



Higher Rates of Processed Pseudogene Acquisition in Humans and Three Great Apes Revealed by Long-Read Assemblies

Xiaowen Feng ^{1,2} and Heng Li ^{*,1,2}

¹Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA, USA

²Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

*Corresponding author: E-mail: hli@ds.dfci.harvard.edu.

Associate editor: Irina Arkhipova

Abstract

LINE-1-mediated retrotransposition of protein-coding mRNAs is an active process in modern humans for both germline and somatic genomes. Prior works that surveyed human data mostly relied on detecting discordant mappings of paired-end short reads, or exon junctions contained in short reads. Moreover, there have been few genome-wide comparisons between gene retrocopies in great apes and humans. In this study, we introduced a more sensitive and accurate method to identify processed pseudogenes. Our method utilizes long-read assemblies, and more importantly, is able to provide full-length retrocopy sequences as well as flanking regions which are missed by short-read based methods. From 22 human individuals, we pinpointed 40 processed pseudogenes that are not present in the human reference genome GRCh38 and identified 17 pseudogenes that are in GRCh38 but absent from some input individuals. This represents a significantly higher discovery rate than previous reports (39 pseudogenes not in the reference genome out of 939 individuals). We also provided an overview of lineage-specific retrocopies in chimpanzee, gorilla, and orangutan genomes.

Key words: processed pseudogene, bioinformatics, humans and apes.

Introduction

Active human LINE-1s are responsible for various retrotransposons in the genome (Ostertag et al. 2003; Kazazian 2011; Mandal et al. 2013), such as SVA, Alu, and processed pseudogenes (Esnault et al. 2000), with a preference of L1 RNAs over non-L1 templates (Wei et al. 2001; Pavlíček et al. 2002). Although processed pseudogene formation usually leads to nontranscribed retrocopies because of the absence of promoter regions, in some cases, the retrocopies could encode proteins (Pink et al. 2011), regulate their parent genes (Cheetham et al. 2020), and ultimately result in significant functional implications such as carcinogenesis (Cooke et al. 2014; Poliseno et al. 2015). The mechanism behind parent gene preference has not been fully revealed yet (Podlaha and Zhang 2009; Kazazian 2011; Richardson et al. 2015).

Processed pseudogene formation in the human genome has remained an active process both in germline and somatic tissues, and nonreference events are described by Schrider and Navarro et al. under the term retroCNVs (Schrider et al. 2013), and later as gene retrocopy insertion polymorphisms (GRIPs). The term does not assume whether a given retrocopy is functional or not. We will also use “processed pseudogene” and “gene retrocopy” interchangeably without any implication about functionality. The total number of processed pseudogenes in the human genome has been estimated to range roughly from around 2,000 to more than

10,000 and settled down on the higher end, depending on the criteria and discovery methods used (Marques et al. 2005; Zhang et al. 2006; Molineris et al. 2010; Navarro and Galante 2015; Frankish et al. 2019). Ewing et al. (2013) found 39 GRIPs representing 36 parent genes in 939 samples from 1000 Genome Project (a gene may have multiple retrocopies), and 26 GRIPs from 85 tumor–normal pairs from TCGA data set, where the two sets overlapped for 17 GRIPs. Cooke et al. (2014) further examined 660 cancer samples, and found a total of 42 somatic events in 17 samples (Wei et al. 2001).

Technologies such as Oxford Nanopore and Pacific Biosciences (PacBio) have enabled the sequencing of reads of kilobases in length, which could further be assembled into contigs of tens of megabases long. Given that 96% human transcripts annotated in Gencode are shorter than 10 kb, the longest one being 109 kb and the medium length being 2.9 kb, we expect long-read-based assemblies to reveal the most processed pseudogenes. In this study, we introduced a novel processed pseudogene discovery approach which was more sensitive and accurate than short-read-based methods, compared the findings with established results, and analyzed the L1 hallmarks as well as sequence landscapes around the retrocopies. Our results hinted that the recent GRIPs among the human population could be much more prevalent than previously suggested, lifting the rate from GRIPs of 39 events (36 parent genes) per 939 individuals to 40 events (36 parent genes) per 22 individuals. We also found 17 events (16 parent

genes) that are in the GRCh38 but not shared by at least one sample. Moreover, we examined three great ape assemblies (chimpanzee, gorilla, and orangutan) and provided an overview of their lineage-specific events.

Results

Processed Pseudogene Polymorphism Surveyed in 22 Samples

We obtained 34 long-read-based assemblies for 22 human samples (Seo et al. 2016; Vollger et al. 2019; Garg et al. 2020; Cheng et al. 2021). Each assembly approximately models a 3-Gb human genome. We have more assemblies than samples because 12 samples (HG00733, HG01109, HG01243, HG02080, HG02723, HG02818, HG03486, HG03492, NA12878, NA24385, NA19240, and PGP1) have two assemblies per sample, representing the two phased haplotypes in a diploid human. We aligned the contigs to GRCh38 and called structural variants (SVs), including insertions and deletions of 50 bp or longer (see Materials and Methods). On average, we called 16,661 long insertions and 11,283 long deletions from each assembly. We extracted the genomic sequences of protein-coding genes (“gene reference,” including introns) using gene coordinates provided in Gencode and sequences of GRCh38, and splice-aligned the long SVs to the gene reference. We found that an average of 6,772 or 24.3% SVs in each assembly was aligned with exon junctions. They were further screened for gene-like structures. Briefly, we required that a retrocopy should contain at least two exons and no more than one intron; retrocopies with only two exons were not allowed to retain any intron (see Materials and Methods). We also manually inspected the selected SVs (fig. 1B, supplementary table S1, Supplementary Material online; see Discussion). We would from here refer to the retrocopies identified from long insertions/deletions as inserted/deleted retrocopies or such processed pseudogenes. Deleted retrocopies were the retrocopies that existed in the GRCh38 and not found in any of the compared assemblies (i.e., ancestral events of GRCh38 not seen in some other assemblies; not deletions in either assemblies), and inserted retrocopies were the ones found only in the assemblies but not the GRCh38. Human/ape lineage-specific copies were also interpreted in this way (fig. 1A).

We found a total of 148 inserted retrocopies and 102 deleted retrocopies derived from 36 inserted parent genes (*CBX3*, *EEF1A1*, *FBRSL1*, *GAPDH*, *GCSH*, *HNRNPC*, *HSPE1*, *KIAA2013*, *MFF*, *MOSMO*, *MT1H*, *NIP7*, *NREP*, *NUDT4*, *PAIP1*, *PDCL3*, *POLR2C*, *PPIA*, *RHEB*, *RPL10*, *RPL21*, *RPL22*, *RPL9*, *RPLP0*, *RPS26*, *RPS28*, *RPS3A*, *SAV1*, *SKA3*, *SLC25A33*, *TDG*, *TYRO3*, *UPF3A*, *UQCR10*, *USP28*, *ZNF664*) and 16 deleted parent genes (*ABHD17A*, *DHFR*, *EEF1A1*, *EIF1AX*, *GCSH*, *GNG10*, *NUDT4*, *RAMAC*, *RHEB*, *RPL21*, *RPL36A*, *RPL9*, *RPS26*, *RPS28*, *SLC25A33*, *UPF3A*). Retrocopies of ten parent genes (*GCSH*, *RHEB*, *SLC25A33*, *NUDT4*, *RPL9*, *RPS26*, *RPS28*, *EEF1A1*, *UPF3A*, *RPL21*) were found in both deletions and insertions at different genomic locations, implying that these genes were relatively more active in retrotranspositions (fig. 2 and supplementary table S4, Supplementary

Material online). Lengths of retrocopies varied from 408 to 4,725 bp, with the mean value of 1,595 bp and the median of 1,149 bp.

With paired-end short reads, Ewing et al. (2013) identified six novel retrocopies in HG00268 (*TDG*, *TYRO3*, *CBX3*, *GCSH*, *RPL10*, and *MCTP2* as parent genes) and three retrocopies in NA19434 (*TDG*, *CBX3*, and *MCTP2*). We observed all but the *MCTP2* retrocopies (both) and the *TDG* retrocopy (NA19434) and in our corresponding assemblies, probably because the two samples only have collapsed assemblies where one of the two parental alleles is randomly dropped. All TSDs reported were identical to those of Ewing et al. (2013). We were able to identify two more inserted retrocopies in HG00268 and two more in NA19434, demonstrating the enhanced power of long-read data.

To confirm the effect of collapsing haplotypes, we collected gene retrocopies from collapsed assembly of NA12878 (GCA_002077035.3), HG00733 (GCA_002208065.1), and NA19240 (GCA_001524155.4) that were assembled independently from the corresponding phased assemblies, and compared the findings. We confirmed that the phased assemblies were able to capture all retrocopies with less erroneous signals during the annotation. Each set of phased assemblies were able to provide an additional three (*FBRSL1*|chr6, *RPS28*|chr17, *HNRNPC*|chr6), two (*SKA3*|chr11, *PDCL3*|chr2), and one (*RPLP0*|chr11) inserted retrocopy(ies), and 2, 3, and 3 deleted retrocopies, respectively. This suggests that using collapsed assemblies, our approach underestimates the abundance of processed pseudogenes because of random haplotype dropout. Detecting processed pseudogenes from raw reads would address the issue. However, due to the technical difficulty in calling long SVs from noisy long reads, read-based discovery is challenging.

L1 Hallmarks, Sequence Identity, and Alternative Splicing in Human Processed Pseudogenes

Since L1-mediated retrotransposition mechanism of mRNAs is target-primed reverse transcription (TPRT) (Cost et al. 2002), target site duplications (TSDs), and endonuclease cleavage sites were expected to be found close to the ends of the retrocopies. All retrocopies reported in this study exhibited at least 6-bp proper target site duplications (TSDs), that is, a pair of TSD is only accepted if it contains no more than 1-bp mismatch or indel, and locates at the immediate flanking regions of the retrocopy sequence (fig. 1; see Materials and Methods), with an average length of 14-bp TSD motif for deleted events and 15 bp for inserted events. The cleavage site motif resembled previous reports (Ostertag and Kazazian 2001a; Crooks et al. 2004) (supplementary table S1, Supplementary Material online). PolyA tracts that were no less than 6 bp and located at the proper position was required for all events with TSD shorter than 10 bp (see Materials and Methods); we identified polyA tails in 37 out of 40 (92%) inserted events, measuring a median length of 21 bp with max length of 61 bp, and in all deleted events with median polyA tail length of 14 bp and max length of 26 bp (supplementary table S2, Supplementary

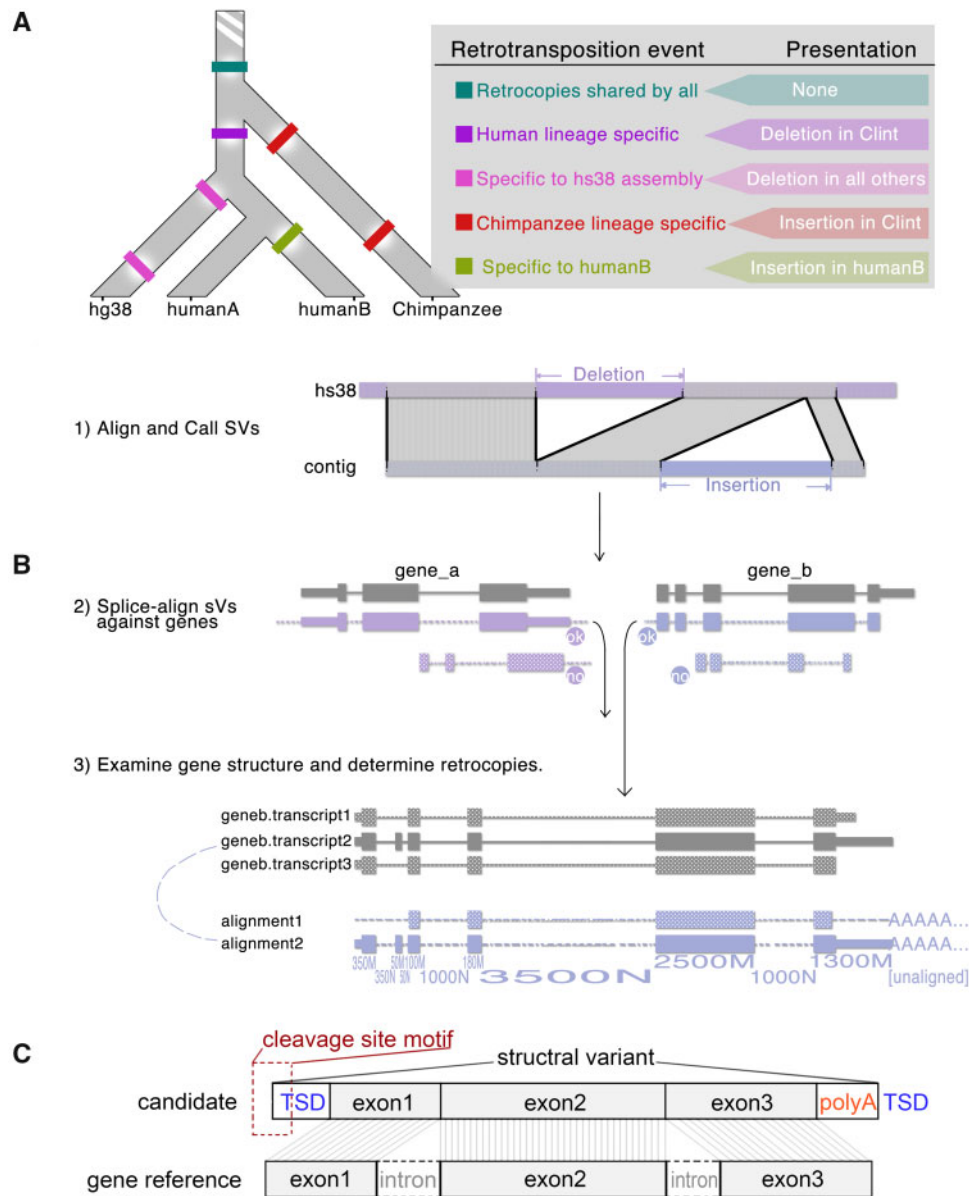


Fig. 1. Overview of the gene retrocopy discovery approach. (A) Lineage-specific and nonreference retrotransposition events. For example, if an ancient event occurred before human–chimpanzee divergence, then theoretically its retrocopy would not be visible during structural variant calling since both GRCh38 and Clint would possess the copy at the same genomic location. On the other hand, as shown in the figure, if humanB’s lineage gained a processed pseudogene after diverging from the lineage of humanA, then the retrocopy was expected to show up as a long inserted SV when using either A (inserted) or B (deleted) as the reference genome. (B) Diagram of detecting gene retrocopies from SVs. See Materials and Methods and Discussion for details and edge cases. (C) The expected location of TSD, polyA, and cleavage site motif. Another possible scenario is having the left TSD outside of the SV and the right TSD within SV (see Materials and Methods).

Material online). The tails were relatively short compared with the polyA tail lengths in mature mRNAs which centered at 50 bp (Chang et al. 2014), although individual cases may demonstrate substantial polyA tails. Grandi and Rosser et al. demonstrated experimentally in mice that polyA-tracts of L1-mediated retrotransposition sequences would shorten rapidly both intraindividually and intergenerationally (Grandi et al. 2013), which might be an explanation to our observations in human gene retrocopy polyA tails. The polyA tails also harbored mutations or indels.

Compared with splice forms of protein-coding genes in Gencode, 45% events experience 5’ truncation. Manual

inspection confirmed that events with polyA tail and 3’ truncation were caused by the corresponding splice forms not registered as protein-coding genes (e.g., retrocopies of *RPS28* took a form similar to ENST00000417088.2, which was annotated as “retained_intron” by Gencode), or noncanonical alternative splicings. We also noticed that inserted retrocopies of three genes (*RPL10*, *SKA3*, *ZNF664*) had noncanonical exon organization. For further comparison, we retrieved polished iso-seq data for validation from PacBio data release (<https://www.pacb.com/blog/data-release-whole-human-transcriptome/>, last accessed May 9, 2020) which provided whole genome full-length transcriptomes for human brain, liver, and

The Sequence Landscape near Processed Pseudogene Insertion Point and the Characteristics of Parent Genes

Insertion sites of our identified events showed no obvious preference of inserting into genes. About 41% and 56% of inserted and deleted events overlapped with protein-coding genes (which cover a total of 1.2 Gb on the GRCh38). One inserted event (retrocopy of *UQCR10*, HG01243-H2 [maternal]) located in protein-coding gene *C1orf194*'s exon. It is tempting to think the chance of a GRIP to be positively correlated with the transcript abundance of the parent gene (Sisu et al. 2020). Although ribosomal genes were more active in the retrotransposition process (Zhang et al. 2006; Zerbino et al. 2018), a substantial portion of GRIPs or deleted retrocopies were contributed by lower-expressed genes and more inactive parent genes. We collected gene-level, tissue-specific expression profiles from GTEx in transcripts per million (TPM) (supplementary table S3, Supplementary Material online). The max TPM values of 15% parent genes were less than 100, and 65% parent genes had average TPM values lower than 100. Some genes, such as *NREP*, despite having high tissue specificity (neurons), exhibited retrocopies. Moreover, many GRIPs originated from parent genes that were rather inactive in ancestral events, that is, the known retrocopies in GRCh38 that were also shared by other assemblies (fig. 3A). Since retrotransposition events in somatic cells would not be inheritable, polymorphism of processed pseudogenes is arguably accumulated events that happened to reproductive cells or during early stages of embryo development. It is not possible for this study to confirm this claim, but Feusier et al. (2019) has reported eight L1 germline retrotransposition events in 437 individuals of a 33 three-generations CEPH pedigrees via blood-derived short-read WGS, suggesting L1-mediated events are active and inheritable. Furthermore, Richardson et al. (2017) applied mouse retrotransposon capture sequencing (mRC-seq) and short-read WGS sequencing to pedigrees of mice, which found an L1 insertion rate of ≥ 1 event per eight births, as well as tracing L1-mediated events back to as early as primordial germ cells.

Lineage-Specific Processed Pseudogenes in Great Apes Revisited

We used three assemblies (chimpanzee, gorilla, and orangutan) (Gordon et al. 2016; Kronenberg et al. 2018) for processed pseudogene discovery in great apes (see Materials and Methods; supplementary table S1, Supplementary Material online). We found 187 inserted and 108 deleted retrocopies (143 and 91 parent genes) in the chimpanzee, 136 and 129 retrocopies (119 and 102 parent genes) for the gorilla, 270 and 299 retrocopies (184 and 206 parent genes) for the orangutan, as well as identifying lineage-specific or shared events illustrated in figure 3B and C (supplementary tables S1 and S6, Supplementary Material online). Note that some parent genes shared by the apes might still give rise to lineage-specific events (see Materials and Methods for glossary clarification). For example, nonreference retrocopies of *PPIA* were seen in the human samples and the three great apes, however

all inserted to different genomic locations (human at chr4, chimpanzee at chr6, gorilla at chr11, and orangutan at chr5; chimpanzee's copy was removed due to ambiguous flanking sequences surrounding it, however). Retrotransposition activity level of gorilla, measured by lineage-specific Alus, has been reported to be lower than that of humans (McLain et al. 2013), possibly explaining the lower number of gene novel retrocopies seen in our gorilla sample; humans also showed notable decline in L1 accumulation compared with chimpanzees (Mathews et al. 2003; Hormozdiari et al. 2013; Li et al. 2020), consistent with the observation here. Our results translate to one retrocopy per 187,685 substitutions for the chimpanzee, 1/320,192 for the gorilla, and 1/317,541 for the orangutan.

The Chimpanzee Sequencing and Analysis Consortium reportedly found 246 lineage-specific retrocopies corresponding to around 173 unique parent genes for the chimpanzee Clint (Chimpanzee Sequencing and Analysis Consortium 2005). Since the exact locations of the consortium's findings were not immediately available, and based on the observation that only a few parent genes were likely to give rise to multiple retrocopies, we compared the parent genes unique to chimpanzee lineage reported by the consortium and ours (see Materials and Methods). About 101 parent genes discovered were shared by both (supplementary table S5, Supplementary Material online). About 42 parent genes were found only by our method, being missed by the consortium because of either 1) the retrocopy presented in the assembly but failed to be recognized by the consortium, 2) the retrocopy did not present in the assembly perhaps due to heterozygosity, or 3) the corresponding region not assembled. About 72 parent genes were found only by the consortium. We failed to recognize these retrocopies due to either 1) the retrocopies were not lineage-specific and the consortium should not have recalled the subsequences, 2) random heterozygosity drop out, 3) the retrocopy only represented one exon which was ambiguous to us, or 4) we found that the SV of interest had long flanking sequences, making the retrocopy candidate invalid. Consortium's take on (3) and (4) was not clear. As quality control, we extracted inserted retrocopies from PanTro5 assembly with respect to GRCh38. We found that 15 events in PanTro5 and 12 events in PTR_Clint were unique to the assemblies, respectively, due to the random haplotype dropout. Overall, we were able to reproduce most of the Chimpanzee Consortium's results, provided better annotation and full retrocopy sequences, and added a significant amount of more accurate new parent gene discoveries.

We evaluated L1 hallmarks for the great apes using the same measurements we applied to human data, and yielded similar observations as in human retrocopies. All events were required to display TSDs of at least 6 bp. The median length of TSDs was 14 bp for both inserted and deleted events. About 87.7% inserted events and 86.2% deleted events examined contained polyA tails, median lengths of which were 14 and 13 bp, respectively. 5' truncation was prevalent, similar to humans, as found in 45.7% events. Note that some cases described above might be noncanonical or lineage-specific alternative splicing in the great apes.

Although the L1 hallmark possession rate of the events reported in this study is higher than previous short-read-based reports, we were not able to confirm the hallmarks for all retrocopies. We speculate that, especially for polyA tails, this might be due to mutational events that have happened later than the retrotransposition.

In addition to the reported discoveries, we here describe two observations. First, we noticed that the presence of exon junctions might not be sufficient for polymorphic processed pseudogene identification. It is desirable to at least require TSDs. One example of this is the “retrocopies” of AK2. All our nonreference human assemblies called “inserted retrocopies” of AK2 (truncated, containing exon 4~7 where exon4 lost its 5' end) originating from one same retrotransposition event, which was not explainable by incomplete lineage sorting; sequence divergence between the retrocopy candidate and the reference parent gene was >3%, hinting the event might have happened before human–chimpanzee divergence. Manual inspections suggested that GRCh38 did not misassemble around the insertion point (chr2:31823409), which is also the end point of the known processed pseudogene AK2P2 (truncated, containing exon 1~4 of AK2). Instead, the illusion of polymorphic, truncated AK2 retrocopies were probably created by a deleted SV on GRCh38 that removed exon 1~4 of AK2P2 and flanking sequences. In false cases, the short flanking region of the “retrocopies” might be able to align to the reference genome elsewhere. Whether short-read-based processed pseudogene discovery methods pick up the false signals would depend on their implementation. Second, some retrocopy candidates displayed TSD pairs, polyA tails, and valid cleavage sites, but the SVs hosting them had extra flanking sequences around the presumed retrocopy sequences. This violates the expectation based on retrotransposition mechanism, that is, the SV should contain a retrocopy with TSDs, and mostly nothing else. We set a threshold of total 100 bp for flanking bases (see Materials and Methods). This removed retrocopy of *TERT* and retrocopy of *NUDT4* (HG00514) from the human's processed pseudogene collection.

Our approach could also be utilized on long SV data sets without long-read-based assemblies, although the yield would depend on SV calling quality and expect less discoveries than reported in this study. We tested on [Hehir-Kwa et al.'s \(2016\)](#) SV data set which were obtained from 769 individuals of 250 Dutch families. A total of 20,494 long inserted SVs were selected, from which we identified 19 parent genes ([supplementary table S7](#), [Supplementary Material](#) online). As more long-read data sets and de novo assemblies are becoming available in the near future, we believe that the polymorphic processed pseudogenes could soon get better studied and cataloged.

Materials and Methods

Data and Data Processing

We obtained the following 34 human assemblies: AK1(GCA_001750385), CHM1(GCA_001297185, haploid sample), HG00268(GCA_008065235), HG00514(GCA_002180035), HG01352(GCA_002209525), HG02059(GCA_003

070785), HG03807(GCA_003601015), HG04217(GCA_007821485), and NA19434(GCA_002872155), and the following haplotype-resolved assemblies: NA12878, NA24385, CHM13 (haploid sample), and NA19240 assembled with hifiasm ([Cheng et al. 2021](#)) v0.8 (doi: 10.5281/zenodo.4420402), HG00733, HG02818, HG03486, HG01109, HG01243, HG02080, HG02723, and HG03492 assembled with hifiasm v0.12, PGP1 assembled with whdenovo ([Garg et al. 2020](#)). Hifiasm v0.8 and hifiasm v0.12 were consistent on the SVs of interest. Assemblies can be found under the accession IDs and at <https://zenodo.org/record/4420402> (last accessed January 6, 2021). Hifiasm and whdenovo used Pacbio Hifi reads. For great apes, we obtained the following three assemblies: GCA_900006655.3 (Susie the gorilla), GCA_002880755.3 (Clint the chimpanzee), and GCA_002880775.3 (Susie the orangutan). The assembly of kamilah the gorilla (GCA_008122165) was also processed and described in supplementary tables, [Supplementary Material](#) online. We aligned the assemblies against GRCh38 with minimap2 (2.17-r974; -xasm5 -c -cs -z10000,200 -r5k for humans, -xasm20 -c -cs -z10000,200 -r5k for great apes), the results of which were then sorted (sort -k6,6 -k8,8n) and called for structural variants (SV) with minimap2's `paftools` (k8 `paftools.js` call). We discarded SVs shorter than 50 bp based on SV length distributions (empirical observations in this data set) and exon length distribution of protein-coding genes in the human genome (Ensembl annotation). SV sequences were aligned against the genomic sequences of human protein-coding genes (minimap2 same as above; -xsplice -c -cs -f10000 -N100 -p0.1; flags other than -p0.1 were not sensitive; genomic sequences were defined by GRCh38 and Ensembl V31), in search of properly spliced alignments which would be the implications of processed pseudogenes.

We noticed in the SV calling that a long deleted SV could reside together with an roughly equally long, sequentially similar inserted SV, especially when the two (or more) were inside a long segmental duplication region (indicated by the UCSC `segdup` track). We confirmed that such SVs raised from misassemblies instead of aberrant structure variations or misalignments. One example is chr7:55736476-55738701 in the sample AK1 (alignment method as described above). Not all retrocopies of great apes were manually inspected due to the large volume, however, we excluded suspicious SVs by requiring that any two long SVs of opposite types to not appear together closer than 500 bp.

Long-Read Assembly-Based Processed Pseudogene Discovery

We required the following criteria for processed pseudogenes candidates: 1) at least two exons of a multiexon protein-coding gene are partially ($\geq 20\%$ exon content mapped, counting mismatches, and deletions) or fully represented, 2) no more than one intron is partially ($\geq 20\%$ or ≥ 100 bp, whichever is positive) or fully presented, however if only two exons are present, no intron retaining is allowed, 3) the SV shall not be called from apparently misassemblies (see above), 4) the SV shall not contain more than 500 bp of sequences other than the retrocopy, unless it is a long insertion and the

retrocopy has low sequence divergence (mismatch identity >98.7% for human, >80% for great apes), and 5) TSD and polyA shall be examined only at their expected location (see below), 6) a TSD pair can only contain one mismatch or one indel (not counting the ≤ 2 -bp shift at the beginning, if any), and shall be at least 6 bp, 7) if polyA tract is shorter than 6 bp or not enriched in adenine bases, the TSD has to be at least 10 bp long and the entry has to display a valid cleavage site (see below), and 8) if even when aligning to the most similar gene isoform, the SV is still considered to contain 100 bp or more flanking bases before and after the retrocopy sequence, this entry will be dropped regardless of L1 hallmarks. A clarification of glossary: “retrocopy” counts every occurrence of retrocopies, regardless of whether they resemble the same parent gene, or are shared between samples (i.e., called from the same genomic location). The “event” counts ancestral retrotransposition events, for example, two samples both have retrocopies of gene X at chr1:10,000, and a third sample has another retrocopy of X at chr5:30,000, then in the overall statistics these observations would be counted as three retrocopies, one parent gene, and two events.

Criteria (4) removes generic polymorphic SVs that happen to contain ancient retrocopies. For criteria (5), the “expected location” is reasoned as the following: in the ideal scenario of a retrotransposition event, minimap2 will place one TSD (1) on the start or the end (depends on contig alignment direction) of the SV, and the other TSD (2) on the immediate flanking of the SV on the other side. If the event’s insertion point has no repeats, (1) will be on the 5’ side with respect to the reference’s strandness; alternatively, if there are repeats, the placement might move to the 3’ side. The polyA is expected to be exactly the segment between the TSD and the start or the end of gene alignment on the SV, or exactly the segment between the start/end of gene alignment and the start/end of the SV (see fig. 1C). In the implementation, we first assume the ideal scenario to look for TSD; if failed, we switch to examine the alternative scenario. For criteria (6) and (7), when measuring TSD length for a TSD pair with indel, the length of the shorter TSD motif is used to compare with the thresholds; when measuring polyA tract, only the start and the end of the polyA tract segment is examined, where only 1-bp mismatches are allowed, for example, “AAAATAATTT” is considered a 7-bp tail, “CCAAAATAATTT” is not (2-bp mismatch at the start), and “ACAATAATAAATT” is a 12-bp tail. For criteria (7), a cleavage site, which is expected to be TTTT/AA, if displays more than one cytosine or guanine base, is considered invalid. This check was not required for retrocopy candidates with unambiguous polyA tails. The polyA tail sequences, TSDs, and full retrocopies are available in [supplementary table S2, Supplementary Material](#) online, for humans and [supplementary table S6, Supplementary Material](#) online, for primates. Primates and human samples used the sample criteria and implementation.

Comparison with the Previous Chimpanzee Retrocopy Discovery

The Chimpanzee Sequencing and Analysis Consortium provided descriptions of the parent genes or their protein family

instead of unique IDs or gene symbols, and a few descriptors differed only in one or two characters (e.g., a space or a dot). We manually curated the list, which yielded 143 unique descriptors. Since these descriptors were still too broad for our comparison purposes (e.g., Ras GTPase superfamily), we extracted the 246 sequences described by the consortium, and processed them by our approach to link each one to a specific parent gene, which yielded 173 parent genes. Note that since this is an incremental annotation process instead of discovery, we did not apply criteria (4)~(7). Parent genes assigned to the retrocopies all matched their gene descriptors provided by the consortium as long as the consortium provided unambiguous annotations ([supplementary table S5, Supplementary Material](#) online).

Supplementary Material

[Supplementary data](#) are available at *Molecular Biology and Evolution* online.

Acknowledgments

The authors would like to thank Chong Chu for his advice on L1 hallmarks and are grateful to the anonymous reviewers whose comments have helped to improve the manuscript. This work was supported by the National Human Genome Research Institute (NHGRI) (Grant Nos. R01 HG010040, U01 HG010961, and U41 HG010972).

Data Availability

The genome assembly and expression data are openly available at the urls described in the article. The commands and code used for processed pseudogene are available at <https://github.com/xfengnefx/PPGfinder> (last accessed February 23, 2021).

References

- Chang H, Lim J, Ha M, Kim VN. 2014. TAIL-seq: genome-wide determination of poly(A) tail length and 3’ end modifications. *Mol Cell*. 53(6):1044–1052.
- Cheetham SW, Faulkner GJ, Dinger ME. 2020. Overcoming challenges and dogmas to understand the functions of pseudogenes. *Nat Rev Genet*. 21(3):191–201.
- Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods*. 18(2):170–175.
- Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437:69–87.
- Cooke SL, Shlien A, Marshall J, Pipinikas CP, Martincorena I, Tubio JMC, Li Y, Menzies A, Mudie L, Ramakrishna M, et al. 2014. Processed pseudogenes acquired somatically during cancer development. *Nat Commun*. 5:3644.
- Cost CJ, Feng Q, Jacquier A, Boeke JD. 2002. Human L1 element target-primed reverse transcription in vitro. *EMBO J*. 21(21):5899–5910.
- Crooks GE, Hon G, Chandonia J-M, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res*. 14(6):1188–1190.
- Esnault C, Maestre J, Heidmann T. 2000. Human LINE retrotransposons generate processed pseudogenes. *Nat Genet*. 24(4):363–367.
- Ewing AD, Ballinger TJ, Earl D, Harris CC, Ding L, Wilson RK, Haussler D, Broad Institute Genome Sequencing and Analysis Program and Platform. 2013. Retrotransposition of gene transcripts leads to structural variation in mammalian genomes. *Genome Biol*. 14(3):R22.

- Feusier J, Watkins WS, Thomas J, Farrell A, Witherspoon DJ, Baird L, Ha H, Xing J, Jorde LB. 2019. Pedigree-based estimation of human mobile element retrotransposition rates. *Genome Res.* 29(10):1567–1577.
- Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, et al. 2019. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* 47(D1):D766–D773.
- Garg S, Fungtammasan A, Carroll A, Chou M, Schmitt A, Zhou X, Mac S, Peluso P, Hatas E, Ghurye J, et al. 2020. Chromosome-scale, haplotype-resolved assembly of human genomes. *Nat Biotechnol.* <http://dx.doi.org/10.1038/s41587-020-0711-0>.
- Gordon D, Huddleston J, Chaisson MJP, Hill CM, Kronenberg ZN, Munson KM, Malig M, Raja A, Fiddes I, Hillier LW, et al. 2016. Long-read sequence assembly of the gorilla genome. *Science* 352(6281):aae0344.
- Grandi FC, Rosser JM, An W. 2013. LINE-1-derived poly(A) microsatellites undergo rapid shortening and create somatic and germline mosaicism in mice. *Mol Biol Evol.* 30(3):503–512.
- Hehir-Kwa JY, Marschall T, Kloosterman WP, Francioli LC, Baaijens JA, Dijkstra LJ, Abdellaoui A, Koval V, Thung DT, Wardenaar R, et al. 2016. A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. *Nat Commun.* 7:12989.
- Hormozdiari F, Konkel MK, Prado-Martinez J, Chiatante G, Herraes IH, Walker JA, Nelson B, Alkan C, Sudmant PH, Huddleston J, et al. 2013. Rates and patterns of great ape retrotransposition. *Proc Natl Acad Sci U S A.* 110(33):13457–13462.
- Kazazian HH. 2011. Mobile DNA transposition in somatic cells. *BMC Biol.* 9(1):62–64.
- Kronenberg ZN, Fiddes IT, Gordon D, Murali S, Cantsilieris S, Meyerson OS, Underwood JG, Nelson BJ, Chaisson MJP, Dougherty ML, et al. 2018. High-resolution comparative analysis of great ape genomes. *Science* 360(6393):eaar6343.
- Li H, Feng X, Chu C. 2020. The design and construction of reference pangenome graphs with minigraph. *Genome Biol.* 21(1):265.
- Mandal PK, Ewing AD, Hancks DC, Kazazian HH Jr. 2013. Enrichment of processed pseudogene transcripts in L1-ribonucleoprotein particles. *Hum Mol Genet.* 22(18):3730–3748.
- Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H. 2005. Emergence of young human genes after a burst of retroposition in primates. *PLoS Biol.* 3(11):e357.
- Mathews LM, Chi SY, Greenberg N, Ovchinnikov I, Swergold GD. 2003. Large differences between LINE-1 amplification rates in the human and chimpanzee lineages. *Am J Hum Genet.* 72(3):739–748.
- McLain AT, Carman GW, Fullerton ML, Beckstrom TO, Gensler W, Meyer TJ, Faulk C, Batzer MA. 2013. Analysis of western lowland gorilla (*Gorilla gorilla gorilla*) specific Alu repeats. *Mob DNA.* 4(1):26.
- Molineris I, Sales G, Bianchi F, Di Cunto F, Caselle M. 2010. A new approach for the identification of processed pseudogenes. *J Comput Biol.* 17(5):755–765.
- Navarro FCP, Galante PAF. 2015. A genome-wide landscape of retrocopies in primate genomes. *Genome Biol Evol.* 7(8):2265–2275.
- O'Grady T, Wang X, Höner Zu Bentrup K, Baddoo M, Concha M, Flemington EK. 2016. Global transcript structure resolution of high gene density genomes through multi-platform data integration. *Nucleic Acids Res.* 44(18):e145.
- Ostertag EM, Goodier JL, Zhang Y, Kazazian HH Jr. 2003. SVA elements are nonautonomous retrotransposons that cause disease in humans. *Am J Hum Genet.* 73(6):1444–1451.
- Ostertag EM, Kazazian HH. 2001a. Biology of mammalian L1 retrotransposons. *Annu Rev Genet.* 35(1):501–538.
- Ostertag EM, Kazazian HH Jr. 2001b. Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition. *Genome Res.* 11(12):2059–2065.
- Pavlíček A, Paces J, Elleder D, Hejnar J. 2002. Processed pseudogenes of human endogenous retroviruses generated by LINES: their integration, stability, and distribution. *Genome Res.* 12(3):391–399.
- Pink RC, Wicks K, Caley DP, Punch EK, Jacobs L, Francisco Carter DR. 2011. Pseudogenes: pseudo-functional or key regulators in health and disease? *RNA* 17(5):792–798.
- Podlaha O, Zhang J. 2009. Processed pseudogenes: the “fossilized footprints” of past gene expression. *Trends Genet.* 25(10):429–434.
- Poliseno L, Marranci A, Pandolfi PP. 2015. Pseudogenes in human cancer. *Front Med.* 2:68.
- Richardson SR, Doucet AJ, Kopera HC, Moldovan JB, Garcia-Perez JL, Moran JV. 2015. The influence of LINE-1 and SINE retrotransposons on mammalian genomes. *Mob DNA III.* 30:1165–1208.
- Richardson SR, Gerdes P, Gerhardt DJ, Sanchez-Luque FJ, Bodea G-O, Muñoz-Lopez M, Jesuadian JS, Kempen M-JHC, Carreira PE, Jeddeloh JA, et al. 2017. Heritable L1 retrotransposition in the mouse primordial germline and early embryo. *Genome Res.* 27(8):1395–1405.
- Schrider DR, Navarro FCP, Galante PAF, Parmigiani RB, Camargo AA, Hahn MW, de Souza SJ. 2013. Gene copy-number polymorphism caused by retrotransposition in humans. *PLoS Genet.* 9(1):e1003242.
- Seo J-S, Rhie A, Kim J, Lee S, Sohn M-H, Kim C-U, Hastie A, Cao H, Yun J-Y, Kim J, et al. 2016. De novo assembly and phasing of a Korean human genome. *Nature* 538(7624):243–247.
- Sisu C, Muir P, Frankish A, Fiddes I, Diekhans M, Thybert D, Odom DT, Flicek P, Keane TM, Hubbard T, et al. 2020. Transcriptional activity and strain-specific history of mouse pseudogenes. *Nat Commun.* 11(1):3695.
- Vollger MR, Dishuck PC, Sorensen M, Welch AE, Dang V, Dougherty ML, Graves-Lindsay TA, Wilson RK, Chaisson MJP, Eichler EE. 2019. Long-read sequence and assembly of segmental duplications. *Nat Methods.* 16(1):88–94.
- Wei W, Gilbert N, Ooi SL, Lawler JF, Ostertag EM, Kazazian HH, Boeke JD, Moran JV. 2001. Human L1 retrotransposition: cis preference versus trans complementation. *Mol Cell Biol.* 21(4):1429–1439.
- Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhaj J, Billis K, Cummins C, Gall A, Girón CG, et al. 2018. Ensembl 2018. *Nucleic Acids Res.* 46(D1):D754–D761.
- Zhang Z, Carriero N, Zheng D, Karro J, Harrison PM, Gerstein M. 2006. PseudoPipe: an automated pseudogene identification pipeline. *Bioinformatics* 22(12):1437–1439.