lie on diagonal offsets $\Delta_{i\to i'}$, $\Delta_{i'\to i''}$, $\Delta_{i\to i''}$, we have

$$\Delta_{i\to i''} = \Delta_{i\to i'} + \Delta_{i'\to i''}. \tag{1}$$

We keep consistent intron contexts that contain all n homologs, with at most one bracketed exon junction in each. Note that $\Delta_{i\to i'}$ is defined by the anchors only if i contains an intron: we extend the notation in consistent contexts to include $\Delta_{i'\to i} = -\Delta_{i\to i'}$ and $\Delta_{i'\to i''} = \Delta_{i'\to i} + \Delta_{i\to i''}$ where i', i'' have no introns and gene i is an arbitrary intron-bearing gene.

A.4 Putative reannotation and ancestral reconstruction

Given a consistent intron context, diagonal offsets $\Delta_{i\to i'}$ are defined upstream and downstream for all gene pairs i, i', and all triples satisfy (1). Using the diagonals, a genomic position $x_j^{(i)}$ is projected to another gene i' as $f_{i\to i'}\bigl(x_j^{(i)}\bigr) = x_j^{(i)} + \Delta_{i\to i'}$. When gene i contains an intron, the corresponding donor and acceptor sites are projected onto every other homolog i', and marked if gene i' has a possible splicing motif there. More precisely, we project the two intronic positions immediately next to the splice site using upstream diagonals for donor, and downstream diagonals for acceptor sites, and we inspect the 2-nt genomic sequence motif found there. Candidate donor motifs are $\bigl\{\mathsf{GT}, \mathsf{GC}, \mathsf{AT}\bigr\}$, and candidate acceptor motifs are $\bigl\{\mathsf{AG}, \mathsf{AC}, \mathsf{TG}\bigr\}$. Candidate introns are formed by pairing annotated sites and projected sites with proper motifs in all donor-acceptor combinations provided they are in matching phases, do not introduce premature stop codons, and have a minimum length of 24 nt.

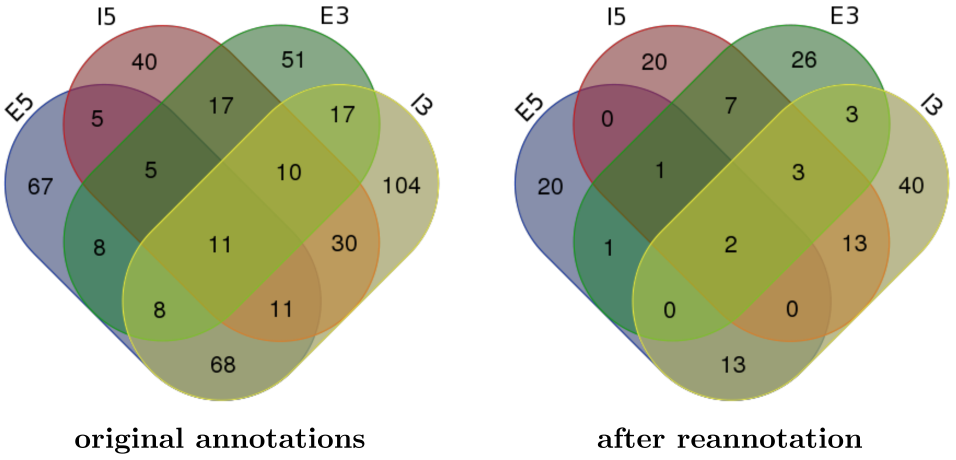
Given a phylogeny, we label every tree node either with pairs of donor-acceptor site coordinates (d, a), or with $\emptyset$ for no intron. Labels are encoded using donor-side and acceptor-side coordinate projections onto the same (arbitrary) reference gene $i = 1$; i.e., a 2-nt donor motif at (projected or annotated) positions $x'..x'+1$ in gene i' is encoded as $d = f_{i'\to 1}(x') = x' + \Delta_{i'\to 1}$. Likewise, acceptor site coordinates are encoded by using the projections defined by the downstream diagonal offsets. A terminal node with an annotated intron can be labeled as $\emptyset$ only if the intron length is a multiple of three, and the sequence introduces no stop codon. For ancestral nodes, we consider the label set comprising all phase-matched pairs formed by donor and acceptor sites at the terminal nodes, and $\emptyset$. At terminal nodes, reannotation of each site, or the "exonification" of an intron (label $\emptyset$) is penalized by unit penalty. The reconstruction uses the following edge penalties (which were chosen so that two losses are favored over one gain, and up to five splice sites can be reannotated for the price of a loss or a shift):

Event	Parent label	Child label	penalty
Loss	(d, a)	$\emptyset$	5
Gain	$\emptyset$	(d, a)	12
5' site shift	(d, a)	$(d', a); d' \neq d$	5
3' site shift	(d, a)	$(d, a'); a' \neq a$	5
intron sliding	(d, a)	$(d', a'); d' \neq d, a' \neq a$	10

The score for a complete labeling is the sum of edge penalties, plus the reannotation penalties at terminal nodes. A minimum-score labeling is found by adapting Sankoff's dynamic programming algorithm (Sankoff and Rousseau, 1975) for the context-specific label set.

Table (i): Gene structure statistics. The ortholog set has higher intron density than average genes in the underlying genomes do, but otherwise the ortholog selection did not introduce bias in length and phase.

Organism	Genes	Number of introns			Typical	Average
		phase 1	phase 2	phase 0	intron length	introns/gene
Complete genomes						
albu	13804	7363 (25%)	7477 (26%)	14370 (49%)	56	2.12
hyal	14321	3862 (28%)	3771 (27%)	6211 (45%)	77	0.97
phca	19805	6389 (28%)	5356 (23%)	11123 (49%)	63	1.15
phci	26131	7011 (25%)	7030 (25%)	13896 (50%)	72	1.07
phin	17787	8761 (28%)	8298 (26%)	14300 (46%)	66	1.76
phpa	20822	6241 (26%)	6151 (25%)	11943 (49%)	65	1.17
phra	15605	7087 (29%)	6256 (25%)	11433 (46%)	67	1.59
phso	18969	9693 (28%)	9132 (27%)	15362 (45%)	71	1.80
pyul	15322	5930 (24%)	6138 (25%)	12561 (51%)	82	1.61
Analyzed ortholog families						
albu	1917	1102 (24%)	1118 (24%)	2389 (52%)	53	2.40
hyal		696 (26%)	682 (26%)	1265 (48%)	75	1.38
phca		814 (25%)	798 (25%)	1635 (50%)	66	1.69
phci		809 (24%)	894 (26%)	1722 (50%)	71	1.79
phin		971 (25%)	976 (25%)	1896 (49%)	69	2.00
phpa		733 (23%)	808 (26%)	1582 (51%)	66	1.63
phra		1002 (26%)	1012 (26%)	1900 (49%)	73	2.04
phso		1007 (26%)	1010 (26%)	1875 (48%)	75	2.03
pyul		992 (24%)	1053 (25%)	2162 (51%)	82	2.19



original annotations **after reannotation**

Figure (i): Intron site histories with splice site shifts. The displayed numbers correspond to the number of intron context histories with at least one 5'-exonization (**E5**), 5'-intronization (**I5**), 3'-exonization (**E3**) and 3'-intronization (**I3**). (Diagram drawn with the Venn tool from Yves van de Peer's group at `http://bioinformatics.psb.ugent.be/webtools/Venn/`.)

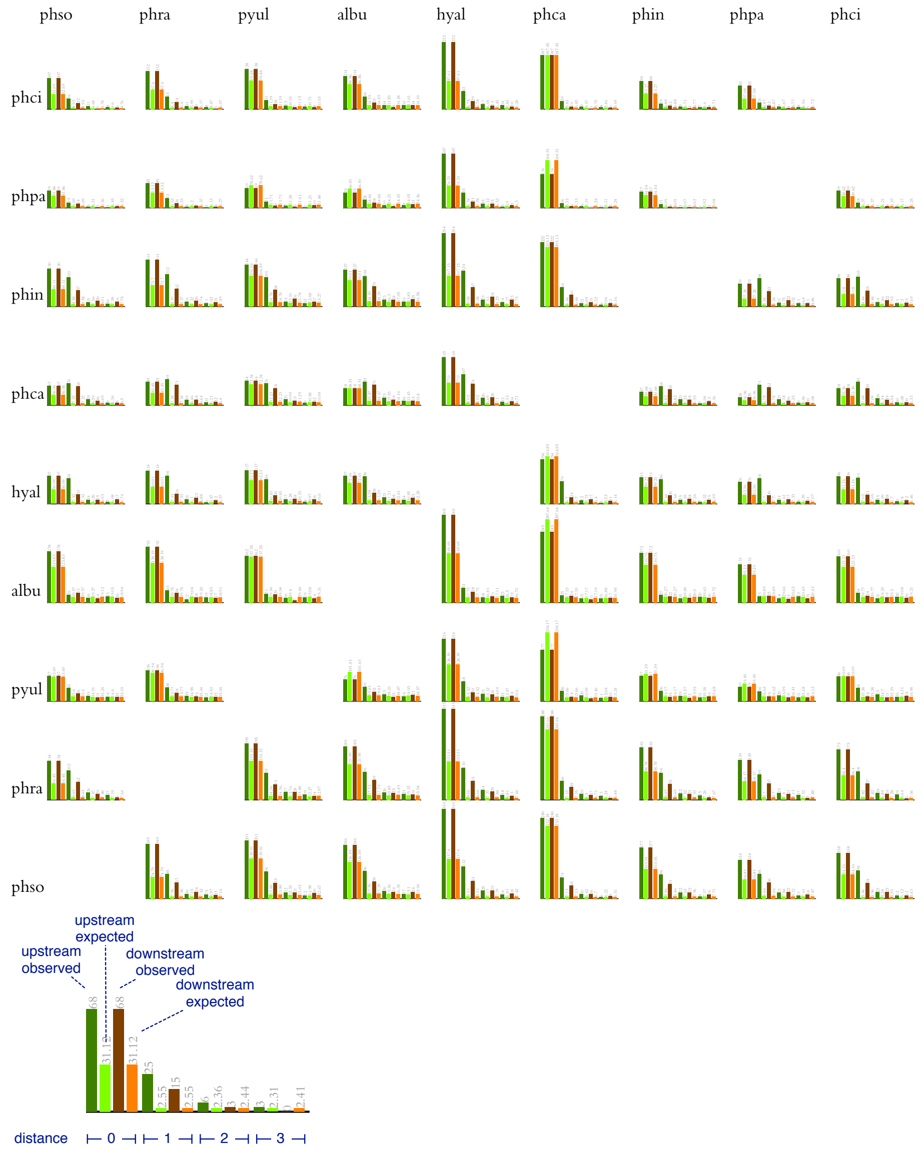


Figure (ii): Distance between phase-1 splice sites (of the row's proteins) and closest upstream or downstream indel in pairwise alignments (with the column's ortholog). [Zoom in for numbers]

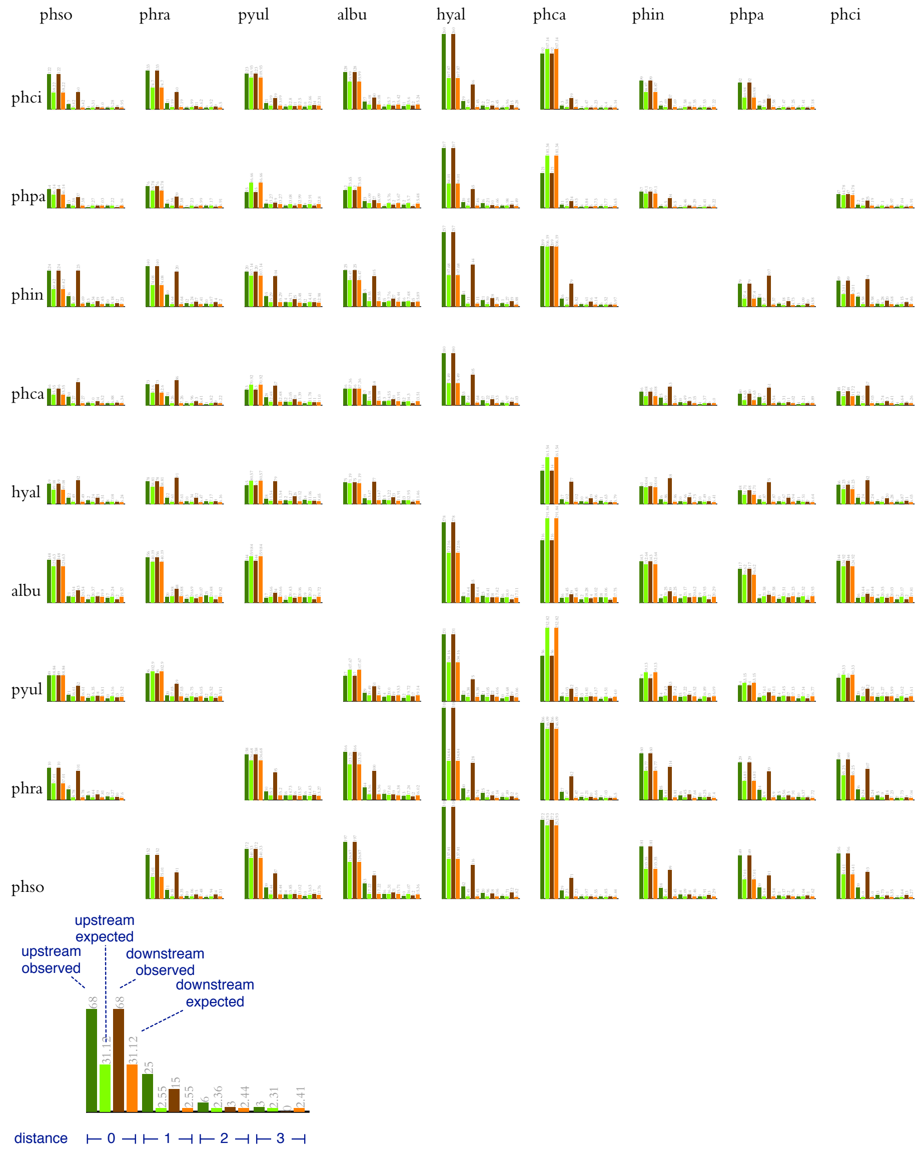


Figure (iii): Distance between phase-2 splice sites (of the row's proteins) and closest upstream or downstream indel in pairwise alignments (with the column's ortholog). [Zoom in for numbers]

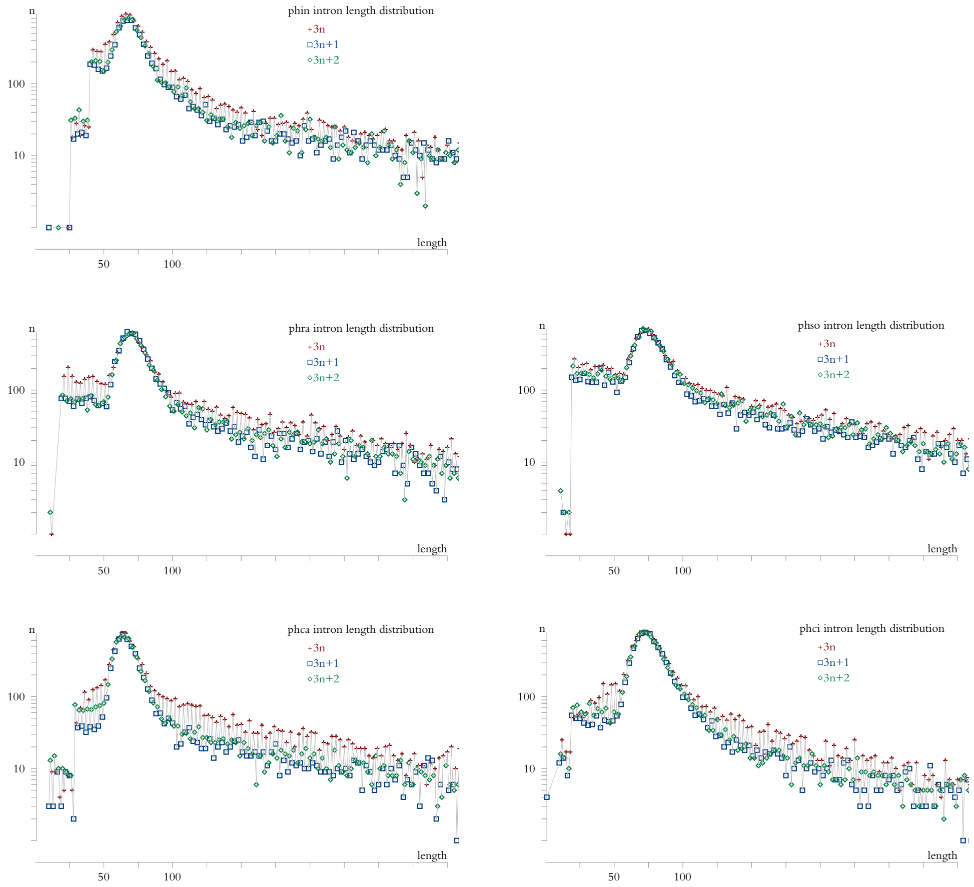


Figure (iv): Intron length distributions with 3*n*-length excess. The log-scaled Y axis plots the number of introns with the given length on the linearly scaled X axis.

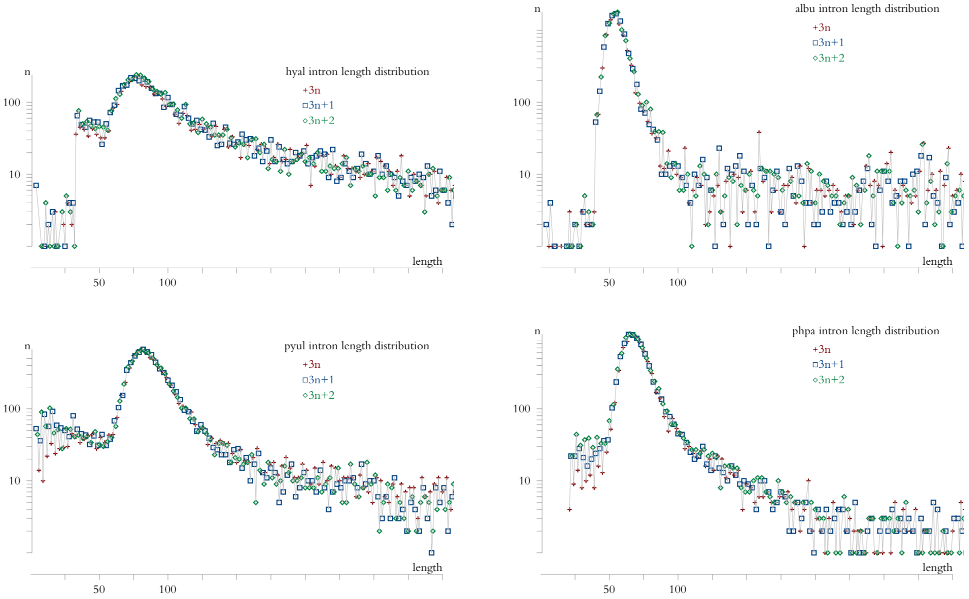


Figure (v): Intron length distributions without conspicuous excess of $3n$-introns.

References

Bentley, J. (1984). Programming pearls: algorithm design techniques. *Communications of the ACM*, **27**(9), 865–873.

Cartwright, R. A. and Graur, D. (2011). The multiple personalities of Watson and Crick strands. *Biology Direct*, **6**, 7.

Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, UK.

Ghiurcuta, C. G. and Moret, B. M. E. (2013). Evaluating synteny for improved comparative studies. *Bioinformatics*, **30**, i9–i18.

Sankoff, D. and Rousseau, P. (1975). Locating the vertices of a Steiner tree in arbitrary metric space. *Mathematical Programming*, **9**, 240–246.