Identification of Multiple Forms of RNA Transcripts Associated with Human-Specific Retrotransposed Gene Copies

Saori Mori[1], Masaaki Hayashi[1], Shun Inagaki[1], Takuji Oshima[1], Ken Tateishi[1], Hiroshi Fujii[2], and Shunsuke Suzuki[1,2,*]

[1]Epigenomics Division, Frontier Agriscience and Technology Center, Faculty of Agriculture, Shinshu University, Kami-Ina, Nagano, Japan

[2]Department of Interdisciplinary Genome Sciences and Cell Metabolism, Institute for Biomedical Sciences, Interdisciplinary Cluster for Cutting Edge Research, Shinshu University, Kami-Ina, Nagano, Japan

*Corresponding author: E-mail: ssuzuki@shinshu-u.ac.jp.

Accepted: June 27, 2016

Abstract

The human genome contains thousands of retrocopies, mostly as processed pseudogenes, which were recently shown to be prevalently transcribed. In particular, those specifically acquired in the human lineage are able to modulate gene expression in a manner that contributed to the evolution of human-specific traits. Therefore, knowledge of the human-specific retrocopies that are transcribed or their full-length transcript structure contributes to better understand human genome evolution. In this study, we identified 16 human-specific retrocopies that harbor 5′ CpG islands by in silico analysis and showed that 12 were transcribed in normal tissues and cancer cell lines with a variety of expression patterns, including cancer-specific expression. Determination of the structure of the transcripts associated with the retrocopies revealed that none were transcribed from their 5′ CpG islands, but rather, from inside the 3′ UTR and the nearby 5′ flanking region of the retrocopies as well as the promoter of neighboring genes. The multiple forms of the transcripts, such as chimeric and individual transcripts in both the sense and antisense orientation, might have introduced novel post-transcriptional regulation into the genome during human evolution. These results shed light on the potential role of human-specific retrocopies in the evolution of gene regulation and genomic disorders.

Introduction

Duplicated genes are abundant in eukaryotic genomes and thus are presumed to play important roles in evolution. One main mechanism underlying the production of duplicated genes is retrotransposition, which is associated with the reverse transcription of processed mRNA and integration into a new genomic locus. The retrocopies that have lost the ability to code protein due to the accumulation of multiple mutations are called "processed pseudogenes" (Mighell et al. 2000; Balakirev and Ayala 2003) while "retrogenes" are termed for the retrocopies that are expressed retaining a protein-coding function similar or identical to that of the parent genes (McCarrey and Thomas 1987; Charrier et al. 2012).

It is estimated that there are thousands of retrocopies in the human genome and recent studies have revealed that a large number of retrocopies and pseudogenes are expressed in various human tissues and cell lines (Vinckenbosch et al. 2006; Baertsch et al. 2008; Kalyana-Sundaram et al. 2012; Zhang 2013; Guo et al. 2014; Navarro and Galante 2015; Carelli et al. 2016). Interestingly, accumulating evidence shows that these retrocopy or pseudogene transcripts are able to function as noncoding RNAs and modulate their parent genes by post-transcriptional mechanisms. For example, expression of the phosphatase and tensin homolog (*PTEN*) tumor suppressor gene is controlled by the transcription of its processed pseudogene *PTENP1*, which functions as a competing endogenous RNA (ceRNA) that acts as a molecular sponge directly competing for *PTEN*-targeting microRNAs (miRNAs) (Poliseno et al. 2010). Another study in snails showed that neuronal expression of the neuronal nitric oxide synthase (*nNOS*) protein was suppressed by an antisense RNA transcribed from an *NOS* pseudogene establishing a stable RNA–RNA duplex with the parent *NOS* mRNA (Korneev et al. 1999). Furthermore, the transcribed retrocopies are known to be involved in the production of endogenous small interfering RNAs (siRNAs) in mouse oocytes (Tam et al. 2008; Watanabe et al. 2008). These siRNAs are formed by hybridization of the retrocopy transcripts to their complementary parent mRNA and subsequent digestion by Dicer. Thus, retrocopies have the potential to introduce novelty into the control of gene expression in diverse species, including humans.

A large number of studies have reported differences in gene expression across primates, particularly between humans and chimpanzees (Enard et al. 2002; Cheng et al. 2005; Loisel et al. 2006; Warner et al. 2009; Pai et al. 2011). Comparisons of the genomic sequences between humans and other primates have revealed certain characteristic features of the human genome that may explain such gene expression differences, such as the hundreds of large indels (McLean et al. 2011) and the human accelerated regions, that is genomic regions that have acquired significantly high number of nucleotide substitutions specifically in the human lineage (Pollard et al. 2006). However, the impact of human-specific retrocopies that duplicated only after the divergence of the chimpanzee so as to introduce novel mechanisms for the regulation of gene expression during human evolution is not yet fully understood. A number of recent large-scale evolutionary studies of retrocopies including primate genomes described important potential contributions of transcribed retrocopies for species-specific evolution of primate genomes (Zhang 2013; Navarro and Galante 2015; Carelli et al. 2016). Navarro and Galante also described a large set of human-specific retrocopies (Navarro and Galante 2015). Other recent studies reported retrocopy number variations resulting from both gain and loss of retrocopy insertions within natural populations of humans, illuminating potential evolutionary, and functional relevance (Abyzov et al. 2013; Ewing et al. 2013; Schrider et al. 2013; Richardson et al. 2014; Kabza et al. 2015). However, more information is still needed to extend our knowledge of human-specific retrocopies that are actually transcribed and the structure of their full-length transcripts.

In this study, we focused on human-specific retrocopies that harbor 5′ CpG islands (CGIs) in light of the possibility that the emergence of novel CGIs by retrotransposition might have altered gene regulation as well (Suzuki et al. 2011). Because transcription start sites (TSSs) tend to be interspersed in a CGI rather than positioned at one or a few specific sites (Yamashita et al. 2005; Okamura and Nakai 2008), we presumed retrocopies that have a 5′ CGI may be transcribed from the TSSs remaining in their CGIs. Extracting the human CGIs that are not found in the orthologous chromosomes of other primates, we identified 16 human-specific retrocopies and demonstrated that they were mostly transcribed. Furthermore, the determination of the structure of the retrocopy-associated transcripts by rapid amplification of cDNA end (RACE) experiments clarified their TSSs. Multiple forms of the transcripts potentially function as noncoding RNAs involved in post-transcriptional regulation, such as ceRNAs and antisense RNAs.

Results

Identification of the 16 Human-Specific Retrocopies

To identify human-specific retrocopies harboring 5′ CGIs, we extracted newly emergent human CGIs that did not share sequence similarity within any orthologous chromosomes with the chimpanzee, orangutan or gorilla using a BLAST-like alignment tool (BLAT) (Kent 2002). This filtering provided 214 candidate loci for potentially human-specific CGIs. After the manual removal of false-positive calls, such as CGIs regarded as human specific not because of retrotransposition but because of the presence of sequencing gaps in the genomic sequences of all three of these nonhuman primates along with small chromosomal translocation and tandem repeat creation in the human genome, such that 46 CGIs were identified as specifically human. Removing SVA retrotransposons and CpG-rich minisatellites, it appeared that 16 of the 46 CGIs were 5′ CGIs of human-specific retrocopies (fig. 1*a* and *b*, table 1). The chromosomal location of these loci did not exhibit any obvious proximal–distal preference or clear pattern of accumulation to any specific chromosome (fig. 1*a*). Intriguingly, more than half (10/16) of the retrocopies were inserted into gene introns and the sense-antisense orientation of the insertion into host genes had no clear bias (six for sense, four for antisense) (table 1). These intragenic insertions may imply the possibility that they exert an effect on host gene expression. Since these retrocopies are very young, as they

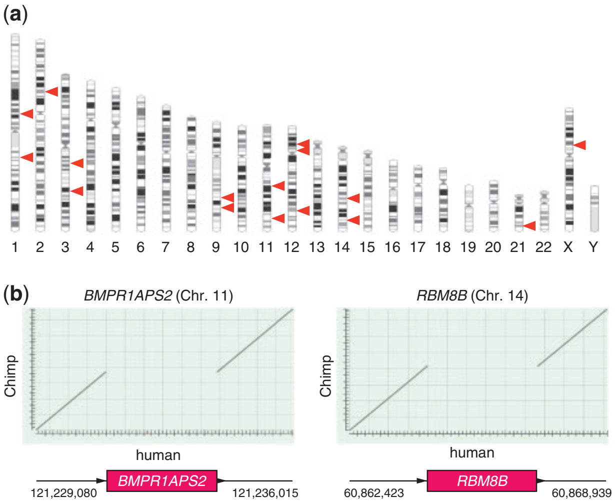

Fig. 1.—Genomic locations of the 16 human-specific retrocopies. (*a*) The locations of the 16 human-specific retrocopies are indicated on the right of the human chromosome by the red arrowheads. The chromosome numbers are also indicated. (*b*) Dot plot analyses between human (*x*-axis) and chimp (*y*-axis) indicate human-specific retrocopy insertions. The chromosome positions are based on GRCh37 and CHIMP2.1.4. The arrowheads indicate the target site duplication (TSD) sequences flanking the insertion sites. Dot plot analyses for other loci are shown in supplementary figure S1, Supplementary Material online.

Table 1

List of 16 Human-Specific Retrocopies

Retrocopy	Location	Length (bp)	Full-length ORF	K_a/K_s	Parent Gene	Host Gene
GCSHP5	chr1: 168024571–168025743	1,173	Yes (no aa change)	–	*GCSH*	*DCAF6* (AS)
C1orf106-r	chr3: 111902199–111904845	2,647	Noncoding	–	*C1orf106*	*SLC9C1* (S)
TMEM183A-r	chr3: 149699451–149701166	1,726	Yes (6 aa change)	–	*TMEM183A*	*PFN2* (S)
ZNF322P1	chr9: 99957623–99962428	4,806	Yes (3 aa change)	–	*ZNF322*	*ANKRD18CP* (S)
RALGAPA1P	chr9: 108282022–108290003	7,982	No	0.912	*RALGAPA1*	*FSD1L* (S)
RPS28P7	chr11: 82400563–82400978	416	Yes (no aa change)	–	*RPS28*	*lincRNA* (AS)
UHRF1-r	chr12: 20704359–20707315	2,957	No	0.182*	*UHRF1*	*PDE3A* (S)
RPL18AP3	chr12: 104656954–104661687	643	Yes (1 aa change)	–	*RPL18A*	*TXNRD1* (S)
RBM8B	chr14: 60864423–60867249	2,827	No	1.807	*RBM8A*	*C14orf39* (AS)
H2BFS	chr21: 44985061–44986086	1,026	Yes (2 aa change)	0.330	*H2BK*	*HSF2BP* (AS)
PHKA1P1	chr1: 91356714–91359574	2,861	No	–	*PHKA1*	Intergenic
EIF3FP3	chr2: 58478566–58479819	1,254	Yes (8 aa change)	0.308*	*EIF3F*	Intergenic
BMPR1APS2	chr11: 121231080–121234015	2,936	No	0.010*	*BMPR1A*	Intergenic
AK4P3	chr12: 31766166–31769527	3,362	Yes (2 aa change)	0.199	*AK4*	Intergenic
Novel-r	chr14: 90991749–90992338	590	Noncoding	–	*Novel*	Intergenic
NUS1P1	chrX: 47369025–47373053	4,029	Yes (1 aa change)	0.106*	*NUS1*	Intergenic

Note—The table summarises the human-specific retrocopies. The chromosome positions of the inserted sequences were based on GRCh37. The retrocopy and parental gene names are from HGNC. The genes without any established name were tentatively named (indicated by gray letters). The K_a/K_s ratio was not able to be calculated at seven loci due to the lack of synonymous or nonsynonymous site. The retro-insertion locations are indicated as intergenic or by the host gene name along with the orientation (S: sense or AS: antisense) to host genes.

*Statistically significant (Fisher's exact test, $P < 0.05$).

became duplicated just after the divergence between the human and chimpanzee lineages, 9 out of 14 copies of the protein-coding genes unsurprisingly still retain an intact open reading frame (ORF) with only a few amino acid replacements (table 1). The K_a/K_s ratio between the retrocopies and their parent genes were less than one in most cases, implying purifying or stabilizing selection for protein-coding sequence of these retrocopies (table 1). Statistical significance (Fisher's exact test, $P < 0.05$) was attained in four retrocopies, *UHRF1-r*, *EIF3FP3*, *BMPR1APS2*, and *NUS1P1*, but not in others unsurprisingly considering that only a small number of nucleotide substitutions were analyzed. This suggests that these retrocopies potentially produce proteins with functions that are mainly conserved but slightly distinct from those of their parents upon transcription. We therefore next analyzed the DNA methylation status in the CGIs located in the 5′ region of the retrocopies.

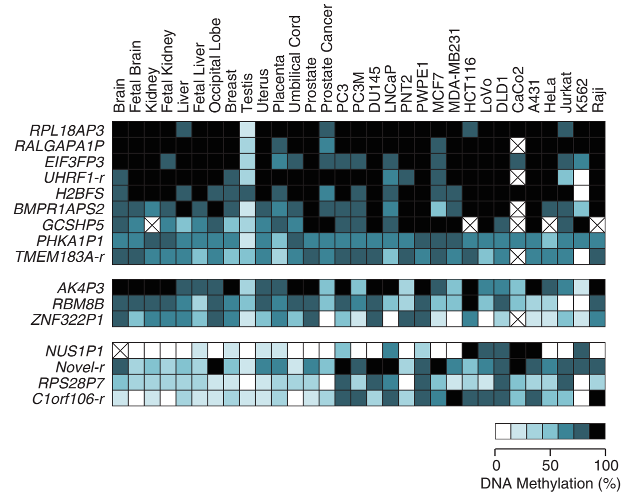

Fig. 2.—Methylation status of the CGIs located in the human-specific retrocopies. Heat map of the methylation status of the 16 human-specific retrocopies. The percentages of methylated DNA in each sample were measured by COBRA. The unmethylated status is shown in white, highly methylated in black and intermediate in a blue gradient.

5′ CpG Island Methylation of the Human-Specific Retrocopies

Although genome-wide DNA methylation data obtained using microarray or next-generation sequencing (NGS) are available in public databases, the extremely close sequence similarity between the retrocopies and their parent genes raises the potential that the 5′ CGIs in the retrotransposed loci have not been clearly distinguished from the original loci in genome-wide data. We therefore performed combined bisulphite and restriction analysis (COBRA), designing one of a pair of PCR primers in the 5′ flanking sequences of the retrocopies to correctly determine the 5′ CGI methylation levels of the retrocopies. The 5′ CGIs in the nine retrocopies were highly methylated in almost all of the tissues and cancer cell lines examined except the testis, consistent with the notion that retrotransposed elements are subject to DNA hypermethylation (fig. 2, upper group). Three other CGIs were highly methylated in normal tissues, but relatively reduced methylation levels were observed in multiple cancer cell lines (fig. 2, the middle group). Interestingly, the four remaining CGIs were hypomethylated in normal tissues, but were highly methylated in certain cancer cell lines (fig. 2, lower group). These results show that not all retrocopies are highly methylated, suggesting that at least some have the potential to be transcribed. Hence, we next performed RT-PCR to determine whether they are indeed actually transcribed.

Transcription of Human-Specific Retrocopies in Normal Tissues and Cancer Cell Lines

To detect the retrocopy transcripts separately from their parent genes or other paralogues, we designed PCR primers to make the 3′-terminus contain a single nucleotide replacement site only in the retrocopies. Performing PCR with genomic DNA using these primer pairs under stringent annealing conditions, the specific amplification of 12 out of the 16 retrocopies was confirmed by restriction fragment length polymorphism (RFLP) analysis or direct sequencing of the PCR products containing the internal retrocopy-specific sequence. In addition, the RT-PCR products were also analyzed in the same manner in order to confirm the retrocopy-specific detection. The specific primer sets were not able to be designed for four retrocopies, *GCSHP5*, *RALGAPA1P*, *RPS28P7*, and *RPL18AP3*, due to the lack of retrocopy-specific sequences. The results of RT-PCR showed that all of the 12 retrocopies examined were transcribed and displayed expression patterns that varied in a tissue and cell line-specific manner. The expression of retrocopies *AK4P3*, *BMPR1APS2*, *H2BFS*, and *EIF3FP3* was clearly higher in the multiple cancer cell lines than in normal tissues (fig. 3). In particular, in the cases of *BMPR1APS2* and *H2BFS* the expression was almost entirely cancer-specific, while *EIF3FP3*, interestingly, was clearly expressed only in the malignant prostate cancer cell lines PC3 and PC3M. On the other hand, *UHRF1-r* was ubiquitously expressed in normal tissues, but seemed to be repressed in many cancer cell lines. These results suggest that these transcribed human-specific retrocopies function as novel genetic elements specifically acquired during human evolution, while dysregulation of these retrocopies expression is associated with certain human diseases, including cancer. However, we wondered if these retrocopies were in fact transcribed from their 5′ CGIs, because no correlation was found between their expression and the 5′ CGI methylation levels. The expression and methylation data from normal tissues may not be fully comparable, because it was not obtained from matched DNA and RNA sets, but the cancer cell line data is not subject to this concern. This prompted us to determine the structure of the transcripts associated with the retrocopies.

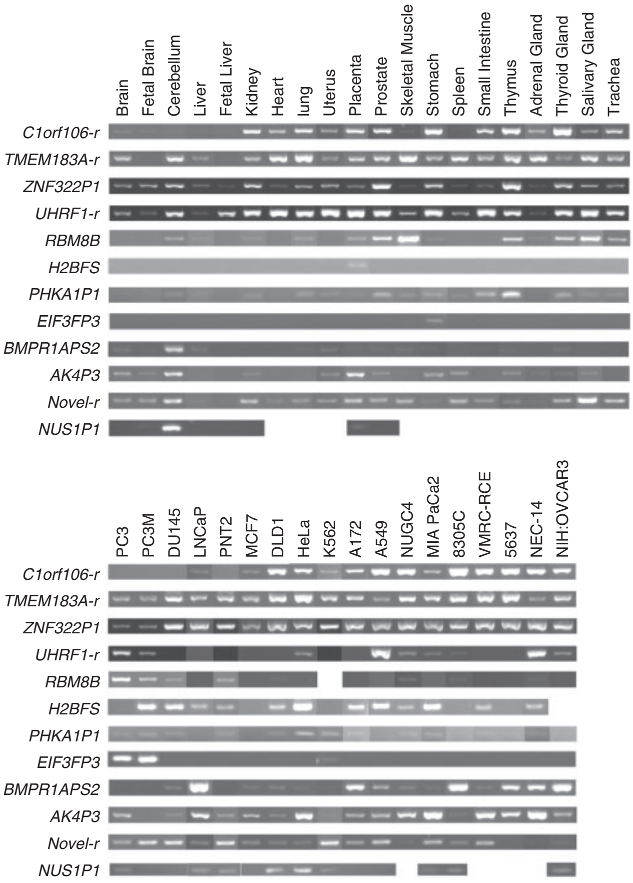

Fig. 3.—Expression pattern of human-specific retrocopies. Expression of human-specific retrocopies was examined by RT-PCR in various tissues and cell lines. GelRed-stained PCR products were run in agarose gel. The primer sequences are shown in supplementary table S2, Supplementary Material online. Gel images were sorted on the basis of tissues and cell lines. (Blank: data not available).

Determination of the Structure of the Transcripts Associated with the Human-Specific Retrocopies

RACE experiments were performed to the structure of the retrocopy-associated transcripts. To avoid any amplification of highly similar parent transcripts, we designed gene-specific primers for RACE to include single nucleotide substitution sites at their 3′ ends, as described in the RT-PCR section. Considering that retrocopy transcription does not necessarily occur in sense orientation, the gene-specific primers were designed in both directions for each RACE experiment. It turned out that three retrocopies were expressed as chimeric transcripts in that the retrocopies were connected to their neighboring genes in the sense or antisense orientation by a pattern of splicing conforming to the GT-AG rule (fig. 4*a*–*c*).

The chimeric transcript associated with *H2BFS* was formed by splicing the connecting exon 8 of *HSF2BP* and the 3′ UTR of *H2BFS* in the antisense orientation (fig. 4*a*). The splice

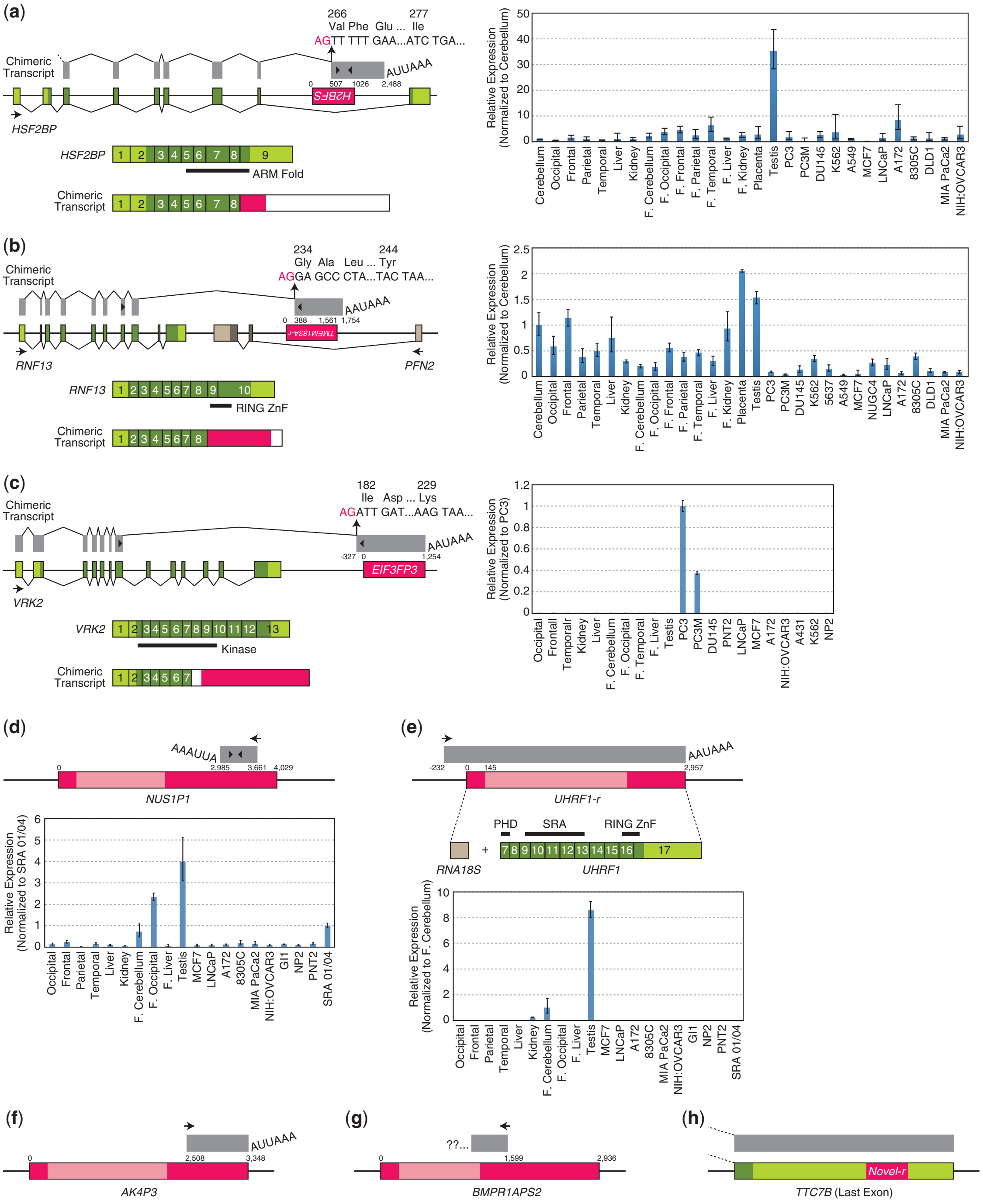

Fig. 4.—Identification of the structure of the retrocopy-associated trasncripts and their expression pattern. Schematic representations of the genomic loci of human-specific retrocopies, (*a*) *H2BFS*, (*b*) *TMEM183A-r*, (*c*) *EIF3FP3*, (*d*) *NUS1P1*, (*e*) *UHRF1-r*, (*f*) *AK4P3*, (*g*) *BMPR1AS2* and (*h*) *Novel-r*. The coding exons are represented as filled boxes and nontranslated sequences are indicated as open boxes. The exons are not drawn to scale. The regions included in the retrocopy-associated transcripts are shown in gray solid boxes. The solid line the major transcript structure and the retrocopy-associate transcript structure. Splicing acceptor site sequences and the corresponding amino-acid sequences in the flanking exon-intron junctions are shown over the extent of the schematic diagrams. The polyA addition signals (AAUAAA or AUUAAA) in retrocopy-associate transcripts are shown. The functional domains found in the chimeric transcripts or parental genes are shown under the schematic diagrams. The arrowheads indicate the position of the primers used in the quantitative RT-PCR. The mRNA levels of the retrocopy-associated transcripts were determined by quantitative RT-PCR. *GAPDH* was used as a control. The primer sequences are shown in supplementary table S2, Supplementary Material online. The error bars represent standard deviations of the relative expression values (*n* = 3–6).

acceptor site in *H2BFS* was located 507 bp inside the 3′ end of *H2BFS* and the chimeric transcript was terminated by a polyadenylation signal sequence located approximately 1,460 bp upstream of the 5′ end of *H2BFS.* In the chimeric transcript, the ORF in exon 9 of *HSF2BP*, encoding part of the armadillo-type fold domain, was truncated, and the sting of 12 amino acids encoded by antisense sequence of *H2BFS* was added. The expression level of the chimeric transcript was remarkably high in the testis compared with other tissues and cell lines, where the expression was nevertheless still detectable (fig. 4*a*).

The chimeric transcript associated with *TMEM183A-r* was formed by splicing the connecting exon 8 of *RNF13* with the 3′ UTR of *TMEM183A-r* in the antisense orientation (fig. 4*b*). The splice acceptor site in *TMEM183A-r* was located 388 bp inside of the 3′ end of *TMEM183A-r* and the chimeric transcript was terminated by a polyadenylation signal sequence located approximately 190 bp upstream of the 5′ end of *TMEM183A-r*. In the chimeric transcript, the ORF in exon 9 and 10 of *RNF13*, encoding a ring-type zinc finger domain, was truncated and the string of 11 amino acids encoded by antisense sequence of *TMEM183A-r* was added. Expression of the chimeric transcript was detected in all the tissues and cancer cell lines examined, but the expression level in cancer cell lines was relatively low (fig. 4*b*).

The chimeric transcript associated with *EIF3FP3* was formed by splicing the connecting exon 7 of *VRK2* with the 5′ flanking region of *EIF3FP3* in the sense orientation (fig. 4*c*). The splice acceptor site in *EIF3FP3* was located 327 bp upstream from the 5′ end of *EIF3FP3* and the chimeric transcript was terminated by a polyadenylation signal sequence located at the 3′ end of *EIF3FP3* itself. In the chimeric transcript, the ORF in exon 8–14 of *VRK2*, encoding part of a kinase domain and transmembrane region, was truncated and the string of 48 amino acids encoded by 5′ flanking region of *EIF3FP3* was added. Interestingly, the chimeric transcript was exclusively expressed in the malignant prostate cancer cell lines PC3 and PC3M (fig. 4*c*).

The four other retrocopies were solely transcribed from the novel TSSs in their 3′ UTRs or the nearby 5′ flanking region (fig. 4*d–g*). *NUS1P1* produced a 977 bp antisense transcript that was transcribed from the 3′ UTR, and its expression level was relatively high in the fetal brain and adult testis (fig. 4*d*). *UHRF1-r* was found to be a retrocopy of the chimeric transcript that consists of *RNA18S* and *UHRF1* lacking exon 1–6 (fig. 4*e*). The transcription of *UHRF1-r* started 232 bp upstream of the 5′ end of the integrated sequence and was terminated by the polyadenylation signal located at the 3′ end. *AK4P3* was transcribed in the sense orientation from the novel TSS in the 3′ UTR, producing a 841 bp RNA (fig. 4*f*). Only the 5′ end was determined for the transcript associated with *BMPR1APS2* and its antisense transcription from the 3′ UTR was shown (fig. 4*g*). The retrocopy of a novel unannotated transcript transcribed from the 5′ region of *PCMT1* in a head-to-head manner was incorporated into the 3′ UTR of *TTC7B* according to the Ensemble database (fig. 4*h*).

Discussion

In the present study, we have identified 16 human-specific retrocopies and determined the transcript structure for 8 of the 12 retrocopies that were confirmed to be transcribed. The form of their transcription varied among the loci, but none of them was transcribed from their 5′ CGIs. One retrocopy was incorporated into the 3′ UTR of the host gene, so it should be transcribed as part of the host gene transcript. Four others were transcribed as sole transcripts in the sense and antisense orientation from novel TSSs in the 3′ UTR and in the nearby 5′ flanking region of the retrocopies. The remaining three were transcribed as chimeric transcripts in which the retrocopies were fused to their neighboring genes. These observations provided good agreement with a number of recent works that comprehensively analyzed transcription and other features of retrocopies in human and other species (Vinckenbosch et al. 2006; Baertsch et al. 2008; Kalyana-Sundaram et al. 2012; Guo et al. 2014; Carelli et al. 2016). In the dataset of expressed retrocopies by Carelli et al., 7 of the 16 retrocopies, *GCSHP5*, *TMEM183A-r*, *RALGAPA1P*, *RPS28P7*, *EIF3FP3*, *AK4P3*, and *NUS1P1*, were also included. The dataset by Baertsch et al. contained 5 of the 16 retrocopies, *RALGAPA1P*, *RPS28P7*, *UHRF1-r*, *RPL18AP3*, and *RBM8B*, and two of them, *UHRF1-r* and *RPL18AP3*, also exist in the dataset by Guo et al. and *RALGAPA1P* and *UHRF1-r* were in the dataset by Kalyana-Sundaram et al. Also the dataset by Vinckenbosch et al. contained 5 of the 16 retrocopies, *GCSHP5*, *TMEM183A-r*, *RPL18AP3*, *RBM8B*, and *NUS1P1.* Interestingly, *GCSHP5* and *RPL18AP3* were appeared in the retrocopy number variation dataset by Schrider et al., indicating the insertions of these two retrocopies are polymorphic (Schrider et al. 2013). In total, the transcription of 10 of the 16 retrocopies has been detected in the other studies based on RNA-Seq and EST data while six cases were only detected in our RT-PCR-based study. We also agreed with the previous studies about the retrocopy transcription from the nearby 5′ flanking region of the retrocopies as well as the promoter of neighboring genes (Vinckenbosch et al. 2006; Baertsch et al. 2008; Carelli et al. 2016). The transcription of *TMEM183A-r* has been previously reported as the origination of a human-specific transmembrane protein gene (Yu et al. 2006). However, the present results show that *TMEM183A-r* was transcribed in the antisense orientation as part of a chimeric transcript (fig. 4*b*), suggesting the previous study also detected this antisense transcription of *TMEM183A-r*. Only human-specific retrocopies that have a 5′ CGI were subject for the detection as we initially presumed these retrocopies might be transcribed from their 5′ CGIs. However, all the retrocopy-associated transcripts identified in this study were indeed transcribed from TSSs independent from their 5′ CGIs.

Hence the high detection rate of the transcripts associated with human-specific retrocopies unlikely due to the presence of 5′ CGI.

Since the retrocopy-associated transcripts presented here are all human specific, their functionalities are potentially related to the acquisition of novel genomic regulation in the evolution of the human genome. At the same time, dysregulation of these transcripts expression can be regarded as potential of diseases including cancer by altering gene regulation. Recent studies have shown that transcribed retrocopies function as ceRNAs that act as molecular sponges or decoys for miRNAs via their preserved miRNA binding sites, thereby upregulating the target genes of miRNAs, including their parent genes. The three retrocopy-associated transcripts identified in this study contained the 3′ UTRs of retrocopies in the sense orientation (fig. 4*c*, *e*, and *f*). These transcripts therefore have the potential to function as ceRNAs to regulate their parent and other genes that share common miRNA binding sites. One such parent gene, *EIF3F*, exhibits a deubiquitinase activity that regulates Notch activation in mice (Moretti et al. 2010). UHRF1 recruits DNMT1 to hemi-methylated regions during DNA replication, maintaining the proper DNA methylation status (Bostick et al. 2007). *AK4* is a progression-associated gene in human lung cancer that promotes metastasis (Jan et al. 2012). Because human-specific retrocopies have an extremely close similarity to their parent genes due to their recent duplication, the miRNA binding sites are well preserved in the retrocopies so there is an increased likelihood that these young human-specific retrocopies function as ceRNAs. Indeed, analysis of the pseudoMap database (http://pseudomap.mbc.nctu.edu.tw, last accessed July 11, 2016) predicted several miRNAs that target both human-specific retrocopies and their parent genes (data not shown).

Antisense transcription of the four retrocopies was also observed (fig. 4*a*, *b*, *d*, and *g*). These transcripts containing antisense retrocopy sequences potentially form an RNA–RNA duplex by hybridization with their complementary parent mRNAs. Consequently, it is possible that the translation of parent genes is inhibited or endogenous siRNAs are produced from these double-stranded RNAs. The parent gene *H2BK* encodes a replication-dependent histone that is a member of the histone H2B family. NUS1, also known as Nogo-B receptor or NgBR, binds farnesylated Ras and recruits Ras to the plasma membrane, a critical step that is required for the activation of Ras signaling in human breast cancer cells and tumorigenesis (Wang et al. 2013). BMPR1A mediates TGF-beta and activin signal transduction, while the function of TMEM183A is currently unknown.

In addition to regulation of the expression of parent genes, the formation of chimeric transcripts regulates their fusion with partner genes. Because these chimeric transcripts are transcribed from the promoter of partner genes, their production leads to a reduction in the normal transcripts of partner genes by partly capturing their transcriptional activity, albeit the expression levels of the chimeric transcripts were not very high. Alternatively, these chimeric transcripts, as well as the individual sense and antisense transcripts, may simply function as long noncoding RNAs (lncRNAs) that positively or negatively regulate their neighboring genes. Indeed, the fusion partner genes of the chimeric transcripts identified in this study are known to have important functions. The E3 ubiquitin ligase gene *RNF13*, the chimeric partner of *TMEM183A-r*, plays a role in spatial learning and assembly of the soluble *N*-ethylmaleimide-sensitive factor-attachment protein receptor (SNARE) complex that controls synaptic function in mice (Zhang et al. 2013). Vaccinia related kinase 2 (VRK2), the chimeric partner of *EIF3FP3*, phosphorylates and stabilises p53, and its binding to the C-terminus region of JNK1 suppresses apoptosis, consistent with the finding that knockdown of VRK2 promotes apoptosis (Blanco et al. 2008; Monsalve et al. 2013). Another potential regulatory effect might result from the production of truncated proteins. As all the chimeric transcripts identified in this study contained incomplete ORFs of partner genes due to the fusion of retrocopies, translation of the chimeric transcripts would result in truncated partner gene proteins lacking the C-terminus regions that contain a number of domains important for protein function. These truncated proteins therefore might exert dominant-negative effects in preventing the activities of wild-type proteins translated from the full-length transcripts of partner genes.

Of course, it should also be considered that some of these retrocopy-associated transcripts may be the product of transcriptional noise and/or not yet under natural selection pressure because these retrocopy insertions are young events. Further investigation by gain-of-function and loss-of-function experiments would clarify precise function of individual transcripts and whether the expression in the cancer cells is directly involved in cause of cancer progression or consequence of transcriptional dysregulation in cancer.

We here report 16 human-specific retrocopies that possess 5′ CGIs and 12 transcripts associated with them. Although their expression levels are low, their potential biological significance as lncRNAs would not depend on the expression level because it is becoming clearer that numerous lncRNAs do have important role in gene regulation despite their much lower expression levels than that of protein-coding genes. There are many more human-specific retrocopies in the human genome (Navarro and Galante 2015). The high detection rate of the retrocopy-associated transcripts supports the hypothesis that the acquisition of human-specific retrocopies has helped create novel regulatory activities in the human genome that are related to evolution and/or health and disease in human-beings. The investigation of the actual roles of human-specific retrocopies in the regulation of the human genome is now open for cell-based functional studies utilising RNA interference (RNAi) and genome editing technologies.

Materials and Methods

Identification of Human-Specific Retrocopies Harboring 5′ CGI

The list of human CGI sequences was downloaded from the table browser of the UCSC Genome Bioinformatics Site (http://genome.ucsc.edu/cgi-bin/hgTables, last accessed July 11, 2016) setting assembly to GRCh37/hg19 and track to CpG Islands. Using BLAT (blatSrc35 parameters: -minIdentity = 90), we checked similar sequences to each human CGI among all of the sequences in the orthologous chromosomes of nonhuman primate species, chimpanzee, orangutan and gorilla. Each human CGI sequence as is in the list was used for query and the entire sequence of the specific orthologous chromosome of nonhuman primate species downloaded from Ensembl (http://www.ensembl.org, last accessed July 11, 2016) was used for database. The 214 human CGIs that did not share sequence similarity with any of the orthologous chromosomes of these three nonhuman primates were extracted as candidate human-specific CGIs. After the manual removal of false-positive calls, such as CGIs regarded as human specific not because of retrotransposition but because of the presence of sequencing gaps in the genomic sequences of all three of these nonhuman primates along with small chromosomal translocation and tandem repeat creation in the human genome, such that 46 CGIs were identified as specifically human. Removing SVA retrotransposons and CpG-rich minisatellites, it appeared that 16 of the 46 CGIs were 5′ CGIs of human-specific retrocopies.

K_a/K_s Analysis

DnaSP (ver. 5.10.1) (Librado and Rozas 2009) was used for the analysis. K_a and K_s were calculated by the Nei–Gojobori method.

DNA Methylation Analysis by COBRA

Genomic DNA was extracted from human cell lines using Trizol (Life Technologies) or a DNeasy Blood and tissue kit (QIAGEN). Human tissue genomic DNA was obtained from BioChain (D1234035, D1234062, D1234086, D1234142, D1234149, D1234260, D1234274, D1234200, D1244035, D1244142, D1244149, D1244272). Genomic DNA was treated with sodium bisulphite solution, as described previously (Frommer et al. 1992; Raizis et al. 1995). After the bisulphite treatment of the genomic DNA, 30 to 35 cycles of PCR were carried out using the primer pairs listed in supplementary table S1, Supplementary Material online. The primers were designed using MethPrimer (Li and Dahiya 2002). The PCR products were digested with 1–10 units of restriction enzymes for 1–2 h at the appropriate temperature for each enzyme (supplementary table S1, Supplementary Material online). The intensity of the cut and uncut bands was quantified with ATTO CS Analyzer 3 software (ATTO).

RT-PCR

Total RNA was extracted from human cell lines using Trizol (Life Technologies) as instructed by the manufacturer. Normal human tissue total RNA was obtained from Clontech (636643, 636527, 636584) and BioChain (R1234142-50, R1234062-50, R1234051-50, R1234149-50, R1234078-50, R1234066-50, R1244039-50, R1244078-50, R1244051-50, R1244062-50, R1244066-50, R1244149-50). Total RNA was treated with DNase I to remove genomic DNA (Promega, M6101). cDNA was synthesised using a Transcriptor First Strand cDNA Synthesis Kit (Roche) with an oligo dT primer. No detection of *β-ACTIN* amplification from the minus RT controls was confirmed for any of the RNA samples examined in this study. Thirty to thirty five cycles of PCR amplification were carried out in 10 μl total volume with 10 ng cDNA using 0.2 U TaKaRa Ex Taq HS (TaKaRa), 4 pmol of each primer and 2 nmol of each dNTP mixture under the following cycle conditions: 96°C for 15 s, 60–71°C for 30 s and 72°C for 15–60 s. PCR products were resolved by gel electrophoreses. The primer sequences are listed in supplementary table S2, Supplementary Material online.

5′ and 3′ RACE

To determine the complete structure of the retrocopy-associated transcripts, we performed 5′ and 3′ RACE using a SMARTer RACE 5′/3′ Kit (Clontech) according to the manufacturer's instructions. The gene specific primer sequences and the derivation of the RNAs that were used for RACE experiments are listed in supplementary table S3, Supplementary Material online. The PCR products were cloned using a pTAC-2 vector (BioDynamics Laboratory) and ECOS-competent *Escherichia coli* DH5α (NIPPON GENE). Plasmids were purified using FastGene Plasmid Mini (NIPPON Genetics) and sequenced.

Quantitative RT-PCR

cDNA was prepared as described in the RT-PCR section. Quantitative real-time polymerase chain reaction (qPCR) was carried out in triplicate in 10 μl volumes containing 25 ng cDNA, 5 nM of the primers and FastStart Essential DNA Green Master (Roche) using LightCycler 96 (Roche). The amplification efficiency was calculated from the standard curve. *GAPDH* was used as the reference gene and the data was analyzed Microsoft Excel. The primer sequences and PCR conditions are listed in supplementary table S2, Supplementary Material online.

Supplementary Material

Supplementary tables S1–S3 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

We thank to Yasuko Hayashi for technical assistance and Nami Suzuki-Hashido for assistance to the K_a/K_s analysis. This work was supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) in Japan, Program to Disseminate Tenure Tracking System (individual selection type) to S.S. and the Ministry of Agriculture, Forestry and Fisheries (MAFF) in Japan, Integration Research for Agriculture and Interdisciplinary Fields to H.F. and S.S. Pacific Edit reviewed the manuscript prior to submission.

Literature Cited

Abyzov A, et al. 2013. Analysis of variable retroduplications in human populations suggests coupling of retrotransposition to cell division. Genome Res. 23:2042–2052.

Baertsch R, Diekhans M, Kent WJ, Haussler D, Brosius J. 2008. Retrocopy contributions to the evolution of the human genome. BMC Genomics 9:466.

Balakirev ES, Ayala FJ. 2003. Pseudogenes: are they "junk" or functional DNA? Annu Rev Genet. 37:123–151.

Blanco S, Sanz-García M, Santos CR, Lazo PA. 2008. Modulation of interleukin-1 transcriptional response by the interaction between VRK2 and the JIP1 scaffold protein. PLoS One 3:e1660.

Bostick M, et al. 2007. UHRF1 plays a role in maintaining DNA methylation in mammalian cells. Science 317:1760–1764.

Carelli FN, et al. 2016. The life history of retrocopies illuminates the evolution of new mammalian genes. Genome Res. 26:301–314.

Charrier C, et al. 2012. Inhibition of SRGAP2 function by its human-specific paralogs induces neoteny during spine maturation. Cell 149:923–935.

Cheng Z, et al. 2005. A genome-wide comparison of recent chimpanzee and human segmental duplications. Nature 437:88–93.

Enard W, et al. 2002. Intra- and interspecific variation in primate gene expression patterns. Science 296:340–343.

Ewing AD, et al. 2013. Retrotransposition of gene transcripts leads to structural variation in mammalian genomes. Genome Biol. 14:R22.

Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, Molloy PL, Paul CL. 1992. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc Natl Acad Sci U S A. 89:1827–1831.

Guo X, Lin M, Rockowitz S, Lachman HM, Zheng D. 2014. Characterization of human pseudogene-derived non-coding RNAs for functional potential. PLoS One 9:e93972.

Jan YH, et al. 2012. Adenylate kinase-4 is a marker of poor clinical outcomes that promotes metastasis of lung cancer by downregulating the transcription factor ATF3. Cancer Res. 72:5119–5129.

Kabza M, et al. 2015. Inter-population differences in retrogene loss and expression in humans. PLoS Genet. 11:e1005579.

Kalyana-Sundaram S, et al. 2012. Expressed pseudogenes in the transcriptional landscape of human cancers. Cell 149:1622–1634.

Kent WJ. 2002. BLAT–the BLAST-like alignment tool. Genome Res. 12:656–664.

Korneev SA, Park JH, O'Shea M. 1999. Neuronal expression of neural nitric oxide synthase (nNOS) protein is suppressed by an antisense RNA transcribed from an NOS pseudogene. J Neurosci. 19:7711–7720.

Li LC, Dahiya R. 2002. MethPrimer: designing primers for methylation PCRs. Bioinformatics 18:1427–1431.

Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25:1451–1452.

Loisel DA, Rockman MV, Wray GA, Altmann J, Alberts SC. 2006. Ancient polymorphism and functional variation in the primate MHC-DQA1 5′ cis-regulatory region. Proc Natl Acad Sci U S A. 103:16331–16336.

McCarrey JR, Thomas K. 1987. Human testis-specific PGK gene lacks introns and possesses characteristics of a processed gene. Nature 326:501–505.

McLean CY, et al. 2011. Human-specific loss of regulatory DNA and the evolution of human-specific traits. Nature 471:216–219.

Mighell AJ, Smith NR, Robinson PA, Markham AF. 2000. Vertebrate pseudogenes. FEBS Lett. 468:109–114.

Monsalve DM, et al. 2013. Human VRK2 modulates apoptosis by interaction with Bcl-xL and regulation of BAX gene expression. Cell Death Dis. 4:e513.

Moretti J, et al. 2010. The translation initiation factor 3f (eIF3f) exhibits a deubiquitinase activity regulating Notch activation. PLoS Biol. 8:e1000545.

Navarro FC, Galante PA. 2015. A genome-wide landscape of retrocopies in primate genomes. Genome Biol Evol. 7:2265–2275.

Okamura K, Nakai K. 2008. Retrotransposition as a source of new promoters. Mol Biol Evol. 25:1231–1238.

Pai AA, Bell JT, Marioni JC, Pritchard JK, Gilad Y. 2011. A genome-wide study of DNA methylation patterns and gene expression levels in multiple human and chimpanzee tissues. PLoS Genet. 7:e1001316.

Poliseno L, et al. 2010. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature 465:1033–1038.

Pollard KS, et al. 2006. An RNA gene expressed during cortical development evolved rapidly in humans. Nature 443:167–172.

Raizis AM, Schmitt F, Jost JP. 1995. A bisulfite method of 5-methylcytosine mapping that minimizes template degradation. Anal Biochem. 226:161–166.

Richardson SR, Salvador-Palomeque C, Faulkner GJ. 2014. Diversity through duplication: whole-genome sequencing reveals novel gene retrocopies in the human population. Bioessays 36:475–481.

Schrider DR, et al. 2013. Gene copy-number polymorphism caused by retrotransposition in humans. PLoS Genet. 9:e1003242.

Suzuki S, Shaw G, Kaneko-Ishino T, Ishino F, Renfree MB. 2011. The evolution of mammalian genomic imprinting was accompanied by the acquisition of novel CpG islands. Genome Biol Evol. 3:1276–1283.

Tam OH, et al. 2008. Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. Nature 453:534–538.

Vinckenbosch N, Dupanloup I, Kaessmann H. 2006. Evolutionary fate of retroposed gene copies in the human genome. Proc Natl Acad Sci U S A. 103:3220–3225.

Wang B, et al. 2013. Expression of NgBR is highly associated with estrogen receptor alpha and survivin in breast cancer. PLoS One 8:e78083.

Warner LR, et al. 2009. Functional consequences of genetic variation in primates on tyrosine hydroxylase (TH) expression in vitro. Brain Res. 1288:1–8.

Watanabe T, et al. 2008. Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. Nature 453:539–543.

Yamashita R, Suzuki Y, Sugano S, Nakai K. 2005. Genome-wide analysis reveals strong correlation between CpG islands with nearby transcription start sites of genes and their tissue specificity. Gene 350:129–136.

Yu H, et al. 2006. Origination and evolution of a human-specific transmembrane protein gene, c1orf37-dup. Hum Mol Genet. 15: 1870–1875.

Zhang Q. 2013. The role of mRNA-based duplication in the evolution of the primate genome. FEBS Lett. 587:3500–3507.

Zhang Q, et al. 2013. E3 ubiquitin ligase RNF13 involves spatial learning and assembly of the SNARE complex. Cell Mol Life Sci. 70: 153–165.

Associate editor: Michelle Meyer