Is human the only lucky species that is favored by new retrogenes?

Deng Pan[1] and Liqing Zhang[1,2*]

[1]Department of Computer Science, Virginia Tech,
2050 Torgerson Hall, Blacksburg, VA 24061-0106, USA
[2]Program in Genetics, Bioinformatics, and Computational Biology
[*]To whom correspondence should be addressed; E-mail: lqzhang@vt.edu

Abstract

A previous study finds a burst of young retrogenes in human and thus emphasizes the importance of retrogenes in human origin. However, available evidence suggests no dramatic difference in retroposition dynamics among most mammals, casting doubts on the claim of the human uniqueness of the observed patterns. Here, we examined the retrogene distributions in 8 mammalian genomes using 4 non-mammalian genomes as a contrast. The unanimous distributional patterns of K_s divergence between retrogenes and their parents in all the studied mammals indicate that all these mammals had a recent burst of young retrogenes, which did not occur in all the studied non-mammals. We also examined the retrofamilies (the gene families that contain retrogenes) in these species and found that most of the retrofamilies are shared by more than one species. Phylogenetic tree analyses show that about 20% of these shared retrofamilies and their retrogenes emerged independently from multiple mammalian lineages. All these observations indicate that the fast recruitment of retrogenes is not a unique phenomenon to human but a shared one by all the surveyed mammals. Furthermore, the sharp contrast in retrogene dynamics between mammals and non-mammals suggests that the recruitment of the specific L1 retrotransposons in mammals might have been an important evolutionary event for the split of mammals and

non-mammals, and retroposition has played and is continuing to play an important and active role in shaping the mammalian genomes, as compared to being rather inert in non-mammals.

Retroposition has been considered to be one of the major mechanisms of gene duplication (**??**). Retroposed gene copies (retrocopies) lack many of their parental genes' genetic features, such as introns and regulatory elements. Most of the retrocopies become pseudogenes (also known as processed pseudogenes) (**????**) and some of them may happen to recruit upstream regulatory elements and become a true gene (aka. retrogene, (**?**)). As the survival rate of retrocopies is low, retrocopies have long been viewed as evolutionary dead ends with little functional significance (**?**). Lately, however, a significant number of retrogenes have been identified in the genomes of mammals and insects (**?????**).

? found a burst of retroposition in human that gave rise to many young retrogenes and thus claimed that retrogenes significantly contribute to the formation of new human genes. The importance of retrogenes in human seems to suggest yet another exciting viewpoint of human origin. However, our recent study (**?**) shows that retroposition seems to have generated as many duplicated gene copies in mouse as in human. This led us to think that the burst of retrogenes observed in **?** might be a common phenomenon in mammals, rather than a unique one in humans.

To address the issue, we analyzed the retrogenes in 8 mammalian genomes, including human (*Homo sapiens*), chimp (*Pan troglodytes*), macaca(*Macaca mulatta*), mouse (*Mus musculus*), rat(*Rattus norvegicus*), dog(*Canis familiaris*), cow(*Bos taurus*), and opossum(*Monodelphis domestica*); and 4 non-mammalian outgroup species, including chicken(*Gallus gallus*), zebra fish (*Danio rerio*), fruitfly (*Drosophila melanogaster*), and anopheles(*Anopheles gambiae*) (Figure 1). Similar to previous studies (**?????**), we defined a retro-

gene as an "intact" retrocopy (i.e. no frameshift mutations and no premature stop codons) that has evidence of gene expression. But since not all the studied species have enough expression evidence, we also used a computational criterion of $K_a/K_s \leq 0.5$ (**????**) to gather retrogenes (see supplements for the detailed procedures of collecting retrogenes).

Summary statistics of retrogenes are shown in Table 1. We denote the gene family that has at least one retrogene as *retrofamily*. Table 1 shows that the number of retrogenes and the number of retrofamilies are approximately equal in all the studied species. This approximate one retrogene per retrofamily is partially due to the stringent standards that we used to obtain the data. However, even without the restrictions, such as $K_a/K_s \leq 0.5$ and different chromosomal locations between parental genes and retrogenes, the ratios in almost all species are still significantly less than 2.

We used the K_s distributions of the parental-retrogene pairs as a proxy of the time distribution of retrogene formation events. Figure 2 shows that in the mammals, a high proportion of retrogenes exist within the small K_s regions and at least $\sim 10\%$ of the parental-retrogene pairs have $K_s \leq 0.1$. The K_a distributions also show a similar pattern (results not shown here). In contrast, all the non-mammals seem to have much less "young" retrogenes. This observation suggests that bursts of young retrogenes occurred in all the studied mammals (and not just in humans), but not in non-mammals.

We examined the distribution of gene families that have retrogenes among the studied mammals and non-mammals. We define the retrofamilies that are present in only one lineage as lineage specific retrofamilies (LSRs). Thus, the non-LSR retrofamilies are shared by at least two lineages. Clearly, the

number of LSRs in a certain lineage is mostly affected by its closest related lineage being compared. The higher the divergence between two species, the more LSRs we expect to see in each of the lineages. We mapped the percentages of LSRs onto the species tree (Figure 1). The percentage of LSRs in a particular lineage is calculated as the number of LSRs in the lineage divided by the total number of retrofamilies that the lineage has. It shows that the percentages of LSRs on the external branches of all species except insects are less than 50%, which even holds for some multiple-species lineages, such as the primate lineage (Branch A, 44.5%), the murine lineage (Branch B, 35.2%), and the lineage including cattle and dog (Branch C, 27.4%). Thus, the observation shows that most of the retrofamilies in mammals are shared retrofamilies.

The high extent of retrofamilies shared by multiple mammals suggests that the retroposition events either are ancestral, or alternatively, occurred independently in multiple species. These two possibilities can be differentiated by the shapes of phylogenetic trees. We therefore constructed phylogenetic trees of parental genes and retrogenes in all shared retrofamilies of mammals. Altogether, we found 297 retrofamilies that are shared by at least 2 mammalian species and built 296 trees (one tree is unable to build due to high sequence divergence), among which 57 trees follow strictly the pattern that the retroposition occurred independently. Since human and chimp are closely related, we also considered the two species together as the great ape taxon and obtained 7 additional independently-occurred trees. Thus, about 22% retrogene formation events occurred in multiple mammalian species are independent.

All the above observations suggest that the retrogene burst occurred in all mammals, and therefore is not a unique phenomenon in primates or human. The observations are unlikely to be due to gene conversion because the parental genes and retrogenes in our study are located on different chromosomes and gene conversion has been shown to be rare between genes on different chromosomes in mammals (**?**). Another concern is the possible inclusion of pseudogenes in the data. This problem can be discussed from three aspects. First, we use all the available "Known" genes and require protein evidence to minimize the influence of pseudogenes. Second, under the most stringent criteria, we estimated that at most 30% to 60% of the retrogenes in the studied species (the exact proportions differ among mammals) with $K_a/K_s \leq 0.5$ might be non-functional (see supplement files for details) and removing these proportion of genes in the small Ks regions does not change the overall pattern. We also estimated that the probability that the independently-occurred retrofamilies contain "functional" retrogenes is as high as $0.66 - 0.83$. Third, we performed the same analyses with all Ensembl versions of 36 to 46 and found that although the collection of retrogenes changes slightly among versions, the overall K_s distribution pattern and other observations do not change. Fourth, to prove a gene (including retrogene) that has completely no function is far more different than to prove it to have some functions. The definition of a gene (including retrogene) is still under hot debates (**?**). Contrary to classical definition, up to 20% of pseudogenes (including broken retrocopies) are expressed and maybe have functions (**?**). Since we obtained the data using the same pipeline for all species, and we used the same definition as previous related studies, and our

collection of human retrogenes under this pipeline highly overlaps with that of **?** (except for some genes due to the annotation changes between Ensembl versions), considering the highly unanimous patterns we obtained, no matter how to define a gene (or pseudogene), it is at least safe to say that human is not unique at this point, and thus we still can not support human's special status as **?** suggested.

Therefore, our observation suggests that retroposition has played a unique role not just in the evolution of humans, but in all the studied mammals. In fact, it may be one of the hallmarks of the evolutionary divergence of mammals from non-mammals. Retroposition is believed to be driven by the enzymatic machinery of LINE1 (Long Interspersed Nucleotide Element 1, L1) or similar mechanisms (**?**). It is reported that mammals maintained a small number of active L1 lineages (**??**). Why only a small number of L1 lineages are actively maintained in mammals without being lost (**?**) is still an open question. At the same time, L1s also seem to be the most persistent type of transposable elements in mammalian genomes (**?**). Based on our observation, We hypothesize that it may be because that L1s, acting as an engine of producing retrogenes, are closely related to functional evolution in mammals, so that only a small range of L1 lineages with the highest efficiency of generating retrocopies were selectively maintained during the evolutionary divergence of mammals from non-mammals. It is noted that some group of South American Rodents (**?**) and megabats (**?**) seem to have lost L1s. But since the estimated extinction time of L1s in these species is very recent (about 5 MYA in rodents and 24 MYA in megabat), this suggests that L1s was important for the divergence of mammals and non-mammlas, but no

longer essential for the recent evolution of some mammals.

Acknowledgments

The authors thank Mark Lawson for reading the manuscript. This work was supported by NSF Grant IIS-0710945 and a VPI&SU ASPIRES grant.

Tables

Species	# of retrogenes	# of retrofamilies	# of retrogenes per family
Human	163	150	1.09
Chimp	199	187	1.07
Macaca	275	240	1.15
Mouse	154	144	1.07
Rat	226	202	1.12
Dog	95	90	1.06
Cow	163	148	1.10
Opossum	232	220	1.05
Chicken	99	89	1.11
Zebrafish	165	122	1.15
FruitFly	212	188	1.13
Anopheles	108	101	1.07

Table 1: Statistics of retrogenes and retrofamilies.

Figure legends

Figure 1

The species tree is adapted from **?**. The percentage of LSRs in a particular lineage (shown on each branch) is the ratio of the number of LSRs in the lineage to the total number of retrofamilies that the lineage has. Branch A is the primate lineage; Branch B is the murine lineage; Branch C contains dog and cattle.

Figure 2

Distributions of K_s distances between parental and retrogenes of all species.

Figures

Figure 1: Species tree with retrogene statistics.

Figure 2: Distributions of K_s between members of parental-retrogene pairs