BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

Newly evolved introns in human retrogenes provide novel insights into their evolutionary roles

Li-Fang Kang[1], Zheng-Lin Zhu[1*], Qian Zhao[1], Li-Yong Chen[2*] and Ze Zhang[1]

Abstract

Background: Retrogenes generally do not contain introns. However, in some instances, retrogenes may recruit internal exonic sequences as introns, which is known as intronization. A retrogene that undergoes intronization is a good model with which to investigate the origin of introns. Nevertheless, previously, only two cases in vertebrates have been reported.

Results: In this study, we systematically screened the human (*Homo sapiens*) genome for retrogenes that evolved introns and analyzed their patterns in structure, expression and origin. In total, we identified nine intron-containing retrogenes. Alignment of pairs of retrogenes and their parents indicated that, in addition to intronization (five cases), retrogenes also may have gained introns by insertion of external sequences into the genes (one case) or reversal of the orientation of transcription (three cases). Interestingly, many intronizations were promoted not by base substitutions but by cryptic splice sites, which were silent in the parental genes but active in the retrogenes. We also observed that the majority of introns generated by intronization did not involve frameshifts.

Conclusions: Intron gains in retrogenes are not as rare as previously thought. Furthermore, diverse mechanisms may lead to intron creation in retrogenes. The activation of cryptic splice sites in the intronization of retrogenes may be triggered by the change of gene structure after retroposition. A high percentage of non-frameshift introns in retrogenes may be because non-frameshift introns do not dramatically affect host proteins. Introns generated by intronization in human retrogenes are generally young, which is consistent with previous findings for *Caenorhabditis elegans*. Our results provide novel insights into the evolutionary role of introns.

Background

Retroposition, or RNA-based duplication, is the process by which reverse-transcribed mRNAs are inserted into new genomic positions, which generates retrocopies [1]. Retrocopies are assumed not to carry the regulatory regions, but by chance they may obtain functions by recruiting new regulatory elements, and then become functional retrogenes [2-7]. These newly evolved genes may acquire introns in the untranslated regions by capture of nearby exons into a new genomic environment or fusion with host genes, which is chimerization based on intron gain [3-8]. Such retrogenes are usually considered to be intronless because introns were not inherited from the parents. However, in some circumstances, retrogenes may recruit internal exonic sequences as introns [9,10], which is known as intronization [11].

Since intronization of retrogenes was first reported [9], this kind of evolutionary event has been commonly observed in plants. In *Arabidopsis* and *Populus*, 29 retrogenes have undergone intronization, which represent about 15.3 % of all known retrogenes [10]. In contrast, rare cases are reported in vertebrates [12,13]. Previously, only two retrogenes were found to be intronized in mammals [14]. This frequency is extremely low given the thousands of retrocopies in the human (*Homo sapiens*) genome [15-17]. How general retrogene intronization is remains unknown. In the present study, we scanned the human genome for intronized retrogenes

* Correspondence: zhuzl@cqu.edu.cn; mzkcly@yahoo.com.cn
[1]College of Life Sciences, Chongqing University, Chongqing 400044, China
[2]Department of Anesthesiology, Research Institute of Surgery, Daping Hospital, Third Military Medical University, 10 Changjiang Zhilu, Chongqing 400042, China

and identified nine cases not reported previously. Our results provide new insights into the mechanism of intron gain and expression patterns of retrogenes.

Methods

Scanning for intron gain in retrogenes

The human genome data were downloaded from the UCSC Genome Browser database (release hg19) [18,19]. Then, we used the approach of Zhu et al. [10] to search the data for retrocopies. First, we mapped human protein sequences onto the genome with tBLASTn [20] and used the Pseudopipe package [21] to process the raw alignments with the default settings, including tBLASTn *e*-value cutoff (1e-10), coverage cutoff (70 %) and identity cutoff (40 %). Next, we retained candidates with more than three introns absent or only one or two introns absent but with a small K_s (<2) or other RNA-based duplication evidence, for example, a poly(A) track. Finally, as described previously [10], we set filters to discard possible DNA-based duplication cases. In brief, we discarded all retrocopies in which at least 50 % of the region overlapped with repeats or with flanking genes similar to the parental gene's flanking regions. We also discarded all retrocopies that aligned well with the introns of the parents. Ultimately, we identified 3436 retrocopies.

We wrote a series of PERL programs to look for intron-containing retrogenes on the basis of annotations from ENSEMBL (GRCh37) [22,23]. We identified 54 candidates of intronized retrogenes for further study.

Gene structure validation by transcription evidence

We utilized the mRNA and EST annotations from the UCSC Genome Brower Database to search for transcription evidence of intron gain in retrogenes [18,19]. For each sample, we inspected the annotated intronic region to see whether there were transcripts that support its splicing. If transcripts were present, we mapped them on the human genome with BLAT [24] to check whether these transcripts uniquely correspond to the retroposed region. By this method, eight intron-containing retrogenes were validated (Additional files 1 and 2).

K_a and K_s calculation

We estimated the non-synonymous substitution rate (K_a), synonymous substitution rate (K_s) and K_a/K_s values between the intronic regions of retrogenes and their parental copies, by implementing the codeml program in the PAML package following the Nei-Gojobori method [25,26] and analyzed the results with the likelihood ratio test. We did K_a/K_s estimation between the exonic regions of retrogenes and their parental copies in the same way.

RT-PCR

In order to validate the structure of the retrogenes, we collected samples of 16 human tissues from Daping Hospital, Chongqing, for experiments (Additional file 3). Following the manufacturer's instructions, we used TRIzol Reagent (Invitrogen, Carlsbad, CA) to isolate RNA and digested the contaminating genomic DNA with RNase-free DNase I (Promega, Madison, WI). cDNAs were synthesized with Moloney murine leukemia virus reverse transcriptase (Promega). We performed PCR in a 25 μl reaction volume, and 5 μl of the PCR products were electrophoresed on a 1.2 % agarose gel. To validate whether the smaller-sized bands represented the retrogenes, we cloned and sequenced those PCR products. Ultimately, we identified two samples in which the sequences of the smaller-sized bands belonged to retrocopies and the larger bands to the parental genes (Additional file 4).

Peptide support for intronized retrogenes

To identify whether one retrogene was expressed at the protein level, we sought peptide evidence in the PeptideAtlas [27-29] and PRIDE [30,31] databases using the gene name. Each search result displayed experimental details including the fractionation and sequencing (by mass spectroscopy or other methods) of short peptides. Among the results, we extracted peptides that matched the protein sequence of the intronized retrogene. Given that one peptide may match many proteins, we also used BLASTp [32,33] to ensure that the peptide specifically mapped to the gene we targeted. We only retained peptides for which the best hit was a targeted protein.

Age estimation of the retrogenes

We examined the presence and absence of orthologs in the phylogenetic tree for vertebrates and used the established origination times of all human genes [34] to infer the times of origin of the retrogenes. For comparison we used the same method to estimate the time of origin of 27 retrogenes that recruited introns by chimerization [8]. We mapped the results on the vertebrate phylogeny (Additional files 5, 6 and 7). The timeline and divergence time of species in the phylogeny were reconstructed based on data from the UCSC Genome Browser database and other sources [19,34-40].

Detection of splicing signals

We detected splicing signals of new introns with SROOGLE [41]. For an intron X, if its upstream exon is Y and downstream exon is Z, we used X and Y to detect signals of the 5′ splice site (SS) and X and Z for that of the branch site (BS), polypyrimidine tract (PPT), and 3′ SS. We performed two detections for each intron; one was performed on the parental gene and the other was done

for the retroposed sequence. The former and latter were considered to represent the status before retroposition and the current status, respectively. Finally, for each detection, we recorded the percentile score for constitutive introns, which was obtained from a data set composed of >50,000 constitutive introns [41], because all introns in our data set showed no evidence for alternative splicing (Additional file 8).

Results

Identification of intron gain in retrogenes

We focused on identifying retrogenes that contain introns and scanned the human genome using a published pipeline [10]. We mapped all human proteins onto the genome with tBLASTn [20] and extracted all possible candidates of retrocopies from among the results with PseudoPipe [21]. Then, we set filters to exclude cases that did not fulfill the properties of retroposition and obtained 3436 retrocopies. Finally, we determined that 54 of the 3436 retrocopies contained introns on the basis of gene structure annotations from ENSEMBL [22,23].

We used two methods to validate the existence of retrogene introns. First, we collected information from the UCSC Genome Browser database [18,19] and found eight cases with confident transcriptional evidence (Additional file 2). Next, we performed experiments to validate the existence of the retrogene introns. Given the high similarity in the flanking regions of new introns for most retro-parental alignments, we designed pairs of primers whose products (Additional file 9) spanned the intronic regions for both the retrogenes and their parental genes. Theoretically, the amplified segments from the retrogenes (without the intronic sequences) would be smaller than those of the parental genes (with the intronic sequences). By this method, we confirmed that two retrogenes contained introns (Additional file 4), one of which was one of the eight retrogenes mentioned above. In total, we identified nine retrogenes that evolved introns in the retroposed regions (Table 1). Our data did not include RNF113B and DCAF12, which were reported in a previous study [14], because the parents of these two retrogenes were lost after the divergence of mammals from vertebrates, whereas our pipeline used parental protein sequences as queries to search for retrocopies. In addition, we discarded POM121L2 and ARPM1, which were suggested to be intronized retrogenes previously [8], because the alignment identities of these retrogenes and their respective parents did not fulfill the criteria set in our pipeline (>40 % identity).

Table 1 Nine human retrogenes that gained introns investigated in this study

Retrogene	Parent	Movement	Intron (+)	Intron (–)	Evidence
TMEM14D	TMEM14B	10 ← 6	1	4	A
RPS3AP5	RPS3A	10 ← 4	1	5	B
XXyac-R12DG2.2	RCN1	13 ← 11	2*	5	B
HSP90B2P	HSP90B1	15 ← 12	2	16	B
HSP90AA4P	HSP90AA1	4 ← 14	3	9	A,B
HSP90AA5P	HSP90AA1	3 ← 14	2	7	B
CSMD3	RPL18	8 ← 19	1	5	B
WBP2NL	SLC25A5	22 ← X	1	3	B
AC019016.1	CSNK1A1	15 ← 5	2*	8	B

In the column 'Movement', '10 ← 6' means a new gene on chromosome 10 is retroposed from a gene on chromosome 6, for example. 'Intron (–)' and 'Intron (+)' are the numbers of intron losses and intron gains in retrocopies, respectively. For 'Evidence', 'A', confirmed by RT-PCR; 'B', supported by convincing transcription evidence. '*' means that the newly evolved intronic regions of XXyac-R12DG2.2 and AC019016.1 could be spliced in two patterns, respectively.

Mechanisms of intron gain in retrogenes

To clarify the intron-gain mechanisms of these retrogenes, we produced protein and nucleotide sequence alignments for the retrogenes and their respective parental genes (Additional files 10 and 11). For RPS3AP5, we observed that its intronic region did not have counterparts in the parental gene. This result indicated that this retrogene did not gain the intron by intronization, but rather by insertion of an external sequence (Figure 1A). Using the inserted sequence as a query for a BLAT [24] search against the human genome, we identified more than five paralogous sequences with identity >95 % and coverage >70 %. The new intron may be derived from one of these paralogs. By checking the genome annotations in the UCSC Genome Browser database [18,19], we found that none of these paralogs were annotated as introns. Thus, the new intron may not have originated by 'reverse splicing', the process by which a spliced-out intronic RNA is inserted into a novel site of one RNA gene transcript by reversal of the splicing reaction [12,42]. The intron may have been created by a mechanism not reported previously.

We observed that three retrogenes (XXyac-R12DG2.2, CSMD3 and WBP2NL) were transcribed in the reverse direction relative to that of their parents. For XXyac-R12DG2.2 there are 10 annotated transcription patterns and introns appeared in four of the 10 patterns (Additional file 2). Taking ENST00000379050 as an example, the retrocopy contained a 170 bp intron, and its splicing donor and acceptor sites ('GT' and 'AG') had reverse counterparts ('AC' and 'AT') in the parental gene (Figure 1B, Additional files 10 and 11). Thus, transcription in the reverse orientation led to the origin of the intron splicing sites. For the remaining three transcription patterns (ENST00000522673, ENST00000519494 and ENST00000330825), the newly evolved intron was

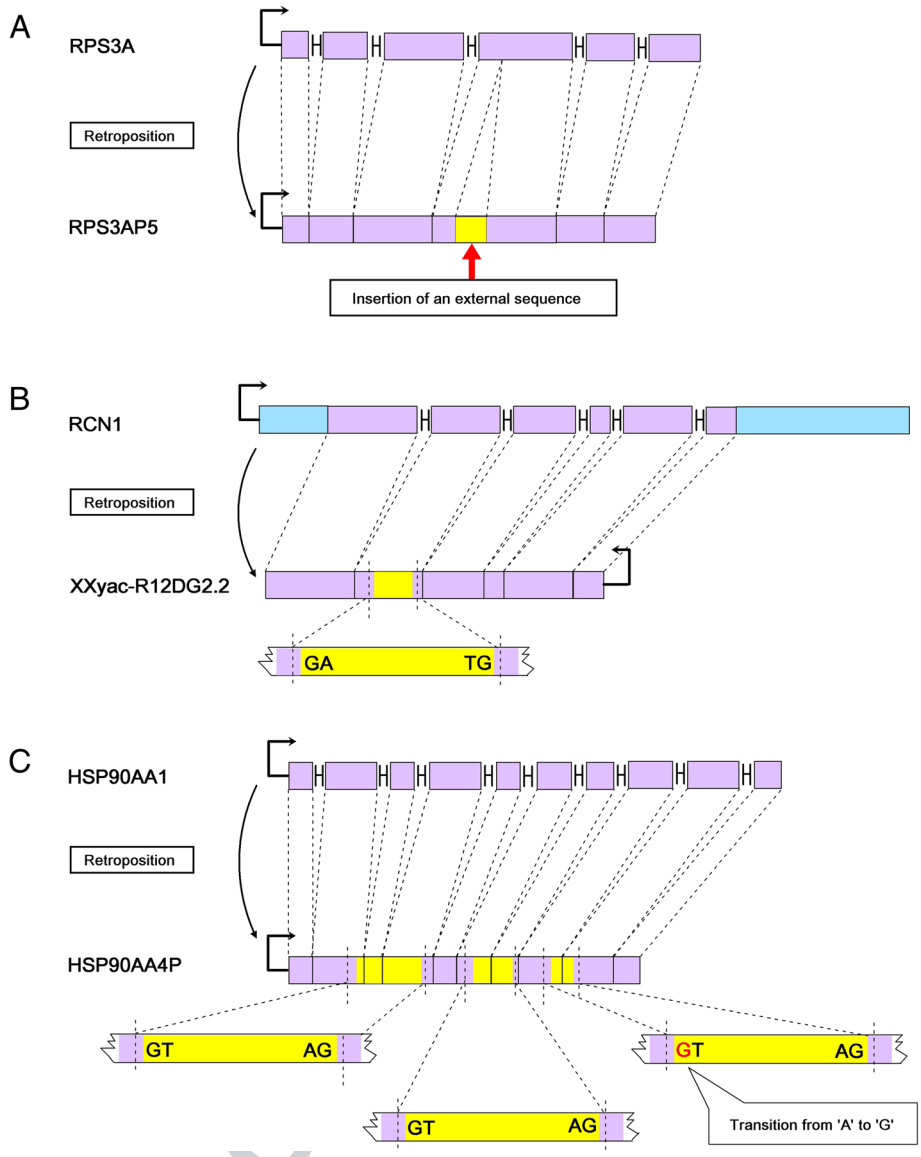

Figure 1 Mechanisms of intron gain in retrogenes. In the parental gene, rectangles represent exons, 'H'-like tags represent introns, the retroposed regions are indicated in purple, and other regions are indicated in blue. In the retrogene, the retroposed region is indicated in purple and the newly evolved intronic regions are indicated in yellow. Semi-rectangle lines with arrows indicate the direction of transcription. (A) The retrogene RPS3AP5 gained an intron by insertion of an external sequence; (B) the retrogene XXyac-R12DG2.2 evolved a new intron after transcription in the opposite orientation compared to the parent; (C) in retrogene HSP90AA4P three new introns were generated by intronization. There is no mutation at the splice sites in the two introns near the 5′ terminus, whereas one transition from 'A' to 'G' (indicated in red) at the splice sites occurred in the intron near the 3′ terminus.

shorter (127 bp) and the retroposed sequence was located near the 3′ end. In addition, the retrocopy is inserted near the 3′ end of a ncRNA gene candidate (LOC 100190939, Additional file 12).

In CSMD3, the retroposed region was located at the 5′ untranslated region (UTR) of the mRNA. Some part of the retrocopy had changed into an intergenic sequence, and some part acted as a portion of an intron (Additional files 2 and 12). The retrogene was located in the first intron of WBP2NL (Additional file 12). Nevertheless, the retrocopy might be transcribed at least some of the time, because an mRNA sequence, BC03789, supports the transcription of this retrogene (Additional file 1 and 2). We did not find evidence for protein-level expression of the three retrogenes that gained an intron after transcription in the reverse orientation. The new introns in these three retrogenes were annotated to be in non-coding regions.

The remaining five retrogenes had gained introns through intronization, which generated 10 new introns. Taking HSP90AA4P as an example, three exonic sequences were changed into introns (Figure 1C). Eight of the 10 introns had the canonical splicing boundaries 'GT-AG'. 80 % (8/10) of the introns arose in ORF and 20 % (2/10) in UTRs.

In total, we observed three mechanisms of intron gain for these retrogenes. In addition to intronization, retrogenes may gain introns after insertion of external sequences or transcription in the opposite orientation compared to the parent (Figure 1).

Non-frameshift introns generated by intronization had greater evolutionary success

For the five retrogenes that underwent intronization, we examined the alignments of retrocopies and their corresponding parental sequences to assess whether these introns had disturbed the frame of putative translation inherited from the parental genes (Additional file 11). If one intron disturbed the frame, we termed it a frameshift intron, otherwise it was considered to be a non-frameshift intron. The lengths of the corresponding sequences of the five retrogenes (70 %) were in multiples of three bases. We performed a manual check for each retrogene. At the location 100 bp upstream of the second intron of HSP90AA4P (from 5′ to 3′, HSP90AA4P-2), we observed an insertion of 23 bases. The length of HSP90AA4P-2 was 83 bp. Thus, compared with the parent, the intron and insertion led to an overall loss of 60 bases (divisible by three) in the transcript. Similarly, for HSP90AA5P we observed an insertion of 22 bases located 1 bp upstream of the intron near the 5′ terminus (HSP90AA5P-1) and a deletion of four bases located 2 bp upstream of the intron near the 3′ terminus (HSP90AA5P-2). The lengths of these two introns were 439 and 254 bp, respectively. As in HSP90AA4P-2, both the indels and intronization shortened the coding sequences by 417 and 258 bp in HSP90AA5P-1 and HSP90AA5P-2, respectively (both numbers are divisible by three). Both were classified as non-frameshift introns. The two alternative spliced introns of AC019016.1 were annotated to be UTR-region introns according to the UCSC database [18,19] and Ensembl [22,23].

In total, eight of the 10 introns created by intronization were non-frameshift introns. This proportion (80 %) is significantly higher than the percentage of frameshift introns generated by chimerization based on intron-gain retrogenes (29.8 %, 16/49) (P-value = 0.017) [8]. From searches of PeptideAtlas [27-29] and PRIDE [30,31], we found that the predicted proteins of HSP90B2P, HSP90AA4P and HSP90AA5P had respective unique matching peptides (Table 2), which indicated the true protein-coding activity of these transcripts. Consistent with findings for *Caenorhabditis elegans* [11], our observations showed that non-frameshift introns had greater evolutionary success.

Retrogenes underwent intronization by cryptic splicing sites

Previous studies showed that most intronizations were caused by base substitutions at the 5′ and 3′ SS [10,11]. However, we observed only four such cases (40 % of all cases) in our data set. By inspecting the EST annotations for the corresponding parental regions of all newly intronized introns, we found that none of these intronized introns was created by inheriting alternative splicing sites from the parental gene. What led to the creation of the other six retrogene introns? Since a retrogene does not contain introns compared with its parental gene, we proposed that the new introns were created by cryptic splice sites in the exonic regions of the parents. That is, cryptic splice sites were silent in the parents, but were activated in the retrogenes after retroposition and the new introns were generated. To test our hypothesis, we used SROOGLE [41] to detect the splicing signals (5′ SS, 3′ SS, the PPT located upstream of the 3′ SS, and the BS located upstream of the PPT) of the retrogene introns and their respective corresponding regions in the parental genes. The splicing signals of introns in four of the six retrogenes were increased, except for those of TMEM14D and HSP90B2P (Figure 2, Table 3). For the latter two retrogenes, in the parental gene the corresponding regions of the retrogene introns had lower splicing signals compared with those of neighboring introns (Additional file 13). It is likely that these cryptic intronic regions were oppressed in the parental genomes and the oppression was released after retroposition. The splice sites of these six new introns pre-existed but were cryptic in the parental genes. After retroposition, the splice sites were activated in the novel genomic environments. In addition, for the four introns that showed base substitutions at their splice sites, the splicing signals increased not only at the 5′ SS and 3′ SS but also at the BS and PPT (Table 3). In addition to point mutation, the change in gene structure after retroposition might also contribute to the evolution of new introns.

Intronization tended to occur in young retrogenes

In *C. elegans*, intronization is reported to be a major contributor to intron creation and most introns generated by this mechanism are young [11]. In our data set, 66.7 % of retrogene introns (10/15) were created by intronization. This finding is consistent with previous studies [11]. We used the established origination times of all human genes to trace the time of origin of intronized retrogenes [34] and examined the presence and absence of the corresponding orthologs in the vertebrates phylogeny (Additional file 6). We found that 80 % (4/5) of the intronized

t2.1 **Table 2 Peptide support for intronized retrogenes**

t2.2 Gene name	Peptide match	Peptide database reference[a]	Location in protein seq	BLASTP hits[b]
t2.3 HSP90B2P	NLNFVKGVVDSGGLSLNVSCETLQQHK	PRIDE: 8670	86	Self (4e-19, 100 %)
t2.4	IEKAMVSQCLTESLCALVASQYGWSGNMER	PRIDE: 8670	270	Self (4e-24, 100 %)
t2.5	AMVSQCLTESLCALVASQYGWSGNMER	PRIDE: 8671; 8668	273	Self (7e-21, 100 %)
t2.6	MAETIQEVEDEYKAFCK	PRIDE: 8672	1	Self (9e-11, 100 %)
t2.7	CVFITDDFRDTMPK	PRIDE: 8669	72	Self (7e-08, 100 %)
t2.8 HSP90AA4P	HNNDEQYAWESSLR	PeptideAtlas: PAp00393519	93	Self (1e-07, 100 %)
t2.9	ADLINNLGTITK	PeptideAtlas: PAp01587648	20	Self (8e-04, 100 %)
t2.10	DQVANSTIVQR	PeptideAtlas: PAp00565957	207	Self (0.005, 100 %)
t2.11 HSP90AA5P	IKEIVKKHSQFIGYPITLFVEKKR	PeptideAtlas: PAp00040955; PAp00423980	33	Self (2e-17, 100 %)
t2.12	HGLEVIYMIELIDKYCVQQLK	PeptideAtlas: PAp00040711	199	Self (2e-15, 100 %)

t2.13 'a', Database name and experiment numbers or identifiers. 'b', BLASTP search against the GenBank non-redundant protein database (e-value and maximum
t2.14 identity of the match are shown in parentheses [32,33]).

retrogenes were primate specific. We also recalculated the ages of 27 chimerizations based on intronized retrogenes with the same method [8] (Additional file 7) and found that only 18.5 % of intronized retrogenes (5/27) were primate specific. This finding indicated that intronization tended to occur in young retrogenes (proportion test, $P = 0.023$). Furthermore, in our data set, no intronized retrogene (0/5) was retroposed from chromosome X ('out-of-X'). The retrogenes from chromosome X were mostly old and evolved after the divergence of eutherian mammals (human or mouse) and marsupials (opossum) [34]. For retrogenes that underwent intron gains by chimerization, the proportion of 'out-of-X' retrogenes was 37 % (Additional file 7). Therefore, the comparison of 0 % and 37 % reinforced the conclusion that intronization tended to occur in young retrogenes.

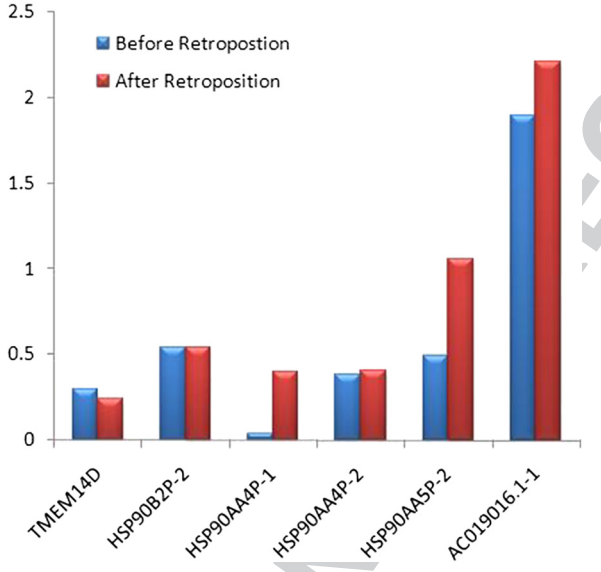

Figure 2 Comparison of splicing signals of retrogene introns (after retroposition) and their corresponding regions in the parental gene (before retroposition). The y-axis is the sum of the percentile scores of four different signals, comprising the branch site (BS), polypyrimidine tract (PPT), and 5′ and 3′ splice sites (Table 3). The higher the score, the stronger the splicing signal. Scores for BS and PPT were calculated with algorithm 'K' [47] and those for the splice sites were calculated with 'M' [48] by SROOGLE [41]. Six different retrogene introns are plotted on the x-axis. If a retrogene evolved only one intron (TMEM14D), we used the gene name to represent the intron, or we marked different introns in one retrogene in the format of the gene name plus a serial number following the hyphen. For example, 'HSP90AA4P-1' represents the first intron (in the direction from 5′ to 3′) in the retrogene HSP90AA4.

Evolutionary rates of intronized retrogenes

To evaluate the evolutionary rates of retrogenes, we calculated K_a, K_s, and K_a/K_s values between the intronic regions of retrogenes and their parental copies as well as between the exonic regions of retrogenes and their parental copies. The K_a values in the intronic regions were higher than those in the exonic regions ($Mean_{intronic} = 0.207$, $Mean_{exonic} = 0.111$, Wilcoxon two-sample test, P-value = 0.098; Table 4). Similarly, K_s values in the intronic regions were higher than those in the exonic regions ($Mean_{intronic} = 0.263$, $Mean_{exonic} = 0.151$, Wilcoxon two-sample test, P-value = 0.194). These findings are consistent with the conclusion that introns evolved faster than exons. T4

In addition, the exonic regions of most intronized retrogenes had K_a/K_s values smaller than 1 (P-value < 0.1), which suggested that the corresponding regions were under negative selection. By checking for evidence of expression, we found that three of the five intronized retrogenes showed evidence for expression at the protein level and the additional two retrogenes showed transcription evidence at the RNA level. This result

Table 3 Percentile scores [40] of splicing signals of retrogene introns (after retroposition) and their corresponding regions in the parental gene (before retroposition)

Intron symbol	Splice sites	After retroposition				Before retroposition			
		BS (K)	PPT (K)	5′ SS (M)	3′ SS (M)	BS (K)	PPT (K)	5′ SS (M)	3′ SS (M)
TMEM14D	GC-AG	0.14	0.06	0.02	0.02	0.14	0.04	0.01	0.11
(HSP90B2P-1)	GT-AG	0.39	0.39	0	0.01	0.39	0.24	0	0
HSP90B2P-2	GT-AG	0.5	0.03	0	0.01	0.5	0.02	0	0.02
HSP90AA4P-1	GT-AG	0.21	0.03	0.04	0.12	0	0	0.04	0
HSP90AA4P-2	GT-AG	0.03	0.06	0.04	0.28	0.03	0.06	0.04	0.25
(HSP90AA4P-3)	GT-AG	0.56	0.33	0	0.03	0.56	0.22	0	0.02
(HSP90AA5P-1)	TT-AG	0.25	0.27	0	0.54	0	0	0	0
HSP90AA5P-2	GT-AG	0.45	0.45	0.01	0.15	0.12	0.15	0.01	0.21
AC019016.1-1	GT-AG	0.91	0.35	0.11	0.84	0.61	0.35	0.11	0.82
(AC019016.1-2)	GT-AG	0.91	0.35	0.47	0.84	0.61	0.35	0	0.82

The higher the score, the stronger the splicing signal is. The scores for BS and PPT were calculated with the 'K' algorithms [47], and those for 5′ SS and 3′ SS were calculated with 'M' [48] by SROOGLE [41]. The intron symbol is in the format of the gene name plus a serial number following the hyphen. For example, 'HSP90B2P-1' indicates the first intron (in the direction from 5′ to 3′) in HSP90B2P. If a retrogene evolved only one intron (TMEM14D), the intron is represented by the gene name. In the column 'Intron symbol', parentheses indicate that the splice sites underwent base substitution.

indicated that most intronized retrogenes were functional and should be under negative selection.

With regard to the three retrogenes that gained introns after transcription in the opposite orientation compared with the parent, they were annotated to be in the non-coding regions of other genes. We observed that CSMD3 and WBP2NL evolved faster than the other retrogenes (Table 4). This finding is consistent with the conclusion that non-coding regions such as UTR regions are under less functional constraint than coding regions. However, XXyac-R12DG2.2 evolved slowly relative to that of CSMD3 and WBP2NL. Thus, XXyac-R12DG2.2 is likely to be under functional constraint.

Discussion

In this study, we systematically searched the human genome for retrogenes that underwent intron gain in the coding region and in total identified 15 retrogene introns. These newly generated introns evolved at a faster rate than neighboring exons. In contrast to the findings in plants [10], we found that intron gain events in retrogenes were rare in humans. In spite of this rarity, the mechanisms of intron creation in these retrogenes are diverse. We found that retrogenes could gain introns in three ways: insertion from an external sequence, transcription in the opposite direction compared with the parent, and intronization. For the latter method, in

Table 4 Substitution rates between the intronic and exonic regions of retrogenes and their corresponding regions of parental genes

Retrogene	Intronic region					Exonic region				
	K_a	K_s	K_a/K_s	*P*-value	Length	K_a	K_s	K_a/K_s	*P*-value	Length
TMEM14D[c]	0.062	0.058	1.074	0.936	105	0.006	0.014	0.440	0.570	237
RPS3AP5[a]	NA	NA	NA	NA	NA	0.017	0.014	1.210	0.172	780
XXyac-R12DG2.2[b]	0.024	0.029	0.830	0.892	129	0.008	0.012	0.643	0.631	813
HSP90B2Pa[c,*]	0.823	0.597	1.379	0.526	144	0.045	0.067	0.678	0.091	2163
HSP90AA4P[c,*]	0.104	0.277	0.374	0.000	744	0.055	0.085	0.656	0.000	1374
HSP90AA5P[c,*]	0.087	0.215	0.406	0.001	672	0.088	0.221	0.400	0.082	897
CSMD3[b]	0.313	0.575	0.544	0.051	291	0.186	0.282	0.659	0.310	225
WBP2NL[b]	0.033	0.088	0.373	0.192	177	0.385	0.377	1.021	0.919	684
AC019016.1[c]	0.083	0.082	1.010	0.978	636	0.081	0.175	0.466	0.084	273

K_a represents the non-synonymous substitution rate and K_s indicates the synonymous substitution rate. The *P*-value was calculated with the likelihood ratio test and the null hypothesis was K_a/K_s =1. NA: not available (the corresponding parental sequence of the new intron in retrogene RPS3AP5 did not exist, because the intron was created by insertion of an external sequence). 'a', The retrogene gained introns by insertion of an external sequence. 'b', The retrogene gained introns after transcription in the opposite orientation compared to the parent. 'c', The retrogene gained introns by intronization. '*', Evidence at the protein level for transcription of the retrogene was obtained.

addition to base substitution, retrogenes also may create introns in exonic regions via cryptic splice sites, which might be activated by the new gene structure after retroposition. Consistent with the findings in *C. elegans* [11], retrogene introns generated by intronization in humans are generally young and are mostly located in the coding region of the new gene. The retrogenes that underwent intronization in coding regions all retained the parental frames of translation and most showed expression evidence at the protein level. The significantly higher percentage of non-frameshift introns implied that this kind of intron possessed a higher likelihood of persistence after intronization. The reason for this may be that frameshift introns mostly have a major effect on the proteins. Thus, non-frameshift introns are more likely to survive. However, non-frameshift introns may be neutral in effect, as proposed previously [43,44]. Furthermore, previous studies have shown that the rate of intron loss is much larger than that of intron gain in mammals [12,13,45]. Consequently, the older the retrogene is, the more probable the retrogene will lose the intronized exon, and this may explain why such introns are mainly observed in young retrogenes.

Some questions arise from careful examination of our observations. For example, for the retrogene RPS3AP5, in which the new intron was created by insertion of an external sequence, the process by which the new intron was created is unknown. In addition, in searches of UCSC [18,19], Ensembl [22,23], PeptideAtlas [27-29] and PRIDE [30,31], we did not obtain evidence of protein-level expression for the three retrogenes that gained introns after transcription in the reverse orientation compared with their parents. The new introns in these three retrogenes were annotated to be in non-coding regions and appeared to be parts of existing intron-containing genes, as described previously [7]. Thus, these retrogenes generally evolved faster than intronized retrogenes (Table 4).

For the eight non-frameshift introns generated by intronization, we examined whether they are under natural selection by checking their genetic variation in different human populations with the 1000 Genomes Browser [46]. However, we did not find insertions, deletions or mutations in splice sites in seven of these retrogenes (file 14), which implied that they are nearly fixed in all populations and may be under negative selection. In addition, there is a possibility that this pattern observed was caused by genetic drift because generation of new introns may be neutral. Finally, what is the importance of producing a shorter protein than the protein from the parent gene? This question may be answered by comparing the functions of the original proteins and that encoded by the retrogenes in the future.

Conclusions

Our results showed that retrogenes may gain introns in three ways: insertion from an external sequence, transcription in the reverse direction compared to that in the parent, and intronization. In addition to base substitution, intronization also may be promoted by cryptic splice sites. For introns generated by intronization, non-frameshift introns might have greater evolutionary success than frameshift introns, because non-frameshift introns have only a small effect on the host proteins or are neutral. Furthermore, intronization tended to occur in young retrogenes.

Additional files

Additional file 1: Transcripts uniquely mapped to retrogenes. This file lists transcripts that spanned the introns of their mapped retrogenes.

Additional file 2: Evidence for transcription of retrogene introns (from the UCSC Genome Browser database). This file contains snapshots from the UCSC Genome Browser database that displays the transcription of retrogenes that gained introns.

Additional file 3: List of human tissues sampled for the experiments. This file lists the human tissues that we used for the experiments to validate the existence of retrogene introns.

Additional file 4: Experimental validation of retrogene introns in TMEM14D and HSP90AA4P. This file shows the experimental results for validating the existence of retrogene introns.

Additional file 5: Phylogenetic tree for vertebrates. A diagram of the phylogenetic tree for vertebrates.

Additional file 6: Chromosome and time of origin of intronized retrogenes. This file shows the origination times of intronized retrogenes.

Additional file 7: Chromosome and time of origin of retrogenes that gained introns by chimerization. This file shows the origination times of retrogenes that gained introns by chimerization.

Additional file 8: Transcription annotations (from the UCSC Genome Browser database) of retrogene introns in the parental gene. This file contains snapshots from the UCSC Genome Browser database displaying transcription annotations of retrogene introns in the parental gene.

Additional file 9: Sequences of primer pairs used to amplify the retrogenes and their parents. A table that lists primer pairs we used to amplify the retrogenes and their parents.

Additional file 10: Protein-level alignments of intron-gain retrogenes ("Sbjct") and their parents ("Query") by GeneWise. This file contains alignments of intron-gain retrogenes and their parents in protein level.

Additional file 11: Nucleotide-level alignments of retrogene introns ('Sbjct', blue and red, splice sites) and parental genes ('Query', program NCBI-BLAST). This file contains alignments of intron-gain retrogens and their parents in DNA level.

Additional file 12: Positions of three retrogenes (XXyac-R12DG2.2, CSMD3 and WBP2NL) in the human genome (from the UCSC Genome Browser database). This file contains snapshots from the UCSC Genome Browser Database displaying the positions of three retrogenes.

Additional file 13: Comparison of splicing signals (percentile score) in the corresponding region of the new intron in the parental gene and neighboring introns. This file shows the results for the comparison of splicing signals in the corresponding region of the new intron in the parental gene and neighboring introns.

Additional file 14: Genetic variation of four retrogenes in different human populations. This file displays alignments of genomes of different human populations in the region of four retrogenes.

Abbreviations

BS: branch site; PPT: polypyrimidine tract; SS: splice site.

Authors' contributions

LFK and ZLZ together carried out the identification of intronized retrogenes and data analysis, and performed the statistical analyses. LFK performed the PCR analysis and helped to draft the manuscript. ZLZ conceived the study, participated in its design and analysis, and drafted the manuscript. QZ helped to perform the data analysis and statistical analyses, participated in the design of the study and helped to draft the manuscript. LYC provided the materials for experiments. ZZ participated in the design of the study and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Many thanks to Prof. Yong Zhang (Institute of Zoology, Chinese Academy of Sciences) and Prof. Tao Sang (Institute of Botany, Chinese Academy of Sciences) for invaluable suggestions and comments, and Dr. Quan-You Yu for help with the experiments. This work was supported by the Fundamental Research Funds for the Central Universities (No. CDJZR11290002) and by Natural Science Foundation Project of CQ CSTC (cstc2012jjB80007).

Received: 30 March 2012 Accepted: 28 July 2012
Published: 28 July 2012

References

1. Brosius J: **Retroposons-seeds of evolution.** *Science* 1991, **251**(4995):753.
2. Betran E, Thornton K, Long M: **Retroposed new genes out of the X in *Drosophila*.** *Genome Res* 2002, **12**(12):1854–1859.
3. Wang W, Brunet FG, Nevo E, Long M: **Origin of sphinx, a young chimeric RNA gene in *Drosophila melanogaster*.** *Proc Natl Acad Sci USA* 2002, **99**(7):4448–4453.
4. Nisole S, Lynch C, Stoye JP, Yap MW: **A Trim5-cyclophilin A fusion protein found in owl monkey kidney cells can restrict HIV-1.** *Proc Natl Acad Sci USA* 2004, **101**(36):13324–13328.
5. Sayah DM, Sokolskaja E, Berthoux L, Luban J: **Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1.** *Nature* 2004, **430**(6999):569–573.
6. Zhang J, Dean AM, Brunet F, Long M: **Evolving protein functional diversity in new genes of *Drosophila*.** *Proc Natl Acad Sci USA* 2004, **101**(46):16246–16250.
7. Baertsch R, Diekhans M, Kent WJ, Haussler D, Brosius J: **Retrocopy contributions to the evolution of the human genome.** *BMC Genomics* 2008, **9**(1):466.
8. Fablet M, Bueno M, Potrzebowski L, Kaessmann H: **Evolutionary origin and functions of retrogene introns.** *Mol Biol Evol* 2009, **26**(9):2147–2156.
9. Lahn BT, Page DC: **Retroposition of autosomal mRNA yielded testis-specific gene family on human Y chromosome.** *Nat Genet* 1999, **21**(4):429–433.
10. Zhu Z, Zhang Y, Long M: **Extensive structural renovation of retrogenes in the evolution of the *Populus* genome.** *Plant Physiol* 2009, **151**(4):1943–1951.
11. Irimia M, Rukov JL, Penny D, Vinther J, Garcia-Fernandez J, Roy SW: **Origin of introns by 'intronization' of exonic sequences.** *Trends Genet* 2008, **24**(8):378–381.
12. Roy SW, Fedorov A, Gilbert W: **Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain.** *Proc Natl Acad Sci USA* 2003, **100**(12):7158–7162.
13. Coulombe-Huntington J, Majewski J: **Characterization of intron loss events in mammals.** *Genome Res* 2007, **17**(1):23–32.
14. Szczesniak MW, Ciomborowska J, Nowak W, Rogozin IB, Makalowska I: **Primate and rodent specific intron gains and the origin of retrogenes with splice variants.** *Mol Biol Evol* 2011, **28**(1):33–37.
15. Emerson JJ, Kaessmann H, Betran E, Long M: **Extensive gene traffic on the mammalian X chromosome.** *Science* 2004, **303**(5657):537–540.
16. Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H: **Emergence of young human genes after a burst of retroposition in primates.** *PLoS Biol* 2005, **3**(11):e357.
17. Vinckenbosch N, Dupanloup I, Kaessmann H: **Evolutionary fate of retroposed gene copies in the human genome.** *Proc Natl Acad Sci USA* 2006, **103**(9):3220–3225.
18. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ: **The UCSC Table Browser data retrieval tool.** *Nucleic Acids Res* 2004, **32**(Database issue):D493–496.
19. Kuhn RM, Karolchik D, Zweig AS, Wang T, Smith KE, Rosenbloom KR, Rhead B, Raney BJ, Pohl A, Pheasant M, Meyer L, Hsu F, Hinrichs AS, Harte RA, Giardine B, Fujita P, Diekhans M, Dreszer T, Clawson H, Barber GP, Haussler D, Kent WJ: **The UCSC Genome Browser Database: update 2009.** *Nucleic Acids Res* 2009, **37**(Database issue):D755–761.
20. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**(17):3389–3402.
21. Zhang Z, Carriero N, Zheng D, Karro J, Harrison PM, Gerstein M: **PseudoPipe: an automated pseudogene identification pipeline.** *Bioinformatics* 2006, **22**(12):1437–1439.
22. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyras E, Gilbert J, Hammond M, Huminiecki L, Kasprzyk A, Lehvaslaiho H, Lijnzaad P, Melsopp C, Mongin E, Pettett R, Pocock M, Potter S, Rust A, Schmidt E, Searle S, Slater G, Smith J, Spooner W, Stabenau A, Stalker J, Stupka E, Ureta-Vidal A, Vastrik I, Clamp M: **The Ensembl genome database project.** *Nucleic Acids Res* 2002, **30**(1):38–41.
23. Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L, Coates G, Fairley S, Fitzgerald S, Fernandez-Banet J, Gordon L, Graf S, Haider S, Hammond M, Holland R, Howe K, Jenkinson A, Johnson N, Kahari A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Lawson D, Longden I, Megy K, Meidl P, Overduin B, Parker A, Pritchard B, Rios D, Schuster M, Slater G, Smedley D, Spooner W, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wilder S, Zadissa A, Birney E, Cunningham F, Curwen V, Durbin R, Fernandez-Suarez XM, Herrero J, Kasprzyk A, Proctor G, Smith J, Searle S, Flicek P: **Ensembl 2009.** *Nucleic Acids Res* 2009, **37**(suppl 1):D690–697.
24. Kent WJ: **BLAT–the BLAST-like alignment tool.** *Genome Res* 2002, **12**(4):656–664.
25. Nei M, Gojobori T: **Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions.** *Mol Biol Evol* 1986, **3**(5):418–426.
26. Yang Z: **PAML: a program package for phylogenetic analysis by maximum likelihood.** *Comput Appl Biosci* 1997, **13**(5):555–556.
27. Desiere F, Deutsch EW, Nesvizhskii AI, Mallick P, King NL, Eng JK, Aderem A, Boyle R, Brunner E, Donohoe S, Fausto N, Hafen E, Hood L, Katze MG, Kennedy KA, Kregenow F, Lee H, Lin B, Martin D, Ranish JA, Rawlings DJ, Samelson LE, Shiio Y, Watts JD, Wollscheid B, Wright ME, Yan W, Yang L, Yi EC, Zhang H, Aebersold R: **Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry.** *Genome Biol* 2005, **6**(1):R9.
28. Deutsch EW, Lam H, Aebersold R: **PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows.** *EMBO Rep* 2008, **9**(5):429–434.
29. Farrah T, Deutsch EW, Omenn GS, Campbell DS, Sun Z, Bletz JA, Mallick P, Katz JE, Malmström J, Ossola R, Watts JD, Lin B, Zhang H, Moritz RL, Aebersold R: **A high-confidence human plasma proteome reference set with estimated concentrations in PeptideAtlas.** *Mol Cell Proteomics* 2011, **10**(9):M110.006353.
30. Jones P, Cote RG, Martens L, Quinn AF, Taylor CF, Derache W, Hermjakob H, Apweiler R: **PRIDE: a public repository of protein and peptide identifications for the proteomics community.** *Nucleic Acids Res* 2006, **34**(Database issue):D659–663.
31. Vizcaino JA, Cote R, Reisinger F, Foster JM, Mueller M, Rameseder J, Hermjakob H, Martens L: **A guide to the Proteomics Identifications Database proteomics data repository.** *Proteomics* 2009, **9**(18):4276–4283.
32. Jenuth JP: **The NCBI. Publicly available tools and resources on the Web.** *Methods Mol Biol* 2000, **132**:301–312.
33. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL: **NCBI BLAST: a better web interface.** *Nucleic Acids Res* 2008, **36(Web Server issue)**:W5–9.
34. Zhang YE, Vibranovski MD, Landback P, Marais GA, Long M: **Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the X chromosome.** *PLoS Biol* 2010, **8**(10): e1000494.

35. Wu NW, Jalkanen S, Streeter PR, Butcher EC: **Evolutionary conservation of tissue-specific lymphocyte-endothelial cell recognition mechanisms involved in lymphocyte homing.** *J Cell Biol* 1988, **107**(5):1845–1851.
36. Trusov YA, Dear PH: **A molecular clock based on the expansion of gene families.** *Nucleic Acids Res* 1996, **24**(6):995–999.
37. Thomas JW, Touchman JW: **Vertebrate genome sequencing: building a backbone for comparative genomics.** *Trends Genet* 2002, **18**(2):104–108.
38. Zhao S, Shetty J, Hou L, Delcher A, Zhu B, Osoegawa K, de Jong P, Nierman WC, Strausberg RL, Fraser CM: **Human, mouse, and rat genome large-scale rearrangements: stability versus speciation.** *Genome Res* 2004, **14**(10A):1851–1860.
39. Falkowski PG, Katz ME, Milligan AJ, Fennel K, Cramer BS, Aubry MP, Berner RA, Novacek MJ, Zapol WM: **The rise of oxygen over the past 205 million years and the evolution of large placental mammals.** *Science* 2005, **309**(5744):2202–2204.
40. Waters PD, Delbridge ML, Deakin JE, El-Mogharbel N, Kirby PJ, Carvalho-Silva DR, Graves JA: **Autosomal location of genes from the conserved mammalian X in the platypus (*Ornithorhynchus anatinus*): implications for mammalian sex chromosome evolution.** *Chromosome Res* 2005, **13**(4):401–410.
41. Schwartz S, Hall E, Ast G: **SROOGLE: webserver for integrative, user-friendly visualization of splicing signals.** *Nucleic Acids Res* 2009, **37**(Web Server issue): W189–192.
42. Cavalier-Smith T: **Selfish DNA and the origin of introns.** *Nature* 1985, **315**:283–284.
43. Castillo-Davis CI, Bedford TBC, Hart DL: **Accelerated rates of intron gain/loss and protein evolution in duplicate genes in human and mouse malaria parasites.** *Mol Biol Evol* 2004, **21**(7):1422–1427.
44. Li W, Tucker AE, Sung W, Thomas WK, Lynch M: **Extensive, recent intron gains in *Daphnia* populations.** *Science* 2009, **326**(5957):1260–1262.
45. Roy SW, Gilbert W: **Rates of intron loss and gain: implications for early eukaryotic evolution.** *Proc Natl Acad Sci USA* 2005, **102**:5773–5778.
46. The 1000 Genomes Project Consortium: **A map of human genome variation from population-scale sequencing.** *Nature* 2010, **467**(7319):1061–1073.
47. Kol G, Lev-Maor G, Ast G: **Human-mouse comparative analysis reveals that branch-site plasticity contributes to splicing regulation.** *Hum Mol Genet* 2005, **14**(11):1559–1568.
48. Schwartz SH, Silva J, Burstein D, Pupko T, Eyras E, Ast G: **Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes.** *Genome Res* 2008, **18**(1):88–103.

doi:10.1186/1471-2148-12-128
Cite this article as: Kang *et al.*: Newly evolved introns in human retrogenes provide novel insights into their evolutionary roles. *BMC Evolutionary Biology* 2012 **12**:128.