

# Cataloging Animal Retrocopies: An Annotation-Independent Methodology

**Shilpa Nadimpalli**

Depts. of Biology & Computer Science  
Tufts University  
[shilpa.nadimpalli@tufts.edu](mailto:shilpa.nadimpalli@tufts.edu)

**Liqing Zhang\***

Dept. of Computer Science  
Virginia Tech  
[lqzhang@vt.edu](mailto:lqzhang@vt.edu)

---

## ABSTRACT

Retrocopies have been shown to play a major role in the origin of novel genes in animal genomes. While their impact on the evolution of distinct genomes has been studied, comparisons of retrocopy occurrence, age, and distribution between genomes have yet to be completed large-scale. In order to assess the impact of retrocopies on inter-species events or generalize their trends in animal genomes on the whole, a standard methodology for identifying retrocopies needs to be established. Because previous methodologies are fine-tuned for specific genomes and rely heavily on existing annotation information, extensive and sensible retrocopy comparison between species is difficult. We present here an annotation-independent methodology for identifying retrogenes, retropseudogenes, and chimeric genes in any animal genome, and summarize some major difficulties in identifying retrocopies in a large-scale study.

---

**Keywords:** *retrotransposable element, retrocopy, chimeric gene, genome annotation, animal genome*

## INTRODUCTION

Retrotransposition, or retroduplication, is a form of gene duplication that utilizes an RNA intermediate. The retrotransposed daughter copies, known as “retrocopies,” lack the genetic features of the parental gene, such as introns and regulatory elements, because they originate from spliced gene transcripts [1,4,5,9,11,13,19]. In addition to a lack of introns, other recognizable traits of retrocopies include the presence of a poly(A)-tail and flanking direct repeats [4,7,9], which can be used to identify retrocopies in genomes. Figure 1 depicts various types of gene duplications, and differentiates retrocopies from other gene duplicates.

Like other gene duplication methods, one significant effect of retrotransposition can be the development of novel genes. While other methods mostly result in new, duplicate genes that are silenced through degenerative

mutations [10], retrotransposition can be largely responsible for the development of original, potentially beneficial genes through a process known as “gene fusion,” or the forming of “chimeric genes” [1,2,9,11,17].

The mRNA transcripts that are reverse transcribed and integrated back into the genome to form retrocopies are often much smaller in length than their parent genes because they lack introns. Subsequently, a retrocopy can easily “land” in a different, existing gene, and be transcribed as another exon or as part of an existing exon, resulting in a chimeric gene [17]. The proteins resulting from this gene can have a vastly different structure and function, and result in an immediately beneficial trait, in contrast to a beneficial trait arising over time after an accumulation of mutations in tandem duplicates.

In addition to providing the raw material for new genes, retrocopies can also provide insight into a gene’s history. As previously mentioned, retrocopies lack introns. This allows us to determine the direction of gene movement. Studies have shown that in certain mammalian genomes, retrocopies are more abundant in the X chromosome than in autosomes, and many of the retrocopies found on autosomes have parental genes on the X chromosome [5,7,9,16]. Such observations based on the movement of retrocopies might model the movement of other genes, too. In addition, some retrocopies may provide evidence for ancient transcripts

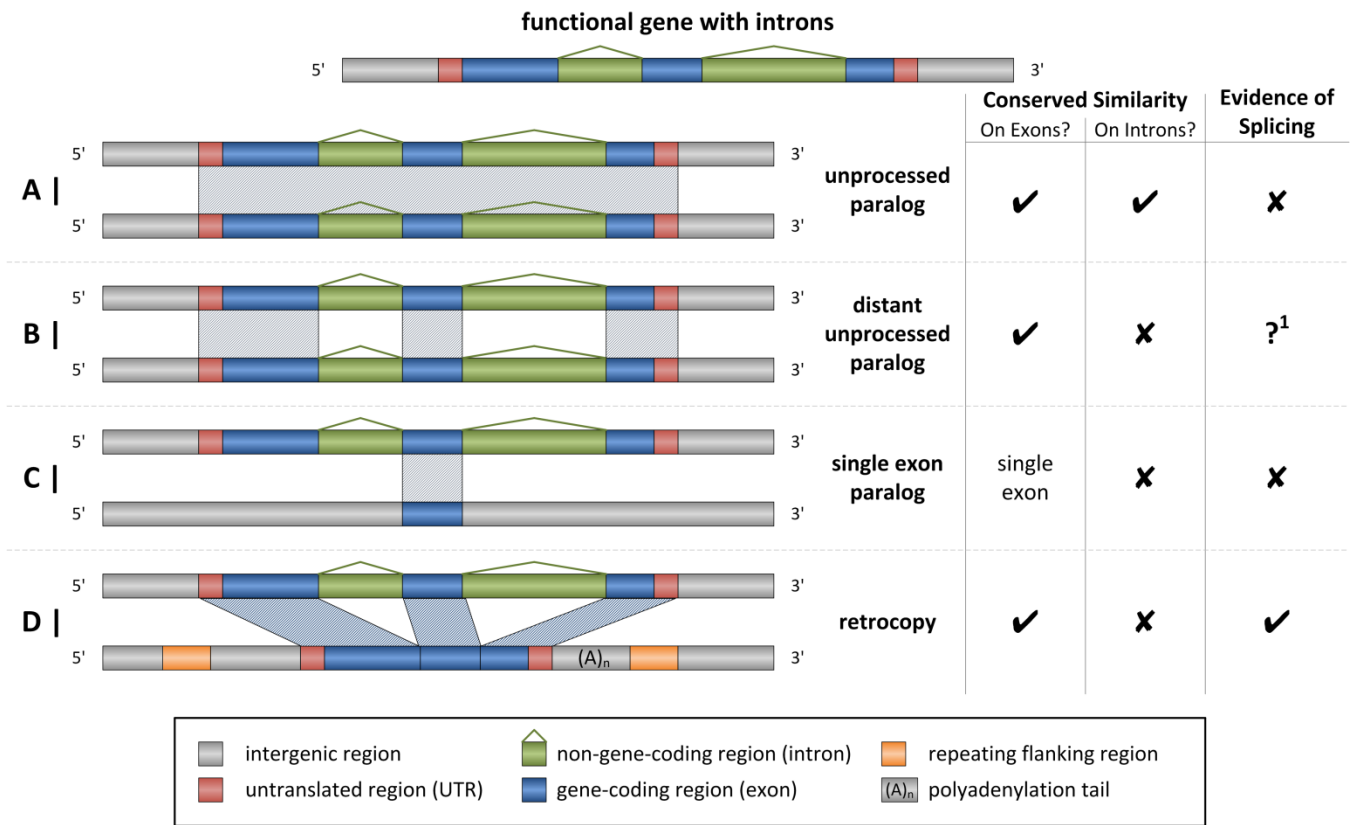
---

\* Corresponding author.

Users are entitled to use, reproduce, and display this article for non-commercial purposes provided that the original authorship is fully and properly attributed with correct citation details given.

Article and publication are at <https://bioinformatics.cs.vt.edu/~shilpa/DREUfinalpaper.pdf>

© 2009, Shilpa Nadimpalli



**Figure 1 |** Types of Duplicated Genes. **(A)** Recent tandem or segmental gene duplication, resulting in a paralogous gene. Similarity is detected on all introns and exons. **(B)** Old but functional tandem or segmental gene duplication. Similarity is detected only on exons; introns have diverged too much to be recognized as homologs. <sup>1</sup>Certain recent studies [8] describe how it is possible for genes or gene segments, including retrocopies, to gain introns, even after a supposed splicing event. **(C)** Old or partial, mostly nonfunctional tandem or segmental gene duplication or retrotransposition. Sequence similarity is only detected on a single exon. **(D)** Retrocopy resulting from reverse transcription of a complete transcript. Sequence similarity is detected on only exons and no introns. (Adapted from Adel et al. [1])

which are no longer transcribed or functional. Age determination of such retrocopies based on flanking repeats or poly(A)-tail degradation could provide evidence for speciation events or major evolutionary advancements in a species.

Because retrocopies have been shown to influence evolution within certain species, a next logical step would be to determine their impact on groups of species as a whole or on the separation of closely-related species. Some recent work has examined trends in retrocopy behavior in certain animal groups [7], yet this research has relied on existing retrocopy data from previous studies, which are largely inconsistent.

The number of retrocopies found in a given genome can vary greatly depending on the identification criteria used by the researchers. For example, the number of retroseudogenes in the human genome can range from 3,664 to 17,759, as reported by Drouin [7]. Using such differing reports to draw conclusions about a species or group of species can provide erroneous information.

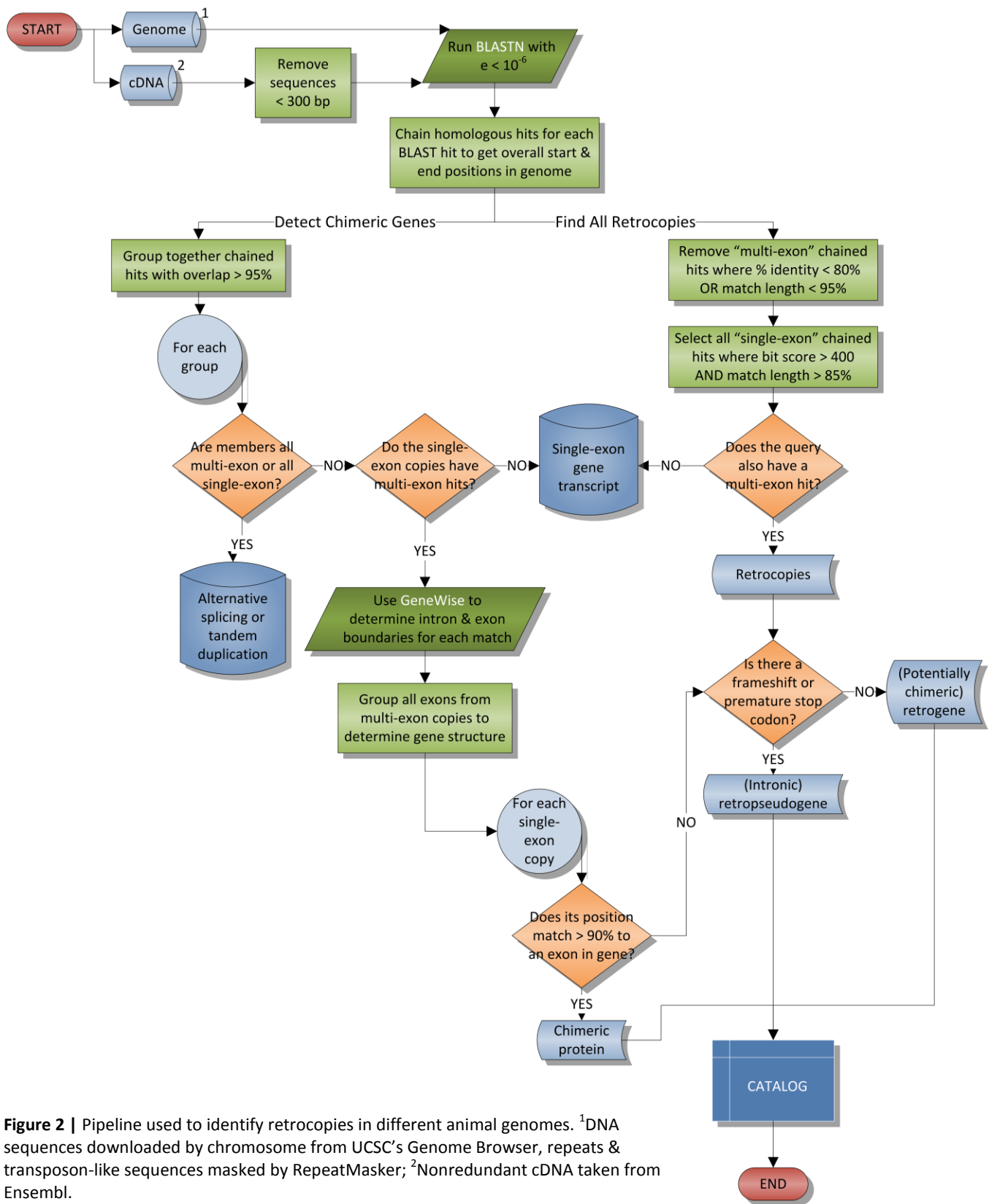
In addition, existing methodologies for identifying and analyzing retrocopies tend to move in two distinct directions: first, identifying all retrogenes or

retroseudogenes in a genome, or second, identifying solely chimeric genes resulting from retrotransposition. Valuable information may be drawn from identifying all three types of retrocopies, and comparing their locations, abundance, age, and other traits. This information can then be used to build a theoretical model of the retrotransposition process.

We present here a standard methodology for identifying retrogenes, retroseudogenes, and chimeric genes resulting from retrotransposition, to be used for analysis of all animal genomes. We will also discuss the challenges in building and using this “universal pipeline,” and suggest methods of improvement and extension.

## METHODOLOGY

In order to identify and classify retrocopies in animal genomes, many steps need to be taken. These steps are illustrated in a flowchart format in Figure 2. A brief summary of the methodology follows, to be expanded in detail later on. First, genome and transcriptome data must be acquired and filtered, then an all-against-all BLASTN will be run to determine gene locations in the



**Figure 2 |** Pipeline used to identify retrocopies in different animal genomes. <sup>1</sup>DNA sequences downloaded by chromosome from UCSC’s Genome Browser, repeats & transposon-like sequences masked by RepeatMasker; <sup>2</sup>Nonredundant cDNA taken from Ensembl.

genome. The major step of the pipeline is to then chain together these BLAST hits to come up with a rough annotation of the genome. Once BLAST hits have been

chained together, a selected set will be fed into GeneWise to determine intron-exon boundaries [3] to identify chimeric genes [17]. Then, each single-exon copy

with at least one corresponding multi-exon copy will be identified, filtered and classified as either retrogene or retroseudogene.

### I. Obtain DNA and cDNA Sequences

Complete genomic builds and corresponding cDNA transcriptomes can be downloaded for use from any site. Our genome builds (one file per chromosome) were obtained from UCSC's Genome Browser ([genome.ucsc.edu](http://genome.ucsc.edu)), where repeating sequences had already been masked by RepeatMasker ([www.repeatmasker.org](http://www.repeatmasker.org)). Sequence fragments that could not be placed on a specific chromosome or at a particular location within a chromosome were discarded.

Ideally, cDNA sequences would have been downloaded from the same database (UCSC's Genome Browser), but only mRNA sequences were available for download from this site. Consequently, non-redundant cDNA transcripts were instead downloaded from the Ensembl Genome Browser ([www.ensembl.org](http://www.ensembl.org)). All cDNA sequences shorter than 300 base-pairs are removed prior to running BLAST in accordance with previous methodologies [1 (80bp), 11 (~150bp), 13 (~150bp), 17 (300bp)].

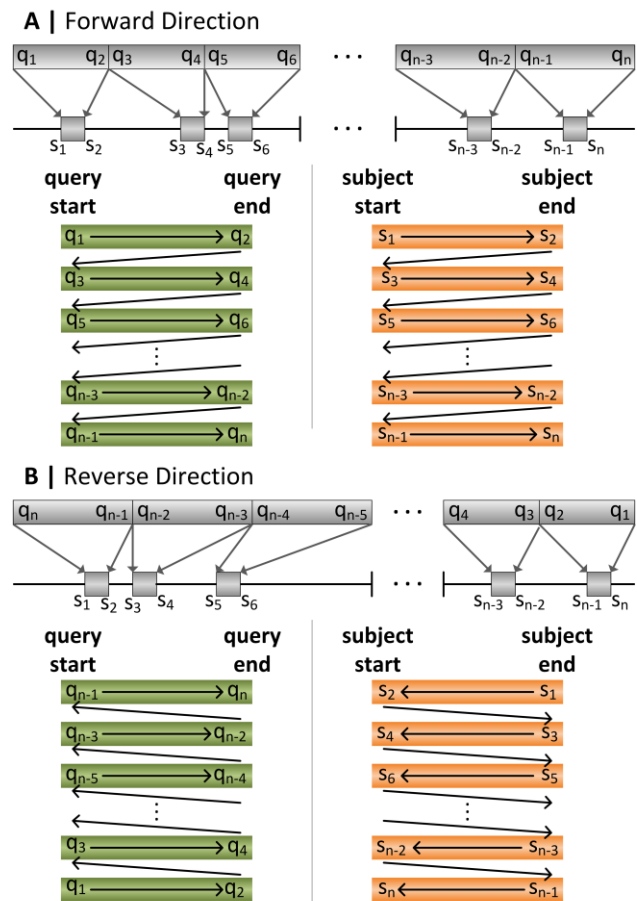
### II. BLAST & Chain Corresponding BLAST Hits

To compare nucleotide query sequences (cDNA) against a nucleotide database (DNA), we used the BLASTN program offered in the BLAST package, using an e-value cutoff of  $10^{-6}$ , approximating cutoff points from existing methodologies [1 ( $10^{-5}$ ), 16 ( $10^{-3}$ ), 17 ( $10^{-8}$ ), 19 ( $10^{-4}$ )]. However, because this process took exceedingly long on a single machine, we found that using mpiBLAST ([www.mpiblast.org](http://www.mpiblast.org)) was the most efficient way to produce desired output quickly. Using 20 nodes in a cluster, mpiBLAST took approximately 2 days to BLAST all human cDNAs against the human genome.

We caught output for the alignment positions on both the cDNA and the chromosome, and also for the following fields: % identity, alignment length, # mismatches, # gap openings, e-value, and bit score. For our purposes, formatting the BLAST output in a tab delineated format (using the "-m8" BLAST running option) allowed us to load the output into a MySQL database, where it could then be easily manipulated.

#### Part 1. Grouping BLAST Output

We will refer to the algorithm meant to appropriately link BLAST output as the "Gluing Algorithm." The original goal of the Gluing Algorithm was to output sections on the chromosome which could be used as input to GeneWise. However, the algorithm was later expanded to become a rough genome annotator to limit the pipeline's reliance on the GeneWise program, for various reasons discussed in Section III.



**Figure 3** | cDNA (q) alignment to a chromosome (s) in the forward and reverse directions. Values  $q_1 \leq q_2 \leq \dots \leq q_n$ , and  $s_1 \leq s_2 \leq \dots \leq s_n$ , where  $\leq$  means less than or within 3-5 base pairs at "conjunctions." Arrows mark the direction of increasing values. **(A)** cDNA aligning to the chromosome in frame +1, +2, or +3, and the corresponding organization of BLAST hits when ordered in ascending order by subject start. **(B)** cDNA alignment in frame -1, -2, or -3, and the corresponding BLAST hit organization.

The cDNAs align to the chromosome(s) in the genome in either a forward or a reverse direction. When ordering BLAST hits by their start position on the chromosome, the two directions need to be differentiated, as illustrated above in Figure 3. The direction of the match is determined as follows. A positive direction is implied if  $(\text{subject end} - \text{subject start}) > 0$  in a single BLAST hit. The negative direction is implied if instead  $(\text{subject end} - \text{subject start}) < 0$ . BLAST hits are combined, assuming they are in the same gene, according to the following criteria:

- Hits overlap on the chromosome  
*If hits do NOT overlap on the chromosome:*
- Sections of the cDNA that match to adjacent hits on the chromosome have overlap < 20%.  
*NOTE:* This case was included to avoid grouping hits resulting from tandem duplication.

- Distance between hits on the chromosome is at most 40,000bp, according to the approximate length of the longest human intron [6].

We stop combining BLAST hits and start a new chromosomal “gene segment” if any of the above-mentioned cases fail. We also start a new segment if the direction changes, the chromosome changes, or the cDNA that is being matched changes.

*Part 2. Determining Exon Count*

The Gluing Algorithm worked quite well at grouping BLAST output into gene segments to be used as input for GeneWise. However, due to the running time of GeneWise, and because it is not necessary to determine the intron-exon boundaries for genes that show no evidence of harboring retrocopies, the Gluing Algorithm was modified to keep an accurate count of the number of exons per gene segment. The only necessary piece of information in regard to gene structure for identifying retrocopies is whether the match is multi- or single-exon.

To avoid miscalculating high-scoring pairs (HSP) as multiple exons, a minimum intron length was established to be 30bp, since the smallest animal introns can be between 20 and 30bp [12]. In addition, sequence divergence could lead to corresponding gap openings in both the cDNA query and chromosome, which would *not* imply an intron. To take these gaps into account, the intron length was calculated as |(distance between adjacent chromosome hits) – (distance between matches to the cDNA query)|. As previously mentioned, the maximum intron length is already set to 40,000bp [6].

*Part 3. Calculating Match Scores*

Although BLAST returns match scores for each individual hit, scores need to be recalculated for the “glued” hits representing genes. These values are used later in the pipeline when checking against cut-off parameters. Average e-value and summed bit scores are used, but the alignment length calculation is slightly more complicated due to the nature of the BLAST hits. In this context, “alignment length” refers to the percent (length) of a cDNA transcript that matched to the chromosome. This value’s use is explained in Section IV.

While chaining together BLAST hits, some hits are glued together because they overlap on the

**Table 1 |** Select BLAST hits for cDNA query “ENST00000279575” on human chromosome 12.

Query Start	Query End	Subject Start	Subject End
794	913	11460136	11460017
132	797	11461819	11461154
306*	658	11461834	11461479
180*	613	11461834	11461401
97	135	11462341	11462303
1	98	11463366	11463269

\* Hits do not correspond to exons, according to Ensembl.

chromosome. However in many cases, these hits also overlap on sections of the cDNA query, meaning an intuitive calculation of percent length (summing cDNA match lengths) would yield a value over 100%. Table 1 shows sample BLAST output where this situation occurs. Although the BLAST hits are ordered by subject start, the corresponding cDNA query matches are not necessarily in order, and may overlap with several other previous cDNA query sections. Consequently, a dynamic approach is taken, where a complete list of “exons” is kept and updated with each new hit encountered. Although this step seems computationally inefficient, a trade-off is the quick discovery of retrocopies later in the pipeline through the use of the “alignment length” value.

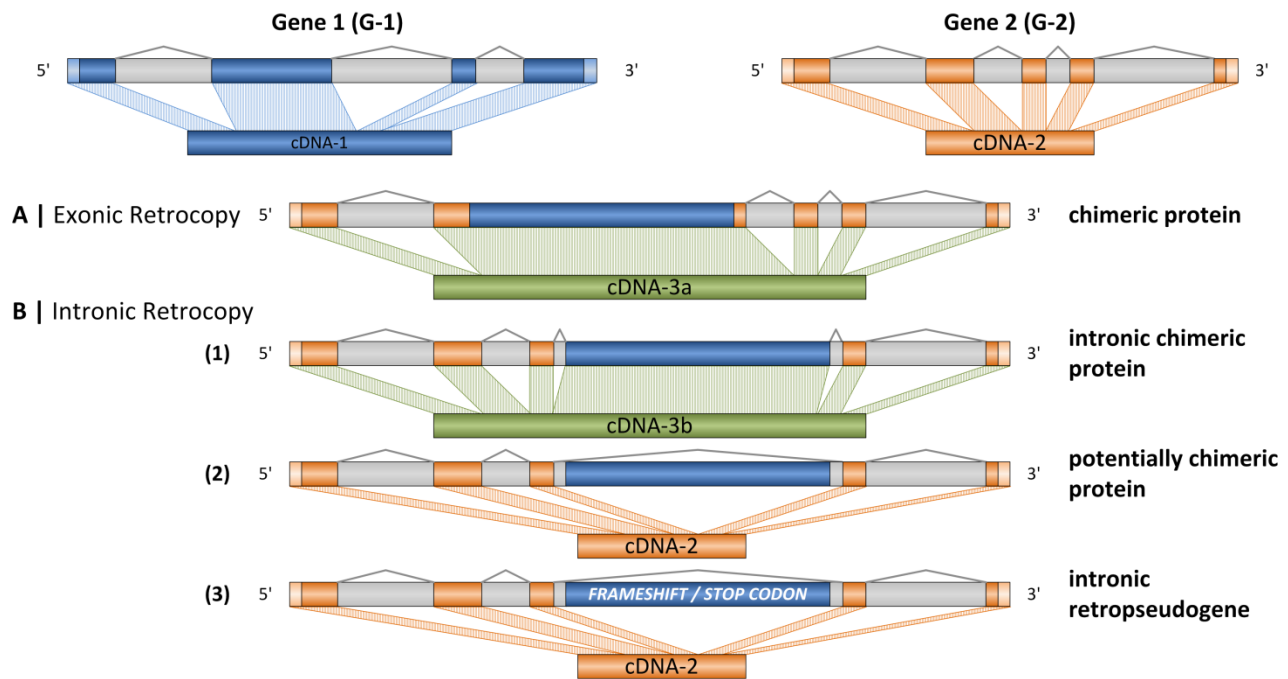
**III. Identify Chimeric Genes with GeneWise**

A large number of previous studies have made extensive use of existing annotation information to identify retrocopies within a genome [1,5,11,13,16,19]. By keeping track of the number of exons in each gene using the Gluing Algorithm, the only part of the pipeline still relying on annotation information is the determination of chimeric genes. In place of using existing, publically available annotation information, the GeneWise gene structure prediction algorithm [3], or “Wise2,” is utilized.

Chimeric genes can result from the retrotransposition of a cDNA transcript into an existing gene region. The four main outcomes of such a retrotransposition can be found in Figure 4. Retrocopies which retrotranspose into an existing exon, or are else transcribed as a new exon in a gene, are chimeric [4,11,17]. If, on the other hand, no transcript exists that gives evidence for the transcription of the retrocopy, the retrocopy is generally considered to be an intronic retroseudogene. However, the cDNA transcriptome available for a genome is often incomplete and fragmented at best. Therefore in our implementation, if a retrocopy exists in the intron of a gene, yet no frameshifts or premature stop codons are found, this retrocopy is marked as potentially chimeric.

*Part 1. Preparing GeneWise Input*

Only those multi-exon genes which harbor single-exon retrocopies can be considered chimeric. Therefore, the first step in identifying potentially chimeric genes involves grouping together output from the Gluing Algorithm based on overlap. All “glued” hits are ordered by position on the chromosome, and hits that overlap by 95% or more are grouped together. Groups containing only multi-exon hits are discarded, because the multiple transcripts in these groups could only have originated through alternative splicing of a multi-exon gene. Also, groups containing solely single-exon hits are discarded because they could not possibly harbor any other retrocopies.



**Figure 4 |** Origination of Chimeric Proteins. There are four possible ways a cDNA originating from a multi-exon gene (cDNA-1 from G-1) may be incorporated once it has retrotransposed into the multi-exonic region of a second gene (G-2). **(A)** Chimeric protein: cDNA-1 retrotransposes into an exon of G-2, and a new transcript (cDNA-3a) with sequence similarity to both cDNA-1 and cDNA-2 is produced. **(B1)** Intronic chimeric protein: cDNA-1 retrocopy inserts into an intron of G-2, yet is now transcribed as a new exon, evident from a new transcript cDNA-3b. **(B2)** Potentially chimeric protein: no evidence (existing sequenced cDNA) shows incorporation of cDNA-1 as a new exon, but a lack of frameshifts or premature stop-codons in the cDNA-1 retrocopy implies the possibility of a chimeric protein. **(B3)** Intronic retropseudogene: no existing cDNA sequence gives evidence that the cDNA-1 retrocopy is transcribed, and frameshifts and/or stop codons also imply a lack of transcription.

Once these hits are filtered out, the single-exon copies of each group are then analyzed further. Cross-referencing back to the results from the Gluing Algorithm, only those single-exon copies are kept where the same cDNA transcript mapped to a multi-exon location elsewhere in the genome. Otherwise, the single-exon copy would not imply retrotransposition, but rather alternative splicing of a single exon from a gene. As previously mentioned, retrocopies are “single-exon” gene transcripts which originate from a multi-exon gene with introns. After these are removed, all groups with solely multi-exon members are again discarded.

The members of all remaining groups eventually become input for GeneWise. The “genewise” version of GeneWise is used, where multiple protein queries are aligned to the same chromosomal DNA segment. The cDNA transcript sequences are translated to amino acid sequences in the forward frames +1, +2, and +3 using the “transeq” program available in the EMBOSS package ([emboss.sourceforge.net](http://emboss.sourceforge.net)). We choose the frame with the longest open reading frame (ORF) to mark as the correct protein sequence and use as GeneWise input. Even though the ORF lengths were originally calculated as the distance between a methionine and a stop codon, we later realized that incomplete transcripts often lack a start codon. As a result, the ORF length is instead calculated as the distance from stop codon to stop

codon. This method returned the correct amino acid sequence for 40/40 random test cases based on Ensembl’s available annotation of cDNA transcripts and their corresponding amino acid sequences.

The second input file contains the entire chromosome region of a group which the multiple cDNAs mapped to. Although the GeneWise documentation recommends adding 15000bp to each end of the gene-containing region prior to running, this step substantially increased the runtime of GeneWise, to the extent where it was infeasible to use the program to determine intron-exon boundaries as needed. Consequently, we tested GeneWise output for five different groups and determined that “flanking” regions of as low as 300bp consistently gave the same results as the flanking regions of 15000bp. To account for cases which were not accurately represented by our test group, we add 600bp regions to each end of the gene segment. This does not result in any major runtime augmentation.

#### Part 2. Using GeneWise Output

The latest stable release of Wise2, version 2.1.20 ([www.ebi.ac.uk/Wise2](http://www.ebi.ac.uk/Wise2)), is run on the two input files (one containing all translated protein sequences which map to a particular gene segment, and the second with that chromosomal DNA segment containing a gene region with an extra 600bp on each end) for each group on a

cluster. The locations of exons and their corresponding bit scores within each region are collected as output.

For each gene region, all exons from multi-exon gene hits are collected and ordered. Often, alternative splicing results in multiple transcripts with considerable overlap in their exons, so only non-duplicate exons were added. This method gives an accurate map of the entire gene structure.

After the gene structure in each region is determined, we compare each single-exon copy to each of the exons in the assembled gene. If a single-exon copy matches >80% to one of the exons, it is marked as chimeric. Otherwise, it is marked as potentially chimeric because by default, it must have mapped to an intron.

Intronic retrocopies are then checked for premature stop codons or frameshift mutations when compared to the original transcript. To calculate frameshift, the original BLAST results are revisited and used as shown in Equation 1. The three stop codons are also searched for by iterating through the entire section of the chromosome which aligned to the cDNA transcript.

(Equation 1)

```
frameshift? {
  true if ( chromosome alignment length
           - cDNA alignment length) mod 3 ∈ {1,2}
  false otherwise
}
```

### Part 3. Observing GeneWise Inadequacies

The method for detecting chimeric genes worked well and as expected in many cases with the GeneWise output that was collected. However, we noticed that in a majority of cases, the GeneWise program returned some unexpected results. For many chromosomal segments, the program returned a fairly complete gene structure, then added single exons in separate “genes” afterwards, sometimes with lengths as few as 2 or 3 base pairs. Such a result can be found in Table 2, where the supposed exon lengths have been calculated to draw attention of

**Table 2** | GeneWise results returned for translated human cDNA “ENST00000000412” on chromosome 12.

Exon Num.	Start Position	End Position	Length
G1, Exon 1	9 102 301	9 102 135	166
Exon 2	9 099 052	9 098 876	176
Exon 3	9 098 231	9 098 065	166
Exon 4	9 096 557	9 096 448	505
Exon 5	9 096 182	9 096 052	130
Exon 6	9 095 189	9 095 063	126
Exon 7	9 094 587	9 094 240	347
G2, Exon 1	9 094 237	9 094 235	<b>2</b>
G3, Exon 1	9 094 224	9 094 111	113
G4, Exon 1	9 094 106	9 094 095	<b>11</b>
G5, Exon 1	9 094 089	9 093 802	287
G6, Exon 1	9 093 796	9 093 788	<b>8</b>
G7, Exon 1	9 093 783	9 093 028	745

example “short” exons. To remove this noise from GeneWise results, we set a minimum exon length of 40 base pairs and did not include redundant exons.

However, we also ran into cases where GeneWise returned incomplete information, in contrast to the previous issue of receiving extra, inaccurate information. We noticed several cases where, given an entire gene region and an appropriate protein sequence, GeneWise returns a single, erroneous (compared to Ensembl’s annotation) exon, whereas the Gluing Algorithm picked up all exons through the combination of BLAST hits.

To check that GeneWise at least returns acceptable output on known cases, we tested the algorithm using protein sequences for ten random genes downloaded from Ensembl, along with their corresponding locations in the human genome. GeneWise returned a correct gene structure for only two out of the ten random cases. In the other eight, exons were either missing or new exons were introduced, and in one extreme case, GeneWise returned one exon out of twenty-four expected, and this “exon” was actually located in an intronic region, as specified by Ensembl’s annotation of the human genome.

Another major problem was the inconsistency of GeneWise results within a given region. Combining results for a single overlap region yielded multiple exons which overlapped sometimes 10-30%, leaving little space for introns. In some cases, a single so-called exon would span an entire 1000-2000 base pair region, where several other exons (from different queries) are also located.

Problems with producing meaningful results stem from these inadequate GeneWise results. One eventual and necessary future improvement would be to eliminate the use of GeneWise and find a better way to determine intron-exon boundaries in a gene region. In addition, these results bring to question the results of previous studies which rely heavily on GeneWise output [17].

## IV. Find Retrocopies from Homolog Groups

Finding all retrocopies is a slightly simpler process than identifying and classifying chimeric genes. To start, all “multi-exon” chained hits (as returned by the Gluing Algorithm) where percent identity > 85% and match length > 95% are collected. Then, all “single-exon” chained hits where bit score > 400 and match length > 85% are also collected and added to the previous set of hits.

Within this newly compiled set, all single-exon copies with at least one corresponding multi-exon hit (generated by the same cDNA query sequence) are marked as retrocopies. To differentiate between retrogenes and retropseudogenes, frameshift mutations and premature stop codons are searched for as described in Section III.

**Table 3** | Counts of cDNA queries and genome matches in the human and chimp genomes at major steps in the pipeline.

Step	Human		Chimp	
	<i>cDNAs</i>	<i>matches</i>	<i>cDNAs</i>	<i>matches</i>
Download from UCSC	54617	--	34623	--
Remove <300bp sequences	52791	--	33442	--
Run mpiBLASTN	52548	2156725	33203	1232690
Glue BLAST hits	52548	662911	33203	329336
Remove hits <40bp	52441	628749	33032	315113
Remove groups with solely multi- or solely single-exon members	39080	595160	23162	291386

## RESULTS

We ran the pipeline in its entirety on both the human and chimp genomes, which are known to contain retrocopies, and are annotated to the extent that our intermediate results could be cross-referenced back to Ensembl for confirmation.

Cutting down both cDNA queries and their hits is an important preprocess because it reduces the input for GeneWise, which is one of the largest runtime contributors in the pipeline. The number of cDNA queries and hits to the genome (BLAST and chained) can be found in [Table 3](#) for each major step taken prior to running GeneWise.

We found 8069 retrogenes and 8070 retroseudogenes in the human genome. Similarly, we found 1859 retrogenes and 2494 retroseudogenes in the chimp genome.

Although the GeneWise results were inaccurate, we still used these results to identify chimeric genes. We found 688 instances where retrocopies retrotransposed into an exon of an existing gene (chimeric genes), and 202 instances where retrocopies retrotransposed into the intronic region of an existing gene. We also found 534 chimeric genes in the chimp genome, and 130 intronic retroseudogenes.

## DISCUSSION

The number of retrocopies we found were approximately equal to the number of retrocopies found in some studies [1,7,19], yet also had large deviations from figures reported in other studies [11,13,16]. Such deviations are likely due to different cut-off values used between the two methodologies [7]. Our count of retrocopies tends to be higher than counts from older studies, and to explain this, it is likely that the genome builds and cDNA transcript libraries used in older studies may have been less complete than current builds.

Another noticeable difference between the retrocopies in the human and chimp genomes is that

there are apparently far fewer retrocopies in the chimp genome. Although the human and chimp genomes are roughly the same size, this difference in retrocopy presence is most likely due to the fact that the chimp genome is not as well sequenced as the human genome. Consequently, fewer full cDNA transcripts are available, and the build of the chimp genome contains more unplaced sequences which were discarded. The first few rows of [Table 3](#) illustrate this

point. There also appear to be more retroseudogenes versus retrogenes in the chimp genome than in the human genome. This again could be due to the incomplete cDNA library used for the chimp study. Missing transcripts result in not only fewer single-exon hits, but also fewer multi-exon hits. Improper builds resulting in poor scoring alignments could also eliminate potential multi-exon copies due to our strict criteria.

The chimeric retrocopy determination relies heavily on GeneWise annotation results. As shown in [Section III](#) of the Methodology section, these results are often highly erroneous. Consequently, analysis based on these results will also be inaccurate.

Specifically, there were far more chimeric genes found than intronic retroseudogenes in both the human and chimp genomes. This result is inconsistent with similar ratios determined in other studies, where it is much more likely for a retrocopy to be an intronic retroseudogene than a chimeric gene [13,17]. By analyzing GeneWise results, it becomes apparent that the high ratio of chimeric genes to intronic retroseudogenes is due to the high number of non-overlapping exons in each gene region. These exons can cover an entire gene region such that there are no apparent introns, as returned by GeneWise.

## FUTURE STEPS

There are many ways to expand and improve this existing retrocopy identification pipeline. Improvements can make the pipeline more computationally efficient and accurate. Certain extensions would allow for the easy analysis of newly identified retrocopies. These improvements and expansions are described below.

### I. Improvement of Gene Annotation Methods

As discussed in the methodology section, the annotation results acquired from GeneWise are often less than satisfactory. Subsequently, methods for classification of chimeric genes that rely on these GeneWise results can provide erroneous results. However, using BLAST results

to approximate a rough annotation of the genome is possible [15,18], at least with concern to linking hits to the genome corresponding to exons. Our preliminary work in this direction has given promising results, but still requires fine-tuning.

Although GeneWise is generally accepted as the “best” annotation software available [3,17,19], it is also worthwhile to examine the accuracy of other available annotation software, such as Sim4, Est2gen, or est\_genome [18]. Another option would be to rely on the available online annotation of a genome; yet in order for this methodology to remain extendable to other, unannotated genomes, this option is undesirable.

## II. Assignment of Parental Genes

The “parental gene” of a given retrocopy refers to the gene that produced the transcript which was later retrotransposed to produce the retrocopy.

Previous studies have been able to match each identified retrocopy to a single multi-exon parent [11,16,17,19]. Taking into consideration the structure of our existing algorithms and data storage, a simple way to accomplish this would be to utilize existing “homolog families” [17]. Within each homolog family, the sequence of each single-exon member can be compared to all other members within the group. If the closest member is another single-exon copy, that “retrocopy” can be eliminated, because it likely arose from the duplication of an existing retrocopy, and not through retrotransposition [1,17]. Otherwise, the closest multi-exon match could be labeled as the parent.

Also importantly, previous studies have not attempted to *group* parental genes in order to determine if a single gene can produce multiple retrocopies. This information would provide insight as to which genes are more likely to produce retrocopies.

## III. Retrocopy Age

Once identified, the age of a retrocopy can be determined based on the amount of accumulated mutation in its sequence. To determine this mutation amount, previous studies compare the retrocopy sequence to that of its parental gene [5,11,13,19]. There are a few other methods of determining retrocopy age, however, that are worth exploring. Such methods that do not require prior knowledge of a parental gene are especially important for dating retrocopies where the parental gene is no longer transcribed or in existence, and therefore cannot be used for comparison.

The standard and commonly used method of determining retrocopy age is to calculate the  $K_s$  value (synonymous mutation at each site) [5,11,13,19], given a rate of silent substitutions per site per year. An example rate is  $1-1.3 \times 10^{-9}$  substitutions per site per year [11].

However, calculating the  $K_s$  value requires the sequence of the transcribed parent gene. Therefore, two remaining alternatives include measuring the decay (length) of the poly-A tail, or calculating sequence similarity between the two identical flanking regions to the retrocopy [14].

Limitations to all of the above-mentioned methods arise when a retrocopy (including its poly-A tail and flanking regions) has decayed to a point beyond recognition. This can be thought of as the “death” of a retrocopy because it can no longer be identified, and subsequently, its age cannot be determined.

## IV. Retrotransposition Model

An interesting further step would be to develop a theoretical model of the retrotransposition process, which could then be used to predict the “life cycle” of a retrocopy. Incorporating the age distribution of retrocopies within a genome or multiple genomes, the assigned parental genes, and location of existing retrocopies, a model could be developed and used to answer several questions:

- Are any direct repeats generated during the insertion of the retrotransposed genes?
- Is it more likely for retrogenes to transpose out of or into a certain chromosome versus others?
- Is there a significant pattern of chimeric proteins resulting from retrotransposition? Do they have a similar purpose or function?

## ACKNOWLEDGEMENTS

Thanks to Ioana Bercea for programming help and input. Thanks to Lenwood Heath for advice and discussions throughout the project duration. Thanks to Rob Hunter for technical assistance. This project was funded by a grant from CRA-W (Computing Research Association’s Committee on the Status of Women in Computing) and CDC (Coalition to Diversify Computing) through the DREU (Distributed Research Experiences for Undergraduates) program.

## REFERENCES

- [1] Adel, Khelifi, Duret Laurent and Mouchiroud Dominique (2005). [HOPPSIGEN: A Database of Human and Mouse Processed Pseudogenes](#). *Nucleic Acids Research* **33**:D59-D66.
- [2] Bergman, Casey M. and Hadi Quesneville (2007). [Discovering and Detecting Transposable Elements in Genome Sequences](#). *Briefings in Bioinformatics* **8**(6):382-392.

- [3] Birney, Ewan, Michele Clamp and Richard Durbin (2004). [GeneWise and Genomewise](#). *Genome Research* **14**:988-995.
- [4] Buzdin, A.A. (2004). [Review: Retroelements and Formation of Chimeric Retrogenes](#). *Cellular and Molecular Life Sciences* **61**:2046-2059.
- [5] Dai, Hongzheng, Toshio F. Yoshimatsu and Manyuan Long (2006). [Retrogene Movement Within- and Between-Chromosomes in the Evolution of \*Drosophila\* Genomes](#). *Gene* **385**:96-102.
- [6] Deutsch, Michael and Manyuan Long (1999). [Intron-Exon Structures of Eukaryotic Model Organisms](#). *Nucleic Acids Research* **27**(15):3219-3228.
- [7] Drouin, Guy (2006). [Processed Pseudogenes Are More Abundant in Human and Mouse X Chromosomes than in Autosomes](#). *Molecular Biology and Evolution* **23**(9):1652-1655.
- [8] Fablet, Marie, Manuel Bueno, Lukasz Potrzebowski and Henrik Kaessmann (2009). [Evolutionary Origin and Functions of Retrogene Introns](#). *Molecular Biology and Evolution* **26**(9):2147-2156.
- [9] Kaessmann, Henrik, Nicolas Vinckenbosch and Manyuan Long (2009). [RNA-Based Gene Duplication: Mechanistic and Evolutionary Insights](#). *Nature Reviews Genetics* **10**:19-31.
- [10] Lynch, Michael and John S. Conery (2000). [The Evolutionary Fate and Consequences of Duplicate Genes](#). *Science* **290**:1151-1155.
- [11] Marques, Ana Claudia, Isabelle Dupanloup, Nicolas Vinckenbosch, Alexandre Reymond and Henrik Kaessmann (2005). [Emergence of Young Human Genes after a Burst of Retroposition in Primates](#). *Public Library of Science Biology* **3**(11):1970-1979.
- [12] Miao, H.E., L.I. Jidong, and Shanghong Zhang (2006). [Statistical Characteristics of Eukaryotic Intron Database](#). *Frontiers of Biology in China* **4**:363-366.
- [13] Pan, Deng and Liqing Zhang (2009). [Burst of Young Retrogenes and Independent Retrogene Formation in Mammals](#). *Public Library of Science ONE* **4**(3).
- [14] SanMiguel, Phillip, Brandon S. Gaut, Alexander Tikhonov, Yuko Nakajima and Jeffrey L. Bennetzen (1998). [The Paleontology of Intergene Retrotransposons of Maize](#). *Nature Genetics* **20**:43-45.
- [15] She, Rong, Jeffrey S.-C. Chu, Ke Wang, Jian Pei, and Nansheng Chen (2009). [genBlastA: Enabling BLAST to Identify Homologous Gene Sequences](#). *Genome Research* **19**:143-149.
- [16] Vinckenbosch, Nicolas, Isabelle Dupanloup and Henrik Kaessmann (2006). [Evolutionary Fate of Retroposed Gene Copies in the Human Genome](#). *Proceedings of the National Academy of Science* **103**(9):3220-3225.
- [17] Wang, Wen, Hongkun Zheng, Chuazhu Fan, Jun Li, Junjie Shi, Zhengqiu Cai, Guojie Zhang, Dongyuan Liu, Jianguo Zhang, Søren Vang, Zhike Lu, Gane Ka-Shu Wong, Manyuan Long and Jun Wang (2006). [High Rate of Chimeric Gene Origination by Retroposition in Plant Genomes](#). *The Plant Cell* **18**:1791-1802.
- [18] Zhang, Hongyu (2003). [Alignment of BLAST High-Scoring Segment Pairs Based on the Longest Increasing Subsequence Algorithm](#). *Bioinformatics* **19**(11):1391-1396.
- [19] Zhang, Zhaolei, Paul M. Harrison, Yin Liu and Mark Gerstein (2003). [Millions of Years of Evolution Preserved: A Comprehensive Catalog of the Processed Pseudogenes in the Human Genome](#). *Genome Research* **13**:2541-2558.