Transposon-Derived and Satellite-Derived Repetitive Sequences Play Distinct Functional Roles in Mammalian Intron Size Expansion

Dapeng Wang1,*, Yao Su1,2,*, Xumin Wang1, Hongxing Lei1,# and Jun Yu1,#

1CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100029, P.R. China. 2Graduate University of Chinese Academy of Sciences, Beijing 100049, P.R. China.

*Corresponding author email: leihx@big.ac.cn; junyu@big.ac.cn

*These authors contributed equally to this work.

Abstract

Background: Repetitive sequences (RSs) are redundant, complex at times, and often lineage-specific, representing significant “building” materials for genes and genomes. According to their origins, sequence characteristics, and ways of propagation, repetitive sequences are divided into transposable elements (TEs) and satellite sequences (SSs) as well as related subfamilies and subgroups hierarchically. The combined changes attributable to the repetitive sequences alter gene and genome architectures, such as the expansion of exonic, intronic, and intergenic sequences, and most of them propagate in a seemingly random fashion and contribute very significantly to the entire mutation spectrum of mammalian genomes.

Principal findings: Our analysis is focused on evolutional features of TEs and SSs in the intronic sequence of twelve selected mammalian genomes. We divided them into four groups—primates, large mammals, rodents, and primary mammals—and used four non-mammalian vertebrate species as the out-group. After classifying intron size variation in an intron-centric way based on RS-dominance (TE-dominant or SS-dominant intron expansions), we observed several distinct profiles in intron length and positioning in different vertebrate lineages, such as retrotransposon-dominance in mammals and DNA transposon-dominance in the lower vertebrates, amphibians, and fishes. The RS patterns of mouse and rat genes are most striking, which are not only distinct from those of other mammals but also different from that of the third rodent species analyzed in this study—guinea pig. Looking into the biological functions of relevant genes, we observed a two-dimensional divergence; in particular, genes that possess SS-dominant and/or RS-free introns are enriched in tissue-specific development and transcription regulation in all mammalian lineages. In addition, we found that the tendency of transposons in increasing intron size is much stronger than that of satellites, and the combined effect of both RSs is greater than either one of them alone in a simple arithmetic sum among the mammals and the opposite is found among the four non-mammalian vertebrates.

Conclusions: TE- and SS-derived RSs represent major mutational forces shaping the size and composition of vertebrate genes and genomes, and through natural selection they either fine-tune or facilitate changes in size expansion, position variation, and duplication, and thus in functions and evolutionary paths for better survival and fitness. When analyzed globally, not only are such changes significantly diversified but also comprehensible in lineages and biological implications.

Keywords: transposable elements, satellite sequences, intron size, mammalian genomes
Introduction
Repetitive sequence (RS) elements are characterized as multi-copied sequences in two broadly defined classes: satellite sequences (SSs), including both micro-satellites and mini-satellites, and transposable elements (TEs) that are characterized based on sequence identity and structure, biogenesis, insertion site preference, and degree of redundancies.\(^1,2\) The RSs are evolutionarily active and show significant influences on the structures of genes and genomes, and are thus highly relevant to biological functions.\(^3,4\)

It has been reported that TE-free regions are negatively selected for certain regulatory elements throughout vertebrate genomes, although the conservation of the sequence contents is often variable.\(^5,6\) Furthermore, TEs have different distributions among exonic, intronic, and intergenic regions.\(^7\)

Indeed, a small number of TE classes are still active, generating population differentiation,\(^8\) and the compositional dynamics of genomic sequences exhibits step-by-step evolutionary changes as a consequence of competitions between host genomes and parasitic sequences.\(^3\) In addition, TE transposition often serves as a driving force for the conversion of introns into exons or gaining novel introns as well as alternatively spliced transcripts.\(^9–11\) Therefore, new sequence integration and the balance of exons and introns in number, length, and ordinal position of a gene provide basic materials for species evolution.\(^12\)

Different subfamilies of TEs have seemingly diverse influences on genes and genomes by changing sequence length to variable extents. Specifically, due to the distinction between “copy-and-paste” of retrotransposons and “cut-and-paste” mostly used by DNA transposons, the former should be a primary player in the event of genome size increase.\(^2\) Introns are considered as the major “warehouse” of TEs\(^11,13\) and certain families of TEs are observed to correlate with functional genes, such as between mammalian interspersed repeats (MIRs) and immune genes.\(^13\) Exploiting the relationship between sequence composition and polymorphism, we noticed that minimal introns (introns in a minimal size range) have unique features distinct from larger introns and demonstrated how these smaller introns escape from TE-driven insertions and also largely free from SS-driven intron expansion.\(^14–16\)

As many vertebrate genomes have now been sequenced, we are able to address more questions on TE- and SS-driven intron expansions in different vertebrate lineages. In particular, we would like to understand how intron expansion relates to gene functions among the three subgroups of mammals—primates, large mammals, and rodents—and what are the roles of mutation and natural selection played in the course of genome evolution.

Results
Intron size increase often involves lineage-specific changes in RS contents in the context of genes
To investigate the relationship between intron size and repeat insertion in a comparable fashion, we divided introns into ten size intervals for the convenience of in-depth analysis since in general introns tend to cluster at certain size ranges (Fig. 1). According to the relationships among shape-variable curves from the three repeat types, retrotransposons, DNA transposons, and satellites, we found that RSs of the twelve mammals fell into two basic patterns. The first pattern is SS-rich, including three rodent species and two primitive mammals, and its repeat abundance ranks as retrotransposon > satellite > DNA transposon. The second pattern, including the rest of the seven mammals, has a repeat content order of retrotransposon > satellite > DNA transposon. The abundance of retrotransposons is lower than those of satellites in zebrafish, frog, and anole, and the abundance of retrotransposons is higher than...
that of satellites but the mode of slope remains the same in chicken and the mode of slope changes into descending in all twelve mammals.

We subsequently tried to find the major TE families that influence intron size in each vertebrate species or lineages by calculating the fraction of introns possessing a particular RS class (Table 1). First, SINEs are supreme in overall abundance among all TEs in mammals. In the primates, Alu and MIR are most abundant. In the two small rodents, mouse and rat, B1, B4, and B2 are most abundant, whereas in guinea pig, the larger rodent of the group, B1 and B4 are most abundant. Second, for the four most abundant TE families in each species, the four large mammals, cow, panda, horse, and elephant, share MIR, L1, and L2, as well as other species-specific TEs that include BovA for cow, tRNA-Lys for panda, and SINE:SINEs that are specific for horse and elephant. MIR is abundant in all twelve mammals; opossum and platypus rank as the top two but the three rodents appear behind all the rest mammals. Third, the three lower vertebrates, chicken, anole, and frog, have CR1, Sauria, and Harbinger as the most abundant TEs, respectively. Zebrafish appears to have the most diverse DNA transposons and they are all quite abundant: DNA:DNA, hAT, hAT-Charlie, TcMar-Tc1, En-Spm, hAT-Ac, and Harbinger. Fourth, concerning satellite sequence classes, we found that all SSs are prevalent in the sixteen vertebrates but mouse, rat, zebrafish, and opossum are more SS-rich among all.

Figure 1. Percentage of introns with retrotransposons, DNA transposons, and satellites.

Notes: The fractions of introns with repeats are displayed over intron length intervals. The ten intervals of intron lengths are defined as: 1, (50–150); 2, (151–300); 3, (301–600); 4, (601–1000); 5, (1001–1400); 6, (1401–2000); 7, (2001–3000); 8, (3001–5000); 9, (5001–10000); and 10, (10001+).
Table 1. Percentage of introns with classified into repetitive families.

| Class/family                  | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|-------------------------------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| DNA:Chapaev-Chap3            | –  | –  | –  | –  | –  | –  | –  | 21%| 19%| 16%| –  | –  | –  | 13%| –  | –  |
| DNA:DNA                      | –  | –  | –  | –  | –  | –  | –  | –  | –  | –  | –  | –  | 35%| –  | –  | –  |
| DNA:En-Spm                   | –  | –  | –  | –  | –  | –  | –  | –  | –  | –  | –  | –  | 14%| –  | –  | –  |
| DNA:Harbinger                | 21%| 19%| 16%| 21%| 12%| 11%| 12%| –  | –  | –  | –  | –  | –  | 21%| 11%| –  | –  |
| DNA:MER1_type                | –  | –  | 21%| 19%| 16%| 21%| 12%| 11%| 12%| 11%| 12%| 13%| –  | –  | –  | –  | –  |
| DNA:T2                       | –  | –  | –  | –  | –  | –  | –  | –  | –  | –  | –  | –  | 16%| –  | –  | –  |
| DNA:ToMar-Tc1                | 10%| –  | –  | –  | –  | –  | –  | –  | –  | –  | –  | –  | –  | 15%| –  | –  | –  |
| DNA:ToMar-Tigger             | 10%| –  | –  | –  | –  | –  | –  | –  | –  | –  | –  | –  | 23%| –  | –  | –  | –  |
| DNA:hAT                      | 21%| –  | –  | 19%| 20%| –  | –  | –  | –  | 17%| –  | –  | 11%| 12%| 20%| –  | –  |
| DNA:hAT-Ac                   | 21%| –  | –  | 19%| 20%| –  | –  | –  | –  | 17%| –  | –  | 11%| 12%| 20%| –  | –  |
| DNA:hAT-Charlie              | 21%| –  | –  | 19%| 20%| –  | –  | –  | –  | 17%| –  | –  | 11%| 12%| 20%| –  | –  |
| LINE:CR1                     | –  | –  | –  | –  | –  | –  | –  | –  | –  | 23%| 19%| –  | –  | –  | –  | –  | –  |
| LINE:L1                      | 27%| 27%| 27%| 27%| 27%| 27%| 27%| 27%| 27%| 27%| 30%| –  | –  | –  | –  | –  | –  |
| LINE:L2                      | 27%| 27%| 25%| 21%| 25%| 29%| 26%| –  | –  | 13%| 36%| 54%| –  | –  | –  | –  | –  |
| LINE:Penelope                | –  | –  | –  | 13%| 20%| –  | –  | –  | –  | 13%| –  | –  | 12%| –  | –  | –  | –  |
| LINE:RTE                     | 14%| –  | –  | 13%| 20%| –  | –  | –  | –  | 13%| –  | –  | 12%| –  | –  | –  | –  |
| LINE:RTE-BovB                | 14%| –  | –  | 13%| 20%| –  | –  | –  | –  | 13%| –  | –  | 12%| –  | –  | –  | –  |
| LTR:ERV1                     | –  | –  | –  | –  | –  | –  | –  | 18%| –  | –  | –  | –  | –  | –  | –  | –  | –  |
| LTR:ERV1-MaLR                | 14%| –  | –  | 13%| 13%| –  | 11%| –  | –  | 19%| 17%| 13%| –  | –  | –  | –  | –  |
| LTR:MaLR                     | –  | 13%| 13%| –  | 11%| –  | 19%| 17%| 13%| –  | –  | –  | –  | –  | –  | –  | –  |
| SINE:Alu                     | 49%| 49%| 51%| –  | –  | –  | –  | 27%| 26%| 27%| 20%| 24%| 17%| 18%| 44%| 40%| 26%|
| SINE:B1                      | –  | –  | –  | –  | –  | –  | 41%| 35%| 37%| –  | –  | –  | –  | –  | –  | –  | –  |
| SINE:B2                      | –  | –  | –  | –  | –  | –  | 31%| 30%| –  | –  | –  | –  | –  | –  | –  | –  | –  |
| SINE:B4                      | –  | –  | –  | –  | –  | –  | 32%| 30%| 26%| –  | –  | –  | –  | –  | –  | –  | –  |
| SINE:Bv-A                    | –  | –  | –  | 36%| –  | –  | –  | –  | –  | 12%| 22%| –  | –  | –  | –  | –  | –  |
| SINE:ID                      | 35%| 36%| 36%| 29%| 33%| 37%| 35%| 15%| 13%| 21%| 62%| 57%| –  | –  | –  | –  | –  |
| SINE:Mir                     | 35%| 36%| 36%| 29%| 33%| 37%| 35%| 15%| 13%| 21%| 62%| 57%| 12%| 13%| 16%| 14%| 34%|
| SINE:SinE                    | 49%| 49%| 51%| –  | –  | –  | 27%| 26%| 27%| 20%| 24%| 17%| 18%| 44%| 40%| 26%| 31%|
| SINE:SinE                    | 49%| 49%| 51%| –  | –  | –  | 27%| 26%| 27%| 20%| 24%| 17%| 18%| 44%| 40%| 26%| 31%|
| SINE:SinE                    | 49%| 49%| 51%| –  | –  | –  | 27%| 26%| 27%| 20%| 24%| 17%| 18%| 44%| 40%| 26%| 31%|
| SINE:Sauria                  | –  | –  | –  | 16%| –  | –  | –  | –  | –  | –  | –  | –  | 15%| –  | –  | –  | –  |
| SINE:Trna-Glu                | –  | –  | –  | 34%| –  | –  | –  | –  | –  | –  | –  | –  | 12%| –  | –  | –  | –  |
| SINE:Trna-Lys                | –  | –  | –  | 34%| –  | –  | –  | –  | –  | –  | –  | –  | 12%| –  | –  | –  | –  |
| Satellite:Satellite          | –  | –  | –  | 34%| –  | –  | –  | –  | –  | –  | –  | –  | 12%| –  | –  | –  | –  |

Notes: The percentages are fractions of introns with the selected repeats over all introns in the listed species and only those greater than 10% are showed in the table. The species codes are: 1, human; 2, orangutan; 3, macaque; 4, cow; 5, panda; 6, horse; 7, elephant; 8, mouse; 9, rat; 10, guinea pig; 11, opossum; 12, platypus; 13, chicken; 14, anole; 15, frog; and 16, zebrafish.
We further identified abundant TE families in each species and have several significant observations (Fig. 2). First, there are near-linear distributions of MIR in introns with a length range of 150 bp–10,000 bp and rapid accumulations of introns over 10,000 bp in the primate and large mammal lineages. In contrast, there is a drastic slowing-down in the rodents, particularly mouse and rat. Aside from this, slowing gains of MIR are also seen in the two primitive mammals. Second, the trends of L1 and L2 insertions over intron sizes are also interesting; the two curves intersect in the large mammals and primates but do not in opossum, where we observe L1 < L2 before and L1 > L2 after the intersections. Third, the distribution of primate-specific Alu repeats has an up-convex curvature, an indication of early saturation and preferred insertions in relatively small introns as compared to LINEs and other SINEs. The rodent-specific B1, in contrast, has a near-linear distribution and is more prevalent than B2 and B4. SINE:ID, unique to mouse and rat, seems more active in rat than in mouse. Fourth, distinctly different from what in other mammals, L2 in platypus behaves similarly to its MIR.

RS-centric intron expansion involves both size and position effects

To look into distinctive effects of TEs and SSs on intron size and position parameters, we divided introns into four basic classes: TS (both RSs), T (TEs), S (SSs), and N (neither TE nor SS). We focused on three essential intron features: fraction, length, and relative position in a gene. We made the

![Figure 2. Percentage of introns with selected repeat families.](image)

Note: The intron length intervals are defined in the same way as what in Figure 1.
following observations (Fig. 3). First, when plotting the percentage of introns in the four classes, we found that the pattern is rather heterogeneous, i.e., the pri-
mates, the large mammals, and platypus are grouped together in a pattern of T > N > TS > S, showing a transposon-dominant pattern, so is opossum that has a pattern of T > TS > N > S. Second, mouse and rat form their own group, as it is noticed that both have more satellite sequences than other mammals: TS > N > T > S. Third, aside from the dominant TS-free group or N, guinea pig (N > T > TS > S), frog (N > T > TS > S), and chicken (N > T > TS > S) all have more transposons in their introns than satellites. Fourth, anole and zebrafish have a pattern of N > TS > T > S, in a similar path as compared to mouse and rat regardless of N. If we pick a single most abundant RS-containing intron group, TS, T, S, and N, for a species, the fractions are 39.6%, 52.7%, 12.8%, and 72% in mouse, platypus, anole, and chicken, respectively.

We also investigated the size relevance of introns according to two simple size intervals: ≤1000 bp and >1000 bp. Obviously, the absolute majority of introns in N are small, ≤1000 bp, as opposed to the fact that the greater majority of introns in TS and T are larger, >1000 bp. When examining the median length, we found that intron length increase is correlated with the complexity of RS insertions: TS > T > S > N (Fig. 4). We also observed that the TS intron group tends to be near the 5′-end of genes as opposed to the

Figure 3. Percentage of the numbers of the four intron classes.
Note: TS, T, S, and N stand for introns with TE and SS, TE only, SS only, and without any of the two basic types, respectively.
N intron group that tends to be near the 3′-end of the genes in primates, large mammals, rodents, opossum, and frog, as well as that the TS intron group tends to be near the 5′-end of the genes in platypus, chicken, and anole (Fig. 5). The extremely biased distributions are seen in mouse, where the transposon-rich introns tend to be near the 3′-end, and in zebrafish, where all four intron groups show no significant bias.

We further examined both length and position effects for four selected transposons: LTR, LINE, SINE, and DNA. Their intron length medians rank as LTR > DNA > LINE > SINE in the primates, the large mammals, and opossum (Fig. 6). In the three rodents, mouse and rat form a unique league themselves with a length order of DNA > LINE > LTR > SINE, but guinea pig stands alone with a similar pattern to other non-rodent mammals: LTR > DNA > LINE > SINE. In addition, the platypus introns with LTR or DNA transposons tend to be larger in size, in comparison with those of LINE- or SINE-containing introns. In contrast, the chicken introns with LINE tend to be smaller, when compared to those with SINE, DNA or LTR. There are other independent patterns such as LTR > SINE > LINE > DNA and LTR > LINE > SINE > DNA in frog and zebrafish, respectively. An exception is unique to anole, where the order becomes LINE > SINE > DNA when LTR is absent. The most likely reason is the lack of well-classified LTR consensus in the RepeatMasker default library due to high diversity of transposable elements in anole, especially

Figure 4. Length comparison of the four intron classes.
Note: The asterisks indicate significant differences between neighbouring data groups based on Wilcoxon rank sum test and cut-off < 0.05.
when compared to mammals. In the primates, the large mammals, and guinea pig, the median position index ranks as LTR < DNA < LINE < 0, and the introns with SINEs in cow, panda, horse, human, and guinea pig have a slight bias toward 5′-end (data not shown). In both mouse and rat, the introns with DNA transposons have the most 5′-end biases and those with SINEs have the least 5′-end biases. In the two primitive mammals, opossum and platypus, their LTRs and DNA transposons tend to be inserted into introns near the 5′-end. The chicken introns harbouring LTRs or DNA transposons have a stronger bias toward insertions at the 5′-end than those with LINE. The order of the median intron position index for anole is LINE < SINE < DNA < 0. The positional preference for the frog introns is the proximity of 5′-end but that of DNA transposon-containing introns is the weakest. In zebrafish, introns with LINE, SINE or LTR have a stronger 5′-end preference, and those with LTR have the least bias.

Intronic RS-abundance and RS-specificity define characteristic gene functions in different mammalian lineages

We first classified genes in a similar way to what we did for introns: (1) TS, genes have both transposons and satellites in their introns; (2) T, genes have only transposons in their introns; (3) S, genes have
only satellites in their introns; (4) N, genes have neither transposons nor satellites in their introns. In general, we observed an order of TS > N > T > S in chicken and anole, but a different order of TS > T > N > S in the rest vertebrates. When compared the same RS classes from different species, the most abundant four classes for TS, T, S, and N are 83.1% in mouse, 33% in horse, 8.32% in chicken, and 28.4% in chicken, respectively (Fig. 7). Furthermore, we considered functional categorization of the four gene classes in the four mammalian lineages: mammals, primates, large mammals, and rodents. We found diverse development- and transcription-related functions in S and/or N genes, including “embryonic skeletal system development” and “transcription regulator activity” in mammals (Table 2), “negative regulation of neuron differentiation” and “gene expression” in primates (Table 3), “midbrain development” and “regulation of transcription” in large mammals (Table 4), and “inner ear morphogenesis” and “regulation of gene expression” in the rodents (Table 5). There are also lineage-specific and tissue-specific profiles for the expression of these genes. For instance, “hormone activity” of N genes is shared by all the major groups of mammals and “pheromone binding” of S genes is unique to the rodents. There are also genes with immunological functions identified in the primate S (eg, “positive regulation of chronic inflammatory response to antigenic stimulus”) and N genes (eg, “MHC class I receptor activity”), in S genes of the large mammals (eg, “antigen processing

Figure 6. Length comparisons of the four TE-containing intron classes. 
Note: The asterisks indicate significant differences between neighbouring data groups based on Wilcoxon rank sum test and cut-off <0.05.
and presentation”), and in N genes of the rodents (eg, “inflammatory response”). In addition, some TS genes are related to fundamental structures and metabolic functions, including “cytoskeleton” and “protein homodimerization activity” in the mammals, “extracellular matrix structural constituent” and “regulation of cell shape” in the primates, “ATP biosynthetic process” in the large mammals, and “acyltransferase activity”, “protein ubiquitination”, and “phosphoinositide binding” in the rodents. There are also rodent TS genes involved in the nervous system and being response to external stimulus or environment. As to T genes, mitochondrial structure related functions are found in both the primates and the large mammals.

The insertion profiles of TEs and SSs are diverse among the vertebrate genomes

We evaluated the expansion strength of TEs and SSs in introns based on the ratio of the repeat length over the corresponding RS-free length (Table 6). We found that zebrafish has the strongest expansion strength among TS, T, and S genes, whereas chicken has the weakest strength in TS and S genes and anole has the weakest strength in T genes. In the mammals, opossum has the strongest strength in TS and S genes but T genes have the most strength in platypus. A striking observation is the fact that the strength of TS genes is greater than the sum of both T and S genes in the
mammals, and we saw the opposite phenomenon in the non-mammalian vertebrates (Table 6).

When integrating the content of intronic repeats in individual genes based on orthology (unique homologous gene in each species), we discovered different topological structures (Fig. 8). The shared clusters between the two trees are the human-orangutan and the mouse-rat clades, the distant relationship to chicken, and the approximation of zebrafish to placental mammals as compared to the other three non-mammalian vertebrates. With regard to TEs, the primates and the large mammals are remarkably distinct from the rest species and are closer to the mouse-rat clade as compared to guinea pig. With regard to SSs, opossum is clustered with the primates as well as the rodents and the four large mammals rather than the other primitive mammal, platypus.

**Discussion**

Other than whole genome duplication, the complexity of vertebrate genomes builds upon many unique sequence and functional features but one of them is genome expansion that compounds with the expansion of gene and intron sizes. There are three essential ways to increase genome sizes. The first is to increase the number of genes through genome and gene duplications. The second and also the foremost important mechanism is gene size expansion through intron size and number increases. The final way is the expansion of intergenic sequences and auxiliary chromosomal structures. With regard to the diversity of RSs and insertion/expansion mechanisms, we classified intron expansion into two categories: TE-driven and SS-driven, and speculated that they may play distinct roles in the intron size expansion of mammalian genomes. First, the profiles of TE insertions can be classified at levels of species and lineages, such as primates, large mammals, and rodents, and we did observe similar modes within lineages and distinctions among lineages. However, exceptions do exist as the rodents are not always cohesive—guinea pig behaves differently from mouse and rat concerning many RS counts. Second, we would like to emphasize the effect of RS expansion event rather than copy number counts, and we hope to see a clear and direct picture that correlates intron size variation with RS insertion.

In general, both TEs and SSs are reported to be non-randomly distributed among eukaryotic genomes. On one hand, there is strong negative selection to protect essential sequences in genomes for the transmission of basic genetic information.
Table 3. Primate-specific GO term enrichment of the four gene classes.

| Class | GO code     | GO name                                                                 | Human | Orangutan | Macaque |
|-------|-------------|-------------------------------------------------------------------------|-------|-----------|---------|
| T     | GO:0005201  | Extracellular matrix structural constituent *                         |       |           |         |
| TS    | GO:0031965  | Nuclear membrane *                                                     |       |           |         |
| TS    | GO:0008360  | Regulation of cell shape *                                             |       |           |         |
| T     | GO:0019882  | Antigen processing and presentation of exogenous peptide antigen via MHC class II | *     |           |         |
| T     | GO:0019886  | Antigen processing and presentation of peptide or polysaccharide antigen via MHC class II | *     |           |         |
| T     | GO:004004   | ATP-dependent RNA helicase activity *                                   |       |           |         |
| T     | GO:0005125  | Cytokine activity *                                                    |       |           |         |
| T     | GO:0022625  | Cytosolic large ribosomal subunit *                                    |       |           |         |
| T     | GO:0010008  | Endosome membrane *                                                    |       |           |         |
| T     | GO:004308   | Exo-alpha-sialidase activity *                                          |       |           |         |
| T     | GO:0031640  | Killing of cells of another organism *                                  |       |           |         |
| T     | GO:0005765  | Lysosomal membrane *                                                   |       |           |         |
| T     | GO:0042613  | MHC class II protein complex *                                          |       |           |         |
| T     | GO:0032395  | MHC class II receptor activity *                                        |       |           |         |
| T     | GO:0005763  | Mitochondrial small ribosomal subunit *                                 |       |           |         |
| T     | GO:000398   | Nuclear mRNA splicing, via spliceosome *                                |       |           |         |
| T     | GO:0005730  | Nucleolus *                                                            |       |           |         |
| T     | GO:0019887  | Protein kinase regulator activity *                                     |       |           |         |
| T     | GO:0003723  | RNA binding *                                                          |       |           |         |
| T     | GO:0008380  | RNA splicing *                                                          |       |           |         |
| T     | GO:0019843  | rRNA binding *                                                         |       |           |         |
| T     | GO:0005681  | Spliceosomal complex *                                                 |       |           |         |
| T     | GO:0006414  | Translational elongation *                                             |       |           |         |
| T     | GO:0017070  | U6 snRNA binding *                                                     |       |           |         |
| S     | GO:0004869  | Cysteine-type endopeptidase inhibitor activity *                        |       |           |         |
| S     | GO:0044424  | Intracellular part *                                                   |       |           |         |
| S     | GO:0045665  | Negative regulation of neuron differentiation *                         |       |           |         |
| S     | GO:0009887  | Organ morphogenesis *                                                  |       |           |         |
| S     | GO:0002876  | Positive regulation of chronic inflammatory response to antigenic stimulus | *     |           |         |
| S     | GO:0002925  | Positive regulation of humoral immune response mediated by circulating immunoglobulin | *     |           |         |
| S     | GO:0010843  | Promoter binding *                                                     |       |           |         |
| S     | GO:0007519  | Skeletal muscle tissue development *                                   |       |           |         |
| S     | GO:0005164  | Tumor necrosis factor receptor binding *                               |       |           |         |
| N     | GO:0002474  | Antigen processing and presentation of peptide antigen via MHC class I | *     |           |         |
| N     | GO:0007267  | Cell-cell signaling *                                                  |       |           |         |
| N     | GO:0009987  | Cellular process *                                                     |       |           |         |
| N     | GO:0010467  | Gene expression *                                                      |       |           |         |
| N     | GO:0008201  | Heparin binding *                                                      |       |           |         |
| N     | GO:0042309  | Homoiothermy *                                                         |       |           |         |
| N     | GO:0050825  | Ice binding *                                                          |       |           |         |
| N     | GO:0048535  | Lymph node development *                                               |       |           |         |
| N     | GO:0032393  | MHC class I receptor activity *                                        |       |           |         |
| N     | GO:0000122  | Negative regulation of transcription from RNA polymerase II promoter   |       |           |         |

(Continued)
Roles of repetitive sequences in mammalian intron evolution

in a relative shorter evolutionary time scale, such as protein-coding sequences or exons. On the other hand, RSs are indispensable as the prime power and raw materials for genomes to evolve for better fitness, to generate complexity and diversity, and to promote speciation and population dynamics. Therefore, RSs have strong influences on gene expression and regulation indirectly through variations in intron length and content. One mechanism shared by all the studied vertebrates is that both TE and SS insertions increase intron size but the strength of the former is much greater than that of the latter. In fact, after eliminating RS insertions in all introns, we observed that the tendency of length increase in the four intron classes remains the same. In other words, the large introns remain large in size even without RS insertions in all four intron classes and so do small introns. However, the introns of anole and chicken genomes are exceptional, where the intron size definitions may shift or not be clearly distinguishable between large and small when RS insertions are removed from the intron sequences (data not shown). We observed a non-random and unbalanced expansion mechanism of intron size evolution: larger introns tend to grow faster than smaller ones when introns are enlarged to a certain size or over a specific threshold. Furthermore, we investigated relationship and mechanism of TE- or SS-driven intron expansions. Satellites can increase intron size at an early or primitive stage as they change intron size in a relatively limited scale, but transposons are capable of increasing intron size in a larger (such as LINEs) and more massive (such as LTRs in multiple insertions) scale and thus have stronger influence on intron size expansion. Most importantly, we observed a synergy between TE-driven and SS-driven insertions, providing a greater degree of intron expansion.

To understand the possible roles of RS families on gene and intron size expansions, we paid special attention on intron length and positioning within a transcript and on functional enrichment in the context of TE- vs. SS dichotomy among species and lineages. For instance, we found that TS-containing introns have a 5′-end bias in all vertebrates but zebrafish and that the RS-free (or the N class) introns have a 3′-end bias in all mammals but platypus. We have recently identified distinct functional profiles of genes at different evolving rates in primates, large mammals, and rodents, and in this study we used a similar classification scheme to investigate protein-coding genes with RS-driven intron expansion. For instance, DNA transposon-containing introns tend to be smaller in fraction, larger in size, and biased toward 5′-end enrichment in mouse and rat. We also pointed out that genes with TE-free introns are enriched in both development and transcription and genes with SS-containing introns are mostly immunity-related in primates and large mammals. We also extracted function categories in nervous systems for mammalian genes possessing SS-containing introns since microsatellite alterations may lead to neurological disorders. Previous studies proposed

---

**Table 3. (Continued)**

| Class | GO code   | GO name                                      | Human | Orangutan | Macaque |
|-------|-----------|----------------------------------------------|-------|-----------|---------|
| N     | GO:0048663 | Neuron fate commitment                       | *     | –         | *       |
| N     | GO:0005184 | Neuropeptide hormone activity                | *     | –         | –       |
| N     | GO:0004522 | Pancreatic ribonuclease activity             | *     | –         | –       |
| N     | GO:0010552 | Positive regulation of gene-specific transcription from RNA polymerase II promoter | –     | –         | *       |
| N     | GO:0045084 | Positive regulation of interleukin-12 biosynthetic process | *     | –         | –       |
| N     | GO:0045944 | Positive regulation of transcription from RNA polymerase II promoter | –     | –         | *       |
| N     | GO:0050826 | Response to freezing                        | –     | –         | *       |
| N     | GO:0016471 | Vacular proton-transporting V-type ATPase complex | *     | –         | –       |

**Note:** The asterisks indicate significant enrichment of GO terms.

---
that microsatellites are unevenly positioned within different regions of protein-coding genes such as UTRs, exons, and introns, and they may play functional roles in regulating gene expression, splicing, mRNA export, and response to external environment. Most SSs that we studied are microsatellites, and we demonstrated that there are functional biases in SS-insertions, such as promoter-related regulatory genes as one of the major categories. In addition, SSs preferentially reside in heterochromatins at or near centromeres and telomeres, where transcriptional activities are rarely discovered. However, if detected, the genes are usually development-related and involved in epigenetic regulation and DNA methylation; the latter two lead to the alteration of chromatin state and may in turn regulate the expression of SS-containing noncoding RNAs. 

We concluded that combined or independent effects of species/lineage-specific TEs and SSs may play an important role in functional differentiations of intron-containing protein-coding genes. At present, the sequence-similarity-based RS library is mostly composed of known TEs, especially the collection of mammal-specific sequences. As increasing number of completed high-quality non-mammalian vertebrate genomes are being sequenced, together with the help of de novo identification technologies, there should be more novel species-specific TEs discovered, adding stronger validation power to the current study.

It is vital for us to track down the precise timing of intron evolution and expansion, such as in a context of lineages, especially the number of introns per gene and the length variation of introns. Spliceosomal introns are the great majority in vertebrate genomes, albeit opposing hypotheses on the origin of introns, “intron-early” and “intron-late”, which argue that introns of this particular type is either more ancient or late comers. Further analyses on genomes based on taxonomy suggested that intron loss is the dominant phenomenon with position-
Table 5. Rodent-specific GO term enrichment of the four gene classes.

| Class | GO code   | GO name                              | Mouse | Rat | Guinea pig |
|-------|-----------|--------------------------------------|-------|-----|------------|
| TS    | GO:0015629| Actin cytoskeleton                   | *     |     |            |
| TS    | GO:0008415| Acalytransferase activity            |       | *   |            |
| TS    | GO:0045177| Apical part of cell                  | *     |     |            |
| TS    | GO:0006915| Apoptosis                            | *     |     |            |
| TS    | GO:0030424| Axon                                 | *     |     |            |
| TS    | GO:0008013| Beta-catenin binding                 |       | *   |            |
| TS    | GO:0059755| Carbohydrate metabolic process       | *     |     |            |
| TS    | GO:0007049| Cell cycle                           | *     |     |            |
| TS    | GO:0051301| Cell division                        | *     |     |            |
| TS    | GO:0042995| Cell projection                      | *     |     |            |
| TS    | GO:0009986| Cell surface                         | *     |     |            |
| TS    | GO:0016568| Chromatin modification               | *     |     |            |
| TS    | GO:0007777| Condensed chromosome kinetochore     | *     |     |            |
| TS    | GO:0016023| Cytoplasmic membrane-bounded vesicle|       | *   |            |
| TS    | GO:0031410| Cytoplasmic vesicle                  | *     |     |            |
| TS    | GO:0009897| External side of plasma membrane     |       | *   |            |
| TS    | GO:0031012| Extracellular matrix                 |       | *   |            |
| TS    | GO:0005925| Focal adhesion                       |       | *   |            |
| TS    | GO:0005525| GTP binding                          |       | *   |            |
| TS    | GO:0005096| GTPase activator activity            | *     |     |            |
| TS    | GO:0004386| Helicase activity                    | *     |     |            |
| TS    | GO:0042802| Identical protein binding            | *     |     |            |
| TS    | GO:0030027| Lamellipodium                        | *     |     |            |
| TS    | GO:0016042| Lipid catabolic process              | *     |     |            |
| TS    | GO:0042470| Melanosome                           |       | *   |            |
| TS    | GO:0008168| Methyltransferase activity           | *     |     |            |
| TS    | GO:0005874| Microtubule                          | *     |     |            |
| TS    | GO:008017 | Microtubule binding                 | *     |     |            |
| TS    | GO:0005739| Mitochondrion                        | *     |     |            |
| TS    | GO:0007067| Mitosis                              | *     |     |            |
| TS    | GO:0006397| mRNA processing                      |       | *   |            |
| TS    | GO:0043066| Negative regulation of apoptosis     |       | *   |            |
| TS    | GO:0043025| Neuronal cell body                   |       | *   |            |
| TS    | GO:0005634| Nucleus                              | *     |     |            |
| TS    | GO:0030165| PDZ domain binding                   |       | *   |            |
| TS    | GO:0048471| Perinuclear region of cytoplasm      | *     |     |            |
| TS    | GO:0005777| Peroxisome                           | *     |     |            |
| TS    | GO:0035091| Phosphoinositide binding             | *     |     |            |
| TS    | GO:0043065| Positive regulation of apoptosis     |       | *   |            |
| TS    | GO:0043123| Positive regulation of I-kappaB kinase/NF-kappaB cascade | * | | |
| TS    | GO:0014069| Postsynaptic density                 |       | *   |            |
| TS    | GO:0006813| Potassium ion transport              |       | *   |            |
| TS    | GO:0042734| Presynaptic membrane                 |       | *   |            |
| TS    | GO:0043234| Protein complex                      | *     |     |            |
| TS    | GO:0032403| Protein complex binding              | *     |     |            |
| TS    | GO:0019904| Protein domain specific binding      |       | *   |            |
| TS    | GO:0046982| Protein heterodimerization activity  |       | *   |            |
| TS    | GO:0019901| Protein kinase binding               |       | *   |            |
| TS    | GO:0008104| Protein localization                 |       | *   |            |
| TS    | GO:0008565| Protein transporter activity          |       | *   |            |

(Continued)
Table 5. (Continued)

| Class | GO code | GO name                                      | Mouse | Rat | Guinea pig |
|-------|---------|----------------------------------------------|-------|-----|------------|
| TS    | GO:0016567 | Protein ubiquitination                       | *     | *   | –          |
| TS    | GO:0045449 | Regulation of transcription                   | *     |     | –          |
| TS    | GO:0006974 | Response to DNA damage stimulus              | –     | *   | –          |
| TS    | GO:0042493 | Response to drug                              | –     |     | –          |
| TS    | GO:001666  | Response to hypoxia                          | –     | *   | –          |
| TS    | GO:0007584 | Response to nutrient                         | –     |     | –          |
| TS    | GO:0014070 | Response to organic cyclic substance         | –     |     | –          |
| TS    | GO:0004871 | Signal transducer activity                   | *     |     | –          |
| TS    | GO:0005625 | Soluble fraction                             | *     | *   | –          |
| TS    | GO:0015293 | Sympporter activity                          | –     | *   | –          |
| TS    | GO:0019717 | Synaptosome                                   | *     |     | –          |
| TS    | GO:0005802 | Trans-Golgi network                           | *     |     | –          |
| TS    | GO:0006511 | Ubiquitin-dependent protein catabolic process| *     |     | –          |
| TS    | GO:0004842 | Ubiquitin-protein ligase activity             | –     |     | *          |
| S     | GO:0009653 | Anatomical structure morphogenesis           | –     | *   | –          |
| S     | GO:0001658 | Branching involved in ureteric bud morphogenesis | * |     | –          |
| S     | GO:0045165 | Cell fate commitment                         | *     |     | –          |
| S     | GO:0042733 | Embryonic digit morphogenesis                 | *     |     | –          |
| S     | GO:0060441 | Epithelial tube branching involved in lung morphogenesis | * |     | –          |
| S     | GO:0042472 | Inner ear morphogenesis                      | *     |     | –          |
| S     | GO:0003676 | Nucleic acid binding                         | –     | *   | –          |
| S     | GO:0048709 | Oligodendrocyte differentiation               | –     | *   | –          |
| S     | GO:001569  | Patternning of blood vessels                  | *     |     | –          |
| S     | GO:0005550 | Pheromone binding                            | –     |     | *          |
| S     | GO:0009074 | Epithelial tube branching involved in lung morphogenesis | * |     | –          |
| S     | GO:0008284 | Positive regulation of cell proliferation    | *     |     | –          |
| S     | GO:0010552 | Positive regulation of gene-specific transcription from RNA polymerase II promoter | – | *   | –          |
| S     | GO:0045666 | Positive regulation of neuron differentiation | *     | *   | –          |
| S     | GO:0010468 | Regulation of gene expression                 | *     | *   | –          |
| S     | GO:0048536 | Spleen development                           | –     | *   | –          |
| S     | GO:0030878 | Thyroid gland development                     | –     | *   | –          |
| S     | GO:0016564 | Transcription repressor activity             | *     |     | –          |
| N     | GO:0006935 | Chemotaxis                                   | *     |     | –          |
| N     | GO:0001533 | Cornified envelope                           | *     |     | –          |
| N     | GO:0006952 | Defense response                             | *     |     | –          |
| N     | GO:0042742 | Defense response to bacterium                | *     |     | –          |
| N     | GO:0005615 | Extracellular space                          | *     |     | *          |
| N     | GO:0006954 | Inflammatory response                        | *     | *   | –          |
| N     | GO:0007389 | Pattern specification process                | *     |     | –          |
| N     | GO:004252  | Serine-type endopeptidase activity           | –     |     | *          |

Note: The asterisks stand for significant enrichment of GO terms.

Phase-specificity in modern mammals and perhaps large amount of intron gains occurred at the early stage of animal evolution, and recent study has found several cases of intron gains happened in the ancestor of placental mammals in transposon-derived domestication-related genes. Moreover, gene length is correlated with gene expression levels and breaths and is affected by RS insertions, such as L1 and MIR. Housekeeping genes are often highly-expressed and harbor smaller introns to reduce the processing cost of transcription, including time and energy. In contrast, tissue-specific genes are often lowly-expressed and harbor larger introns, requiring more effective and complex regulatory elements. Our data, based on a RS-centric stratification approach, showed that intron expansion is strongly influenced by not only RS types but also insertion timing, and the latter is manifested as species-specific propagation of distinct RSs. A comparative study concerning the five teleost genomes indicated...
Table 6. Comparisons of incremental ratio of TEs and SSs.

| Species       | TS   | T   | S   | T + S |
|---------------|------|-----|-----|-------|
| Human         | 0.833| 0.612| 0.080| 0.693 |
| Orangutan     | 0.736| 0.574| 0.067| 0.641 |
| Macaque       | 0.700| 0.569| 0.063| 0.632 |
| Cow           | 0.661| 0.435| 0.065| 0.500 |
| Panda         | 0.512| 0.384| 0.050| 0.434 |
| Horse         | 0.545| 0.426| 0.059| 0.486 |
| Elephant      | 0.660| 0.484| 0.094| 0.578 |
| Mouse         | 0.500| 0.346| 0.071| 0.418 |
| Rat           | 0.458| 0.329| 0.078| 0.406 |
| Guinea pig    | 0.302| 0.244| 0.056| 0.301 |
| Opossum       | 0.867| 0.556| 0.103| 0.660 |
| Platypus      | 0.749| 0.623| 0.091| 0.714 |
| Chicken       | 0.090| 0.183| 0.023| 0.205 |
| Anole         | 0.116| 0.126| 0.035| 0.161 |
| Frog          | 0.330| 0.339| 0.081| 0.420 |
| Zebrafish     | 1.202| 1.207| 0.263| 1.471 |

Note: Incremental ratio is defined as X/(1 – X), where X equals to the median length percentage of repeats in introns.

that zebrafish experienced an ancient large-scale RS-induced intron expansion, and RS profiles of such expansion is rather distinct from the other four fishes with relatively lower insertion frequency.40 Based on these observations, we suspect that the RS content diversity that we observed among vertebrate introns or genes may not be straightforward to characterize with regard to precise timing as the samples we used are still in a limited scope. Insertions of both TEs and SSs should avoid making damages to key regulatory sequences, such as the splice sites, the branch point, the polypyrimidine tract, and other uncharacterized functional elements, and have potential co-evolving patterns with neighbouring sequences;41 and in particular, TEs (eg, SINEs) facilitate the splicing of larger introns via the formation of secondary structure in mammals.42 TE- and SS-derived RSs are forced to cluster or locate in intronic regions and seldom occur in core regulatory regions that are constantly under strong positive or negative selections.

Methods

We obtained RepeatMasker repetitive elements and Ensembl gene structure annotation data from UCSC Genome Database FTP server (ftp://hgdownload.cse.ucsc.edu/), including those from human, orangutan, macaque, cow, panda, horse, elephant, mouse, rat, guinea pig, opossum, platypus, chicken, anole, frog, and zebrafish (Table 7). We excluded genes that do not encode proteins or have very short introns (<50 bp) from our analysis. For each gene, we only keep the longest primary transcript and/or that has the largest number of exons. Concerning the possible overlapping regions in different repeat families or sub-families, we only counted once when a sequence is used multiple times and otherwise indicated. We also collected the gene-transcript-protein relationship, protein sequences, and Gene Ontology (GO) annotations from Ensembl web or FTP sites (http://www.ensembl.org, ftp://ftp.ensembl.org), and used Fisher Exact Test to find the enriched GO terms and adopted the Bonferroni corrections with a cut-off of 0.1 to reduce false positive rate. To compare the major phylogenic groups in
mammals, we regarded the four non-mammalian vertebrates as out-group and considered four divisions (some are obviously lineages and others are not): mammal-specific (occurring only in 12 mammals), primate-specific (occurring only in human, orangutan and/or macaque), non-primate large-mammal-specific (occurring only in cow, panda, horse and/or elephant) and rodent-specific (occurring only in mouse, rat and/or guinea pig). We defined normalized position index as \( (2 \times IO - IN - 1)/IN \), where \( IO \) stands for intron order in a gene along the transcription direction and \( IN \) is total intron number in a gene.

In general, we classified repeat elements into two types of transposons or TE (LTR, LINE, SINE and DNA transposon, in which the former three classes are retrotransposon) and satellites or SS (satellite and microsatellite repeats). We prepared orthologous groups using the inflation parameter \( \lambda = 2 \) in popular MCL algorithm (http://micans.org/mcl/) to cluster gene families after a protein-based all-to-all-blast with a cut-off of 1e-5.43 And then we only selected the groups containing 16 genes and each gene can be assigned a species for phylogenetic analyses. Finally, we used the fraction of number and length of introns in a unit of gene to evaluate the contents of transposons and satellites for 357 orthologous genes, which form a high-dimensional vector for each species. Furthermore, we used the modified cosine of vector included angle to measure the distance of compared species vectors,44 and adopted a way similar to classical UPGMA (Unweighted Pair Group Method with Arithmetic Mean) clustering technology.45 In brief, we began with the twelve initial species and combined the nearest two neighbor species into one cluster and considered the center of the two points in the space as the new vector of the new node and then repeated the process until all nodes came into one cluster. We employed TreeView program to visualize the result of the tree-like structure.46

**Author Contributions**

Conceived and designed the experiments: DW, JY, HL. Collected the data: DW, YS, XW. Analysed the data: DW, YS. Contributed to the writing of the manuscript: DW, JY. All authors reviewed and approved of the final manuscript.

**Funding**

The work was supported by grants from Knowledge Innovation Program of the Chinese Academy of Sciences (KSCX2-EW-R-01-04); Natural Science Foundation of China (90919024); Natural Science Foundation of China (30900831); and the National Basic Research Program (973 Program) from the Ministry of Science and Technology of the People’s Republic of China (2011CB944100).

**Competing Interests**

Authors disclose no potential conflicts of interest.

**Disclosures and Ethics**

As a requirement of publication author(s) have provided to the publisher signed confirmation of compliance with legal and ethical obligations including but not limited to the following: authorship and contributorship, conflicts of interest, privacy and confidentiality and (where applicable) protection of human and animal research subjects. The authors have read and confirmed their agreement with the ICMJE authorship

### Table 7. Species names and the numbers of introns used in this study.

| Short name        | Full name                        | Version | Number of introns |
|-------------------|----------------------------------|---------|-------------------|
| Human             | Homo sapiens, Pongo              | hg19    | 191,918           |
| Orangutan         | pygmaeus abelii                  | ponAbe2 | 108,083           |
| Macaque           | Macaca mulatta                   | rheMac2 | 135,376           |
| Cow               | Bos taurus                       | bosTau4 | 155,350           |
| Panda             | Ailuropoda melanoleuca           | ailMel1 | 136,011           |
| Horse             | Equus caballus                   | equCab2 | 128,897           |
| Elephant          | Loxodonta africana               | loxAfr3 | 127,667           |
| Mouse             | Mus musculus                     | mm9     | 183,175           |
| Rat               | Rattus norvegicus                | r4      | 154,905           |
| Guinea pig        | Cavia porcellus                  | cavPor3 | 119,495           |
| Opossum           | Monodelphis domesticus           | monDom5 | 137,533           |
| Platypus          | Ornithorhynchus anatinus         | ornAna1 | 101,406           |
| Chicken           | Gallus gallus                    | galGal3 | 128,491           |
| Anole             | Anolis carolinensis              | anoCar2 | 122,041           |
| Frog              | Xenopus tropicalis               | xenTro3 | 136,091           |
| Zebrafish         | Danio rerio                      | danRer7 | 207,279           |
and conflict of interest criteria. The authors have also confirmed that this article is unique and not under consideration or published in any other publication, and that they have permission from rights holders to reproduce any copyrighted material. Any disclosures are made in this section. The external blind peer reviewers report no conflicts of interest.

References

1. Toth G, Gaspari Z, Jurka J. Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res. Jul 2000;10(7):967–81.
2. Kidwell MG. Transposable elements and the evolution of genome size in eukaryotes. Genetica. May 2002;115(1):49–63.
3. Hua-Van A, Le Rouzic A, Boutin TS, Filiè J, Capy P. The struggle for life of the genome’s selfish architects. Biol Direct. 2011;6:19.
4. Slotkin RK, Martienssen R. Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet. Apr 2007;8(4):272–85.
5. Simons C, Makunin IV, Makunin IV, Mattick JS. Transposon-free regions in mammalian genomes. Genome Res. Feb 2006;16(2):164–72.
6. Simons C, Makunin IV, Pheasant M, Mattick JS. Maintenance of transposon-free regions throughout vertebrate evolution. BMC Genomics. 2007;8:470.
7. Zhang Y, Romashkin MT, Mager DL. Distributions of transposable elements reveal hazardous zones in mammalian introns. PLoS Comput Biol. May 2011;7(5):e1002046.
8. Mills RE, Bennett EA, Iskow RC, Devine SE. Which transposable elements are active in the human genome? Trends Genet. Apr 2007;23(4):183–91.
9. Sela N, Mersch B, Hotz-Wagenblatt A, Ast G. Characteristics of transposable element exonization within human and mouse. PLoS One. 2010;5(6):e10907.
10. Sela N, Mersch B, Gal-Mark N, Lev-Maor G, Hotz-Wagenblatt A, Ast G. Comparative analysis of transposed element insertion within human and mouse genomes reveals Alu’s unique role in shaping the human transcriptome. Genome Biol. 2007;8(6):R127.
11. Sela N, Kim E, Ast G. The role of transposable elements in the evolution of non-mammalian vertebrates and invertebrates. Genome Biol. 2010;11(6):R59.
12. Zhu L, Zhang Y, Zhang W, Yang S, Chen JQ, Tian D. Patterns of exon-intron architecture variation of genes in eukaryotic genomes. BMC Genomics. 2009;10:47.
13. Sironi M, Menozzi G, Comi GP, et al. Gene function and expression level influence the insertion/fixation dynamics of distinct transposon families in mammalian introns. Genome Biol. 2006;7(12):R120.
14. Yu J, Yang Z, Kibukawa M, Paddock M, Passey DA, Wong GK. Minimal introns are not “junk”. Genome Res. Aug 2002;12(8):1185–9.
15. Zhu J, He F, Wang D, et al. A novel role for minimal introns: routing mRNAs to the cytosol. PLoS One. 2010;5(4):e10144.
16. Wang D, Yu J. Both size and GC-content of minimal introns are selected in human populations. PloS One. 2011;6(3):e17945.
17. Tillot M, Boissinot S. The transposable element profile of the anolis genome: How a lizard can provide insights into the evolution of vertebrate genome size and structure. Mob Genet Elements. Jul 2011;1(2):107–11.
18. Wong GK, Passey DA, Huang Y, Yang Z, Yu J. Is “junk” DNA mostly intron DNA? Genom Res. Nov 2000;10(11):1672–8.
19. Wong GK, Passey DA, Yu J. Most of the human genome is transcribed. Genome Res. Dec 2001;11(12):1975–7.
20. Vinogradov AE. Intron-genome size relationship on a large evolutionary scale. J Mol Evol. Sep 1999;49(3):376–84.
21. Tiširigos A, Rigoutsos I. Alu and b1 repeats have been selectively retained in the upstream and intronic regions of genes of specific functional classes. PLoS Comput Biol. Dec 2009;5(12):e1000610.