The identification and characterization of inter- and intra-species genetic diversity derived from retrotransposons in humans

By

Wanxiangfu Tang, B.Eng.

A Thesis submitted to the Department of Biological Sciences in partial fulfillment of the requirements for the degree of Master of Science

August, 2012

Brock University

St. Catharines, Ontario

Table of Content

1. Abstract 4
2. List of Figures 6
3. List of Tables 7
4. List of abbreviations 9
5. Acknowledgement 10
6. Introduction 11
6.1. Human retrotransposons 15
6.1.1. LINE-1 elements 15
6.1.2. Alu elements 17
6.1.3. SVA elements 17
6.1.4. ERV elements 18
6.2. Impact on genome evolution 19
6.2.1. *Impact on human genome structure* 19
6.2.1.1. Target site duplication 23
6.2.1.2. Transduction 24
6.2.1.3. Insertion mediated deletion 24
6.2.2. Impact on splicing and gene expression 25
6.3. Retrotransposon insertion polymorphisms (RIPs) 26
6.3.1. Methods for identifying and ascertaining RIPs 26
6.3.2. Database documentation of RIPs in dbRIP 30
6.4. Human specific retrotransposons 31
7. Materials and methods 33
7.1. Data collection 33
7.1.1. Documenting reported RIPs 33
7.1.2. Sources for genomic sequences 34
7.2. In silico identification of human specific retrotransposon elements 35
7.2.1. Analysis and preparation of the raw data 36
7.2.2. BLAT based method 38
7.2.3. liftOver based method 41
7.2.4. Determination of optimal BLAT and liftOver criteria for identifying HS-REs 42
7.2.5. Transduction 44
7.2.6. RE insertion mediated deletion 47
7.2.7. Documentation of HS-REs in dbRIP 48
7.2.8. Computational scripts 48
8. Results 50
8.1. Repeat masker input data 50
8.1.1. TE inserting into TEs 50
8.2. Criteria determined by utilizing RIPs training dataset 52
8.3. Human-specific REs 55
8.4. Transduction 58
8.5. RE-insertion mediated deletion 60
8.6. Final HS-RE list 62
8.7. Functional impact assessment 63

8.7.1. *TSD length and integration site sequence motif* 63
8.7.2. Retrotransposition activity of HS-RE subfamilies 66
8.7.3. Distribution of HS-REs throughout the human genome 69
8.7.4. GC content for different HS-RE families and all other members in the same families 72
8.7.5. Genome size contribution by human-specific retrotransposons 73
8.7.6. Gene context of human specific retrotransposons elements 75
9. Discussion 78
9.1. In silico identification of human specific retrotransposon elements 78
9.1.1. Value of using four additional primate genomes in identifying HS-REs. ... 78
9.1.2. Combination of BLAT and liftOver based methods 80
9.1.3. The value of using RIPs as training and testing dataset 81
9.2. HS-RE activity level 82
9.3. The bias of HS-REs for chromosome Y 85
10. Summary and future perspectives 89
11. References 91

1. Abstract

Retrotransposons, which used to be considered as “junk DNA”, have begun to reveal their immense value to genome evolution and human biology due to recent studies. They consist of at least ~45% of the human genome and are more or less the same in other mammalian genomes. Retrotransposon elements (REs) are known to affect the human genome through many different mechanisms, such as generating insertion mutations, genomic instability, and alteration in gene expression. Previous studies have suggested several RE subfamilies, such as Alu, L1, SVA and LTR, are currently active in the human genome, and they are an important source of genetic diversity between human and other primates, as well as among humans. Although several groups had used Retrotransposon Insertion Polymorphisms (RIPs) as markers in studying primate evolutionary history, no study specifically focused on identifying Human-Specific Retrotransposon Element (HS-RE) and their roles in human genome evolution. In this study, by computationally comparing the human genome to 4 primate genomes, we identified a total of 18,860 HS-REs, among which are 11,664 Alus, 4,887 L1s, 1,526 SVAs and 783 LTRs (222 full length entries), representing the largest and most comprehensive list of HS-REs generated to date. Together, these HS-REs contributed a total of 14.2Mb sequence increase from the inserted REs and Target Site Duplications (TSDs), 71.6Kb increase from transductions, and 268.2 Kb sequence deletion of from insertion-mediated deletion, leading to a net increase of ~14 Mb sequences to the human genome. Furthermore, we observed for the first time that Y chromosome might be a hot target for new retrotransposon insertions in general and particularly for LTRs. The data also allowed for the first time the survey of

frequency of TE insertions inside other TEs in comparison with TE insertion into none-TE regions. In summary, our data suggest that retrotransposon elements have played a significant role in the evolution of *Homo sapiens*.

2. List of Figures

Figure 1 the transposable element content of the human genome. 13

Figure 2. A schematic diagram of target-primed reverse transcription (TPRT). 16

Figure 3 Impact of retrotransposon elements on human genome structure. 22

Figure 4. Illustration of PCR as RIP ascertaining method. 30

Figure 5. Screen shots of HS-REs in UCSC genome browser..... 36

Figure 6. An example of a fragmented retrotransposon element. 37

Figure 7. A schematic diagram of the *in silico* comparative genomics approach for identifying HS-REs 40

Figure 8. A schematic diagram of the algorithm for identifying HS-RE with 3' transductions 46

Figure 9. A schematic diagram of *in silico* comparative genomics approach for identifying typical RE insertion mediated deletions. 49

Figure 10. Base composition of insertion motif for different HS-RE families. 66

Figure 11. Subfamily distribution of HS-RE. 69

Figure 12. The density of HS-REs and genes in the human chromosomes. 72

Figure 13. A screenshot of a shared RE elements in UCSC genome browser 80

3. List of Tables

Table 1 dbRIP Statistics Release 2 (hg19) 35

Table 2 percentage of TE inserting into TE region of major TE families 51

Table 3 Percentage of different HOST TE families 51

Table 4. Sample BLAT output 52

Table 5. BLAT criteria based on the RIPs training dataset 53

Table 6. BLAT output of the RIPs training dataset 53

Table 7. liftOver output of the RIPs test dataset* 54

Table 8. Combined results of the RIPs test dataset 54

Table 9. Match pattern for BLAT-based and lifeOver methods 55

Table 10. Combination of BLAT and liftOver results 57

Table 11. Transduction events between different families 59

Table 12. Genes involving HS-RE transduction in the intron regions 60

Table 13. RIMD events between different families 61

Table 14. A list of genes carrying a RMID in intron regions 62

Table 15. Stats of final HS-RE list 63

Table 16. Average length of TSD of each RE family 64

Table 17. Z-test results of average TSD lengths of different RE families 65

Table 18. Subfamily distribution of different HS-RE families 67

Table 19. RE and gene density in the human genome 70

Table 20. GC-content data of different HS-RE families and older members in the RE family 72

Table 21. Z-tests result for different HS-RE families 73

Table 22. Size contribution by HS-RE on different chromosomes 75

Table 23. Gene context information of different RE families 76

Table 24. A list of genes containing HS-REs in exon regions 77

4.List of abbreviations

BLAST-like alignment tool (BLAT);
Copy number variations (CNVs);
Database of retrotransposon insertion polymorphisms (dbRIP);
Double-strand breaks (DSBs);
Duchenne muscular dystrophy(DMD);
Endogenous retroviruses (ERVs);
Endonuclease-independent manner (ENi);
Fukuyama-type congenital muscular dystrophy(FCMD);
Human endogenous retroviruses (HERV);
Human specific retrotransposons (HS-REs);
Next generation sequencing (NGS);
Non-allelic homologous recombination (NAHR);
Open reading frame(ORF);
Polymerase chain reaction (PCR);
Recombination-mediated deletions (RMD);
RE-insertion mediated deletion (RIMD);
Retrotransposon elements (REs);
SINE-R, VNTR, and *Alu*(SVA);
Retrotransposon insertion polymorphisms (RIPs);
Single nucleotide polymorphisms (SNPs);
Target site duplications (TSDs);
Target-primed reverse transcription (TPRT);
The *Practical Extraction and Reporting Language* (PERL);
Transposable element (TE);
University of California at Santa Cruz (UCSC);
Variable number of tandem repeats (VNTR);

5.Acknowledgement

I would first like to thank my supervisor Professor Liang. Without his help, both financially and academically, this work would not be possible. I would also like to thank Professor Bruce, Professor Inglis and Professor Qiu, who are also in my supervisor committee, for their advices on this project. I would then like to thank my parents for their financial support. I would also like to thank my lab mates: Musa Ahmed, Scott Golem and Amanda Bering for all the inspiration I had from our discussion.

6.Introduction

Transposable elements (TEs) are defined as DNA sequences which can change their positions in the genome. Although they were first discovered by McClintock back in the 1940s [1], they had been overlooked and thought to be "junk genes" for about 50 years. It was only after the availability of genome sequences for many species that researchers began to have a better appreciation of their abundance and function in the genomes [2].

TEs can be divided into two major classes: DNA transposons and retrotransposon elements (REs). In the human genome, the vast majority of TEs are retrotransposons, making up at least 51% of the genome, while DNA transposons make up 3.5% of the genome (Figure 1). DNA transposons are able to excise themselves from the genome and travel to new genome sites in the form of DNA, but in human genome they have lost this ability ~37 million years ago [3, 4]. Retrotransposons mobilize to new locations in the genome via an RNA-based duplication process called retrotransposition [4]. Depending on the presence or absence of long-terminal repeats (LTRs), retrotransposons can be divided into LTR retrotransposons and non-LTR retrotransposons. The largest class of LTR-retrotransposons is endogenous retroviruses (ERVs), and they account for ~9.3% of the human genome. Most of these LTRs have lost their mobilization ability and are possessed by all members in the human population. Evidence suggests that very few of them still have ongoing retrotransposition activity [5, 6]. The non-LTR retrotransposons can be further divided into two classes: autonomous retrotransposons and non-autonomous retrotransposons, depending on whether they encode genes necessary for their transposition. The LINE-1(L1), *Alu* and SVA elements, have the highest

retrotransposition ability and are responsible for ~55% of human TEs (Figure 1) [2]. Also, they are shown to be currently active in the human genome. *de novo* insertion are known to be responsible for genetic disorder in more than 80 cases [7-10]. This study will focus on identifying human-specific retrotransposon insertions and evaluating their impact on genome evolution because they are by far the largest and most active TE groups. More detailed descriptions about REs that are known to remain active are provided in the next sections,

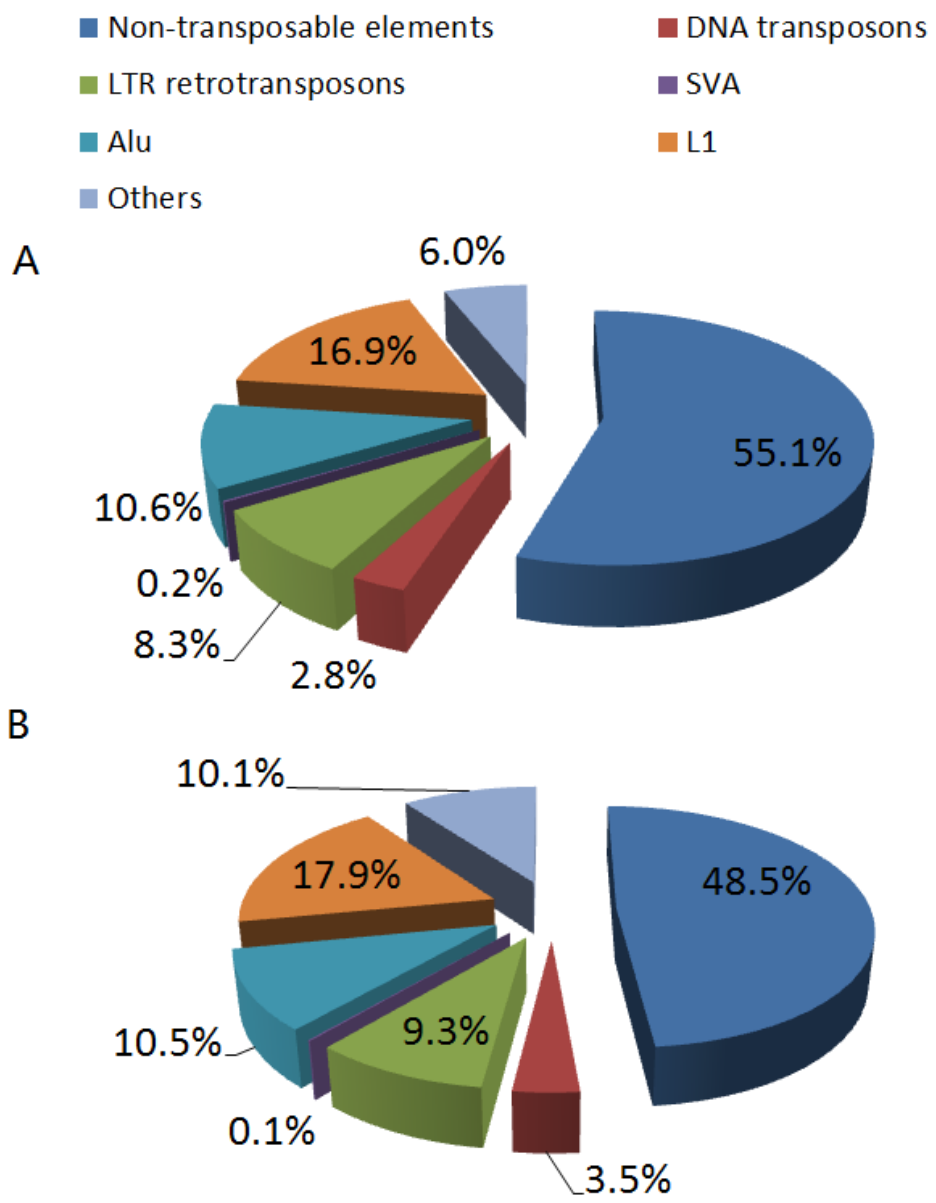

Figure 1. The transposable element content of the human genome.

Panel A is adapted by permission from Macmillan Publisher LTD: Nature Review Genetics, copyright (2009) Figure 1 in "The impact of retrotransposons on human genome evolution" by Cordaux R and Batzer MA [11]. About 45% of the human genome can currently be recognized as being derived from transposable elements, the vast majority of which are non-LTR retrotransposons such as L1, *Alu* and SVA elements. Panel B is developed using data from the RepeatMasker data provided by the UCSC genome website

(http://genome.ucsc.edu), which is more updated and indicates that transposable elements constitute ~51.5% of the human reference genome (hg19).

6.1. Human retrotransposons

6.1.1. LINE-1 elements

Having their own enzymatic machinery for retrotransposition, L1s are the only known autonomous non-LTR retrotransposons that are currently active within the human genome. Their continued mobilization activity in the last 150 million years has led to ~500,000 copies in the human genome [2]. Consisting of ~18% of the human genome, L1s are the most successful TEs in the human genome by sequence length. A full-length L1 is about 6,000 base pairs long and is made of an internal RNA polymerase II promoter, two open reading frames (ORF1 and ORF2) and a polyadenylation signal followed by a homopolymeric tract of adenosines (polyA-tail)[12, 13]. The ORF1 gene encodes a RNA-binding protein and ORF2 encodes a protein with endonuclease and reverse transcriptase activity [14-16]. The process, which provides L1s the ability to mobilize in the human genome, is known as target-primed reverse transcription (TPRT) (Figure 2). Not all members of the L1 family retain this ability; most of them have lost this ability because of truncations, internal rearrangements and mutations [17]. Research suggests that there are no more than 100 copies of L1s that are still active. This is a small number comparing to the total ~ 500,000 copies in the human genome [18]. Also, it was suggested that only a few "hot" L1s are responsible for the bulk of new insertions [18].

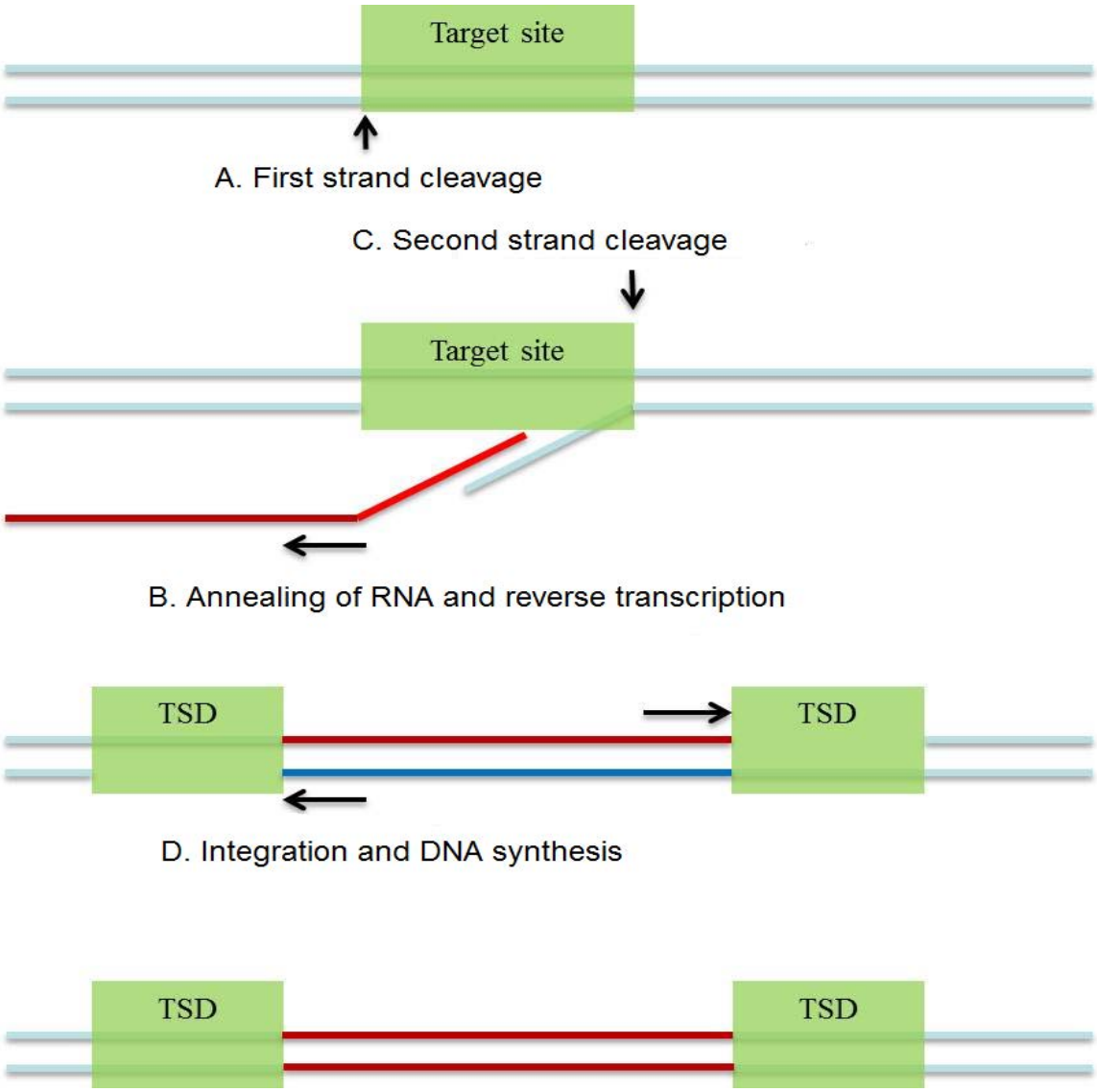

Figure 2. A schematic diagram of target-primed reverse transcription (TPRT).

This figure was adapted by permission from Macmillan Publisher LTD: Nature Review Genetics, copyright (2009), Box 1 in "The impact of retrotransposons on human genome evolution" by Cordaux R and Batzer MA [11]. A: Cleavage of first DNA strand at the target site by the retrotransposon endonuclease (EN); B: The retrotransposon RNA anneals at the nick site and starts reverse transcription by the retrotransposon reverse transcriptase (RT); C: Cleavage of second DNA strand. D: Integration at the double-strand break and removal of RNA and completion of DNA synthesis, leading to the insertion of a new copy of the retrotransposon at the target site and generation of target site duplications (TSDs).

6.1.2. Alu elements

Alu elements have been active in the past ~65 million years (My), which resulted in more than 1 million *Alu* copies present in the current human genomes [2, 19]. This makes *Alu* to be the most abundant REs in number. Considering that *Alu* elements are primate-specific and therefore relatively young, its rate of propagation is even more remarkable. A canonical full-length *Alu* element has ~300 base pairs and a dimeric structure made of two non-identical monomers, which were derived from the *7SL RNA* gene (also known as the signal recognition particle RNA, a component of the signal recognition particle), connected by an adenosine-rich linker[20-22]. *Alu* transcripts may extend into the downstream flanking sequence until a terminator is found, leading to the formation of 3' transduction, because *Alu* lacks RNA polymerase III termination signals [23, 24]. Being non-autonomous retrotransposons, *Alu* elements have to borrow the reverse transcriptase machinery of L1 elements since they lack of any coding capacity [25].

6.1.3. SVA elements

SVA elements, which originated ~25My ago, now have ~3000 copies in the human genome[26]. SVA is a retrotransposon family found only in hominoids (the group of primates comprising human and apes, which diverged from Old World monkey about 25 My ago) [26]. Shen et al. first used the term "SVA" (SINE-R, VNTR, and *Alu*) to describe this class of composite retrotransposons [27]. A full-length SVA element is ~2kb long and has a hexamer repeat region, an *Alu*-like region, a variable number of tandem repeats (VNTR) region, and a human endogenous retroviruses (HERV)-K10-like region followed by a polyadenylation signal ending with a polyA-tail [26, 28]. They might rely

on promoter activity in flanking regions since they lack an internal promoter [26, 28]. Like *Alu* elements, SVA elements are non-autonomous TEs, and they have to borrow the reverse transcriptase machinery from L1s as well [26, 28].

6.1.4. ERV elements

Endogenous retroviruses (ERVs), which closely resemble infectious retroviruses in the human genome, are LTR retrotransposons [29]. They are results of ancestral infection by exogenous retroviruses. The human ERV (HERV) now contributes ~9% of the human genome, and most HERV subfamilies have lost the ability to retrotranspose. A typical full-length HERV consists of the main retroviral genes *gag*, *pol*, and *env*, flanked by long terminal repeats (LTRs). The ERV retrotransposons have the largest size among all four types of retrotransposons (average 7~12kb). Additionally, the ERV retrotransposons can form solo-LTR through homologous recombination between the two LTR sequences that are almost identical. The majority of HERV can be found in other primate genomes, indicating that they might have originated in our ancestors at least 25 My ago[30]. Currently, HERV-K is the only subfamily that has known activity in the human genome [31].

6.2. Impact on genome evolution

Because of their continuous retrotransposition activity over the past tens of My, non-LTR retrotransposons have had a great impact on both genome structure and function which in turn has had a tremendous impact on primate genome evolution. How frequently does retrotransposition happen in germline? This is vital for us to understand how much impact TEs have on genome evolution. Research suggests that the current rate of retrotransposition is around one insertion varies from every ~20 to ~900 births within different types of TEs [32-34]. For example, researchers have estimated the current rate of *Alu* retrotransposition to be one insertion every ~20 births in humans [32, 35]. Yet because of the potential bias brought by the small datasets available, it is very difficult to give an accurate estimation on the amplification rates. In addition to germline, several groups have demonstrated that retrotransposition can occur in somatic tissues as well, with associations ranging from cancer to a possible role in brain development [34, 36]. Provided in the following subsections are descriptions of the major mechanisms for RE's impact on human genome evolution and function.

6.2.1. *Impact on human genome structure*

There are many ways that retrotransposons can affect genome structure; they can either create local genomic instability or cause genome rearrangements such as deletions, duplications and inversions. By inserting themselves into protein-coding or regulatory regions, retrotransposons can affect gene function and expression (Figure 3). These retrotransposons were the first to be detected because they may have a visible impact via phenotype [7]. Up to date, there are more than 100 cases of human genetic disorders

reported to be associated with *de novo* L1, *Alu* and SVA insertions, including haemophilia, cystic fibrosis, Apert syndrome, neurofibromatosis, Duchenne muscular dystrophy, β-thalassemia, hypercholesterolemia and breast and colon cancers [8, 10, 37].

Recent studies suggest that the ORF2 protein of L1s can cause many more DNA double-strand breaks (DSBs) than those caused by the actual L1 insertions (Figure 3) [38]. L1 elements have also been found to be associated with DSBs repair through an endonuclease-independent manner (ENi). ENi L1 insertions show a different pattern than TPRT, therefore suggesting that L1s can integrate into DSBs DNA lesions and repair them [39]. The ENi retrotransposition might be an ancient mechanism of RNA-mediated DNA repair before the acquisition of an endonuclease domain [40, 41]. Additionally, 0.5-0.7% of all L1 and *Alu* insertions has non-canonical structure and possibly derived from ENi retrotransposition [42, 43]. This suggests that ENi might be a general mechanism in non-LTR retrotransposons rather than just L1s. However, the repair of L1-mediated DSB damage does not necessarily need the involvement of L1s. This might suggest that a good proportion of the genomic instability associated with DSBs is contributed by L1 activity. However, it is not clear to what extent this mechanism contributes to human genome instability since most of these data were obtained from in-vitro experiments. Studies on *Alu* elements suggested that non-LTR retrotransposons can create microsatellites at many loci in the human genome because of their large copy numbers and their homopolymeric tracts (Figure 3) [44, 45]. *Alu* elements have two potential sources of microsatellites: the linker region in the middle and the 3’ oligo dA-rich tail. These homopolymeric repeats can experience various mutational like Single Nucleotide Polymorphisms (SNPs), which may create microsatellites of varying length and complexity.

Alu elements have been shown to be able to undergo recombination between their homologous chromosomal alleles, which has been proposed as a mechanism for gene conversion. Gene conversion might also play a role in the evolution of *Alu* elements, as it might be able to inactivate active *Alu* elements or reactivate inactive copies [46, 47].

While inserting themselves into the genome, retrotransposons normally create a type of structural rearrangement called target site duplication (TSD), in addition to the inserted RE sequences. This will be discussed in the next section 6.2.1.1. Instead of TSD, insertion of retrotransposons can sometimes cause the deletion of target genomic sequences simultaneously (Figure 3). Possible mechanisms include endonuclease-dependent and ENi mechanisms. L1 and *Alu* insertion mediated-deletions are shown to be usually shorter and happen with a much lower frequency in the human and chimpanzee genomes than in cultured cells. This might suggest the involvement of a possible negative selection against large retrotransposon insertion-mediated deletions [48, 49].

L1 and *Alu* elements can perform non-allelic homologous recombination (NAHR) to generate structural variation at the post-insertion stage because of their extremely high copy numbers in the human genome and high level of sequence similarity among members from the same family (Figure 3). It can lead to many types of genomic rearrangements such as deletions, duplications and inversions. For example, *Alu* recombination-mediated deletions (RMD) can occur in the human genome. More than 70 *Alu* RMDs have been reported to be associated with various forms of cancer and genetic disorders [8, 9].

L1s and SVA elements can sometimes generate 5’ or 3’ transduction, which means they can carry upstream or downstream flanking sequences with them to new loci in the

human genome (Figure 3). This may be caused by different levels of polyadenylation or promoter signal strength [50, 51]. This is the mechanism causing exon shuffling which will be later discussed in section 6.2.1.2.

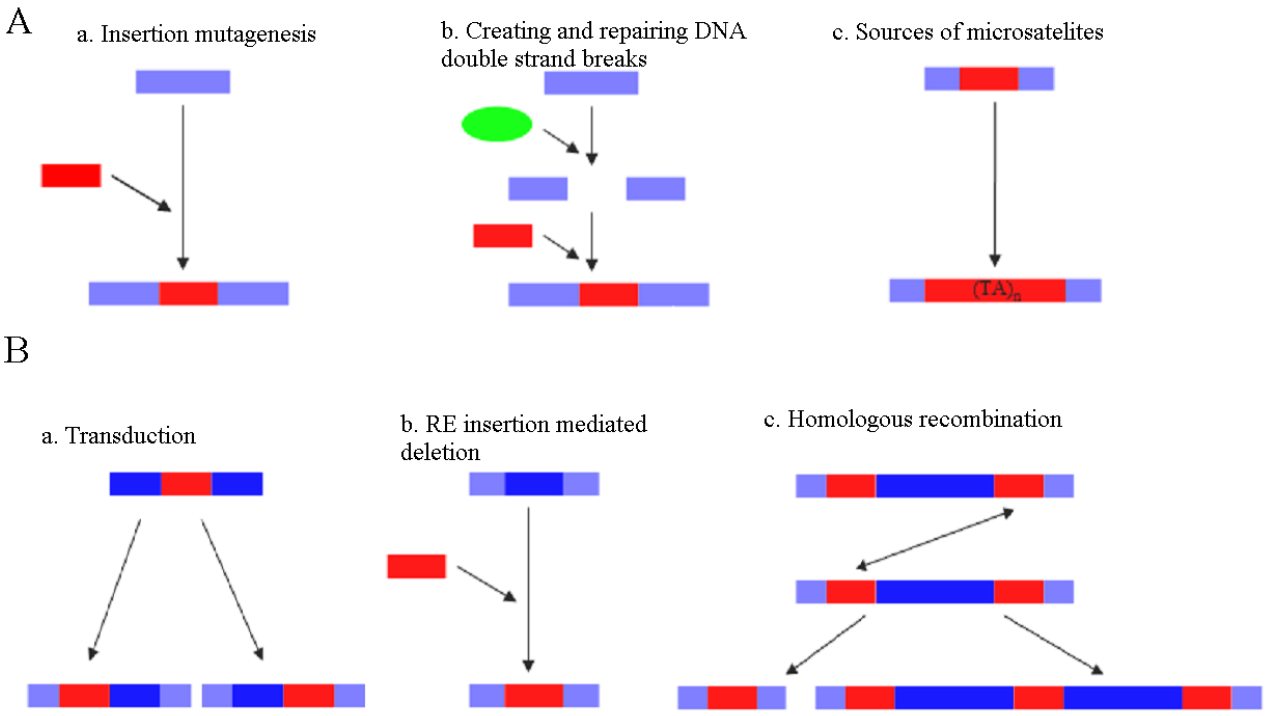

Figure 3 Impact of retrotransposon elements on human genome structure.

This figure was adapted by permission from Macmillan Publisher LTD: Nature Review Genetics, copyright (2009), from Figure 2 in "The impact of retrotransposons on human genome evolution" by Cordaux R and Batzer MA [11].
Panel A: Different RE mechanisms which can generate local genome instability. a: REs (red box) can create insertion mutagenesis depending on genic location of the integration site(light blue boxes); b: L1s have known ability to create DNA double strand breaks by its protein product ORF1 (green circle) as well as repair existing DSBs by non-canonical endonuclease-independent insertion(red box); b: Homopolymeric tracts carried by REs (red box) can generate microsatellites.
Panel B: Several RE mechanisms which contribute to genome rearrangements. a: REs (red box) have known ability of co-retrotransposition of downstream or/and upstream flanking sequences (blue boxes). b: An RE insertion (red box) can sometime result in deletion of the pre-integration site sequences (blue boxes). c: Homologous recombination is possible between non-allelic homologous REs (red boxes). This could result in either deletion or duplication of genomic sequences between those REs (blue boxes).

6.2.1.1. Target site duplication

Unlike other variations such as single nucleotide polymorphisms (SNPs) and copy number variations (CNVs), TE insertions have some unique characteristics. Target site duplication (TSD) is the hallmark of TE insertion.

In the late 1970s, several research groups noticed that the TEs tended to have a pair of short repeats alongside the insertion sequence [1, 52, 53]. The TSDs are generated by the transposition process. Because of the staggered cleavage of the double strand DNA (Figure 2), most TE transposition would leave a pair of short repeats of sequence at the integration site on each side of the insertion. This is the direct evidence of a TE insertion event. As different types of TEs utilize different mechanisms for transposition inside human genome, they have different patterns for TSDs, which vary in length from several bps to several dozen bps and also likely in sequence signature. Therefore, TSDs can be used as an important tool to identify retrotransposon insertion events and to study the mechanism differences between different RE families.

Generation of TSD serves as TE's mechanism second to insertion of REs causing the genome to increase in size. Although each TSD only provides several bps to several dozen bps, their abundant copy number would be responsible for increasing megabases of primate genome. Also, as previously discussed, 0.5-0.7% of all L1 and *Alu* insertions are proved to have non-canonical structure in which case the TSDs are absent [42, 43]. Therefore, the lack of TSDs could be used as evidence to identify non-canonical RE insertion events.

6.2.1.2. Transduction

The 3’ transduction by L1 elements in human genome was first reported in 1994[54]. They discovered that a 3’ transduction alongside the insertion of a L1 insertion into the dystrophin gene caused muscular dystrophy in a single human individual.

In the case of 3’ transduction, a read-through transcript of the RE element transcribes flanking genomic region downstream by virtue of a weak L1 termination and polyadenylation signal. Transduction of adjacent genomic DNA by RE elements may result in the creation of new exons and alteration of gene expression through promoter and enhancer shuffling [55]. It was estimated that ~20% of all L1 elements are associated with 3’ transduction.

5’ transduction occurs commonly in some SVA subfamilies [51] at an estimated rate of ~8%. In the case of 5’ transduction, retrotransposon transcription is likely driven by an external promoter located in the 5’ region of a RE, bringing in extra sequences upstream of the parent RE to the new site. Similar to the 3’ transduction events, the 5’ transduction events have known disease association. As an example, a group reported a disease associated L1 mediated 5’ transduction event in mice [56].

6.2.1.3. Insertion mediated deletion

In the case of retrotransposon insertion-mediated deletion, the RE sequences have a non-canonical structure: lack of TSDs, non-canonical endonuclease nick sites and sometimes the absence of an oligo-dA rich tail. Researchers first reported L1-mediated

deletion in 2002. In vivo and in vitro studies suggested that L1 retrotransposition events could result in significant deletion of genomic sequences spanning from 1 bp to 70,000 bp at a rate of 10% [57-59]. If 10% of all L1 retrotransposons in the human genome induced deletion, millions of bases could have been deleted. If the deleted target site DNA is located in genic regions, it might have potential association with disease. For example, a 1-bp deletion in the DMD gene by a L1 element has been reported to result in Duchenne muscular dystrophy, and a 6-bp deletion also by a L1 insertion in the FCMD gene has been considered to cause Fukuyama-type congenital muscular dystrophy [60, 61].

6.2.2. Impact on splicing and gene expression

In addition to the impact on genome structure, recent studies have shown that retrotransposons can also influence human evolution at the RNA level via various mechanisms [62-77]. Alternative splicing is a widespread phenomenon, which occurs in 40-60% of human genes [2] and leads to human proteome variation by producing more than one type of mRNA from a single gene, encoding different proteins. Both *Alu* and L1 elements can provide new splice sites, which can possibly contribute to exonization and alternative splicing [62, 63].

Research has demonstrated that RNA polymerase II has reduced ability to read through L1 sequences [64]. Therefore, intronic L1 elements can interfere with transcriptional elongation of the host gene. Additionally, gene transcript can be terminated by the polyadenylation signals carried by retrotransposons, which some time

leads to gene splitting [65-67].

Studies have shown that *Alu* elements are able to bind transcriptional factors, which might modulate gene expression [68, 69]. Evidence suggested that sense or anti-sense transcription through nearby genes can be triggered by the functional promoter of L1 and *Alu* elements [70, 71]. In addition, evidence also suggests that editing with *Alu* elements might be enhanced by their dimeric structure and certain intramolecular pairs of *Alus*, and A-to-I editing of pairs of opposite directional *Alus* in the 3' UTR region can suppress expression through nuclear retention of mRNA transcripts [77].

6.3. Retrotransposon insertion polymorphisms (RIPs)

Due to their continuing retrotransposition process, retrotransposons generate genetic diversity among species, as well as within species. Retrotransposon insertion polymorphisms (RIPs) can be defined as the co-existence of presence and absence status of a retrotransposon insertion at a specific site in a population of the same species. RIP represents an important type of genetic diversity due to the important impact of retrotranspsons on genome evolution and function as previous discussed, and therefore it is valuable to identify and document such type of genetic polymorphism in humans, along with other type of genetic polymorphism, such as SNPs and indels.

6.3.1. Methods for identifying and ascertaining RIPs

Many methods have been used to identify RIPs. Genomic library screening with RE

specific probes/primers has discovered a small number of RIPs in the early stage of the study on polymorphic retrotransposon insertions [78-80]. However, this can be very time consuming. The publication of the human genome reference sequence has allowed researchers to develop more fruitful methods. In the first strategy, retrotransposon elements belonging to young subfamilies were first identified as likely candidates for RIPs by computational sequence analysis, and oligonucleotide primers were then designed based on the flanking regions for polymerase chain reaction (PCR)-based assays to ascertain the polymorphism status of these candidates by screening DNA samples from diverse human populations[81-89]. While this strategy is straight forward and proven fruitful, leading to the identification of a total of 881 Alu, 392 L1, and 25 SVAs RIPs [26], the candidates are limited to those TEs that are contained in the reference genome sequence, and it is costly prohibitory to screen all young TEs.

A second and more fruitful type of approaches is computational comparative genomics [90-96]. Wang et al. used the in silico comparative genomics approach at the whole genome level to identify *Alu* insertion polymorphisms in 2006 [90]. They compared the human genome sequences generated by the International Consortium of Human Genome Project with the version generated by the Venter group at Celera [2, 95]. A total of 800 *Alu* insertion polymorphisms were identified, which represented the largest data set of *Alu* insertion polymorphisms found by a single study at that time. These demonstrate that computational comparative genomics can be used as an efficient high-throughput strategy for ascertaining RIPs. However, the lack of complete human genome data limited this method back then.

Recently, a few more approaches have been developed for identifying *de novo*

retrotransposon insertions in individual genomes to utilize of next generation sequencing (NGS) technologies. A few groups have been using NGS to selectively sequence the junction areas between RE insertions and their flanking region. For example, Witherspoon *et al.* identified 483 novel *Alu* RIPs from several Japanese individuals by using this strategy [97]. Similarly, Ewing & Kazazian utilized the NGS sequencing technology to identify novel L1 insertions. By surveying 26 individuals, they were able to identify 367 L1s outside the reference genome, the majority of which are novel polymorphic L1s [94]. Lately, with the availability of personal genome data in large numbers, such as those from the 1000 genome project, the computational comparative genomics approach has been adapted to make the use of these newly available genomes for identification of RIPs. For example, a total of 6209 novel RIPs were identified from the 1000 genome project [98, 99].

As the genome sequence data and the methods are not 100% accurate, validation of computational predictions are needed. PCR is the current gold standard for ascertaining a RIP. The presence and absence of the RE insertion will lead to size difference at a specific locus: the size of the insertion positive allele will be larger than that of the insertion negative allele roughly by the size of the insertion. In this case, a pair of PCR primers based on the flanking region of the insertion can be designed. This method is good for REs with relatively small size and it can distinguish among the three genotypes of an RE insertion. Having only one large product means "+/+" and only one small product means "-/-" while having both the large product and the small product indicates "+/-" (Figure 4A).

However, it will be difficult to obtain a product of the insertion positive allele even with a long range PCR for a full length L1 and HERV as they can be as long as 10 kb. In this case, two additional primers inside the RE, which are oriented outwards, have to be designed. In the presence of the RE insertion, these two primers will work with the two primers in the flanking region to generate two shorter products, while for the insertion negative allele there is only one product from the two primers in the flanking regions. Similarly, the genotypes of the RE insertion can be determined. Two larger products means "+/+" and one short product means "-/-" while having all three products stands for "+/-" situation (Figure 4B).

The other advantage of the PCR assay is that it provides DNA for sequencing, so that the complete RE insertion sequence can be available. This is essential for RIPs outside the reference genome, since the RE insertion and TSD information are usually unavailable in such cases. Among the known RIPs, only a small portion has been PCR verified. Researchers tend to use it to assess the accuracy of the methods, as it might not be feasible to experimentally verify all RIP candidates due to their large numbers and prohibitive cost of validation by PCR. However, it important to experimentally validate all computationally identified RIPs as it will give more reliable and detailed genetic variation data, and it seems to be the solution to provide complete sequences for the RIPs determined by NGS based methods as they tended to give incomplete sequence information.

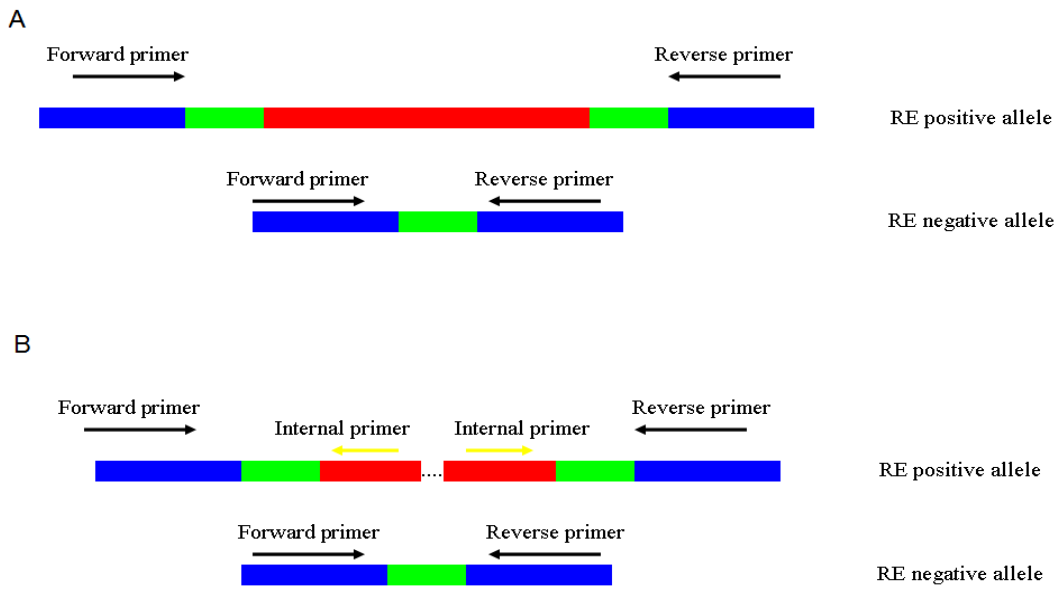

Figure 4. Illustration of PCR as RIP ascertaining method.

This figure was adopted and redrawn from Figure 1 in "Database documentation of retrotransposon insertion polymorphisms" by Liang P and Tang W [100]. Panel A: strategy for ascertaining a short RE insertions (e.g. <=2kb); Panel B: strategy for ascertaining a long RE insertions, such as full-length L1 and HERV insertions.

6.3.2. Database documentation of RIPs in dbRIP

Because of the large number of known RIPs and many more are expected to be identified, it is essential that these data are compiled in a way that is accurate and easy to access. Accuracy here refers not only to the reliability and accuracy of the data in all components of the information, including the sequence of the insertion, location, and classification, but also the completeness of the data. For example, just knowing the presence of the insertion at a specific location does not provide sufficient information about the potential impact of the insertion. It is important also to know the exact sequence

of the insertion, its orientation, and the TSDs or deletion of the flanking sequence and/or the presence of 5' or 3' transduced sequence. Other types of information, such as the source of the polymorphism (i.e. the specific population or individual showing the presence or absence of the insertion), the ascertaining/genotyping method, the insertion allele frequency in the examined populations, the phenotype association, etc. are also very useful. The sample source is very important for future study of rare RIPs. Among the current databases covering RIP data, dbRIP is the only database which was designed specifically to accommodate the special needs of retrotransposon insertion polymorphism data [101]. Before this study, dbRIP contained a total of 2095 unique RIPs, among which are 1625 Alus, 407 L1s and 63 SVAs.

6.4. Human specific retrotransposons

As previously discussed, retrotransposon insertion polymorphisms are an important mechanism of generating both inter-species and intra-species diversity and retrotransposition events happened throughout the entire primate evolutionary history. There were retrotransposition events that happened after the divergence of humans and primates, therefore these retrotransposons were possessed by humans only. It would be valuable to identify these human specific retrotransposons and study their impact on the human genome.

Previous studies on human specific retrotransposons (HS-REs) have either focused on identifying if a RE subfamily is human specific [79] or focused on testing human specific status for other results such as transduction [11]. In addition, most of the studies

have been limited by lack of resources: there were not enough primate genomes available at the time of their studies.

Our hypothesis is that among the current REs in the human genome, a proportion of them arose from RE insertion events, which happened after the human and primate diverging and are human specific. These HS-REs have impacted human genome evolution and function by generating the genetic diversity between human and non-human primates, as well as among human individuals. Therefore, this study was set to first develop a computational method to identify all HS-REs in the human genome by comparing the human genome with 4 other primate genomes, and then to analyze their trend, distribution and assess their impact on the evolution and function of the human genome.

7.Materials and methods

7.1. Data collection

7.1.1. Documenting reported RIPs

In the first full release of dbRIP data in June of 2006[101], there were 2095 non-redundant entries from a total 2897 reported cases, including 1625, 407 and 63 cases of *Alu*s, L1s and SVAs, respectively, and later, the database was extended to include RIPs derived from HERVs. As part of this study, new RIPs reported in recent literature were collected with the sequences of the REs and their flanking region sequences (if available) were collected based on information from the original studies along with other detail information such as genotype, primer and so on. Then the RE was annotated to define the exact RE type, the exact start and end position of the RE insertions and TSDs. The sequence annotation was mostly done by RE-carrying allele sequences with corresponding pre-integration allele sequences using BLAT programs and in-house PERL scripts. A total of 1098 RIPs (803 Alus, 271 L1, 14 SVA, 10 HERVs) were collected and added to dbRIP from this study. In addition to data update, improvements and addition of utilities were made, and these include: i) the dbRIP was updated to support both hg18 and hg19; ii) a new dbRIP ID system was developed: a fixed length (7 digits) number was assigned to each unique entry with the 1st digit indicative of the TE type (1000000~1999999 for Alus; 2000000~2999999 for L1s; 3000000~3999999 for SVAs; 4000000~4999999 for ERVs); iii) a new utility (Position Mapping) in dbRIP was

developed which allows researchers to easily check for overlaps between their RIPs result and RIPs in dbRIP; iv) the existing SearchdbRIP utility was upgraded to allow search by several new criteria; v) several existing errors were fixed and uniformed terminology was applied for several data fields.

As of writing, dbRIP covers a total of 3,254 non-redundant RIP entries [100, 101], including 2569, 598, 77, 10 cases of *Alu*s, L1s, SVAs, and HERVs, respectively (Table 1, collected from over 70 publications. The most updated RIP data were used as training and test data for methods used to identify HS-REs (see details in sections 7.2.4 and 8.2)

7.1.2. Sources for genomic sequences

The human genomic sequence data used in this study was the public version obtained from the University of California at Santa Cruz (UCSC) genomic website (Feb. 2009 hg19, GRCh37) at http://genome.ucsc.edu, and the sequences of the chimpanzee genome (Oct. 2010 panTro3), the orangutan genome (Jul. 2007 ponAbe2) and the rhesus monkey genome (Jan. 2006 rheMac2) were obtained at the same site. The gorilla genome sequence was obtained from the Ensembl site at http://www.ensemble.org. The Repeatmasker file [102] for human genome (rmskRM327 for hg19) which contains all transposable element information was also downloaded from the UCSC genomic website. All DNA sequences in fasta format were downloaded onto local bioinformatics server for further analyses.

Table 1 dbRIP Statistics Release 2 (hg19)

RIP Class	***# of loci (unique/total)***	***# of loci outside hg19***	***# of loci with genotype***	***Loci in gene context (P/E/I/IG*)***	***Disease-related loci***
Alu	2086/2708	858	526	5/7/23/735/1316	33
L1	598/800	299	123	2/2/10/182/402	15
SVA	77/87	18	31	1/1/4/28/43	3
HERV	10/10	2	6	0/0/0/5/5	0
Total	**2771/3605**	**1177**	**686**	**8/10/37/950/1766**	**51**

*: D: downstream up to 1kb; P: promoter up to 1kb; E: exon; I: intron; IG: intergenic region;

7.2. In silico identification of human specific retrotransposon elements

Overview: A strategy was developed to identify all HS-REs in the human genome by comparing sequences at each of the REs in the human genome with the sequences at the corresponding (orthologous) region in the out-group genomes (i.e. the 4 non-primate genomes). Two main bioinformatics tools deployed in this strategy were BLAT and liftOver. RepeatMasker raw data was processed to generate an input list containing the chromosome coordinates of all RE candidates for the BLAT-based and a liftOver-based approach to identify HS-REs (the details of these two methods are provided in section 7.2.2 and 7.2.3). The two lists of output from the two methods are then combined together to form the final HS-RE list. The selection of criteria for each of the two methods was determined based on a training dataset containing all RIPs documented in dbRIP, as described in section 7.2.4, and the accuracy and sensitivity of the approach was assessed by using a dataset containing novel RIPs identified by the 1000 genome project team[98].

Human-specific retrotransposon elements can be classified into two types based on the availability of orthologous regions of their flanking sequences in the out-group

genomes. Type I HS-REs have the pre-integration sequence available in out-group genomes (Figure 5A). Type II HS-REs have the entire region (insertion plus the specified minimal flanking sequences) missing in all out-group genomes (Figure 5B).

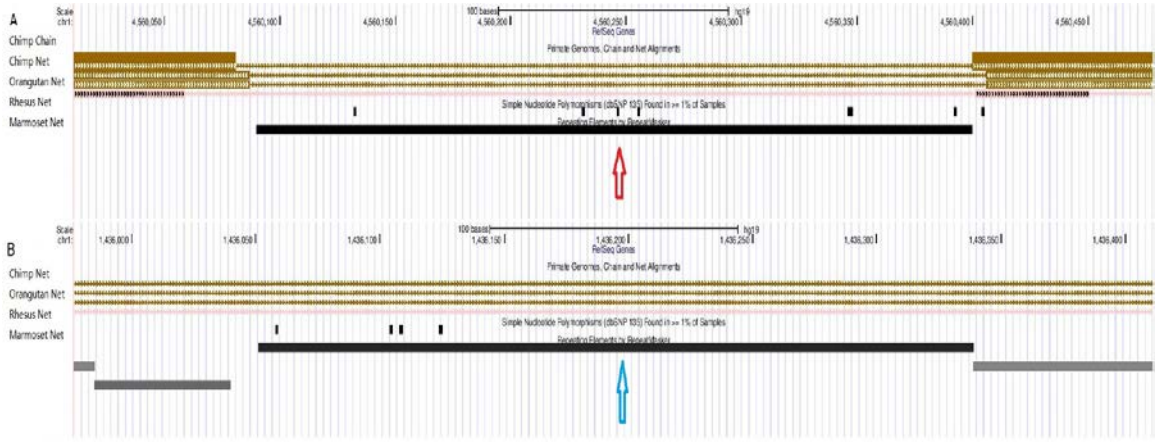

Figure 5. Screen shots of HS-REs in UCSC genome browser.

Panel A: a typical Type I HS-RE. The red arrow points to an RE insertion, while the thin lines in all primates' genomes indicate that the insertion is missing in all out-group genomes. Panel B: a typical Type II HS-RE. The RE labeled by the blue arrow has the insertion and flanking sequences missing in all out-group genomes (thin lines in all out-group genome).

7.2.1. Analysis and preparation of the raw data

RepeatMasker is a program that screens DNA sequences for interspersed repeats and low complexity DNA sequences and it provides the only commonly accepted source for repeat sequences in the human genome. The Repeatmasker file (rmskRM327 for hg19) from UCSC genome website was used as input file. It contains detailed information of all transposable elements in the human genome such as chromosome coordinates, family/subfamily information and alignment against RE consensus sequence. *Alu*, L1, SVA and LTR retrotransposons were selected among all RE families as the focus of our

analysis, as they are known to have retrotransposition activities in the human genome. We also examined REs in L2 family and did not find any HS-REs (data not shown). We noticed that Repeatmasker reports as multiple individual entries for a RE that is interrupted by other sequence such as another RE insertion (see Figure 6 for an example), and a consolidation was performed necessarily for our purpose. RE fragments were grouped together and treated as one entry if they meet the following requirements: 1) they are in the same strand and belong to the same RE subfamily; 2) they are close to each other (with in 50kb); 3) they represent continuous sections on the RE consensus sequence. Also, since the Repeatmasker reports LTR and their internal elements separately, we grouped LTRs with their internal elements as one LTR entry. A unique ID was assigned to each RE after consolidation.

We also looked into the events TE inserting into other TE. For fragmented REs which could be consolidated, we examined if there were TEs inside the gap. As shown in Figure 6, the AluSx would be considered as a TE inserting into TE events because the two adjacent AluJb could be consolidated. This was processed during the consolidation. More detail would be discussed in later section 8.1.1;

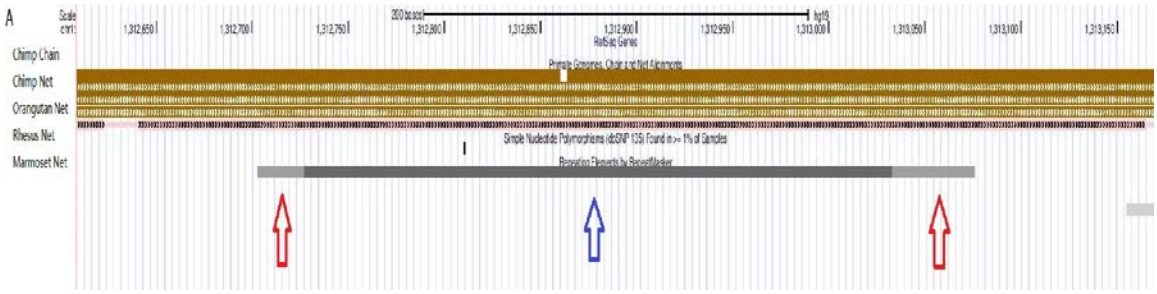

Figure 6. An example of a fragmented retrotransposon element.

Two fragmented AluJb elements (indicated by the red arrows) belong to one complete AluJb, which was later interrupted by an AluSx insertion (indicated by the blue arrow).

7.2.2. BLAT based method

BLAST-like alignment tool (BLAT) [103] was developed specially to perform fast rapid mRNA/DNA and protein alignments at a high sequence similarity level. In the BLAT based method for identifying HS-REs, DNA sequences representing the RE insertions and their flanking regions are retrieved from the human genome (hg19) based on the RE positions provided in the RE input file. Specifically, for each RE, we prepared and used three sequences as marker regions (Figure 7A). Marker 1 consists of 50 bp sequences from each side the RE insertion joined together to represent the pre-integration site marker; Marker 2 consists of 50 bp from 5' flanking region of the RE and 50 bp of the 5' end of the RE to represent the 5' junction area of the RE insertion. Marker 3 consists of 50 bp of 3' the RE and 50 bp of the 3' flanking region to represent the 3' junction area of the RE insertion. These three markers are identical in length at 100 bp. For each marker sequence, four BLAT runs were performed against the four out-group genomes. The match with the best score and sequence identity was selected among the four out-group genomes for that marker sequence. The advantage of using multiple out-group genomes is discussed later in section 8.1.1.

A total of four criteria were used for the BLAT based method: score, identity, coverage and span. The score equals the number of matches minus number of mis-matches and the number of gaps in the target sequences (Score = match – mis-match – T gap count). A higher score means a better match. The identity is a percentage of the similarity between the query sequence and the target sequence. It is calculated as 100.0 *

(match + repeat match (Rep. match)) / (match + mismatch + Rep.match). Span and coverage are used to provide additional information about the match. Span is the total length of the target sequence (Span = Target end – Target start). Since BLAT tolerates gaps in the target sequence, it is able to report matches consisting of fragments with gaps in-between. So there is possibility that an entry could have both high score and identity while having a long gap in the target sequence. Similarly coverage is used to monitor the quality of match for the query sequence: Coverage = (Q end – Q start).

If the BLAT result met all the following criteria: a minimum of score, identity, coverage and span length, a "+" sign was used to indicate that there is a positive match for the marker sequence in at least one of the out-group genomes. Otherwise a "-" sign was assigned to indicate the absence of the marker sequence at the corresponding region in all out-group genomes. The specific optimal criteria for determining a sequence match for each marker sequence, as well as a combinatorial criterion in evaluating all three markers were determined by utilizing a training dataset from dbRIP, the details of which are discussed later in section 7.2.4 and 8.2. In a perfect situation, a typical Type I HS-RE should have a combined pattern of "+/-/-" for the three marker sequences in the order of "pre-integration site/5' junction/3' junction", since if the retrotransposition insertion event happened after the human-chimpanzee separation, only the pre-integration site is present in the out-group genomes, while the two RE-flanking junction sequences will have only partial matches in these genomes (i.e., either to the flanking or RE, but not the two together at the same locus) . Similarly, a "-/-/-" pattern indicates a typical Type II HS-RE, in which the whole genome region is missing in the out-group genome. A typical

non HS-RE would have a pattern of "-/+/+", since the RE is shared by human genome and at least one out-group genome.

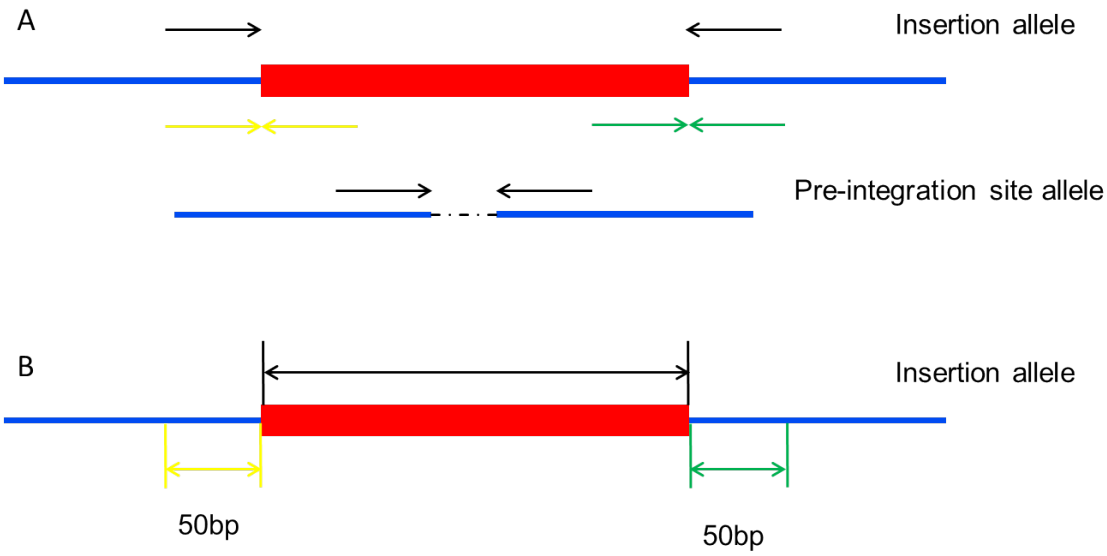

Figure 7. A schematic diagram of the *in silico* comparative genomics approach for identifying HS-REs

Panel A: The BLAT-based method. The red box indicates the RE insertion while the two blue lines indicate the flanking sequences. Each arrow stands for 50 bp sequence; the two black arrows consist of the pre-integration site, while the yellow arrows indicate the 5' junction and the green ones stand for the 3' junction. These sequences were later used for the BLAT based method as input data. Panel B: The liftOver-based method. The red box indicates the RE insertion while the two blue lines indicate flanking sequences. The green and yellow arrows, which are of 50 bp in length, stand for the chromosome coordinates of 5' flanking and 3' flanking regions, respectively. The black arrow indicates the start and end positions of the RE insertion. In the liftOver method, the chromosomal coordinates are used for finding a match in one of the four outgroup genomes.

7.2.3. liftOver based method

LiftOver is a tool originally developed by the UCSC genome team (http://genome.ucsc.edu/cgi-bin/hgLiftOver) to allow conversion of genome coordinates between different assemblies of the same genome, but its use is also extended for finding corresponding/orthologous regions between genomes of closely related species. The alignments between two genomes were constructed with BLASTZ, which is an independent implementation of the Gapped BLAST algorithm specifically designed for aligning two long genomic sequences [104]. These alignments were then chained together to form a chain file, which is a sequence of gapless aligned blocks, where for each region in the genome the alignments of the best quality were used. The resulting chain is used by liftOver to identify orthologous region between two genomes.

For the liftOver based method, the position information of each RE was subtracted from the input file. The starting and ending positions of the 50 bp upstream flanking region, the RE region, and the 50 bp downstream flanking region were lifted onto each of the four out-group genomes (Figure 7B). Similar to the BLAT-based method, the best hit for each maker region among the four out-group genomes was used in the report of that RE. A positive hit, which means the positions can be lifted onto an out-group genome, is marked as "+", while a negative hit is marked as "-". In this case, a Type I HS-RE is expected to have a pattern of "-/+/+" for the three regions in the order of 5' flanking, RE, and 3' flanking, while a typical non HS-REs is expected to have a pattern of "+/+/+". Similar to BLAT-based method, a Type II HS-RE is expected to have a pattern of "-/-/-".

The details on how to determine liftOver criteria would be discussed in section 7.2.4 and 8.2.

7.2.4. Determination of optimal BLAT and liftOver criteria for identifying HS-REs

The most challenging part of method development for identifying HS-REs is the determination of an optimal set of criteria for each of the method we used. Instead of using arbitrarily determined criteria, a training dataset containing RIP data in dbRIP was used. RIPs are results of recent retrotransposition events and show polymorphic status for its presence and absence in the human population. Therefore, human RIPs should be human specific, and it allows us to train our algorithms by checking the sensitivity of the method using criteria at different levels of stringency. A set of criteria is considered to be optimal when it generates a minimum accuracy of 95%. After the criteria for each method is determined, the BLAT and liftOver pattern for each RE can be obtained, and we classify all REs into four classes of REs based on the BLAT and liftOver patterns.

Class I: These REs are shown as Type I HS-REs supported by both BLAT and liftOver based methods with the match pattern being "+/-/-" and "-/+/+", respectively. This class of HS-REs has the highest confident level due to the presence of a reliable pre-integration site in the out-group genomes. Class II: These REs are shown as Type I HS-RE by BLAT or liftOver based method but not both, thus the confidence level would be lower. Class III: These entries have a pattern of "-/-/-" (Type II HS-REs) by both BLAT and liftOver, and thus are HS-REs with a high level of confidence. Class IV: Similar to

Class II, these entries show a pattern of "-/-/-" entries by one of the two methods, and thus are HS-REs with a low level of confidence. We chose to leave out Class II and IV for better quality.

In addition to typical Class I and typical Class III entries, there are non-canonical Class I and non-canonical Class III entries, which we name as Class Ib and Class IIIb, respectively. Class Ib entries are those that have one of its flanking sequences missing in the out-group genomes. For example, if an entry showed "+/-/-" for the BLAT based method and "-/+/-" for the liftOver based method, it would be listed as a Class Ib entry missing the 5' flanking region. Similarly if a RE has a "-/-/+" pattern for both BLAT- and liftOver-based method, it is a ClassIb entry missing the 3' flanking region. These can happen if a HS-RE insertion occurred right next to another HS-RE, resulting the absence of a match for that side of flanking sequence in the out-group genome. Class IIIb entries have a pattern of "+/-/-" pattern for the BLAT method and a pattern of "-/-/-" for the liftOver method, and this situation can happen if a HS-RE inserted into an HS-RE. In this case, despite the absence of a true orthologous pre-integration site in the out-group genomes, the BLAT search can find many high quality matches due to the presence of many sequences highly similar to the RE that was inserted into. More detailed discussion is provided later in section 8.3 and 9.1.2. Since Class Ib and IIIb represent legitimate HS-REs by definition, we included them in our final list of HS_REs.

HS-REs that involve transduction or insertion mediated deletion also have different BLAT and liftOver patterns from the canonical class I and III HS-REs. They are discussed separately in section 7.2.5 and section 7.2.6, since their identification requires additional steps of computational analysis.

7.2.5. Transduction

As previously discussed in section 6.2.1.2, the transduction events have known functional impact such as exon shuffling and disease association. For a typical 3' transduction event, the BLAT based method would return a "-/-/+" pattern and the liftOver based method would return a "-/+/-" pattern (Figure 8). For BLAT match, the pre-integration site marker would return a negative hit because of the presence of the transduced sequence in the human reference genome, while the 3' junction marker sequence would return a positive hit because the combination of the particular RE and transduced sequence appear at a different location in the out-group genomes. The liftOver based method should be able to return a "+" for the 5' flanking marker, therefore its pattern would be "-/+/-". Similarly, a typical 5' transduction event would have "-/+/-" for the BLAT pattern and "-/-/+" for the liftOver pattern. However, these patterns may also be generated due to some other reasons. Therefore, entries with these BLAT and liftOver patterns were used as the candidates for HS-REs carrying transduction and subjected to further filtering processes.

To identify HS-REs carrying transduction, it is necessary to determine the length of transduced sequence. An iterating search approach is used to identify the longest possible transduced sequence (Figure 9). For example, for a 3' transduction event, the 5' flanking sequence was located in the human reference genome and retrieved, along with 50bp from the 3' flanking (transduced sequence) was retrieved. The combination of two flanking sequences was then compared with the orthologous sequence retrieved from the

out-group genome to see if there was a positive match. If so, the entry would be stored for further validation, otherwise the script would continue down the 3' flanking for another 50bp until it reaches the arbitrary cutoff set at 20Kb.

To further validate the transduction, the flanking sequences and the transduced sequence retrieved from the human reference genome were compared against the orthologous sequences retrieved from the out-group genome again to identify the TSDs. Several BLAT runs were performed using a combination of the RE sequence and the transduced sequence against the human genome and the out-group genomes. For an authentic HS-RE involving transduction, there should be one or more copies of junction marker sequence carrying transduction elsewhere in the human genome (one for the current insertion and one for the parent copy) and one copy in one or more of the out-group genomes representing the parent RE.

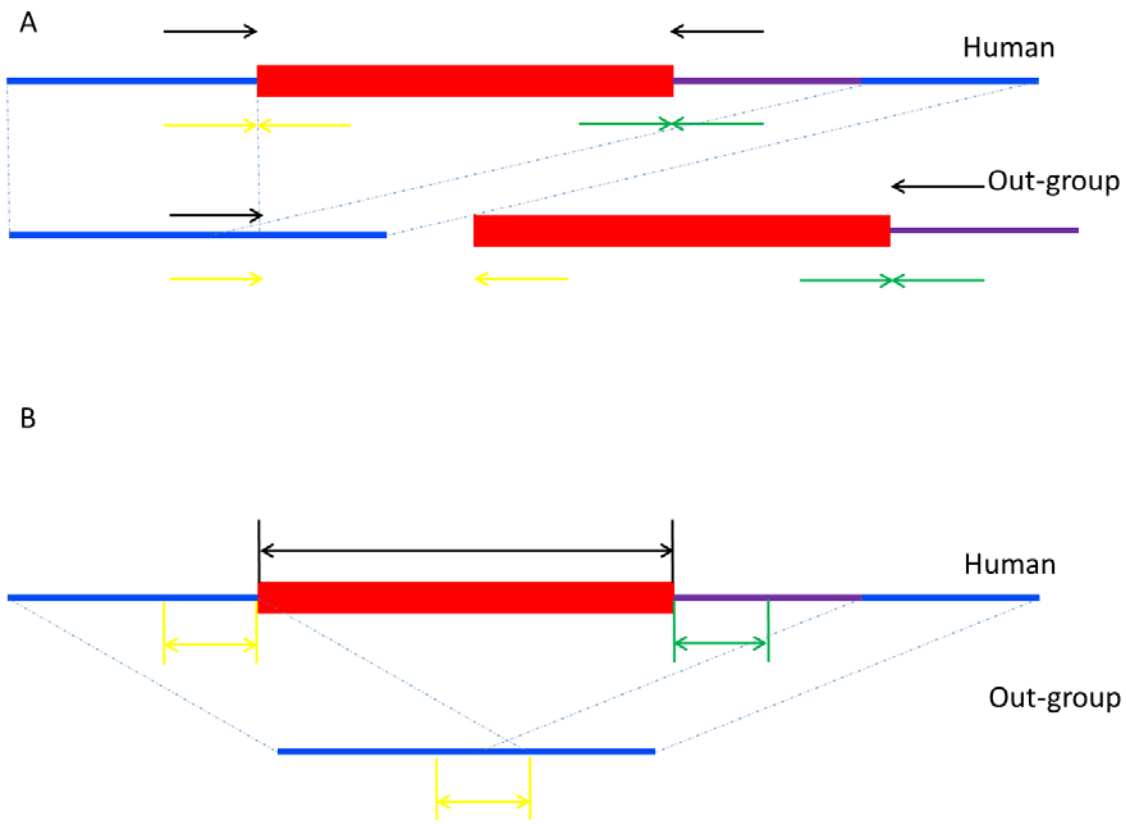

Figure 8. A schematic diagram of the algorithm for identifying HS-RE with 3' transductions

Panel A: The BLAT-based method. The upper boxes indicate RE insertion in the human reference genome. The lower boxes indicate the orthologous regions in the out-group genomes. Red box indicates the RE insertion while the blue lines indicate flanking sequences and their orthologous regions in the out-group genome. The purple line indicates the transduced sequence. The BLAT run using the pre-integration site which was represented by the two black arrows would return a negative match as the two black arrows were quite distant. Similarly, the BLAT output for 5' junction would be a "-" because the two yellow arrows are quite distant. However the BLAT run for 3' junction would return a positive result because the RE and the transduced sequence were present at a different location in the genome. Panel B: The liftOver-based method. The upper boxes indicate the RE insertion in the human reference genome while the lower boxes indicate the orthologous region the out-group genome. The red box indicates the RE insertion, the two blue lines indicate flanking sequences and the purple line stands for the transduced sequence. liftOver would only be able to find a positive match for the start and end position of the 5' flanking sequences while the start and end positions of the RE and the 3' flanking sequences were absent in the orthologous region in the out-group genome.

7.2.6. RE insertion mediated deletion

The RE-insertion mediated deletion (RIMD) was also investigated in this study, as it has known functional impact to the human genome (previously discussed in section 6.2.1.3). In such an event, instead of generating a TSD, the RE insertion causes the deletion of the region at the insertion site, leaving the apparent "pre-integration" site marker sequence without a full-length match in the out-group genomes. For this reason, in the case of a RIMD HS-RE, the BLAT based method would return a "-/-/-" pattern, because none of the three marker sequences will have a full-length match in the outgroup genomes (Figure 9). The liftOver based method would return a "-/+/+" pattern, as the two flanking regions have corresponding sequences in the out-group genomes despite of the deletion, but with a gap between the two matching regions at a distance equal to the length of the deletion, which ranges from 1bp to 70kb (previously discussed in section 6.2.1.3).

A looping strategy similar to the identification of transduction was used to identify the maximal length of deletion. For each entry, the flanking sequences in human reference genome were located. The combination of two flanking sequences is then compared with the flanking sequence retrieved from the out-group genome to see if there is a positive match. If so, the entry would be stored for further validation, otherwise the script would continue to retrieve another 50bp until it reaches the arbitrary cutoff of 70Kb.

To further validate the candidates from above step, the pre-integration site sequence retrieved from the human genome was compared against the orthologous sequences

retrieved in the out-group genome to look for evidence of TSDs. If there is a sign of TSDs, the entry would be considered as a false positive.

7.2.7. Documentation of HS-REs in dbRIP

HS-REs identified in this study have been deposited into dbRIP as a separate data set labeled using a special type of entry ID and a note. This data covers 6 different classes: Class I, Class Ib, Class III, Class IIIb, Transduction and RE insertion mediated deletion. RE sequence (labeled as red font in dbRIP) plus 400bp flanking sequence from each end (labled as blue font in dbRIP) are included. Additionally, for Class I, transduction and RE insertion mediated deletion entries, the orthologous sequences of the pre-integration site sequence in the out-group genomes is also provided and the available TSD sequences indicated in green color. The search utility in dbRIP has been modified to enable searching for HS-RE of any specific Class or subfamily.

7.2.8. Computational scripts

All computer procedures described in earlier sections were performed mostly using in-house computer scripts implemented in PERL (The *Practical Extraction and Reporting Language*), which was initially developed by Lary Wall in 1987[105]. A total of 609 scripts were developed for projects related to this thesis.

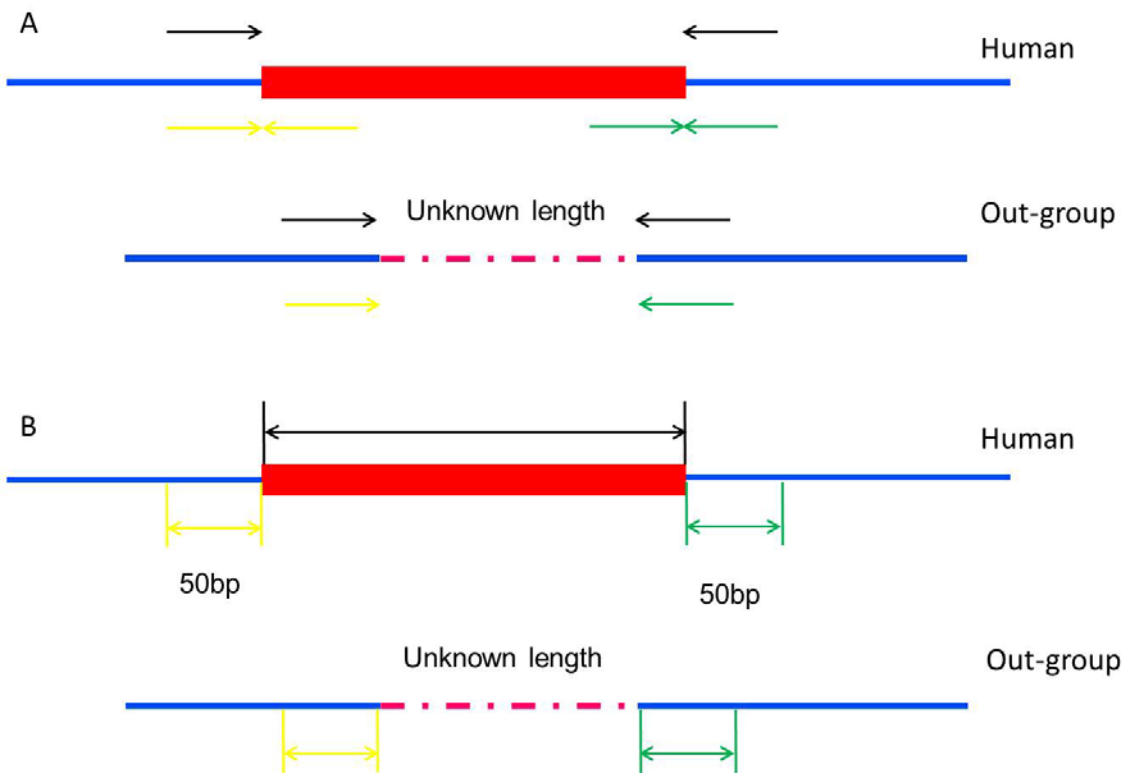

Figure 9. A schematic diagram of *in silico* comparative genomics approach for identifying typical RE insertion mediated deletions.

Panel A: The BLAT-based method. The upper boxes indicate RE insertion in the human reference genome. The lower boxes indicate the orthologous region in the out-group genomes. Red box indicates the RE insertion while the two blue lines indicate flanking sequences. The purple dashed line indicates the deleted sequence. The BLAT run using the pre-integration site sequence would return a negative match because of the deleted sequence between the two black arrows. Also BLAT results for the 5' and 3' junction would be "-" as well. Panel B: The liftOver-based method. The upper boxes indicate RE insertion in the human reference genome. The lower boxes indicate the orthologous region in the out-group genomes. The red box indicates the RE insertion while the two blue lines indicate flanking sequences. The dashed purple indicates the deleted sequence. The green and yellow arrows stand for the start and end positions of 5' flanking and 3' flanking, respectively. The black arrows stand for the start and end positions of the RE insertion. The liftOver based method would return a negative match for the insertion and positive matches for the 5' and 3' flanking. However, because the length of the deleted sequence was unknown, liftOver might only be able to locate one of the two flanking which gives a pattern of "-/+/-" or "-/-/+".

8. Results

8.1. Repeat masker input data

REs that belong to the *Alu*, L1, SVA and LTR families were retrieved from the human genome reference sequence based on RE annotation by Repeatmasker, with each entry assigned with a unique ID after consolidation for fragmented REs, and they were used as the starting point for identification of human-specific REs. Specifically, a total of initial 2,774,171 RE entries were selected out of 5,298,130 of all TEs. The initial REs were consolidated into 2,132,960 complete and non-redundant entries. This list is used as input data for all downstream analyses.

8.1.1. TE inserting into TEs

All TEs reported by RepeatMasker were retrieved from the reference genome hg19. 4,335,409 of all TEs were selected after consolidation, while 1290877(~30%) of them have inserted into anther TE comparing to 7739 HS-REs (~42%) have inserted into a TE. As previously discussed, TEs comprise ~45% of the human genome (Figure 1). Assuming TEs have equal chance inserting into TE region and non-TE region, there will be ~45% of the HS-RE inserted into a TE which is close to what we have observed. This suggests that HS-REs have no bias inserting into another TE. Assuming TEs in general have shared the same non-bias pattern, the ~30% of TE inserting into TE regions suggests that TEs comprises an average of ~30% of the human genome through the evolutionary history. Further analysis suggested that there is no TE family in particular is strongly

biased on inserting into TE regions except SVA (Table 2). SVA has a higher percentage of inserting into TE region because it is the newest RE family which originated ~25 My ago. DNA transposons which had lost their transposition abilities ~37 My ago, have the lowest percentage of inserting into TE regions. This suggests that at that time, TEs consisted of ~23% of the human genome.

Table 2 percentage of TE inserting into TE region of major TE families

Family	*TE inserting into TE region*	*Total number*	*Percentage*
DNA	87434	379846	23.0%
LINE	259489	945612	27.4%
LTR	153641	502748	30.6%
SVA	1593	3530	45.1%
SINE	527127	1692494	31.1%

We then examine the host TEs to see if any TE families are strongly biased to be host TEs (Table 3). LINE is the most successful host TE Class as ~68% host TEs are LINE elements. Statistical analysis suggests that the success of LINE elements as host TE is contributed by its large size in the human genome: a Pearson correlation test was performed which suggested that the host TE number is significantly correlated to their size in the human genome hg19 (correlation coefficient = 0.8915; two tailed P value = 0.0422; the correlation is significant).

Table 3 Percentage of different HOST TE families

Family	*Host TE number*	*Total size in hg19(Mb)*	*Percentage*
DNA	78493	98.4	6.97%
LINE	765516	632.0	68.01%
LTR	131927	264.4	11.72%
SVA	93	4.0	0.01%
SINE	149508	391.8	13.28%

8.2. Criteria determined by utilizing RIPs training dataset

To determine the optimal criteria for our methods to identify HS-REs, a training dataset containing 2331 previously documented typical RIPs with complete sequence information in dbRIP database was used. The previously described BLAT and liftOver based methods were then applied to this training dataset.

For individual RIPs, the 100 bp sequence of each of the three markers were obtained from dbRIP and were then used for BLAT against the four out-group genome. A sample BLAT output is shown in Table 4. The BLAT would align the query sequence and the target sequence and return all information related to the search such as numbers of matching and mismatching bases. It also indicates whether there is any gap in the alignment. Such information is used to determine the nature of the sequence match.

Table 4. Sample BLAT output

match	*mis-match*	*Rep. match*	*N's*	*Q gap counts**	*Q gap bases*	*T gap counts***	*T gap bases*	*Q start*	*Q end*	*T start*	*T end*
100	0	0	0	0	0	1	173	0	100	5536	5809
98	2	0	0	0	0	2	174	0	100	3979	4253

*Q: Query sequence was the 100bp sequences used for the BLAT input.
**T: Target sequence was the targeted sequence BLAT found in the out-group genomes.

A normal probability plot, which is a graphical technique to assess if a dataset is approximately normally distributed, was done for each criterion using all loci in the RIPs training set (Data not shown). Values which allow a 95% probability were selected. After manual adjustment, the optimal BLAT criteria were determined (Table 5). The pre-integration site had the lowest score cut-off as expected since the presence of TSDs (an extra copy of a small sequence) cause a gap for the alignment between Pre-integration

site sequence (a joint of the 5' and 3' flanking sequences) and the actual pre-integration site sequence from the out-group genomes. Also, the 3' junction area had lower cut-offs than the 5' junction because REs often have high mutation rate at the 3' end (poly-A tail).

Table 5. BLAT criteria based on the RIPs training dataset

	score	*Identity*	*coverage*	*Span*
Pre-integration site Area	76	92	92	110
5' Junction area	86	90	90	110
3' Junction area	82	90	85	110

When these criteria were applied to the RIPs training dataset, 1722 Type I HS-RE and 501 Type II HS-RE were identified from the total 2331 entries (Table 6). The accuracy (sensitivity) is 2223/2331 or 95.4%.

Table 6. BLAT output of the RIPs training dataset

Pattern	*count*
+/-/-*	1722**
-/-/-*	501**
-/-/+	38
-/+/-	17
-/+/+	13
+/-/+	17
+/+/-	18
+/+/+	5

* Patterns indicate HS-RE
** A sensitivity of 95.36%

The liftOver based method was also applied to the RIPs test dataset. The liftOver utilizes a criterion called minMatch, which represents the minimum ratio of bases that must map on to the new genome. For example, if the minMatch ratio is set to be 0.5, liftOver would report a positive hit when 50% of the region can be mapped on to the new

genome. The positions of each of the three markers are collected from dbRIP and are then used for liftOver. The liftOver based method returns a sensitivity ratio of 95.8% when the minMatch ratio is set to be 0.1(Table 7).

Table 7. liftOver output of the RIPs test dataset*

Pattern	*number*
-/+/+**	2207***
-/-/-**	26***
-/-/+	56
-/+/-	40
+/+/+	2

* minMatch ratio of 0.1

** Patterns indicate HS-REs

*** A sensitivity of 95.6%

Finally, the results from both methods were combined as previously described, i.e., a locus has to be called as positive for both methods using their respective parameters, and the resulted sensitivity for the RIPs test dataset was tested is 95.1%, providing an error rate slightly below the commonly accepted 5% (Table 8). Therefore, the criteria are appropriate for generating the HS-RE list.

Table 8. Combined results of the RIPs test dataset

RE Class	*number*
Class I*	1698**
Class Ib*	504**
Class III*	14**
Class IIIb*	5**
Class II	103
Class IV	7

* Class belongs to HS-REs
** A sensitivity of 95.06%

8.3. Human-specific REs

The 2,774,171 RE entries were processed using established optimal BLAT and liftOver criteria as described above and the results were shown in Table 9. We were able to identify 12156 typical Type I HS-RE and 12982 typical Type II HS-RE using the BLAT based method. As expected, the most common pattern for BLAT based result was "-/+/+" for shared REs, which was 87.1% of all REs. 28273 typical Type I HS-REs were identified by using the liftOver based method, which is significantly more than that of the BLAT based method. Also, the liftOver based results contained 5999 "-/-/+" and 5815 "-/+/-" entries, which represent Type I HS-RE with one missing flanking. The liftOver based method was able to identify 7320 RE entries showing "-/-/-" pattern. Therefore, overall, the BLAT based method appeared to higher sensitivity, perhaps also higher false positive rate, in identifying the Type II HS-REs, while the liftOver based method seems to have better sensitivity for typical Type I HS-REs, but again perhaps also higher false positive here, since we do not have a method for ascertaining the identified HS-REs without performing bench experiments.

Table 9. Match pattern for BLAT-based and lifeOver methods

Pattern	*BLAT counts*	*liftOver counts*
-/-/-	12982**	7320**
-/-/+	49054	5999***
-/+/-	75389	5815***
-/+/+	1859883	28273*
+/-/-	12156*	18
+/-/+	6897	3121
+/+/-	10960	3421
+/+/+	105639	2078993

*: Typical Type I HS-RE
**: Typical Type II HS-RE
***: Type I HS-RE missing one flanking

To assess the reliability of these HS-RE candidates, we grouped then into 4 classes, depending whether they are supported by one or both methods as described in the related method section (7.2.4). Class I consists of entries showing the "+/-/-" pattern by BLAT and "-/+/+" by liftOver for Type I HS-REs in both methods. Class III contains entries showing "-/-/-" pattern in both methods. Some of the Type II HS-REs by BLAT based method may represent false positive as a result of high level of sequence divergence between the out-group genomes and the human genome. For example, the pre-integration site can be quite diverged due to SNPs, deletions, or most likely a nearby HS-RE insertion. In this case, the BLAT based method would probably have difficulty finding the match as it is based on similarity of the relative short sequence marker, therefore showing a "+/-/-" or "-/-/-" pattern. The liftOver based method works fine for this situation because it is based on the alignment of two genomes, so it would allow partial match by showing a "-/+/-" or "-/-/+" pattern. Therefore, the Class Ib would contain entries with incomplete pre-integration site information but are still supported as human specific by both methods (Table 10).

Because of the abundance of RE elements in the human genome, there were many cases where an RE was inserted into another RE. For similar reasons, the BLAT based method would return a false positive pre-integration site marker because of the high sequence similarity (BLAT pattern "+/-/-"), while the liftOver would report it as typical Type II HS-REs (liftOver pattern "-/-/-"). Entries showing this pattern were labeled as Class IIIb.

Table 10. Combination of BLAT and liftOver results

HS-RE Class	*BLAT result pattern*	*liftOver result pattern*	*Number of entries*
Class I*	+/-/-	-/+/+	10095
Class Ib*/***	-/-/-	-/+/+	2014
Class Ib*	-/-/-	-/-/+	1944
Class Ib*	-/-/-	-/+/-	1795
Class Ib*	+/-/-	-/+/-	333
Class Ib*	+/-/-	-/-/+	321
Class III*	-/-/-	-/-/-	1981
Class IIIb*	+/-/-	-/-/-	265
Class II	-/+/+	-/+/+	9254
Class II**	-/+/-	-/+/+	2094
Class II**	-/-/+	-/+/+	1885
Class II	+/+/+	-/+/+	1591
Class II	-/+/+	-/+/-	1217
Class II	-/+/+	-/-/+	1183
Class II	+/-/-	+/+/+	1077
Class II**	-/-/+	-/-/+	1040
Class II**	-/-/+	-/+/-	994
Class II**	-/+/-	-/-/+	952
Class II**	-/+/-	-/+/-	908
Class II	+/+/-	-/+/+	800
Class II	+/-/+	-/+/+	540
Class II	+/+/+	-/+/-	256
Class II	+/+/+	-/-/+	237
Class II	+/+/-	-/-/+	172
Class II	+/-/+	-/+/-	157
Class II	+/+/-	-/+/-	155
Class II	+/-/+	-/-/+	150
Class II	+/-/-	+/-/+	33
Class II	+/-/-	+/+/-	32
Class IV	-/-/-	+/+/+	4756
Class IV	-/+/+	-/-/-	2889
Class IV	-/-/+	-/-/-	736
Class IV	-/+/-	-/-/-	656
Class IV	+/+/+	-/-/-	527
Class IV	-/-/-	+/-/+	245
Class IV	-/-/-	+/+/-	236
Class IV	+/+/-	-/-/-	145
Class IV	+/-/+5	-/-/-	121
Class IV	-/-/-	+/-/-	11

*: Entries picked for final result list
**: Entries picked as potential candidates for transduction event
***: Entries picked as potential candidates for RE insertion mediated deletion.

In total, 10095, 6252, 1981, and 265 entries were obtained for Class I, IIb, III, and IIIb, respectively (Table 10), making a grand total of 18593 entries as HS-RE candidates As a way to assess the accuracy of the result, a set of 779 novel RIPs generated from the 1000 Genome Project [98], which are not currently included in dbRIP, are collected and cross-matched with the list of 18593 HS-RE candidates. 749 out of the 779 RIP entries

were covered by the HS-RE list, providing a sensitivity of 96.1%, which we think is satisfactory.

8.4. Transduction

7873 Class II entries with the previously discussed patterns were picked as potential candidates of transduction events. Among there, 1885 entries showing "-/-/+" for BLAT and "-/+/+" for liftOver were selected as potential candidates for 3' transduction, while 2094 entries showing "-/+/-" and "-/+/+" were selected as potential candidates for 5' transduction event. In addition, 3894 entries with non-typical patterns not matching above two patterns were selected (Table 10 entries labeled as **) in order to ensure best accuracy. After the 1st round process using the loop strategy, 3766 possible cases of transduction events were obtained, from which 1241 entries were retained after filtering with a bl2seq-based script. A final list of 112 transduction events was obtained after manual examination in the USC genome browser using BLAT.

Among the final 112 entries, 38 (34%) are 5' transduction events, while the remaining 74 (76%) are 3' transduction events. Therefore, the 3' transduction event seems to be more common than the 5' transduction. As previously discussed, the mechanisms differ between 5' transduction and 3' transduction: 3' transduction is generated by a read-through transcription beyond the end of the RE element into the downstream region due to a weak termination signal within the RE, while 5' transduction events are usually associated with external promoters upstream of the RE. It is interesting to see that the ratio of the two transduction events differ significantly among

different REs. L1s had more 3' transduction events than 5' transduction events with ~80% being 3' transduction (Table 11). However, 5' transduction events were more common in Alus and SVAs: 50% of the Alu transductions and ~43% of SVA transductions were 5' transduction events. This may suggest that Alu and SVA have weaker internal promoters than L1, which means that Alu and SVA are more likely to be transcribed as a carry-on event of transcription involving external promoters located in their upstream regions. As previously mentioned in section 6.1.2 and 6.1.3, Alu and SVA are non-autonomous retrotransposons and have to hijack L1 machinery.

The transduction events among HS-REs have contributed to the size increase of human genome. SVA is the most successful family in terms of size contribution via transduction which has transduced ~42kb of genomic sequences. In total, the 112 cases of transduction events have brought an increase of 71kb sequence in addition to the inserted RE sequence to the human genome.

Table 11. Transduction events between different families

Family	*5' transduction*	*5' transduction size(bp)*	*3' transduction*	*3' transduction size(bp)*
Alu	11	784	11	2136
L1	9	3355	38	21513
SVA	17	21940	23	20006
LTR	1	1155	2	337
Total	38	27234	74	43992

RE elements have been reported to have the ability to impact human genome evolution through exon/gene shuffling mediated by transduction [106]. RE-mediated transduction inserted into intronic regions could possibly carry functional splice acceptor

sites, which could potentially lead to alternative splicing. Among the 112 cases, 27 cases have inserted into intron regions while the remaining 85 entries have inserted into intergenic regions (Table 12).

Table 12. Genes involving HS-RE transduction in the intron regions

Family	*Gene name*	*Accession number*	*NO. of intron*
L1	CTNND2	NM:001332	1/22
L1	ETF1	NM:004730	2/11
L1	MAGI1	NM:004742	2/23
L1	OXR1	NM:018002	2/16
L1	TNRC6B	NM:015088	13/21
L1	GABBR2	NM:005458	7/19
L1	ENOX2	NM:006375	2/15
L1	KCNMB2	NM:005832	1/6
SVA	VAPB	NM:004738	1/6
SVA	MAP4	NM:002375	1/19
SVA	PTPRA	NM:080841	12/23
SVA	C11orf49	NM:001003676	3/8
SVA	ABHD6	NM:020676	8/9
SVA	TTLL11	NM:001139442	6/9
SVA	ZAN	NM:003386	7/48
SVA	CEP170L	NR:003135	2/8
SVA	CASP8	NM:033355	9/10
SVA	PDXDC2	NR:003610	1/26
SVA	SNX8	NM:013321	5/11
SVA	C5orf25	NM:198567	3/9
SVA	CPA6	NM:020361	8/11
SVA	ARFGEF1	NM:006421	1/39
SVA	PTPN4	NM:002830	2/27
Alu	AS3MT	NM:020682	9/11
Alu	C1orf125	NM:144696	16/26
Alu	CPNE4	NM:130808	1/16
Alu	FAM135B	NM:015912	12/20

8.5. RE-insertion mediated deletion

2014 entries showing "-/-/-" pattern for BLAT and "-/+/+" for liftOver were picked as potential candidates for RE-insertion mediated deletion (Table 10 entries labeled as

***). After applying the loop strategy, 716 entries were filtered out and the remaining 1298 entries were subjected to further validation, and finally 155 entries were identified as RIMD events by manual examination using BLAT within the UCSC genome browser.

Among the 155 final RIMD entries, Alu contributed over 74 cases, close to 50% of all RIMDs (Table 13). Among the remaining 81 cases, 59 were for L1, and they contributed to the largest amount of deletion (~150kb). In total, the 155 entries deleted ~270k of the genomic sequences.

Table 13. RIMD events between different families

Family	*Number of RIMD*	*Deletion size(bp)*	*Average deletion Size (bp)*
Alu	74	76968	1040
L1	59	149519	2534
SVA	6	13958	2326
LTR	16	27759	1734
Total	155	268204	1730

As previously discussed, if the deleted target site DNA is located in genic regions, it might have potential to impact the function of the gene. Therefore, the locations of these RIMD entries were examined. Among the 155 entries, 37 cases had inserted into intron regions while the remaining 118 cases were in intergenic regions.

Table 14. A list of genes carrying a RMID in intron regions

Family	*Gene name*	*Accession number*	*NO. of intron*	*Deletion Size(bp)*
L1	ELANE	NM:001972	3/5	54
L1	SRGAP1	NM:020762	1/22	1416
L1	GPC6	NM:005708	2/9	40
L1	PTPN11	NM:002834	1/16	39
L1	MAML2	NM:032427	2/5	25
L1	PRKY	NR:028062	6/8	113
L1	ANAPC5	NM:016237	15/17	10981
L1	ANKS1B	NM:152788	12/26	3311
L1	THSD7A	NM:015204	6/27	42
LTR	IMPG1	NM:001563	1/17	851
SVA	HSD17B6	NM:003725	3/5	5736
SVA	RNU5D	NR:002755	1/2	15
Alu	PDE7B	NM:018945	2/13	158
Alu	MIB1	NM:020774	1/21	667
Alu	FRMD5	NM:032892	1/14	140
Alu	FANCA	NM:000135	5/43	313
Alu	NFAT5	NM:138714	2/16	33
Alu	EML5	NM:183387	21/44	2615
Alu	ERBB4	NM:005235	20/28	788
Alu	TTBK2	NM:173500	1/15	75
Alu	EYS	NM:001142800	26/43	381
Alu	CNTN5	NM:014361	6/25	46
Alu	ANKS6	NM:173551	6/15	40
Alu	TRIP12	NM:004238	1/41	597
Alu	USP32	NM:032582	27/34	27
Alu	STX8	NM:004853	4/8	27
Alu	C10orf11	NM:032024	4/6	80
Alu	DIS3L2	NM:152383	1/21	2674
Alu	SEMA3A	NM:006080	4/17	155
Alu	SCARB1	NM:005505	9/13	35
Alu	RYR2	NM:001035	10/105	113
Alu	PCSK2	NM:002594	2/12	1014
Alu	CLEC16A	NM:015226	11/23	30
Alu	GLCE	NM:015554	2/5	27
Alu	CEP63	NM:025180	2/16	299
Alu	UBE2G1	NM:003342	1/6	44
Alu	GPC5	NM:004466	7/8	900

8.6. Final HS-RE list

The transduction and RIMD entries were added to the final list, making a total of 18860 HS-REs, and this is the list used for functional impact assessment. Most of HS-

REs were Class I and Ib entries with 10095 and 6252 entries, respectively (Table 15). Grouping of these HS-REs by the family allows us to examine the transposition activity level of different RE families (see more details in section 8.7.2). In addition to those from Alu, L1, and SVA, which are known to be highly active, we also found a large number of HS-LTRs. Among the total 783 HS-LTRs, 222 were full length entries, representing ~28% among all HS-LTRs, which is significantly higher than the 9.3% full-length entries among non-HS-LTRs, suggesting that the formation of solo-LTR occurs in a time-dependent fashion.

Table 15. Stats of final HS-RE list

RE class	*Class I*	*Class Ib*	*Class III*	*Class IIIb*	*RIMD*	*Transduction*	*All HS-RE*
Alu	8372	2160	887	149	74	22	**11664**
L1	1239	2846	615	81	59	47	**4887**
SVA	329	926	216	9	6	40	**1526**
LTR*	155(34)	320(90)	263(87)	26(7)	16(3)	3(1)	**783(222)**
Total	**10095**	**6252**	**1981**	**265**	**155**	**112**	**18860**

*: total number (full length entry number)

8.7. Functional impact assessment

8.7.1. *TSD length and integration site sequence motif*

The TSD is the hallmark of an RE insertion event. It also represents a type of genomic variation generated by RE-insertion in addition to the RE insertion itself, and its sequence pattern reflects the retrotransposition mechanism involved. The pre-integration site sequence for the 10095 Class I entries were retrieved from the out-group genomes and used to compare with the corresponding human sequences. By utilizing a BLAST2sequence tool (bl2seq), we were able to identify TSDs for 9222 entries.

The average lengths of the TSD were examined as it is an important characteristic of RE insertion (Table 16). *Alu* family had the longest average TSD length at 18.7bp with standard deviation being 14.8 bp, while L1 and SVA have similar length of TSD at 15.5bp and 16.4bp, respectively. The LTR had the shortest TSD pattern among all families, only 8.6bp on average. Although not significant, the average TSD length does seem to have a weak negative correlation with the average RE length of the RE family (correlation coefficient is -0.19; one tailed P value = 0.4041).

Table 16. Average length of TSD of each RE family

Type	*Copy number*	*Average TSD length(bp)*	*Standard deviation(bp)*	*Average RE length(bp)*
HS-*Alu*	7723	18.7	14.8	312.2
HS-L1	1061	15.5	9.8	6146.8
HS-SVA	305	16.4	6.9	1383.8
Hs-LTR	133	8.6	5.8	2272.6
HS-RE	9222	18.1	14.1	1047.2

Overall, HS-*Alu*, HS-L1 and HS-SVA have TSD length around 16 bp. This confirms the existing theory that *Alu* and SVA hijack L1's machinery for retrotransposition, and so they share the same mechanism for transposition. While the utilization of Alu transposition via L1 machinery seems to be supported by experimental data, the use of L1 machinery for SVA transposition is purely hypothetical so far. Therefore, our data here provides the first evidence supporting this hypothesis. The fact that LTRs have significantly shorter TSD length than the three non-LTR families (Table 17) agrees with

the fact that LTRs use a different retrotransposition mechanism and the hypothesis that the pattern of TSD (length and sequence characteristics) is determined by the retrotransposition mechanism.

Table 17. Z-test results of average TSD lengths of different RE families

2-tailed P value	*HS-Alu*	*HS-L1*	*HS-SVA*	*HS-LTR*
HS-*Alu*	1	< 0.0001	0.0002	< 0.0001
HS-L1	< 0.0001	1	0.0699	< 0.0001
HS-SVA	0.0002	0.0699	1	< 0.0001
HS-LTR	< 0.0001	< 0.0001	< 0.0001	1

We also examined and compared the sequence motif of the integration sites among different RE families by taking the 15bp sequence on each side of insertion sites. The HS-*Alu*, HS-L1 and HS-*Alu* show tt-AAAA pattern at the nick site, which agrees with results from previous studies [90](Figure 10). However, our data suggested that there is a low but evident bias on the 7^{th} and 8^{th} bp for "A", and this extends the nick site pattern to "tt-AAAAAA". The HS-LTR, however, did not show a visible motif at the nick site, and this further proves that *Alu*, SVA and L1 share the same machinery while LTRs utilize a different machinery for retrotransposition.

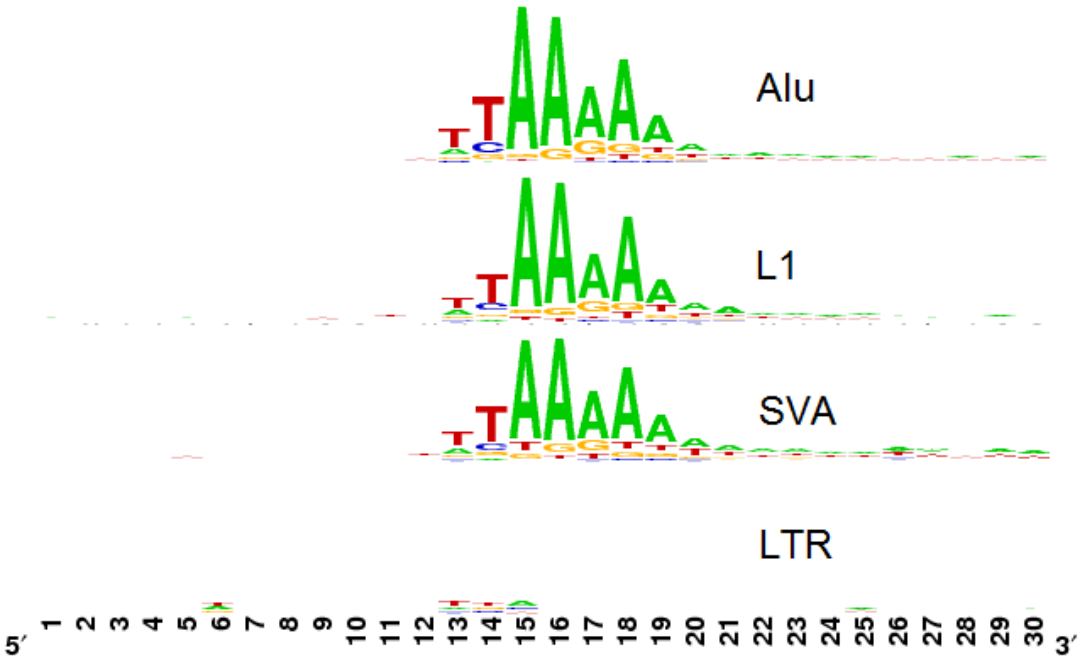

Figure 10. Base composition of insertion motif for different HS-RE families.

Sequence logo generated by using the weblogo tool (http://weblogo.berkeley.edu/logo.cgi). The numbers on the x-axis indicate the base position starting from the 15th base on the 5' side of the pre-integration site. The height of the base letter indicates the frequency of the most dominant base at the position.

8.7.2. Retrotransposition activity of HS-RE subfamilies

The availability of a complete set of HS-REs allowed us to provide an accurate assessment of the activity of RE families in the human genome. For this, we use the human-specific ratio (HS-ratio), which is defined as the ratio between the number of HS-REs and the total number of members for a subfamily in the human genome sequences (Table 18). Since HS-REs are products of recent retrotransposition events, having a higher HS-ratio means there are more members relatively contributed by recent

retrotransposition events. Therefore, we used the HS-ratio as an indicator for the activity level of different RE families.

Table 18. Subfamily distribution of different HS-RE families

Class	*Family*	*Subfamily*	*Total copy number in hg19*	*Copy number of HS-REs*	*HS-ratio**	*size contribution (kb)*
SINE	Alu	AluYa5	3857	3218	83%	996.4
SINE	Alu	AluYb8	2820	2279	81%	710.8
SINE	Alu	AluYb9	327	258	79%	72.1
SINE	Alu	AluYd8	225	151	67%	44.8
SINE	Alu	AluYg6	784	386	49%	119.3
SINE	Alu	AluYk12	212	80	38%	23.0
SINE	Alu	AluYa8	325	118	36%	23.0
SINE	Alu	AluYf4	1347	298	22%	92.3
SINE	Alu	AluYc5	45	6	13%	1.6
SINE	Alu	AluYh9	265	29	11%	8.1
SINE	Alu	AluYk11	1026	83	8%	23.1
SINE	Alu	AluYc3	567	41	7%	12.2
SINE	Alu	AluYc	8268	257	3%	39.0
SINE	Alu	AluYk4	1839	41	2%	12.1
SINE	Alu	AluY	117010	2446	2%	729.0
SINE	Alu	AluYf5	180	3	2%	1.0
		All Alu	**1083462**	**11664**	**1%**	**3377.4**
LINE	L1	L1HS	1484	1135	76%	2574.4
LINE	L1	L1PA2	4699	1977	42%	3811.6
LINE	L1	HAL1-2a_MD	1281	197	15%	37.3
LINE	L1	L1P1	2869	209	7%	148.9
LINE	L1	HAL1-3A_ME	1824	103	6%	11.9
LINE	L1	L1PA3	10072	348	3%	676.8
LINE	L1	L1P4e	121	3	2%	0.2
LINE	L1	L1M4b	2412	49	2%	25.2
LINE	L1	L1P2	1245	21	2%	9.0
LINE	L1	L1P3b	65	1	2%	0.6
LINE	L1	L1M2a1	72	1	1%	0.1
LINE	L1	L1MA5	2609	26	1%	4.3
		All L1	**546108**	**4887**	**1%**	**7653.7**
Other	Other	SVA_E	226	155	69%	176.6
Other	Other	SVA_D	1329	817	61%	1142.9
Other	Other	SVA_F	986	494	50%	506.6
Other	Other	SVA_C	278	33	12%	43.4
Other	Other	SVA_B	458	20	4%	25.7
Other	Other	SVA_A	253	7	3%	8.9
		All SVA	**3530**	**1526**	**43%**	**1904.1**
LTR	LTR	ERVK	7351	218	3%	420.9
LTR	LTR	ERV1	107223	344	0%	663.5
LTR	LTR	ERVL	109629	103	0%	51.9
LTR	LTR	ERVL-MaLR	257325	115	0%	11.9
LTR	LTR	Gypsy	8356	3	0%	0.2
		All LTR	**499860**	**783**	**0%**	**1148.4**
		All RE	**2132960**	**18860**	**1%**	**14083.6**

*Subfamilies with HS-RE ratio >= 1% are listed except HS-LTR, for which anything above 0% are listed.

Most of the HS-REs belong to the *Alu* family and a lot of those HS-*Alu*s are within the *Alu*Y linage. This made HS-*Alu* the most successful and active RE family in terms of copy number. Our data confirms the previous observations that Ya5 and Yb8 are very active *Alu* subfamilies within the human lineage (83.43% and 80.82% HS-ratio), while Yb9 and Yd8 are also very active having 78.90% and 67.11% as HS-ratio. Additionally, *Alu* subfamilies Yg6, Yk12, Ya8, Yf4 and Yc5 have higher retrotransposition activity than average.

Most of L1 subfamilies have very low level of activity while the L1HS, L1PA2 and HAL1-2a_MD subfamilies seem to have maintained a significant level of retrotransposition activity as their HS-ratio are close to most of *Alu* and SVA subfamilies. L1s are the most successful HS-REs in terms of size contribution (Figure 12).

As previously reported, SVA subfamilies are very active. Overall, it has higher activity among all HS-RE families. Among the 6 subfamilies, SVA_D, SVA_E and SVA_F have a significant higher level of activity than the rest. This agrees with the fact that they are the newest group of REs in the human genome [26].

The LTR family has the lowest activity level as expected. Our data confirm the previous observation that the HERV-K subfamily is the only current active LTR subfamily [107] as it has significant higher HS-ratio, more than 10 times higher than the rest of LTR subfamilies. This might indicate that the current activity of HERV-K subfamily has been underestimated: the HERV-K subfamily had a 2.97% HS-ratio which was higher than the AluY family.

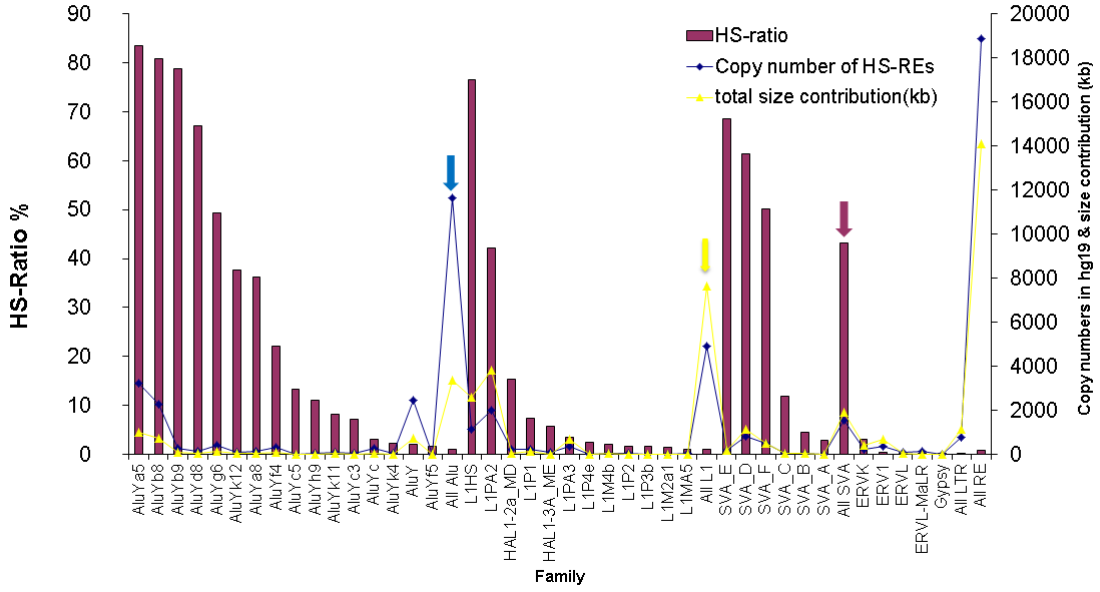

Figure 11. Subfamily distribution of HS-RE.

The x-axis indicates different RE subfamilies; The left y-axis indicates HS-ratio in percentage; The right y-axis indicates RE copy numbers in human genome hg19 and their size contribution (kb); Purple bars stand for HS-ratio for each RE subfamily; Blue line indicates copy number for each HS-RE subfamily; Yellow line indicates total size contribution of each HS-RE subfamily; Blue arrow indicates the peak of copy numbers of HS-REs in the human genome; Yellow arrow indicates the peak of size contribution of HS-L1 in the human genome; Purple arrow indicates the highest overall

8.7.3. Distribution of HS-REs throughout the human genome

The distribution pattern of HS-REs throughout the human genome was examined to check whether there is any bias for specific chromosomes, the relationship between gene density and RE density, and the relationship with HS-REs and shared REs. By analyzing the RE density data and gene density data, we were able to identify the relationship between RE density and gene density within each RE family (Table 19). The result of Pearson correlation test suggests that HS-Alu density is positively related to gene density, but the correlation is not significant (correlation coefficient= 0.07; the two tailed P value = 0.7402). Shared Alu density is significantly related to gene density (correlation

coefficient= 0.95; two tailed P value < 0.0001). HS-L1 density show a negative correlation with gene density but the correlations is insignificant (correlation coefficient = -0.13; two tailed P vale = 0.5392). The shared L1 density is negatively correlated to gene density while the correlation appears to be significant (correlation coefficient= -0.48; two tailed P vale = 0.0146). The HS-SVA density and shared SVA density are significantly correlated to the gene density with a correlation coefficient of 0.83 (two tailed P value <0.0001) and 0.95 (two tailed P value <0.0001), respectively. HS-LTR density is showing an insignificant negative correlation with gene density while shared LTR density is showing a significant negative correlation with gene density (correlation coefficient = -0.05 and -0.44, two tailed P value = 0.8038 and 0.0296, respectively).

Table 19. RE and gene density in the human genome

Chromosome	*Gene density (per Mb)*	*Alu density (per Mb)*		*L1 density (per 500Kb)*		*SVA density (per 100Kb)*		*LTR density (per 100Kb)*	
		HS-Alu	*Shared Alu*	*HS-L1*	*Shared L1*	*HS-SVA*	*Shared SVA*	*HS-LTR*	*Shared LTR*
chr1	14.8	3.8	414.2	2.7	335.5	7.1	8.5	1.2	161.0
chr2	8.4	3.7	322.8	3.4	378.2	5.1	6.5	0.9	173.6
chr3	9.2	3.8	314.7	3.1	381.3	5.4	6.1	1.4	176.7
chr4	6.5	3.7	268.2	4.2	373.3	3.7	5.4	1.0	200.1
chr5	8.0	3.8	296.1	3.6	382.3	4.6	6.4	1.5	183.6
chr6	10.0	3.6	319.5	3.1	374.8	5.3	5.9	1.6	170.2
chr7	10.2	3.6	413.2	3.0	384.2	5.6	6.9	0.9	167.6
chr8	8.4	3.4	316.4	3.2	377.0	3.8	5.5	1.5	186.1
chr9	10.4	4.2	374.3	2.7	393.3	5.8	7.8	1.6	171.7
chr10	10.4	3.3	392.4	2.4	375.3	4.8	6.9	1.2	164.1
chr11	15.3	3.7	336.4	3.4	359.2	5.4	7.5	2.0	161.3
chr12	12.7	3.7	414.1	3.1	359.4	5.2	7.3	2.0	176.4
chr13	5.7	4.0	277.5	2.9	394.8	3.8	4.4	1.7	192.0
chr14	12.3	3.5	376.3	2.9	365.4	5.3	7.8	1.4	178.1
chr15	13.8	3.7	433.7	2.5	390.8	5.0	7.8	1.8	147.5
chr16	16.3	4.3	587.6	2.6	357.7	5.3	7.0	1.5	179.4
chr17	24.4	4.6	669.3	1.8	332.8	9.1	10.4	2.2	134.3
chr18	6.1	3.7	301.1	3.4	385.2	2.8	5.9	1.1	177.5
chr19	40.9	8.0	930.3	2.9	327.6	12.0	15.2	6.5	161.9
chr20	15.4	3.7	443.5	3.1	379.8	8.9	7.6	0.8	201.3
chr21	12.2	3.5	330.6	2.6	379.8	2.0	5.7	1.7	229.7
chr22	21.9	4.1	635.8	3.3	313.7	10.0	9.5	2.9	142.1
chrX	12.1	4.4	286.2	5.5	512.8	4.2	7.3	3.9	178.5
chrY	9.9	27.3	327.1	26.0	398.6	1.6	4.7	123.2	172.1
Whole genome	11.7	4.1	374.6	3.4	378.3	5.3	7.0	2.7	174.4

When the density of HS-REs and shared REs are compared among chromosomes (Figure 12), it is interesting to see that HS-Alu, HS-L1, and HS-LTR all have their density peaks at chromosome Y, which are much higher than other chromosomes. More interestingly, it seems from the graph that the HS-Alu density and HS-LTR density are correlated to gene density except at chromosome Y even though previous Pearson correlation tests suggest otherwise. Therefore, another set of Pearson correlation tests were done by excluding the data from chromosome Y. By these tests, HS-Alu density is showing a significant correlation with gene density (correlation coefficient = 0.85; the two tailed P value <0.0001). The HS-LTR is showing a significant correlation with gene density (correlation coefficient = 0.79; the two tailed P value <0.0001), while correlation level did not change for the rest comparisons. The unusual high density of HS-REs in Y chromosome in unexpected, and we suspect that it may be due to a higher rate of false positive as a result of insufficient sequence data for Y chromosomes in the other primate genomes. The fact a higher ratio of type II HS-REs is seen in Chromosome Y than in other chromosomes is also an indication for this possibility.

We also suspect that the gap on chimpanzee chromosome Y, which is the only available out-group chromosome Y, is the reason for high numbers of Type II HS-RE on the human chromosome Y. If a RE is located within the homologous regions of such gap regions on the human chromosome Y, it would be reported as a Type II HS-RE which could potentially be false positive. In order to find out how many Type II HS-REs were contributed by those gap region, we identified the homologous regions of those gap regions on the human Y chromosome by using liftOver. Surprisingly, none of our

identified HS-REs are located within these homologous regions. This suggests that chromosome Y bias of HS-REs might be valid.

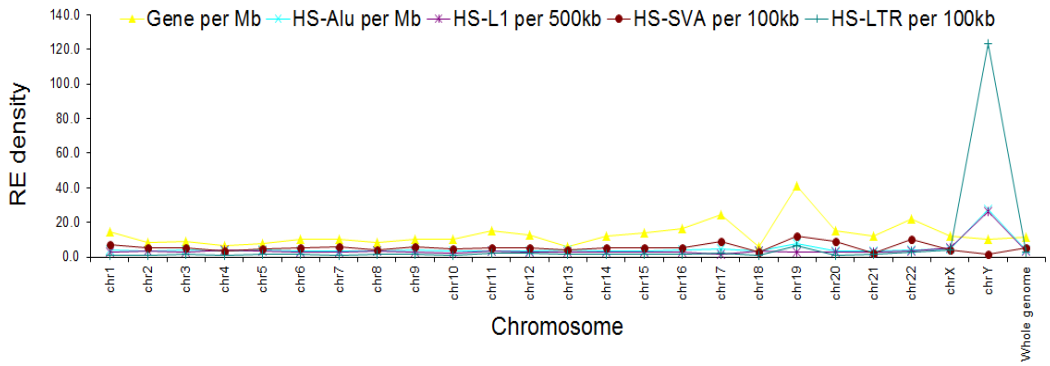

Figure 12. The density of HS-REs and genes in the human chromosomes.

The y axis indicates the RE and gene density. The x axis indicates different chromosomes.

8.7.4. GC content for different HS-RE families and all other members in the same families

Based on previous studies, L1s tend to insert into GC-poor region and Alus tend to insert into GC-rich region [108]. So, it would be interesting to find out if HS-REs share the same pattern with the older RE members in their families. Different sizes of the flanking region were tested to find out the optimal cut-off. According to the results, 1500bp from both upstream and downstream of the RE insertions were used for calculating the GC-content (Table 20).

Table 20. GC-content data of different HS-RE families and older members in the RE family

	HS-RE			*Older RE*		
	Copy numbers	*GC content*	*Standard deviation*	*Copy number*	*GC content*	*Standard deviation*
Alu	11664	40.5%	6.1%	1071798	42.1%	5.8%
L1	4887	38.4%	5.4%	541221	39.0%	5.4%
SVA	1526	45.9%	8.1%	2004	44.9%	8.1%
LTR	783	41.6%	6.5%	499077	40.3%	5.6%

All RE	18860	40.5%	6.4%	2114100	40.8%	5.7%

Z-test was performed to test if the patterns of GC-content of different HS-RE families are significantly different from each other, as well as from older members in their families (Table 21). The results indicate that each HS-RE family has a unique GC-content pattern when comparing to older members in the family. Also, the GC-content pattern for each HS-RE family is significantly different than other HS-RE families. HS-Alu and HS-L1 have lower GC-content than the older REs in their families, while HS-SVA and HS-LTR have higher GC-content than their older counterparts. Among the four HS-RE families, HS-SVA has the highest GC-content, while HS-L1 has the lowest GC-content.

Table 21. Z-tests result for different HS-RE families

2-tailed p value	*HS-Alu*	*HS-L1*	*HS-SVA*	*HS-LTR*
Older RE	< 0.0001	< 0.0001	0.0003	< 0.0001
HS-Alu	1	< 0.0001	< 0.0001	< 0.0001
HS-L1	< 0.0001	1	< 0.0001	< 0.0001
HS-SVA	< 0.0001	< 0.0001	1	< 0.0001
HS-LTR	< 0.0001	< 0.0001	< 0.0001	1

8.7.5. Genome size contribution by human-specific retrotransposons

Genome size is a property characteristic of a species. Transposable elements are a major factor impacting the genome size. Here, we examined the impact of HS-REs on the human genome size with breakdowns into individual chromosomes (Table 22). In total, the HS-REs have contributed to ~ 14 Mb size increases, which represent ~0.49% of the entire human genome. T-test indicates that the overall size increased ratios of most chromosomes are not so different from the size increased ratio of the whole genome. Yet the size increase ratio of the chromosome Y is significantly larger than that of the other chromosomes (5.36% compare to a whole genome average of 0.49), and this was mostly contributed by HS-LTRs and HS-Alus as previously described. Interestingly, HS-LTR on chromosome Y has contributed to a 725.6kb increases which consists of ~63.2% of its total contribution to the whole genome.

Based on our result, it could be concluded that because of the bias of HS-RE (which would later be discussed in section 9.2.2), chromosome Y is receiving significantly more size increase than chromosome X and autosomes: a one tailed one sample T test is performed utilizing increase size ratio data (Table 22). The null hypothesis is that the average size increase ratio in the human genome is significantly smaller than the size increase ratio on chromosome Y. Based on the T test result (one tailed P value< 0.0001; alpha = 0.05), we failed to reject the null hypothesis. This proves that chromosome Y has significantly higher size increase ratio than any other chromosomes in the human genome.

Table 22. Size contribution by HS-RE on different chromosomes

Chr	*Non gap length(Mb)*	*Total size increase by HS-RE(kb)*	*Increased size ratio %*
chr1	225.3	1040.0	0.46
chr2	238.2	1108.4	0.47
chr3	194.8	1002.8	0.51
chr4	187.7	983.2	0.52
chr5	177.7	864.4	0.49
chr6	167.4	710.8	0.42
chr7	155.4	622.9	0.40
chr8	142.9	588.3	0.41
chr9	120.1	460.8	0.38
chr10	131.3	461.2	0.35
chr11	131.1	611.8	0.47
chr12	130.5	556.7	0.43
chr13	95.6	406.2	0.42
chr14	88.3	391.5	0.44
chr15	81.7	294.6	0.36
chr16	78.9	284.9	0.36
chr17	77.8	291.7	0.37
chr18	74.7	346.4	0.46
chr19	55.8	313.4	0.56
chr20	59.5	250.1	0.42
chr21	35.1	109.6	0.31
chr22	34.9	136.8	0.39
chrX	151.1	872.7	0.58
chrY	25.7	1374.5	5.36
Whole genome	2861.3	14083.6	0.49

8.7.6. Gene context of human specific retrotransposons elements

As a way of assessing the potential impact of HS-REs, we examined their distribution in relation to genes by categorizing them into different genic regions. While most of HS-REs have inserted into intron and intergenic region, there are 85 of them have inserted into exon and promoter regions (Table 23). This suggests that HS-REs have potential impact on gene expression in the human genome.

Table 23. Gene context information of different RE families

	Gene context loci				
Family	exon	promoter	downstream	intron	intergenic
Alu**	25/4586	36/46	26/4223	4150/486227	7427/609437
L1	8/1986	4/2493	6/1319	1342/307310	3527/620489
SVA	5/6	3/6	4/4	606/751	908/1315
LTR	1/1520	3/1743	0/1196	114/134808	665/463452
All RE	39/8098	46/11852	36/6742	6212/929096	12527/1694693

*: downstream up to 1kb; promoter up to 1kb;

**: HS-RE/non HS-RE

The genes with HS-RE insertions in their exon regions are investigated individually (Table 24). Among the 39 cases, 3 REs inserted into coding sequences (CDS); 2 REs inserted into 5' untranslated region (5' UTR); 31 cases have inserted into 3' UTR region while the rest 3 REs have inserted into non-coding region.

As previously discussed in section 6.2.2, many REs contain in their sequences the binding sites for transcriptional factors, thus HS-REs in the promoter regions can have the potential to alter gene expression by creating new sites for transcriptional factors, aside from interrupting existing transcription factor binding sites by insertions. REs inserted into 3' UTR region can down regulate gene expression. For example, A-to-I editing of pairs of opposite directional *Alus* in the 3' UTR region can suppress expression through nuclear retention of mRNA transcripts [77]. In the case of BBS5, a gene

encoding a protein that has been directly linked to Bardet-Biedl syndrome, an HS-SVA inserted at the end of last exon forms a long 3' UTR region that is unique to human.

Table 24. A list of genes containing HS-REs in exon regions

Region of insertion	*Official Gene Symbol*	*Inserted RE type*	*Accession Number*
CDS	OVCH2	AluYa5	NM_198185
CDS	RPGR	HAL1-3A_ME	NM_001034853
3UTR	AAK1	L1PA3	NM_014911
3UTR	ABCC9	L1PA2	NM_020297
3UTR	BBS5	SVA_D	NM_152384
3UTR	BCDIN3D	AluYa5	NM_181708
3UTR	C5orf36	AluYb8	NM_001145678
3UTR	CCDC122	AluYb8	NM_144974
3UTR	DNAJC21	AluYa5	NM_194283
3UTR	DSG3	AluYb8	NM_001944
3UTR	FAM119A	SVA_E	NM_001127395
3UTR	FAM119A	SVA_E	NM_001127395
3UTR	FOXK1	AluY	NM_001037165
3UTR	GBP4	AluYg6	NM_052941
3UTR	GNB5	AluYc3	NM_016194
3UTR	HIST2H2BF	AluYa5	NM_001024599
3UTR	IGFL4	L1PA2	NM_001002923
3UTR	KIAA0101	AluYb8	NM_014736
3UTR	KIAA0319L	AluSc8	NM_024874
3UTR	LRRC58	AluY	NM_001099678
3UTR	MBOAT1	AluYa5	NM_001080480
3UTR	NEK5	SVA_D	NM_199289
3UTR	OPHN1	L1HS	NM_002547
3UTR	PDDC1	MLT1A0	NM_182612
3UTR	RAB21	AluY	NM_014999
3UTR	RAB3B	AluYb8	NM_002867
3UTR	SEMA3E	AluYb9	NM_012431
3UTR	SLC13A1	AluYa8	NM_022444
3UTR	TBC1D15	AluYa5	NM_022771
3UTR	UTP11L	AluYb8	NM_016037
3UTR	WIPF2	AluYc	NM_133264
3UTR	ZNF543	AluYa5	NM_213598
5UTR	CHRM3	L1HS	NM_000740
5UTR	TRIB3	AluY	NM_021158
NC	C14orf23	AluYb8	NR_026731
NC	OR6W1P	SVA_E	NR_002140

9. Discussion

9.1. In silico identification of human specific retrotransposon elements

In this study, we have presented a comparison between the human genome and four other primate genomes towards identifying and characterizing human specific RE insertions. The computational approach used in this study represents an efficient and accurate method to identify all HS-REs in the human genome. In our study, a total of four out-group genomes were used to ensure maximum accuracy. Furthermore, two different computational tools were used, and the methods were trained with known RIPs in dbRIP and their sensitivity validated using novel RIPs outside dbRIP. In total, 18860 or ~0.9% of all RE elements in the human genome were conservatively identified as human specific. Our study represents the most comprehensive analysis of human specific retrotransposons to date, providing new insights on the degree of retrotransposition in the human genome and the impact on genome evolution and function

9.1.1. Value of using four additional primate genomes in identifying HS-REs.

Studies have suggested that REs had played an important role in the primate evolutionary history [106, 109-111]. Although several groups have focused on identifying REs that are uniquely present in the human genome ever since the human genome sequence was published in 2002, they were either focusing on either a particular RE subfamily/family or a particular out-group family which was limited by the lack of resources at the time of their studies [112-115]. This is the first comprehensive study

utilizing the human reference genome with a larger number of closely related primate out-group genomes.

The biggest advantage to utilize four out-group genomes is that it ensures maximum accuracy. As variations accumulate during the long primate evolutionary history, the out-group genomes can be quite diverged compared with the human genome for certain regions in the genomes. This would affect the quality of the results if only one out-group genome is used. For example, although there is no known mechanism to specifically remove a RE element, the whole region could be deleted in the out-group genome. If a locus is completely deleted in an out-group genome, it will generate false-positive of a HS-RE based on that genome. However, the chances for all four out-group genomes to share a deletion at the same locus are very small. As an example for this, a RE element is present in both Orangutan genome and Rhesus monkey genome but is absent in the Chimpanzee genome (Figure 13). This is truly a shared insertion and the absence in the chimpanzee genome could have been caused by either a deletion in the chimpanzee genome or an error during sequencing/annotation of the chimpanzee assembly. Even though the chimpanzee genome is the least diverged when compared with the human genome among all the out-group genomes, the use of other out-group genomes helps to eliminate this and other types of false positives.

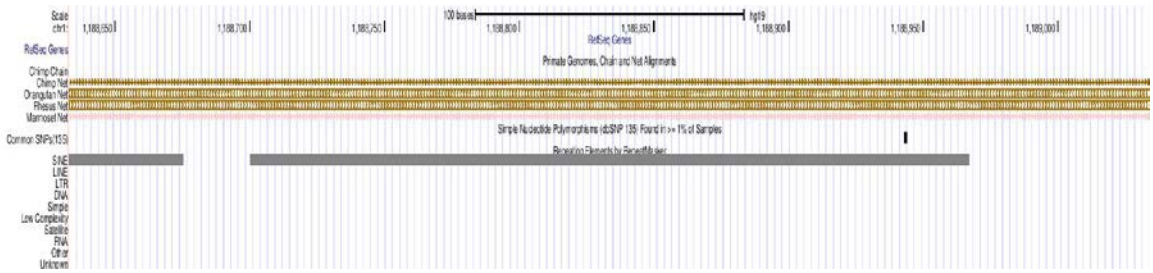

Figure 13. A screenshot of a shared RE elements in UCSC genome browser

The black box which is labeled as SINE indicates an Alu insertion. Brown boxes indicate the orthologous sequence is present in the out-group genome while open line suggests that it is missing.

9.1.2. Combination of BLAT and liftOver based methods

The reason for using BLAT based method and liftOver based method is that they both have limitations. The main principle of the BLAT based method is to examine the human specific status of a RE by searching for similar sequences of the REs junction areas. However, as previously discussed, mutations can easily occur during the long primate evolutionary history. So if the junction area is too diverged, BLAT won't be able to report it as a positive hit, this will result in false-positive HS-REs. One the other end, a sequence can be due to the presence of a tandem duplication of the whole insertion region or an RE inserted into same type of REs.

The liftOver based method relies on the previous generated chain file. Unlike the BLAT based method, which utilize short local sequences, the chain files are generated by constructing alignments between two genomes where the examined region are much longer. However, it has its own drawback. As the liftOver is based on a non-overlap chain, each group of positions could have only one hit on the out-group genome. If tandem duplications of the region containing REs happen after the human-primate

divergence, liftOver would treat the duplicated REs as HS-RE while they are true negatives. Whereas the BLAT based method can find a good hit of the original RE since it's based on sequence similarity. Also, there are mistakes in the liftOver chain file as well. For example, for entry 00644388, the liftOver based method returns a "-/+/+" pattern indicating it is a Type I HS-RE. However, the BLAT based method suggests that it is truly a shared RE ("-/-/+"). Manual review has confirmed the BLAT based result. The liftOver based method has lifted the 5' flanking of the RE insertion onto a gap region on the chimpanzee genome.

The combination of BLAT and liftOver based results have provided a more accurate list of HS-REs. Among all the classes, Class I and Class III have the highest confidence level since they are supported by both BLAT and liftOver based methods. Class I cases can be a valuable resource for studying mechanism of retrotransposition since they provide detailed information of the pre-integration site that are useful for identifying TSDs, while there is no accurate pre-integration site information for most of the Class III entries.

9.1.3. The value of using RIPs as training and testing dataset

The biggest challenge in the method development of this study is how to determine the optimal criteria for identifying HS-REs. Instead of arbitrarily assigning each criterion, a training dataset containing known RIPs are used to generate a simulation in the human genome.

As previously discussed, the RIPs in the human genome represent RE insertions that are polymorphic within humans as a result of RE insertions that have occurred after the separation of humans and chimpanzees, thus they should be human specific. This unique characteristic of RIPs has provided an ideal training dataset for identifying HS-REs. Therefore, we choose to use a set of parameters for the BLAT based and liftOver based methods that provide a sensitivity of at least 95% in detecting RIPs as HS-REs. The specific criteria are as previously discussed in section 7.2.4 and 8.2.

To assess the quality final HS-RE list, a validation dataset is also used. As described in section 7.3, a total of 779 novel RIPs, not included in dbRIP, are used to examine the sensitivity of our analysis method, and the test indicated a sensitivity of 96.1%, suggesting that our HS-RE list should have covered the most of such REs in the human genome. Unfortunately, unless we experimentally validate these HS-REs by confirming their absence in the orthologous regions in the other primate genomes, we will not be able to assess the false positive rate of our data. With experimental validation, it would be very difficult if possible to confirm Type II HS-REs due to the absence of the flanking genome regions in the other genomes. However, with the further improvement of reference genome sequences for all involved species, we can expect that some of the false positives may be identified and eliminated.

9.2. HS-RE activity level

Recent studies have suggested that the majority of known polymorphic *Alu* insertions belong to the close related Y subfamilies [116-118]: among the 2569 *Alu* RIPs in dbRIP, 2491 of which are members from *AluY* subfamily [100]. The Ya5 and Yb8 are the major subfamilies among the Y family, containing 809 and 918 entries in dbRIP, respectively [100]. Data suggested that L1HS is the major subfamily of known polymorphic L1 insertion: 446 entries among 598 total L1 RIPs in dbRIP [100]. SVA_E SVA_D and SVA_F made up the majority of known polymorphic SVA insertions. HERV-K is the only subfamily, which is known to have current activity in the human genome [107].

In these studies, researchers have mainly used RIPs as a tool to evaluate the activity level of certain subfamilies during human evolution [90]. The human and chimpanzee divergence happened ~6 My ago and modern humans and Neanderthals likely diverged in Africa ~300,000–700,000 yr ago[119]. This suggests that during evolutionary history, there would be a proportion of REs in the human genome, which is unique to human. Although RIPs would be human specific, they represent only a small portion of all HS-REs, and it would be insufficient for illustrating the retrotransposition activity level of a certain subfamily during human evolutionary history. Since our method imposes no bias to identify a specific subfamily, our results provide the first large and unbiased data set suitable for assessment of RE insertion activities after the human and chimpanzee divergence.

While our data confirms the previous observation that Ya5 and Yb8 are very active *Alu* subfamilies within the human lineage, it also suggests that other *Alu* subfamilies such

as Yb9, Yd8, Yg6, Yk12, Ya8, Yf4, Yc5 and Yh9 also have very high retrotransposition activity. The *AluY* is the most successful *Alu* family within the human genome. As confirmed by our result, L1HS is the most successful L1 subfamily. Moreover, our data suggests that L1PA2 and HAL1-2a_MD are very also active in the human lineage. Based on our result, it could be estimated that the current RIPs from the L1PA2 subfamily are only a proportion of what is in the human genome: among the 1135 L1HS reported as human specific by our result, 448 of which are reported RIPs; while for L1PA2, there are only 24 reported RIPs up to date despite the fact it has more human specific entries (1977 HS-L1PA2 compared to 4699 all L1PA2). Similar conclusions could be made for the SVA subfamily: despite their high retrotransposition activity level suggested by our results (1526 HS-SVA in total), there are only 77 reported RIP cases to date. Being the youngest RE family, which originated ~25My ago [27], SVA is the most active family in the human genome, and therefore we could expect to identify more RIPs from this subfamily. Our result confirms that HERVK subfamily is the most active LTR subfamily in the human lineage. However, the fact that HERVK has a ~3% HS-ratio may suggest that the current 9 entries in dbRIP represent only a proportion of the HERVK insertion polymorphism in the human genome. Also, the ERV1 family is showing appreciable retrotransposition activity: with 344 HS-ERV1 comparing to 218 HS-HERVK. We can thus expect to find more ERV1 RIPs in the human genome.

9.3. The bias of HS-REs for chromosome Y

As shown in section 8.7.3, the HS-RE density, as well as the HS-RE and gene density ratio on chromosome Y is significantly higher than all other chromosomes. Also, as previously discussed in section 8.7.5, the HS-LTR on chromosome Y has contributed to over 60% of its total size contribution. Therefore, it is clear that HS-REs are most successful on chromosome Y. Three of the four RE families examined (Alu, L1 and LTR) have their HS-REs showing strong relative bias towards chromosome Y.

As previously discussed (section 8.7.3), we suspected such strong bias towards chromosome Y could have contributed by the poor quality of out-group chromosome Y. However, our results suggested that the gap regions on the chimpanzee chromosome Y didn't cause any false positives. The absence of other out-group chromosome Y sequence data could also contribute to false positive Type II HS-REs. There could be some REs that are shared by humans and other primates but not chimpanzee. However, chimpanzees are the closet related primate to humans. In theory, other primates should be less related to human and therefore contain less shared REs with humans on their chromosome Y. Based on these, we would consider that there is a chromosome Y bias in HS-REs. However future studies are needed to further validate such bias with the availability of chromosome Y sequence data from other primates.

Several groups have focused on examining the distribution pattern of REs between the sex chromosomes and autosomes. Based on their findings, the young members in the Alu family (mainly AluYs) have two times higher density on chromosome Y than on autosomes [120]. Recent studies suggest that the young L1 subfamily (primate-specific

L1Ps) have a significantly higher density on the sex chromosomes when compared to autosomes [121, 122]. These finding are supported by our results that HS-Alu and HS-L1, which are mainly young members in the family, are significantly biased to chromosome Y: the HS-Alu density on chromosome Y is almost six times higher than the density of autosomes; the HS-L1has an eight times higher density on chromosome Y. In addition, our results showed for the first time that human specific LTRs are also significantly biased to chromosome Y, with a ~45 times higher density on chromosome Y than autosomes. How did this bias happen during evolutionary history? There are no currently known mechanisms contribute to this chromosome Y bias. However, several hypotheses have been proposed.

According to the male germline hypothesis, the densities of REs on each chromosome types should positively correlate with the amount of time spent in the male germline. Assuming cell division is mutagenic and the number of germ cell divisions is larger in males than in females (as is likely in most mammals), male is expected to be main source of mutations. The autosomes have a ½ chance carried by the male while chromosome Y is always carried by the male. Chromosome X, however, has a 1/3 chance carried by the male. In theory, as per generation, the maximum mutation frequency between autosomes and sex chromosomes would be Y:A:X = 6:3:2[123]. This hypothesis is supported by the observed Y>A>X density of young Alus[120].

Another hypothesis is that the RE density on chromosome Y is facilitated by the low recombination-driven deletion in chromosome Y. As chromosome Y does not recombine with chromosome X outside of the pseudo-autosomal regions, it would have higher RE density than the autosomes. However, this hypothesis is not supported by our results that

the HS-RE density and shared RE density are higher on the pseudo-autosomal regions than the male specific regions on chromosome Y (Data not shown).

Natural selection might also be contributing to the chromosome Y bias: chromosome Y has a lower frequency of DNA recombination due to lack of homologous regions [122]. DNA recombination-mediated deletion can serve as a mechanism for removing RE insertions in a way biased for those with negative functional impact via selection. Because chromosome Y has much fewer genes than any other chromosomes, it is facing a lower level of selection pressure and lower rate of insertion removal.

However, none of these mentioned hypotheses could explain why this bias is observed for HS-RE but not for shared REs, nor can one explain why HS-SVA is not showing such chromosome Y bias. Although genetic decay may be the principal dynamic in the evolution of chromosome Y, recent studies suggested that remodeling and regeneration have dominated chimpanzee and human male specific chromosome Y evolution during the past 6 million years[124]. This could potentially explain the success of HS-REs on chromosome Y.

Assuming this bias is valid, how would it affect chromosome Y? Recently, there has been a heated debate among the science community about whether chromosome Y is disappearing or not [125]. Our result could serve as evidence that “chromosome Y has not disappeared yet” [125] and HS-REs may have contributed to the observed fast evolving pattern on chromosome Y after the human and chimpanzee divergence [124]. However, the potential functional impact about the preferential insertion of HS-RE on chromosome Y is still unknown.

10. Summary and future perspectives

Transposable elements play important roles in genome evolution and function in most higher organisms, including humans. Although several groups have looked into the genetic diversity contributed by retrotransposons between human and other primates, they have focused on either a specific family of retrotransposons or a specific primate, leaving the total number of human specific transposable elements largely undetermined. In this study, we took advantage of the availability of the human genome sequences and a few non-human primate genome sequences to study the impact of retrotransposons on the human genome from several aspects.

By computationally comparing the human genome to 4 primate genomes, we identified a total of 18,860 HS-REs, among which are 11,664 Alus, 4,887 L1s, 1,526 SVAs and 783 LTRs (222 full length entries), representing the largest and most comprehensive list of HS-REs generated to date. Together, these HS-REs contributed a total of 14.2Mb sequence increase from the inserted REs and Target Site Duplications (TSDs), 71.6Kb increase from transductions, and 268.2 Kb sequence deletion of from insertion-mediated deletion, leading to a net increase of ~14 Mb sequences to the human genome. Furthermore, we observed for the first time that Y chromosome might be a hot target for new retrotransposon insertions in general and particularly for LTRs. Many of these HS-REs insert into gene regions, including exon, promoter, and intron regions, with high potential for direct impact on gene regulation, splicing and protein coding. All these

data suggest that retrotransposon elements have played a significant role in the evolution of *Homo sapiens*.

Many directions of future study may be explored to further characterize the specific impact of transposable elements on human genome evolution and function. First, with the availability of personal genome data from the 1000 genome project and many large personal genome projects alike, it is possible to focus on identifying population specific REs and their contribution to the phenotypic differences observable among different population. Second, the functional impact of all human specific transposable elements on the associated genes can be examined experimentally, perhaps by using a combination of animal model, in vitro study, and disease association study, and by giving those in located in the genic region higher priority. Last, but not least, similar study can be applied to each of the other primate species to provide an assessment of transposable elements' functional impact on evolution of the species and and the unique biology.

11. References

1. MCCLINTOCK B: **Controlling elements and the gene.** Cold Spring Harb Symp Quant Biol 1956, **21**:197-216.

2. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K: **Initial sequencing and analysis of the human genome.** Nature 2001, **409**(6822):860.

3. Pace II JK, Feschotte C: **The evolutionary history of human DNA transposons: Evidence for intense activity in the primate lineage.** Genome Res 2007, **17**(4):4-4.

4. Herron PR: **Mobile DNA II.** Heredity 2004, **92**(5):476-476.

5. Tassabehji M, Strachan T, Anderson M, Campbell RD, Collier S, Lako M: **Identification of a novel family of human endogenous retroviruses and characterization of one family member, HERV-K(C4), located in the complement C4 gene cluster.** Nucleic Acids Res 1994, **22**(24):5211-5217.

6. Belshaw R, Dawson ALA, Woolven-Allen J, Redding J, Burt A, Tristem M: **Genomewide Screening Reveals High Levels of Insertional Polymorphism in the Human Endogenous Retrovirus Family HERV-K(HML2): Implications for Present-Day Activity.** J Virol 2005, **79**(19):12507-12514.

7. Kazazian HH,Jr, Wong C, Youssoufian H, Scott AF, Phillips DG, Antonarakis SE: **Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man.** Nature 1988, **332**(6160):164-166.

8. Deininger PL, Batzer MA: **Alu repeats and human disease.** Mol Genet Metab 1999, **67**(3):183-193.

9. Callinan PA, Batzer MA: **Retrotransposable elements and human disease.** Genome Dyn 2006, **1**:104-115.

10. Belancio VP, Hedges DJ, Deininger P: **Mammalian non-LTR retrotransposons: for better or worse, in sickness and in health.** Genome Res 2008, **18**(3):343-358.

11. Cordaux R, Batzer MA: **The impact of retrotransposons on human genome evolution.** Nat Rev Genet 2009, **10**(10):691-703.

12. Babushok DV, Kazazian HH,Jr: **Progress in understanding the biology of the human mutagen LINE-1.** Hum Mutat 2007, **28**(6):527-539.

13. Skowronski J, Fanning TG, Singer MF: **Unit-length line-1 transcripts in human teratocarcinoma cells.** Mol Cell Biol 1988, **8**(4):1385-1397.

14. Feng Q, Moran JV, Kazazian HH,Jr, Boeke JD: **Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition.** Cell 1996, **87**(5):905-916.

15. Jurka J: **Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons.** Proc Natl Acad Sci U S A 1997, **94**(5):1872-1877.

16. Mathias SL, Scott AF, Kazazian HH,Jr, Boeke JD, Gabriel A: **Reverse transcriptase encoded by a human transposable element.** Science 1991, **254**(5039):1808-1810.

17. Szak ST, Pickeral OK, Makalowski W, Boguski MS, Landsman D, Boeke JD: **Molecular archeology of L1 insertions in the human genome.** Genome Biol 2002, **3**(10):research0052.

18. Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, Moran JV, Kazazian HH,Jr: **Hot L1s account for the bulk of retrotransposition in the human population.** Proc Natl Acad Sci U S A 2003, **100**(9):5280-5285.

19. Batzer MA, Deininger PL: **Alu repeats and human genomic diversity.** Nat Rev Genet 2002, **3**(5):370-379.

20. Weiner AM, Deininger PL, Efstratiadis A: **Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information.** Annu Rev Biochem 1986, **55**:631-661.

21. Roy-Engel AM, Batzer MA, Deininger PL: **Evolution of Human Retrosequences: Alu.** In *eLS.* Edited by Anonymous John Wiley & Sons, Ltd; 2001:.

22. Kriegs JO, Churakov G, Jurka J, Brosius J, Schmitz J: **Evolutionary history of 7SL RNA-derived SINEs in Supraprimates.** Trends Genet 2007, **23**(4):158-161.

23. Shaikh TH, Roy AM, Kim J, Batzer MA, Deininger PL: **cDNAs derived from primary and small cytoplasmic Alu (scAlu) transcripts.** J Mol Biol 1997, **271**(2):222-234.

24. Comeaux MS, Roy-Engel AM, Hedges DJ, Deininger PL: **Diverse cis factors controlling Alu retrotransposition: what causes Alu elements to die?** Genome Res 2009, **19**(4):545-555.

25. Dewannieux M, Esnault C, Heidmann T: **LINE-mediated retrotransposition of marked Alu sequences.** Nat Genet 2003, **35**(1):41-48.

26. Wang H, Xing J, Grover D, Hedges DJ, Han K, Walker JA, Batzer MA: **SVA elements: a hominid-specific retroposon family.** J Mol Biol 2005, **354**(4):994-1007.

27. Shen L, Wu LC, Sanlioglu S, Chen R, Mendoza AR, Dangel AW, Carroll MC, Zipf WB, Yu CY: **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.

28. Ostertag EM, Goodier JL, Zhang Y, Kazazian HH,Jr: **SVA elements are nonautonomous retrotransposons that cause disease in humans.** Am J Hum Genet 2003, **73**(6):1444-1451.

29. Gifford R, Tristem M: **The evolution, distribution and diversity of endogenous retroviruses.** Virus Genes 2003, **26**(3):291-315.

30. Moyes D, Griffiths DJ, Venables PJ: **Insertional polymorphisms: a new lease of life for endogenous retroviruses in human disease.** Trends Genet 2007, **23**(7):326-333.

31. Turner G, Barbulescu M, Su M, Jensen-Seaman MI, Kidd KK, Lenz J: **Insertional polymorphisms of full-length endogenous retroviruses in humans.** Curr Biol 2001, **11**(19):1531-1535.

32. Xing J, Zhang Y, Han K, Salem AH, Sen SK, Huff CD, Zhou Q, Kirkness EF, Levy S, Batzer MA, Jorde LB: **Mobile elements create structural variation: analysis of a complete human genome.** Genome Res 2009, **19**(9):1516-1526.

33. Kazazian HH,Jr: **An estimated frequency of endogenous insertional mutations in humans.** Nat Genet 1999, **22**(2):130.

34. Muotri AR, Chu VT, Marchetto MC, Deng W, Moran JV, Gage FH: **Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition.** Nature 2005, **435**(7044):903-910.

35. Cordaux R, Hedges DJ, Herke SW, Batzer MA: **Estimating the retrotransposition rate of human Alu elements.** Gene 2006, **373**:134-137.

36. Kano H, Godoy I, Courtney C, Vetter MR, Gerton GL, Ostertag EM, Kazazian HH,Jr: **L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism.** Genes Dev 2009, **23**(11):1303-1312.

37. Chen JM, Stenson PD, Cooper DN, Ferec C: **A systematic analysis of LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease.** Hum Genet 2005, **117**(5):411-427.

38. Gasior SL, Wakeman TP, Xu B, Deininger PL: **The human LINE-1 retrotransposon creates DNA double-strand breaks.** J Mol Biol 2006, **357**(5):1383-1393.

39. Morrish TA, Gilbert N, Myers JS, Vincent BJ, Stamato TD, Taccioli GE, Batzer MA, Moran JV: **DNA repair mediated by endonuclease-independent LINE-1 retrotransposition.** Nat Genet 2002, **31**(2):159-165.

40. Morrish TA, Garcia-Perez JL, Stamato TD, Taccioli GE, Sekiguchi J, Moran JV: **Endonuclease-independent LINE-1 retrotransposition at mammalian telomeres.** Nature 2007, **446**(7132):208-212.

41. Gladyshev EA, Arkhipova IR: **Telomere-associated endonuclease-deficient Penelope-like retroelements in diverse eukaryotes.** Proc Natl Acad Sci U S A 2007, **104**(22):9352-9357.

42. Sen SK, Huang CT, Han K, Batzer MA: **Endonuclease-independent insertion provides an alternative pathway for L1 retrotransposition in the human genome.** Nucleic Acids Res 2007, **35**(11):3741-3751.

43. Srikanta D, Sen SK, Huang CT, Conlin EM, Rhodes RM, Batzer MA: **An alternative pathway for Alu retrotransposition suggests a role in DNA double-strand break repair.** Genomics 2009, **93**(3):205-212.

44. Arcot SS, Wang Z, Weber JL, Deininger PL, Batzer MA: **Alu repeats: a source for the genesis of primate microsatellites.** Genomics 1995, **29**(1):136-144.

45. Jurka J, Pethiyagoda C: **Simple repetitive DNA sequences from primates: compilation and analysis.** J Mol Evol 1995, **40**(2):120-126.

46. Kass DH, Batzer MA, Deininger PL: **Gene conversion as a secondary mechanism of short interspersed element (SINE) evolution.** Mol Cell Biol 1995, **15**(1):19-25.

47. Roy AM, Carroll ML, Nguyen SV, Salem AH, Oldridge M, Wilkie AO, Batzer MA, Deininger PL: **Potential gene conversion and source genes for recently integrated Alu elements.** Genome Res 2000, **10**(10):1485-1495.

48. Callinan PA, Wang J, Herke SW, Garber RK, Liang P, Batzer MA: **Alu retrotransposition-mediated deletion.** J Mol Biol 2005, **348**(4):791-800.

49. Han K, Sen SK, Wang J, Callinan PA, Lee J, Cordaux R, Liang P, Batzer MA: **Genomic rearrangements by LINE-1 insertion-mediated deletion in the human and chimpanzee lineages.** Nucleic Acids Res 2005, **33**(13):4040-4052.

50. Hancks DC, Ewing AD, Chen JE, Tokunaga K, Kazazian HH,Jr: **Exon-trapping mediated by the human retrotransposon SVA.** Genome Res 2009, **19**(11):1983-1991.

51. Damert A, Raiz J, Horn AV, Lower J, Wang H, Xing J, Batzer MA, Lower R, Schumann GG: **5'-Transducing SVA retrotransposon groups spread efficiently throughout the human genome.** Genome Res 2009, **19**(11):1992-2008.

52. Grindley ND: **IS1 insertion generates duplication of a nine base pair sequence at its target site.** Cell 1978, **13**(3):419-426.

53. Allet B: **Mu insertion duplicates a 5 base pair sequence at the host inserted site.** Cell 1979, **16**(1):123-129.

54. Holmes SE, Dombroski BA, Krebs CM, Boehm CD, Kazazian HH,Jr: **A new retrotransposable human L1 element from the LRE2 locus on chromosome 1q produces a chimaeric insertion.** Nat Genet 1994, **7**(2):143-148.

55. Moran JV, DeBerardinis RJ, Kazazian HH,Jr: **Exon shuffling by L1 retrotransposition.** Science 1999, **283**(5407):1530-1534.

56. Chen J, Rattner A, Nathans J: **Effects of L1 retrotransposon insertion on transcript processing, localization and accumulation: lessons from the retinal degeneration 7 mouse and implications for the genomic ecology of L1 elements.** Hum Mol Genet 2006, **15**(13):2146-2156.

57. Gilbert N, Lutz-Prigge S, Moran JV: **Genomic deletions created upon LINE-1 retrotransposition.** Cell 2002, **110**(3):315-325.

58. Symer DE, Connelly C, Szak ST, Caputo EM, Cost GJ, Parmigiani G, Boeke JD: **Human l1 retrotransposition is associated with genetic instability in vivo.** Cell 2002, **110**(3):327-338.

59. Kazazian HH,Jr, Goodier JL: **LINE drive. retrotransposition and genome instability.** Cell 2002, **110**(3):277-280.

60. Kondo-Iida E, Kobayashi K, Watanabe M, Sasaki J, Kumagai T, Koide H, Saito K, Osawa M, Nakamura Y, Toda T: **Novel mutations and genotype-phenotype relationships in 107 families with Fukuyama-type congenital muscular dystrophy (FCMD).** Hum Mol Genet 1999, **8**(12):2303-2309.

61. Narita N, Nishio H, Kitoh Y, Ishikawa Y, Ishikawa Y, Minami R, Nakamura H, Matsuo M: **Insertion of a 5' truncated L1 element into the 3' end of exon 44 of the dystrophin gene resulted in skipping of the exon during splicing in a case of Duchenne muscular dystrophy.** J Clin Invest 1993, **91**(5):1862-1867.

62. Belancio VP, Hedges DJ, Deininger P: **LINE-1 RNA splicing and influences on mammalian gene expression.** Nucleic Acids Res 2006, **34**(5):1512-1521.

63. Belancio VP, Roy-Engel AM, Deininger P: **The impact of multiple splice sites in human L1 elements.** Gene 2008, **411**(1-2):38-45.

64. Han JS, Szak ST, Boeke JD: **Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes.** Nature 2004, **429**(6989):268-274.

65. Perepelitsa-Belancio V, Deininger P: **RNA truncation by premature polyadenylation attenuates human mobile element activity.** Nat Genet 2003, **35**(4):363-366.

66. Lee JY, Ji Z, Tian B: **Phylogenetic analysis of mRNA polyadenylation sites reveals a role of transposable elements in evolution of the 3'-end of genes.** Nucleic Acids Res 2008, **36**(17):5581-5590.

67. Chen C, Ara T, Gautheret D: **Using Alu elements as polyadenylation sites: A case of retroposon exaptation.** Mol Biol Evol 2009, **26**(2):327-334.

68. Shankar R, Grover D, Brahmachari SK, Mukerji M: **Evolution and distribution of RNA polymerase II regulatory sites from RNA polymerase III dependant mobile Alu elements.** BMC Evol Biol 2004, **4**:37.

69. Polak P, Domany E: **Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes.** BMC Genomics 2006, **7**:133.

70. Speek M: **Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes.** Mol Cell Biol 2001, **21**(6):1973-1985.

71. Borchert GM, Lanier W, Davidson BL: **RNA polymerase III transcribes human microRNAs.** Nat Struct Mol Biol 2006, **13**(12):1097-1101.

72. Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, Schroder K, Cloonan N, Steptoe AL, Lassmann T, Waki K, Hornig N, Arakawa T, Takahashi H, Kawai J, Forrest AR, Suzuki H, Hayashizaki Y, Hume DA, Orlando V, Grimmond SM, Carninci P: **The regulated retrotransposon transcriptome of mammalian cells.** Nat Genet 2009, **41**(5):563-571.

73. Kim DD, Kim TT, Walsh T, Kobayashi Y, Matise TC, Buyske S, Gabriel A: **Widespread RNA editing of embedded alu elements in the human transcriptome.** Genome Res 2004, **14**(9):1719-1725.

74. Athanasiadis A, Rich A, Maas S: **Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome.** PLoS Biol 2004, **2**(12):e391.

75. Levanon EY, Eisenberg E, Yelin R, Nemzer S, Hallegger M, Shemesh R, Fligelman ZY, Shoshan A, Pollock SR, Sztybel D, Olshansky M, Rechavi G, Jantsch MF: **Systematic identification of abundant A-to-I editing sites in the human transcriptome.** Nat Biotechnol 2004, **22**(8):1001-1005.

76. Li JB, Levanon EY, Yoon JK, Aach J, Xie B, Leproust E, Zhang K, Gao Y, Church GM: **Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing.** Science 2009, **324**(5931):1210-1213.

77. Chen LL, DeCerbo JN, Carmichael GG: **Alu element-mediated gene silencing.** EMBO J 2008, **27**(12):1694-1705.

78. Batzer MA, Rubin CM, Hellmann-Blumberg U, Alegria-Hartman M, Leeflang EP, Stern JD, Bazan HA, Shaikh TH, Deininger PL, Schmid CW: **Dispersion and insertion polymorphism in two small subfamilies of recently amplified human Alu repeats.** J Mol Biol 1995, **247**(3):418-427.

79. Batzer MA, Deininger PL: **A human-specific subfamily of Alu sequences.** Genomics 1991, **9**(3):481-487.

80. Roy AM, Carroll ML, Kass DH, Nguyen SV, Salem AH, Batzer MA, Deininger PL: **Recently integrated human Alu repeats: finding needles in the haystack.** Genetica 1999, **107**(1-3):149-161.

81. Carroll ML, Roy-Engel AM, Nguyen SV, Salem AH, Vogel E, Vincent B, Myers J, Ahmad Z, Nguyen L, Sammarco M, Watkins WS, Henke J, Makalowski W, Jorde LB, Deininger PL, Batzer MA: **Large-scale analysis of the Alu Ya5 and Yb8 subfamilies and their contribution to human genomic diversity.** J Mol Biol 2001, **311**(1):17-40.

82. Callinan PA, Hedges DJ, Salem AH, Xing J, Walker JA, Garber RK, Watkins WS, Bamshad MJ, Jorde LB, Batzer MA: **Comprehensive analysis of Alu-associated diversity on the human sex chromosomes.** Gene 2003, **317**(1-2):103-110.

83. Carter AB, Salem AH, Hedges DJ, Keegan CN, Kimball B, Walker JA, Watkins WS, Jorde LB, Batzer MA: **Genome-wide analysis of the human Alu Yb-lineage.** Hum Genomics 2004, **1**(3):167-178.

84. Garber RK, Hedges DJ, Herke SW, Hazard NW, Batzer MA: **The Alu Yc1 subfamily: sorting the wheat from the chaff.** Cytogenet Genome Res 2005, **110**(1-4):537-542.

85. Otieno AC, Carter AB, Hedges DJ, Walker JA, Ray DA, Garber RK, Anders BA, Stoilova N, Laborde ME, Fowlkes JD, Huang CH, Perodeau B, Batzer MA: **Analysis of the human Alu Ya-lineage.** J Mol Biol 2004, **342**(1):109-118.

86. Roy-Engel AM, Carroll ML, Vogel E, Garber RK, Nguyen SV, Salem AH, Batzer MA, Deininger PL: **Alu insertion polymorphisms for the study of human genomic diversity.** Genetics 2001, **159**(1):279-290.

87. Xing J, Salem AH, Hedges DJ, Kilroy GE, Watkins WS, Schienman JE, Stewart CB, Jurka J, Jorde LB, Batzer MA: **Comprehensive analysis of two Alu Yd subfamilies.** J Mol Evol 2003, **57 Suppl 1**:S76-89.

88. Salem AH, Kilroy GE, Watkins WS, Jorde LB, Batzer MA: **Recently integrated Alu elements and human genomic diversity.** Mol Biol Evol 2003, **20**(8):1349-1361.

89. Salem AH, Ray DA, Hedges DJ, Jurka J, Batzer MA: **Analysis of the human Alu Ye lineage.** BMC Evol Biol 2005, **5**(1):18.

90. Wang J, Song L, Gonder MK, Azrak S, Ray DA, Batzer MA, Tishkoff SA, Liang P: **Whole genome computational comparative genomics: A fruitful approach for ascertaining Alu insertion polymorphisms.** Gene 2006, **365**:11-20.

91. Wheelan SJ, Scheifele LZ, Martinez-Murillo F, Irizarry RA, Boeke JD: **Transposon insertion site profiling chip (TIP-chip).** Proc Natl Acad Sci U S A 2006, **103**(47):17632-17637.

92. Beck CR, Collier P, Macfarlane C, Malig M, Kidd JM, Eichler EE, Badge RM, Moran JV: **LINE-1 retrotransposition activity in human genomes.** Cell 2010, **141**(7):1159-1170.

93. Iskow RC, McCabe MT, Mills RE, Torene S, Pittard WS, Neuwald AF, Van Meir EG, Vertino PM, Devine SE: **Natural mutagenesis of human genomes by endogenous retrotransposons.** Cell 2010, **141**(7):1253-1261.

94. Ewing AD, Kazazian HH,Jr: **High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes.** Genome Res 2010, **20**(9):1262-1270.

95. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski

S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X: **The sequence of the human genome.** Science 2001, **291**(5507):1304-1351.

96. Huang CR, Schneider AM, Lu Y, Niranjan T, Shen P, Robinson MA, Steranka JP, Valle D, Civin CI, Wang T, Wheelan SJ, Ji H, Boeke JD, Burns KH: **Mobile interspersed repeats are major structural variants in the human genome.** Cell 2010, **141**(7):1171-1182.

97. Witherspoon DJ, Xing J, Zhang Y, Watkins WS, Batzer MA, Jorde LB: **Mobile element scanning (ME-Scan) by targeted high-throughput sequencing.** BMC Genomics 2010, **11**(1):410.

98. Stewart C, Kural D, Stromberg MP, Walker JA, Konkel MK, Stutz AM, Urban AE, Grubert F, Lam HY, Lee WP, Busby M, Indap AR, Garrison E, Huff C, Xing J, Snyder MP, Jorde LB, Batzer MA, Korbel JO, Marth GT, 1000 Genomes Project: **A comprehensive map of mobile element insertion polymorphisms in humans.** PLoS Genet 2011, **7**(8):e1002236.

99. 1000 Genomes Project Consortium, Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA: **A map of human genome variation from population-scale sequencing.** Nature 2010, **467**(7319):1061-1073.

100. Liang P, Tang W: **Database documentation of retrotransposon insertion polymorphisms.** Front Biosci (Elite Ed) 2012, **4**:1542-1555.

101. Wang J, Song L, Grover D, Azrak S, Batzer MA, Liang P: **dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans.** Hum Mutat 2006, **27**(4):323-329.

102. Smit A, Hubley R, Green P: **RepeatMasker Open-3.0.** 2004, :.

103. Kent WJ: **BLAT--the BLAST-like alignment tool.** Genome Res 2002, **12**(4):656-664.

104. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W: **Human-mouse alignments with BLASTZ.** Genome Res 2003, **13**(1):103-107.

105. Barry P: **Programming Perl (Book Review).** Linux Journal 2000, (80):62.

106. Xing J, Wang H, Belancio VP, Cordaux R, Deininger PL, Batzer MA: **Emergence of primate genes by retrotransposon-mediated sequence transduction.** Proc Natl Acad Sci U S A 2006, **103**(47):17608-17613.

107. Kulski JK, Shigenari A, Shiina T, Ota M, Hosomichi K, James I, Inoko H: **Human endogenous retrovirus (HERVK9) structural polymorphism with haplotypic HLA-A allelic associations.** Genetics 2008, **180**(1):445-457.

108. Abrusan G, Krambeck HJ: **The distribution of L1 and Alu retroelements in relation to GC content on human sex chromosomes is consistent with the ectopic recombination model.** J Mol Evol 2006, **63**(4):484-492.

109. Liu GE, Alkan C, Jiang L, Zhao S, Eichler EE: **Comparative analysis of Alu repeats in primate genomes.** Genome Res 2009, **19**(5):876-885.

110. Liu G, NISC Comparative Sequencing Program, Zhao S, Bailey JA, Sahinalp SC, Alkan C, Tuzun E, Green ED, Eichler EE: **Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome.** Genome Res 2003, **13**(3):358-368.

111. Vincent BJ, Myers JS, Ho HJ, Kilroy GE, Walker JA, Watkins WS, Jorde LB, Batzer MA: **Following the LINEs: an analysis of primate genomic variation at human-specific LINE-1 insertion sites.** Mol Biol Evol 2003, **20**(8):1338-1348.

112. Rhesus Macaque Genome Sequencing and Analysis Consortium, Gibbs RA, Rogers J, Katze MG, Bumgarner R, Weinstock GM, Mardis ER, Remington KA, Strausberg RL, Venter JC, Wilson RK, Batzer MA, Bustamante CD, Eichler EE, Hahn MW, Hardison RC, Makova KD, Miller W, Milosavljevic A, Palermo RE, Siepel A, Sikela JM, Attaway T, Bell S, Bernard KE, Buhay CJ, Chandrabose MN, Dao M, Davis C, Delehaunty KD, Ding Y, Dinh HH, Dugan-Rocha S, Fulton LA, Gabisi RA, Garner TT, Godfrey J, Hawes AC, Hernandez J, Hines S, Holder M, Hume J, Jhangiani SN, Joshi V, Khan ZM, Kirkness

EF, Cree A, Fowler RG, Lee S, Lewis LR, Li Z, Liu YS, Moore SM, Muzny D, Nazareth LV, Ngo DN, Okwuonu GO, Pai G, Parker D, Paul HA, Pfannkoch C, Pohl CS, Rogers YH, Ruiz SJ, Sabo A, Santibanez J, Schneider BW, Smith SM, Sodergren E, Svatek AF, Utterback TR, Vattathil S, Warren W, White CS, Chinwalla AT, Feng Y, Halpern AL, Hillier LW, Huang X, Minx P, Nelson JO, Pepin KH, Qin X, Sutton GG, Venter E, Walenz BP, Wallis JW, Worley KC, Yang SP, Jones SM, Marra MA, Rocchi M, Schein JE, Baertsch R, Clarke L, Csuros M, Glasscock J, Harris RA, Havlak P, Jackson AR, Jiang H, Liu Y, Messina DN, Shen Y, Song HX, Wylie T, Zhang L, Birney E, Han K, Konkel MK, Lee J, Smit AF, Ullmer B, Wang H, Xing J, Burhans R, Cheng Z, Karro JE, Ma J, Raney B, She X, Cox MJ, Demuth JP, Dumas LJ, Han SG, Hopkins J, Karimpour-Fard A, Kim YH, Pollack JR, Vinar T, Addo-Quaye C, Degenhardt J, Denby A, Hubisz MJ, Indap A, Kosiol C, Lahn BT, Lawson HA, Marklein A, Nielsen R, Vallender EJ, Clark AG, Ferguson B, Hernandez RD, Hirani K, Kehrer-Sawatzki H, Kolb J, Patil S, Pu LL, Ren Y, Smith DG, Wheeler DA, Schenck I, Ball EV, Chen R, Cooper DN, Giardine B, Hsu F, Kent WJ, Lesk A, Nelson DL, O'brien WE, Prufer K, Stenson PD, Wallace JC, Ke H, Liu XM, Wang P, Xiang AP, Yang F, Barber GP, Haussler D, Karolchik D, Kern AD, Kuhn RM, Smith KE, Zwieg AS: **Evolutionary and biomedical insights from the rhesus macaque genome.** Science 2007, **316**(5822):222-234.

113. Chimpanzee Sequencing and Analysis Consortium: **Initial sequence of the chimpanzee genome and comparison with the human genome.** Nature 2005, **437**(7055):69-87.

114. Locke DP, Hillier LW, Warren WC, Worley KC, Nazareth LV, Muzny DM, Yang SP, Wang Z, Chinwalla AT, Minx P, Mitreva M, Cook L, Delehaunty KD, Fronick C, Schmidt H, Fulton LA, Fulton RS, Nelson JO, Magrini V, Pohl C, Graves TA, Markovic C, Cree A, Dinh HH, Hume J, Kovar CL, Fowler GR, Lunter G, Meader S, Heger A, Ponting CP, Marques-Bonet T, Alkan C, Chen L, Cheng Z, Kidd JM, Eichler EE, White S, Searle S, Vilella AJ, Chen Y, Flicek P, Ma J, Raney B, Suh B, Burhans R, Herrero J, Haussler D, Faria R, Fernando O, Darre F, Farre D, Gazave E, Oliva M, Navarro A, Roberto R, Capozzi O, Archidiacono N, Della Valle G, Purgato S, Rocchi M, Konkel MK, Walker JA, Ullmer B, Batzer MA, Smit AF, Hubley R, Casola C, Schrider DR, Hahn MW, Quesada V, Puente XS, Ordonez GR, Lopez-Otin C, Vinar T, Brejova B, Ratan A, Harris RS, Miller W, Kosiol C, Lawson HA, Taliwal V, Martins AL, Siepel A, Roychoudhury A, Ma X, Degenhardt J, Bustamante CD, Gutenkunst RN, Mailund T, Dutheil JY, Hobolth A, Schierup MH, Ryder OA, Yoshinaga Y, de Jong PJ, Weinstock GM, Rogers J, Mardis ER, Gibbs RA,

Wilson RK: **Comparative and demographic analysis of orang-utan genomes.** Nature 2011, **469**(7331):529-533.

115. Scally A, Dutheil J, Hillier L, Jordan G, Goodhead I, Herrero J, Hobolth A, Lappalainen T, Mailund T, Marques-Bonet T, McCarthy S, Montgomery S, Schwalie P, Tang Y, Ward M, Xue Y, Yngvadottir B, Alkan C, Andersen L, Ayub Q, Ball E, Beal K, Bradley B, Chen Y, Clee C, Fitzgerald S, Graves T, Gu Y, Heath P, Heger A, Karakoc E, Kolb-Kokocinski A, Laird G, Lunter G, Meader S, Mort M, Mullikin J, Munch K, O'Connor T, Phillips A, Prado-Martinez J, Rogers A, Sajjadian S, Schmidt D, Shaw K, Simpson J, Stenson P, Turner D, Vigilant L, Vilella A, Whitener W, Zhu B, Cooper D, de Jong P, Dermitzakis E, Eichler E, Flicek P, Goldman N, Mundy N, Ning Z, Odom D, Ponting C, Quail M, Ryder O, Searle S, Warren W, Wilson R, Schierup M, Rogers J, Tyler-Smith C, Durbin R: **Insights into hominid evolution from the gorilla genome sequence.** Nature 2012, **483**(7388):169-175.

116. Bennett EA, Keller H, Mills RE, Schmidt S, Moran JV, Weichenrieder O, Devine SE: **Active Alu retrotransposons in the human genome.** Genome Res 2008, **18**(12):1875-1883.

117. Battilana J, Fagundes NJ, Heller AH, Goldani A, Freitas LB, Tarazona-Santos E, Munkhbat B, Munkhtuvshin N, Krylov M, Benevolenskaia L, Arnett FC, Batzer MA, Deininger PL, Salzano FM, Bonatto SL: **Alu insertion polymorphisms in Native Americans and related Asian populations.** Ann Hum Biol 2006, **33**(2):142-160.

118. Gibbons R, Dugaiczyk LJ, Girke T, Duistermars B, Zielinski R, Dugaiczyk A: **Distinguishing humans from great apes with AluYb8 repeats.** J Mol Biol 2004, **339**(4):721-729.

119. Noonan JP: **Neanderthal genomics and the evolution of modern humans.** Genome Res 2010, **20**(5):547-553.

120. Jurka J, Krnjajic M, Kapitonov VV, Stenger JE, Kokhanyy O: **Active Alu elements are passed primarily through paternal germlines.** Theor Popul Biol 2002, **61**(4):519-530.

121. Abrusan G, Giordano J, Warburton PE: **Analysis of transposon interruptions suggests selection for L1 elements on the X chromosome.** PLoS Genet 2008, **4**(8):e1000172.

122. Boissinot S, Entezam A, Furano AV: **Selection against deleterious LINE-1-containing loci in the human lineage.** Mol Biol Evol 2001, **18**(6):926-935.

123. Miyata T, Hayashida H, Kuma K, Mitsuyasu K, Yasunaga T: **Male-driven molecular evolution: a model and nucleotide sequence analysis.** Cold Spring Harb Symp Quant Biol 1987, **52**:863-867.

124. Hughes JF, Skaletsky H, Pyntikova T, Graves TA, van Daalen, Saskia K. M., Minx PJ, Fulton RS, McGrath SD, Locke DP, Friedman C, Trask BJ, Mardis ER, Warren WC, Repping S, Rozen S, Wilson RK, Page DC: **Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content.** Nature 2010, **463**(7280):536-539.

125. Griffin DK: **Is the Y chromosome disappearing?--both sides of the argument.** Chromosome Res 2012, **20**(1):35-45.