HOPPSIGEN: a database dedicated to comparative analysis of homologous processed pseudogenes

Khelifi Adel[1*], Mouchiroud Dominique[1]

1. Laboratoire de Biométrie et Biologie Évolutive, UMR CNRS 5558, Université Claude Bernard – Lyon 1, 43 bd. du 11 Novembre 1918, 69622 Villeurbanne Cedex, France

* To whom correspondence should be addressed. Phone:+33 472 44 80 00 (836 59); F ax: +33 478 89 27 19; Email: khelifi@biomserv.univ-lyon1.fr

Keywords: processed pseudogenes, database, web server, pseudogene identification.

ABSTRACT

Processed pseudogenes result from reverse transcribed mRNAs. Because they lack a promoter, they are nonfunctional once integrated into the genome. Subsequently, they are not subjected to the functional constraints exerting on protein-coding genes. Due to these characteristics, processed pseudogenes represent good neutral markers. In this work, we screened both mouse and human complete genomes from ENSEMBL [1] to find processed pseudogenes generated from functional genes with introns. We found 6321 processed pseudogenes in human genome and 5228 in mouse genome. So we have estimated their number to ~9400 copies in human genome (2.89 copies/Mb) and ~7400 copies in mouse genome (2.72 copies/Mb). To facilitate the use of processed pseudogenes in studying genome structure and evolution, DNA sequences from processed pseudogenes, and their functional reverse transcribed homologues are now stored in a nucleic acid database: HOPPSIGEN. It can be browsed on the PBIL (Pôle Bioinformatique Lyonnais) World-Wide Web server (http://pbil.univ-lyon1.fr/) or fully downloaded (by request) for local install. We present here few applications of the database in two fields: comparative genomic and molecular evolution.

INTRODUCTION

Processed pseudogenes arise by reverse transcription of mRNAs and integration of resulting cDNAs into the genome [2, 3]. Hence, they lack introns, possess relics of the poly-(A) tail at their 3' ends, are flanked by target-site duplications and show especially a strong similarity with the mRNAs they come from. Whereas their number was estimated from 22,000 to 75,000 [4; 5, 6, 7, 8, 9] in the human genome. But only few of them are available in nucleic databases. Moreover, available sequences of processed pseudogenes are generally poorly annotated and there is no database dedicated to that kind of sequence. In this context and considering the amount of data available, we looked for processed pseudogenes in whole human and mouse ENSEMBL genomes [1] and built the HOPPSIGEN (HOmologous Processed Pseudogenes) database. We will use these data to answer different questions, concerning particularly comparative genome analysis between mouse and human and molecular evolution of processed pseudogenes. Do these processed pseudogenes show the same insertion pattern as other close transposable elements like LINE or SINE? Which substitution pattern are these sequences subjected of? Is this pattern similar between human and mouse genomes?

MATERIAL AND METHODS.

1.Processed pseudogenes identification

Processed pseudogenes were found by searching similarity between coding sequences of functional genes having introns and whole genome sequence. We excluded intron-less genes because it is difficult to determine if processed pseudogenes generated by such genes are real processed pseudogenes (produced by reverse transcription) or duplicated genes. Moreover, some intron-less genes could be in fact processed pseudogenes wrongly annotated. This work was done using the human and mouse genome from ENSEMBL (release 8.3 and release 8.3c). For each species, we selected a dataset of non-redundant CDS (coding sequences) of genes with introns: 21913 for human and 22427 for mouse.

We made a similarity search between whole genome sequences and the dataset of genes with introns using TBLASTX [10] to find DNA sequences similar to functional CDS. To determine if a positive match was a processed pseudogenes or not, we used the lack of introns as main criteria. If a selected match was associated to more than one parent CDS, we selected the best CDS candidate as the one showing the highest similarity and the longest sequence size. All positive matches at this stage were considered as processed pseudogenes. Then using

SIM [11] to compute exact alignments, we have also identified processed pseudogenes features like polyadenylation regions or repeats.
Following this protocol, we identified 12224 human processed pseudogenes and 11372 mouse processed pseudogenes. But some of these pseudogenes are falling into a gene annotation. So after removing these cases, we kept 6321 human and 5228 mouse processed pseudogenes, associated with 2977 human CDS and 2375 mouse CDS. These sequences and all associated information were used to build the database HOPPSIGEN.
We found less processed pseudogenes than expected in both genomes. But we focused only on processed pseudogenes generated by genes with introns (approximately 85% of whole genes), and we do not have all genes. Moreover, we choose to remove all processed pseudogenes falling into a gene annotation, so according to our data, approximately 6000 ENSEMBL genes in mouse and human genome could be in fact processed pseudogenes. Indeed, we hope for the next release of the database to add rat processed pseudogenes to HOPPSIGEN and to work on better annotated genomes.

2.Database browsing

HOPPSIGEN can be accessed through PBIL (Pôle Bioinformatique Lyonnais) World-Wide Web server at (http://pbil.univ-lyon1.fr) using the WWW-query system to make requests [12]. HOPPSIGEN can be fully downloaded (by request) for local installation. Thus HOPPSIGEN database can be efficiently accessed with the Query [13] or Query_win [14] programs allowing complex queries. All information related to HOPPSIGEN installation and use can be found at http://pbil.univ-lyon1.fr/databases/hoppsigen.html.

RESULTS AND DICUSSION

1.Lineage specific processed pseudogenes.

Lineage specific repeats like SINE or LINE are more frequent in mouse genome than in human genome highlighting the fact that these two elements were more active in mouse lineage [15]. Considering the values of neutral substitutions rates proposed by Waterston and colleagues [15], we have estimated that 10% of divergence between processed pseudogenes and their functional homologous genes should correspond to sequences insertion about 45 Myr ago in human genome and 22 Myr ago in mouse genome. To be sure to select only recent processed pseudogenes, and assuming a speciation time between human and mouse of 75 Myr [15], we used this limit of 10% divergence (90% identity) to select lineage specific processed pseudogenes. Using HOPPSIGEN, we have selected human and mouse processed pseudogenes showing a similarity more than 90% with their functional homologous gene (Figure 1) and we compared the frequency of these processed pseudogenes between the two species.

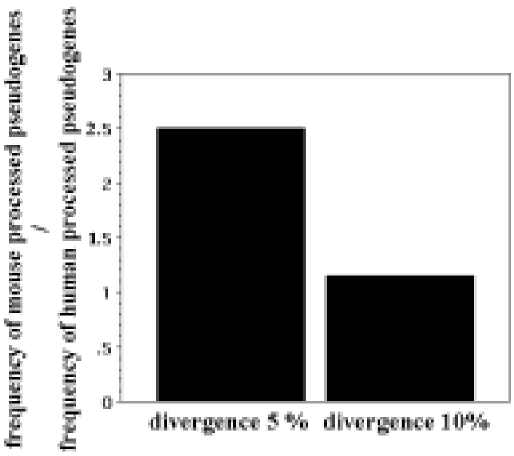

Fig. 1: Relative frequency of processed pseudogenes in human and mouse genome for a divergence of 10% and a divergence of 5%.

There are 1178 recent processed pseudogenes in mouse lineage (22.5%) compared to 1028 in human lineage (16.3%). The difference between the two frequencies is significant ($p<2.2x10^{-16}$) and it is even greater for very recent processed pseudogenes (<5% of divergence). So recent processed pseudogenes, like recent LINE and SINE elements, are more frequent in mouse lineage. May be it could have resulted of a higher reverse transcriptase activity in mouse lineage or a lower activity in human lineage, or both [15].

2.Substitution pattern.

Indeed there are few exceptions [16], most processed pseudogenes are non functional sequences for two reasons. They lack promoters so they cannot be transcribed, and they freely accumulate substitutions and indels so they cannot produce a functional protein. It has been proposed that they could be used as neutral markers for studying genome evolution [17]. We are particularly interested by the relation between substitution pattern and recombination rate. Recently, Montoya-Burgos and colleagues demonstrated that the mouse Fxy gene had increased its G+C content when translocated in a highly recombinant region [18]. Using our data we looked for the relation between substitution pattern in human processed pseudogenes and recombination pattern calculated by Kong and colleagues [19] in human genome. We needed to estimate substitution pattern of processed pseudogenes in human genome. For that purpose, we firstly built a protein alignment with CLUSTALW [20], using a functional gene, reverse transcribed extracted from HOPPSIGEN and one functional orthologous gene (from mouse), extracted from ENSEMBL 8.3. Then, the nucleotidic alignment deduced from the previous protein alignment was used as a profile in CLUSTALW to align each homologous processed pseudogene. We selected substitutions in processed pseudogenes corresponding to a conserved site between mouse and human functional gene (Figure 2). Moreover, to be parsimonious, only position 1 and 2 of a codon were retained to calculate substitutions.

```
Human Functional             ATG|CTA|AGG|CCC|ATG
Mouse Functional             AGG|CTG|ATG|CAC|ATG
Human processed pseudogene   AGG|TTA|GGT|GCC|AGC
                                 *   *  - *    *-
```

Fig. 2: Estimation of substitution pattern in processed pseudogenes. In this example, we have selected sites identical between the two functional genes and different in the human processed pseudogene. The " * " sites were selected, the " – " sites were not selected because they are in the third position of a codon. Only " * " sites were used to infer substitution pattern.

This method is very conservative, and we probably loose a great number of sites, but we kept only real substitutions. The recombination pattern is not the same for different lineage, so if we compare substitution and recombination pattern, we have to work only with recent processed pseudogenes. Moreover, very recent processed pseudogenes could be transcribed. So to get the substitution pattern, we selected processed pseudogenes showing a divergence between 1 and 6% with their functional genes (insertion date estimated to 4,5 –27 Myr). We studied 160 processed pseudogenes (1.8% of informative sites). We found that AT->GC and GC->AT substitution patterns in processed pseudogenes are strongly dependant to recombination rate (Chi2, p=0.001871) (table 1).

Tab 1.: Relation between the substitution pattern and the recombination rate in processed pseudogenes. The three categories of recombination rates were defined by dividing the data into three equal parts.

	Low recombination rate		Intermediate recombination rate		High recombination rate	
substitutions	number	percentage	number	percentage	number	percentage
AT -> GC	178	37.2	232	44.8	198	48.6
GC -> AT	301	62.8	286	55.2	209	51.4
total	479	100	518	100	407	100

So according to our data, the expected GC content at equilibrium increases with recombination rate in human processed pseudogenes. Using a similar method, we found the same relation for mouse processed pseudogenes (data not shown). It could be explained by selection in favor of GC substitutions in GC rich regions or by biased gene conversion (BGC). The latter is a neutral mechanism occurring during DNA-DNA mismatches. It causes DNA to be repaired preferentially toward GC, so according to this mechanism recombination increasing mismatches should also increase AT->GC mutations [21, 22]. But like selection, the impact of BGC on the fixation process depends on the population size. Moreover, it is difficult to distinguish between the selective and the BGC hypotheses because the predictions for the two models are the same. Using polymorphism data should help us to test selective hypotheses. In this case, we should find a link between polymorphism and recombination.

REFERENCES

[1] Clamp,M., Andrews,D., Barker,D., Bevan,P., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T., Durbin,R., Eyras,E., Gilbert,J., Hammond,M, Hubbard,T., Kasprzyk,A., Keefe,D., Lehvaslaiho,H., Iyer,V., Melsopp,C., Mongin,E., Pettett,R., 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., Birney,E. (2003) Ensembl 2002: accommodating comparative genomics. *Nucleic. Acids Res.,* **31**, 38-42.

[2] Vanin,E.F. (1985) Processed pseudogenes: characteristics and evolution. *Annu. Rev. Genet*., **19**, 253-272.

[3] Esnault,C., Maestre,J. and Heidmann,T. (2000) Human LINE retrotransposons generate processed pseudogenes. *Nat. Genet*., **24**, 363-367.

[4] Crollius,H.R., Jaillon,O., Dasilva,C., Ozouf-Costaz,C., Fizames,C., Fischer,C., Bouneau,L., Billault,A., Quetier,F., Saurin,W., Bernot,A., and Weissenbach,J. (2000) Characterization and repeat analysis of the compact genome of the freshwater pufferfish Tetraodon nigroviridis. *Genome Res.*, **10**, 939-949.

[5] Ewing,B. and Green,P. (2000) Analysis of expressed sequence tags indicates 35,000 human genes. *Nat. Genet.,* **25**, 232-234.

[6] Gonçalves,I., Duret,L. and Mouchiroud,D. (2000) Nature and structure of human genes that generate retropseudogenes. *Genome Res*., **10**, 672-678.

[7] International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature,* **409**, 860-921.

[8] Harrison,P.M., Hegyi,H., Balasubramanian,S., Luscombe,N.M., Bertone,P., Echols,N., Johnson,T., Gerstein,M.. (2002) Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. *Genome Res.*, **12**, 272-280.

[9] Chen,C., Gentles,A.J, Jurka,J., Karlin,S. (2002) Genes, pseudogenes, and Alu sequence organization across human chromosomes 21 and 22. *Proc Natl. Acad. Sci. U S A.* **99** , 2930-

[10] Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic. Acids Res*., **25**, 3389-3402.

[11] Huang, X. and Miller, W. (1991) A Time-Efficient, Linear-Space Local Similarity Algorithm. *Advances in Applied Mathematics*., **12**, 337-357.

[12] Perriere,G. and Gouy,M. (1996) WWW-query: an on-line retrieval system for biological sequence banks. *Biochimie*, **78**, 364-390.

[13] Gouy,M., Gautier,C., Attimonelli,M., Lanave,C. and Di Paolo,G. (1985) ACNUC-- a portable retrievial system for nucleic acid sequence databases: logical and physical designs and usage. *Comput. Appl. Biosci.,* **1**, 167-172.

[14] Perriere,G., Gouy,M. and Gojobori,T. (1994) NRSub: a non-redundant data base for the Bacillus subtilis genome. *Nucleic. Acids Res.,* **22**, 5525-5559.

[15] Mouse Genome Sequencing Consortium (2003) Initial sequencing and comparative analysis of the mouse genome. *Nature,* **420**, 520-562.

[16] Lee, J.T. (2003) Complicity of gene and pseudonyme. *Nature,* **423**, 26-28.

[17] Francino, M.P., and Ochman, H. (1999) Isochores result from mutation, not selection. *Nature,* **400**, 30-31.

[18] Montoya-Burgos, J.I., Boursot, P., Galtier, N. (2003) Recombination explains isochores in mammalian genomes. *Trends in genetics,* **19**, 128-130.

[19] Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., Shlien, A., Palsson, S.T., Frigge, M.L., Thorgeirsson, T.E., Gulcher, J.R., Stefansson, K. (2002) A high-resolution recombination map of the human genome. *Nat Genet.,* **31**, 241-247.

[20] Thompson, J.D., Higgins, D.G., and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic. Acids Res.,* **22**, 4673-4680.

[21] Eyre-Walker, A. (1993) Recombination and mammalian genome evolution. *Proc R Soc Lond B Biol Sci.,***252 (1335)**, 237-243.

[22] Galtier, N., Piganeau, G., Mouchiroud, D., Duret, L. (2001) GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics*., **159**, 907-911.