UNIVERSITY OF CALIFORNIA

SANTA CRUZ

IMPROVING GENE ANNOTATION AND UNDERSTANDING OF MAMMALIAN EVOLUTION BY IDENTIFYING RETROCOPIES OF MRNA TRANSCRIPTS

A dissertation submitted in partial satisfaction
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

BIOINFORMATICS

By

Robert D. Baertsch

June 2010

The Dissertation of Robert D. Baertsch
Is approved:

_____
Professor David Haussler, Chair

_____
W. James Kent

_____
Professor John Tamkun

Table of Contents

Abstract

Improving gene annotation and understanding of mammalian evolution
by identifying of retrocopies of mRNA transcripts

by

Robert Baertsch

Understanding gene creation is essential to the study of human evolution. Duplication followed by specialization has been suggested as an important mechanism for evolving new functions in multi-cellular organisms. In this work, I present a new method for recognizing non-viral retrotransposition of mRNAs as well as analysis of how this mechanism has influenced mammalian and specifically primate evolution.

My work has centered on two areas: 1) developing an automated method to annotate retroposed pseudogenes and retrogenes; 2) analyzing the resulting sets of data in human and mouse to increase our understanding of this mechanism in two mammalian lineages. This work has resulted in a number of papers in collaborations with external groups in the areas of: improved gene annotation via screening pseudogenes, pseudogene annotation and retrogene annotation.

In the analysis of the dataset, I found a set of human and mouse retrogenes that arose after their divergence. Surpisingly, these were not just simple duplication events rather a extremely diverse set of cases where “anything goes”. I classify these events into broad groupings: 1) duplication events; 2) exon shuffling events, and 3) novel genes.

In the set of putative human retrogenes that showed evidence of expression, eight were determined to by specific to humans. We performed a series experiments to determine if they were fixed in the population and also to verify RNA expression in selected cases.

Finally, I show that the previously reported burst in retrotransposition in mammals, which was suppressed in primates, *has not decreased* the rate of gene duplication via retrotransposition in primates. Perhaps this evolutionary mechanism can help explain the rapid change in pheotypes during primate evolution.

Dedication

This is dedicated to my wife Elizabeth who is an indomitable force for good in my life.

Acknowledgements

I would like to thank my committee for helping me achieve this degree. Special thanks to my advisor David Haussler for steering me in the right direction, despite my repeated attempts to go off in many different directions. His unique style combining a spark of creativity and deep knowledge has allowed his entire team to blossom in many unexpected ways. I would like to thank Jeurgen Brosius for invaluable help and insight in getting my first manuscript created. I would like to thank Jim Kent for creating the UCSC Genome Browser without which my research (and thousands of others) would not be possible. I would like to thank Sofie Salama and Bryan King for help in the wet lab. Marsha Bundman for helpful comments on the manuscript (chapter 2); the UCSC Browser team for cross species alignments; Martin Kiefmann for (re)sequencing some of the primate loci and Chuck Sugnet, Jerzy Jurka and Tom Pringle for helpful discussions. I thank Rachel Harte for her help in editing my second manuscript and hours of time testing and running my retroFinder pipeline to verify that I was not the only one who could run it. Finally, I would like to thank Mark Diekhans for getting my started on bioinformatics and his constant encouragement throughout my time at UCSC without whose inspiration I would not have finished.

1. Background

Introduction to Transposable Elements

The study of evolution consists of tracing a series of DNA rearrangements and mutations in genomes over time. Recombination of DNA is one of the major mechanisms that result in divergence between individuals in a population that leads to speciation. General recombination results in rearrangements between similar DNA sequences. This results in large-scale changes in structure but leaves gene order and intervening sequence intact. These blocks of contiguous DNA between species are called orthologous or syntenic regions (Flores-Rozias & Kolodner 2000). Transposable elements are derived from a different type of recombination, known as site-specific recombination. This type of recombination results in rearrangements from mobile genetic elements between non-homologous regions within the genome. Site-specific recombination events from transposable elements are typically smaller than general recombination events and can be replicated on different chromosomes in many places throughout the genome. In some cases, this has resulted in a greatly expanded size of the genome relative to species that do not have transposable elements.

The importance of transposable elements was first recognized by Barbara McClintock. She described the linkage between the presence of certain pigments expressed in maize kernels and the change in copy number of specific “mutable

elements" (McClintock 1950). This early work demonstrated the importance of transposable elements in gene regulation, which we are only now beginning to understand.

There are two distinct mechanisms, using different enzymes and target sites, for creating mobile genetic elements: 1) transpositional site-specific recombination which does not require specific sequences at the target site and 2) conservative site-specific recombination which involves formation of a short heteroduplex joint which requires a short matching sequence between the donor and target DNA that is not required in the transpositional type.

Transposable elements

Transpositional site-specific recombination events are also called transposable elements or transposons. There are three major classes of transposons that share the property of being nearly random in their insertion point back into the genome. The first major class is DNA-only transposons, where the DNA is excised and rejoined by a transposase enzyme at the target site (Wicker et al., 2007). The signature of this class is short inverted repeats at each end of the target site that are used to form the DNA loop used in the cut and paste operation. The other two classes use an RNA intermediate in the copy process leaving the donor DNA intact. The second class, called retroviral-like retrotransposon, uses RNA polymerase to transcribe the DNA

sequence into RNA. A double stranded break is formed at the target site by an integrase (reverse transcriptase), followed by a reinsertion of the DNA, leaving short gaps. The DNA repair mechanism fills in these gaps leaving short direct repeats at the ends of the inserted DNA. The third class, known as a non-retroviral transposons, also uses an RNA intermediate, endonuclease and a reverse transcriptase to reinsert the DNA copy at the target site. However this class forms a single stranded break in the DNA at the target site and also reinserts the polyA tail present at the 3' end of the RNA intermediate, creating a unique signaure. The DNA repair mechanism fills in the second strand and the short gaps at each end of the insertion point, known as target-site duplications. This last class of transposon makes up 42% of the human genome (Lander et al., 2001) and forms the largest portion of duplicated DNA in the genome.

Types of non-retroviral retrotransposons

Non-retroviral retrotransposons can be further classified into autonomous elements called LINEs (long interspersed elements) and non-autonomous SINEs (short interspersed elements) with about 850,000 (21% of the human genome) and 1.5 million copies (13% of the human genome) respectively in the human genome (Lander et al., 2001). SINE elements do not have reverse transcriptase or endonuclease enzymes so they use the LINE machinery for propagation around the genome.

Processed pseudogenes and retrogene

Processed psuedogenes, also a type of non-retroviral retrotransposon, are formed when a mature mRNA, used as a template for protein synthesis, is picked up by the LINE machinery and reinserted into the genome. These gene-like transcripts are devoid of introns since they are usually inserted after splicing has occurred (Vanin 1985). Another distinguishing feature is the presence of the polyA tail at the 3' end of the pseudogene (Vanin 1985). Various groups have estimated the number of processed pseudogenes to be between 8,000 and 20,000 copies in the human genome (Zhang et al 2003, Sakai et al 2007, Khelifi et al 2005). Most of the pseudogenes are "dead on arrival" after insertion because they need to acquire a promotor and regulatory control from their new context in order to be transcribed (Zheng and Gerstein 2007). Processed pseudogenes that aquire a promotor and become functional are known as retrogenes. Young genes are difficult to distinguish from pseudogenes since they have not had time to accumulate sufficient mutations to establish evidence of selection. Transcript evidence alone is not sufficient to establish function. Gene knockouts or other experimental techniques are required to verify function; this can be difficult in primates. Several groups have made estimates of retrogenes ranging from several hundred (Harrison et al., 2005, Vinckenbosch 2006) up to 726 (Baertsch et al., 2008). Out of the 12,000 human processed pseudogenes that we studied, 726 show evidence of expression and are candidate functional genes called "retrogenes" (Baertsch et al., 2008). Not all of these cases are

due to simple duplication events; occasionally retrocopies were incorporated into existing transcripts or alternative splice variants of existing transcripts. Retroposed mRNAs provide a rich source of protein-coding sequence that is sometimes used by evolution in unexpected ways. We found that the UTR can be used a source for novel coding sequence and even a few cases of anti-sense transcript showing evidence of expression (Baertsch et al., 2008).

Dating of retrotransposition events

In order to determine the date of an insertion event, orthology was used to determine the approximate date. Orthology was chosen because molecular methods can be difficult to use when looking at insertions of a few hundred bases.

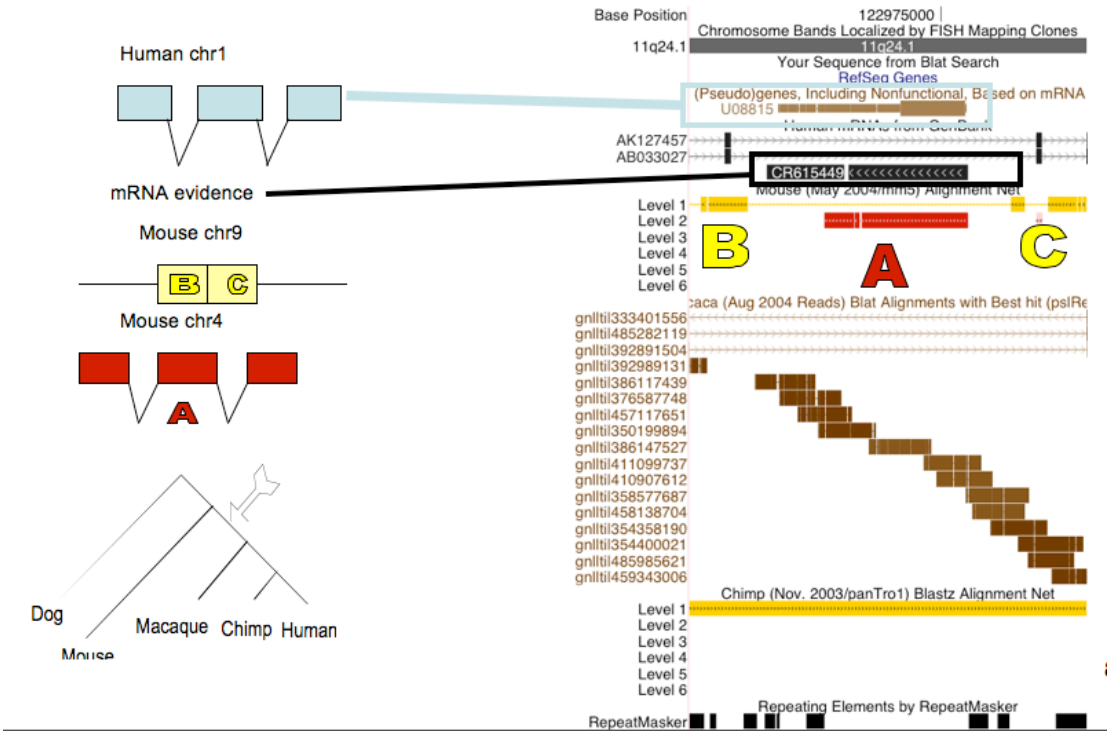


Figure 1 – Processed pseudogene derived from parent gene SF3A3 (light blue – human, red - mouse). Tree on left shows show species tree of mammals used to date insertion event. Breaks in the orthlogous nets (from UCSC genome browser) were used to identify appoximate time of insersion (show with arrow). The pseudogene is not present in mouse since the flanking DNA (yellow) is contiguous in mouse. Macaque whole genome shotgun reads that were aligned to the human genome (brown) show orthologous DNA and therefore presence of the processed pseudogene in macaque.

Overview of thesis

In Chapter 2, I present a published paper on the method and results of detailed analysis of retrocopies in the human genome. Chapter 3 presents a second publication with a refined method that has been turned into a software package that is now used the UCSC Genome Browser staff to generate retroposed gene annotations. It also presents results of retrocopy annotation on the mouse genome compared with human. It also compares my method to other methods for detecting of pseudogenes along with the strengths and weakness of different methods. I describe class of cases that others do not detect. Chapter 4 describes the results of wet lab experiments that I performed in search of a novel human specific retrogene. Appendix 1 is a publication resulting from a collaboration with a group doing gene prediction. As a result, the gene finder, Augustus, increased in specificity after adding retrocopies as hints. Appendix 2 presents results of analyzing transcribed pseudogenes as part of the The Encycopedia of DNA Elements (ENCODE) project. Yale, Sanger and UCSC compared our pseudogene data sets and determined a set of common predictions that were tested for expression using Rapid Amplification of cDNA Ends (RACE). Appendix 3 covers work that done in conjunction with the CCDS group which is

a set that is common to annotation from two collaborating centers: RefSeq (NCBI) and HAVANA/Ensembl (WTSI/EBI).

2. Retrocopy contributions to the evolution of the human genome

Published in *BMC Genomics 2008, 9:466.*

Robert Baertsch[1] , Mark Diekhans[1] , W James Kent[1] , David Haussler[1] and Jürgen Brosius[2]

1. Center for Biomolecular Science and Engineering, University of California Santa Cruz, Santa Cruz, California 95064, USA

2. Institute of Experimental Pathology, ZMBE, University of Münster, Von-Esmarch-Str. 56, D-48149, Münster, Germany

Abstract

Background

Evolution via point mutations is a relatively slow process and is unlikely to completely explain the differences between primates and other mammals. By contrast, 45% of the human genome is composed of retroposed elements, many of which were inserted in the primate lineage. A subset of retroposed mRNAs (retrocopies) shows strong evidence of expression in primates, often yielding functional retrogenes.

Results

To identify and analyze the relatively recently evolved retrogenes, we carried out BLASTZ alignments of all human mRNAs against the human genome and scored a set of features indicative of retroposition. Of over 12,000 putative retrocopy-derived genes that arose mainly in the primate lineage, 726 with strong evidence of transcript expression were examined in detail. These mRNA retroposition events fall into three categories: I) 34 retrocopies and antisense retrocopies that added potential protein coding space and UTRs to existing genes; II) 682 complete retrocopy duplications inserted into new loci; and III) an unexpected set of 13 retrocopies that contributed out-of-frame, or antisense sequences in combination with other types of transposed elements (SINEs, LINEs, LTRs), even unannotated sequence to form potentially novel genes with no homologs outside primates. In addition to their presence in human, several of the gene candidates also had potentially viable ORFs in chimpanzee, orangutan, and rhesus macaque, underscoring their potential of function.

Conclusion

mRNA-derived retrocopies provide raw material for the evolution of genes in a wide variety of ways, duplicating and amending the protein coding region of existing genes as well as generating the potential for new protein coding space, or non-protein coding RNAs, by unexpected contributions out of frame, in reverse orientation, or from previously non-protein coding sequence.

Background

While it is said that the human and chimpanzee genomes share anywhere from 95 to 98.5% similarity in their DNA sequences, base exchanges and small indels alone are unlikely to completely explain the differences among diverging primates and between other mammals. Comparative genomics identifies functional elements by searching for conserved DNA across species and is an excellent method for identifying highly conserved biological functions (Berjerano et al. 2006, Bejerano et al. 2004, Lowe et al. 2007, Nishihara et al. 2006). However, sequence conservation is not sufficient to identify *newly* evolved functions. Although point mutations in combination with selection can explain changes in transcriptomes and proteomes, the high number of retroposition events along the primate lineage must also be considered to understand the phenotypic difference between primates and other mammals. Our aim is to understand how one class of these, retroposed mRNA (which we call retrocopies) influenced gene evolution in primates.

Primates experienced a large increase in retroposed insertions, and at least forty-five percent of the human genome is composed of retroposed elements (Lander et al, 2001). The majority of random sequences in the human genome are also retroposition-derived, but are too old to be recognized as such (Brosius 1999). The subclass of retroposed copies of spliced mRNAs are referred to as retrocopies, irrespective of their potential functionality (Vinckenbosch et al. 2006). Retrogenes are then those retrocopies that do not decay in the genome but have been exapted

into a variant or novel function. Discernible retrocopies contribute only about 1% of the human genome, yet they have a much more varied protein coding space than other more numerous types of retroposed elements, including the long interspersed elements (LINEs, 21%) and short interspersed elements (SINEs, 13%). The latter lack protein coding regions altogether. Yet, part of Alu and MIR SINEs contributed to newly evolved exons (Zhang et al. 2006, Sorek et al. 2002, Krull et al. 2005, Krull et al. 2007, Sela et al. 2007).

Authors differ widely in their assessment of the number of functional retrogenes (Gotea and Makalowski 2006, Zheng and Gerstein 2007, Suyama et al. 2006, Sakai et al. 2007, Gray et al 2006, Balakirev and Ayala 2003). Following the first reports of functional retrogenes in the late eighties (McCarrey and Thomas 1987, Dahl et al 1990, Boer et al 1987), it was thought that these were isolated cases, as mRNA retrocopies lack their own promoter elements. Despite this apparent handicap, research in the last 20 years has shown that gene duplication via retroposition – as opposed to segmental duplication of entire genes including their regulatory elements- is able to equip retrogenes with regulatory elements that are different from those in the parent gene and often leads to different expression patterns (Long et al 2003, Brosius 1991, Brosius and Gould 1992). Formation of new genes via retroposition followed by positive selection and/or adaptive evolution during modification of genes via exon acquisition has been increasingly reported (Vinckenbosch et al 2006, Long et al 2003, Sorek et al 2004, Wang et al 2006). In addition to simple

duplication events, gene fusions involving retrocopies that contribute protein domains with sequence similarities to existing genes were also reported both in animals (Vinckenbosch et al 2006 , Wang et al 2002, Long and Langley 1993) and in plants (Wang et al 2006). Transcribed retrogenes or retrocopies also have the potential to become functional as non-protein coding RNAs (reviewed by Zheng and Gerstein 2007, Sasidharan and Gerstein 2008).

The number of new genes that have arisen since mouse and human split from their common ancestor is small compared with the total number of human genes (Waterston et al 2002) and apparently, most proteins arose by duplication and divergence of existing ones (Chothia et al. 2003). Yet, it would be surprising if evolution would not have grasped opportunities of genes (or parts therof) "out of the blue", i.e., from loci that previously were intergenic or intronic. Early studies suggested "overprinting" of a protein coding region in a different reading frame as a means of generating new protein sequence space (Grasse 1997, Ohno 1984, Keese and Gibbs 1992). This concept has been revisited one and a half decades later in alternatively spliced genes (Liang and Landweber 2006). The idea of recruiting novel protein sequence space out of random intronic sequences dates back to Wally Gilbert's suggestion three decades ago (Gilbert 1978) with experimental evidence initially accumulating slowly (Krull et al 2007, Alekseyenko at al 2007, Buttice et al 1990, Golding et al 1994). Interestingly, it has been shown experimentally that an arbitrary sequence can evolve towards acquiring a biological function (Hayashi et al

2003). Young genes or young parts of more ancient genes are a unique set to examine because we can see both the putative successes and apparent failures of natural selection, before the latter are erased by mutations. Although our study focuses on these more recent events, we also find evidence for more ancient events. This, along with the fact that 12–15% of mammalian genes are intronless (Sakharkar et al 2005), one of the hallmarks of retrocopy retroposition, suggests that the process of retroposition and the idiosyncratic variations of potential novel protein sequence acquisition have been important for billions of years in generating novel protein-sequence space (Gilbert 1978, Patthy 1991, Dorit et al 1990). Our results document that there are many mechanisms beyond segmental duplication and point mutations by which genomes generate new genes or variants of existing ones. Retrocopies provide the means for modifying splice patterns of genes (Tress et al 2007), potentially adding entirely new protein coding sequences, and contributing non-protein coding RNA or regulatory sequences, thereby expanding the possibilities to shape gene evolution. From here on, when we mention retrocopy-derived protein sequences, ORFs, and/or exons, we assert that they are potential, hypothetical or theoretical only. The main rationale of this publication is to delineate the multitude of possible ways in that mRNA retrocopies, once exapted, can contribute to novel protein sequences over evolutionary time.

Results

Retrocopies with strong evidence of expression

To determine how many retrocopies are potentially functional, we used BLASTZ to align all human mRNAs to the human genome, which resulted in several hundred thousand alignments. These matches were then scored for a set of features, including the number of processed introns; the absence of conserved splice sites; breaks in orthology with mouse, dog, and rhesus monkey; the presence, position, and length of the poly(A) tail; and sequence similarity and fraction of the parent mRNA that is represented in the retrogene (see Methods for full description), indicating evidence of the likelihood of recent retroposition. From this we obtained a set of 12,801 candidates that are likely retroposed copies of intron-containing parent genes. In order to set our score threshold, we compared our set to the manually curated Vega of processed pseudogenes (retrocopies of mRNAs that may or may not be functional). When we found disagreements between the sets, we either improved our feature set or discovered problems with the Vega annotation. This resulted in improvements of both. In order to determine if the retrocopies are expressed, we looked for overlap with mRNA or EST evidence. Following filtering to eliminate 6413 cases without mRNA or EST evidence, we found that 6,287 retrocopies showed evidence of expression by at least one EST or one mRNA [See Additional File 1]. For our analysis we used more stringent requirements for expression (see below) than previous work. We chose not to use Ka/Ks analysis to look for evidence of natural selection, mostly due to the short length of retrogene sequences. Many of

those could be functional parts of genes; however, for more recent events, final proof might be difficult.

Additional file 1. Entire list of 726 candidate expressed retrocopies with strong evidence.

Format: XLS Size: 179KB

Major categories of retrocopy contributions

To evaluate the types of events that led to new functional gene candidates or modifications of existing genes, and to reduce the possibility that a given transcript resulted from genomic priming, we increased the stringency factor for evidence of expression and examined in more detail a reduced set of 726 cases that overlapped at least five ESTs and one mRNA or annotated as a gene in RefSeq or UCSC KnownGenes (derived from Swiss-Prot).

We examined, in detail, all cases that were not purely duplication events, specifically those events that exhibited evidence of exon acquisition (see Methods): 1) cases with multiple coding exons, and 2) cases that showed evidence of contributions from retrocopy ORFs in the antisense direction. In general, we found that the retroposition events can be described predominantly by the following three categories: Type I: exon acquisition, in which part of the retrocopy was included into an existing gene transcript, in particular, in which a portion of the retrocopy could potentially serve as

a protein coding exon. Type II: retropositional gene duplication, in which apparently no pre-existing host gene at the site of insertion was altered. Type II events, in order be functional, would require recruitment of resident regulatory elements at the site of insertion, such as promoters and/or enhancers, and the process may have been accompanied by intron generation, for example, to reduce the size of 5' UTRs (Brosius and Gould 1992). Finally, Type III retrocopy events occur, in contrast to Type I and II events, when the retrocopy contributed a sequence that is largely out-of-frame, derived from a UTR, or in the opposite orientation with respect to the retrocopy's parent. Other flanking DNA sequences, including those derived from other transposed elements (SINEs, LINEs, endogenous retroviruses, and DNA transposons), may also be co-opted into the structure of these *ab initio* gene candidates. Hence, the Type III genes, if functional, have a protein sequence that is mostly novel.

Comparison with other datasets

We found good agreement between our candidate set of transcribed retrocopies and the major transcribed retrocopy datasets produced by other groups (Vinckenbosch et al 2006, Harrison et al 2005, Zhang et al 2003). Of the 223 transcribed retrocopies reported by Harrison (Harrison et al 2005, Zhang et al 2003) we agreed with 189 cases. Randomly selected cases that were missing from our set were cases that relied soley on scarce EST data and fell below our threshold. Of the ten randomly selected cases in our set that were missing from the Harrison set, 30% were present in the

Kaessmann set (Vinckenbosch et al 2006). Of the remaining cases not present in our set or Kaessman's, 20% have weak expression evidence but were nevertheless classified as retrocopies by Harrison [See Additional File 1]. In contrast, we found many examples in the HOPPSIGEN dataset (Khelifi et al 2005) that were not present in our dataset or Kaessman's set. A random sample of ten were all found to be either segmental duplications or inactive LINE elements that we assume were false positives in their data. Kaessmann reported 1,080 expressed retrocopies with at least one EST (Vinckenbosch et al 2006). We agreed with 936 of these. Most of the cases missing from our set had mitochondrial, immunoglobin or zinc finger genes as parent genes. These were systematically excluded from our dataset because they are frequently generated by a different mechanism, i.e., segmental duplication. We reported 936 cases that were not present in Kaessman's set. Most were due to the smaller starting gene set that his pipeline used and exclusion of parental UTRs from the analysis.

Although the functional potential of Type II retrogenes was discussed early on (Brosius 1991, Brosius and Gould 1992) and overwhelmingly substantiated over the past 10 years (Vinckenbosch et al 2006, Balakirev et al 2003, Long et al 2003, Yu et al 2007), only a few Type I exon-acquisition events have been reported (Dupuy et al 2002) as well as de-novo gene evolution (Levine et al 2006, Begun et al 2007, Zhou et al 2008). The significant number of Type I and Type III events that we report demonstrates the extent of the contribution of retrocopies to the evolutionary

processes that test, reject, and retain novel amino acid encoding sequence space. All the retrogene candidates fall along a continuum from a large degree of similarity (Type II) to little similarity (Type III) to the original sequences in their respective parent genes. Many of the putative, novel retrogenes, potentially encoding proteins with no similarities to other existing proteins, may have been missed by methods relying on protein alignments, as protein-based screening methods cannot find antisense insertions and also are not able to align UTR regions of retrogenes. Protein alignment methods miss Type III retrogenes entirely. The retrocopies involved in generation of Type III gene candidates made relatively small, but potentially 'seeding', contributions to the formation of novel genes. Of course, it must be emphasized that most of the Type II and Type III mRNA retrocopy-derived "retrogenes" described in this study are putative genes for which no proteins have as yet been documented. While some of these new transcripts may code for proteins, others may serve as non-protein coding RNAs, possibly involved in cellular regulation (Zheng and Gerstein 2007, Sasidharan and Gerstein 2008) or in chromatin remodeling (Shen et al 2000).

Overview of types of events whose expression was strongly supported

Of the 726 candidate retrocopies whose expression was supported by many transcripts, 624 were composed of single protein coding exons and 102 contained multiple protein coding exons. The 102 cases came from a set of manually examined cases that overlapped known genes with more than one exon (about 500 cases),

single exon cases transcribed in the reverse orientation based on EST and mRNA evidence (32 cases), and a random sample of the other single exon cases that slipped through our initial screen due to alternative splicing (see Table 1). We compared: 1) the phylogenetic conservation of the ORF in the various species listed in Methods, 2) the relative contribution of the retrocopy to the new gene, 3) the relative contribution of the host gene (where applicable), 4) contribution from other types of transposed elements, 5) whether the retrocopy inserted in the sense or antisense orientation, and 6) we compared the parent ORF to the retrocopy ORF looking for frameshifts and mutations. Conclusions based on phylogenetic analysis are to be treated with caution as the non-human primate sequences contain a sufficient percentage of mistakes, erroneously indicating lack or presence of phylogenetic conservation in predicted ORFs.

Table 1. Description and distribution of expressed retrocopy events

Type of event	Parent gene contribution	Count	Percentage
Type I – Exon acquisition (host gene modified by retrocopy)	New 5' exon (UTR and/or N-terminal protein coding)	10	1%
	New 3' exon (UTR and/or C-terminal protein coding)	18	2%
	New internal exon	6	1%
Type II – duplication (no host gene involved)	Single exon	624	86%
	Exons/introns generated, post insertion	55	8%
Type III – novel genes (no host gene involved)	Antisense, majority of ORF out-of-frame wrt to parent, and other cases (e.g., from non-genic regions)	13	2%
Total		726	

Type I: retrocopy inserted into or near an existing gene. A portion of the retrocopy contributes, mostly by alternative splicing, a new sequence to a pre-existing mRNA. Type I events can be divided into cases that add new N- or C-terminal encoding exons or internal

exons. Type II: duplicated gene inserted at a locus where no prior gene existed. Type II events often acquired 5' or 3' UTR portions from the locus of integration after the insertion. Type III: novel gene sequence, whose encoded protein has little or no amino acid sequence similarity to that of the retrocopy's parent. Frequently, Type III events include SINEs, LINEs, LTRs etc., as well as unannotated sequences as additional contributors to gene candidates.

Type I – genes modified by exon acquisition

Of all cases with strong evidence of expression that we inspected, we identified 5% as being potential gene fusions, or exon-acquisition events. It is generally assumed that inserted retrocopies decay without affecting the structure of the host gene. However, we found several examples in which part of a retrocopy ORF integrated into the host gene (Figure 1A, categories 1–4, 6; Table 2), and often led to alternative mRNA splicing (Figure 1A, categories 1, 4, 5). We cannot be sure of the duration of time between the retrotranposition event and the start of alternative splicing. The new splice sites were either fortuitously present in the ORF of the retrocopy, or they arose subsequent to the integration by base changes over time. We did not observe splice sites in retrogenes that coincided with the splice sites in the parent gene. This is not surprising, as important intronic parts of splice sites are removed on the processed mRNA templates prior to retroposition. Therefore, Figure 1 shows generic splice sites as dotted white vertical lines that do not coincide with splice sites used in the novel gene context. The six categories in Figure 1A are defined as follows: 1) Part of protein coding sequence from parent is used as alternatively spliced exon of the host gene. 2) Retrocopy contributes new 3' exon to host gene (mostly in-frame, magenta, and partially out-of-frame, dark red, with respect to parent gene). 3) In-

frame contribution (magenta) combined with out-of-frame contribution (dark red) form a new N-terminal encoding region. A short 5' UTR (medium size bar, dark red) has been generated from the ORF of the retrocopy.

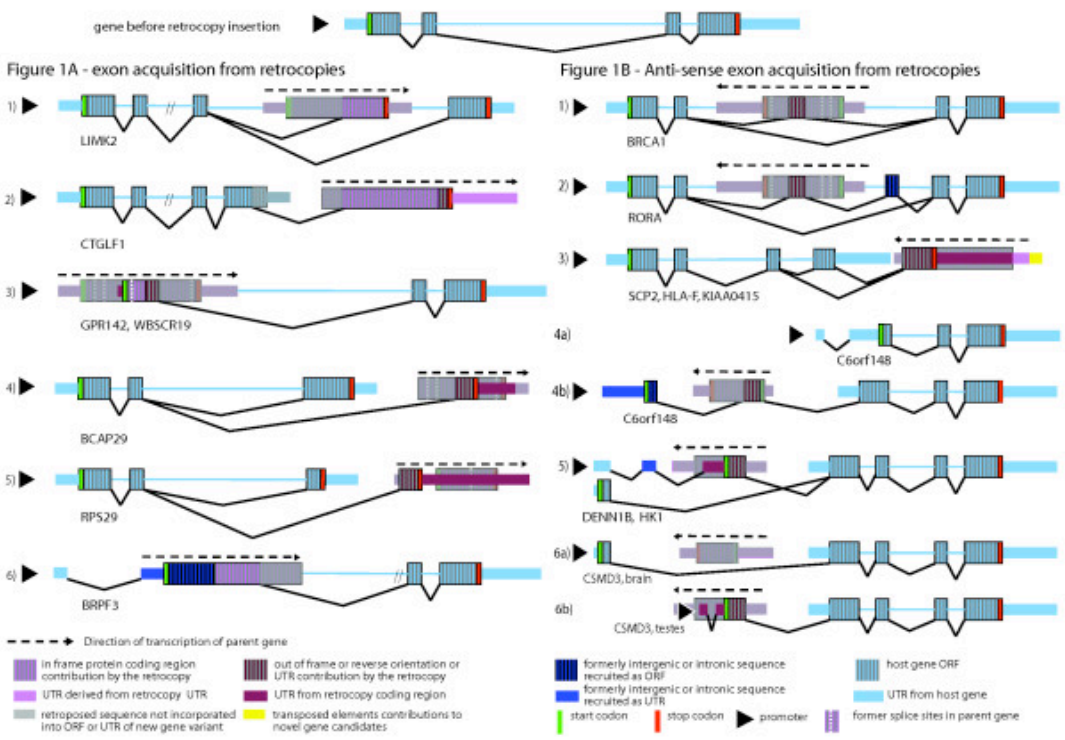


Figure 1. Categories of Type I retrocopy events. A. Examples of Type Ia exon acquisitions contributed by "same orientation" of retrocopies (in magenta or dark red) with respect to host gene (light blue); not drawn to scale, splice events are marked by angled lines, open reading frames are depicted as vertically striped thick bars, UTRs by medium size bars, introns in the host gene as light blue lines (for symbols and colors, see also keys below). When parts of retrocopies are described they correspond to what they used to be in the parent gene. The retrocopy's start and stop codons are shown by green and red vertical bars, respectively. Retrogene parts apparently not recruited as functional modules are overlayered with gray. B, Examples of Type Ib exon acquisitions contributed by "reverse orientation" retrocopies. For detailed descriptions, see text.

We also found several examples in which a putative novel exon had been exapted from an ORF, but the reading frame is now different from that of the parent gene (Figure 1A, category 4) which results in a shorter transcript. Also, putative novel exons were exapted entirely from sequences that correspond to UTRs of the parent gene, here alternatively sliced (Figure 1A, category 5). In other instances insertion of a retrocopy and exaptation of an exon from that sequence triggered recruitment of an additional exon entirely from intronic space. For example, in Figure 1A, category 6, the retrocopy contributes in-frame protein coding region (magenta) combined with unannotated intergenic sequence (dark blue) to form a new N-terminal encoding exon for the host gene. In turn, portions of the intronless retrocopy's protein coding region became an intronic sequence in the host gene (overlayed in grey). An interesting example of retrocopy mediated domain shuffling is the CTGLF1 gene (Figure 1A category 2), which started as a cyclin gene, and then had three domains (PH, ArfGap and Ankyrin) contributed by insertion of a CENTG2-derived retrocopy. The mouse version of this gene, AK132782, has only a cyclin domain and represents the ancestral form before the retrocopy insertion. These observations underscore the fact that natural selection exapts novel sequence space in addition to slowly modifying existing sequence space.

Type I exon acquisition in reverse orientation

While at least one group has reported on the existence of sense retrocopy integrations into existing genes, with coding contributions (Vinckenbosch et al 2006), this is the

first report of mRNA retrocopy integrations in the antisense orientation. The existence of this category of retrocopy events, if functional, supports the idea that natural selection has no preference with respect to the origin of novel sequences. In this category, novel exons were recruited from retrocopies that inserted into or adjacent to host genes in the opposite orientation to the retrocopy's parent gene. As in the case of Type Ia 'sense' retrocopies, the splice sites in the 'antisense' retrocopies, of course, do not correspond to those present in the parent genes. Of the Type Ib examples that we manually inspected, polypyrimidine tracts inserted by the retrocopy – used for recognition of splice sites – were frequently derived from antisense oligopurine tracts in the parent gene. These sequences are often rich in codons for lysine, glutamic acid, and glycine, as well as certain codons of arginine in the parent gene. A few Type Ib examples are described in detail below.

1) Internal exon (dark red) added to host gene in the opposite orientation relative to the parent of the retrocopy For example, the BRCA1 gene has an alternatively spliced internal cassette exon (potentially encoding 22 aa) contributed by RPL21 in the antisense direction (Figure 1B, example 1; Table 2). The insertion occurred after the New World monkey split and the reading frame is open in chimpanzee, orangutan, and rhesus monkey.

2) Internal exon added to host gene triggered recruitment of an additional protein coding exon from formerly intronic sequence (dark blue). For example, RORA

acquired an internal cassette (encoding 25 aa – PDB structures 1N83 and 1S0X and Swiss-Prot P35398) from an antisense retrocopy of CYCS. Interestingly, a second exon (encoding 27 aa) appeared in conjunction with the retrocopy-derived exon, apparently derived from an intronic sequence, that maintains the frame of the gene (Figure 1B, example 2). The open reading frame is maintained in orangutan, rhesus monkey, and marmoset, but there is an early in-frame stop in chimpanzee (confirmed by our re-sequencing, unpublished) – an example in which one lineage did not retain such an innovation (see discussion).

3) Recruitment of a 3' exon including novel ORF and 3' UTR generated from ORF of the retrocopy; it extends the ORF of the host gene (Figure 1B, example 3). Examples are SCP2, HLA-F, and KIAA0415 with potentially functional variants that have alternatively spliced 3' ends that are derived from antisense retrocopy insertions of RRAS2, RPL23, and FLJ10324, respectively. The SCP2 variant that includes the retrocopy-derived exon has a shorter transcript (potentially encoding 338 aa instead of 547 aa) and is only present in chimpanzee and human. In HLA-F the insertion generated a longer ORF (potentially encoding 442 aa instead of 362 aa; Figure 1B, example 3). Importantly, the retro-derived variants of SCP2 and HLA-F are reviewed NCBI Refseq genes.

4) Portion of the retrocopy contributes a potentially alternatively spliced protein coding exon in conjunction with a novel protein coding exon generated from

intergenic sequence (dark blue). For C6orf148, we detected two mRNA variants; the first, depicted in Figure 1B (example 4a), presumably also represents the ancestral status prior to the retrocopy insertion. The second putative variant has an alternative, upstream promoter, a new first exon from an unknown source, and a second protein coding exon (Figure 1B, example 4b) derived from the EIF3S6 retrocopy in reverse orientation. Surprisingly, the third putative coding exon in the second variant is also longer than the corresponding N-terminal coding region in the original variant. Part of the EIF3S6 UTR was potentially exapted as a protein coding sequence (example 4a). Both splice forms and open reading frames coexist in chimpanzee and rhesus monkey.

5) Alternative splicing or alternative translation after retrocopy insertion. The first putative protein coding exon becomes one of the 5' UTR exons (light blue); a second 5' UTR exon is recruited from unknown sequences. The first putative protein coding exon (dark red) is recruited from the retrocopy (Figure 1B, example 5). As in the aforementioned examples, DENN1B and HK1 also exhibit mRNA variants with and without their respective retrocopies. Their promoters are shared by both variants and both have putative alternative translation starts. Interestingly, in both cases the version with retrocopy contribution does not start transcription in the first exon, but instead, includes a second UTR exon before splicing to the retrocopy-derived putative antisense coding exon. The next protein coding exon is shared by both variants. The ORF for HK1 is open in chimpanzee, orangutan, and rhesus monkey.

DENN1B has valid ORFs only in human and chimpanzee, but the retrocopy-derived portion is present with disruptions in orangutan and rhesus monkey.

6) Two 5' UTR exons, intron, and N-terminal encoding exon are recruited from the protein coding region of the retrocopy. For example, one variant of CSMD3 is expressed in the brain and contains a sequence encoding a potential N-terminal 79 aa exon (Figure 1B, example 6a). The other putative variant of CSMD3 is expressed in testes (based on mRNA and EST evidence), and instead of the 79 aa exon uses part of an antisense RPL18-derived retrocopy; the event might also have led to use of a new promoter for the gene (Figure 1B, example 6b). In addition, a 5' UTR exon and a small intron were derived from the protein coding region of the retrocopy. Only the human version of the retrocopy-containing putative variant has an intact ORF.

Type II duplication events

Of the Type II events, 60 of them contained one or more 5' and/or 3' untranslated exons acquired along with regulatory elements derived from the flanking region of the insertion site [see Additional File 2, categories 2 and 3] as predicted previously (Brosius and Gould 1992). A few examples in which introns also arose in flanking UTRs were reported recently (Vinckenbosch et al 2006). In our numerous examples, we found no indication that the UTR introns came from the parent gene. Occasionally, a 5' or 3' exon recruited from the locus provided not only a UTR, but also the first or last protein coding exon [see Additional File 3]. We also observed

shorter and longer N- or C-terminal encoding parts of genes in separate lineages [see Additional File 2, categories 4–8,10,11], representing one of the mechanisms that might explain the frayed ends of many protein homologs, even orthologs (Weiner et al 2006). In addition, we found cases where an intron arose within the coding region after retroposition [see Additional File 2, category 8–11, Additional File 3]. These and the aforementioned examples (categories 2 and 3) underscore the notion that even intron-containing genes, especially those with large exons and relatively intron-impoverished with respect to their parent genes, can be derived from retroposition. Similarly, one or several introns of a gene can be lost by recombination with the corresponding retrocopies (Krasnov et al 2005).

Additional file 2. Classes of Type II retrogenes.

Format: PDF Size: 29KB

Additional file 3. Details on additional examples and Additional methods.

Format: DOC Size: 97KB Download file

Type III novel gene candidates

In a small fraction of the cases (16) we examined, a putative new gene with no known homologs included a retrocopy (usually only part thereof) that inserted into the genome and possibly provided protein coding sequence either 1) out-of-frame (Figure 2A–F, Additional File 4) or 2) antisense with respect to the parent gene

(Figure 2G–K, Additional File 4, Table 3 and Additional File 3).

Additional file 4. More novel Type III gene candidates.

Table 3. Type III novel retrogenes that are out of frame or reverse sense with respect to the parent gene

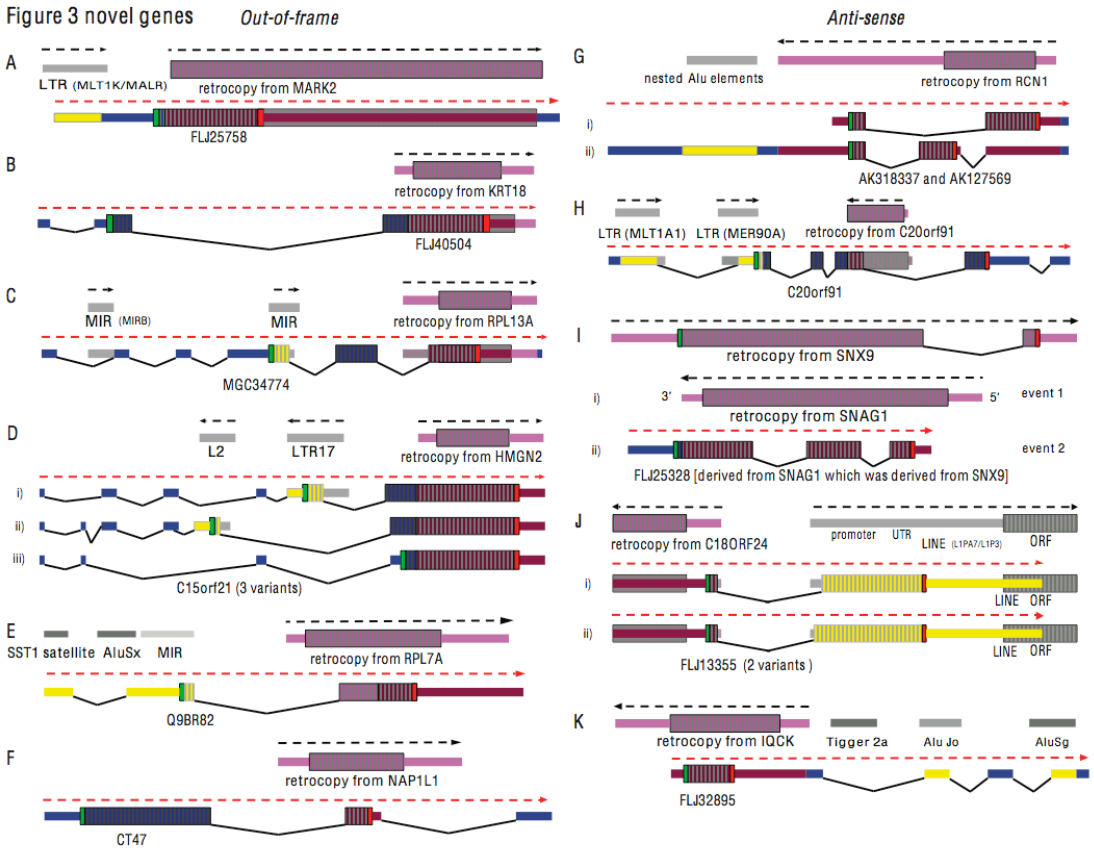


Figure 2. Novel protein-sequence space generated by parts of retrocopies combined with other transposons or unusual events. For each part of the figure, the spliced parent mRNA is shown first (before retroposition) and the resulting gene(s) are shown below. New sequence space was triggered by a combination of retrogene insertions, recruitment of non-genic regions including retroposons, whereby the contribution of the retrocopy's original in-frame ORF is very small (see text and legend to Fig. 1 including color key for further details). Yellow boxes with grey vertical stripes and yellow medium size bars correspond to retroposed element contributions to ORFs and UTRs, respectively. For detailed descriptions see text.

Examples are briefly described as follows: A) The novel candidate gene FLJ25758 was generated from MARK2-derived retrogene featuring protein coding sequence completely out-of-frame with respect to the parent. There is a potential LTR contribution of promoter and the remaining UTR sequences of FLJ25758 gene were

derived from flanking sequence shown as blue bars. B) The novel candidate gene FLJ40504 was generated from a KRT18-derived retrocopy out-of-frame with respect to the parent, the remainder of the protein coding sequence was derived from flanking region. C) The first protein coding exon of candidate gene MGC34774 was derived from a MIR element (yellow); the second exon from intergenic region (blue) and the final protein coding exon from the 5' UTR and ORF of L13A-derived retrocopy out-of-frame (dark red). D) The novel gene candidate C15orf21 is processed into three alternatively spliced transcripts, two of which use retroposed sequences as first protein coding exon (yellow). The second protein coding exon, in part from unknown sequences, was fused to an HMG14-derived retrocopy completely out-of-frame. E) The RPL7A-derived retrocopy contributes half of the ORF in-frame (magenta) and half out-of-frame (dark red) yielding novel gene candidate Q9BR82. Upstream exons were contributed by repeats (yellow). F) CT47 gene was generated, in part, from a NAPL1-derived retrocopy, which contributed the C-terminal encoding region out-of-frame. Most of the ORF (blue) arose from intergenic sequence. G) Two alternatively spliced mRNAs originated from an RCN1-derived retrocopy (3' UTR and ORF) in reverse orientation. H) LTRs contributed UTR and first protein coding exons (yellow); other exons were derived from intergenic sequence (blue); candidate gene C20orf91 contains part of a LOC16236-derived retrocopy in opposite orientation. I) Two events, retroposition of SNX9 followed by a second retroposition of SNAG1 or segmental duplication

formed the intron containing candidate gene FLJ25328 in reverse orientation to SNAG1, presumably with generation of two introns out of ancestral ORF sequences. J) A novel gene candidate FLJ13355 (2 splice variants) was formed from a C18orf24-derived retrocopy (retrocopy ORF contributed 5' UTR and retrocopy 5' UTR contributed N-terminal encoding region). The second, larger exon including a large part of ORF was contributed by the internal promoter and 5' UTR region of a LINE element (yellow). K) An IQCK-derived retrocopy contributed the protein coding exon in reverse orientation (dark red) yielding novel gene candidate FLJ32895. Downstream UTR exons were derived from Alu elements and intergenic sequences.

One can summarize the following: Although there is no evidence that a protein is produced, the size of the putative new ORFs ranged from 81 to 259 aa in human, and seven maintained open reading frames in human, chimpanzee and orangutan. Only four cases also had open reading frames in rhesus monkey and only two also in marmoset (Table 3). Suprisingly, all but three had multiple putative exons, lending further weight to the notion that we were not observing random transcription. While fusions between existing genes and mobile elements have been described (Cordaux et al 2006), we also observed exons that were generated, in conjunction with the retrocopy, by other types of transposed elements and/or unannotated sequences. For example, most of the CT47 gene (Chen et al 2006) which has evidence of protein coding sequence (Swiss-Prot Q5JQC4) arose initially from unannotated sequence

(see below), amplified by segmental duplication and has a valid ORF in human, chimpanzee, orangutan, rhesus macaque, and marmoset. Other cases in which a gene arose from unannotated sequences have been described in flies (Levin et al 2006, Begun et al 2007, Zhou et al 2008). Out of the 16 primate cases, seven had putative protein coding regions that included repetitive elements [shown in yellow tall boxes, Additional File 4]. One was composed of a chimeric fusion of two retrocopies [see Additional File 4, Additional File 3].

Functional categories of source genes

We looked at GO annotations of the parent genes that spawned retrocopies. For both the expressed on non-expressed sets of retrocopies we found no statistically significant enrichment from a normally distributed set.

Discussion

The more than 700 instances of evidently transcribed retrocopies in the human genome indicate that this process has contributed significantly to our transcriptome, and may have contributed novel protein coding segments more often than previously appreciated. How these changes altered the content and regulation of the repertoire of protein coding genes remains to be experimentally investigated. However, it is evident that the retrocopy mechanism provides an extremely versatile means to tinker with gene structure, and many of the topologically possible kinds of novel exon recruitment events seem to have been explored.

We observed retroposition events that happened prior to vertebrate diversification, those that date from after the *Homo sapiens* lineage diverged from that of the chimpanzees, and those that occurred at times between these events. Each of these events was/is initially a chance event – probably neutral at best. The combination of neutral forces and natural selection are responsible of the future trajectories, whether such a gene with retrocopy contribution persists or is abandoned in all or some of the subsequent lineages after a "trial period", which can be very short or last for tens of millions of years. This does not differ from genes that arose entirely by segmental duplication. In either case, a certain proportion is eventually discarded in all or some lineages, after being active for some time, while others persist.

Does segmental duplication or retroposition lead to more functional duplicated genes? The fact that some retrocopies lack their own promoters might be construed as a disadvantage. However, the high percentage (as much as 70%) of the genome being transcribed (Pheasant and Mattick 2007) should help neutralize this apparent disadvantage. Moreover, recruitment of a novel promoter might be an advantage over that of an amplified gene that was generated via segmental duplication. In fact, many retrogenes exhibit expression patterns that are drastically different from those of the parent genes (Brosius 1991). Perhaps the increase in retroposition in primates has allowed greater regulatory flexibility to evolve in a relatively short time.

The uncertainty of a long life after the birth of a gene is not much different from the

exonizations of novel sequence domains in "established" host genes. Krull examined the history of five Alu element exonizations by phylogenetic analyses and found that many of these events did not persist in all lineages in which they were exonized (Krull et al 2005). In contrast, of five MIR element exonizations analyzed, four are present and expressed as mRNA in all mammals examined (Krull et al 2007, Sela et al 2007), and there are likely many more – thus far unproven – MIR exonizations. When comparing the older MIR exonizations with the younger Alu exonizations, it is apparent that over the past billion years, most exonization were transient and, due to low levels or lack of evolutionary pressure, did not persist. As an aside, it appears that exonizations can occur at any time following the SINE insertion. There is one apparently old and previously neutrally evolving MIR that only "recently" was exapted as a protein coding exon in the alpha1 nicotinic cholinergic receptor gene in great apes (Krull et al 2007). Likewise, it is conceivable that a retrocopy or part thereof can be exapted at any stage of decay (see C20orf91 example). At the same time, it is clear that some of the aforementioned Alu exonizations have persisted in some lineages and not in others. The inability to predict whether an event will persist is equally impossible for younger Type II retrocopy insertion events.

Novel exon recruitment often occurs at the ends of genes. Eleven of the 36 Type I (exon acquisition) mRNA retrocopy events added exons at the 3' or 5' ends of the transcripts, while only four transcripts contained coding exons inserted internally, the remaining cases have new UTR exons added to existing genes. This makes sense if

one considers that only one splice site needs to be added to extend the ORF at the end of the gene. We also observed that Type II retrogenes, presumably during phases of no or little purifying selection and/or during periods of positive selection, frequently changed in different lineages, mostly with respect to the N- and C-terminal encoding parts of the ORFs (i.e., frayed ends). For example, in one lineage the C-terminal encoding part of the ORF is truncated compared to the orthologous region in another lineage. It might be reasoned that changes at protein extremities are better tolerated than elsewhere in the protein. This is supported by large-scale sequence comparisons of orthologous proteins, in which the terminals vary more than do the rest of the proteins (Weiner et al 2006, Kalamegham et al 2007).

Apart from the acquisition of existing protein domains (Type I events), the acquisition of novel protein sequence space via retroposition (Type III events) plays a role in ab initio formation of genes. Apart from hitherto neutrally evolving sequences (see below under iv), these putatitvely novel genes can include i) ORF exons from what corresponds to 5' and 3' UTRs of the parent genes; ii) ORF exons out-of-frame with respect to the parent gene; iii) ORF exons from any retrocopy part, inserted in the antisense orientation; and iv) ORF exons from intronic sequences or intergenic regions adjacent and in addition to the co-opted retrocopy parts. Extreme cases are those examples in Figure 2, in which the retrocopy did not contribute much novel or pre-existing sequence space, but contributed to the formation of potential genes out of unannotated intergenic regions or retroposons.

Conclusion

Examining the births of these, as yet, putative retrogenes provides us with important ideas concerning the evolution of older, known genes – or parts therof. Of course, for young events the "gene" status is hard to prove short of experimental verification of a functional protein product, and for older events the history of the gene modules is more difficult to discern. These points notwithstanding, whether or not the examples presented will stand the test of time – either on an evolutionary scale or in the lab (if experiments can be devised that are conclusive), this is clearly another way that novel genes or gene variants can arise. We see evidence of the diversity of tinkering that occurs with genes beyond simple point mutations. The generation of new transcripts with protein coding potential derived from anti-sense retrocopies would be an unexpected contribution to protein evolution, if function can be shown. Thus, retroposition is certainly a mechanism that can help explain the differences that we see in phenotypes between species.

Methods

Algorithm for finding candidate retrocopies in the human genome

The search for and identification of retrocopies and their corresponding parent genes have been confounded by the existence of gene duplications generated by other evolutionary processes, such as segmental duplication. To avoid such difficulties we first aligned all human mRNAs (with poly-A tails removed) to the human genome

using BLASTZ, and looked for sites where mRNAs aligned in more than one location (Schwartz et al 2003), indicating that one or more gene copies have been made. If one of the locations was annotated as a known gene (referred to as the "parent gene"), we then assigned a confidence score, based on the analysis of a feature vector, to each of the other alignment hits to determine if a retroposition event had occurred. For each putative retrocopy locus, we constructed a feature vector from a number of features (listed below) and used a score function to assign a weight to each feature associated with a retroposition event. A schematic of the entire 'pipeline' is presented in Figure 3.

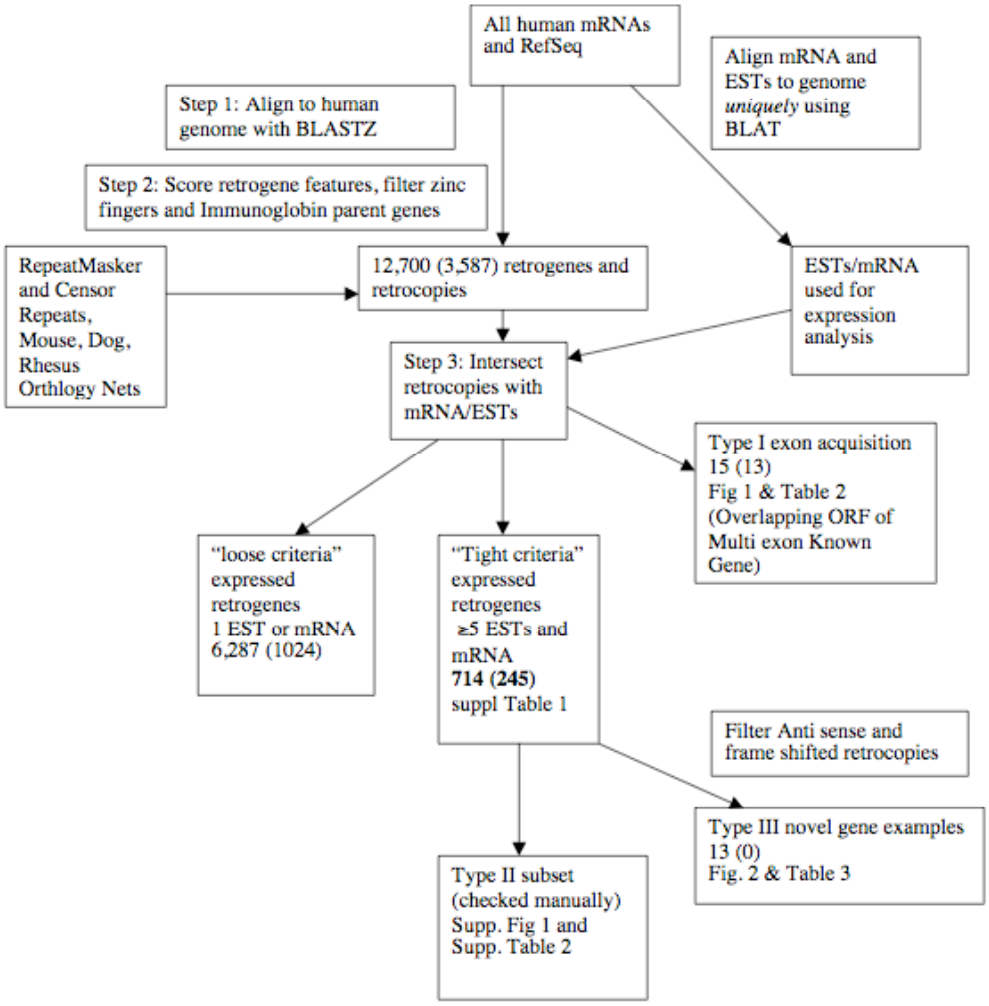


Figure 3. The retroFinder pipeline for annotating retrocopies. Alignments of all human mRNAs that aligned more than once to the genome were scored for a set of features (see Methods). Number of strict ESTs, mRNAs, and size of ORF were applied to determine evidence of expression. Retrocopies that partially overlapped the protein coding region of annotated multi exon Refseq genes were classified as exon acquisition events. Numbers in parenthesis were reported previously (Vinckenbosch et al 2006)

Score Function

To exclude segmental duplications (without retrocopy origin) and other non-retroposition related events, we used a score function based on a weighted linear combination of each feature (described below) to evaluate whether a retrocopy event had occurred. For each alignment above, we generated a set of features and applied the score function [see Additional File 3] with a threshold. We trained the weights and the threshold on a set of curated processed pseudogenes (non-functional retrocopies) from the Vega dataset (Ashurst et al 2005). Vega was chosen since they have a set of carefully manually curated set of pseudogenes using standards used by Sanger in the Vega pseudogene annotation. We used a set of 2,838 processed Vega pseudogenes as positive examples and 303 Vega non-processed psuedogenes as negative examples to set the weights. Chr18 was held out from the Vega training set so that we could check the weights against a set that was not used for training. We then checked high scoring false positives against the Vega set of processed pseudogenes. In some cases we found misannotations in their dataset and they updated their procedure for curation. They discovered that there were differences in how the curators defined pseudogenes that were rectified in subsequent Vega releases (J. Harrow, personal communication).

The following score function was used to combine the features defined below. We

used a threshold of 650 to filter the set of 12,801 retrogenes. The threshold was set based on with the Vega processed pseudogene set.

$$retrocopyScore = \sum_{i=0..7} w_i f_i(x_i)$$

where functions normalized to scale from 0 to 1000 and weights are between -1 and 1,

$f_0(x)$ = percent identity to parent gene, $w_0 = +0.3$,

$f_1(x) = \log_2(\text{exons removed}+1)*200$, $w_1 = +0.85$,

$f_2(x) = \log_2(\text{chained alignment score})*170 - 1000$, $w_2 = +0.7$,

$f_3(x) = \log_2(\text{length poly A tail} +2)*200$, $w_3 = +0.4$,

$f_4(x) = \max(\text{percent coverage of ortholog in mouse/dog}) * 10$, $w_4 = +0.3$,

$f_5(x) = \text{sqrt}(\text{count of introns})*750$, $w_5 = -1$,

$f_6(x) = \text{percent coverage of parent}*(1-\text{percent truncated 3'}) *300$, $w_7 = 1$,

$f_7(x) = \text{percent overlap repeatMasker (SINES or LINES)} *10$, $w_8 = -1$,

Description of Features

For each putative retrocopy alignment we extracted the following set of features:

• The most obvious sign of retroposition is the presence of multiple contiguous exons with introns removed. This signal can be weakened by any insertions, deletions, and substitutions that occur after retroposition. We counted the number of contiguous processed exons in the retrocopy and compared that to the parent gene. We did not count any recent Alu/LINE insertions as introns, as that has been a problem with other methods (Zhang, D; personal communication). When we aligned the mRNA to both parent gene and putative retrocopy loci, we were able to map the location of the breaks in the alignment back to the mRNA coordinates. For the parent gene, most of these insertions (larger than 35 bp) corresponded to introns. We made the assumption that if the insertion was larger than 35 bp and it occurred within 7 bp of the splice site in the parent gene, then it was a spliced intron at the retrocopy location. If no splice site was found at this location in the retrocopy and the mRNA was alignable at this location, then it was counted as a "spliced exonn. This feature had the heaviest weight assigned as it is the strongest signal of retrotransposition [see Additional File 3].

• Conserved splice sites were counted by looking at the position of the splice sites in the parent gene in cDNA coordinates. These positions were mapped to the retrogene and any break in the alignment that was larger than 35 bp and within 15 bp of the splice site in the parent gene were considered a retained splice site and reduced the count of the spliced exons. Most retrocopies should not have any conserved splice sites, but occasionally they have retained an intron due to incomplete processing

prior to insertion. Therefore this feature's weight was lowered to allow a small number of conserved splice sites provided other signals were strong.

• We counted the number of introns in the retrocopy by looking at gaps in the alignment of the parent cDNA to the retrocopy locus that were larger than 35 bp. If there was no corresponding gap on the cDNA side of the alignment, or if the break in the alignment on the genome side was at least three times larger than the break in the alignment on the cDNA side, then it was counted as an intron and assigned a large negative weight. We masked out Alu/LINE repeat insertions before calculating this feature to avoid counting recent transposons as introns.

• The "relative orthology with mouse" feature took advantage of the fact that retrocopies inserted since the mouse/human divergence show a break in the human-mouse genomic alignment as defined by the UCSC syntenic alignment nets (Kent et al 2003). Relative orthology is defined as the ratio of the size of the putative retrocopy to the size of the genomic insertion – defined by the break in the alignment "net" (from by the UCSC Browser). Ratios close to one represent possible retroposition events and ratios close to zero are most likely segmental duplications. Non-processed pseudogenes tend to score low because they are often generated via large segmental duplication events. We used the same relative orthology feature with the dog/human and rhesus monkey/human alignment nets to avoid false assignments due to deletions in mouse. The weight on this feature was less than the previous

feature since older events will not show a break in orthology.

• The poly(A) tail feature measured the length of the poly(A) tail that was inserted into the genome during retroposition (Vanin 1984). For more recent insertions, this signal distinguishes retrocopies from non-processed pseudogenes. The poly(A) tail was determined to be the largest scoring segment in a window of 70 bases near the end of the retrocopy. A's were scored +1 and the remaining three bases were scored -1. The weight was rather low, because poly(A) tails can also arise by chance or secondary retroposon insertions, thus, they did not add much weight to the score unless they were quite long.

Since we used a score function to classify retrocopies, the absence of one or two features did not exclude a given candidate. In this way we identified retrocopies that did not show orthology breaks in mammals, which also enabled us to identify older events. Likewise, absence of a poly(A) tail (e.g., in older or truncated retrocopies) did not lead to exclusion. To evaluate the transcriptional competence of the retrocopy set, in addition to the normal criteria used in the UCSC Genome Browser (Kent et al 2002, Hinrichs et al 2006, Furey et al 2004), we only used cDNA (mRNA) and EST evidence that uniquely mapped to the retrocopy and not to the parent gene in at least five nucleotide positions.

Filtering Alignments

We removed any candidate retrocopies that overlapped by more than 50% with

repeats identified by RepeatMasker (Smit et al 2004) and Tandem Repeat Finder (Benson 1999). Because RepeatMasker appeared to be overly aggressive in misclassifying pseudogene insertions as repeats, we corrected this by performing a base-by-base intersection with CENSOR (Kohany et al 2006, Jurka et al 2005) and eliminated only masking regions where RepeatMasker and CENSOR agreed. We also removed low percentage identity alignments (below 75%) that overlapped Alu elements. We found that recent independent Alus that were close to the parent and the retroposed segment were included in the alignments to the parent genes, generating false alignments, which we discarded manually.

Determining the Parent Loci

To find the locus of the originating parent gene, we looked for places where the parent mRNA aligned at least twice in the genome and defined the parent location as the best genome hit defined by the UCSC Genome Browser mRNA and refSeq tracks. The other non-overlapping hits define the locations of the retrocopies.

Resolving Conflicts from Multiple Parent Genes

To resolve conflicts posed by multiple potential parent genes, we initially took the simple approach of labeling as the most likely true parent the gene with the highest percent identity to the retrocopy. However, there were many cases in which retrocopies subsequently were copied via segmental duplications that occurred relatively recently during primate evolution (Bailey et al 2004, Bailey et al 2002). To

handle these cases, we selected the parental gene with the highest retrocopy score. Our aim here was not to unequivocally determine the correct parent gene among the segmentally duplicated copies, but merely to look for functional elements arising from retroposition regardless of whether they occurred before or after other duplications.

Filtering out Zinc Finger, Mitrochondrial and Immunoglobin Genes

We found a high number of potentially false positives in our initial gene set due to the large size of the zinc finger, mitrochonrial and immunoglobin gene families. Since many of these copies were generated via mechanisms other than retroposition, we excluded all of these cases from our dataset.

Determining Expression using mRNA and EST Evidence

The candidate retrocopies identified by the above process were then screened for an overlap with BLAT mRNA and EST alignments (Kent et al 2003). We used BLAT instead of BLASTZ since it is aware of splice sites and is better at aligning mRNAs to the genome. In cases where the retrocopy was very similar to the parent gene, we found that it was necessary to look at individual bases that were different between the two genomic locations. We required that the mRNA align better to the retrocopy location than did the parent gene at a minimum of 5 positions. We measured this by counting the number of sites (excluding SNPs) where an mRNA base differed from the genomic base to which it was aligned. A few such differences can occur in the

alignment of the parent gene's mRNA to the DNA of the parent gene due to polymorphism, or to errors in the mRNA sequences (Furey et al 2004). If we could not uniquely identify the genomic locus for an mRNA or EST, then it was not considered evidence of expression. We used the program bestOrf (Solovyev 1994) to score putative retrogenes for protein coding potential. Cases scoring less than 50 were removed. The program outputs potential CDS positions produced taking into account probabilities of each potential start codon, as well as longest ORF positions, extending the of CDS upstream from start codon). We then categorized the resulting set of expressed candidate retrogenes two ways: first, by type of evolutionary event, and second, by the strength of the evidence of expression. The entire pipeline is shown in Figure 3 (see additional files 4567 for numbers) .

Classification of Retrocopies according to Evolutionary Event

Finally, we classified the retrocopies found in our initial set of genes into three types. The Type I retrocopies contributed domains, in either the sense or antisense orientation, to known multiple coding exon genes (Hsu et al 2006) or to those found in the Refseq database (Pruitt et al 2005). Type II retrocopies consisted of duplications of one or more coding exons derived from the parent gene that formed new genes independent of host genes. Those cases with additional UTR exons or small coding exons derived from intergenic regions were also considered Type II. All remaining cases were considered Type III, defined as retrocopy contributions, in either the sense or antisense orientation, to novel genes out-of-frame with respect to

the retrocopy parent, combined with major contributions from other types of retroposed elements and protein coding segments derived from the unannotated genomic environment. Retrogenes containing Type III retrocopies were defined as having no significant BLAST alignment to any other protein coding genes (using an e-value threshold of 0.01).

Species Comparisons

In order to determine the age of the retrocopies, we used outgroup analysis. We checked for the presence and/or absense of all 726 candidate retrocopy genes in the following species (UCSC Genome browser databases http://genome.ucsc.edu *webcite (shown in parenthesis) :Homo sapiens* (hg18), *Pan troglodytes* (panTro2), *Macaca mulatta* (rheMac2), *Mus musculus* (mm8), and *Canis familaris* (canFam2). For the recent events (based on their presence in human and rhesus monkey and absence in mouse and dog) we selectively looked at the trace archives of *Pongo pygmaeus abelii* (orangutan) and *Callithrix jacchus* (marmoset). We manually assembled traces orangutan and marmoset from the NCBI trace archive to determine if the reading frame was open in the cases shown in tables 2 and 3. In a few selected cases of older events (present in all of the above species), we also examined chicken (galGal3) and *C. elegans* (ce2).

Authors' contributions

RB developed the algorithm, carried out the molecular genetic studies, performed the sequence alignment and drafted the manuscript. MD participated in the data analysis and methods development, JK participated in the sequence alignment and algorithm development, DH conceived of the study, participated in its design and edited the manuscript, JB participated in analysis of data, interpretation of the results and drafting of the manuscript. All authors read and approved the final manuscript.

Additional file 5. Type II retrogenes – selected cases.

Format: PDF Size: 237KB Download file

Additional file 6. TXNDC2 chimeric retrocopy.

Format: PDF Size: 31KB Download file

Additional file 7. DGCR13 and TSSK2 alignments in multiple primates.

Format: PDF Size: 87KB Download file

3. Burst of exapted human retrogenes not found in mouse and improved methods for identifying retrocopies.

Robert Baertsch, Rachel Harte, Mark Diekhans, James Kent, David Haussler

Abstract

To identify and analyze the processed pseudogenes and potentially functional retrogenes, we carried out BLASTZ alignments of all mouse mRNAs against the mouse genome and scored a set of features indicative of retroposition, using a modified method that we previously developed (Baertsch et al, 2008). To validate our improved method, we compared our results with pseudogene data sets from Yale, Sanger and Univ of Virginia. We performed a detailed analysis of the cases and list the strengths and weaknesses of the different methods.

In order to understand the effect of retrotransposition on the human and mouse lineages we compared two data sets: 1) processed pseudogene retrocopies; and 2) annotated retrogenes. In each set, we looked at three evolutionary timepoints: A) at least since the mouse human divergence; B) at least since the rat divergence in mouse and C) at least since the rhesus macacque split in human. After correcting for the faster evolutionary rate in mouse in set 1 above, we saw a drop off in retrotransposition in human relative to mouse, probably due to the expansion of the APOBEC3 gene family, as previously reported. Remarkable, this effect was weaker

in set 2 above (expressed retrgenes. After considering the effects of possible biases from the greater number of mRNAs and ESTs in human, we conclude that either a population bottleneck occurred after the divergence with macaque that reduced selection on deleterious retrocopies or the lower rate of retrotransposition induced by APOBEC3 somehow led to the doubling of the rate of fixation of retrogenes.

Background

Motivation

Understanding gene creation is essential to the study of human evolution. There are three known mechanisms of gene duplication: large segmental duplication (intrachromomal duplications greater than 20 kilobases), tandem duplication of genes (also known as tandem array duplication) and duplication by retrotransposition. Although recent segmental duplication accounts for more than 10% of the human genome (Bailey et al, 2002), it has been shown that the number of genes created by segmental duplication is smaller than the other mechanisms (Hahn et al. 2007). Although the ratio of retrogenes to tandem array duplicated genes (TAG), is close to one, it has also been shown that retrogenes are twice as likely to duplicate singletons when compared with tandem duplications (Pan and Zheng 2007). Accurately classifying different types of duplication events is an important precursor to identifying rates of these different events.

In the interest of functional gene annotation, we have collaborated with other groups to improve the annotation by screening out pseudogenes. In the process, we have devised improved methods for pseudogene annotation and that have been applied to annotated gene sets (Zheng et al 2007). As the number of sequenced genomes expands, it will become crucial to have good automated methods for annotation.

One of the other motivations for looking at retrogenes is that they are associated with some dramatic phenotypes. A recent publication showed that an expressed copy of FGF4 a growth factor in dogs is responsible for chondrodysplasia which is the cause of short legs in dogs. This polymorphic retrogene occurs on chromosome 18 in many types of dogs and has been maintained via selective breeding. Our aim is to provide a rich dataset for experimentalists to search for lineage-specific or polymorphic retrogenes.

Definitions

Retrotransposition is a process involving the copying of DNA by a group of enzymes that have the ability to reverse transcribe spliced mRNAs, resulting in single-exon copies of genes. By removing introns and flanking DNA, this results in copies of genes that lack much of their original regulatory sequence (Vanin 1984, Ohno 2004). Most of these retrotransposition events result in non-functional genes, known as processed

pseudogenes. Occasionally, the retrocopy will pick up the regulatory sequence of the region it gets inserted into and becomes an expressed gene.

Previous work on pseudogene and retrogene annotation

Numerous studies have looked at the rate of gene creation via retrotransposition. One study identified over 1,000 transcribed retrocopies in the human genome of which, 120 have evolved as bonafide genes using a fairly restrictive definition (Vinckenbosch et al. 2006). By considering cases with two introns removed from the parent transcript, they excluded a potentially large group of retrocopies. A group at Yale published a catalogue of processed pseudogenes showing that at least 8,000 pseudogenes in the human genome (Zhang et al 2003). The used protein alignments in their algorithm, excluding UTRs and other non-coding parent sequence from the resulting set. As part of the Vega project, the HAVANA group (Wellcome Trust Sanger Institute (WTSI) is manually annotating both genes and pseudogenes resulting in a higher quality than a automatically generated dataset. To date, they have annotated over 3,000 pseudogenes in the human genome. They have also classified them according to type: processed, unprocessed, unitary, polymorphic and immunoglobin pseudogenes. HAVANA also identifies transcribed pseudogenes, but not expressed retroposed genes. A recent study of positive selection on recent duplication looked at both retrogenes and other mechanisms of gene duplication, their method did not

consider cases where chimeric retrogenes result from incorporation of a retrocopy into an existing gene (Hahn et al 2007).

Previous work on rate of gene duplication in primates and non-primates

A recent study inferred a accelerated rate of gene gain/loss in primates, resulting in a gain of least 434 genes in the human genome and the loss of 740 genes in the chimpanzee genome since their split 5–6 MYA (Hahn et al. 2007, Han et al. 2009). These results imply that ~5% (1,174/22,000) of all human genes do not have a one-to-one ortholog in chimpanzee. Hahn also reported that in the human, macaque, mouse, and rat genomes, 10% of lineage specific duplications showed evidence of positive selection (Han et al. 2009), which was not true of the ancestral copy. This contrasts to a previous study where only 1.7% of single-copy orthlogous genes showed evidence of positive selection (Gibbs et al. 2007), implying that gene duplication plays in important role in explaining differences among mammals.

In order to tease apart the three main types of gene duplication, a group at Virginia Tech has estimated the rate of gene duplication via retrotransposition and tandem array gene duplication (TAG). They found that the percentage of genes duplicated via these two mechanism is 60% to 70% in human and mouse respectively (Pan and Zhang 2007). The remaining genes are duplicated via segmental duplication and other, as yet, unidentified

mechanisms. Thus Pan's group estimated between 585 and 700 human-specific retrogenes and between 727 and 857 that arose in mouse, not including retrocopies with multiple exons, which their method does not detect. It was also reported that the mouse genome gene duplication rate was about two to three times faster compared to human and also that TAGs were more active in multigene families while the rate of recent retrogene duplication was *10x higher* in single copy genes in both mouse and human (Hahn et al. 2007). To date, there is a rough agreement on numbers but there has not been a consensus set of pseudogenes and retrogenes making gene annotation a more difficult task.

To further complicate the story, previous work has shown that not all gene duplications are simply copies of their parent gene. Retrogenes can also be inserted into introns or intragenic regions and be incorporated into existing genes creating chimeric new genes. Kaessman identified 36 such events in the human genome and we also identified about the same number in a previous paper (Vinckenbosch et al. 2007, Baertsch et al. 2008). In this paper, we describe an automated pipeline to identify processed pseudogenes, transcribed as well as chimeric retrogenes genes in both human and mouse and we have therefore been able to increase the amount of genes annotated by this method.

This paper will describe the methods and tools used to identify retrogenes as well as look at the different types of impacts that that retrocopies have on existing genes These events

include: duplication of genes; adding protein-coding domains to a pre-existing gene; novel genes that are created from the UTR; and antisense insertion of retrocopies. A track in the UCSC Genome Browser describes the data and provides a visual means to see differences between the parent and retrogene. We compare our methods to the other retrogene sets; Yale, Vega, Swiss Bioinformatics Institute and the Virginia Tech data set. Finally our work is also being used by the Encode and Gencode projects to annotate human genes and this method also has application for annotating reconstructed ancestral genomes.

Algorithm

Overview

In section 2.2, we describe the algorithm, the features used to recognized retrocopies, and the function used to combine features and assign a score that represents the confidence that the candidate locus is a retroposed copy of an mRNA. In section 2.3, we document the tools and provide a step-by-step description of the pipeline. We also show how to display the dataset in the UCSC genome browser using track details pages and summary html tables that result from the analysis step.

We first aligned all human mRNAs (with poly-A tails removed) to the human genome using LASTZ (derived from BLASTZ), and searched for sites where

mRNAs aligned in more than one location (Schwartz 2002), indicating that one or more gene copies have been made. If one of the locations was annotated as a known gene (referred to as the "parent gene"), we then assigned a confidence score, based on the analysis of a weighted feature vector (described in the following section), to each of the other alignment hits to determine if a retroposition event had occurred.

Features

Exons-spliced feature

The most obvious sign of retroposition is the presence of multiple contiguous exons with introns removed. This signal can be weakened by any insertions, deletions, and substitutions that occur after retroposition. We counted the number of contiguous processed exons in the retrocopy and compared that to the parent gene. We did not count any recent Alu/LINE insertions as introns, as that has been a problem with other methods (Zhang, D; personal communication). When we aligned the mRNA to both parent gene and putative retrocopy loci, we were able to map the location of the breaks in the alignment back to the mRNA coordinates. For the parent gene, most of these insertions (larger than 35 bp) corresponded to introns. We made the assumption that if the insertion was larger than 35 bp and it occurred within 7 bp of the splice site in the parent gene, then it was a spliced intron at the retrocopy location. If no

splice site was found at this location in the retrocopy and the mRNA was alignable at this location, then it was counted as a "spliced exon". This feature had the heaviest weight assigned to it because it is the strongest signal of retrotransposition.

Conserved splice-site feature

Conserved splice sites were counted by looking at the position of the splice sites in the parent gene in cDNA coordinates. These positions were mapped to the retrogene and any break in the alignment that was larger than 35 bp and within 15 bp of the splice site in the parent gene were considered a retained splice site and reduced the count of the spliced exons. Most retrocopies should not have any conserved splice sites, but occasionally they have retained an intron due to incomplete processing prior to insertion. Therefore this feature's weight was lowered to allow a small number of conserved splice sites provided other signals were strong.

Intron feature

We counted the number of introns in the retrocopy by looking at gaps in the alignment of the parent cDNA to the retrocopy locus that were larger than 35 bp. If there was no corresponding gap on the cDNA side of the alignment, or if the break in

the alignment on the genome side was at least three times longer than the break in the alignment on the cDNA side, then it was counted as an intron and assigned a large negative weight. We masked out Alu/LINE repeat insertions before calculating this feature to avoid counting recent transposons as introns.

Break-in-orthology feature

The "relative orthology" feature takes advantage of the fact that retrocopies inserted since divergence with another species show a break in the pairwise genomic alignment as defined by the UCSC syntenic alignment nets (Kent WJ et al. 2003). Relative orthology is defined as the ratio of the size of the putative retrocopy to the size of the genomic insertion – defined by the break in the alignment "net" (Downloaded from http://hgdownload.cse.ucsc.edu/goldenPath/hg18/vsMm9/). Ratios close to one represent possible retroposition events and ratios close to zero are most likely segmental duplications. Non-processed pseudogenes tend to score low because they are often generated via large segmental duplication events. We used the same relative orthology feature with the dog/human and rhesus monkey/human alignment nets to avoid false assignments due to deletions in mouse. The weight on this feature was less than the intron feature since older events will not show a break in orthology.

Poly(A) feature

The poly(A) tail feature measures the length of the poly(A) tail that was inserted into the genome during retroposition. For more recent insertions, this signal distinguishes retrocopies from non-processed pseudogenes. The poly(A) tail was determined to be the largest scoring segment in a window of 70 bases near the end of the retrocopy. A's were scored +1 and the remaining three bases were scored -1. The weight was rather low, because poly(A) tails can also arise by chance or secondary retroposon insertions, thus, they did not add much weight to the score unless they were quite long.

Score function

We used the same score function as our previous paper (Baertsch et al 2008) with improved weights that were obtained by using the increased size of our training set on Vega. The “introns processed” feature weight was changed from 0.85 to 1.0. The chained alignment score weight was reduced from 0.7 to 0.1 since it was largely duplicated by the percent identity feature, and only added alignment length as an useful feature. Other improvements to the program were made to improve the detection of orthology breaks.

Tools

Source code and scripts

The code and scripts to run the retroFinder pipeline can be downloaded from http://compbio.soe.ucsc.edu/retrogene/ The included README file contains detailed instructions for running the pipeline.

Web pages for displaying Results

The final step, ucscRetroStep6.sh, generates web pages for both duplicated retrocopies and chimeric retrogenes. It categorizes the retrocopies by age using the breaks in orthology feature (see section 2.2.4). Comparing other datasets can be also quickly generated with the compareRetroset.sh scripts (see section 4 or http://hgwdev.cse.ucsc.edu/~baertsch/retro for example web pages). The variable ROOTDIR in the DEF file defines the directory where the HTML pages will be created. The SPECIES parameter in the DEF defines which web pages will be generated; entries should contain mysql database names of the genomes to be displayed. Counts are displayed on summary pages with click through to each individual example.

The retrogene track in the UCSC Genome browser

If we open the retrogene track in the UCSC genome browser (http://genome.ucsc.edu), the browser will display the mismatches between the parent and the retrogene. Matching bases between the parent gene and the retrogene are visualized in blue. Vertical bars of red

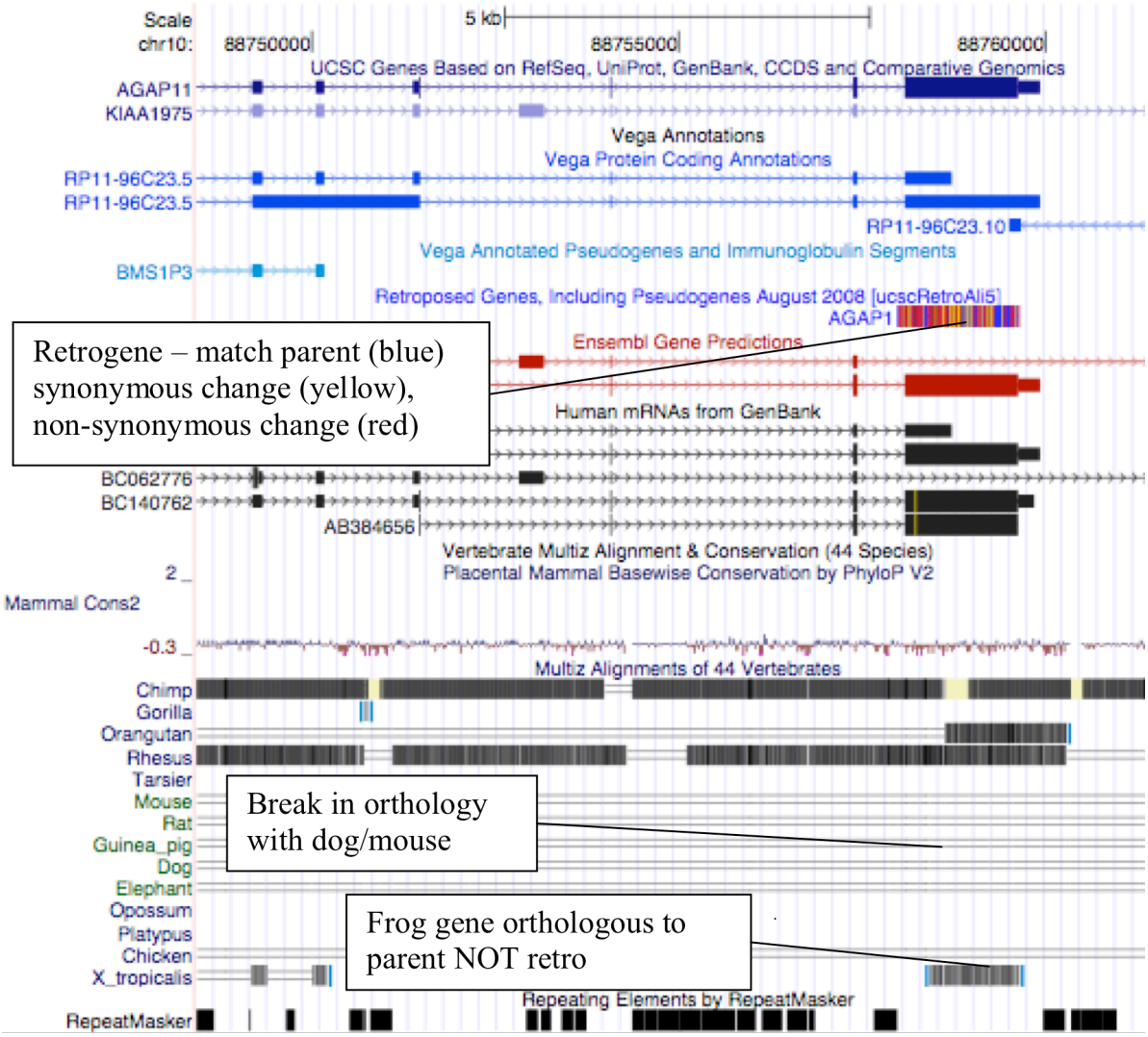
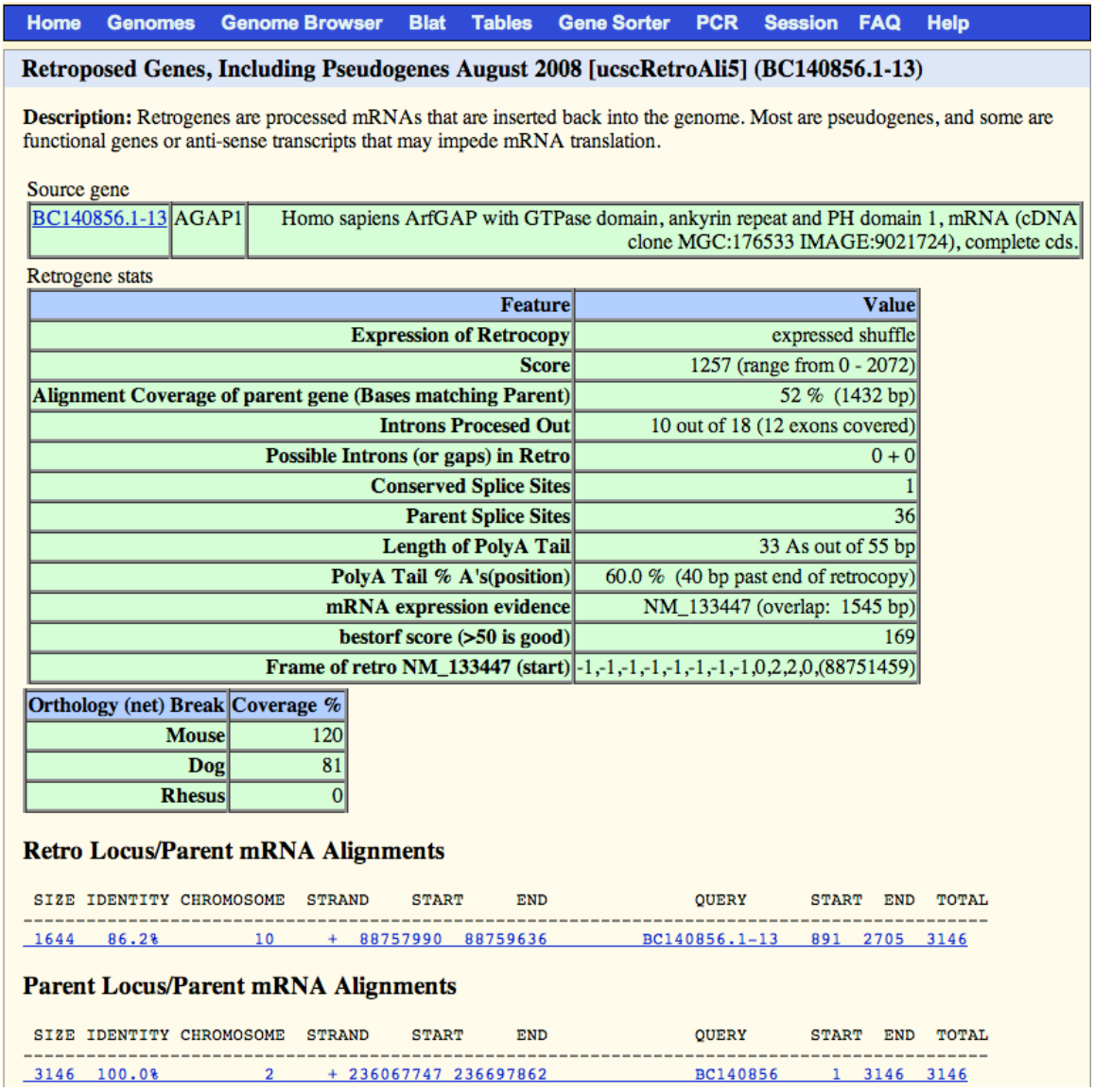


Figure 1 – UCSC Genome Browser with retrogene track displayed. AGAP1 (ankyrin repeat ArfGAP with GTPase domain, and PH domain 1) is the parent gene that was reinserted into the genome. Several upstream coding exons were presumably acquired later to form AGAP11. Bases that match the parent gene are colored blue, synonymous mismatches are yellow (positive BLOSUM62 score) and non-synonymous are marked red (negative BLOSUM62 score). Frog DNA is flanked by blue vertical bars indicated a break in the alignment thus denoting a lack of retrogene at this genomic locus in frog.

represent non-synonomous changes in the protein sequence and yellow is a synonymous mismatch. Clicking on the alignment of a retrogene in the track leads to the details pages for that item; the details page shows the alignment to the retrogene and parent as well as the values of each of the features used in the score function.



Home Genomes Genome Browser Blat Tables Gene Sorter PCR Session FAQ Help

Retroposed Genes, Including Pseudogenes August 2008 [ucscRetroAli5] (BC140856.1-13)

Description: Retrogenes are processed mRNAs that are inserted back into the genome. Most are pseudogenes, and some are functional genes or anti-sense transcripts that may impede mRNA translation.

Source gene

BC140856.1-13	AGAP1	Homo sapiens ArfGAP with GTPase domain, ankyrin repeat and PH domain 1, mRNA (cDNA clone MGC:176533 IMAGE:9021724), complete cds.

Retrogene stats

Feature	Value
Expression of Retrocopy	expressed shuffle
Score	1257 (range from 0 - 2072)
Alignment Coverage of parent gene (Bases matching Parent)	52 % (1432 bp)
Introns Procesed Out	10 out of 18 (12 exons covered)
Possible Introns (or gaps) in Retro	0 + 0
Conserved Splice Sites	1
Parent Splice Sites	36
Length of PolyA Tail	33 As out of 55 bp
PolyA Tail % A's(position)	60.0 % (40 bp past end of retrocopy)
mRNA expression evidence	NM_133447 (overlap: 1545 bp)
bestorf score (>50 is good)	169
Frame of retro NM_133447 (start)	-1,-1,-1,-1,-1,-1,-1,-1,0,2,2,2,0,(88751459)

Orthology (net) Break	Coverage %
Mouse	120
Dog	81
Rhesus	0

Retro Locus/Parent mRNA Alignments

SIZE	IDENTITY	CHROMOSOME	STRAND	START	END	QUERY	START	END	TOTAL
1644	86.2%	10	+	88757990	88759636	BC140856.1-13	891	2705	3146

Parent Locus/Parent mRNA Alignments

SIZE	IDENTITY	CHROMOSOME	STRAND	START	END	QUERY	START	END	TOTAL
3146	100.0%	2	+	236067747	236697862	BC140856	1	3146	3146

Figure 2 – Details page from UCSC Genome Browser after clicking on a retrogene. Features used to score the retrogene shown (see Methods). Alignment of mRNA to

retrogene as well as mRNA to parent genomic loci can be displaying using links near bottom.

Results and Discussion

Datasets for benchmarking

In order to test the accuracy of our method we used Vega Build 35, a set of manually annotated pseudogenes that was produced by the HAVANA group at WTSI (Wilming et al 2007). The Vega set is annotated processed and non-processed pseudogenes as well as protein coating genes. We compared retroFinder predictions with a subset of the Vega manually curated set of processed pseudogenes as well as transcribed_processed_pseudogenes (downloaded build35 Aug 2009 from ftp://ftp.sanger.ac.uk/pub/vega) and Yale processed pseudogene predictions (downloaded from http://pseudogene.org).

Comparison with Vega processed pseudogenes

After checking for overlapping predictions in the Vega and retroFinder sets, we found that 4,309 cases in human and 3,029 in mouse overlap at least 30% of the bases. (See http://hgwdev.cse.ucsc.edu/~baertsch/retro/vegaProc/). However Vega missed 5,016 cases in human and 11,941 in mouse that were predicted by retroFinder, using the same threshold. The Vega team produces high quality manually curated datasets and this is an ongoing process so they have not completely annotated all regions of both genomes. RetroFinder missed 849 cases in human and

254 in mouse. Manual inspection of these cases revealed that they are mostly missed by retroFinder due to stricter criteria when classifying processed versus non-processed pseudogenes. The most common example is single exon parent genes that do not have any introns to remove. We don't annotate these cases because they could be a result of some other evolutionary mechanism.

		Human (vega build35)	**Mouse (vega build35)**
a	**Agreement with Vega processed pseudogenes**	4,306	3,029
b	**Missed by Vega (processed only)**	5,016	11,941
c	**Missed by RetroFinter (before filtering)**	914	279
d	**Missed by RetroFinder (after filtering repeats and knownGenes)**	849	254

Table 1 – Comparison of Yale processed pseudogenes with retroFinder. Counts of predictions from retrofinder without evidence of expression compared with the Yale processed pseudogene set as well as the Yale processed pseudogene combine with cases marked as "ambiguous". See section 4.1.2 for explanation of differences.

Comparison with Yale processed pseudogenes

In human, 6,589 cases (see Table 2a) were present in both retroFinder and the Yale datasets. There were 8,124 in human and 9,825 in mouse (see Table 2b) predicted by retroFinder that were not identified as processed pseudogenes by the Yale pipeline. On closer inspection of these predictions, about 5,500 in human and 3,000 in mouse (see Table 2c) were classified as ambiguous, implying that the parent gene has one exon. However, the retroFinder pipeline requires multiple parent exons to be aligned

to the retrocopy, so we think a majority of these cases should be classified as processed pseudogenes. After screening the Yale dataset to remove predictions that are repeats or annotated genes, 821 cases in human and 743 cases in mouse (see Table 2d) were not found by the retroFinder pipeline. In this set, we found over 3,000 cases in human and 1,000 cases in mouse where more than 75% of the putative processed pseudogene overlapped with repeats annotated by RepeatMasker or Tandem Repeat Finder in the Yale set (data not shown). We inspected 20 randomly selected cases and found that 17 of these cases are in segmental duplicated regions or are tandemly duplicated genes. Looking at the common pseudogenes that Yale missed, we saw cases where processed pseudogenes were interrupted with transposon insertions (See Figure 3). Apparently they considered these insertions were introns instead of LINE or SINEs. Insertion of

Figure 3. Processed pseudogene (red and blue) recognized by RetroFinder that was missed by other groups. Alignment to parent gene CNOT1 shown in grey. LINE element and two LTRs that interrupt processed pseudogene shown in black.

pseudogenes within pseudogenes also creates problems in a few cases that we looked at. Further analysis is required to resolve these differences.

		Human (Build 55)	Mouse (Build 56)
a	**Agreement with Yale processed pseudogenes**	6,589	6,631
b	**Missed by Yale (processed only)**	8,124	9,825
c	**Missed by Yale (processed and ambiguous)**	2.621	5,890
d	**Missed by RetroFinder (after filtering repeats and knownGenes)**	1,199	925

Table 2– Comparison of Yale processed pseudogenes with retroFinder. Counts of predictions from retrofinder without evidence of expression compared with the Yale processed pseudogene set as well as the Yale processed pseudogene combine with cases marked as "ambiguous". See text for explanation of differences. Detail cases are listed at http://hgwdev.cse.ucsc.edu/~baertsch/retro/yale/ and http://hgwdev.cse.ucsc.edu/~baertsch/retro/yaleAmbig.

Comparison with Virgina Tech retrogenes

We also examined the Virginia Tech retrogene dataset and compared with our dataset; they appear to have a lot of false positives particularly resulting in cases where there were non-processed pseudogenes were annotated as processed and, in fact, they identified a supposed retrogene with 13 exons. For this reason, we did no further analysis on this set. In order to distinguish between introns and just gaps in sequence alignments, one method we employ is to align the splice junctions of the parent mRNA and compare this to the retrogene and if the splice structures from the parent mRNA map to alignment breaks in the retrogene, we downweight the score in the retroFinder score function to eliminate a lot of false positives (see sections 2.2.1 and 2.2.2).

Results compared with UCSC Known Genes and expression

Of the 12,101 human retrocopies and 16,419 mouse retrocopies, 1,370 and 909 respectively show evidence of expression as annotated by the Known Gene track (now known as UCSC Genes) on the UCSC genome browser (See Tables 3 and 4) Of those, 612 in human and 168 in mouse respectively had evidence of being protein coding according to UniProt.

	Est. time mya	Non-expressed and expressed retrocopies		Expressed retrocopies		Annotated retrogenes	
		Human	Mouse	Human	Mouse	Human	Mouse
In common Placental ancestor		1,901	669	664	338	913	406
At least since dog split	92	489	561	97	73	37	56
At least since human/ mouse split	80	**7,389**	**7,642**	**690**	**496**	**254**	**284**
At least since Rat/Mouse Split	40		7,546		1,275		162
At least since Rhesus split	25	2,060		699		147	
At least since chimp split	6	261		205		19	
Total		**12,101**	**16,419**	**2,355**	**2,182**	**1,370**	**909**

Table 3 - Counts of human & mouse retrocopies categorized by the time of insertion (rows) and by level of expression (columns). Dating was accomplished by looking at breaks in orthology between the human retrocopy and the corresponding genome position in dog, mouse, rhesus macaque and chimpanzee. Expressed retrocopies required at least 5 ESTs that map uniquely or one mRNA. UCSC known genes was used as the gene annotation.

There were also 400 retrogenes annotated as non-coding genes, 200 annotated as near coding and 54 annotated as antisense. These are further split out into genes that were created along the primate lineage, which was determined byfinding retrocopies that have a

break in orthology with respect to dog, mouse, rhesus macaque and chimpanzee (Table 3 rows 2,3,5). Ancient retrogenes, defined as having orthology with all three genomes analyzed were separately tabulated (Table 3 row 1). Similarly, mouse has 561 pseudogenes and retrogenes since the split with dog; 669 on the ancient common ancestor; 7,642 at least since the split with human ; and 7,546 at least since the split with rat (Table 3 row 4). To analyze the transcription of these retrocopies we used both mRNA and EST evidence as well as the known gene annotation track. To avoid the problem of ESTs that map to multiple places in the genome, we used a stricter filter than that used for alignments in the UCSC genome browser, and for this purpose, we developed a utility, pslCDnaGenomeMatch (see above).

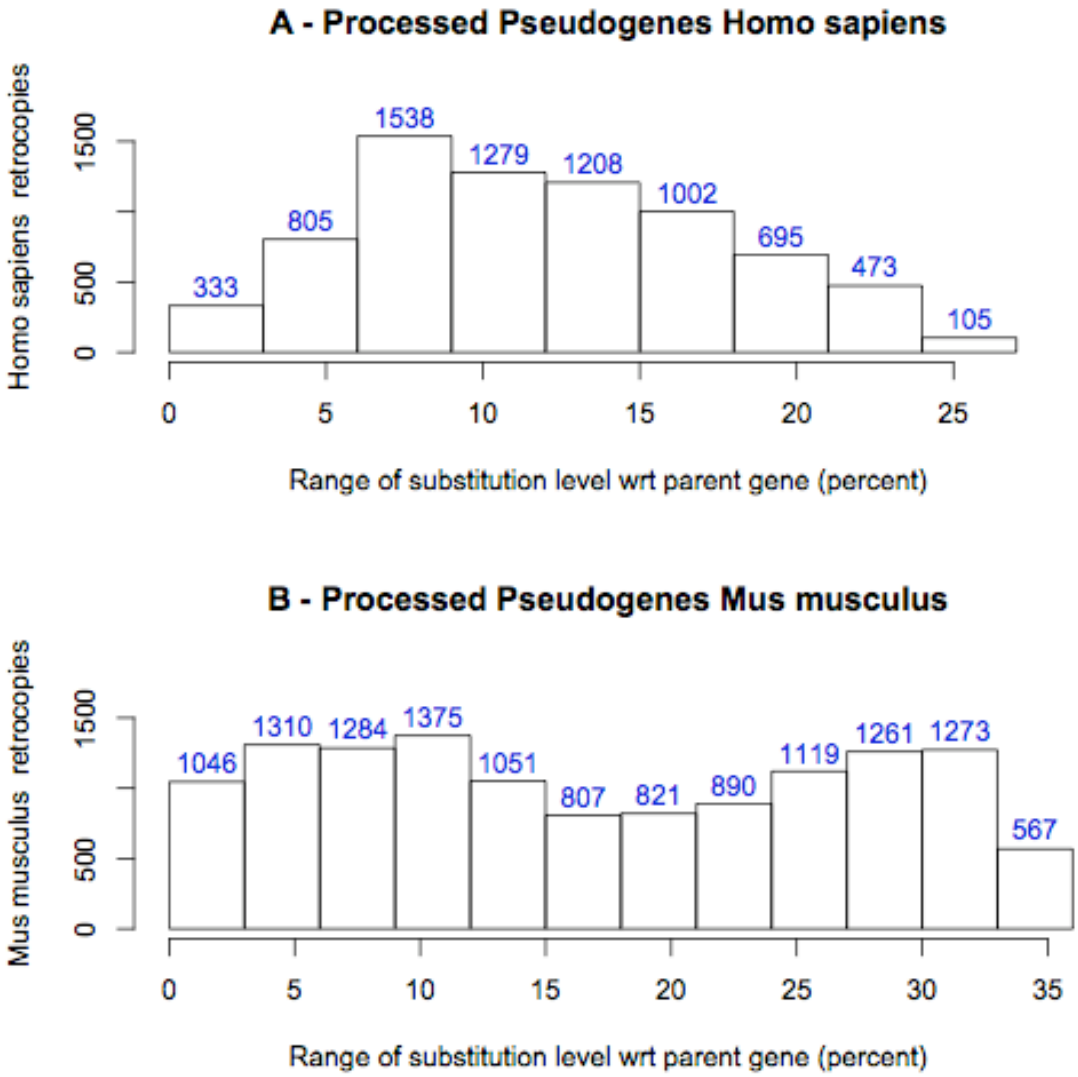


Figure 4. Age distribution of processed pseudogenes in the (a) human and (b) mouse genomes. Retrocopies were sorted by their divergence from the parent gene (which approximates the sequence at the time of insertion). There is a different correspondence between substitution levels and time periods because of the faster substitution rate in mouse (Mouse Genome Consortium 2002).

Age Distribution

Comparing the age distribution of retrocopies in human and mouse genome provides a record of the background rate of insertion over time. The age of the retrocopy can be inferred by two methods: 1) divergence with respect to the parent gene (assuming the parent gene changes much more slowly than the retrocopy) and 2) by breaks in orthology that are present in one branch but not the other. Figure 4 shows the data grouped in bins based on the percent identity compared with the parent gene. We used substitution as a proxy for age with non-expressed pseudogenes since they are apparently not under selective pressure. In the mouse genome, the rate of retrotransposition is relatively unchanged over all four periods, where corresponds to rate of L1 retrotransposon activity reported earlier (Mouse genome consortium 2002). The sharp drop off in human in the last two 25 mya periods is expected given the corresponding drop in L1 activity in recent primate evolution (see discussion for explanation). We could not use this measure for expressed retrogene since the effect of negative selection biases the result. Table 4 shows the age distribution or expressed retrogene (according to the UCSC known genes track) using the orthlogy method for dating retrocopies that annotated as protein-coding genes. Despite the drop in L1 activity shown in Figure 4A, there is no corresponding drop in exaptation of retrocopies in human, instead we see a doubling of the rate of retrogene birth if you normalize using the rate of processed pseudogene insertion. This is surprising

because one would expect to see a corresponding 4 fold drop in the recent retrogenes to match the drop in the rate of processed pseudogene formation.

	Rate of Retrogene formation (Ratio of Retrogenes to Processed Pseudogenes)		
Time period	Human		Mouse
At least since Mouse	0.033	At Least since Human	0.037
At least since Rhesus	0.068	At Least since Rat	0.021
At least since Chimp	0.072	<na>	

Table 4 - Age distribution of expressed retrocopies and annotated retrogenes in the human and mouse genomes. Retrogene time period assigned by orthology based on presence or absence from the outgroup species shown. Retrogenes were defined as retrocopies that overlapped at least 30% with the UCSC genes track with category "coding" in the UCSC Genome Browser.

	Human	Mouse
Expressed Retrogenes	422	446
Total mRNAs sequenced	287,000	230,000
Total ESTs sequenced	800,000	400,000

Table 5 – Summary of mRNA and EST expression evidence for human and mouse in the UCSC Genome Browser

Discussion

There has been a decrease in retrotransposition activity in recent primate evolution due to multiple copies of APOBEC3B and APOBEC3F (Stenglein and Harris 2006). If we compare the number of retrocopies since the split with macaque (2,884) about 25 MYA with the number inserted since chimp, about 6 MYA, we would expect to see about 600-700 retrocopies since chimp. We observed 974 (Table 3 row 6), which is not surprising. On the other hand, the 305 expressed retrocopies since chimp are

about half of the number found since macaque (690). Looking at annotated genes, the contrast is even more striking, 132 retrocopies since chimp compared with 165 since macaque. Despite the suppression of L1 mediated retrotransposition by APOBEC3 (Figure 4A), more retroposed transcripts are still being maintained than we would expect by just looking at the rate of insertion (Figure 5A,C). In mouse, there has been no decrease in retroposition (Figure 4B), rather an increase shown by 9,598 retrocopies since rat versus 6,161 since the human split. However, unlike in human, the ratio of the transcribed retrocopies to all retrocopies, shows no enrichment (Table 4, rows 3 and 4). If APOBEC3 was evolution's response to HIV and other viral infections, could the benefit of gene duplication via retrotransposition still be operating? If APOBEC evolved to fight off the hiv virus, other types of endogenous retroviruses (eg retrogenes) could also be deleterious to fitness in primates.

Effect of Population size

Since mouse has a much larger effective population size than humans, you would expect that if a retrocopy is deleterious, so any deleterious effect would be selected against more strongly. Since the population size of mouse is constant over time (or possibly slightly increasing) we would not expect a dramatic change in the rate of retrogene fixation over time (Table 4, column 4). There is a slight drop in mouse retrogene formation. In order to

be sure that the rate in humans is not faster due to a smaller population size, we looked at three time points. The rate of retrogene formation (computed as a ratio of count of retrogenes to count of processed pseudogenes) more than doubled at least since the divergence from Rhesus macaque (Table 4, column 2). This could be due to a population bottleneck in the apes or some other selective advantage to retrogene formation.

Further bias are possible due to the large numbers of mRNAs and ESTs sequence in human (table 5), however, since there are roughly 10 mRNAs and 40 ESTs per mouse gene, the likelihood of errors larger than those identified above are small.

Conclusion and Future Work

The retroFinder pipeline provides a very accurate method for classifying processed pseudogenes and expressed retrocopies. It can reliably distinguish between the two other mechanisms of gene duplication, namely tandem gene duplication and large segmental duplication. It also provides an automatic method for finding chimeric retrogenes that other methods do not find. By using DNA alignment methods we are able to increase sensitivity from traditional protein based methods. We also generate long predictions by using UTR as well as the coding sequence. Since UTR is a

functional portion of genes, it is important to include UTR in pseudogene as well as gene annotations.

Application to gene prediction

This pipeline has been used to distinguish genes from pseudogenes in three separate projects: CCDS (Pruitt et al., 2009), ENCODE (Encode Project Consortium 2007) and Gencode (Harrow et al., 2006). By incorporating the output as additional features to a genefinder, Augustus, it has been shown to improve genefinding performance (Stanke et al. 2008).

Studying unique evolutionary events

Numerous papers calculate gene birth and gene death by calculating rates of tandem array gene duplication (TAG), segmental gene duplication and retrotransposition. These papers generally using single methods that are prone to misclassification. Better classification of types of gene duplication events will lead to improved estimates of gene birth rates and further our understanding of evolution as well as possible disease mechanisms.

4. Wet Lab Experiments to confirm novel human-specific retrogenes

Computational screening

I screened the set of human retrocopies (from Chapter 1) for cases that did not exist in chimp, rhesus, or mouse. This was done using the orthology feature identified in algorithm. Any retrocopy that had a break orthology with respect to all four species was a candidate. Then I looked for cases there had at least 3 mRNAs from different tissues or cell lines as evidence of expression. Out of the eight candidates cases found, RAP1B-h was selected for experiments because of its compelling function. RAP1B (the parent gene) is expressed in the brain and is involved in regulating neuronal polarity in mouse (Li and Werner 2008). The hypothesis was that humans have an extra copy of this gene RAP1B-h that was somehow related to human brain development. CDNAs have been sequenced in brain and testes so my goal was to replicate expression independently and verify that humans indeed have two copies of this gene: one with introns and an intronless retrogene.

Experimental setup

PCR Strategy

We wanted to see if the 8 human specific genes were polymorphic in the population so I designed primers two sets of primers for all 8 genes to look for presence or absence of the genomic DNA for the gene. If the retrogene is present, primer pair B+C would show a band of 445 bp and also A+C would show a band of 1245 bp. If the retrogene is missing, then primer pair A+C would show a band at 245 bp (See Figure 4-1).

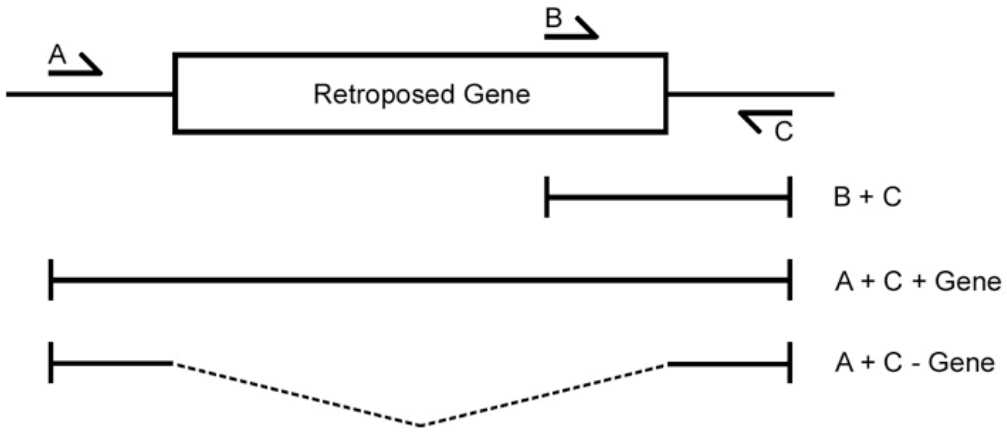


Figure 4-1 Primer design for looking for presence / absence of human specific retrogenes. 20-mer primers where designed to hybridize about 200 bp upstream and downstream from the 5' and 3' ends of the gene.

RT-PCR Strategy

I designed primer to prime at the point of mismatch between the retro RAP1B-h RNA and the parent RNA RAP1B. I looked for expression in placenta and Ramos B cell line since they were previously sequenced and deposited in genbank. I also looked in testes since many genes are expressed in testes. I repeated the experiment several times in each tissue with different primers.

Results

Bryan King ran the first set of experiments and found that all eight human specific genes were present in all 24 individuals. One of these, hx.7 is shown in Figure 4-2.

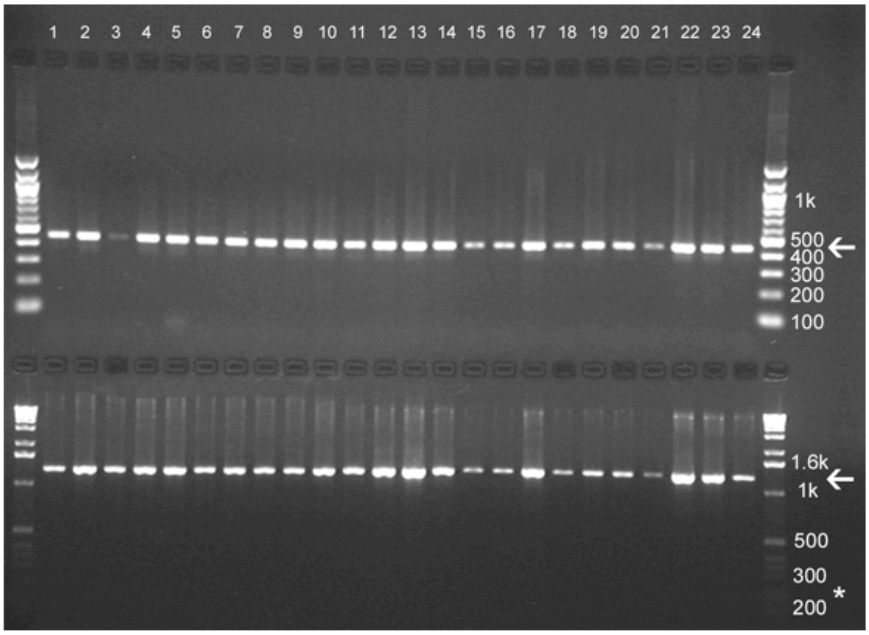


Figure 4-2 hx.7 LAMR1h2 on the NHGRI Human Diversity Panel. Top panel: Primer pair B+C, arrow shows length of product from retrogene (445 bp). Bottom panel: Primer pair A+C, arrow shows length of product from parent gene (1245 bp). Asterix marks expected product length of 245 bp A+C, if gene is absent.

We were able to not able to get any RNA expression from placenta or b ce ll lines or testes.

Discussion

We noticed that all 5 genbank entries for cDNAs for this gene were deposited by the same lab, Invitrogen. We were not able to contact the original researcher as he had

left the company and no one was responsible for maintaining the data. Some of the entries associated with this dataset appear to have been aligned to the reference genome and the sequences subsequently modified to match the genome. Subsequently all cDNA entries from this dataset have been marked with a warning in the UCSC Genome Browser. A recent paper came out looking at two retrogenes from RAP1A in mouse. One of these, mRap1A-retro1, solely stimulates cell spreading but not adhesion, points to a unique property of this isoform suggesting it may be undergoing functionalization (Zemojtel et al., 2010). More work is needed to determine the role of these genes but clearly latent retrogenes, could cause powerful phenotypes if activated by mutations or other assaults on the cell.

5. Conclusion

In this dissertation I have developed a pipeline for automatically recognizing retrocopies that can be run on any sequenced genome that is alignable to mammals.The retroFinder pipeline provides a very accurate method for classifying processed pseudogenes and expressed retrocopies. It can reliably distinguish between the two other mechanisms of gene duplication, namely tandem gene duplication and large segmental duplication. It also provides an automatic method for finding chimeric retrogenes that other methods do not find. By using DNA alignment methods we are able to increase sensitivity from traditional protein based methods. Unlike other pseudogene annotations, I also generate long predictions by using UTR as well as the coding sequence. Since UTR is a functional portion of genes, it is important to include UTR in pseudogene as well as gene annotations.

This pipeline has been used to distinguish genes from pseudogenes in three separate projects: CCDS (Pruitt et al., 2009), ENCODE (Encode Project Consortium 2007) and Gencode (Harrow et al., 2006). By incorporating the output as additional features to a genefinder, Augustus, it has been shown to improve genefinding performance (Stanke et al. 2008).

Numerous papers calculate gene birth and gene death by calculating rates of tandem array gene duplication (TAG), segmental gene duplication and retrotransposition. These papers generally using single methods that are prone to

misclassification. Better classification of types of gene duplication events will lead to improved estimates of gene birth rates and further our understanding of evolution as well as possible disease mechanisms.

6. References

Flores-Rozias H & Kolodner RD. Links between replication, recombination and genetic stability in eukaryotes. Trends Biochem. Sci. (2000) 25, 169-200.

McClintock B. The origin and behavior of mutable loci in maize. Proc Natl Acad Sci U S A. (1950) Jun;36(6):344-55.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al.: Initial sequencing and analysis of the human genome. Nature 2001, 409:860-921.

Wicker T, Sabot F, Hua-Van A, et al. (December 2007). "A unified classification system for eukaryotic transposable elements". Nat. Rev. Genet. 8 (12): 973–82.

Vanin EF (1985). "Processed pseudogenes: characteristics and evolution". *Annu. Rev. Genet.* **19**: 253–72

Zhang Z, Harrison PM, Liu Y, Gerstein M: Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. Genome research 2003, 13:2541-2558.

Sakai H, Koyanagi KO, Imanishi T, Itoh T, Gojobori T: Frequent emergence and functional resurrection of processed pseudogenes in the human and mouse genomes. Gene 2007, 389:196-203.

Khelifi A, Duret L, Mouchiroud D: HOPPSIGEN: a database of human and mouse processed pseudogenes. Nucleic Acids Res 2005, 33:D59-66.

Zheng D, Gerstein MB: The ambiguous boundary between genes and pseudogenes: the dead rise up, or do they? Trends Genet 2007, 23:219-224.

Harrison PM, Zheng D, Zhang Z, Carriero N, Gerstein M: Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability. Nucleic acids research 2005, 33:2374-2383.

Vinckenbosch N, Dupanloup I, Kaessmann H: Evolutionary fate of retroposed gene copies in the human genome. Proceedings of the National Academy of Sciences of the United States of America 2006, 103:3220-3225

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

Li YH, Werner H, Püschel AW. Rheb and mTOR regulate neuronal polarity through Rap1B. J Biol Chem. 2008 Nov 28;283(48):33784-92. Epub 2008 Oct 8.

Zemojtel T, Duchniewicz M, Zhang Z, Paluch T, Luz H, Penzkofer T, Scheele JS, Zwartkruis FJ. Retrotransposition and mutation events yield Rap1 GTPases with differential signalling capacity. BMC Evol Biol. 2010 Feb 19;10:55.

Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, Kent WJ, Haussler D: A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature 2006, 441:87-90.

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D: Ultraconserved elements in the human genome. Science (New York, NY) 2004, 304:1321-1325.

Lowe CB, Bejerano G, Haussler D: Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. Proceedings of the National Academy of Sciences of the United States of America 2007, 104:8005-8010.
Nishihara H, Smit AF, Okada N: Functional noncoding sequences derived from SINEs in the mammalian genome. Genome research 2006, 16:864-874.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al.: Initial sequencing and analysis of the human genome. Nature 2001, 409:860-921.

Brosius J: RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. Gene 1999, 238:115-134.

Zhang XH, Chasin LA: Comparison of multiple vertebrate genomes reveals the birth and evolution of human exons. Proceedings of the National Academy of Sciences of the United States of America 2006, 103:13427-13432.

Sorek R, Ast G, Graur D: Alu-containing exons are alternatively spliced. Genome research 2002, 12:1060-1067.

Krull M, Brosius J, Schmitz J: Alu-SINE exonization: en route to protein-coding function. Molecular biology and evolution 2005, 22:1702-1711.

Krull M, Petrusma M, Makalowski W, Brosius J, Schmitz J: Functional persistence of exonized mammalian-wide interspersed repeat elements (MIRs). Genome research 2007, 17:1139-1145.

Sela N, Mersch B, Gal-Mark N, Lev-Maor G, Hotz-Wagenblatt A, Ast G: Comparative analysis of transposed element insertion within human and mouse genomes reveals Alu's unique role in shaping the human transcriptome. Genome Biol 2007, 8:R127.

Gotea V, Makalowski W: Do transposable elements really contribute to proteomes? Trends Genet 2006, 22:260-267.

Zheng D, Gerstein MB: The ambiguous boundary between genes and pseudogenes: the dead rise up, or do they? Trends Genet 2007, 23:219-224.

Suyama M, Harrington E, Bork P, Torrents D: Identification and analysis of genes and pseudogenes within duplicated regions in the human and mouse genomes. PLoS computational biology 2006, 2:e76.

Sakai H, Koyanagi KO, Imanishi T, Itoh T, Gojobori T: Frequent emergence and functional resurrection of processed pseudogenes in the human and mouse genomes. Gene 2007, 389:196-203.

Gray TA, Wilson A, Fortin PJ, Nicholls RD: The putatively functional Mkrn1-p1 pseudogene is neither expressed nor imprinted, nor does it regulate its source gene in trans. Proceedings of the National Academy of Sciences of the United States of America 2006, 103:12039-12044.

Balakirev ES, Ayala FJ: Pseudogenes: are they "junk" or functional DNA? Annual review of genetics 2003, 37:123-151.

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

Dahl HH, Brown RM, Hutchison WM, Maragos C, Brown GK: A testis-specific form of the human pyruvate dehydrogenase E1 alpha subunit is coded for by an intronless gene on chromosome 4. Genomics 1990, 8:225-232.

Boer PH, Adra CN, Lau YF, McBurney MW: The testis-specific phosphoglycerate kinase gene pgk-2 is a recruited retroposon. Molecular and cellular biology 1987, 7:3107-3112.

Long M, Betran E, Thornton K, Wang W: The origin of new genes: glimpses from the young and old. Nature reviews 2003, 4:865-875.

Brosius J: Retroposons – seeds of evolution. Science 1991, 251:753.

Brosius J, Gould SJ: On "genomenclature": a comprehensive (and respectful) taxonomy for pseudogenes and other "junk DNA". Proceedings of the National Academy of Sciences of the United States of America 1992, 89:10706-10710.

Sorek R, Shamir R, Ast G: How prevalent is functional alternative splicing in the human genome? Trends Genet 2004, 20:68-71.
Wang W, Zheng H, Fan C, Li J, Shi J, Cai Z, Zhang G, Liu D, Zhang J, Vang S, et al.: High rate of chimeric gene origination by retroposition in plant genomes. The Plant cell 2006, 18:1791-1802.

Wang W, Brunet FG, Nevo E, Long M: Origin of sphinx, a young chimeric RNA gene in Drosophila melanogaster. Proceedings of the National Academy of Sciences of the United States of America 2002, 99:4448-4453.

Long M, Langley CH: Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila. Science (New York, NY) 1993, 260:91-95.

Sasidharan R, Gerstein M: Genomics: protein fossils live on as RNA. Nature 2008, 453:729-731.

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al.: Initial sequencing and comparative analysis of the mouse genome. Nature 2002, 420:520-562.

Chothia C, Gough J, Vogel C, Teichmann SA: Evolution of the protein repertoire. Science (New York, NY) 2003, 300:1701-1703.

Grassé PP: Evolution of living organisms. New York: Academic Press; 1977.

Ohno S: Birth of a unique enzyme from an alternative reading frame of the preexisted, internally repetitious coding sequence. Proceedings of the National Academy of Sciences of the United States of America 1984, 81:2421-2425.

Keese PK, Gibbs A: Origins of genes: "big bang" or continuous creation? Proceedings of the National Academy of Sciences of the United States of America 1992, 89:9489-9493.

Liang H, Landweber LF: A genome-wide study of dual coding regions in human alternatively spliced genes. Genome research 2006, 16:190-196.

Gilbert W: Why genes in pieces? Nature 1978, 271:501.

Alekseyenko AV, Kim N, Lee CJ: Global analysis of exon creation versus loss and the role of alternative splicing in 17 vertebrate genomes. RNA (New York, NY) 2007, 13:661-670.

Buttice G, Kaytes P, D'Armiento J, Vogeli G, Kurkinen M: Evolution of collagen IV genes from a 54-base pair exon: a role for introns in gene evolution. Journal of molecular evolution 1990, 30:479-488.

Golding GB, Tsao N, Pearlman RE: Evidence for intron capture: an unusual path for the evolution of proteins. Proceedings of the National Academy of Sciences of the United States of America 1994, 91:7506-7509.

Hayashi Y, Sakata H, Makino Y, Urabe I, Yomo T: Can an arbitrary sequence evolve towards acquiring a biological function? Journal of molecular evolution 2003, 56:162-168.

Sakharkar MK, Chow VT, Ghosh K, Chaturvedi I, Lee PC, Bagavathi SP, Shapshak P, Subbiah S, Kangueane P: Computational prediction of SEG (single exon gene) function in humans. Front Biosci 2005, 10:1382-1395.

Patthy L: Exons – original building blocks of proteins? Bioessays 1991, 13:187-192.

Dorit RL, Schoenbach L, Gilbert W: How big is the universe of exons? Science (New York, NY) 1990, 250:1377-1382.

Tress ML, Martelli PL, Frankish A, Reeves GA, Wesselink JJ, Yeats C, Olason PL, Albrecht M, Hegyi H, Giorgetti A, et al.: The implications of alternative splicing in the ENCODE protein complement. Proc Natl Acad Sci USA 2007, 104:5495-5500.

Harrison PM, Zheng D, Zhang Z, Carriero N, Gerstein M: Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability. Nucleic acids research 2005, 33:2374-2383.

Zhang Z, Harrison PM, Liu Y, Gerstein M: Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. Genome research 2003, 13:2541-2558.

Khelifi A, Duret L, Mouchiroud D: HOPPSIGEN: a database of human and mouse processed pseudogenes. Nucleic Acids Res 2005, 33:D59-66.

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

Dupuy D, Duperat VG, Arveiler B: SCAN domain-containing 2 gene (SCAND2) is a novel nuclear protein derived from the zinc finger family by exon shuffling. Gene 2002, 289:1-6.

Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ: Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. Proceedings of the National Academy of Sciences of the United States of America 2006, 103:9935-9939. |

Begun DJ, Lindfors HA, Kern AD, Jones CD: Evidence for de novo evolution of testis-expressed genes in the Drosophila yakuba/Drosophila erecta clade. Genetics 2007, 176:1131-1137.
Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, Zhan Z, Li X, Ding Y, Yang S, Wang W: On the origin of new genes in Drosophila. Genome Res 2008, 18:1446-1455.

Shen X, Mizuguchi G, Hamiche A, Wu C: A chromatin remodelling complex involved in transcription and DNA processing. Nature 2000, 406:541-544.

Weiner J 3rd, Beaussart F, Bornberg-Bauer E: Domain deletions and substitutions in the modular protein evolution. The FEBS journal 2006, 273:2037-2047.

Krasnov AN, Kurshakova MM, Ramensky VE, Mardanov PV, Nabirochkina EN, Georgieva SG: A retrocopy of a gene can functionally displace the source gene in evolution. Nucleic acids research 2005, 33:6654-6661. | |

Cordaux R, Udit S, Batzer MA, Feschotte C: Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. Proceedings of the National Academy of Sciences of the United States of America 2006, 103:8101-8106. | |

Chen YT, Iseli C, Venditti CA, Old LJ, Simpson AJ, Jongeneel CV: Identification of a new cancer/testis gene family, CT47, among expressed multicopy genes on the human X chromosome. Genes, chromosomes & cancer 2006, 45:392-400. |

Pheasant M, Mattick JS: Raising the estimate of functional human sequences. Genome research 2007, 17:1245-1253. |

Kalamegham R, Sturgill D, Siegfried E, Oliver B: Drosophila mojoless, a retroposed GSK-3, has functionally diverged to acquire an essential role in male fertility. Molecular biology and evolution 2007, 24:732-742.

Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W: Human-mouse alignments with BLASTZ. Genome Res 2003, 13:103-107. | |

Ashurst JL, Chen CK, Gilbert JG, Jekosch K, Keenan S, Meidl P, Searle SM, Stalker J, Storey R, Trevanion S, et al.: The Vertebrate Genome Annotation (Vega) database. Nucleic Acids Res 2005, 33:D459-465. | |

Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D: Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad Sci USA 2003, 100:11484-11489. | |

Vanin EF: Processed pseudogenes. Characteristics and evolution. Biochimica et biophysica acta 1984, 782:231-241. |

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

Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, et al.: The UCSC Genome Browser Database: update 2006. Nucleic Acids Res 2006, 34:D590-598. | |

Furey TS, Diekhans M, Lu Y, Graves TA, Oddy L, Randall-Maher J, Hillier LW, Wilson RK, Haussler D: Analysis of human mRNAs with the reference genome sequence reveals potential errors, polymorphisms, and RNA editing. Genome research 2004, 14:2034-2040.

Smit A, Hubley R, Green P: RepeatMasker Open-3.0. [http://www.repeatmasker.org] webcite 2004.

Benson G: Tandem repeats finder: a program to analyze DNA sequences. Nucleic acids research 1999, 27:573-580.

Kohany O, Gentles AJ, Hankus L, Jurka J: Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics 2006, 7:474.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: Repbase Update, a database of eukaryotic repetitive elements. Cytogenetic and genome research 2005, 110:462-467.

Bailey JA, Baertsch R, Kent WJ, Haussler D, Eichler EE: Hotspots of mammalian chromosomal evolution. Genome biology 2004, 5:R23.

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE: Recent segmental duplications in the human genome. Science 2002, 297:1003-1007.

Solovyev V: Structure, Properties and Computer Identification of Eukaryotic Genes. In Bioinformatics – From Genomes to Drugs Edited by: Lengauer PDT. 1994.

Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D: The UCSC Known Genes. Bioinformatics (Oxford, England) 2006, 22:1036-1046.

Pruitt KD, Tatusova T, Maglott DR: NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic acids research 2005, 33:D501-504.

Vanin EF: Processed pseudogenes. Characteristics and evolution. Biochimica et biophysica acta 1984, 782:231-241.

Ohno S: Birth of a unique enzyme from an alternative reading frame of the preexisted, internally repetitious coding sequence. Proceedings of the National Academy of Sciences of the United States of America 1984, 81:2421-2425.

Balakirev ES, Ayala FJ: Protogenes PSEUDOGENES: Are They “Junk” or Functional DNA? Annual Review of Genetics 2003 37, 123-151

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

Vinckenbosch N, Dupanloup I, Kaessmann H: Evolutionary fate of retroposed gene copies in the human genome. Proceedings of the National Academy of Sciences of the United States of America 2006, 103:3220-3225.

Pan D, Zhang L, Quantifying the major mechanisms of recent gene duplications in the human and mouse genomes: a novel strategy to estimate gene duplication rates Genome Biology 2007, 8:R158doi:10.1186/gb-2007-8-8-r158

Hahn, Matthew W., Demuth, Jeffery P., Han, Sang-Gook. Accelerated Rate of Gene Gain and Loss in Primates. Genetics 2007 177: 1941-1949

Han MV, Demuth JP, McGrath CL, Casola C, Hahn MW. Adaptive evolution of young gene duplicates in mammals. Genome Res. 2009 May;19(5):859-67.

Chimpanzee Sequencing and Analysis Consortium (2005). Initial sequence of the chimpanzee genome and comparison with the human genome . Nature 437: 69–87.

Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 2008 Mar 1;24(5):637-44. Epub 2008 Jan 24.

Bailey JA, Baertsch R, Kent WJ, Haussler D, Eichler EE: Hotspots of mammalian chromosomal evolution. Genome biology 2004, 5:R23.

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE: Recent segmental duplications in the human genome. Science 2002, 297:1003-1007.

Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W: Human-mouse alignments with BLASTZ. Genome Res 2003, 13:103-107.

Zhang, Z., Harrison, P., Gerstein, M. (2002) Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome *Genome Res.*, **12**, 1466–14482 Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, Lagarde J, Gilbert JG, Storey R, Swarbreck D, Rossier C, Ucla C, Hubbard T, Antonarakis SE, Guigo R. GENCODE: producing a reference annotation for ENCODE. Genome Biol. 2006;7 Suppl 1:S4.1-9. Epub 2006 Aug 7.

Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, Searle S, Farrell CM, Loveland JE, Ruef BJ, Hart E, Suner MM, Landrum MJ, Aken B, Ayling S, Baertsch R, Fernandez-Banet J, Cherry JL, Curwen V, Dicuccio M, Kellis M, Lee J, Lin MF, Schuster M, Shkeda A, Amid C, Brown G, Dukhanina O, Frankish A, Hart J, Maidak BL, Mudge J, Murphy MR, Murphy T, Rajan J, Rajput B, Riddick LD, Snow C, Steward C, Webb D, Weber JA, Wilming L, Wu W, Birney E, Haussler D, Hubbard T, Ostell J, Durbin R, Lipman D. The consensus coding

sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. Genome Res. 2009 Jul;19(7):1316-23. Epub 2009 Jun 4.

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

ENCODE Project Consortium Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007 Jun 14;447(7146):799-816.

Zhang Z, Harrison PM, Liu Y, Gerstein M (2003) Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. Genome Res 13: 2541–58.

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

Kent WJ, Baertsch R , Hinrichs A, Miller W and Haussler D. Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes PNAS, September 30, 2003, vol. 100,no. 20 pp 11484-11489

Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, et al.: The UCSC Genome Browser Database: update 2006. Nucleic Acids Res 2006, 34:D590-598.

Furey TS, Diekhans M, Lu Y, Graves TA, Oddy L, Randall-Maher J, Hillier LW, Wilson RK, Haussler D: Analysis of human mRNAs with the reference genome sequence reveals potential errors, polymorphisms, and RNA editing. Genome research 2004, 14:2034-2040.

Smit A, Hubley R, Green P: RepeatMasker Open-3.0. [http://www.repeatmasker.org]

Benson G: Tandem repeats finder: a program to analyze DNA sequences. Nucleic acids research 1999, 27:573-580.

D Zheng, A Frankish, R Baertsch, P Kapranov, A Reymond, SW Choo, Y Lu, F Denoeud, S Antonarakis, M Snyder, Y Ruan, C Wei, T Gingeras, R Guigó, J Harrow, and MB Gerstein. Pseudogenes in the ENCODE regions: Consensus annotation, analysis of transcription, and evolution. Genome Res. 2007 June; 17(6): 839–851.

Stenglein MD, Harris RS. APOBEC3B and APOBEC3F Inhibit L1 Retrotransposition by a DNA Deamination-independent Mechanism. J. Biol. Chem. 2006 281: 16837-16841.