5-2023

Identification of SVA-Mediated 3’ and 5’ Transductions in Human Genome Sequences

Emily C. Golba

Identification of SVA-Mediated 3' and 5' Transductions in Human Genome Sequences

Emily C. Golba

Abstract

Transposable elements (TEs) are sequences of DNA that can move, or transpose, within a genome. Retrotransposons are TEs that propagate via a "copy and paste" mechanism where the elements are transposed to a new genomic location via an RNA intermediate. Short interspersed element (SINE)-VNTR-*Alu* elements (SVAs) are non-autonomous retrotransposons that use long interspersed element 1 proteins to mobilize. SVAs are currently active in the human genome and often are characterized by the mobilization of sequences adjacent to the 3' and 5' ends of insertions, known as transduction events (TDs). TDs were a focus of this study due to their contribution to genome expansion, exon shuffling, and gene duplication. We analyzed 547 SVA elements across 35 genetically diverse individuals from the Human Genome Structural Variation Consortium. We identified 187 full-length elements, 12 of which carried 3' TDs, 13 harbored 5' TDs, and five held both 3' and 5' TDs. Four TDs contained exons of genes, including *HGSNAT*, *UQCRC1*, and *RP11-137H2.4*. Utilizing TDs, we screened the dataset for source elements, i.e., active TEs capable of producing offspring, by querying the human reference genome (GRCh38) using BLAT to identify the origin of the TDs and SVA source elements. We found that Chromosomes 3, 10, and 11 showed enrichment for active SVA elements. Novel TDs were identified and can be used to study genome expansion due to their ability to take additional sequences from each genomic location. Source elements can cause insertions that drive disease and genome expansion, exemplifying their significance.

Introduction

Transposable elements

Transposable elements (TEs) are genetic elements that have the ability to move, or "transpose," within a genome. TEs can have significant effects on the genome. They can disrupt genes or regulatory regions, cause gene duplications, alter gene expression patterns, and contribute to genome evolution (Ayarpadikannan et al., 2014). However, TEs can also have negative effects, such as causing mutations that lead to diseases such as leukemia and Fukuyama muscular dystrophy or altering the regulation of genes in a harmful way, i.e., nonsense-mediated decay (Rebollo et al., 2012). There are two main types of TEs, class 1: retrotransposons and class 2: DNA transposons. TEs are highly prevalent in the human genome, comprising over 50% of DNA content. Of this, the majority are retrotransposons, which make up about 47% of the genome, while DNA transposons account for only about 3% (Lander et al., 2001).

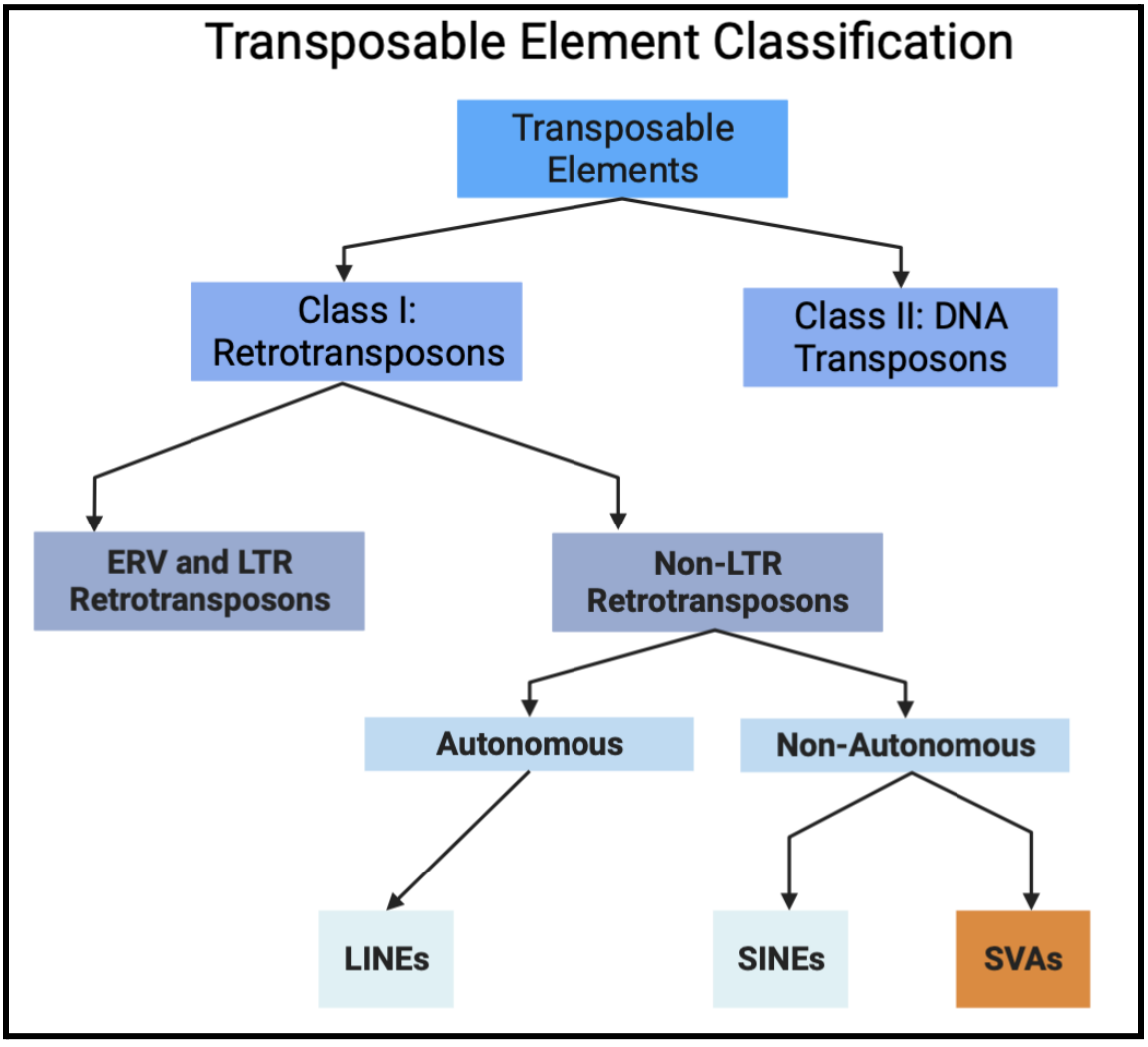


Figure 1: Transposable element classification. TEs are Categorized based on transposition mechanisms as DNA transposons or retrotransposons. LINEs are the only active autonomous retrotransposon in the human genome. SINEs and SVAs are types of currently propagating non-autonomous retrotransposons.

DNA transposons move by a "cut-and-paste" mechanism, where the transposon is excised from one location in the genome and inserted into another location (Pace & Feschotte, 2007). The last evidence of DNA transposons being active in the human genome was ~37 million years ago (Pace & Feschotte, 2007). Retrotransposons are typically divided into two main groups, long terminal repeat (LTR) retrotransposons and non-LTR retrotransposons. LTR retrotransposons have long terminal repeats at both ends of their DNA sequence, while non-LTR retrotransposons do not. PolyA-tail and target site duplications (TSDs) are features that characterize non-LTR retrotransposons. TSDs are identical sequences that are found on both ends of a retrotransposon, which arise from the insertion of the retrotransposon (Pray, 2008). Non-LTR retrotransposons can be further classified into two types, autonomous and non-autonomous elements. An autonomous non-LTR retrotransposon is a TE that can move within the human genome using its own enzymatic machinery without requiring the help of other TEs. Long interspersed element 1s (L1s) are autonomous non-LTR retrotransposons that are characterized by two open reading frames (ORFs) that encode proteins required for retrotransposition in mammalian genomes. ORF1 encodes a nucleic acid-binding protein, while ORF2 encodes a polyprotein that has endonuclease and reverse transcriptase activities. These proteins work together to insert a copy of the retrotransposon into a new location in the genome via a "copy-and-paste" mechanism (Cordaux & Batzer, 2009). Non-autonomous retrotransposons include short interspersed elements (SINEs) and SINE-VNTR-Alu (SVAs), which rely on the enzymatic machinery of autonomous elements to move within the genome (Pray, 2008).

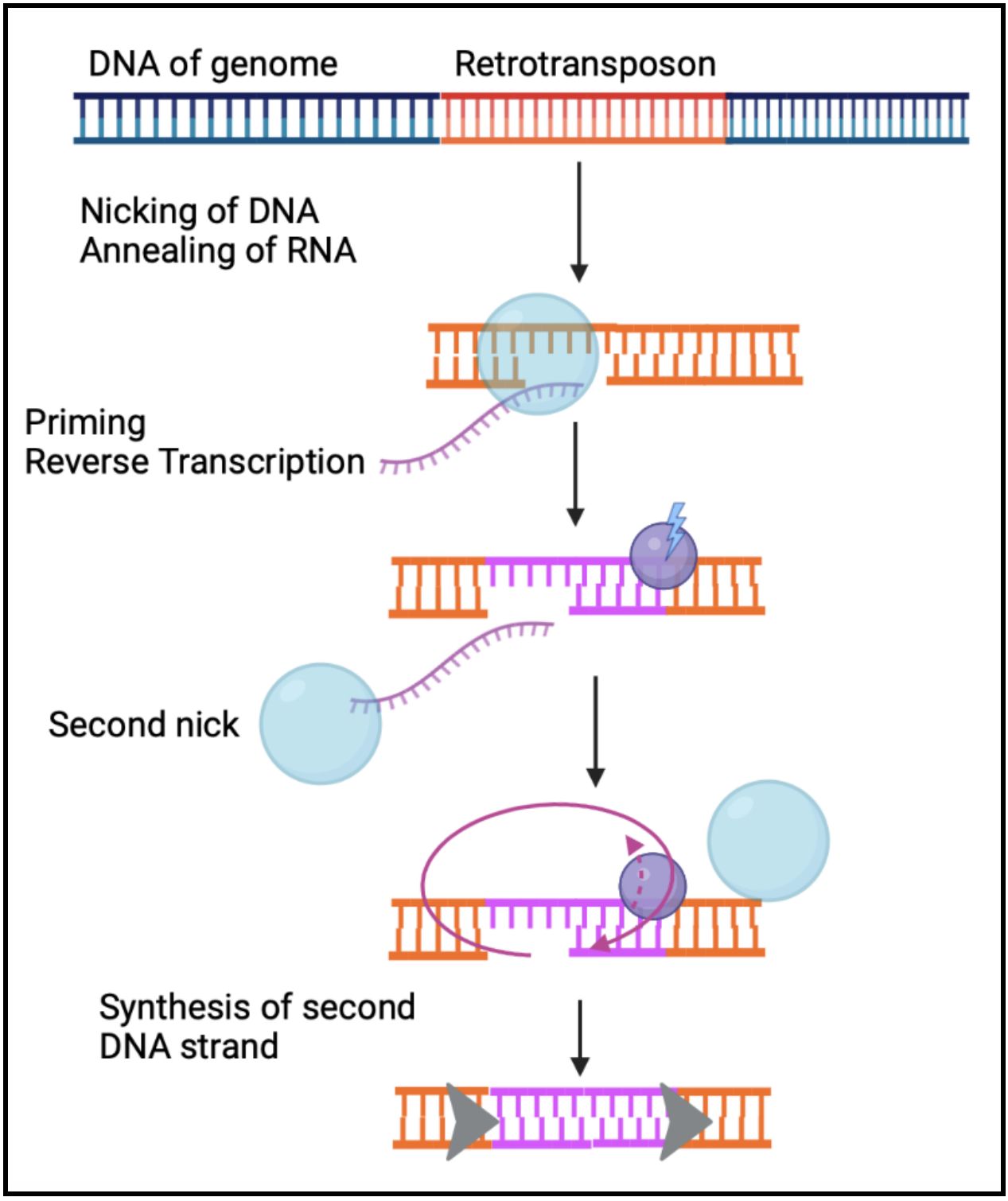


Figure 2: Reverse transcription mechanism for retrotransposons. Blue circles represent ORF2p, the purple circle is the unknown protein responsible for the formation of the second nick. Gray triangles represent the TDS. The pink represents the new insert.

Non-LTR Retrotransposons move by a "copy-and-paste" mechanism via an RNA intermediate. An endonuclease protein will nick the target site to initiate reverse transcription. Within the nucleus, the RNA molecule serves as a template for reverse transcription to produce a complementary DNA strand. This process is carried out by the reverse transcriptase enzyme, which is encoded by the retrotransposon itself or, in the case of a non-autonomous element, it is provided by L1 proteins. The resulting DNA intermediate can then integrate into the genome inside the nucleus (Figure 2) (Bourque et al., 2018).

LINEs

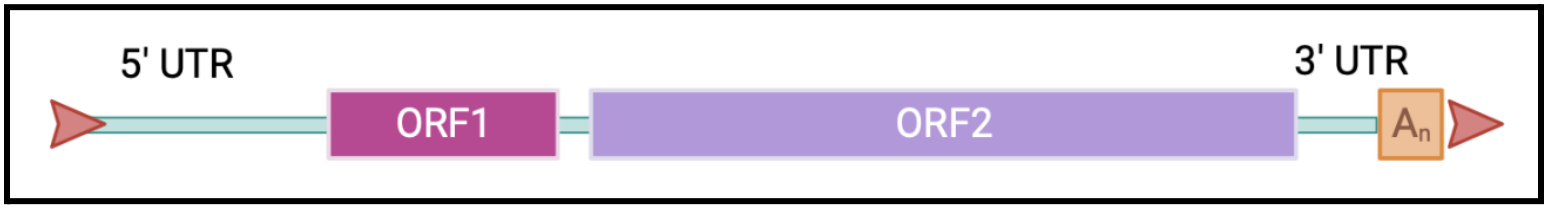


Figure 3: LINE 1 structure. 5’ untranslated region (5’ UTR), open reading frame 1 (ORF1), ORF2, 3’ UTR, polyA tail (A_n), target site duplications (red arrow)

The only active autonomous non-LTR retrotransposon is L1, the most proliferative TE in the human genome, accounting for approximately 20% of DNA content (Lander et al., 2001). L1s have played an essential role in shaping the evolution of the genomes of many eukaryotic organisms, including humans, and are still active in some lineages today (Beck et al., 2011). There are ~500,000 copies of L1 in the human haploid genome, but many are no longer active. It is estimated that each human has ~100 active L1s (Bourque et al., 2018). L1 elements are composed of an upstream promoter region (5’ UTR) that drives transcription and a downstream region that contains the coding sequences required for retrotransposition. In the coding region, there are two reading frames, ORF1 and ORF2, which are required for the retrotransposition process (Beck et al., 2011). L1 elements can cause genomic instability, and they can also

contribute to genetic diversity and expansion. This can be from chromosomal deletion, 5' inversions, or even having significantly longer and more frequent polyA- tails (Symer et al., 2002)

In contrast to autonomous retrotransposons, which carry all the necessary machinery for transposition, non-autonomous retrotransposons rely on other TEs, typically autonomous retrotransposons, for their mobility. When a non-autonomous retrotransposon is transcribed, it produces an RNA molecule that is then reverse-transcribed into DNA and inserted into the genome by the machinery of an autonomous retrotransposon i.e. L1 (Hancks, Kazazian, 2016). Although non-autonomous retrotransposons do not encode the proteins necessary for their own transposition, they can still have significant effects on genome expansion by creating new copies of themselves in different locations of the genome (Havecker et al., 2004). Examples of non-autonomous retrotransposons include SINEs, and SVA elements.

SINE-VNTR-Alus

Figure 4: Full-length SVA Structure. It is composed of several domains in order from the 5' end: 1) CCCTCT hexameric repeats, 2) an *Alu*-like domain with two antisense *Alu* fragments, with black arrows indicating directionality, 3) a Variable Number Tandem Repeat domain derived from the ancestral SVA2 element, 4) a SINE-R domain that shares homology with the 3'end of the HERV-K10 env gene, and ends with a polyA tail (An), with the whole SVA insertion flanked by TSDs (gray arrows).

SVA retrotransposons are non-LTR retrotransposons found in the genomes of humans and other primate species, including gorillas, bonobos, gibbons, orangutans, and common

chimpanzees (Hancks, Kazazian, 2010). From 5' to 3', full-length SVA elements consist of the hexameric region (CCCTCT), an *Alu-like* domain, a Variable Number Tandem Repeat (VNTR) domain, SINE-R, and a poly-a tail (Shen et al., 1994). Full-length elements are ~2kb in length (Gianfrancesco et al., 2017). They are known to be highly variable in size, with different copies exhibiting different patterns of VNTR length, as well as variable flanking sequences known as TSD. TSDs arise as a result of staggered double-strand breaks at the target site and range from 4bp to 25bp (Linheiro et al., 2012). Gianfrancesco et al., 2019 found that younger SVA subfamilies are over-represented at zinc finger loci, and these subfamilies have a greater ability to regulate nearby gene expression due to longer VNTR length and higher GC content. Despite their relatively low copy number (~2700 copies) in the human genome, SVAs are a young and active class of TEs (Hancks, Kazazian, 2010). They are thought to be *trans*-mobilized by the use of L1 machinery to insert into the human genome. SVAs have been implicated in various human diseases and disorders, including Parkinson's disease, Fukuyama muscular dystrophy, and Leukemia (Makino et al., 2007)(Kobayashi et al., 1998)(Takasu et al., 2007). They have also been found to regulate gene expression, with SVAs containing regulatory elements that can affect nearby genes depending on their proximity. This is shown from data of a *PARK7* SVA element where it was determined the element contained multiple regulatory domains (Quinn, Bubb, 2014). SVAs are a relatively recent addition to the hominoid genome during its ~6 My evolution and are still being studied to understand how they arose and their biological functions (Damert et al., 2009).

Transduction events

SVA-mediated transductions (TDs) refer to a process by which SVA retrotransposons can facilitate the transfer of genetic material from one part of the genome to another, often between different or non-homologous chromosomes. This process can result in the generation of novel gene fusions or the duplication and amplification of existing genes (Damert, 2022). The SVA-mediated TDs can sometimes include protein-coding genes or regulatory elements. Around 16% of human SVA elements are involved in sequence transduction events (Hancks, Kazazian, 2010).

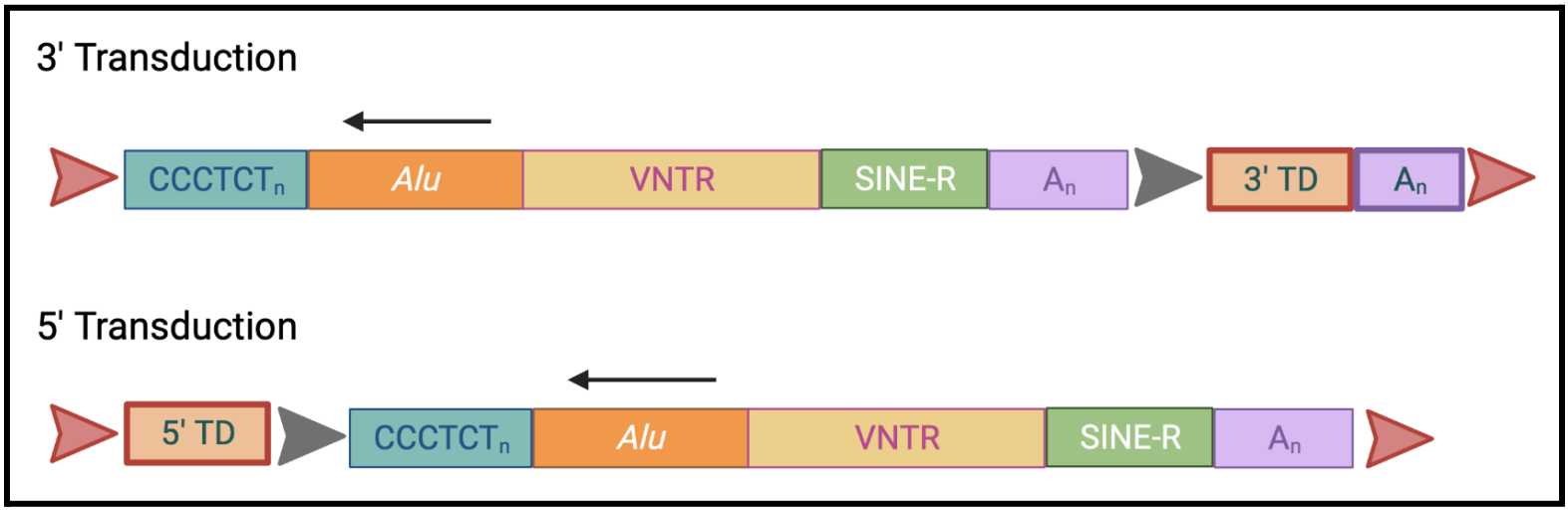


Figure 5: Structure of SVA with 3' and 5' transduction. Target site duplications (red arrow), initial target site duplications (gray arrow)

A 3' transduction (3' TD) occurs when SVA transcription bypasses its initial termination site due to a weak polyadenylation signal. The 3' TD is then retrotransposed while carrying sequences from the previous location. In a 5' transduction (5' TD) event, an upstream transcriptional start site is used to transcribe the sequence to a downstream SVA element with an

additional 5’ sequence (Hancks, Kazazian, 2010). Both transductions will have TSDs flanking them as they are hallmarks of the integration process (Gianfrancesco et al., 2017). This research studied SVAs because they are currently active in the human genome and can mobilize sequences adjacent to the 3’ and 5’ ends of insertions, known as TDs. TDs were a focus due to their contribution to genome expansion, exon shuffling, and gene duplication (Hancks, Kazazian, 2010).

Source elements

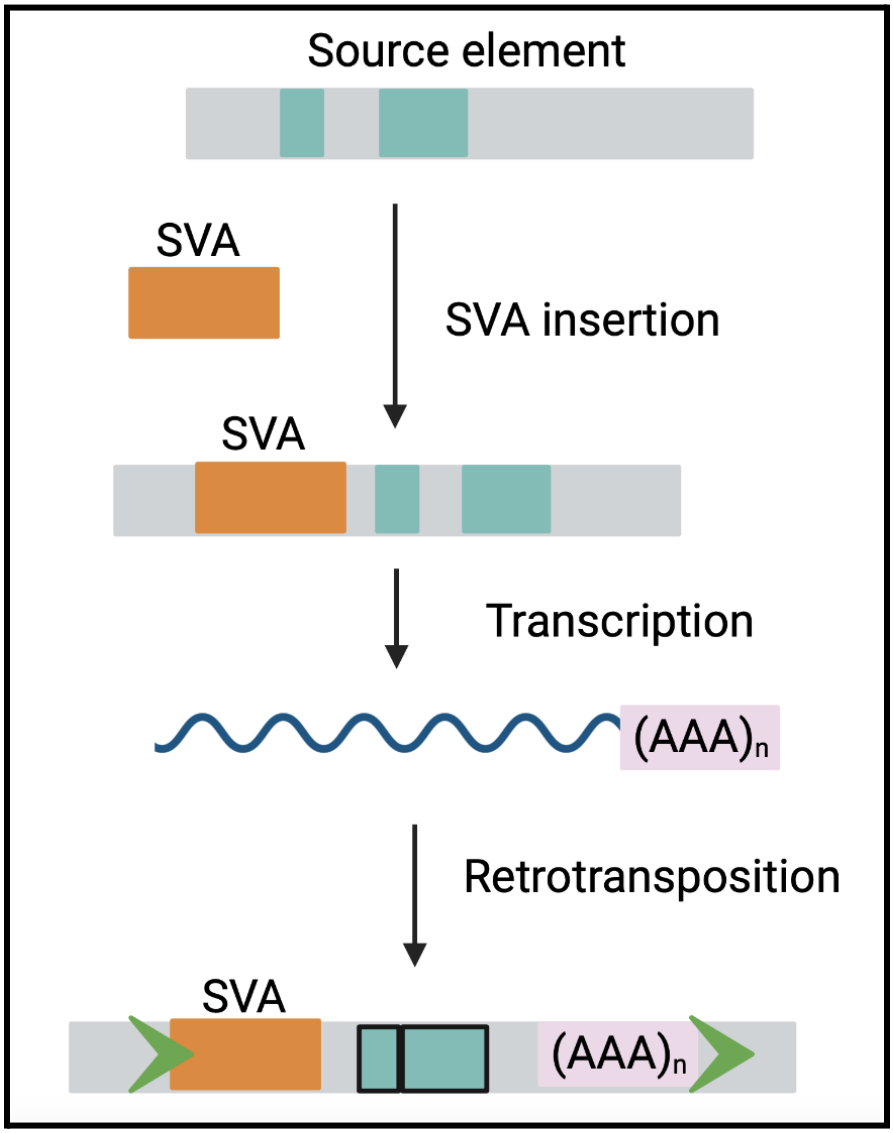


Figure 6: Source element schematic. An active SVA element was inserted upstream of a gene locus and transcribed the full-length gene sequence through the process of transcriptional transduction. RNA processing machinery removed the intron of the gene during retrotransposition. The resulting intronless gene sequence, along with the SVA element, retrotransposed into new genomic locations.

Source elements refer to the original or ancestral TE that initiates the transposition process by producing an RNA intermediate, which is then reverse-transcribed and integrated into a new genomic location. They are capable of producing offspring, and they are often able to be identified by comparing the TDs' TSD to the source element (Hancks, Kazazian, 2010). Another way you can identify them is by examining the sequence similarity between the TD in the original location and the TD in the new location and determine whether ans SVA is present. If there is a high degree of sequence similarity between TD and reference sequence, it suggests that the TD was likely initiated by that source element (Damert, 2022).

Methods

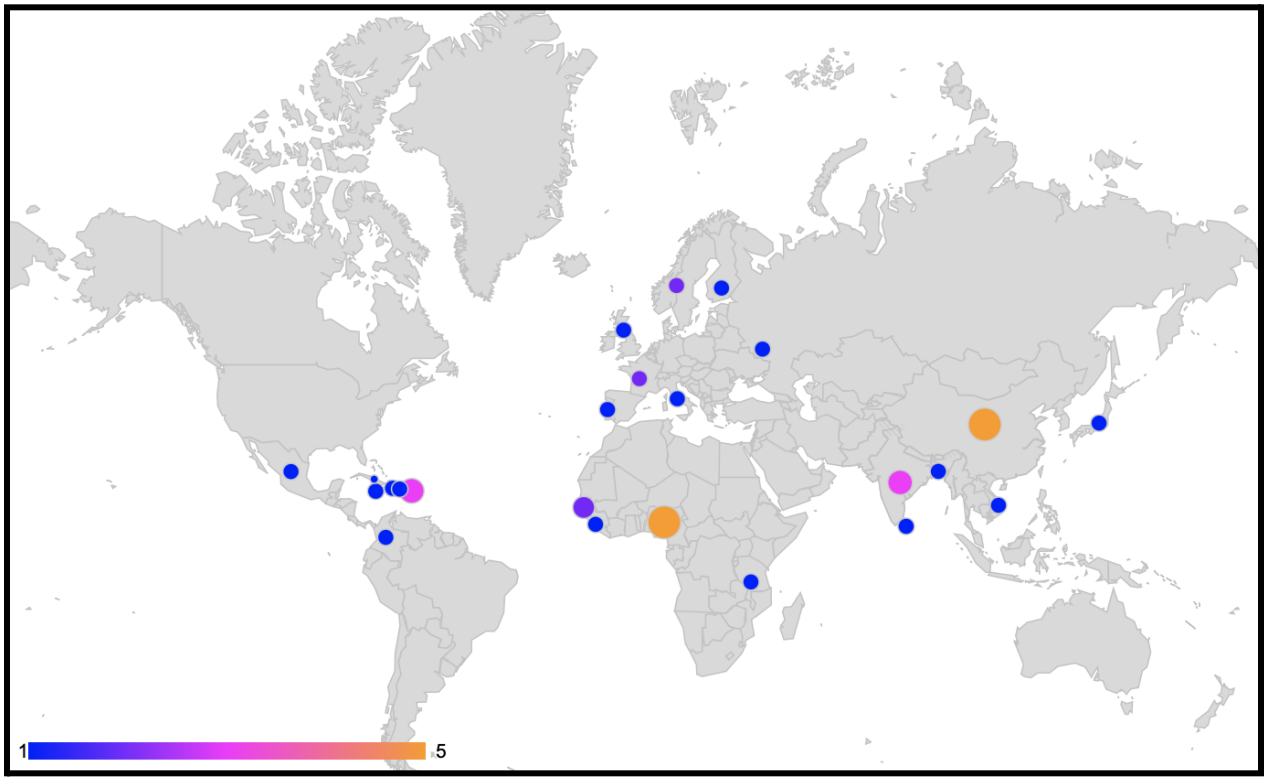
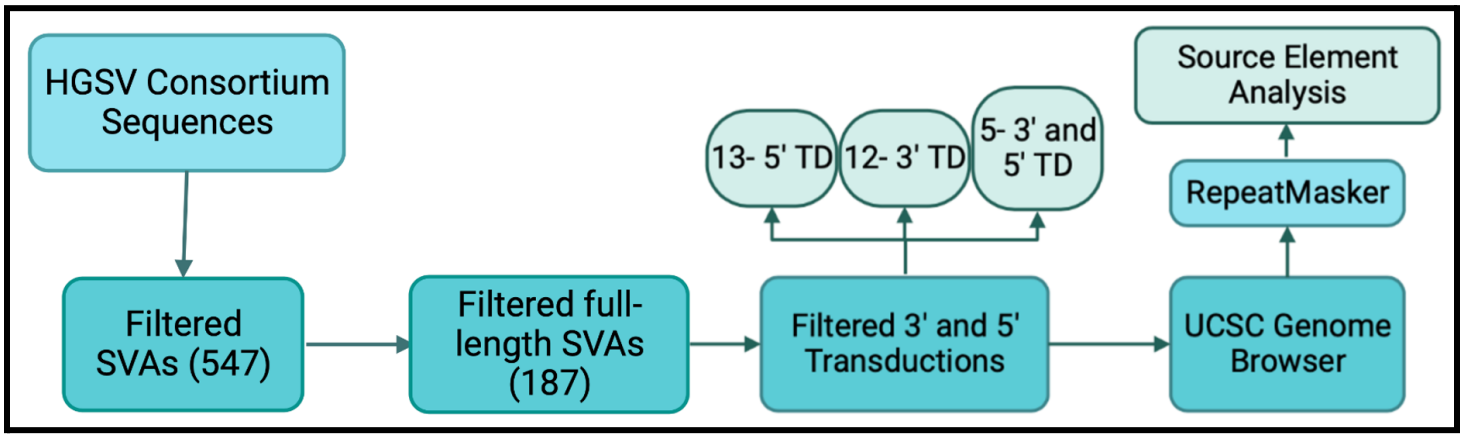


Figure 7: Ethnicity of sample individuals. There are 35 individuals represented by the points. The larger the point means more samples from that location.

The SVA samples are from a freeze4 polymorphic dataset of 35 individuals generated by the Human Genome Structural Variation Consortium (HGSV) (Ebert et al., 2021). Long-read technology and single-cell template strand sequencing were used to generate fully phased diploid genome assemblies. There were 32 individuals in the dataset that were unrelated, and these individuals represented 25 diverse human populations. Three children included in the study were part of parent-child trios whose ethnicities are Southern Han Chinese, Puerto Rican, and Yoruba. The genome assemblies were created in order to discover previously unknown structural variations. This made it possible to determine the frequency of SVA insertions within the population and uncover potential disease loci by querying the sequences in BLAT through the UCSC Genome Browser (Kent et al., 2002).

Methodology



Flow Chart 1: Schematic of the methods for filtering the data.

The HGSV sequences contained other structural variants other than SVAs. These were sorted into different categories based on the structural variants, and the sequences with SVAs were used. There were 547 loci that contained SVA elements that were filtered further. The data

was then filtered to contain only full-length elements, meaning the loci had to contain the hexameric region, an *Alu-like* domain, a VNTR domain, SINE-R, and a poly-a tail.

There were 187 full-length SVAs that were then screened for 3' and 5' TDs. This was done by first aligning the sequences in AliView (Larsson, 2014) with the use of Muscle (Edgar, 2021) and then determining whether there was an additional sequence upstream of the hexameric region or after the polyA-tail. These loci with TDs were then compared with the human reference genome (GRCh38) in the UCSC Genome Browser (Kent et al., 2002). To do this, TD sequences were queried using BLAT, and they were aligned with the reference sequences. This allowed for visualization of the location of the loci and the ability to identify source elements of the 3' and 5' TDs by determining if there was an SVA present in GRCh38. Repeating elements in the reference genome were identified by RepeatMasker (Smit et al., 2013-2015), which confirmed if an SVA was present and would be used for further analyses i.e., TD length, which samples it was present in, and to determine the location of source elements.

Results

The HGSV Consortium human genome sequences were analyzed for SVA elements. Out of 547 SVA sequences, we found 187 full-length elements. After filtering, 30 TDs were identified and were present across all samples (Figure 8). The TD count varies between individuals, but the minimum was two TDs, and the maximum was 9 TDs (Figure 9). There were also ten TDs that only occurred in one individual and four TDs that occurred in over 20 individuals (Figure 8). Out of the 30 identified loci, 12 carried 3' TDs, 13 harbored 5' TDs, and five held both 3' and 5' TDs. On average, the 3' TDs were longer but less frequent than the 5' TDs. The length of 3' TDs ranged from 20bp to 1396bp, while the length of 5' TDs ranged from

28bp to 992bp. The average length of 3' TDs was 360bp, while the average length of 5' TDs was 141bp (Figure 10). Indicating that the 3' TDs were longer than the 5' TDs.

Four SVA loci contained TDs harboring exons of genes (*RP11-137H2.4, HGSNAT, UQCRC1)*. Since the HGSV Consortium freeze4 data includes information for both alleles of the diploid genome, it is possible to determine the zygosity of individuals for a TD. Zygosity analysis was performed on the data for the four TD events that carried exons to see how common the TD was. Out of these, a total of 11 individuals were homozygous, and 18 were heterozygous for TDs carrying exon 1 out of 2 in the *RP11-137H2.4* gene. However, since two SVA loci in the dataset had the same donor location, the same exon was listed twice in Table 1. Consequently, one locus had four heterozygous individuals with no homozygotes, while another had 14 heterozygotes and 11 homozygotes. Two individuals were heterozygous for a TD of exon 10 of 18 in the *HGSNAT* gene, and one individual was heterozygous for a TDwith exon 2 of 13 in the *UQCRC1* gene.

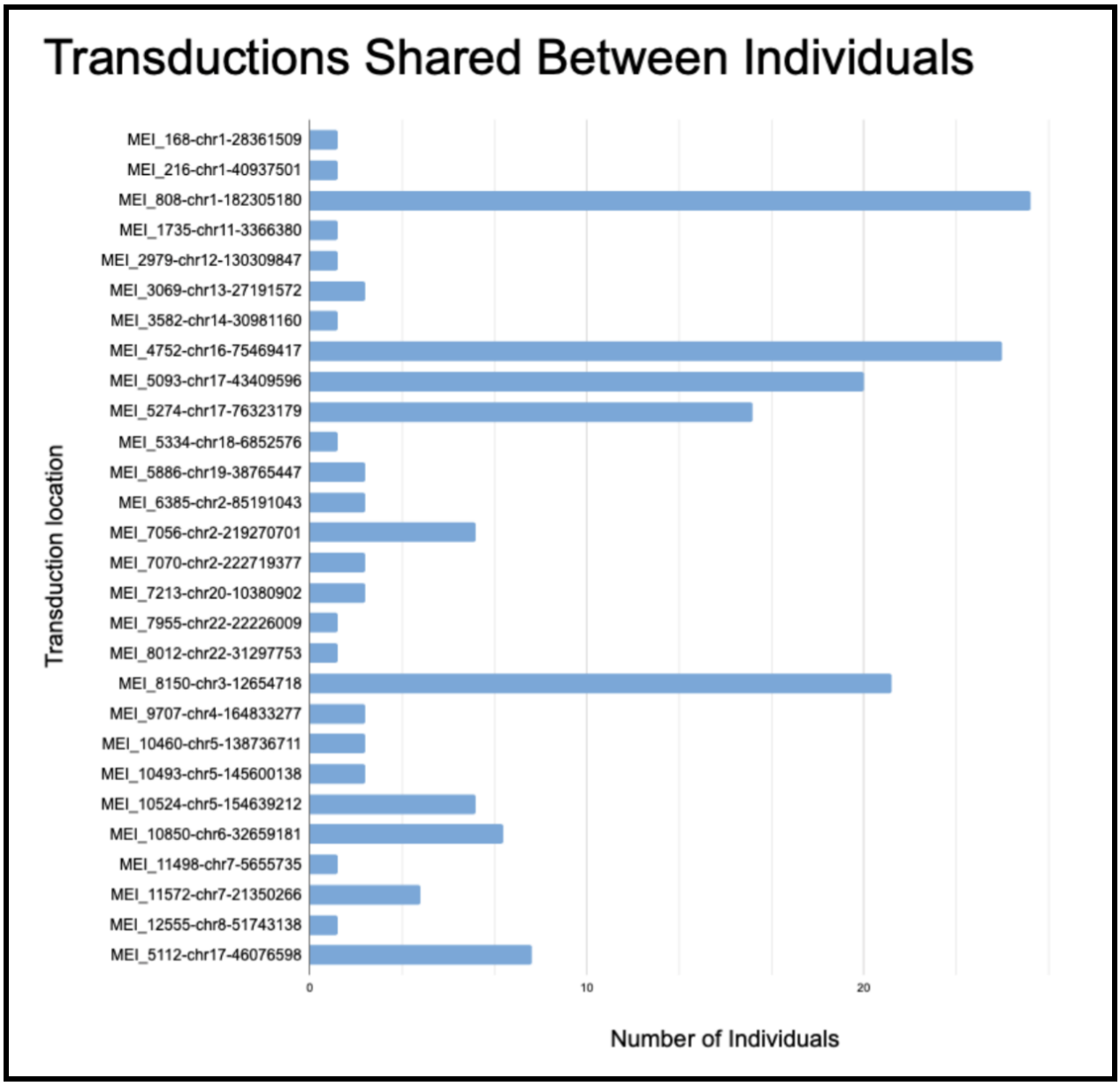


Figure 8: Transductions shared between individuals. There were ten singleton TDs and four TDs that were present in over 20 individuals.

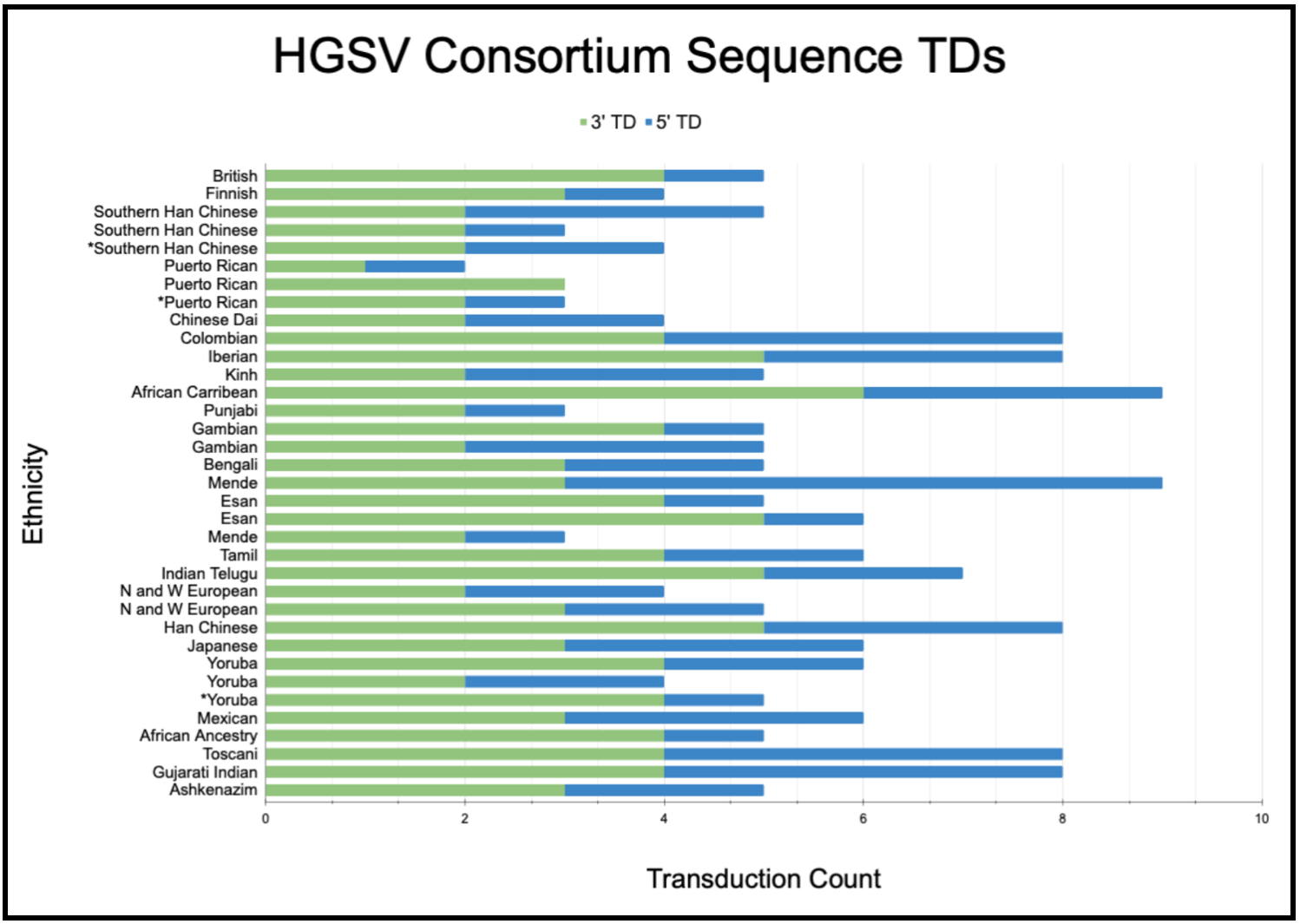


Figure 9: Count of 3' and 5' transductions for the 35 individuals. " *" denotes the child of the parent-child trios. The range of TDs is from 2 to 9 in each individual and is spread across all ethnicities.

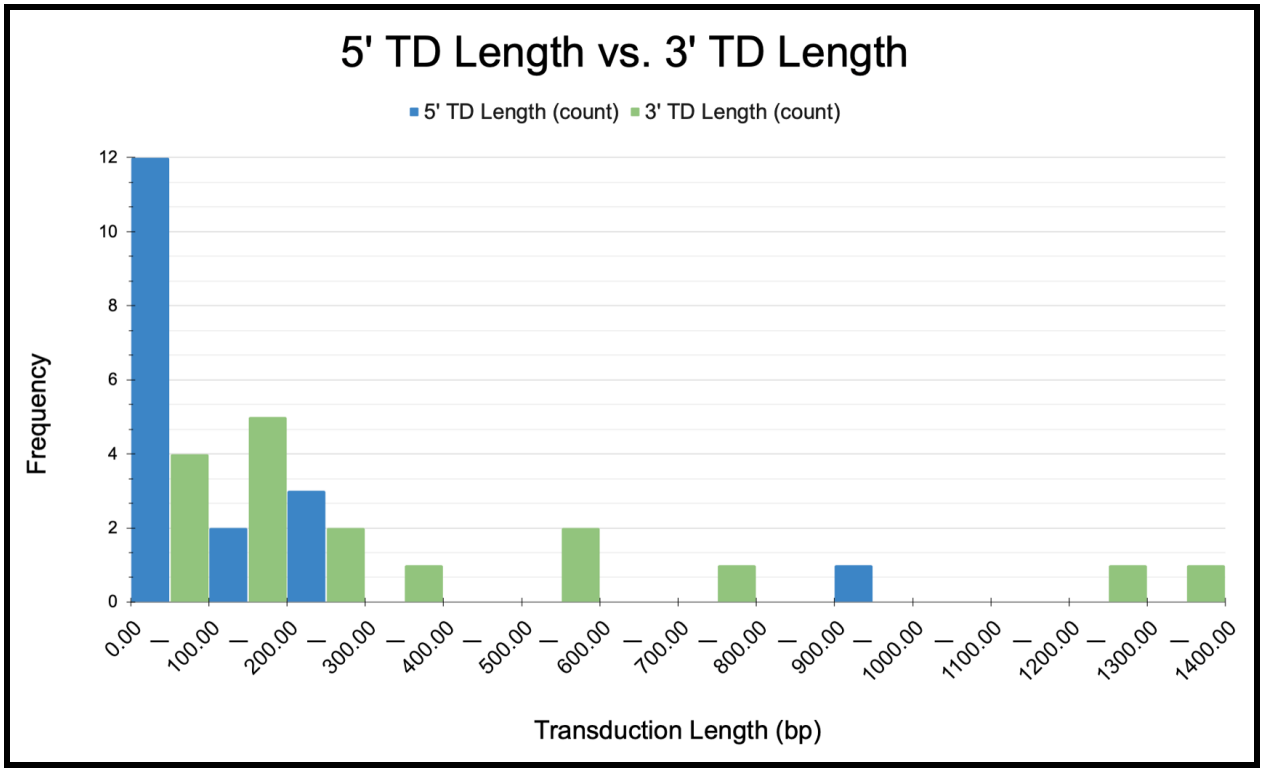


Figure 10: Length of 3' and 5' transductions in each locus. The range of TDs was from 20bp - 1396bp, and they were sorted in bins of 100. The 3' TDs were, on average, longer (360bp).

Gene Name	Homozygous	Heterozygous
HGSNAT	0	2
UQCRC1	0	1
RP11-137H2.4	0	4
RP11-137H2.4	11	14

Table 1: Genes with exons present in TDs with their zygosity. One locus with the RP11-137H2.4 gene had four heterozygous individuals, another had 14 heterozygotes and 11 homozygotes. The HGSNAT gene locus had two individuals were heterozygous for a TD of exon 10 of 18. One individual was heterozygous for a TD with exon 2 of 13 in the UQCRC1 gene.

By utilizing TDs, we screened the dataset for source elements through the UCSC Genome Browser. The TDs were queried using BLAT and aligned with the human reference genome (GRCh38) to determine their similarities to identify their source. The reference sequence

was retrieved through the UCSC Genome browser by pulling the DNA from the We found 27 source elements for the 30 SVA loci since there were three occurrences where two SVA loci shared one source element. On chromosomes 1, 3, and 10, source elements had two offspring in the SVA loci data. Furthermore, TDs can appear multiple times at various positions in the genome through multiple transposition events. The source element for the *RP11-137H2.4* gene, which produced two of our SVA loci, is on chromosome 10. Through the visualization of the TD of the RP11-137H2.4 exon, we can see that the source element produced multiple offspring (Figure 11).

Additionally, we found that Chromosomes 3, 10, and 11 contained the most SVA elements involved in TDs (Figure 12). The TDs that are mediated by SVAs have the ability to hold whole exons and move them to a new location. Another source element that was identified for an SVA locus was one located on chromosome 8. This is the SVA that has mobilized the complete exon 10 of the HGSNAT gene to a new genomic location at chromosome 5 (Figure 13).

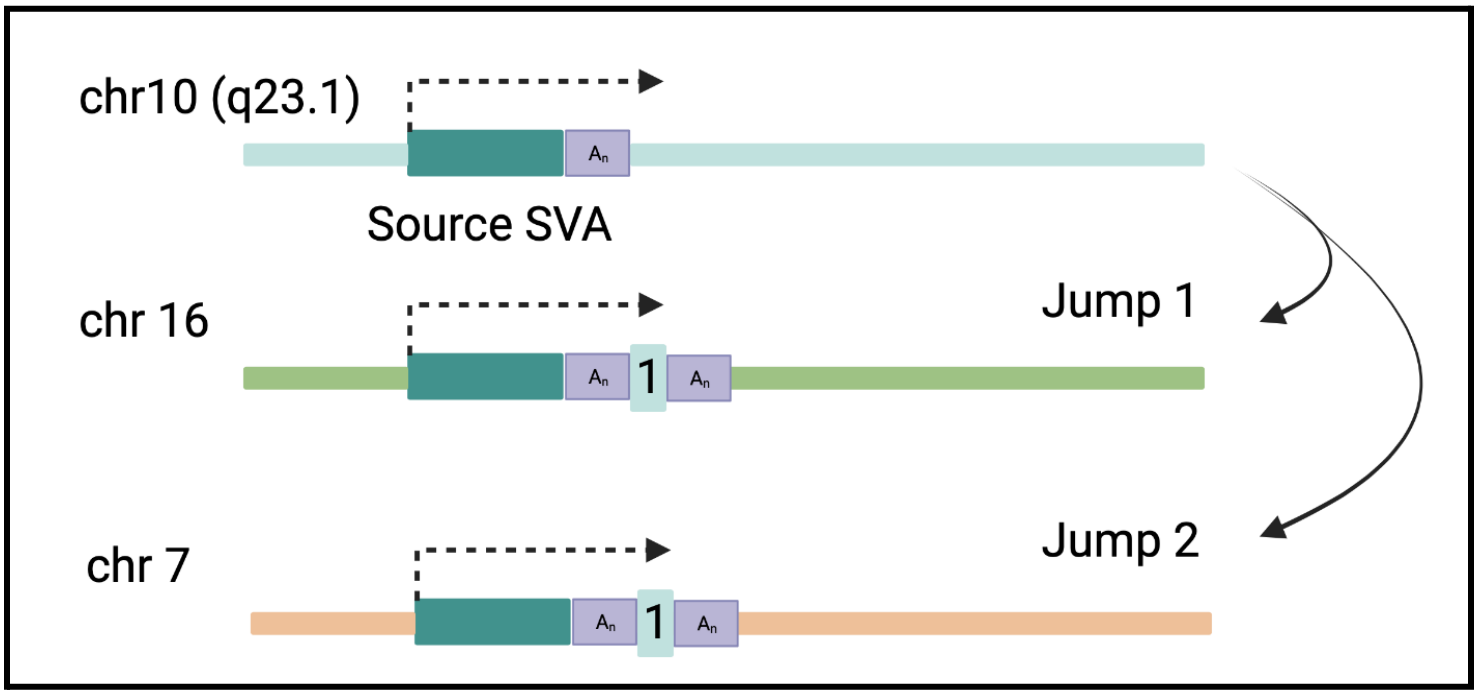


Figure 11: Source element tracing for the RP11-137H2.4 gene. The source element is on chromosome 10 and it produces two offspring with 3' TDs, one on chromosome 16 and the other on chromosome 7.

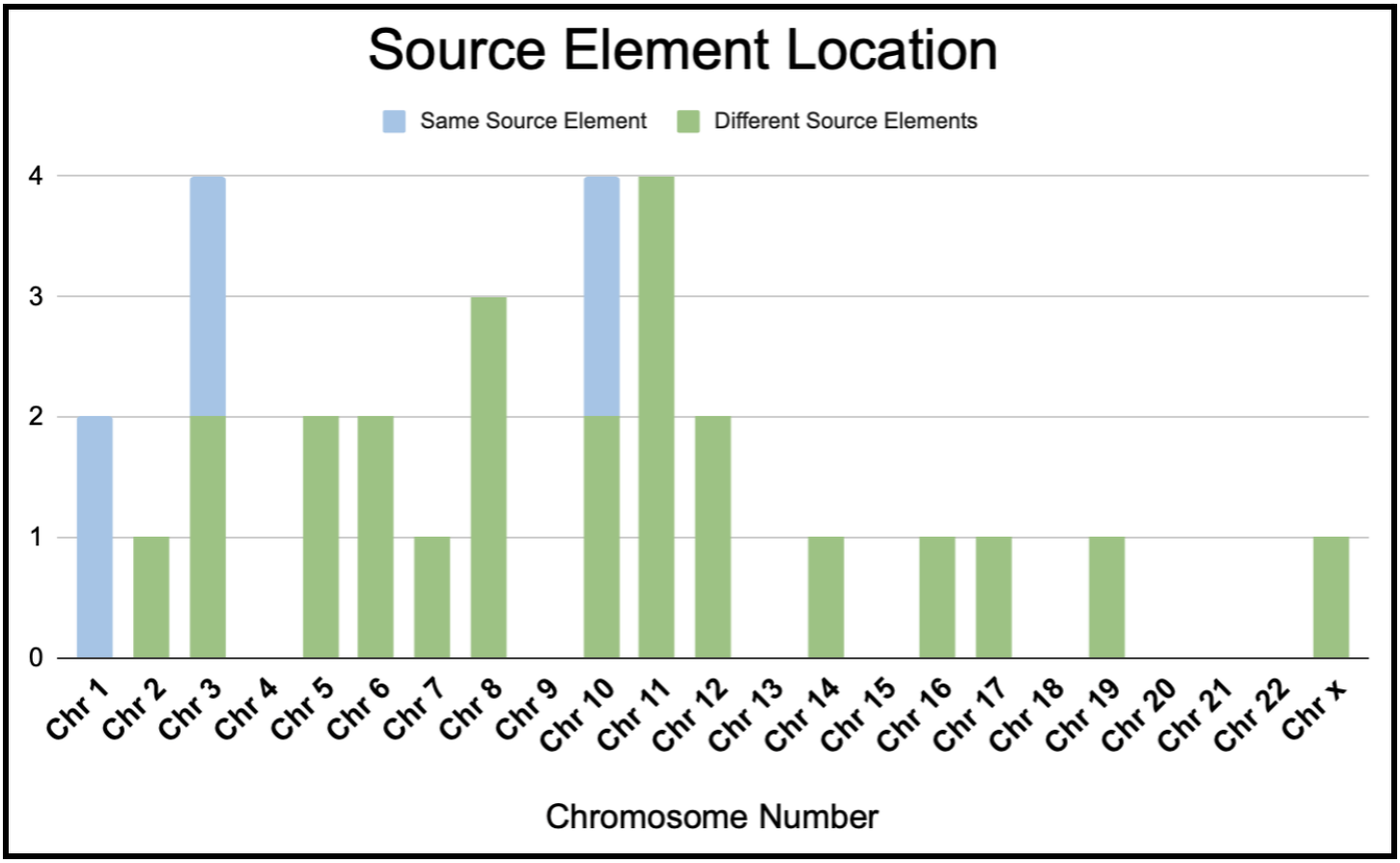


Figure 12: Chromosomal location of source elements. 27 source elements were identified with the most being on chromosomes 3, 10, and 11.

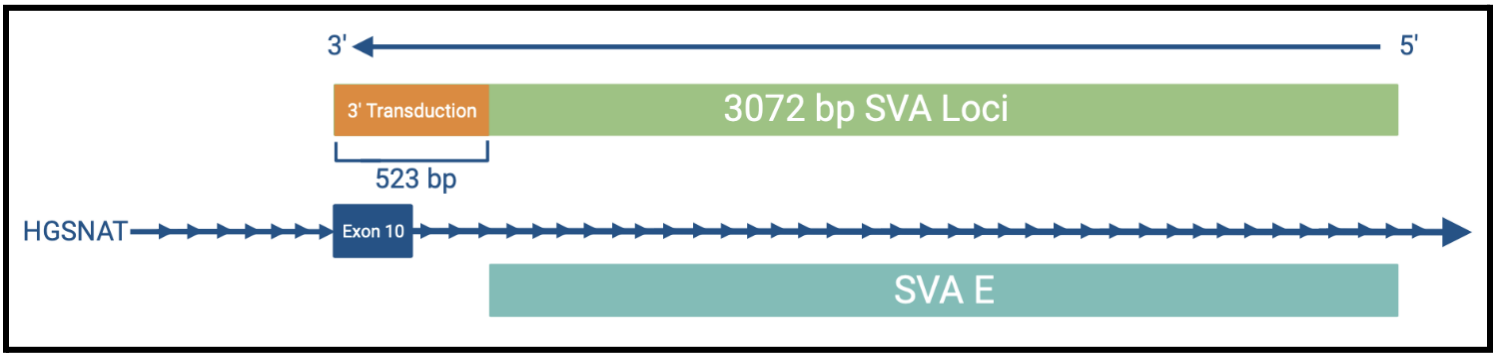


Figure 13: Source element example. A 523bp 3' TD holding an exon of the HGSNAT gene. The SVA sequence has the same length as the reference genome, indicating it is the source element.

Discussion

SVA retrotransposons show an evolutionary importance facilitated by 3' and 5’ TDs. SVA-mediated TDs are responsible for the formation of the AMAC1 gene family, which includes four copies in the human genome (Xing et al., 2006). TDs are facilitated by SVAs, which pick up on average 250 base pairs from a specific genomic location and then mediate the transposition of this sequence to a new location. Upstream transcriptional start sites and weak polyadenylation signals play a factor in this occurrence (Hancks & Kazazian, 2016). Out of 187 full-length elements, the study found 30 TDs that were mediated by SVAs. These events were not specific to any one population as they were found equally across all ethnicities. Not all SVA elements were identified however there were six novel 3’ TDs, seven novel 5’ TDs, and three novel 3’ and 5’ TDs compared to the dataset of Ebert et al., 2021. Four TDs contained exons of genes, including HGSNAT, UQCRC1, and RP11-137H2.4, with only the HGSNAT gene being identified in Ebert. The most common chromosomes that harbored source elements for TDs were chromosomes 3, 10, and 11.

Conclusion

This study aimed to identify TDs within SVA retrotransposons in the human genome. Through filtering and analysis of the HGSV Consortium data, 187 full-length SVA elements were screened for 3’ and 5’ TDs. A total of 30 Ts were identified across multiple individuals and populations. The presence of TDs in multiple individuals suggests that SVA-mediated TDs have occurred throughout human evolutionary history. Through the analysis, novel TDs, like the exon of the RP11-137H2.4 gene, were identified and can be used to study genome expansion due to their ability to take additional sequences from each genomic location. Source elements can cause

insertions that drive disease and genome expansion, exemplifying their significance. Overall, this study provides insight into the diversity and complexity of the human genome and highlights the importance of studying SVA mechanisms.

Future Directions

Further studies could investigate the potential impact of these TDs on human genome function and evolution. This could be done by studying different populations along with more samples.

Acknowledgments

This work is supported in part by the "Center of Biomedical Research Excellence (COBRE) in Human Genetics" 1P20GM139769, Clemson Creative Inquiry Program, and the National Institute of General Medical Sciences (NIGMS). Clemson University is acknowledged for the Palmetto cluster resource.

Special Regards

I would also like to acknowledge my PI: **Miriam K. Konkel,** and the members of the CI Tangible Genomics: **Ashley Kirby**, Mark Loftus, Ashtyn Hill, Gianni Martino, Kassandra Roemer, Olivia Duffy, and Gillian Barnard.

References

Ali, A., Han, K., & Liang, P. (2021). Role of Transposable Elements in Gene Regulation in the Human Genome. *Life (Basel, Switzerland)*, *11*(2), 118. https://doi.org/10.3390/life11020118

Ayarpadikannan, S., & Kim, H. S. (2014). The impact of transposable elements in genome evolution and genetic instability and their implications in various diseases. *Genomics & informatics*, *12*(3), 98–104. https://doi.org/10.5808/GI.2014.12.3.98

Beck, C. R., Garcia-Perez, J. L., Badge, R. M., & Moran, J. V. (2011). LINE-1 elements in structural variation and disease. *Annual review of genomics and human genetics*, *12*, 187–215. https://doi.org/10.1146/annurev-genom-082509-141802

Bourque, G., Burns, K.H., Gehring, M. *et al.* Ten things you should know about transposable elements. *Genome Biol* 19, 199 (2018). https://doi.org/10.1186/s13059-018-1577-z

Cordaux, R., & Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nature reviews. Genetics*, *10*(10), 691–703. https://doi.org/10.1038/nrg2640

Damert A, Raiz J, Horn AV, Löwer J, Wang H, Xing J, Batzer MA, Löwer R, Schumann GG. 5'-Transducing SVA retrotransposon groups spread efficiently throughout the human genome. Genome Res. 2009 Nov;19(11):1992-2008. doi: 10.1101/gr.093435.109. Epub 2009 Aug 3. PMID: 19652014; PMCID: PMC2775593.

Damert A, SVA Retrotransposons and a Low Copy Repeat in Humans and Great Apes: A Mobile Connection, *Molecular Biology and Evolution*, Volume 39, Issue 5, May 2022, msac103, https://doi.org/10.1093/molbev/msac103

Ebert P, Audano PA, Zhu Q, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*. 2021;372(6537):eabf7117. doi:10.1126/science.abf7117

Edgar, RC (2021), MUSCLE v5 enables improved estimates of phylogenetic tree confidence by ensemble bootstrapping, *bioRxiv* 2021.06.20.449169. https://doi.org/10.1101/2021.06.20.449169.

Gianfrancesco, O., Bubb, V. J., & Quinn, J. P. (2017). SVA retrotransposons as potential modulators of neuropeptide gene expression. *Neuropeptides*, *64*, 3–7. https://doi.org/10.1016/j.npep.2016.09.006

Gianfrancesco O, Geary B, Savage AL, Billingsley KJ, Bubb VJ, Quinn JP. The Role of SINE-VNTR-Alu (SVA) Retrotransposons in Shaping the Human Genome. *International Journal of Molecular Sciences*. 2019; 20(23):5977. https://doi.org/10.3390/ijms20235977

Goodier, J.L., Kazazian, H.H., Retrotransposons Revisited: The Restraint and Rehabilitation of Parasites. *The Cell* 1, 135 (2008). https://doi.org/10.1016/j.cell.2008.09.022

Ha, H., Loh, J.W. & Xing, J. Identification of polymorphic SVA retrotransposons using a mobile element scanning method for SVA (ME-Scan-SVA). *Mobile DNA* 7, 15 (2016). https://doi.org/10.1186/s13100-016-0072-x

Halabian, R., & Makałowski, W. (2022). A Map of 3' DNA Transduction Variants Mediated by Non-LTR Retroelements on 3202 Human Genomes. *Biology*, *11*(7), 1032. https://doi.org/10.3390/biology11071032

Han, J.S. Non-long terminal repeat (non-LTR) retrotransposons: mechanisms, recent developments, and unanswered questions. *Mobile DNA* 1, 15 (2010). https://doi.org/10.1186/1759-8753-1-15

Hancks, D. C., & Kazazian, H. H., Jr (2010). SVA retrotransposons: Evolution and genetic instability. *Seminars in cancer biology*, *20*(4), 234–245. https://doi.org/10.1016/j.semcancer.2010.04.001

Hancks, D.C., Kazazian, H.H. Roles for retrotransposon insertions in human disease. *Mobile DNA* 7, 9 (2016). https://doi.org/10.1186/s13100-016-0065-9

Havecker, E.R., Gao, X. & Voytas, D.F. The diversity of LTR retrotransposons. *Genome Biol*5, 225 (2004). https://doi.org/10.1186/gb-2004-5-6-225

Kent WJ. BLAT - the BLAST-like alignment tool. *Genome Res.* 2002 Apr;12(4):656-64.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. *Genome Res.* 2002 Jun;12(6):996-1006.

Kobayashi K, Nakahori Y, Miyake M, et al. An ancient retrotransposal insertion causes Fukuyama-type congenital muscular dystrophy. *Nature*. 1998;394(6691):388-392. doi:10.1038/28653

Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome [published correction appears in Nature 2001 Aug 2;412(6846):565] [published correction appears in Nature 2001 Jun 7;411(6838):720. Szustakowki, J [corrected to Szustakowski, J]]. Nature. 2001;409(6822):860-921. doi:10.1038/35057062

Larsson, A. (2014). AliView: a fast and lightweight alignment viewer and editor for large data sets. *Bioinformatics*30(22): 3276-3278. http://dx.doi.org/10.1093/bioinformatics/btu531

Linheiro, R. S., & Bergman, C. M. (2012). Whole genome resequencing reveals natural target site preferences of transposable elements in Drosophila melanogaster. *PloS one*, *7*(2), e30008. https://doi.org/10.1371/journal.pone.0030008

Makino S, Kaji R, Ando S, et al. Reduced neuron-specific expression of the TAF1 gene is associated with X-linked dystonia-parkinsonism. *Am J Hum Genet*. 2007;80(3):393-406. doi:10.1086/512129

Pace, J. K., 2nd, & Feschotte, C. (2007). The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage. *Genome research*, *17*(4), 422–432. https://doi.org/10.1101/gr.5826307

Pfaff, A.L., Bubb, V.J., Quinn, J.P. *et al.* Reference SVA insertion polymorphisms are associated with Parkinson's Disease progression and differential gene expression. *npj Parkinsons Dis.* 7, 44 (2021). https://doi.org/10.1038/s41531-021-00189-4

Pourrajab, F., & Hekmatimoghaddam, S. (2021). Transposable elements, contributors in the evolution of organisms (from an arms race to a source of raw materials). *Heliyon*, *7*(1), e06029. https://doi.org/10.1016/j.heliyon.2021.e06029

Pray, L. (2008) Transposons: The jumping genes. *Nature Education* 1(1):204
Quinn, J. P., & Bubb, V. J. (2014). SVA retrotransposons as modulators of gene expression. *Mobile genetic elements*, *4*, e32102. https://doi.org/10.4161/mge.32102

Rebollo, R., Farivar, S. & Mager, D.L. C-GATE - catalogue of genes affected by transposable elements. *Mobile DNA* 3, 9 (2012). https://doi.org/10.1186/1759-8753-3-9

Sanders, A., Falconer, E., Hills, M. *et al.* Single-cell template strand sequencing by Strand-seq enables the characterization of individual homologs. *Nat Protoc* 12, 1151–1176 (2017). https://doi.org/10.1038/nprot.2017.029

Shen L, Wu LC, Sanlioglu S, et al. Structure and genetics of the partially duplicated gene RP located immediately upstream of the complement C4A and the C4B genes in the HLA class III region. Molecular cloning, exon-intron structure, composite retroposon, and breakpoint of gene duplication. *J Biol Chem*. 1994;269(11):8466-8476.

Smit, AFA, Hubley, R & Green, P. *RepeatMasker Open-4.0*.

2013-2015 <http://www.repeatmasker.org>.

Symer DE, Connelly C, Szak ST, et al. Human l1 retrotransposition is associated with genetic instability in vivo. *Cell*. 2002;110(3):327-338. doi:10.1016/s0092-8674(02)00839-5

Takasu M, Hayashi R, Maruya E, et al. Deletion of entire HLA-A gene accompanied by an insertion of a retrotransposon. *Tissue Antigens*. 2007;70(2):144-150. doi:10.1111/j.1399-0039.2007.00870.x

Xing, J., Wang, H., Belancio, V. P., Cordaux, R., Deininger, P. L., & Batzer, M. A. (2006). Emergence of primate genes by retrotransposon-mediated sequence transduction. *Proceedings of the National Academy of Sciences*, *103*(47), 17608-17613.https://doi.org/10.1073/pnas.0603224103