RetrogeneDB—A Database of Animal Retrogenes

Michał Kabza,[1] Joanna Ciomborowska,[1] and Izabela Makałowska*[,1]
[1]Labolatory of Bioinformatics, Faculty of Biology, Adam Mickiewicz University, Poznań, Poland
***Corresponding author:** E-mail: izabel@amu.edu.pl.
Associate editor: Naruya Saitou

Abstract

Retrocopies of protein-coding genes, reverse transcribed and inserted into the genome copies of mature RNA, have commonly been categorized as pseudogenes with no biological importance. However, recent studies showed that they play important role in the genomes evolution and shaping interspecies differences. Here, we present RetrogeneDB, a database of retrocopies in 62 animal genomes. RetrogeneDB contains information about retrocopies, their genomic localization, parental genes, ORF conservation, and expression. To our best knowledge, this is the most complete retrocopies database providing information for dozens of species previously never analyzed in the context of protein-coding genes retroposition. The database is available at http://retrogenedb.amu.edu.pl.

Key words: **retroposition, gene duplication, retrogene, database.**

Brief Communication

Fast Track

Retrogenes, for a long time considered to be not important copies of parental genes are nowadays called "seeds of the evolution," because they made a significant contribution to genomes evolution (Brosius 1991). It has been shown that they play very important role in the diversification of transcriptomes and proteomes and may be responsible for the wealth of species-specific features (Betrán et al. 2002; Balasubramanian et al. 2009; Szcześniak et al. 2011). As duplicates of their parental genes, they evolve relatively fast, so these genes may acquire novel functions. Retrocopies of protein-coding genes are also known to be involved in many diseases (Prendergast 2001; Ciomborowska et al. 2013).

Analyses of retroduplications have been mostly limited to the few mammalian model species (mainly human and mouse) and fruit fly (Kaessmann et al. 2009). Nonmammalian vertebrates have been largely overlooked in retrocopies studies, and our knowledge of their evolution in other animals is even more limited. Although retrocopies are annotated in major genomic databases (Ensembl [Flicek et al. 2014], UCSC Genome Browser [Meyer et al. 2013], National Center for Biotechnology Information Gene [Maglott et al. 2011]), they are often annotated just as "pseudogenes," the same way as duplicates originated via DNA-based mechanisms. The same problem refers to more specialized database Pseudogene.org (www.pseudogene.org, last accessed January 2014). The most complete retrocopies' annotations are in Ensembl database; although they are very good for human and mouse, the quality is very poor for remaining genomes. There are only two databases fully dedicated to retrocopies: RCPedia (Navarro and Galante 2013) and HOPPSIGEN (Khelifi et al. 2005). However, the first one contains data only for a few primate species, and the latter is limited to human and mouse.

We have analyzed genomes of 62 animal species to identify retrocopies. The search was done based on the similarities between reference genomic sequence and proteins coded by multiexon genes in a given species. To increase accuracy, we applied several criteria to call a genomic region a retrocopy: Length of the alignment at least 150 bp, minimum of 50% coverage of parental gene, minimum of 50% identity, and loss of at least two introns among others (for details see supplementary file S1, Supplementary Material online). Resulting data set was additionally manually inspected to exclude potential false positives, especially copies of transposons annotated as protein-coding genes, which in some genomes totaled for as many as few thousands. Our strategy led to identification of 84,808 retrocopies, including 6,277 protein-coding genes not recognized previously as retrogenes. A total of 64,225 retrocopies identified by us are not present in the Ensembl database, this includes 139 retrocopies in the human and as many as 2,205 in the mouse genome, which belong to the best annotated. Because of our stringent requirements, applied in the order to generate a high-quality data set, the number of identified retrocopies in a given species is considerably lower than in most other databases. However, this method gave consistently good results in both, well and poorly annotated, low-coverage genomes, for example, alpaca or dolphin.

The number of retrocopies differs significantly even between closely related species, for example, 4,927 in human vs. 3,285 in chimpanzee. This may be resulting from differences in annotations and from species-specific retroposition events. In addition, retrocopies are polymorphic and higher number of retrocopies in human (vs. chimpanzee) may reflect a large amount of human population data (Abyzov et al. 2013).

Retrocopies, as a second copy of the existing gene, evolve relatively quickly and accumulate mutations. However, many of them gain functionality and become subjected to purifying selection (Vinckenbosch et al. 2006; Yu et al. 2007). We compared retrocopies with their progenitors to single out those with conserved ORF, that is, without internal stop codons or frameshifts over the entire alignment. Conserved ORFs in

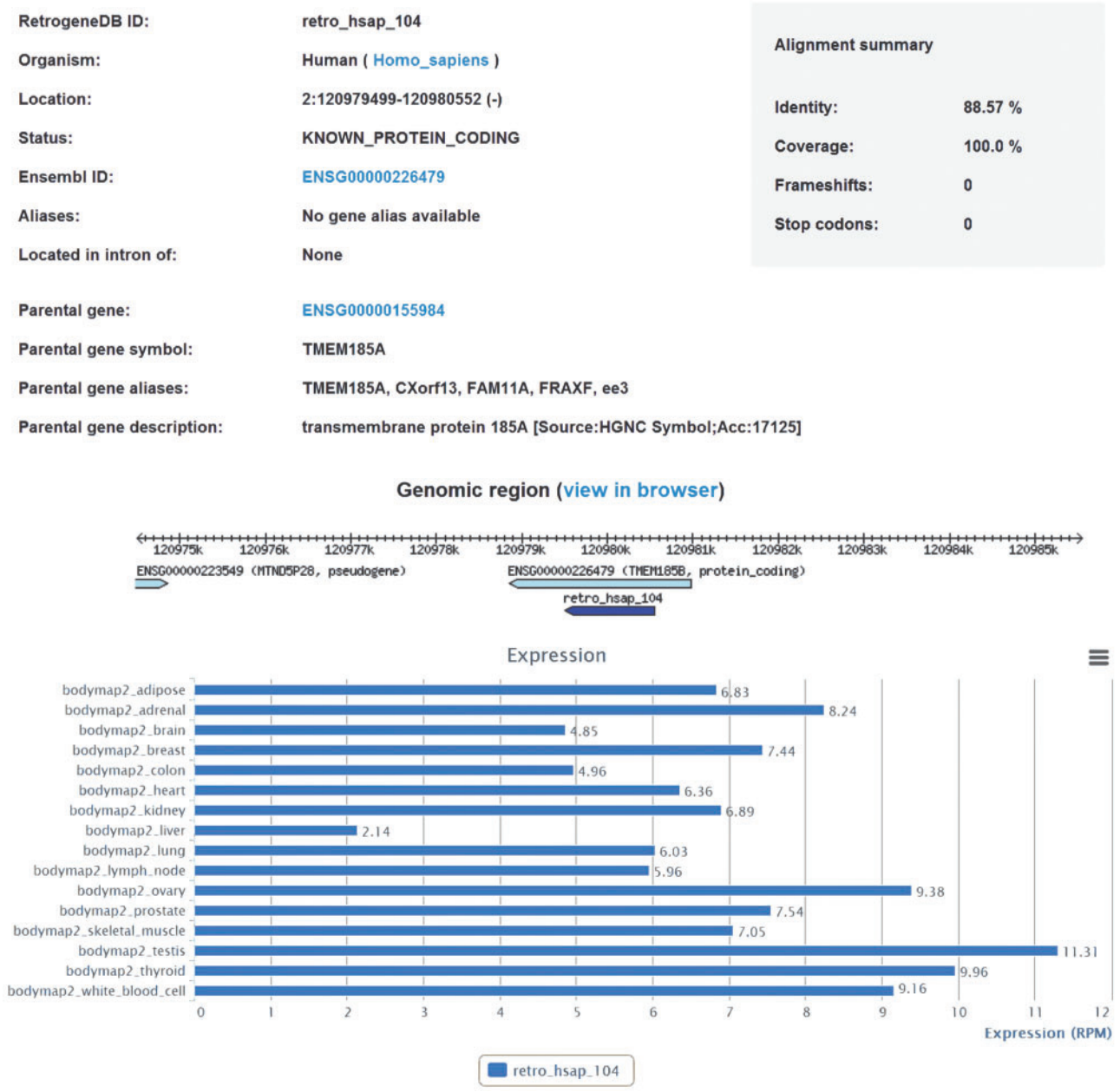

Fig. 1. Example of RetrogeneDB record with selected data.

mammals account for 10–25% of retrocopies. In nonmammalian animals, the fraction is much higher, considerably over 50% and in some species close to 100. However, the conservation of the ORF over the length of alignment does not automatically imply that a retrocopy is efficiently translated, even if it is expressed. In selected species, we also identified expressed retrocopies based on the RNA-seq data. Because of the high similarity to parental genes, in the process of reads mapping, we made sure they uniquely and perfectly map to retrocopies (supplementary file S1, Supplementary Material online). This led to the underestimation of retrocopies expression level but prevented false-positive predictions of expressed retrocopies. Approximately 10–20% of mammalian retrocopies are expressed in at least one library at minimal level of 1 RPM (reads per million mapped). In lizard, this number is higher with almost 40% of expressed retrocopies. Majority of expressed retrocopies in marsupials, egg-laying mammals, and nonmammalian species have conserved ORFs. However, in placental mammals, the fraction of expressed retrocopies with conserved ORF is lower, from only 30% in human up to 65% in horse.

All the data are stored in MySQL database (www.mysql.com, last accessed September 2013), and the web interface was developed using Django framework (www.djangoproject.com, last accessed January 2014). The database is available at http://retrogenedb.amu.edu.pl (last accessed April 26, 2014) and can be searched either from the retrocopy or the parental gene perspective. The retrocopy search can be done based on

the genomic localization, key words, parental gene name, and retrocopy ID, and results can be filtered based on the retrocopy type, ORF conservation, or expression. In addition, a JBrowse genome browser was implemented allowing retrocopy inspection in the genomic context (fig. 1). The search from parental gene perspective enables to identify all retrocopies of a given gene or all orthologs, which were retroposed in any other species. Users can also perform sequence-based search using BLAST tool.

Supplementary Material

Supplementary file S1 is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by the National Science Centre of Poland grant number 2013/09/N/NZ2/01221 to M.K. and grant number 2011/01/N/NZ2/01701 to J.C. and European Union grant PIRSES-GA-2009-247633 to I.M.

References

Abyzov A, Iskow R, Gokcumen O, Radke DW, Balasubramanian S, Pei B, Habegger L; 1000 Genomes Project Consortium, Lee C, Gerstein M. 2013. Analysis of variable retroduplications in human populations suggests coupling of retrotransposition to cell division. *Genome Res.* 23(12):2042–2052.

Balasubramanian S, Zheng D, Liu Y-J, Fang G, Frankish A, Carriero N, Robilotto R, Cayting P, Gerstein M. 2009. Comparative analysis of processed ribosomal protein pseudogenes in four mammalian genomes. *Genome Biol.* 10:R2.

Betrán E, Wang W, Jin L, Long M. 2002. Evolution of the phosphoglycerate mutase processed gene in human and chimpanzee revealing the origin of a new primate gene. *Mol Biol Evol.* 19:654–663.

Brosius J. 1991. Retroposons—seeds of evolution. *Science* 251:753.

Ciomborowska J, Rosikiewicz W, Szklarczyk D, Makałowski W, Makałowska I. 2013. "Orphan" retrogenes in the human genome. *Mol Biol Evol.* 30:384–396.

Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, et al. 2014. Ensembl 2014. *Nucleic Acids Res.* 42:D749–D755.

Kaessmann H, Vinckenbosch N, Long M. 2009. RNA-based gene duplication: mechanistic and evolutionary insights. *Nat Rev Genet.* 10: 19–31.

Khelifi A, Duret L, Mouchiroud D. 2005. HOPPSIGEN: a database of human and mouse processed pseudogenes. *Nucleic Acids Res.* 33: D59–D66.

Maglott D, Ostell J, Pruitt KD, Tatusova T. 2011. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* 39: D52–D57.

Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, Wong M, Sloan CA, Rosenbloom KR, Roe G, Rhead B, et al. 2013. The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res.* 41:D64–D69.

Navarro FC, Galante PA. 2013. RCPedia: a database of retrocopied genes. *Bioinformatics* 29:1235–1237.

Prendergast GC. 2001. Actin' up: RhoB in cancer and apoptosis. *Nat Rev Cancer.* 1:162–168.

Szcześniak MW, Ciomborowska J, Nowak W, Rogozin IB, Makałowska I. 2011. Primate and rodent specific intron gains and the origin of retrogenes with splice variants. *Mol Biol Evol.* 28:33–37.

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.

Yu Z, Morais D, Ivanga M, Harrison PM. 2007. Analysis of the role of retrotransposition in gene evolution in vertebrates. *BMC Bioinformatics* 8:308.