Identification and Annotation of Recombinant Repeats In Mammals Indicates They Are Experimental Products For Creating Novel Transposable Element Families

Sim Lin Lim

A thesis submitted for the degree of Doctor of Philosophy

Discipline of Genetics

School of Molecular and Biomedical Science

The University of Adelaide

October 2013

Table of Contents

Abstract

About 40-50% of mammalian genomes are made up of repetitive elements, primarily transposable elements. Transposable elements' activities not only drive genome evolution, they contribute to the creation of novel recombinant repeats. Recombinant repeats have largely remained uncharacterized due to their complexity. Initially, I developed a pipeline for the genome wide identification of recombinant repeats in four different mammals: human, mouse, cow and horse. The pipeline identified 1,336,824 copies, but only 37,830 sequences were able to be clustered into 6,116 families. The majority of the recombinant repeats were simple recombinant repeat families and only a small proportion were complex recombinant repeat families. My analysis showed that recombinant repeat families only covered a small fraction of the genomes examined (0.30% in human, 0.13% in mouse, 0.217% in horse and 0.464% in cow), indicating most of the recombinant repeats were singletons. Further analysis has shown that both classes of RR were created via transposon-into-transposon events, indicating that novel transposable elements are likely to be created via this mechanism. I found that simple recombinant repeats were probably retrotranspositionally active because they contained polyA tails and target site duplications, showing that they integrated into the genome via retrotransposition events. However, complex recombinant repeat families were only replicated via segmental duplications. My analysis showed that complex recombinant repeat families are excellent candidates for the identification of genome segmental duplication regions that cannot be found through standard methods. In addition, I used the RR identification pipeline to annotate possible RR in pig genome. I discovered a

novel RR family (LTR2i_SS) that contained > 1,000 copies. Repeat annotation showed that it was a chimeric LTR2_SS that contained ~300bp of un-annotated sequence, only found in the pig genome. Further investigation revealed that some LTR2i_SS flanked β3 proviruses, but these proviruses were unable to replicate autonomously as they did not encode a functional, complete polyprotein. My phylogenetic tree analysis of the LTR2i_SS and LTR2_SS familis suggested that LTR2i_SS was the ancestral form of LTR2_SS. In conclusion, I was able to identify the recombinant repeat distributions in different mammals and determine their most probable origin as TinT events. I have shown that recombinant repeats could serve as an important model to explain the origin of novel transposable elements in genomes, or could be used as markers to identify structural variations, or segmental duplications in different species. However, my data have also shown that we have to be cautious when annotating novel recombinant repeats in genomes, as they could be the ancestral form of other known transposable elements rather than novel forms generated through TinT.

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In additon, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicatble, any partner institution responsible for the joint-award of this degree.

I give consent to this copy of my thesis, when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

The author acknowledges that copyright of published works contained within this thesis (as listed below) resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library catalogue, and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Signed……..……....Date…….

Acknowledgements

I would like to express my sincere gratitude to the following people:

Professor David Adelson, for supervision and guidance, and for giving me the opportunity and support to do the PhD. I am so lucky to have Dave to be my mentor. The knowledge, lessons and experiences that I learned from you is not only valuable for my PhD, but it is an invaluable experiences for my future career.

Professor Hamish Scott, my co-supervisor, for providing me the sequencing data for analysis.

Dr Dan Kortchak, who guided me in solving computational and bioinformatic questions, read my manuscripts and always provided valuable suggestions; Joy Raison, for helping me to solve statistical questions, and all other members of the Adelson lab, past and present, for making it such a supportive and enjoyable environments to work.

Dong Wang in the Timmis lab, and everyone else in the MLS building that has helped me along my PhD research.

My wife, Sanny and my parents, for always encouraging and supporting me through my PhD, and for their endless love throughout my life. My sister and brother who have always providing me supports and encouragements.

Chapter 1: Transposable Elements And Recombinant Repeats: Characteristics And Impact On Mammalian Genomes

1. Introduction

Transposable elements (TE) are mobile DNA elements that can move from one location to another location in the host genome, or laterally transfer to another unrelated species [1]. The first TE was discovered by B. McClintock where she observed an unusual colour pattern in maize [2]. Her work led her to conclude that the maize colour mutations were caused by a "controlling element" [3]. Her work however, brought skepticism amongst people until the 1960's. Since then, TE have been found in every known eukaryote, bacteria or archea. Ohno described TE as genomic parasites or "junk" DNA [4]. He believed that TE had no role, no function and they are genetic burdens to the host genomes. Early works also suggested that TE shared similar behavior with retroviruses, where they used their full potential to replicate for survival purposes and had detrimental effects on the hosts [5]. Nevertheless, such simplistic views about TE have been recently challenged [6-10]. Numerous genome assembly projects, advanced molecular genetic experiments and next-generation sequencing (NGS) technologies have shown that 40~80% of typical eukaryotic genomes are composed of TE [6,11-15]. Current work has revealed that TE activities can create structural variations (SV), new genes, alternative splicing, and exert regulatory effects on gene expression [5,16-19]. TE have also played an important role in shaping eukaryotic genome structures through evolutionary processes and later leading to speciation events [20-22].

TE can be allocated into different classes based on their transposition intermediates, structures and their homologues to known repetitive elements.

There are two major TE classes: Class 1 retrotransposon and Class 2 DNA transposon. Both TE classes have distinctive structures and transpositional mechanisms. Meanwhile, a novel repetitive element class, recombinant repeats (RR) have been reported in different studies [23-26]. RR are chimeric repeats composed by various repetitive elements and they are capable of retrotransposing in genomes.

We begin this review with an introduction of TE classes, transposition methods, TE impacts on mammalian genomes and SV. We also discuss the potential mechanisms that create novel TE families in mammals. Finally, we give an overview of RR and discuss their biological impacts based on currently available knowledge.

2. Transposable Element Types And Structures

2.1 Class 2 DNA Transposon

2.1.1 DNA Transposons Structure and Transposition Mechanism

DNA transposons are ~5kbp TE that transpose over the genome in the form of DNA intermediates (Figure 1) [27]. DNA transposons are considered the oldest TE to date and are present in all genomes [28]. DNA transposons contain a single open-reading frame (ORF) that encodes a transposase, and they are flanked by terminal inverted repeats (TIR) [29]. During the transposition events, they are excised from their original site as double strand DNA intermediates, and inserted into another genomic region of the host (Figure 2) [28]. This mechanism is commonly referred as “cut and paste” mechanism. However, there are exceptional cases where DNA transposons do not use double-stranded DNA intermediates for transposition. For example, Helitrons mobilize as single-stranded DNA through a rolling-circle like mechanism [30,31].

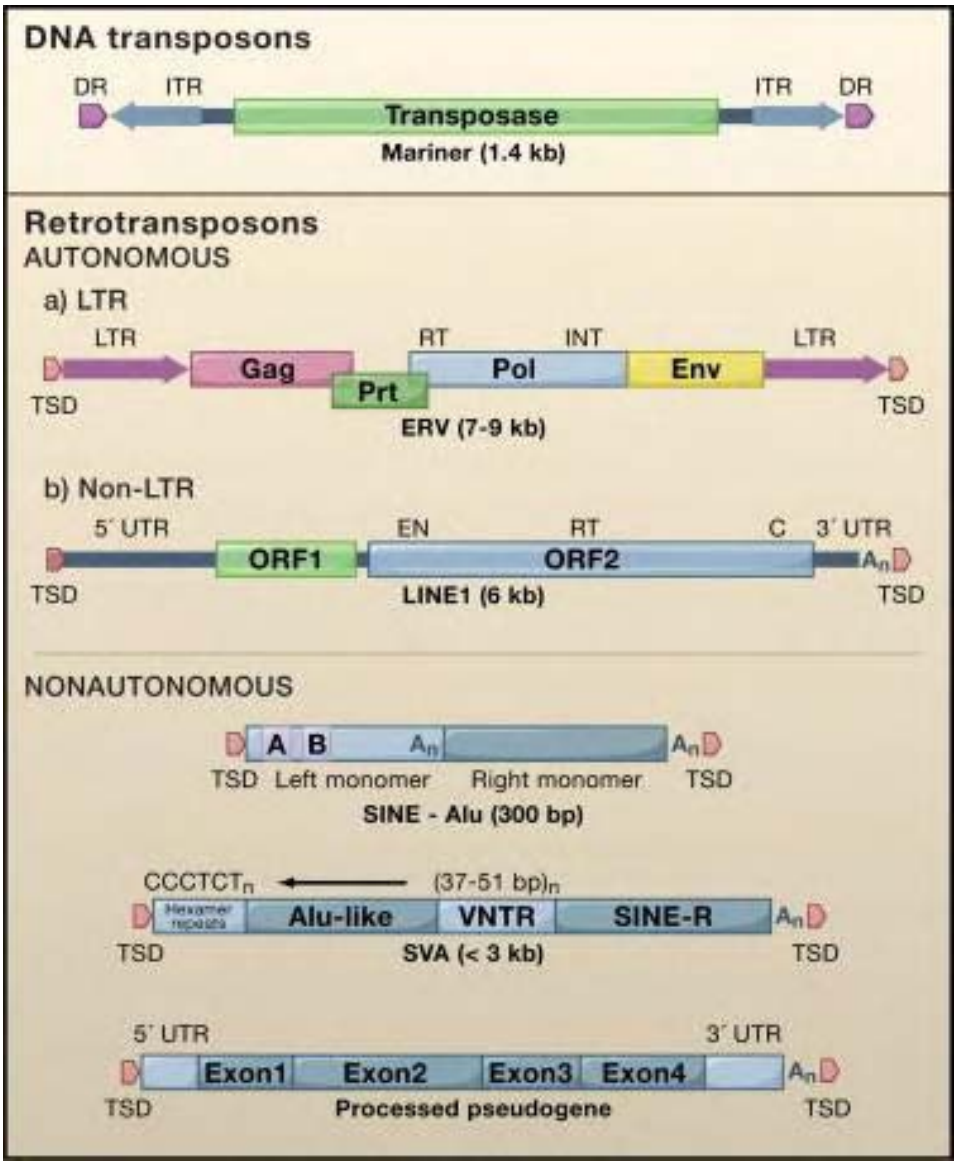


Figure 1. Different Classes and structures of TE in mammalian genomes. They can be divided into 2 types: The class I Retrotransposons and class II DNA transposons (Adapted and modified from [5]).

2.1.2 DNA Transposon Impacts On Genomes And Molecular Research

DNA transposons have limited movements in genomes because they are non-replicable sequences, thus, they are smaller components of mammalian genomes [7-10]. Current genome assemblies have revealed that DNA transposons in mammals are immobile because they have accumulated large numbers of mutations [28]. Although DNA transposon impacts on genome

evolution are unclear, their unique transposition mechanism and specific site insertion preferences have seen applications in biotechnology [32,33]. For example, Sleeping Beauties are modified DNA transposons that are used for gene-rescue and gene knock- out experiments [34-39].

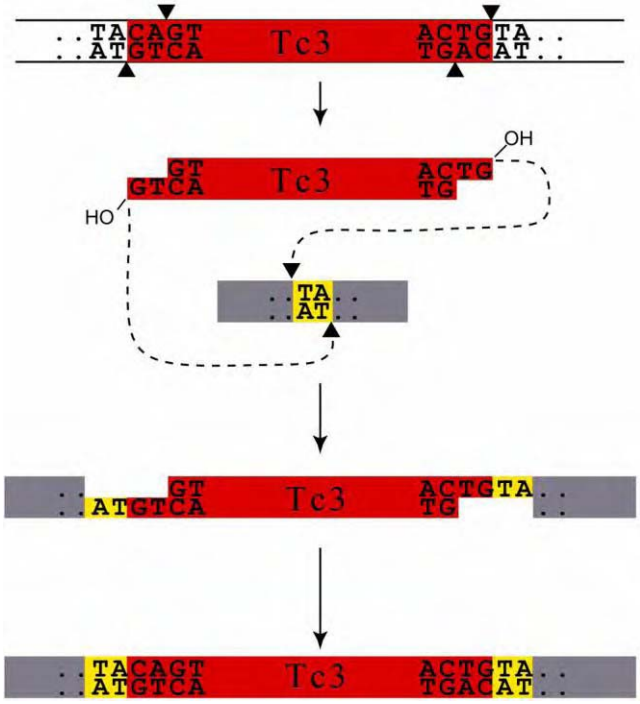


Figure 2: The DNA transposon's cut and paste mechanism (*C. elegans* DNA transposon as model). First, the transposase excises the TC3 DNA transposon, causing a DNA-double strand break in the original location that will later be repaired. The excised DNA transposon will integrate into another genomic location. The fixation of the DNA transposon will result in duplication of the TA nucleotide at each end (Adapted and modified from http://www.wormbook.org/chapters/www_transposons/transposonsfig2.jpg, Date 30/9/2013).

2.2 Class 1 Retrotransposon

Retrotransposons are TE that replicate through RNA copies [20,40], also a process known as "copy and paste" mechanism. This mechanism ensures the

retrotransposon ‘master copy’ remains in the original genomic location, and its new replicates can retrotranspose into other genomic locations. This effective mechanism enables the retrotransposons to account for large fractions of mammalian genomes. Retrotransposons can be further divided into two subclasses: a) Long terminal repeat (LTR) retrotransposons, and b) non-LTR retrotransposons [5,40]. Both subclasses have distinctive structures and mobility, but they use the same mechanism for retrotransposition.

2.2.1 LTR Retrotransposons

LTR retrotransposons are 4~12kbp Endogenous Retrovirus-like (ERV) retroelements flanked by 200 to 600bp LTRs (Figure 1) [41-44]. They contain multiple regulatory elements that act as a single transcriptional unit. LTR retrotransposon regulatory elements share homologues with known endogenous retroviruses [43,44]: a) They have *gag* genes that encode structural proteins associated with nucleic acid binding activities [45], and b) *pol* genes that encode polyproteins with protease, reverse transcriptase, RnaseH and integrase [41,44]. However, they do not have the retroviral *env* genes, hence they are limited to replicate and spread in an intracellular fashion.

The mechanism of LTR retrotransposon retrotransposition has been described using yeast Ty elements [46-49]. First, LTR retrotransposon DNA is transcribed into an RNA intermediate (Figure 3) [50,51]. The RNA intermediate is circularised and retrotransposed back into the genome. However, the detail of LTR reverse transcription processes are unknown as no study has been able to identify how LTR retrotransposon RNA intermediates are associated with protein complexes or reintegrated into the genome [52]. Previous work has revealed that LTR retrotransposon RNA intermediates are able to undergo intra-element

homologous recombination via their LTRs [53]. This results in the excision of their structural genes and leads to singleton LTR being integrated into the genome.

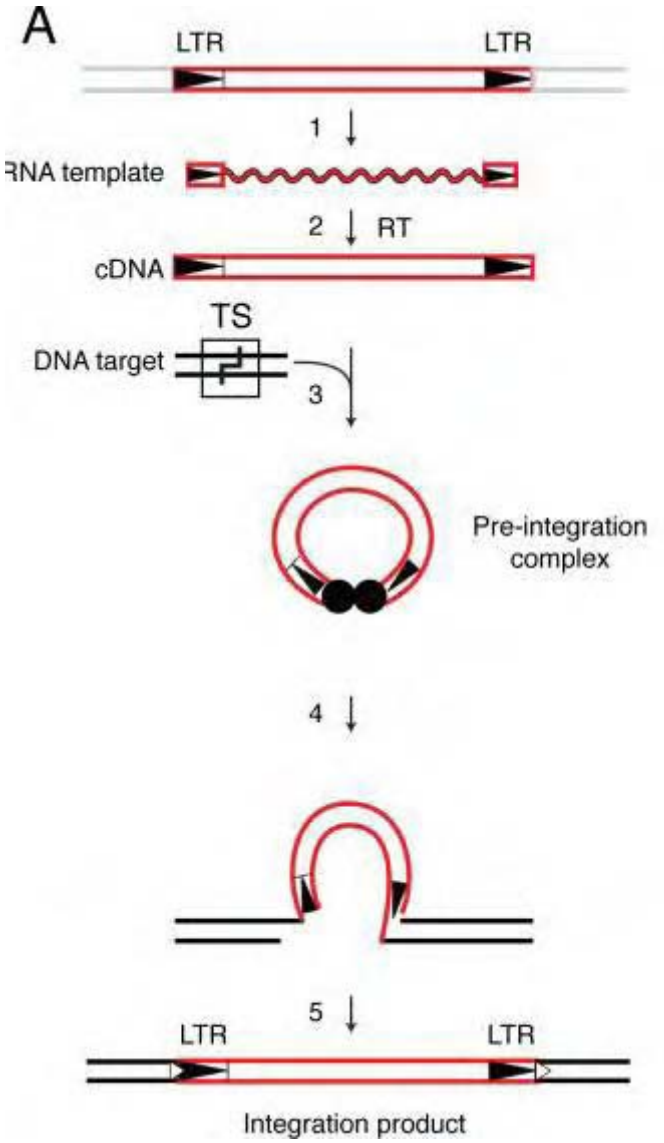


Figure 3. The LTR retrotransposition pathway. The LTR retrotransposon is colored in red. First, it is transcribed into an intact RNA template with LTR (Process 1, then reverse-transcribed into cDNA (Process 2) to form a pre-integration complex (process 3). The preintegration complex interacts with a genomic target site and integrates into the genome (Process 4 and 5) (Adapted and modified from [54]).

2.2.2 Non-LTR Retrotransposons

Non-LTR retrotransposons can be divided into two subclasses: a) Autonomous non-LTR retrotransposons, Long Interspersed Nuclear Elements (LINE) that contain protein coding genes required for retrotransposition, and b) Non-automonous non-LTR retrotransposon, Short Interspersed Nuclear Elements (SINE) that lack protein coding genes, and rely on LINE proteins to assist their retrotransposition [5].

2.2.2a Long Interspersed Elements (LINE)

LINE are 5~6kbp sequences that consist of an internal promoter in the 5' untranslated region (5' UTR), two open reading frames (ORF1 and ORF2), and a short 3' UTR terminating in an AATAA polyadenylation signal or polyA tails (Figure 1) [55-59]. The LINE ORF1 encodes an RNA binding protein that contains a coiled-coiled domain (CC), a non-canonical RNA recognition motif (RRM) domain and a basic C-terminal domain (CTD). The ORF1 RNA binding proteins (ORF1p) are believed to act as chaperones to assist LINE reverse-transcription [60-67], and previous work has revealed that ORF1 deletions or knock-out in active LINE resulted in lower retrotransposition rates, but did not halt retrotransposition [64,65,67]. The ORF2 encodes a protein that has reverse transcriptase and endonuclease activities, a Z domain and a Cys-rich domain [57,60,68-70]. ORF2 proteins (ORF2p) are critical in successful retrotransposition because previous *in vitro* studies showed that ORF2 mutations in active LINE permanently deactivated LINE reverse transcription activity [68-71].

During LINE retrotransposition (Figure 4), the LINE ORF1 and ORF2 translate into ORF1p and ORF2p, that form a ribonucleoprotein particle (RNP) [60,65,66]. This RNP includes the LINE RNA intermediate and forms a complex RNA particle. When it is imported back to the nucleus [72,73], it reverse-

transcribes the attached LINE RNA back into DNA. The DNA integrates back into the genome via a target primed reverse transcription (TPRT) process [57,74-76] [51,69-71]. However, most retrotransposed copies are 5'-truncated sequences that have lost 5' UTR internal promoters, indicating they are no longer capable of retrotransposition [45]. These "dead on arrival" LINE are commonly observed in mammalian genomes, composing ~20% of a typical genome sequence [7-10,77-82].

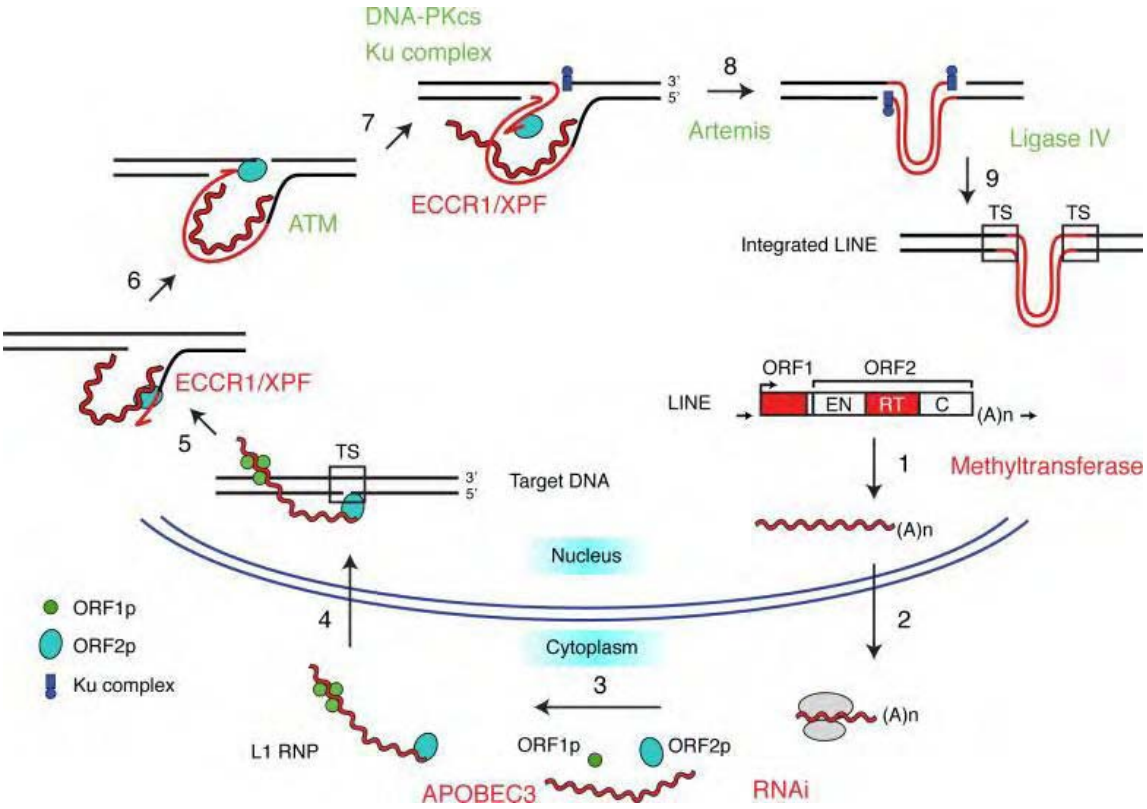


Figure 4. The overall mechanism of the autonomous non-LTR retrotransposon (LINE) mechanism. Initially, a LINE is transcribed into RNA (Process 1). The RNA copy is exported to the cytoplasm (Process 2). The ORF1p and ORF2p are translated from the RNA copy, and form a complex RNP (Process 3). The RNP complex interacts with the LINE RNA copy and is imported back to the nucleus (Process 4). The LINE RNA is reverse-transcribed and integrated into the genome

via a target-primed retrotransposition (TPRT) mechanism (Process 5-8) (Adapted and modified from [54]).

2.2.2 Interspersed Elements (SINE)

Non-autonomous SINE are short sequences (300~ 500bp) without any functional protein coding genes. They contain 5' tRNA or 7SL-like sequences that act as an RNA polymerase III promoter, and a 3' end associated with an oligo (A) tail or A-rich stretch (Figure 1) [20,40,42,83]. SINE retrotransposition relies on LINE retrotransposition via two methods: a) LINE 3' transduction events that can mobilize a SINE at their 3' end during the transcription process [84-86], or b) Hijacking LINE reverse transcriptase for replication purposes (Figure 5) [87]. SINE also integrate into the genome via a TPRT mechanism, but most of them remain as intact sequences. Although SINE make up 10~15% of typical mammalian genomes, their copy numbers are significantly higher than other TE in mammals [7-10].

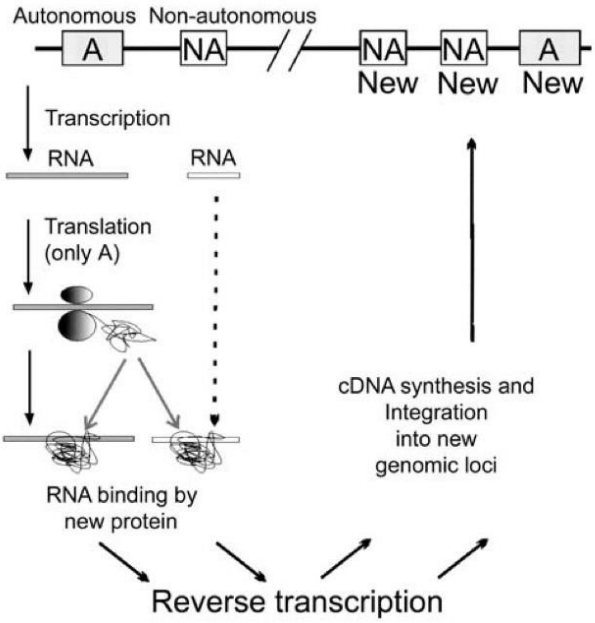


Figure 5. Non-autonomous retrotransposons (SINE) replication mechanism. SINE lack functional protein-coding genes and must hijack the LINE reverse transcriptase (or RNP complex) to initiate reverse transcription and integrate into the genome (Adapted and modified from [16]).

2.2.3 The Biological Impact Of Retrotransposons On Genomes

It is accepted that retrotransposon activity has affected genome structure, both directly and indirectly [5,16,62,88]. However, these impacts are limited to retrotransposition events that occur in germ cells or early embryonic development that can be passed to subsequent generations and can ultimately be fixed in populations [11,89,90]. One major impact is that they contribute to speciation events through genome evolutionary processes. Previous work has shown that over a million copies of Alu (SINE) are exclusive to primate families, and thousands of them presumably differentiate human from chimpanzee [91-96]. Retrotransposons are able to exert regulatory effects if they insert into flanking or genic functional sequences – promoters, enhancers, exons, etc. For example, LTR retrotransposons can act as transcriptional repressors, competing with the gene promoter for binding of transcriptional factors [97-99], or interact with hormone receptors and transcriptional regulators in host [100-111]. Some non-LTR retrotransposon insertions in 3' UTR or intronic regions can affect gene expression or produce alternative transcripts [97,112-114]. Retrotransposons have not only provided abundant raw materials for the formation of new genes, they are able to repair double-stranded DNA breaks by pasting a new TE into a damaged locus of chromosome [115]. In short, retrotransposons are the major players that have shaped eukaryotic genome structures, and create variations in individual genomes [116].

3. TE And Genome Structural Variations

3.1 Structural Variation Definitions

Structural variations (SV) have been a subject of research since the publication of the human genome assembly in 2001 [8]. Structural variation is defined as: the rearrangement or shuffling of DNA sequences in chromosomes that create variations between individuals [117,118]. SV are ubiquitous in different species and it is believed they contribute to genome evolution and speciation events. SV can be divided into 3 types. A) single nucleotide polymorphisms (SNP), where there is a single nucleotide difference (A, T, C or G) in an individual genome. The current human HapMap Project had identified more than 3.1 million SNP in human populations [119-122]. B) Short insertions and deletions (indels) (≤1kbp) that are common in genomes [123-128]. C) Copy number variations (CNV), defined by different copy numbers of identical sequences across individuals genome [129-133]. However, CNV are not as clearly defined compared to indels or SNP. For example, tandem repeats, TE and segmentally duplicated gene families share common features with CNV, but their mobility and characteristics are distinctive.

In the past, SV could not be identified through experimental or bioinformatic methods due to the limited test samples and incomplete reference genomes [134]. At that time, various 'experimental' technology platforms and data processing algorithms were developed to study SV in human, but the results contained high false-positive/false-negative rates and remained controversial [126,134-138]. However, genome sequencing techniques and molecular experimental methods have been improved, and more genome assemblies have been released for further study. Current researchers are not limited to studying SV in a single genome, but are able to compare SV across species and study their impacts on genome evolution [15,116,139-145].

3.2 Retrotransposon Roles In Human SV

The roles of retrotransposons in SV have been a topic of debate. Previous work has proposed that retrotransposons were junk DNA and did not provide beneficial effects to the host [11,72]. However, four independent publications in 2010 have suggested a strong relationship between TE activity and SV in the human genome (Figure 6) [6,12,13,146]. They showed that 'hot' L1 element retrotranspositional rates are much higher than anyone previously expected. Huang et al. identified high copy number new human L1 element insertional polymorphisms by performing transposon insertion profiling microarray (TIP-chip) [13], showing that the rate of novel L1 insertions is twice as high as previously thought. They estimate that there is at least one new L1 insertion in every 108 births in the human population. Iskow et al. combined element/locus junction-specific PCR with next generation 454 DNA sequencing to study L1 activities in 76 individuals [146]. They demonstrated that L1 insertions are active and favour insertion into intronic regions. Beck et. al. used a LINE experimental assay and identified 68 novel 'hot' L1 from 5 individuals from different human populations [6]. Ewing and Kazazian used high-throughput DNA sequencing to investigate human-specific L1 insertion sites in 25 individuals [12], and found that the L1 insertions are structural dimorphisms. They concluded that L1 retrotransposition rates in human are between 1/95 or 1/270 births. These independent studies indicate that retrotransposons are significant contributor to SV in the genome, and they have the ability to alter or shape the genome through evolutionary processes.

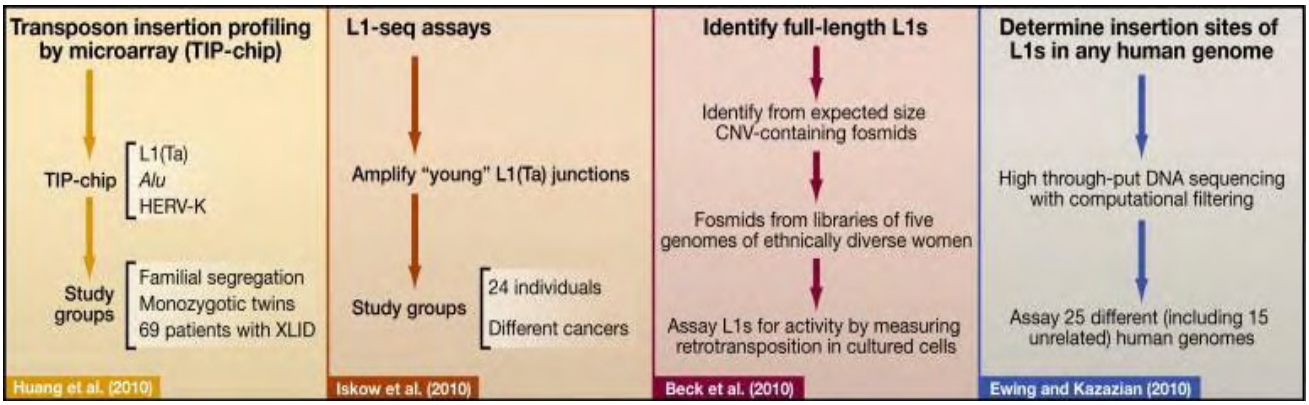


Figure 6. The overall experimental methods to prove the important role of LINE in human SV. The results concluded that the number of human L1 insertional events in human are random and underestimated (Adapted and modified from [116]).

4. Recombinant Repeat Structures And Characteristics

4.1 Current Understanding Of Recombinant Repeats

Recombinant repeats (RR) are sequences composed by combinations of well-characterized repetitive elements or non-repetitive element sequences. They have previously been described as nested repeats or chimeric repeats [23-26,83,147]. RR cannot be classified by normal TE annotation methods because they span multiple TE sequences[16,147,148]. They can be made up by different types of repetitive elements or fragments, making simple structures such as simple recombinant repeat (SRR) composed by 2 (bipartite) (Figure 7) or 3 TE (tripartite), or complex recombinant repeat (CRR) that are products of more than 4 TE [23,25,26,149]. For example, SVA elements are well-known hominid-specific SRR composed of Alu, variable number tandem repeats (VNTR), and SINE. Computational analyses revealed that human RR chimera could be traced back to 47~100 million years ago [149], and are products generated by active transposable elements and found in specific genomes [150,151]. To date, RR have only observed in certain mammals, birds, fungal and plant genomes, but

they have not yet been described in the genomes of invertebrates, amphibia and fish [25,26,147,151-154]. Based on current knowledge, it appears that many RR are lineage specific, for example, F. Sabot and H. Schulman discovered the chimeric LTR retrotransposons in Triticeae, but absent in Arabidopsis [153]. The chimeric structures in RR imply that independent fusion events between different repetitive elements have occurred during evolution. Some studies have shown that RR are created by “TE inserted in to other TE” processes, where a currently active TE can insert into new, older, inactive or fossil TE and create RR [16,23-25,147,148,154,155]. RR can also be used as genetic markers to resolve primate, rodent and rabbit phylogenetic trees [24,152,155], indicating that RR are potential tools for studying evolution of species.

4.2 RR Share Similar Characteristics With Retrotransposed TE

RR share similar characteristics with retrotransposed TE or processed pseudogenes (Figure 7). These RR usually contain: a) 2 target-site duplications (TSDs) or known as “direct repeats” (12~20 base pairs) in their flanking regions, b) poly A tails at the 3' end of RR and c) TTAAAA, a motif observed 5’ of RR’s TSD that is preferentially recognized by L1 nicking endonuclease [148]. These properties indicate that RR mRNA are capable of hijacking L1 RNP to initiate retrotransposition. RR do not contain any protein-coding genes, but they are transcribed and expressed in a tissue-specific manner in *in vitro* experiments [147]. This raises the possibility that RR may have functional roles, perhaps as non-coding RNAs.

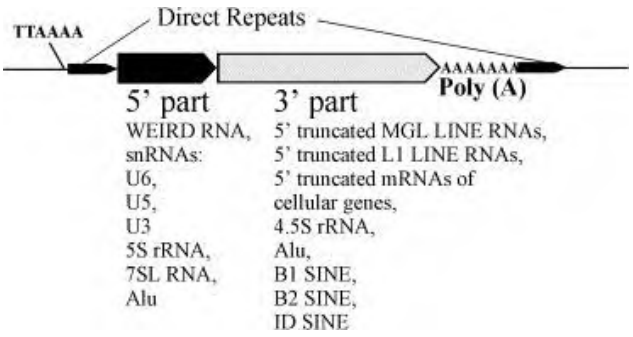


Figure 7. An example of bipartite Recombinant repeat structure. It is a fusion of different classes of repetitive elements, or other non-repetitive elements (pseudogenes) that show retrotransposition characteristics, such as polyA tail and direct repeats/target site duplications (TSD) (Adapted and modified from [148]).

5. RR As Models To Explain The Evolution of Novel TE Families

Although TE transposition methods and structures have been well studied, the way novel TE arise in a genome is still not fully understood. For example, the origin of SINE is still an open question. SINE have a composite or modular structure: a 5' part with similarity to 5' tRNA promoters, a tRNA-unrelated region and a 3' end with shared similarity to the 3' tail of LINEs [83]. It is critical to understand how they have been created as they have had profound impacts on genome expansion and SV. However, we are so far unable to conduct *in vivo* or *in vitro* experiments to test or observe novel TE creation processes.

The discovery of RR with retrotransposon characteristics suggests that they could be novel TE [23-25,147,154]. The unique structure of RR have allowed the research community to use them as a model to study the origin of TE and explain the modular structure of SINE [83]. Five mechanisms have been proposed to explain how novel TE families arise in the genome.

5.1 Transposon-into-Transposon (TinT) Insertions

TinT insertions are common occurrences in genomes, and the insertions can be random or site-specific [6,11-14,156-158]. TinT events are usually the result of newer TE inserting into older TE, and interrupt the old TE, thereby destroying the retrotranspositional activity of the old TE . However, TinT also allows TE to interact and undergo 'dimerization' /'trimerization' process that are able to create novel sequences (Figure 8) [159,160]. In Churakov et al. analysis of rodent TinT, multiple copies of novel dimeric/trimeric SINEs (SP-D-Geo, tri-Spe, and twinID- Spe) were discovered in rodent species [24]. Their analysis not only showed that these chimeric SINE are possibly created via TinT but they also have the advantage of better retropositional efficiency.

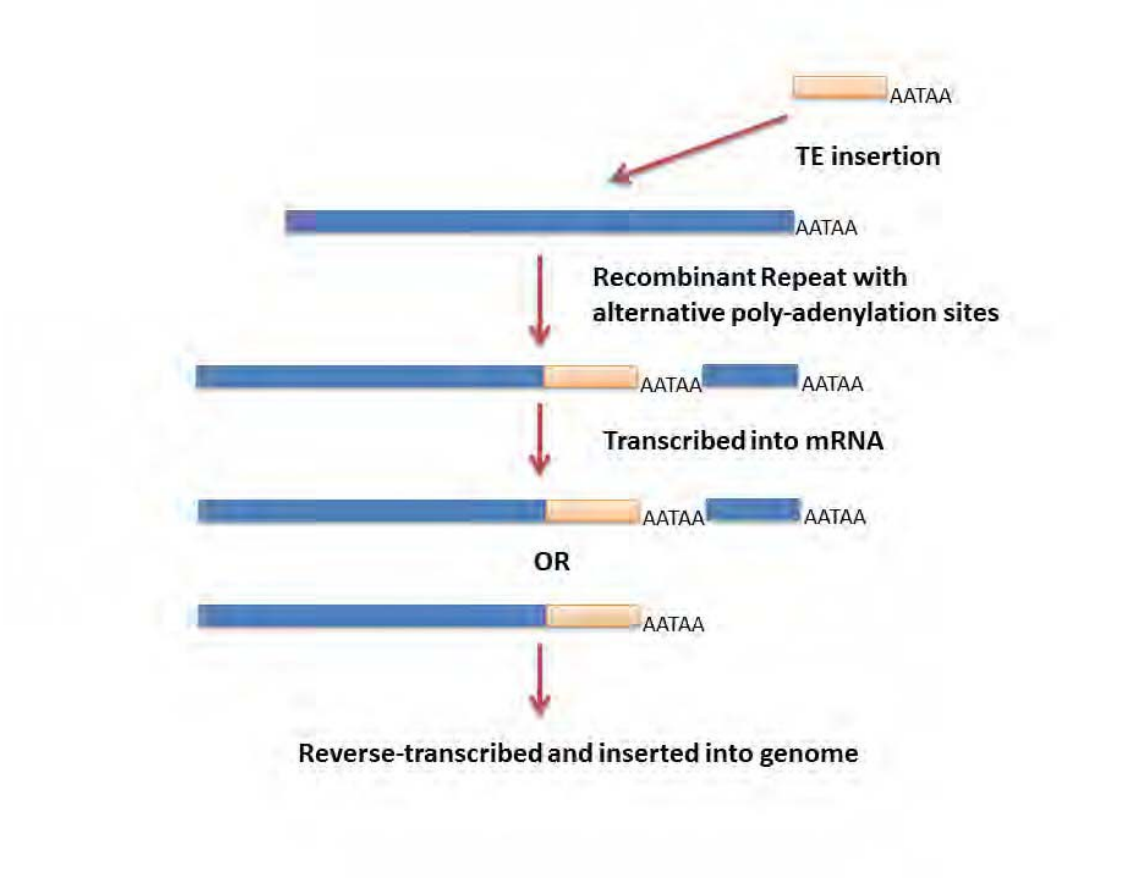


Figure 8. Novel TE/RR generated via TinT mechanism. During the active transposition event, the active TE inserts into another TE, thus creating a chimeric

TE. If the promoter still remained intact, it can be transcribed into 2 possible forms through different polyadenylation signals (AATAA), and retrotransposed back into the genome.

5.2 Alternative Splicing

Alternative splicing is a common process in eukaryotes. It allows a single gene to encode multiple protein variants. In this process, specific exons may be included or excluded from the final gene transcript, producing alternatively spliced mRNA [161]. TE insertions into protein-coding gene regions enable the production of novel transcript variants (containing partial TE sequences) via alternative splicing [162-167]. If the 3' ends of these novel transcripts contain sequence motifs that are recognized by reverse transcriptases, the novel transcripts can retrotranspose into genome. This process was demonstrated by Bantysh et al. and Hancks et al., who identified a human-specific RR family consisting of MAST2 gene sequences fused with an SVA element (Figure 9) [168,169]. This family was created by alternative splicing of a fusion of the first exon of the MAST2 gene with an SVA inserted into the first intron of the MAST2 gene. These transcript variants were not only expressed, but were capable of retrotransposition. Identification of novel TE created via alternative-splicing of coding gene transcripts has been reported from analysis of EST data [170-172].

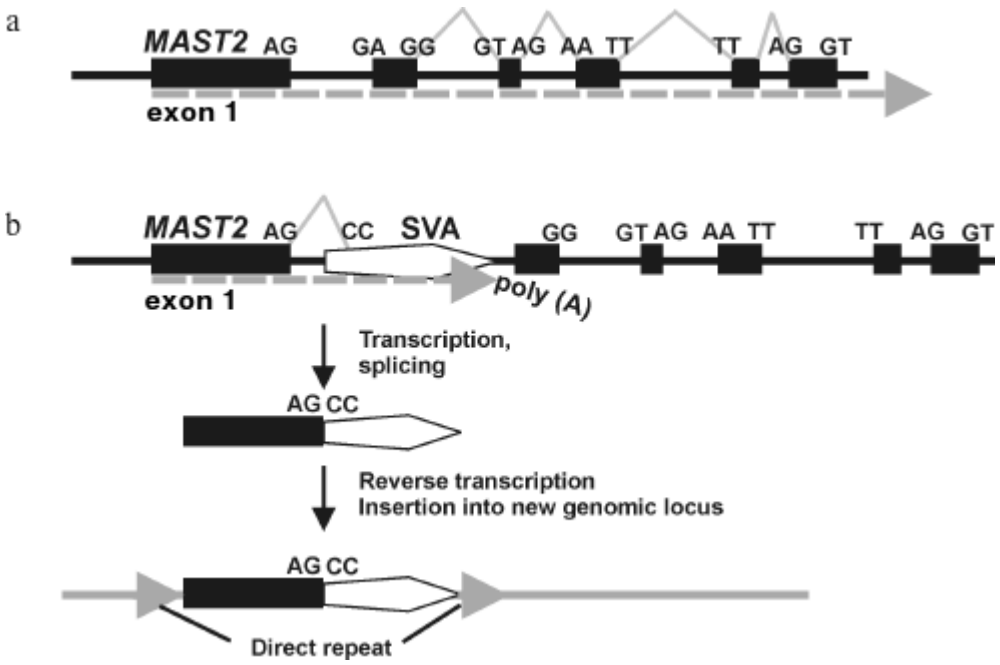


Figure 9. Novel TE created via alternative splicing mechanism. In this diagram, SVA was inserted into the first intronic region of the MAST2 gene. During the transcription event, it produces a novel transcript where the first MAST2 exon will fuse with SVA via alternative splicing. This novel transcript could retrotranspose back to genome, and forms a novel TE family (Adapted and modified from [168].

5.3 DNA Repair

Formation of RR via DNA repair was first shown by Han et al. using *in vivo* experiments [173,174]. They found that if a double strand break (DSB) occurred in a region containing 2 similar TE, DNA DSB repair was initiated and could induce SV via 2 different methods: a) non-homologous end-joining (NHEJ) repair or b) non-allelic homologous recombination (NAHR) repair (Figure 10) [175]. During the NHEJ DNA repair mechanism process, both 'broken' TE were joined and repaired via sequence microhomology [176-178]. NAHR on the other hand allowed the broken TE to be repaired through homologous recombination [179-181]. DSB repair processes not only contributed to DNA deletion, but

created novel RR [173,174]. Furthermore, it also has been observed in Arabidopsis that insertion of plastid genome sequences into the nuclear genome, results in complex arrays of plastid, repetitive and other sequences that appear to be the result of double strand break repair [182]. Although there is no evidence to show these novel TE/RR are expressed or retrotransposed, they provide a potential mechanism to explain how novel TE families can originate.

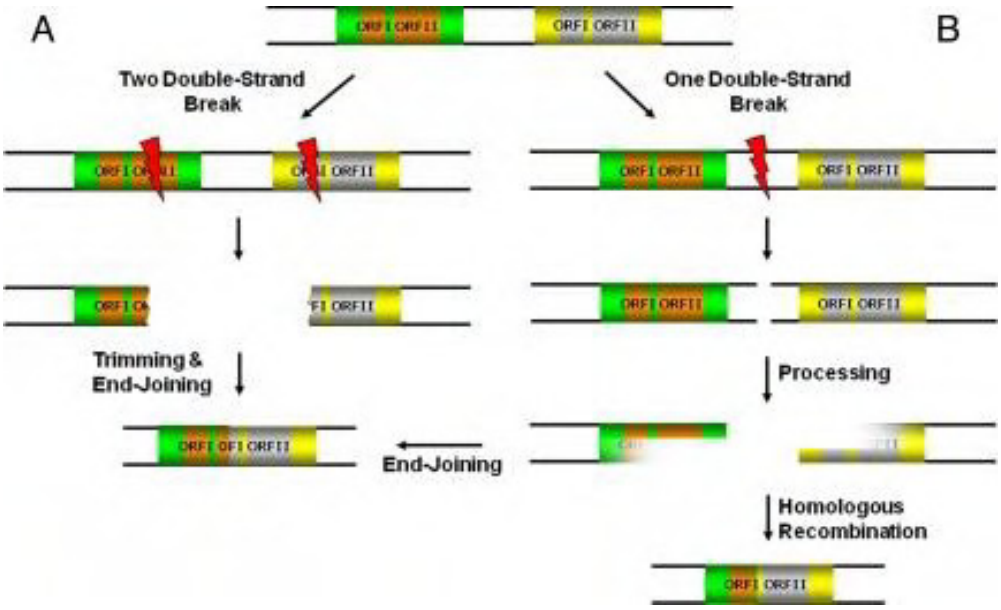


Figure 10. Chimeric TE formation via DNA double-strand break repair, non-homologous end joining (NHEJ) and non-allelic homologous recombination (NAHR). As a result of a DNA double strand break event between two similar TE, the nucleus will activate emergency DNA repair. This process can result in novel TE that differ from the original sequences (Adapted and modified from [173].

5.4 Transduction

It is known that LINE without strong polyadenylation signals are capable of transducing the genomic sequences near their 3'-ends [59,84,183-187]. During transcription, RNA polymerase produces run on transcripts of LINE and 3' flanking sequences [84]. This chimeric RNA can then be retrotransposed back into the

genome (Figure 11) [184,188]. This transduction mechanism is not limited to LINE, as it has been shown that the SVA element (SINE) insertions contain 5' and 3' transduction sequences [186,189]. If these transductions retain retrotransposon activity they can create novel TE families.

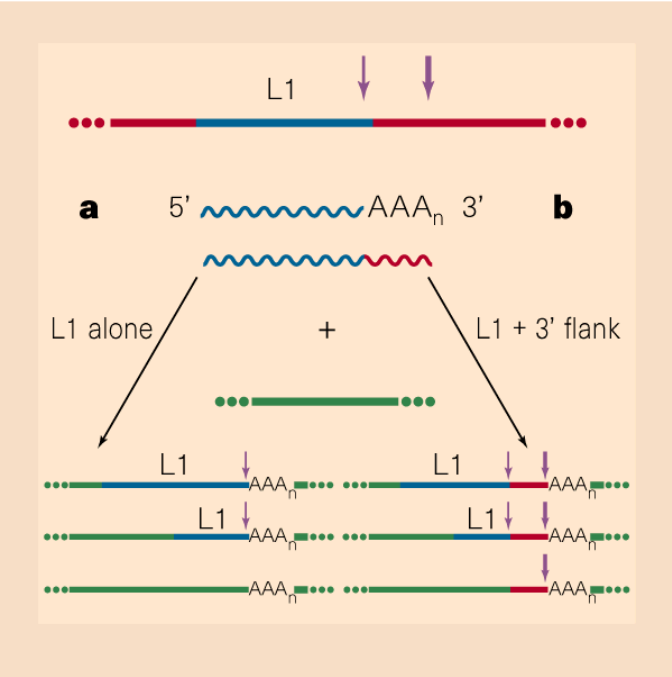


Figure 11. Potential novel TE created via transduction. During non-LTR retrotransposon transcription, the polymerase is able to transcribe the L1 into two forms: a) an intact L1 mRNA copy, or b) an intact L1 mRNA copy with extra sequences derived from its 3' flanking region. If the L1 continues to replicate the '3'-extended' copies, they have potential to become novel TE family. (Adapted and modified from [190]).

5.5 Template Switching

During retrovirus infection, retrovirus reverse transcriptase is able to switch from one RNA template to another RNA template via the RNA recombination process known as template switching [158,191,192]. This mechanism increases the retroviral genome diversity and shapes the mosaic structure of most retroviruses.

This process is not only limited to retroviruses, it can occur in TE as well [23,26,147,149,193,194]. Buzdin et. al. found an unusual TE family composed of U6 RNA fused with 3'-truncated L1 in human [19]. They concluded that during the reverse- transcription process, the LINE RNP complex was able to switch its LINE RNA to another RNA template via RNA recombination, and create a chimeric cDNA that was retrotransposed back into genome (Figure 12) [26,148]. TE created via template switching have been found in vertebrates and plants, indicating that it is a potential pathway to create RR and novel TE families [26,153].

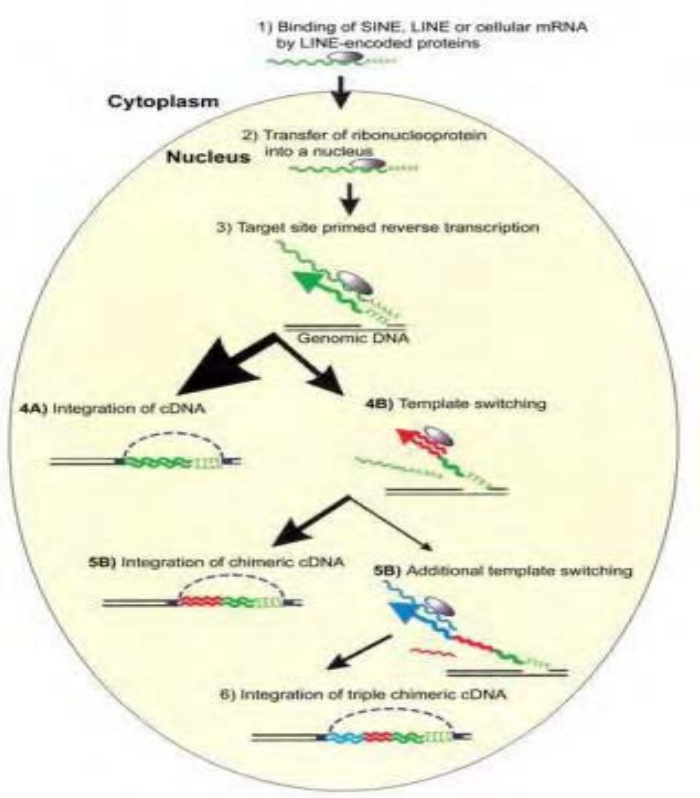


Figure 12. Novel TE created via template-switching (TS). During the retrotransposition process, the LINE, SINE or other mRNA is bound to a protein complex (Process 1). The RNA complex is imported from the cytoplasm to the nucleus (Process 2). The RNA is primed to a genomic location via TPRT, and starts to reverse-transcribe (Process 3). An intact or truncated copy of the cDNA

integrates into genome (Process 4A), or TS happens and the original RNA template is switched to a nearby mRNA transcript (Process 4b), forming bipartite (Process 5B, left), or tripartite (Process 5B, right) chimeric repeats that integrate into the genome (Process 6) (Adapted and modified from [16]).

6. Genomic Distribution Of RR Is Uncharacterized

Although RR have been identified in multiple studies, the exact RR copy number and their distribution in various genomes are unknown. Current software tools are ineffective in annotating RR because their chimeric structure makes them distinct from consensus TE sequences [195,196]. Homology-based search tools such as RepeatMasker and Censor are unable to identify intact RR because they rely on the available consensus sequences in the Repbase library (http://www.repeatmasker.org, [197]). *De novo* identification tools such as PALS/PILER and RepeatScout are effective methods to identify RR but they are time-consuming and their output results are heavily dependent on the parameter values selected by the user [198-200]. There is a need for methods to identify and annotate RR in order to determine their prevalence and significance.

7. Conclusion

TE are rich resources for the creation of novel functions or SV in genomes. They are important players that can shape species diversity and evolution. TE characteristics and their replication machinery are well understood, but the origin of TE remains unknown. RR serve as important models in understanding how novel TE might arise via different mechanisms, but the majority of RR presently remain uncharacterized. Thorough characterization of RR will help us understand the potential mechanisms that create novel RR families and their role(s) in genome

evolution.

8. References

1. Walsh AM, Kortschak RD, Gardner MG, Bertozzi T, Adelson DL (2013) Widespread horizontal transfer of retrotransposons. Proc Natl Acad Sci U S A 110: 1012-1016.
2. McClintock B (1950) The origin and behavior of mutable loci in maize. Proc Natl Acad Sci U S A 36: 344-355.
3. McClintock B (1956) Controlling elements and the gene. Cold Spring Harbor symposia on quantitative biology 21: 197-216.
4. S O (1972) So much "junk" DNA in our genome. Brookhaven Symp Biol 23: 366-370.
5. Goodier JL, Kazazian HH (2008) Retrotransposons revisited: The restraint and rehabilitation of parasites. Cell 135: 23-35.
6. Beck CR, Collier P, Macfarlane C, Malig M, Kidd JM, et al. (2010) LINE-1 Retrotransposition Activity in Human Genomes. Cell 141: 1159-U1110.
7. Elsik CG, Tellam RL, Worley KC, Gibbs RA, Muzny DM, et al. (2009) The genome sequence of taurine cattle: a window to ruminant biology and evolution. Science 324: 522-528.
8. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921.
9. Wade CM, Giulotto E, Sigurdsson S, Zoli M, Gnerre S, et al. (2009) Genome sequence, comparative analysis, and population genetics of the domestic horse. Science 326: 865-867.
10. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520-562.
11. Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, et al. (2003) Hot L1s account for the bulk of retrotransposition in the human population. Proc Natl Acad Sci U S A 100: 5280-5285.
12. Ewing AD, Kazazian HH, Jr. (2010) High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. Genome Res 20: 1262-1270.
13. Huang CR, Schneider AM, Lu Y, Niranjan T, Shen P, et al. (2010) Mobile interspersed repeats are major structural variants in the human genome. Cell 141: 1171-1182.
14. Levy A, Schwartz S, Ast G (2010) Large-scale discovery of insertion hotspots and preferential integration sites of human transposed elements. Nucleic acids research 38: 1515-1530.
15. Xing J, Zhang Y, Han K, Salem AH, Sen SK, et al. (2009) Mobile elements create structural variation: analysis of a complete human genome. Genome research 19: 1516-1526.
16. Buzdin AA (2004) Retroelements and formation of chimeric retrogenes. Cell Mol Life Sci 61: 2046-2059.
17. Wessler SR (1998) Transposable elements and the evolution of gene expression. Symp Soc Exp Biol 51: 115-122.
18. Chen Y, Zhou Z, Zhao G, Li X, Song L, et al. (2014) Transposable Element rbg Induces the Differential Expression of opaque-2 Mutant Gene in Two Maize o2 NILs Derived from the Same Inbred Line. PLoS One 9: e85159.
19. Tufarelli C, Cruickshanks HA, Meehan RR (2013) LINE-1 activation and epigenetic silencing of suppressor genes in cancer: Causally related events? Mob Genet Elements 3: e26832.
20. Kazazian HH, Jr. (2004) Mobile elements: drivers of genome evolution. Science 303: 1626-1632.
21. Nadeau NJ, Whibley A, Jones RT, Davey JW, Dasmahapatra KK, et al. (2012) Genomic

islands of divergence in hybridizing Heliconius butterflies identified by large-scale targeted sequencing. Philos Trans R Soc Lond B Biol Sci 367: 343-353.
22. Yaakov B, Ben-David S, Kashkush K (2013) Genome-wide analysis of Stowaway-like MITEs in wheat reveals high sequence conservation, gene association, and genomic diversification. Plant Physiol 161: 486-496.
23. Buzdin A, Ustyugova S, Gogvadze E, Vinogradova T, Lebedev Y, et al. (2002) A new family of chimeric retrotranscripts formed by a full copy of U6 small nuclear RNA fused to the 3' terminus of l1. Genomics 80: 402-406.
24. Churakov G, Sadasivuni MK, Rosenbloom KR, Huchon D, Brosius J, et al. (2010) Rodent evolution: back to the root. Molecular biology and evolution 27: 1315-1326.
25. Giordano J, Ge Y, Gelfand Y, Abrusan G, Benson G, et al. (2007) Evolutionary history of mammalian transposons determined by genome-wide defragmentation. PLoS computational biology 3: e137.
26. Gogvadze E, Barbisan C, Lebrun MH, Buzdin A (2007) Tripartite chimeric pseudogene from the genome of rice blast fungus Magnaporthe grisea suggests double template jumps during long interspersed nuclear element (LINE) reverse transcription. Bmc Genomics 8.
27. Reznikoff WS (2003) Tn5 as a model for understanding DNA transposition. Mol Microbiol 47: 1199-1206.
28. Feschotte C, Pritham EJ (2007) DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet 41: 331-368.
29. Pritham EJ (2009) Transposable elements and factors influencing their success in eukaryotes. J Hered 100: 648-655.
30. Kapitonov VV, Jurka J (2001) Rolling-circle transposons in eukaryotes. Proc Natl Acad Sci U S A 98: 8714-8719.
31. Coates BS, Hellmich RL, Grant DM, Abel CA (2012) Mobilizing the genome of Lepidoptera through novel sequence gains and end creation by non-autonomous Lep1 Helitrons. DNA Res 19: 11-21.
32. Ivics Z, Hackett PB, Plasterk RH, Izsvak Z (1997) Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. Cell 91: 501-510.
33. Plasterk RH (1993) Molecular mechanisms of transposition and its control. Cell 74: 781-786.
34. Carlson CM, Largaespada DA (2005) Insertional mutagenesis in mice: new perspectives and tools. Nat Rev Genet 6: 568-580.
35. Dupuy AJ (2010) Transposon-based screens for cancer gene discovery in mouse models. Semin Cancer Biol 20: 261-268.
36. Ivics Z, Izsvak Z (2005) A whole lotta jumpin' goin' on: new transposon tools for vertebrate functional genomics. Trends Genet 21: 8-11.
37. Jacob HJ, Lazar J, Dwinell MR, Moreno C, Geurts AM (2010) Gene targeting in the rat: advances and opportunities. Trends Genet 26: 510-518.
38. Aronovich EL, Hall BC, Bell JB, McIvor RS, Hackett PB (2013) Quantitative analysis of alpha-L-iduronidase expression in immunocompetent mice treated with the Sleeping Beauty transposon system. PLoS One 8: e78161.
39. Sjeklocha LM, Wong PY, Belcher JD, Vercellotti GM, Steer CJ (2013) beta-Globin sleeping beauty transposon reduces red blood cell sickling in a patient-derived CD34(+)-based in vitro model. PLoS One 8: e80403.
40. Gogvadze E, Buzdin A (2009) Retroelements and their impact on genome evolution and functioning. Cellular and molecular life sciences : CMLS 66: 3727-3742.
41. Cheng Z, Menees TM (2004) RNA branching and debranching in the yeast retrovirus-like element Ty1. Science 303: 240-243.
42. Leib-Mosch C, Seifarth W (1995) Evolution and biological significance of human retroelements. Virus Genes 11: 133-145.

43. Urnovitz HB, Murphy WH (1996) Human endogenous retroviruses: nature, occurrence, and clinical implications in human disease. Clin Microbiol Rev 9: 72-99.
44. Wilhelm M, Wilhelm FX (2001) Reverse transcription of retroviruses and LTR retrotransposons. Cell Mol Life Sci 58: 1246-1262.
45. Eickbush TH, Jamburuthugoda VK (2008) The diversity of retrotransposons and the properties of their reverse transcriptases. Virus Res 134: 221-234.
46. Eichinger DJ, Boeke JD (1988) The DNA intermediate in yeast Ty1 element transposition copurifies with virus-like particles: cell-free Ty1 transposition. Cell 54: 955-966.
47. Mellor J, Malim MH, Gull K, Tuite MF, McCready S, et al. (1985) Reverse transcriptase activity and Ty RNA are associated with virus-like particles in yeast. Nature 318: 583-586.
48. Muller F, Laufer W, Pott U, Ciriacy M (1991) Characterization of products of TY1-mediated reverse transcription in Saccharomyces cerevisiae. Mol Gen Genet 226: 145-153.
49. Xu H, Boeke JD (1990) Localization of sequences required in cis for yeast Ty1 element transposition near the long terminal repeats: analysis of mini-Ty1 elements. Mol Cell Biol 10: 2695-2702.
50. Goodwin TJ, Poulter RT (2004) A new group of tyrosine recombinase-encoding retrotransposons. Mol Biol Evol 21: 746-759.
51. Poulter RT, Goodwin TJ (2005) DIRS-1 and the other tyrosine recombinase retrotransposons. Cytogenet Genome Res 110: 575-588.
52. Katoh I, Kurata SI (2013) Association of Endogenous Retroviruses and Long Terminal Repeats with Human Disorders. Front Oncol 3: 234.
53. Belshaw R, Watson J, Katzourakis A, Howe A, Woolven-Allen J, et al. (2007) Rate of recombinational deletion among human endogenous retroviruses. J Virol 81: 9437-9442.
54. Beauregard A, Curcio MJ, Belfort M (2008) The take and give between retrotransposable elements and their hosts. Annu Rev Genet 42: 587-617.
55. Cost GJ, Feng Q, Jacquier A, Boeke JD (2002) Human L1 element target-primed reverse transcription in vitro. The EMBO journal 21: 5899-5910.
56. Dombroski BA, Mathias SL, Nanthakumar E, Scott AF, Kazazian HH, Jr. (1991) Isolation of an active human transposable element. Science 254: 1805-1808.
57. Feng Q, Moran JV, Kazazian HH, Jr., Boeke JD (1996) Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell 87: 905-916.
58. Singer MF (1982) SINEs and LINEs: highly repeated short and long interspersed sequences in mammalian genomes. Cell 28: 433-434.
59. Szak ST, Pickeral OK, Makalowski W, Boguski MS, Landsman D, et al. (2002) Molecular archeology of L1 insertions in the human genome. Genome Biol 3: research0052.
60. Alisch RS, Garcia-Perez JL, Muotri AR, Gage FH, Moran JV (2006) Unconventional translation of mammalian LINE-1 retrotransposons. Genes Dev 20: 210-224.
61. Gu W, Ray DA, Walker JA, Barnes EW, Gentles AJ, et al. (2007) SINEs, evolution and genome structure in the opossum. Gene 396: 46-58.
62. Han JS, Boeke JD (2005) LINE-1 retrotransposons: modulators of quantity and quality of mammalian gene expression? Bioessays 27: 775-784.
63. Hohjoh H, Singer MF (1996) Cytoplasmic ribonucleoprotein complexes containing human LINE-1 protein and RNA. EMBO J 15: 630-639.
64. Kolosha VO, Martin SL (2003) High-affinity, non-sequence-specific RNA binding by the open reading frame 1 (ORF1) protein from long interspersed nuclear element 1 (LINE-1). J Biol Chem 278: 8112-8117.
65. Kulpa DA, Moran JV (2006) Cis-preferential LINE-1 reverse transcriptase activity in

ribonucleoprotein particles. Nat Struct Mol Biol 13: 655-660.
66. Martin SL (2006) The ORF1 protein encoded by LINE-1: structure and function during L1 retrotransposition. J Biomed Biotechnol 2006: 45621.
67. Martin SL, Cruceanu M, Branciforte D, Wai-Lun Li P, Kwok SC, et al. (2005) LINE-1 retrotransposition requires the nucleic acid chaperone activity of the ORF1 protein. J Mol Biol 348: 549-561.
68. Fanning T, Singer M (1987) The LINE-1 DNA sequences in four mammalian orders predict proteins that conserve homologies to retrovirus proteins. Nucleic Acids Res 15: 2251-2260.
69. Mathias SL, Scott AF, Kazazian HH, Jr., Boeke JD, Gabriel A (1991) Reverse transcriptase encoded by a human transposable element. Science 254: 1808-1810.
70. Moran JV, Holmes SE, Naas TP, DeBerardinis RJ, Boeke JD, et al. (1996) High frequency retrotransposition in cultured mammalian cells. Cell 87: 917-927.
71. Clements AP, Singer MF (1998) The human LINE-1 reverse transcriptase:effect of deletions outside the common reverse transcriptase domain. Nucleic Acids Res 26: 3528-3535.
72. Babushok DV, Ostertag EM, Courtney CE, Choi JM, Kazazian HH (2006) L1 integration in a transgenic mouse model. Genome Research 16: 240-250.
73. Kubo S, Seleme MC, Soifer HS, Perez JL, Moran JV, et al. (2006) L1 retrotransposition in nondividing and primary human somatic cells. Proc Natl Acad Sci U S A 103: 8036-8041.
74. Jurka J (1997) Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proc Natl Acad Sci U S A 94: 1872-1877.
75. Luan DD, Korman MH, Jakubczak JL, Eickbush TH (1993) Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell 72: 595-605.
76. Wei W, Gilbert N, Ooi SL, Lawler JF, Ostertag EM, et al. (2001) Human L1 retrotransposition: cis preference versus trans complementation. Mol Cell Biol 21: 1429-1439.
77. Adelson DL, Raison JM, Edgar RC (2009) Characterization and distribution of retrotransposons and simple sequence repeats in the bovine genome. Proceedings of the National Academy of Sciences of the United States of America 106: 12855-12860.
78. Adelson DL, Raison JM, Garber M, Edgar RC (2010) Interspersed repeats in the horse (Equus caballus); spatial correlations highlight conserved chromosomal domains. Animal genetics 41 Suppl 2: 91-99.
79. Burke WD, Malik HS, Jones JP, Eickbush TH (1999) The domain structure and retrotransposition mechanism of R2 elements are conserved throughout arthropods. Mol Biol Evol 16: 502-511.
80. Furano AV (2000) The biological properties and evolutionary dynamics of mammalian LINE-1 retrotransposons. Prog Nucleic Acid Res Mol Biol 64: 255-294.
81. Moran JV (1999) Human L1 retrotransposition: insights and peculiarities learned from a cultured cell retrotransposition assay. Genetica 107: 39-51.
82. Yang J, Eickbush TH (1998) RNA-induced changes in the activity of the endonuclease encoded by the R2 retrotransposable element. Mol Cell Biol 18: 3455-3465.
83. Kramerov DA, Vassetzky NS (2005) Short retroposons in eukaryotic genomes. International review of cytology 247: 165-221.
84. Dewannieux M, Esnault C, Heidmann T (2003) LINE-mediated retrotransposition of marked Alu sequences. Nature genetics 35: 41-48.
85. Ohshima K, Hamada M, Terai Y, Okada N (1996) The 3' ends of tRNA-derived short interspersed repetitive elements are derived from the 3' ends of long interspersed repetitive elements. Mol Cell Biol 16: 3756-3764.

86. Raiz J, Damert A, Chira S, Held U, Klawitter S, et al. (2012) The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery. Nucleic Acids Research 40: 1666-1683.
87. Eickbush TH (1992) Transposing without ends: the non-LTR retrotransposable elements. New Biol 4: 430-440.
88. Kazazian HH, Jr., Moran JV (1998) The impact of L1 retrotransposons on the human genome. Nat Genet 19: 19-24.
89. Hickey DA (1982) Selfish DNA: a sexually-transmitted nuclear parasite. Genetics 101: 519-531.
90. Hickey DA (1992) Evolutionary dynamics of transposable elements in prokaryotes and eukaryotes. Genetica 86: 269-274.
91. Britten RJ (2010) Transposable element insertions have strongly affected human evolution. Proc Natl Acad Sci U S A 107: 19945-19948.
92. Roy-Engel AM, Carroll ML, El-Sawy M, Salem AH, Garber RK, et al. (2002) Non-traditional Alu evolution and primate genomic diversity. J Mol Biol 316: 1033-1040.
93. Salem AH, Kilroy GE, Watkins WS, Jorde LB, Batzer MA (2003) Recently integrated Alu elements and human genomic diversity. Mol Biol Evol 20: 1349-1361.
94. Xing J, Salem AH, Hedges DJ, Kilroy GE, Watkins WS, et al. (2003) Comprehensive analysis of two Alu Yd subfamilies. J Mol Evol 57 Suppl 1: S76-89.
95. Schneider H, Sampaio I (2013) The systematics and evolution of New World primates - A review. Mol Phylogenet Evol.
96. Ahmed M, Li W, Liang P (2013) Identification of three new Alu Yb subfamilies by source tracking of recently integrated Alu Yb elements. Mob DNA 4: 25.
97. Lee JY, Ji Z, Tian B (2008) Phylogenetic analysis of mRNA polyadenylation sites reveals a role of transposable elements in evolution of the 3'-end of genes. Nucleic Acids Res 36: 5581-5590.
98. Diehl WE, Johnson WE, Hunter E (2013) Elevated rate of fixation of endogenous retroviral elements in Haplorhini TRIM5 and TRIM22 genomic sequences: impact on transcriptional regulation. PLoS One 8: e58532.
99. Kremer D, Schichel T, Forster M, Tzekova N, Bernard C, et al. (2013) Human endogenous retrovirus type W envelope protein inhibits oligodendroglial precursor cell differentiation. Ann Neurol 74: 721-732.
100. Bieche I, Laurent A, Laurendeau I, Duret L, Giovangrandi Y, et al. (2003) Placenta-specific INSL4 expression is mediated by a human endogenous retrovirus element. Biol Reprod 68: 1422-1429.
101. Boronat S, Richard-Foy H, Pina B (1997) Specific deactivation of the mouse mammary tumor virus long terminal repeat promoter upon continuous hormone treatment. J Biol Chem 272: 21803-21810.
102. Caricasole A, Ward-van Oostwaard D, Zeinstra L, van den Eijnden-van Raaij A, Mummery C (2000) Bone morphogenetic proteins (BMPs) induce epithelial differentiation of NT2D1 human embryonal carcinoma cells. Int J Dev Biol 44: 443-450.
103. Chene L, Nugeyre MT, Barre-Sinoussi F, Israel N (1999) High-level replication of human immunodeficiency virus in thymocytes requires NF-kappaB activation through interaction with thymic epithelial cells. J Virol 73: 2064-2073.
104. Conley AB, Piriyapongsa J, Jordan IK (2008) Retroviral promoters in the human genome. Bioinformatics 24: 1563-1567.
105. de Parseval N, Alkabbani H, Heidmann T (1999) The long terminal repeats of the HERV-H human endogenous retrovirus contain binding sites for transcriptional regulation by the Myb protein. J Gen Virol 80 (Pt 4): 841-845.
106. Dunn CA, van de Lagemaat LN, Baillie GJ, Mager DL (2005) Endogenous retrovirus long terminal repeats as ready-to-use mobile promoters: the case of primate beta3GAL-T5. Gene 364: 2-12.

107. Knossl M, Lower R, Lower J (1999) Expression of the human endogenous retrovirus HTDV/HERV-K is enhanced by cellular transcription factor YY1. J Virol 73: 1254-1261.
108. Medstrand P, Landry JR, Mager DL (2001) Long terminal repeats are used as alternative promoters for the endothelin B receptor and apolipoprotein C-I genes in humans. J Biol Chem 276: 1896-1903.
109. Schon U, Seifarth W, Baust C, Hohenadl C, Erfle V, et al. (2001) Cell type-specific expression and promoter activity of human endogenous retroviral long terminal repeats. Virology 279: 280-291.
110. Sjottem E, Anderssen S, Johansen T (1996) The promoter activity of long terminal repeats of the HERV-H family of human retrovirus-like elements is critically dependent on Sp1 family proteins interacting with a GC/GT box located immediately 3' to the TATA box. J Virol 70: 188-198.
111. Suntsova M, Gogvadze EV, Salozhin S, Gaifullin N, Eroshkin F, et al. (2013) Human-specific endogenous retroviral insert serves as an enhancer for the schizophrenia-linked gene PRODH. Proc Natl Acad Sci U S A 110: 19472-19477.
112. Kim DS, Hahn Y (2011) Identification of human-specific transcript variants induced by DNA insertions in the human genome. Bioinformatics 27: 14-21.
113. Holdt LM, Hoffmann S, Sass K, Langenberger D, Scholz M, et al. (2013) Alu elements in ANRIL non-coding RNA at chromosome 9p21 modulate atherogenic cell functions through trans-regulation of gene networks. PLoS Genet 9: e1003588.
114. Liang KH, Yeh CT (2013) A gene expression restriction network mediated by sense and antisense Alu sequences located on protein-coding messenger RNAs. BMC Genomics 14: 325.
115. Pace JK, 2nd, Sen SK, Batzer MA, Feschotte C (2009) Repair-mediated duplication by capture of proximal chromosomal DNA has shaped vertebrate genome evolution. PLoS Genet 5: e1000469.
116. Lupski JR (2010) Retrotransposition and structural variation in the human genome. Cell 141: 1110-1112.
117. Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. Nat Rev Genet 7: 85-97.
118. Sharp AJ, Cheng Z, Eichler EE (2006) Structural variation of the human genome. Annu Rev Genomics Hum Genet 7: 407-442.
119. Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, et al. (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature 407: 513-516.
120. Hinds DA, Kloek AP, Jen M, Chen X, Frazer KA (2006) Common deletions and SNPs are in linkage disequilibrium in the human genome. Nat Genet 38: 82-85.
121. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, et al. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409: 928-933.
122. International HapMap C (2005) A haplotype map of the human genome. Nature 437: 1299-1320.
123. Bhangale TR, Rieder MJ, Livingston RJ, Nickerson DA (2005) Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes. Hum Mol Genet 14: 59-69.
124. Bhangale TR, Stephens M, Nickerson DA (2006) Automating resequencing-based detection of insertion-deletion polymorphisms. Nat Genet 38: 1457-1462.
125. McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, et al. (2006) Common deletion polymorphisms in the human genome. Nat Genet 38: 86-92.
126. Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, et al. (2006) An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res 16: 1182-1190.
127. Larue BL, Lagace R, Chang CW, Holt A, Hennessy L, et al. (2014) Characterization of

114 insertion/deletion (INDEL) polymorphisms, and selection for a global INDEL panel for human identification. Leg Med (Tokyo) 16: 26-32.
128. Ribeiro-Dos-Santos AM, de Souza JE, Almeida R, Alencar DO, Barbosa MS, et al. (2013) High-throughput sequencing of a South american amerindian. PLoS One 8: e83340.
129. Cooper GM, Nickerson DA, Eichler EE (2007) Mutational and selective effects on copy-number variants in the human genome. Nat Genet 39: S22-29.
130. Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, et al. (2006) Copy number variation: new insights in genome diversity. Genome Res 16: 949-961.
131. Locke DP, Sharp AJ, McCarroll SA, McGrath SD, Newman TL, et al. (2006) Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. Am J Hum Genet 79: 275-290.
132. Dweep H, Georgiou GD, Gretz N, Deltas C, Voskarides K, et al. (2013) CNVs-microRNAs Interactions Demonstrate Unique Characteristics in the Human Genome. An Interspecies in silico Analysis. PLoS One 8: e81204.
133. Shishido E, Aleksic B, Ozaki N (2013) Copy-number variation in the pathogenesis of autism spectrum disorder. Psychiatry Clin Neurosci.
134. Scherer SW, Lee C, Birney E, Altshuler DM, Eichler EE, et al. (2007) Challenges and standards in integrating surveys of structural variation. Nat Genet 39: S7-15.
135. Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK (2006) A high-resolution survey of deletion polymorphism in the human genome. Nat Genet 38: 75-81.
136. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, et al. (2004) Detection of large-scale variation in the human genome. Nat Genet 36: 949-951.
137. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, et al. (2004) Large-scale copy number polymorphism in the human genome. Science 305: 525-528.
138. Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, et al. (2005) Fine-scale structural variation of the human genome. Nat Genet 37: 727-732.
139. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, et al. (2007) Paired-end mapping reveals extensive structural variation in the human genome. Science 318: 420-426.
140. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, et al. (2006) Global variation in copy number in the human genome. Nature 444: 444-454.
141. Hormozdiari F, Konkel MK, Prado-Martinez J, Chiatante G, Herraez IH, et al. (2013) Rates and patterns of great ape retrotransposition. Proc Natl Acad Sci U S A 110: 13457-13462.
142. Kahyo T, Tao H, Shinmura K, Yamada H, Mori H, et al. (2013) Identification and association study with lung cancer for novel insertion polymorphisms of human endogenous retrovirus. Carcinogenesis 34: 2531-2538.
143. Schrider DR, Navarro FC, Galante PA, Parmigiani RB, Camargo AA, et al. (2013) Gene copy-number polymorphism caused by retrotransposition in humans. PLoS Genet 9: e1003242.
144. Witherspoon DJ, Zhang Y, Xing J, Watkins WS, Ha H, et al. (2013) Mobile element scanning (ME-Scan) identifies thousands of novel Alu insertions in diverse human populations. Genome Res 23: 1170-1181.
145. Guliyev M, Yilmaz S, Sahin K, Marakli S, Gozukirmizi N (2013) Human endogenous retrovirus-H insertion screening. Mol Med Rep 7: 1305-1309.
146. Iskow RC, McCabe MT, Mills RE, Torene S, Pittard WS, et al. (2010) Natural mutagenesis of human genomes by endogenous retrotransposons. Cell 141: 1253-1261.
147. Buzdin A, Gogvadze E, Kovalskaya E, Volchkov P, Ustyugova S, et al. (2003) The human genome contains many types of chimeric retrogenes generated through in vivo RNA recombination. Nucleic Acids Research 31: 4385-4390.
148. Buzdin A, Gogvadze E, Lebrun MH (2007) Chimeric retrogenes suggest a role for

the nucleolus in LINE amplification. FEBS letters 581: 2877-2882.
149. Hasnaoui M, Doucet AJ, Meziane O, Gilbert N (2009) Ancient repeat sequence derived from U6 snRNA in primate genomes. Gene 448: 139-144.
150. Hancks DC, Kazazian HH, Jr. (2010) SVA retrotransposons: Evolution and genetic instability. Seminars in cancer biology 20: 234-245.
151. Wang H, Xing J, Grover D, Hedges DJ, Han K, et al. (2005) SVA elements: a hominid-specific retroposon family. Journal of molecular biology 354: 994-1007.
152. Kriegs JO, Zemann A, Churakov G, Matzke A, Ohme M, et al. (2010) Retroposon insertions provide insights into deep lagomorph evolution. Molecular biology and evolution 27: 2678-2681.
153. Sabot F, Schulman AH (2007) Template switching can create complex LTR retrotransposon insertions in Triticeae genomes. BMC genomics 8: 247.
154. Kriegs JO, Matzke A, Churakov G, Kuritzin A, Mayr G, et al. (2007) Waves of genomic hitchhikers shed light on the evolution of gamebirds (Aves: Galliformes). BMC evolutionary biology 7: 190.
155. Churakov G, Grundmann N, Kuritzin A, Brosius J, Makalowski W, et al. (2010) A novel web-based TinT application and the chronology of the Primate Alu retroposon activity. BMC evolutionary biology 10: 376.
156. Gohl DM, Freifeld L, Silies M, Hwa JJ, Horowitz M, et al. (2013) Large-Scale Mapping of Transposable Element Insertion Sites Using Digital Encoding of Sample Identity. Genetics.
157. Nasri S, Abdollahi Mandoulakani B, Darvishzadeh R, Bernousi I (2013) Retrotransposon insertional polymorphism in Iranian bread wheat cultivars and breeding lines revealed by IRAP and REMAP markers. Biochem Genet 51: 927-943.
158. Shin W, Lee J, Son SY, Ahn K, Kim HS, et al. (2013) Human-specific HERV-K insertion causes genomic variations in the human genome. PLoS One 8: e60605.
159. Giles KE, Caputi M, Beemon KL (2004) Packaging and reverse transcription of snRNAs by retroviruses may generate pseudogenes. Rna-a Publication of the Rna Society 10: 299-307.
160. Nishihara H, Smit AFA, Okada N (2006) Functional noncoding sequences derived from SINEs in the mammalian genome. Genome Research 16: 864-874.
161. Black DL (2003) Mechanisms of alternative pre-messenger RNA splicing. Annu Rev Biochem 72: 291-336.
162. Michel D, Chatelain G, Mauduit C, Benahmed M, Brun G (1997) Recent evolutionary acquisition of alternative pre-mRNA splicing and 3' processing regulations induced by intronic B2 SINE insertion. Nucleic Acids Res 25: 3228-3234.
163. Wu M, Li L, Sun Z (2007) Transposable element fragments in protein-coding regions and their contributions to human functional proteins. Gene 401: 165-171.
164. Del Arco A (2005) Novel variants of human SCaMC-3, an isoform of the ATP-Mg/P(i) mitochondrial carrier, generated by alternative splicing from 3'-flanking transposable elements. Biochem J 389: 647-655.
165. Papamichos SI (2013) An Alu exonization event allowing for the generation of a novel OCT4 isoform. Gene 512: 175-177.
166. Salem IH, Hsairi I, Mezghani N, Kenoun H, Triki C, et al. (2012) CAPN3 mRNA processing alteration caused by splicing mutation associated with novel genomic rearrangement of Alu elements. J Hum Genet 57: 92-100.
167. Taniguchi-Ikeda M, Kobayashi K, Kanagawa M, Yu CC, Mori K, et al. (2011) Pathogenic exon-trapping by SVA retrotransposon and rescue in Fukuyama muscular dystrophy. Nature 478: 127-131.
168. Bantysh OB, Buzdin AA (2009) Novel family of human transposable elements formed due to fusion of the first exon of gene MAST2 with retrotransposon SVA. Biochemistry Biokhimiia 74: 1393-1399.

169. Hancks DC, Ewing AD, Chen JE, Tokunaga K, Kazazian HH, Jr. (2009) Exon-trapping mediated by the human retrotransposon SVA. Genome Res 19: 1983-1991.
170. Maher CA, Palanisamy N, Brenner JC, Cao X, Kalyana-Sundaram S, et al. (2009) Chimeric transcript discovery by paired-end transcriptome sequencing. Proc Natl Acad Sci U S A 106: 12353-12358.
171. Unneberg P, Claverie JM (2007) Tentative mapping of transcription-induced interchromosomal interaction using chimeric EST and mRNA data. PLoS One 2: e254.
172. Belancio VP, Hedges DJ, Deininger P (2006) LINE-1 RNA splicing and influences on mammalian gene expression. Nucleic Acids Res 34: 1512-1521.
173. Han K, Lee J, Meyer TJ, Remedios P, Goodwin L, et al. (2008) L1 recombination-associated deletions generate human genomic variation. Proc Natl Acad Sci U S A 105: 19366-19371.
174. Sen SK, Han K, Wang J, Lee J, Wang H, et al. (2006) Human genomic deletions mediated by recombination between Alu elements. American journal of human genetics 79: 41-53.
175. Moore JK, Haber JE (1996) Cell cycle and genetic requirements of two pathways of nonhomologous end-joining repair of double-strand breaks in Saccharomyces cerevisiae. Mol Cell Biol 16: 2164-2173.
176. Budman J, Chu G (2005) Processing of DNA for nonhomologous end-joining by cell-free extract. EMBO J 24: 849-860.
177. Guirouilh-Barbat J, Huck S, Bertrand P, Pirzio L, Desmaze C, et al. (2004) Impact of the KU80 pathway on NHEJ-induced genome rearrangements in mammalian cells. Mol Cell 14: 611-623.
178. Wilson TE, Lieber MR (1999) Efficient processing of DNA ends during yeast nonhomologous end joining. Evidence for a DNA polymerase beta (Pol4)-dependent pathway. J Biol Chem 274: 23599-23609.
179. Beckmann JS, Estivill X, Antonarakis SE (2007) Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. Nat Rev Genet 8: 639-646.
180. Steinmann K, Cooper DN, Kluwe L, Chuzhanova NA, Senger C, et al. (2007) Type 2 NF1 deletions are highly unusual by virtue of the absence of nonallelic homologous recombination hotspots and an apparent preference for female mitotic recombination. Am J Hum Genet 81: 1201-1220.
181. Robberecht C, Voet T, Zamani Esteki M, Nowakowska BA, Vermeesch JR (2013) Nonallelic homologous recombination between retrotransposable elements is a driver of de novo unbalanced translocations. Genome Res 23: 411-418.
182. Lloyd AH, Timmis JN (2011) The origin and characterization of new nuclear genes originating from a cytoplasmic organellar genome. Mol Biol Evol 28: 2019-2028.
183. Ejima Y, Yang LC (2003) Trans mobilization of genomic DNA as a mechanism for retrotransposon-mediated exon shuffling. Human Molecular Genetics 12: 1321-1328.
184. Goodier JL, Ostertag EM, Kazazian HH, Jr. (2000) Transduction of 3'-flanking sequences is common in L1 retrotransposition. Human molecular genetics 9: 653-657.
185. Hancks DC, Goodier JL, Mandal PK, Cheung LE, Kazazian HH (2011) Retrotransposition of marked SVA elements by human L1s in cultured cells. Human Molecular Genetics 20: 3386-3400.
186. Xing J, Wang H, Belancio VP, Cordaux R, Deininger PL, et al. (2006) Emergence of primate genes by retrotransposon-mediated sequence transduction. Proceedings of the National Academy of Sciences of the United States of America 103: 17608-17613.
187. Belancio VP, Whelton M, Deininger P (2007) Requirements for polyadenylation at the 3' end of LINE-1 elements. Gene 390: 98-107.

188. Solyom S, Ewing AD, Hancks DC, Takeshima Y, Awano H, et al. (2012) Pathogenic orphan transduction created by a nonreference LINE-1 retrotransposon. Hum Mutat 33: 369-371.
189. Damert A, Raiz J, Horn AV, Lower J, Wang H, et al. (2009) 5'-Transducing SVA retrotransposon groups spread efficiently throughout the human genome. Genome Res 19: 1992-2008.
190. Boeke JD, Pickeral OK (1999) Retroshuffling the genomic deck. Nature 398: 108-109, 111.
191. Temin HM (1993) Retrovirus variation and reverse transcription: abnormal strand transfers result in retrovirus genetic variation. Proc Natl Acad Sci U S A 90: 6900-6903.
192. Zuniga S, Cruz JL, Sola I, Mateos-Gomez PA, Palacio L, et al. (2010) Coronavirus nucleocapsid protein facilitates template switching and is required for efficient transcription. J Virol 84: 2169-2175.
193. Hara T, Hirai Y, Baicharoen S, Hayakawa T, Hirai H, et al. (2012) A novel composite retrotransposon derived from or generated independently of the SVA (SINE/VNTR/Alu) transposon has undergone proliferation in gibbon genomes. Genes Genet Syst 87: 181-190.
194. Moisy C, Blanc S, Merdinoglu D, Pelsy F (2008) Structural variability of Tvv1 grapevine retrotransposons can be caused by illegitimate recombination. Theor Appl Genet 116: 671-682.
195. Lerat E (2010) Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. Heredity (Edinb) 104: 520-533.
196. Saha S, Bridges S, Magbanua ZV, Peterson DG (2008) Empirical comparison of ab initio repeat finding programs. Nucleic Acids Res 36: 2284-2294.
197. Kohany O, Gentles AJ, Hankus L, Jurka J (2006) Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics 7: 474.
198. Edgar RC (2007) PILER-CR: fast and accurate identification of CRISPR repeats. BMC Bioinformatics 8: 18.
199. Edgar RC, Myers EW (2005) PILER: identification and classification of genomic repeats. Bioinformatics 21 Suppl 1: i152-158.
200. Price AL, Jones NC, Pevzner PA (2005) De novo identification of repeat families in large genomes. Bioinformatics 21 Suppl 1: i351-358.

Statement of Authorship

Title of Paper	
Publication Status	○ Published, ○ Accepted for Publication, ○ Submitted for Publication, ○ Publication style
Publication Details	

Author Contributions

By signing the Statement of Authorship, each author certifies that their stated contribution to the publication is accurate and that permission is granted for the publication to be included in the candidate's thesis.

Name of Principal Author (Candidate)			
Contribution to the Paper			
Signature		Date	

Name of Co-Author			
Contribution to the Paper			
Signature		Date	

Name of Co-Author			
Contribution to the Paper			
Signature		Date	

Name of Co-Author			
Contribution to the Paper			
Signature		Date	

Chapter 2

Evolution of Novel Transposable Elements: Experimental Products of Recombinant Repeats?

Sim Lin Lim and David L. Adelson

School of Molecular and Biomedical Science
The University of Adelaide
North Terrace
Adelaide, 5005
South Australia
Australia

RESEARCH ARTICLE

Evolution of Novel Transposable Elements: Experimental Products of Recombinant Repeats?

Sim L. Lim, David L. Adelson*

School of Molecular and Biomedical Science
The University of Adelaide
North Terrace
Adelaide, 5005
South Australia
Australia

*Corresponding Author: email: david.adelson@adelaide.edu.au,
Tel: +61 (0)8 8313 7555,
Fax: +61 (0) 8313 4362

Running title: Evolution of Novel SINE

Abstract

About 40-50% of mammalian genomes are made up of repetitive elements, primarily retrotransposons. Repetitive elements' activities not only drive genome evolution and speciation, they also create unique structural variation in individual genomes as well. Retrotransposons can give rise to nested transposable elements, or recombinant repeat sequences in mammalian genomes. These recombinant repeats can act as specific diagnostic elements for identifying different species. Recombinant repeats also serve as important resources to study the possible mechanisms by which new active transposable elements arise. Due to the complex structure of recombinant repeats, they have largely remained unclassified. In this report, we have identified simple bipartite and tripartite recombinant repeats in four mammalian genomes (human, mouse, cow and horse). While most recombinant repeats were singletons. we were able to classify a small proportion into families and subfamilies based on their fragment compostion and likely mode of origin. Some of these recombinant repeats are ancestral sequences, indicating that recombinant repeat formation has been occurring since before the mammalian radiation. Our analysis has shown that these classified recombinants repeats were primarily generated through transposon-into-transposon processes and not through a template-switching mechanism. Our analysis also showed that some recombinant repeats are likely to be retrotranspositionally active and some may be generated through alternate polyadenylation signals. Our results provided strong support for the notion that recombinant repeats are experimental evolutionary byproducts derived from processes that create novel transposable elements in mammalian genomes.

1. Introduction

Repetitive DNA sequences are found in almost all eukaryotic genomes and are prevalent in metazoan genomes. In the past, many research studies considered these repetitive elements as “Junk DNA” or “genomic parasites”; as their presence appeared not to confer any advantages on their host genome. However, recent large scale DNA sequencing projects have shown that repetitive elements are important resources as they can cause structural variation [1]. Repetitive elements not only define species clades and taxa, they are also excellent resources for creating new functional protein-coding genes via exaptation.

Repetitive elements can be divided into two major types: A) Tandem repeats where two or more nucleotides are repeated, such as satellite or microsatellite DNA. Most tandem repeats are located in the telomeres and centromeres and have been demonstrated to have important roles in the control of chromosome structure and stability [2]. B) Transposable elements (TEs), the largest class of DNA sequences in many genomes, account for ~40-50% of mammalian genomes and up to 90% of some plant genomes. Transposable elements can be divided into two fundamental classes based on their transposition mechanism: DNA transposons that rely on a “cut and paste” mechanism, or LTR retrotransposons and non-LTR retrotransposons (SINE and LINE) that use “copy and paste” mechanisms [3].

Recombinant repeats arise from combinations of well-characterized repetitive elements and have been described as nested repeats, in different research studies [4-6]. They are the products generated by active transposable elements and they are found in various genomes [7-13]. Recombinant repeats have structures that are difficult to characterize because they can be made up from

different types of repetitive elements. Characteristics of recombinant repeats remain largely unknown, but they serve as important tools in studying species evolution. Some studies have used recombinant repeats as diagnostic elements to resolve rodent and rabbit evolutionary trees [8,12], and to study the timeline of active Alu elements in primate genomes [7]. Recombinant repeats can be used as highly informative species-specific or genus-specific genetic markers [6-8,14].

Recombinant repeats have also been used as evolutionary models to explain how novel transposable elements arise within a genome, especially short non-LTR retrotransposons (SINE) families that are genus-specific [15-17]. How SINEs originate is still an open question. They have a composite or modular structure: a 5' end with similarity with 5' tRNA, 7SL RNA or 5S rRNA promoters, a RNA-unrelated region and a 3' end with shared similarity to the 3' tail of LINEs [18]. The most accepted views on SINE origins rely on the proposed template-switching mechanism of Buzdin et al. to explain this structure [1,4,15,17,19]. This template-switching mechanism is based on the study of recombinant repeat structure, where the L1 reverse transcriptase switches from its own L1 mRNA to other nearby mRNA sequences through an RNA-RNA recombination process, thus creating new recombinant repeats or pseudogenes (and possibly SINEs) during L1 insertion [4,10,19,20]. However, there are other studies that have suggested direct transposon into transposon (TinT) insertion events as an alternative mechanism to create novel transposable elements by "accident" [5,6,21]. Until now, no detailed comparative study has been carried out to analyze recombinant repeats in order to determine which of these mechanisms predominates.

Because recombinant repeats are complex and difficult to characterize, current software tools often fail to identify them effectively [7,22-24]. RepeatMasker

software relies on existing Repbase repetitive element data and is unable to identify new recombinant repeats. Sequence alignment tools such as PALS/PILER and RepeatScout are alignment programs that can be used to search for new recombinant repeats during *de novo* repeat identification. However, these programs are time-consuming and still rely on annotation by RepeatMasker to resolve the origin of repeat sequences in recombinant repeats and lack a framework for classification of these repeats [25,26]. An alternative tool is the TinT application, which has been effective for the identification of new recombinant repeats, but at present, this is used to study recombinant repeats that created by single type TE only [7,14,27]. There is a need for methodologies to identify and study recombinant repeats in order to determine their prevalence, significance and likely mode of origin.

In this report, we describe a pipeline that we have used to identify and classify simple bipartite and tripartite recombinant repeats into families, in the following mammalian genomes: human (*Homo sapiens*), mouse (*Mus musculus*), cow (*Bos taurus*) and horse (*Equus Callabus*). Our methods have allowed us to systematically study classified recombinant repeats, resulting in the discovery of ancestral recombinant repeats that were present prior to the mammalian radiation. We have also shown that some recombinant repeats are very likely retrotranspositionally active. Finally, our comparative genomic analysis indicated that TinT insertion was the major mode of origin for new recombinant repeats and possibly novel transposable elements in mammalian genomes.

2. Materials and Methods

Database

The data used in this research was downloaded from public databases. The genome assemblies of horse (EqCab 2.0), cow (BosTau4.0), human (hg19), and mouse (mm9) were downloaded from the UCSC Genome Browser [28-31]. The haplotype sequences and unmapped sequences in the genome assembly were discarded in this analysis The species-specific repetitive element sequences (as of February 2011) were downloaded from Repbase (http://www.girinst.org/repbase/).

Software for recombinant repeat identification and classification

PERL, Postgres-SQL, R, Bedtools, RepeatMasker and TSDscan used in this pipeline are open source tools (Supplementary information 1). All of them are stand-alone versions running under the Linux environment.

Recombinant Repeats Identification Pipeline

The overview of the pipeline is shown in Supplementary information 1. Simple recombinant repeat characteristics (tripartite and bipartite) are shown in Figure 1. Mammalian genomes were masked with RepeatMasker (http://www.repeatmasker.org/) using species-specific libraries of repetitive elements from Repbase. Simple tripartite recombinant repeats made up of two types of element were identified based on 5 conditions: 1) Could only be composed by LINE, SINE, LTR or DNA transposon elements, 2) Could not be part of a single transposable element family, 3) Transposable element A contained an inserted transposable element B, 4) Sequences for A and B could not overlap for more than 10bp, 5) No gap between elements A and B longer than 10bp in length.

In cases where three elements were involved condition 3 was modified such that three types of transposable element could occur within a recombinant repeat and flanking regions of 500bp were repeat free.

Tripartite recombinant repeats that met the criteria were extracted and mapped back to the genome to recombinant repeats located in repeat rich regions which otherwise would have been missed.

The initial simple bipartite recombinant repeat identification was similar to simple tripartite recombinant repeats conditions 1, 2, 4, and 5, but associated with one additional condition: The recombinant repeat was made from 2 transposable elements (A,B) associated with 500bp flanking region free of repetitive elements.

The bipartite recombinant repeats that met the criteria were extracted. These bipartite recombinant repeats were mapped back to the genome to search for 'missed-out' recombinant repeats located in repetitive element rich region that would have been missed because of the 500bp non-repetitive flanking sequence condition.

Recombinant Repeat Classification Pipeline

PERL scripts were used to classify tripartite and bipartite recombinant repeats into families (Supplementary Information 1). The following criteria were used for clustering 1) Clustered recombinant repeats must be derived from single structure type (bipartite or tripartite). 2) Clustered recombinant repeats must share the same transposable element fragments. 3) Clustered recombinant repeat fragments must have similar orientations. 4) Individual repeat regions within recombinant repeats could not overlap for more than 10bp and could have no non-repetitive region between them longer than 10bp. 5) Each cluster had to contain at least three members.

Recombinant repeat length was ignored in the clustering process because of the prevalence of 5' truncated repeats.

Recombinant Repeat Analysis

Coordinates for classified recombinant repeats were stored in a Postgres-SQL database. SQL queries and PERL scripts were used to determine recombinant repeat sizes (bp) and carry out analyses. Repbase annotations were used to determine recombinant repeat classes. PERL scripts were used to examine the transposable elements annotated within recombinant repeats. Recombinant repeats were also divided into different types (LINE-SINE, LTR-SINE etc) based on this analysis.

Identifying Recombinant Repeats With Apparent Transpositional Activity

The full-length L1 (≥ 6kbp) and AluY (≥ 300bp) 50bp flanking regions were analysed using TSDscan [32,33]. We searched for potential TSD sites (≥ 5bp) based on TSDscan recommendation settings. The non-LTR TEs perfect TSD matches (without any mismatch in alignment) were extracted to plot non-LTR TE TSD distributions. Potential TSDs were identified in each classified LINE-SINE recombinant repeat's (human, cow, horse) external and internal 50-bp flanking regions.

Identifying potential CpG Islands and protein coding gene promoters flanking recombinant repeats

1kbp intervals flanking LINE-SINE recombinant repeats with perfect external TSD pairs and imperfect internal TSD pairs were identified for analysis. The protein-coding gene promoters were downloaded from UCSC genome browser (http://genome.ucsc.edu/) and CpG island intervals were identified using the criteria in Takai et. al. [34]. We used Bedtools to intersect the recombinant repeat 1kbp genomic intervals with the protein coding gene promoters and CpG island intervals in the relevant assembly

Distinguishing Template Switching Products From TinT Products

A Perl script was written to distinguish the LINE-SINE bipartite recombinant repeats generated via TinT or template switching mechanism (Supplementary Information 1). A bipartite recombinant repeat was considered to have originated via template switching if: 1) both transposable elements were in the same 5' to 3' orientation, and 2)the transposable element B was not located near the100bp of the transposable element A 3' end, as it could be potential TinT event. Bipartite recombinant repeats that did not meet these conditions were considered to have originated via TinT events.

Distinguishing Bipartite Recombinant Repeats From Truncated Tripartite Recombinant Repeats

We used Postgres-SQL queries and PERL to store and interrogate our data using scripts to identify possible bipartite recombinant repeats that shared similar structures with tripartite recombinant repeats (Supplementary Information 1). Bipartite repeats were counted as specific truncated tripartite recombinant repeat if the bipartite recombinant repeat fragments were exactly matched with any two adjacent tripartite recombinant repeat fragments, if the bipartite recombinant repeat shared the same boundaries with a tripartite recombinant repeat without overlap or decay for more than 10bp, and if three or more bipartite recombinant repeats shared the same structure with a single tripartite recombinant repeat.

3. Results

Simple Recombinant Repeat Identification Pipeline And General Characteristics

We identified bipartite and tripartite recombinant repeats from four mammalian genomes (human, cow, horse and mouse), by developing a computational pipeline based on publicly available software and PERL scripts as shown in

Supplementary Table 1 and Supplementary Figure 1. In this report we have not attempted to analyse recombinant repeats of higher order because the classifications for those increased combinatorially. To identify simple recombinant repeats, we used RepeatMasker output as the input data for the pipeline. We identified tripartite and bipartite recombinant repeats based on criteria summarised in Figure 1 (see material and methods for more detailed identification criteria). We did not include recombinant repeats created by a single type of transposable element class in order to avoid potential false positives caused by faulty RepeatMasker annotation. Non-LTR retrotransposons (LINE and SINE) were considered as distinct classes in this analysis. We were able to identify a total of 118,211 recombinant repeats from the human, mouse, cow and horse genomes (Table 1).

We classified the recombinant repeats into groups based on traditional transposable element classification as shown in Figure 1 (see material and methods for more detail). A recombinant group was established if 3 or more recombinant repeats shared exactly the same transposable element fragments, same strand orientation and similar boundaries. As a result of our classification process, 36,234 recombinant repeats were grouped into 3,314 bipartite families and 2,412 tripartite families. 81,977 recombinant repeats were singletons that could not be grouped. Most recombinant repeats (83.42% or 30,229 copies) were shorter than 1kbp, with about two thirds of these between 300-700bp long (Figure 2). However, there were 95 longer than 5kbp (82 LINE-SINE, 7 LTR-LINE, and 6 complex tripartite recombinant repeats).

Table 1 revealed that recombinant repeat copy number and genome coverage were somewhat variable across our four mammalian genomes, with recombinant

repeat copy number and coverage lowest in the mouse compared to human, horse and cow. When we examined recombinant repeat distribution in each chromosome our data showed that the genomic coverage of recombinant repeats was similar in human, horse and cow as judged by the distribution of distances between recombinant repeats (Figure 3). Recombinant repeat copy numbers in human, cow and horse were positively correlated with chromosome size (Supplementary File 1), but mouse recombinant repeat inter-repeat genomic distances were not similar to the other mammals, and their copy number did not correlate with chromosome size. The Y chromosome for human and mouse had very low recombinant repeat copy numbers and horse and cow chromosome Y data were not available, so the Y chromosome was not included in this analysis.

Recombinant Repeats: Content And Evolutionary Timeline

We divided recombinant repeats into 12 different types based on their transposable element content: LINE-SINE, LINE-LTR, LINE-DNA, LTR-SINE, LTR-LINE, LTR-DNA, DNA-SINE, DNA-LINE, DNA-LTR, SINE-LINE, SINE-LTR and SINE-DNA as shown in Figure 4. Tripartite recombinant repeats that contained 3 different types of transposable element (LINE-LTR-SINE, or LINE A-LTR-LINE-B), were categorized as Unknown tripartite recombinant repeats. SINE retrotransposons were the dominant contributors as they were present in 93.54% (33,892 copies) of the recombinant repeats (not including complex recombinant repeats). In fact, LINE-SINE recombinant repeats were the dominant type in human, cow and horse followed by LTR-SINE and DNA-SINE recombinant repeats. However, this was not true in mouse, where LINE-SINE recombinant repeats were present at much lower copy number. On the other hand, the cow had an exceptional number of LINE-SINE recombinant repeats compared with the

other mammals. About a third of the cow LINE-SINE recombinant repeats (29.02%, 3426 copies) were composed of clade specific BovB LINE that are not present in the other genomes.

In addition to our twelve types of recombinant repeats we divided them into the following three age groups. “Ancestral” recombinant repeats, that contained only transposable elements that stopped transposing before the mammalian radiation (LINE L2 or SINE MIR recombinant repeats). “Mixed” recombinant repeats, containing one ancestral transposable element, and one clade specific transposable element (ie, L2-AluSx recombinant repeats). “Lineage-specific” recombinant repeats, that contained only clade-specific transposable elements. About two thirds (59.46%, 21,545 copies) of recombinant repeats were lineage-specific, followed by mixed (31.95%, 11,577 copies) and ancestral (8.59%, 3,112 copies). While recombinant repeats have been around since before the mammalian radiation, and specific recombinant types were created in all four species, we found no ancestral recombinant repeats in mouse.

Potential Transpositional Ability of Recombinant Repeats

Previous studies suggested that active transposable elements generate Terminal Sequence Duplications (TSDs) from 5 to 20 bp long at their 5’ and 3’ ends and could be associated with 3’ polyA tails [33,35-39]. We therefore analysed the TSD sites of tripartite LINE-SINE recombinant repeats in order to infer their retrotranspositional potential. We expected for recombinant repeats that had retrotransposed that we would find perfect TSDs (without any mismatch) at their 5’ and 3’ ends, because retrotransposed sequences with perfect TSDs are indicative of recent insertion [38,40]. In addition, because recombinant repeats could be created by TinT, we expected that the internal fragment boundaries would be

flanked by TSDs as well, but that these internal TSD should be older and therefore less well conserved (1 or more mismatches). We identified perfect TSD pairs in the 50bp internal and external flanking regions (LINE-SINE) as shown in Figure 5.. Our results (Figure 6) showed that 2093 LINE-SINE recombinant repeats (663 in human, 764 in cow, 666 in horse) contained perfect internal SINE TSD pairs, indicating that recombinant repeats are created via SINE insertion. There were however, 454 recombinant LINE-SINEs with perfect external TSDs, albeit with a different length distribution compared to intact LINE TSDs. 384 of these, originating from 311 families, appeared to have been retrotransposed because they had imperfect internal TSDs. Further investigation showed that 161 of these families (51.8%) had members that could be transcribed and retrotransposed because they were within 1kbp of CpG islands (305 family members) or protein coding gene promoters (74 family members) (Table 2) [41-46]. It is worth noting that 10 ancestral recombinant repeats (with perfect external TSD pairs) appeared to have been retrotransposed based on our TSDscan results, indicating that the fossil recombinant repeat sequences were capable for retrotranspositions. We were unable to identify potentially retrotransposed LINE-SINE recombinant repeats in mouse, and this is consistent with the low copy number of mouse LINE-SINE recombinant repeats.

The Origin Of Recombinant Repeats

Previous studies have shown that there are two possible mechanisms to create novel transposable elements or recombinant repeats: TinT insertion and Template Switching (TS) [4,6]. In order to understand the origin of recombinant repeats, we examined the bipartite LINE-SINE recombinant repeat fragments and orientations to see which mechanism they were consistent with. Current literatures suggested

that non-LTR transposable elements are able to undergo template switching during reverse-transcription events in mammalians [4,47]. Therefore we only examined the LINE-SINE bipartite recombinant repeats for potential template switching events. Bipartites were selected because tripartite recombinant repeats were likely created via TinT events because their random TE class combinations and TE orientation combinations were inconsistent with template switching [6-8]. Figure 7 shows the breakdown of recombinant repeats consistent with either TinT or TS mechanisms and indicates that the vast majority (87.83%, 31,825 copies) of the classified recombinant repeats were consistent with having been created via TinT. This result was consistent across all four genomes.

Connection of Tripartite and Bipartite Recombinant Repeats

It has been demonstrated that SINE insertions into exonic or intronic regions of genes can create novel polyadenylation sites. These polyadenylation sites can affect transcription and translation, and contribute to splice-variants, pseudogenes, and possibly new transposable elements [1,48,49]. It is therefore critical to understand the potential of tripartite recombinant repeat to generate truncated copies (bipartite) via internal polyA sites/tails. We identified bipartite recombinant repeats that shared similar matches (boundaries, sequences direction and transposable element fragments) with tripartite recombinant repeats. A tripartite recombinant repeat was considered able to generate a bipartite if there were at least 3 bipartite recombinant repeats that met the above criteria. We found 1,290 bipartite recombinant repeats that shared exact structures and similar boundaries with 221 tripartite recombinant repeats belonging to LINE-SINE, LTR-SINE and DNA-SINE families. The SINE sequences of these recombinant repeats were full-length and associated with poly-A tails. This result demonstrates that if tripartite

recombinant repeats were transcribed, they had the potential to generate full-length (tripartite) or truncated (bipartite) recombinant repeats through the use of a premature polyadenylation signal [50].

4. Discussion

While a number of studies have shown that recombinant repeats provide valuable information for studying genome evolution, their computational identification and classification can be challenging [22,23]. The current methods used to identify recombinant repeats are not as effective as one might like, because they are not optimized to search the recombinant repeats [7,24-26,51]. RepeatMasker software is able to identify whole or partial repetitive elements based on consensus sequences. REPET, PALS and PILER are used for *de novo* identification of repeats in genome sequence. These programs are of limited use for finding recombinant repeats because they can only identify repeats that have >85% similarity, therefore missing older, more divergent repeats. RepeatScout uses a unique l-mer seed algorithm and is able to identify novel recombinant repeats, but it generates many false positive recombinant repeat sequences. The TinT identification pipeline is effective in identifying recombinant repeats, but can only detect recombinant repeats composed by single-type TE as currently implemented [7,14,27] Our recombinant repeat identification pipelines allow us to identify recombinant repeats without being affected by length considerations, which is a limitation of other recombinant repeat identification methods [5,7]. Furthermore, our pipeline is able to classify two different types of recombinant repeats (bipartite and tripartite) into different groups based on their repetitive element content and borders. Our pipeline also allows us to detect transposable element insertion polymorphism, search for new phylogenetic markers based on

repetitive elements, and identify possible new transposable element classes. Genome sequences from related species (eg primates) are useful for discovering novel recombinant repeats insertions, and can help identify recombinant repeat copy number variation within a same species. However, our pipeline may miss some of the simple recombinant repeats because some of these recombinant repeats hide in repeat-rich regions. For example, our pipeline cannot extract bipartite recombinant repeats in repeat-rich regions because it generates unlimited bipartite recombinant repeat patterns that increase false-positive results. Therefore we extract the bipartite recombinant repeats in repeat-free flanking regions, map them to the genome and pull the potential bipartite recombinant repeats. Besides that, the pipeline is unable to search higher complexity of recombinant repeats (more complex than tripartite) in genomes. Therefore the complex recombinant repeats have been excluded from our analysis. It will require additional effort and improved methods to identify and classify complex recombinant repeats.

Previous studies have used repetitive elements as evolutionary study tools in order to study genome evolution and speciation processes. [6-8,12,13]. Our analyses not only provides new recombinant repeats for studying genome evolution, but has allowed us to discover the presence of novel recombinant repeats that can be assigned to the evolutionary timeline. The existence of ancestral, mixed, and recent recombinant repeats indicate that specific repetitive elements started to shuffle and generate recombinant repeats during their active period. These experimental ‘mix-and-match’ recombinant repeats may provide the raw material to generate new lineage-specific transposable elements such as SVA retroelements. It is likely that recombinant repeats are a significant product of TE

evolutionary processes not only in mammalian genomes, but it is a common processes in other vertebrate genomes as well, where Sauria-SINE-Helitron recombinant repeats in reptilian genome (*Anolis carolinensis*) are more frequently recognized than Sauria-SINEs itself [52]. They help explain the presence of lineage specific repetitive elements.

Some mammalian retrotransposons (LINE L2 and SINE MIR) were active and then subsequently became inactive prior to the mammalian radiation. These ancestral repeats are therefore considered molecular fossils in the genomes of eutherian mammals, but have remained active in monotremes [53]. Studies showed that after the mammalian radiation, most of the ancestral repetitive elements became inactive [54,55]. In the past, it had been demonstrated that an inactive ancestral TE “Sleeping Beauty” had been “revived” via mutations [56]. Our analyses have shown that some ancestral recombinant repeats appeared to be retrotransposed based on the basis of TSD analysis. It indicates that some of the ancestral repeats have been brought back from the dead via recombinant repeat form.While some ancestral repeats have been shown to have been exapted into regulatory regions such as enhancers and undergo purifying selection [50,57-59], some ancestral repeats may be able to resist sequence decay not just through purifying selection, but by retrotransposition.However, we did not observe ancestral recombinant repeat families in mouse. We suspected that most ancestral recombinant repeats were disrupted by higher rate of newer TE insertions due to the mouse short-generation time (faster evolving rate) before they had a chance to undergo retrotranspositions.

Our results show that recombinant repeats have interesting characteristics that have not been previously described. First, the recombinant repeats classified into

families were low copy repeats (LCR), but they were not chromosome specific. Second, some of these recombinant repeats may not have been generated through TinT, because a small number of recombinant repeats have imperfect internal TSD but perfect external TSD suggesting that they could have been retrotransposed [38,40]. These recombinant retrotransposons could have been transcribed into mRNA because even though they might lack an internal promoter, they had family members with a flanking promoter. However, the TSD analysis should be used with caution and the TSDscan software might misannotated some recombinant repeat TSD and provide false-interpretations

Previous studies suggested two different mechanisms to create novel recombinant repeats. Template switching is the first proposed mechanism, and is based on observations of chimeric retroviruses [60]. The second mechanism is TinT (transposon-into-transposon), based on TEs natural characteristics, where they insert into the genome via Double-Strand Break repair or Non-Homologous End Joining (NHEJ) repair [61,62]. Our study supports the TinT activities are the the primary mechanism that contribute to the creation of novel TEs.

Previous studies have shown that specific TE insertion in exons or introns can generate alternative mRNA transcripts via the addition of TE polyadenylation sites [48,50,63,64]. We have shown that bipartite recombinant repeats may be created via alternate polyadenylation sites present in tripartite repeats. This is the first evidence to indicate that nested recombinant repeats have the potential to generate recombinant repeat variants and contribute to the evolution of new TEs as shown in Figure 8. However, this hypothesis is based on our current computational analysis only. We require further experimental validation to support this observation..

Various evolutionary models for transposable elements have been proposed in order to explain how novel transposable elements are created [15,17,65-67]. Our results suggest that the TinT mechanism can readily account for new transposable element (SINE) formation, and that new TE formation via template switching is a much less likely event. We can explain the formation of novel SINE retrotransposons by the insertion of 5' truncated LINE sequence into a tRNA gene 3' of the tRNA Pol III promoter. This TinT insertion would create a tripartite recombinant repeat that contains two possible polyadenylation signal, one from original tRNA, and the other one from 5' truncated LINE. During reverse transcription event, the LINE polymerase recognized the 5' truncated LINE polyadenylation signal and initiate the reverse-transcription process. It contributes to the creation of novel bipartite SINE, followed by integration into the genome. As a result of sequence decay, the original tripartite sequence might be degraded, but the novel SINEs would be retrotranspositionally active and amplify within the genome. This model not only accounts for the sudden rise of Alu monomers in primate genomes; it also explains the unusual structure of SVA elements which may arise via multiple TinT events [17,19,65,68]. Our results and proposed model are also consistent with the proposed evolutionary hypothesis for LTR retrotransposons, where LTR retrotransposons were chimeric repetitive elements created through the fusion of a non-LTR retrotransposon and a DNA transposon [67].

Conclusion

Recombinant repeats are a useful resource for studying the molecular evolutionary mechanisms that affect mammalian genomes. Our results provide an alternative framework to understand how novel repeats arise, in particular how

novel clade specific SINEs may arise. Further experimental work is needed to confirm our conclusions and experimentally validate the retrotransposition of novel recombinant repeats.

Acknowledgments

Thanks to Reuben Buckley for helpful discussion and comments and to Joy Raison for CpG island coordinates.

References

1. Gogvadze E, Buzdin A (2009) Retroelements and their impact on genome evolution and functioning. Cellular and molecular life sciences : CMLS 66: 3727-3742.
2. Plohl M, Luchetti A, Mestrovic N, Mantovani B (2008) Satellite DNAs between selfishness and functionality: structure, genomics and evolution of tandem repeats in centromeric (hetero)chromatin. Gene 409: 72-82.
3. Jurka J, Kapitonov VV, Kohany O, Jurka MV (2007) Repetitive sequences in complex genomes: structure and evolution. Annual review of genomics and human genetics 8: 241-259.
4. Buzdin A, Ustyugova S, Gogvadze E, Vinogradova T, Lebedev Y, et al. (2002) A new family of chimeric retrotranscripts formed by a full copy of U6 small nuclear RNA fused to the 3' terminus of l1. Genomics 80: 402-406.
5. Giordano J, Ge Y, Gelfand Y, Abrusan G, Benson G, et al. (2007) Evolutionary history of mammalian transposons determined by genome-wide defragmentation. PLoS computational biology 3: e137.
6. Kriegs JO, Matzke A, Churakov G, Kuritzin A, Mayr G, et al. (2007) Waves of genomic hitchhikers shed light on the evolution of gamebirds (Aves: Galliformes). BMC evolutionary biology 7: 190.
7. Churakov G, Grundmann N, Kuritzin A, Brosius J, Makalowski W, et al. (2010) A novel web-based TinT application and the chronology of the Primate Alu retroposon activity. BMC evolutionary biology 10: 376.
8. Churakov G, Sadasivuni MK, Rosenbloom KR, Huchon D, Brosius J, et al. (2010) Rodent evolution: back to the root. Molecular biology and evolution 27: 1315-1326.
9. Gilbert N, Lutz S, Morrish TA, Moran JV (2005) Multiple fates of L1 retrotransposition intermediates in cultured human cells. Molecular and cellular biology 25: 7780-7795.
10. Gogvadze E, Barbisan C, Lebrun MH, Buzdin A (2007) Tripartite chimeric pseudogene from the genome of rice blast fungus Magnaporthe grisea suggests double template jumps during long interspersed nuclear element (LINE) reverse transcription. BMC genomics 8: 360.
11. Hasnaoui M, Doucet AJ, Meziane O, Gilbert N (2009) Ancient repeat sequence derived from U6 snRNA in primate genomes. Gene 448: 139-144.
12. Kriegs JO, Zemann A, Churakov G, Matzke A, Ohme M, et al. (2010) Retroposon insertions provide insights into deep lagomorph evolution. Molecular biology and evolution 27: 2678-2681.

13. Nilsson MA, Churakov G, Sommer M, Tran NV, Zemann A, et al. (2010) Tracking marsupial evolution using archaic genomic retroposon insertions. PLoS biology 8: e1000436.
14. Suh A, Paus M, Kiefmann M, Churakov G, Franke FA, et al. (2011) Mesozoic retroposons reveal parrots as the closest living relatives of passerine birds. Nature communications 2: 443.
15. Kramerov DA, Vassetzky NS (2005) Short retroposons in eukaryotic genomes. International review of cytology 247: 165-221.
16. Kriegs JO, Churakov G, Jurka J, Brosius J, Schmitz J (2007) Evolutionary history of 7SL RNA-derived SINEs in Supraprimates. Trends in genetics : TIG 23: 158-161.
17. Ohshima K, Okada N (2005) SINEs and LINEs: symbionts of eukaryotic genomes with a common tail. Cytogenetic and genome research 110: 475-490.
18. Piskurek O, Jackson DJ (2012) Transposable Elements: From DNA Parasites to Architects of Metazoan Evolution. Genes 3: 409-422.
19. Gilbert N, Labuda D (2000) Evolutionary inventions and continuity of CORE-SINEs in mammals. Journal of Molecular Biology 298: 365-377.
20. Szafranski K, Dingermann T, Glockner G, Winckler T (2004) Template jumping by a LINE reverse transcriptase has created a SINE-like 5S rRNA retropseudogene in Dictyostelium. Molecular genetics and genomics : MGG 271: 98-102.
21. Ichiyanagi K, Nakajima R, Kajikawa M, Okada N (2007) Novel retrotransposon analysis reveals multiple mobility pathways dictated by hosts. Genome research 17: 33-41.
22. Lerat E (2010) Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. Heredity (Edinb) 104: 520-533.
23. Saha S, Bridges S, Magbanua ZV, Peterson DG (2008) Empirical comparison of ab initio repeat finding programs. Nucleic Acids Res 36: 2284-2294.
24. Flutre T, Duprat E, Feuillet C, Quesneville H (2011) Considering transposable element diversification in de novo annotation approaches. PloS one 6: e16526.
25. Edgar RC, Myers EW (2005) PILER: identification and classification of genomic repeats. Bioinformatics 21 Suppl 1: i152-158.
26. Price AL, Jones NC, Pevzner PA (2005) De novo identification of repeat families in large genomes. Bioinformatics 21 Suppl 1: i351-358.
27. Matzke A, Churakov G, Berkes P, Arms EM, Kelsey D, et al. (2012) Retroposon insertion patterns of neoavian birds: strong evidence for an extensive incomplete lineage sorting era. Molecular biology and evolution 29: 1497-1501.
28. Elsik CG, Tellam RL, Worley KC, Gibbs RA, Muzny DM, et al. (2009) The genome sequence of taurine cattle: a window to ruminant biology and evolution. Science 324: 522-528.
29. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921.
30. Wade CM, Giulotto E, Sigurdsson S, Zoli M, Gnerre S, et al. (2009) Genome sequence, comparative analysis, and population genetics of the domestic horse. Science 326: 865-867.
31. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520-562.
32. Terai G, Yoshizawa A, Okida H, Asai K, Mituyama T (2010) Discovery of short pseudogenes derived from messenger RNAs. Nucleic acids research 38: 1163-1171.

33. Szak ST, Pickeral OK, Makalowski W, Boguski MS, Landsman D, et al. (2002) Molecular archeology of L1 insertions in the human genome. Genome Biol 3: research0052.
34. Takai D, Jones PA (2002) Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc Natl Acad Sci U S A 99: 3740-3745.
35. Ostertag EM, Kazazian HH, Jr. (2001) Biology of mammalian L1 retrotransposons. Annual review of genetics 35: 501-538.
36. Babushok DV, Ostertag EM, Courtney CE, Choi JM, Kazazian HH, Jr. (2006) L1 integration in a transgenic mouse model. Genome Res 16: 240-250.
37. Lucier JF, Perreault J, Noel JF, Boire G, Perreault JP (2007) RTAnalyzer: a web application for finding new retrotransposons and detecting L1 retrotransposition signatures. Nucleic Acids Res 35: W269-274.
38. Damert A, Raiz J, Horn AV, Lower J, Wang H, et al. (2009) 5'-Transducing SVA retrotransposon groups spread efficiently throughout the human genome. Genome Res 19: 1992-2008.
39. Cost GJ, Feng Q, Jacquier A, Boeke JD (2002) Human L1 element target-primed reverse transcription in vitro. The EMBO journal 21: 5899-5910.
40. Dewannieux M, Esnault C, Heidmann T (2003) LINE-mediated retrotransposition of marked Alu sequences. Nature genetics 35: 41-48.
41. Lavie L, Maldener E, Brouha B, Meese EU, Mayer J (2004) The human L1 promoter: variable transcription initiation sites and a major impact of upstream flanking sequence on promoter activity. Genome Res 14: 2253-2260.
42. Lee SH, Cho SY, Shannon MF, Fan J, Rangasamy D (2010) The impact of CpG island on defining transcriptional activation of the mouse L1 retrotransposable elements. PLoS One 5: e11353.
43. Yu F, Zingler N, Schumann G, Stratling WH (2001) Methyl-CpG-binding protein 2 represses LINE-1 expression and retrotransposition but not Alu transcription. Nucleic Acids Res 29: 4493-4501.
44. Minakami R, Kurose K, Etoh K, Furuhata Y, Hattori M, et al. (1992) Identification of an internal cis-element essential for the human L1 transcription and a nuclear factor(s) binding to the element. Nucleic Acids Res 20: 3139-3145.
45. Thayer RE, Singer MF, Fanning TG (1993) Undermethylation of specific LINE-1 sequences in human cells producing a LINE-1-encoded protein. Gene 133: 273-277.
46. Mathias SL, Scott AF (1993) Promoter binding proteins of an active human L1 retrotransposon. Biochem Biophys Res Commun 191: 625-632.
47. Buzdin A, Gogvadze E, Lebrun MH (2007) Chimeric retrogenes suggest a role for the nucleolus in LINE amplification. FEBS letters 581: 2877-2882.
48. Kim DS, Hahn Y (2011) Identification of human-specific transcript variants induced by DNA insertions in the human genome. Bioinformatics 27: 14-21.
49. Zhang Z, Harrison PM, Liu Y, Gerstein M (2003) Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. Genome research 13: 2541-2558.
50. Lee JY, Ji Z, Tian B (2008) Phylogenetic analysis of mRNA polyadenylation sites reveals a role of transposable elements in evolution of the 3'-end of genes. Nucleic acids research 36: 5581-5590.
51. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, et al. (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenetic and genome research 110: 462-467.

52. Piskurek O, Nishihara H, Okada N (2009) The evolution of two partner LINE/SINE families and a full-length chromodomain-containing Ty3/Gypsy LTR element in the first reptilian genome of Anolis carolinensis. Gene 441: 111-118.
53. Warren WC, Hillier LW, Graves JAM, Birney E, Ponting CP, et al. (2008) Genome analysis of the platypus reveals unique signatures of evolution (vol 453, pg 175, 2008). Nature 455: 256-256.
54. Lovsin N, Gubensek F, Kordis D (2001) Evolutionary dynamics in a novel L2 clade of non-LTR retrotransposons in deuterostomia. Molecular Biology and Evolution 18: 2213-2224.
55. Smit AFA, Riggs AD (1995) Mirs Are Classic, Transfer-Rna-Derived Sines That Amplified before the Mammalian Radiation. Nucleic Acids Research 23: 98-102.
56. Ivics Z, Hackett PB, Plasterk RH, Izsvak Z (1997) Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. Cell 91: 501-510.
57. Bowen NJ, Jordan IK (2007) Exaptation of protein coding sequences from transposable elements. Genome Dyn 3: 147-162.
58. Gombart AF, Saito T, Koeffler HP (2009) Exaptation of an ancient Alu short interspersed element provides a highly conserved vitamin D-mediated innate immune response in humans and primates. Bmc Genomics 10.
59. Piriyapongsa J, Rutledge MT, Patel S, Borodovsky M, Jordan IK (2007) Evaluating the protein coding potential of exonized transposable element sequences. Biology Direct 2.
60. Bowman RR, Hu WS, Pathak VK (1998) Relative rates of retroviral reverse transcriptase template switching during RNA- and DNA-dependent DNA synthesis. Journal of Virology 72: 5198-5206.
61. Srikanta D, Sen SK, Huang CT, Conlin EM, Rhodes RM, et al. (2009) An alternative pathway for Alu retrotransposition suggests a role in DNA double-strand break repair. Genomics 93: 205-212.
62. Suzuki J, Yamaguchi K, Kajikawa M, Ichiyanagi K, Adachi N, et al. (2009) Genetic Evidence That the Non-Homologous End-Joining Repair Pathway Is Involved in LINE Retrotransposition. Plos Genetics 5.
63. Bantysh OB, Buzdin AA (2009) Novel family of human transposable elements formed due to fusion of the first exon of gene MAST2 with retrotransposon SVA. Biochemistry Biokhimiia 74: 1393-1399.
64. Chen C, Ara T, Gautheret D (2009) Using Alu elements as polyadenylation sites: A case of retroposon exaptation. Molecular biology and evolution 26: 327-334.
65. Hancks DC, Kazazian HH, Jr. (2010) SVA retrotransposons: Evolution and genetic instability. Seminars in cancer biology 20: 234-245.
66. Malik HS, Burke WD, Eickbush TH (1999) The age and evolution of non-LTR retrotransposable elements. Molecular biology and evolution 16: 793-805.
67. Malik HS, Eickbush TH (2001) Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses. Genome research 11: 1187-1197.
68. Wang H, Xing J, Grover D, Hedges DJ, Han K, et al. (2005) SVA elements: a hominid-specific retroposon family. Journal of molecular biology 354: 994-1007.

Figure Captions

Fig. 1. **A description of bipartite and tripartite recombinant repeats, and their classification**.

Fig. 2. Length distributions of classified recombinant repeats. 't' is the classified recombinant repeat's size (kb). 3A) more than 85% of the recombinant repeats in mammalian genomes are less than 1kb. 3B) The majority of recombinant repeats are 300bp-700bp in length.

Fig. 3. Boxplot of inter-repeat distances of classified recombinant repeats

Fig. 4. Identification scheme for recombinant repeats. It allocates into 9 different classes: LINE-SINE, LINE-LTR, LINE-DNA, LTR-SINE, LTR-LINE, LTR-DNA, DNA-SINE, DNA-LINE and DNA-LTR.

Fig. 5. External and Internal TSDs in Tripartite and Bipartite Recombinant repeats

Fig. 6. Analysis of human LINE, AluY and mammalian recombinant repeat perfect TSD length distributions. 6A) full-length human LINE and AluY perfect TSD length distribution, 6B) human, cow and horse recombinant LINE-SINE perfect external TSD length distribution and 6C) human, cow and horse recombinant LINE-SINE perfect internal TSD length distribution. The y-axis is the recombinant repeat fraction (%), and the x-axis shows the perfect TSD pair length.

Fig. 7. Frequency of recombinant repeats generated by the Transposon-into-Transposon (TinT) mechanism or the Template Switching (TS) mechanism.

Fig. 8. How new bipartite or tripartite recombinant repeat can be created through a TinT event and multiple polyadenylation signals.

Table 1. An overview of recombinant repeats from four mammalian genomes: human, mouse, cow and horse.

	Human (*H. sapiens*)	**Mouse (*M. musculus*)**	**Cow (*B. taurus*)**	**Horse (*E. caballus*)**
Total Recombinant Repeats*	43,478	11,425	40,588	22,720
Total Recombinant Repeats (base pairs)	38,600,856	9,977,723	34,979,177	21,689,310
Genome Coverage of Recombinant Repeats (%)	1.25	0.38	1.33	0.92
Total Classified Recombinant Repeats**	10,935	2,867	15,724	6,708
Total Classified recombinant repeat (base pairs)	7,291,148	1,841,245	12,170,532	4,875,244
Genome Coverage of Classified Recombinant Repeats (%)	0.24	0.07	0.46	0.21
Type of Classified Recombinant Repeat				
LINE-SINE	4,879	695	11,803	4,128
LINE-LTR	79	203	322	87
LINE-DNA	103	14	48	133
LTR-SINE	2,943	1,485	2,079	965
LTR-LINE	3	51	0	0
LTR-DNA	93	118	8	26
DNA-SINE	2,546	243	833	1,293
DNA-LINE	5	8	0	0
DNA-LTR	6	8	0	3
SINE-LINE	0	0	0	0
SINE-LTR	0	3	0	0
SINE-DNA	0	0	0	0
Unknown	278	39	631	73
Bipartite Recombinant	341/60	31/7	799/129	119/25

Repeats derived from Tripartite Recombinant repeats (Bipartite/Tripartite)				
Classified Ancestral Recombinant repeats	1048	0	300	1764
Classified Mixed Recombinant repeats	3826	138	5,205	2408
Classified Lineage-specific Recombinant repeats	6,061	2,729	10,219	2,536

* Total Recombinant Repeats are all the sequences found to contain 2 (bipartite) or 3 (tripartite) TE fragments.

** Classified Recombinant Repeats are bipartite and tripartite sequences (≥ 3 copies) that could be clustered and classified into families based on their TE content.

Table 2. Potentially retrotransposed recombinant repeat families with CpG islands or protein coding gene promoters in their 1kbp flanking regions.

	Human (*H. sapiens*)	**Cow (*B. taurus*)**	**Horse (*E. caballus*)**
Recombinant repeats with perfect external TSDs and imperfect internal TSDs	64[a]/56[b] (468[c])	245[a]/188[b] (2,810[c])	75[a]/67[b] (766[c])
Recombinant repeats (with perfect external TSDs and imperfect internal TSDs). Families contained members located in 1kbp flanking region next to protein coding gene promoter or Cpg Island	(26[d]) (50[e]) [34[f]]	(48[d]) (166[e]) [94[f]]	(0[d]) (89[e]) [33[f]]

[a] Number of simple recombinant repeats with perfect external TSDs and imperfect internal TSDs

[b] Number of families of simple recombinant repeats with perfect external TSDs and imperfect internal TSDs mentioned in [a]

[c] All recombinant repeats from families mentioned in [b]

[d] Number of family members located next to protein coding gene promoter in 1kbp flanking region

[e] Number of family members located next to Cpg island in 1kbp flanking region

[f] Number of families contained members that located next to protein coding gene promoter or Cpg island in 1kbp flanking region

Supporting Information Legends

Supplementary File 1. Genomic location of simple recombinant repeat in different mammals

Supplementary Information 1. The overview of software, custom PERL script and guide of simple recombinant repeat identification pipelines.

Figure 1:

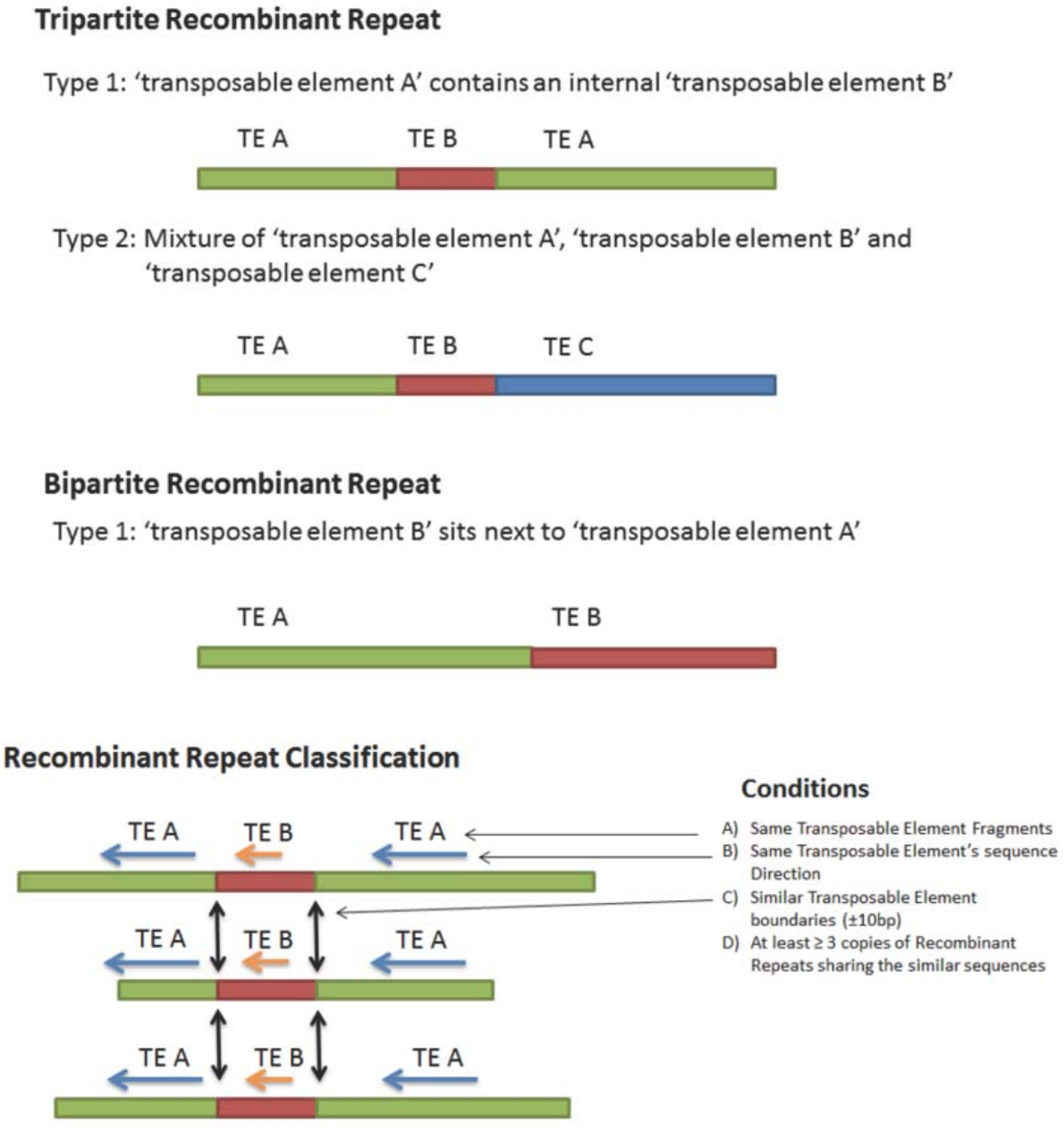

Tripartite Recombinant Repeat
Type 1: 'transposable element A' contains an internal 'transposable element B'
TE A
TE B
TE A
Type 2: Mixture of 'transposable element A', 'transposable element B' and 'transposable element C'
TE A
TE B
TE C
Bipartite Recombinant Repeat
Type 1: 'transposable element B' sits next to 'transposable element A'
TE A
TE B
Recombinant Repeat Classification
Conditions
A) Same Transposable Element Fragments
B) Same Transposable Element's sequence Direction
C) Similar Transposable Element boundaries (±10bp)
D) At least ≥ 3 copies of Recombinant Repeats sharing the similar sequences
TE A
TE B
TE A
TE A
TE B
TE A
TE A
TE B
TE A

Figure 2:

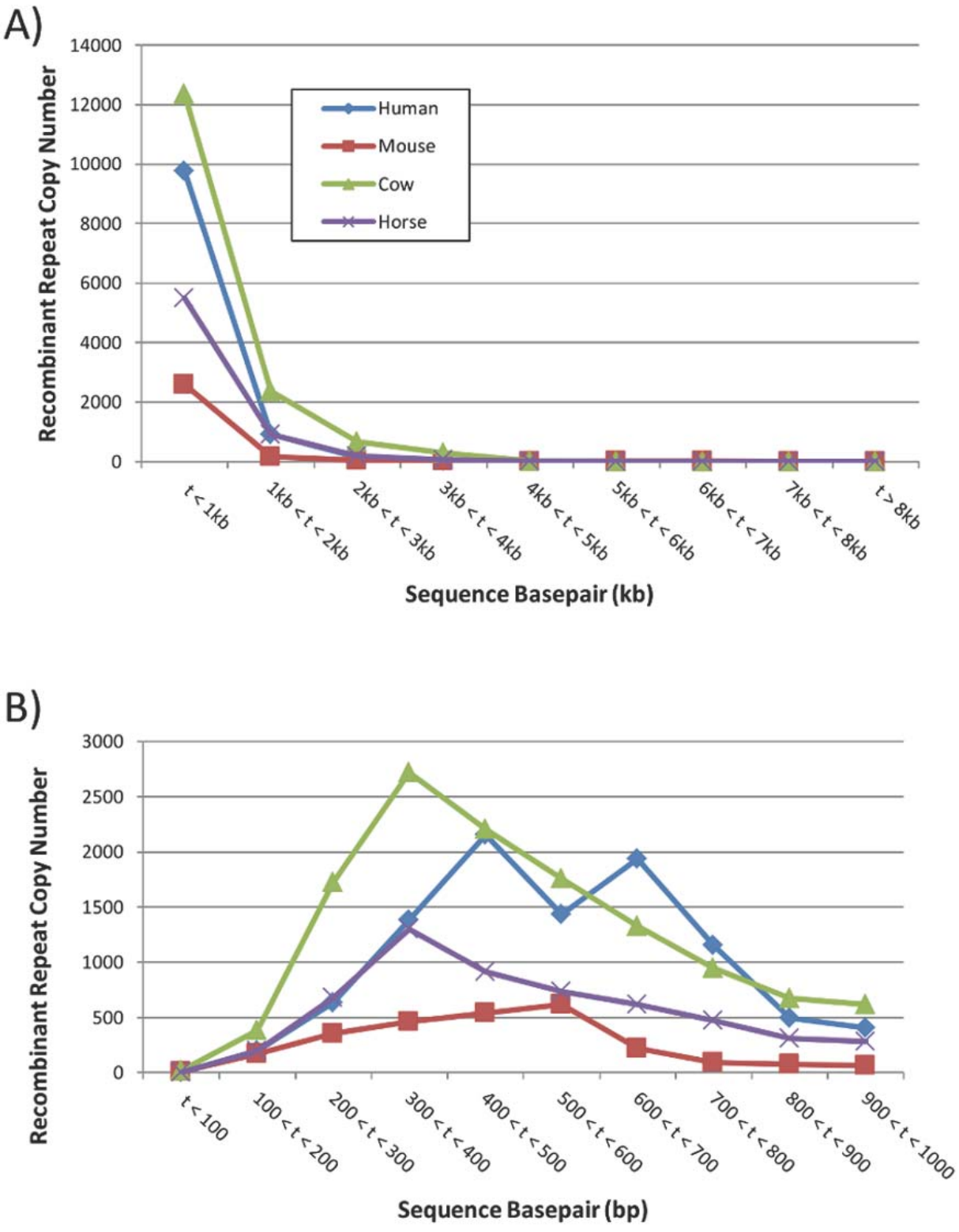

A)
Recombinant Repeat Copy Number
14000
12000
10000
8000
6000
4000
2000
0
Human
Mouse
Cow
Horse
t < 1kb
1kb < t < 2kb
2kb < t < 3kb
3kb < t < 4kb
4kb < t < 5kb
5kb < t < 6kb
6kb < t < 7kb
7kb < t < 8kb
t > 8kb
Sequence Basepair (kb)
B)
Recombinant Repeat Copy Number
3000
2500
2000
1500
1000
500
0
t < 100
100 < t < 200
200 < t < 300
300 < t < 400
400 < t < 500
500 < t < 600
600 < t < 700
700 < t < 800
800 < t < 900
900 < t < 1000
Sequence Basepair (bp)

Figure 3:

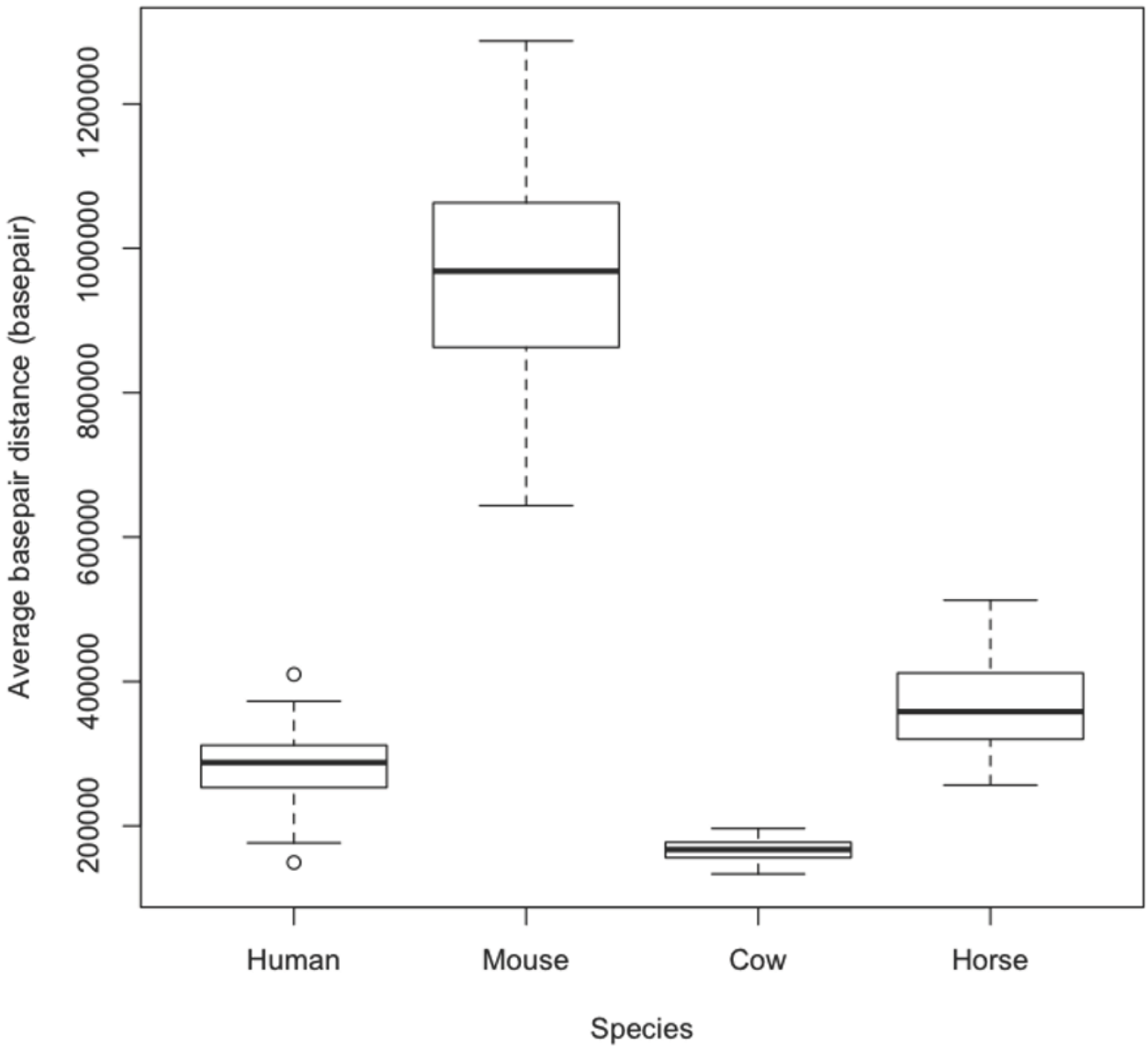

Average basepair distance (basepair)
200000
400000
600000
800000
1000000
1200000
Human
Mouse
Cow
Horse
Species

Figure 4:


Recombinant Repeat Type
Tripartite
Bipartite
LINE-SINE
LINE A SINE LINE B
LINE SINE
LINE-LTR
LINE A LTR LINE B
LINE LTR
LINE-DNA
LINE A DNA LINE B
LINE DNA
LTR-SINE
LTR A SINE LTR B
LTR SINE
LTR-LINE
LTR A LINE LTR B
Non-Available
LTR-DNA
LTR A DNA LTR B
LTR DNA
DNA-SINE
DNA A SINE DNA B
LTR SINE
DNA-LINE
DNA A LINE DNA B
Non-Available
DNA-LTR
DNA A LTR DNA B
Non-Available
SINE-LINE
SINE A LINE SINE B
Non-Available
SINE-LTR
SINE A LTR SINE B
Non-Available
SINE-DNA
SINE A DNA SINE B
Non-Available

Figure 5:

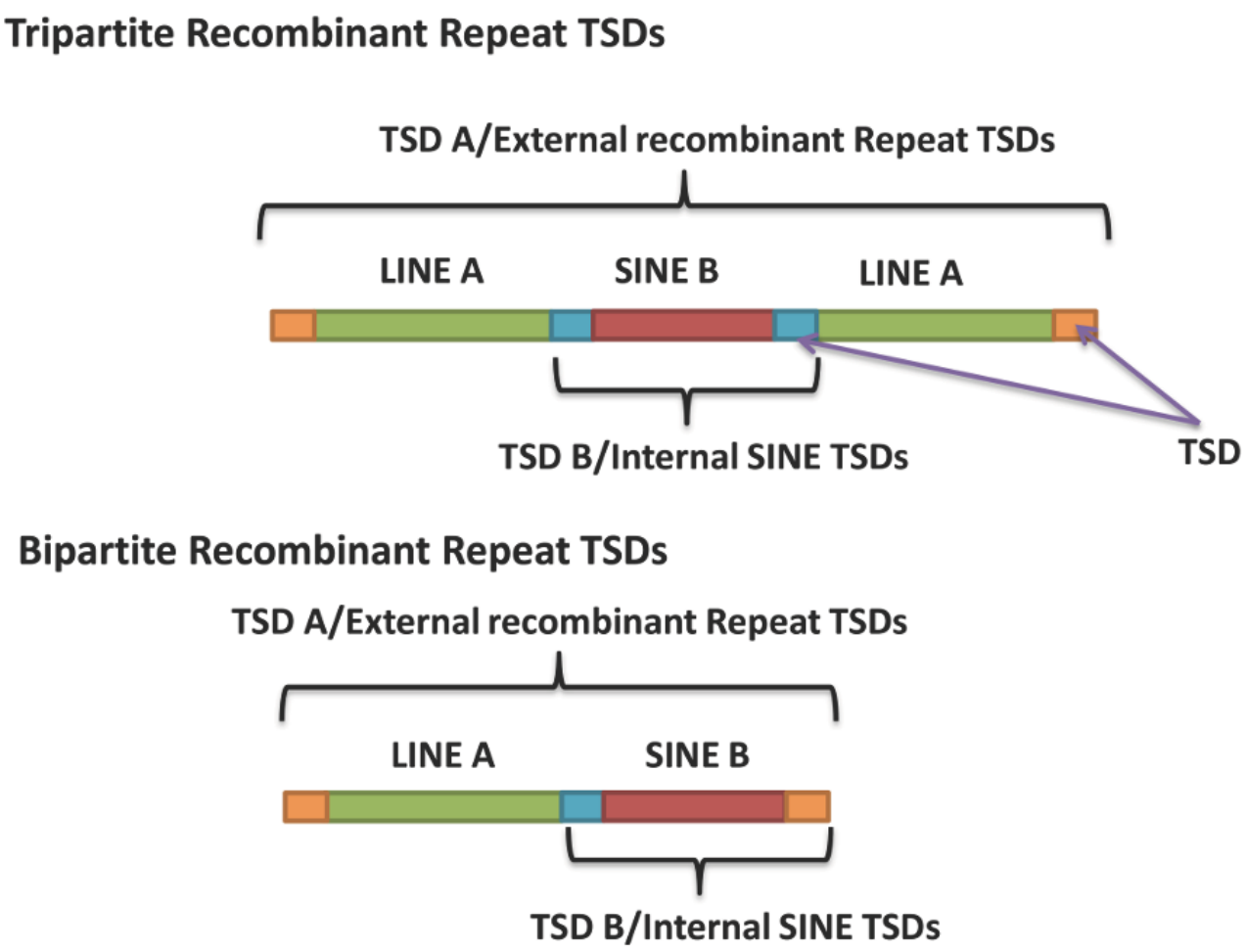
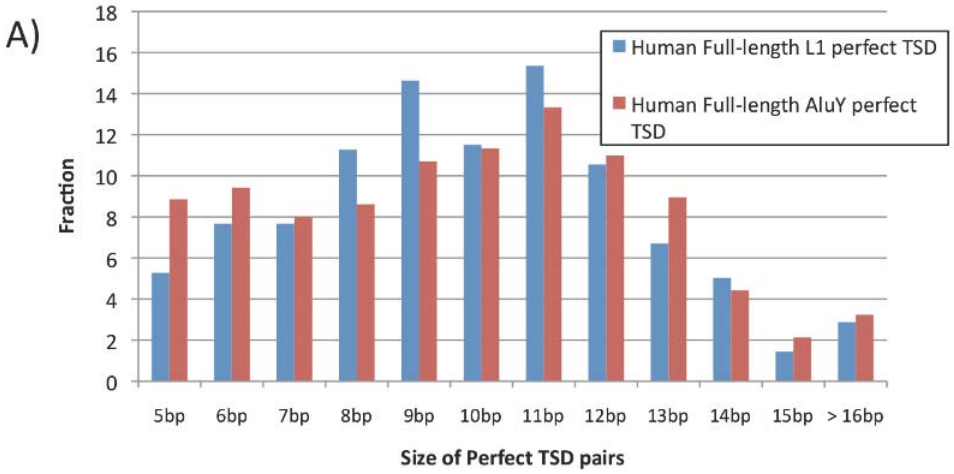
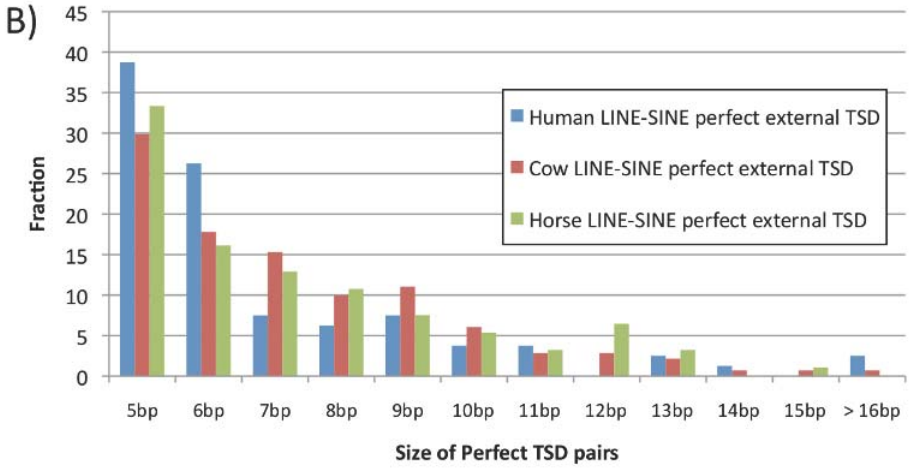

Tripartite Recombinant Repeat TSDs
TSD A/External recombinant Repeat TSDs
LINE A
SINE B
LINE A
TSD B/Internal SINE TSDs
TSD
Bipartite Recombinant Repeat TSDs
TSD A/External recombinant Repeat TSDs
LINE A
SINE B
TSD B/Internal SINE TSDs

Figure 6:


A)
Fraction
18
16
14
12
10
8
6
4
2
0
5bp
6bp
7bp
8bp
9bp
10bp
11bp
12bp
13bp
14bp
15bp
> 16bp
Size of Perfect TSD pairs
Human Full-length L1 perfect TSD
Human Full-length AluY perfect TSD


B)
Fraction
45
40
35
30
25
20
15
10
5
0
5bp
6bp
7bp
8bp
9bp
10bp
11bp
12bp
13bp
14bp
15bp
> 16bp
Size of Perfect TSD pairs
Human LINE-SINE perfect external TSD
Cow LINE-SINE perfect external TSD
Horse LINE-SINE perfect external TSD

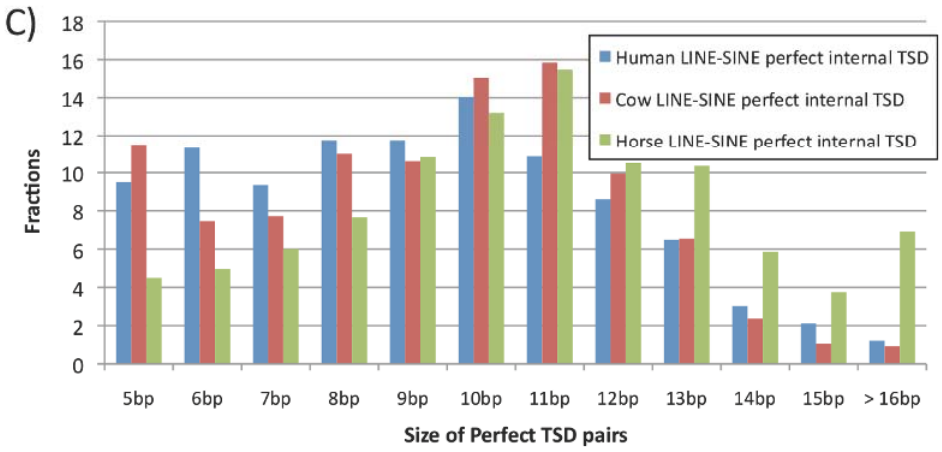

C)
Fractions
18
16
14
12
10
8
6
4
2
0
5bp
6bp
7bp
8bp
9bp
10bp
11bp
12bp
13bp
14bp
15bp
> 16bp
Size of Perfect TSD pairs
Human LINE-SINE perfect internal TSD
Cow LINE-SINE perfect internal TSD
Horse LINE-SINE perfect internal TSD

Figure 7:

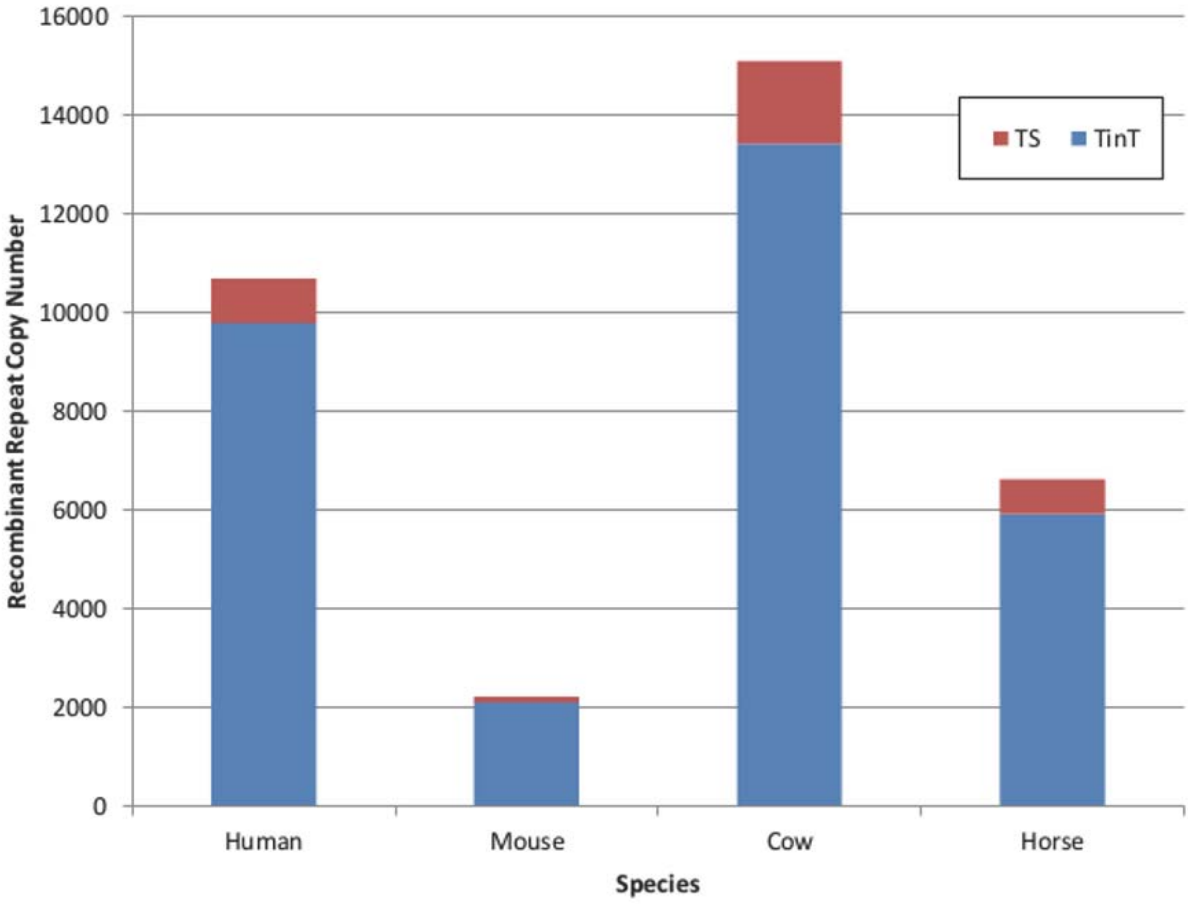
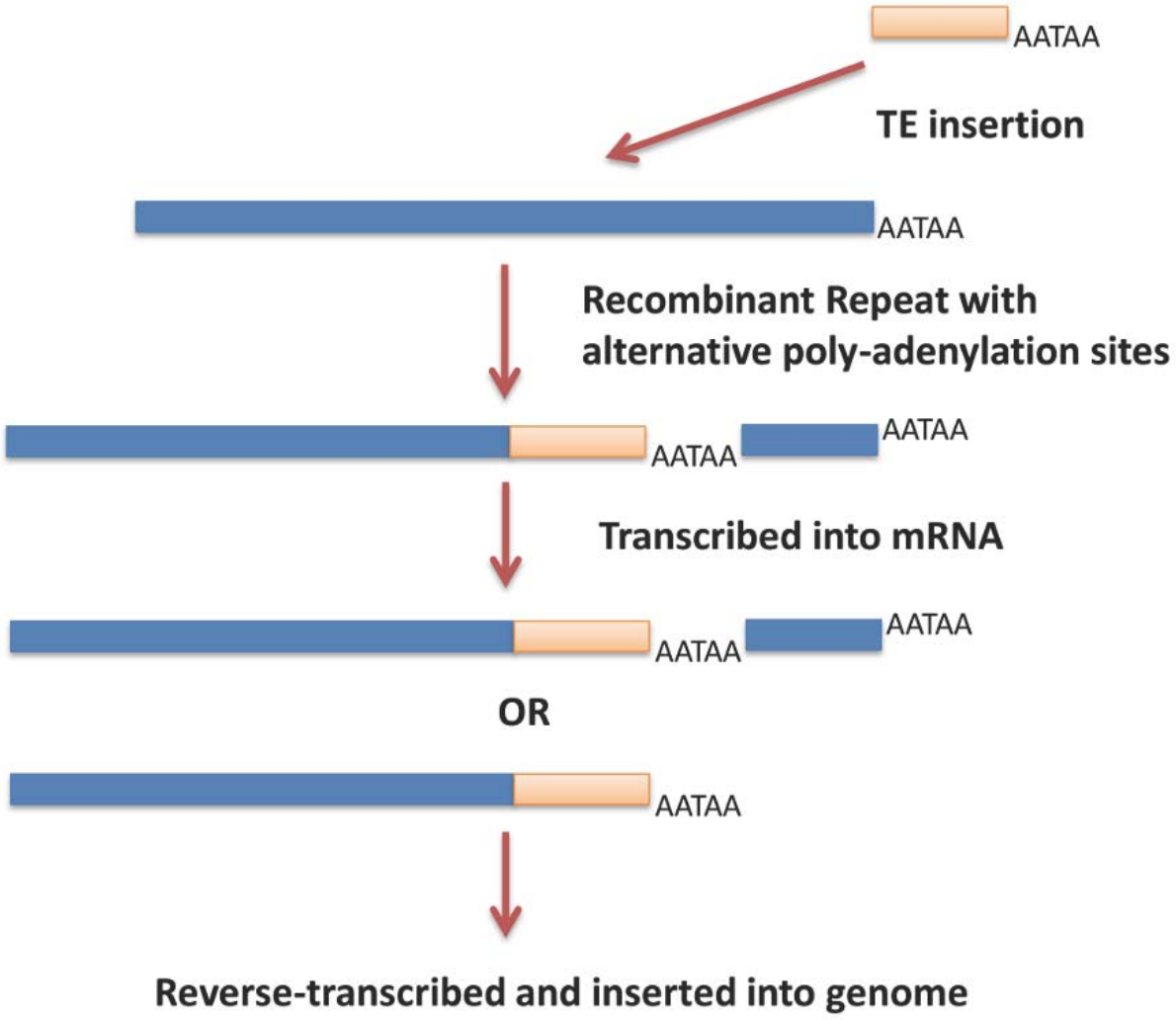

16000
14000
12000
10000
8000
6000
4000
2000
0
Recombinant Repeat Copy Number
TS
TinT
Human
Mouse
Cow
Horse
Species

Figure 8:


AATAA
TE insertion
AATAA
Recombinant Repeat with
alternative poly-adenylation sites
AATAA
AATAA
Transcribed into mRNA
AATAA
AATAA
OR
AATAA
Reverse-transcribed and inserted into genome

Statement of Authorship

Title of Paper	
Publication Status	○ Published, ○ Accepted for Publication, ○ Submitted for Publication, ○ Publication style
Publication Details	

Author Contributions

By signing the Statement of Authorship, each author certifies that their stated contribution to the publication is accurate and that permission is granted for the publication to be included in the candidate's thesis.

Name of Principal Author (Candidate)			
Contribution to the Paper			
Signature		Date	

Name of Co-Author			
Contribution to the Paper			
Signature		Date	

Name of Co-Author			
Contribution to the Paper			
Signature		Date	

Name of Co-Author			
Contribution to the Paper			
Signature		Date	

Chapter 3

Segmental Duplication Events Can Be Detected Using Repetitive Elements.

Sim Lin Lim, R. Daniel Kortschak and David L. Adelson

School of Molecular and Biomedical Science
The University of Adelaide
North Terrace
Adelaide, 5005
South Australia
Australia

RESEARCH ARTICLE

Segmental Duplication Events Can Be Detected Using Repetitive Elements.

---

[1]Sim L. Lim, R. [1]Daniel Kortschak [1,2]David L. Adelson[*]

[1]School of Molecular and Biomedical Science

[2]Centre for Molecular Pathology

The University of Adelaide

North Terrace

Adelaide, 5005

South Australia

Australia

***Corresponding Author: email: david.adelson@adelaide.edu.au,**

Tel: +61 (0)8 8313 7555,

Fax: +61 (0) 8313 4362

Abstract

Mammalian genomes contain many repetitive elements (~40% of genomic sequence) largely originating from retrotransposons. Repetitive elements not only drive exaptation, speciation, and genome evolution, they contribute significantly to structural variation in individuals. Transposable elements can insert into other repetitive elements to create chimeric, or recombinant repeats. Recombinant repeats have been used to study genome evolution and composition. Recombinant repeats are also the major source of novel transposable elements, but previous work has focused exclusively on simple recombinant repeats in eukaryotic genomes. In this report, we modified *de novo* repeat identification methods to identify complex recombinant repeat families in four mammalian genomes (human, mouse, cow and horse). Most complex recombinant repeats are present as singletons and account for 18%~25% of a typical mammalian genome. Complex recombinant repeat families were chromosome-specific and overlap significantly with segmental duplications, so were probably generated through duplication events. Complex recombinant repeats are therefore inactive sequences that do not retrotranspose. Because segmental duplication annotation currently depends on repeat free genome alignments, complex recombinant repeats provide an additional tool for identifying segmental duplications made up of mostly repetitive sequences. Our comparative analysis has shown that complex recombinants repeats are useful for estimating the rate of segmental duplication formation in primate genomes.

Introduction

Transposable elements (TE) are mobile sequences that are copied or move from one location to another within a genome or, occasionally from one genome to another [1,2]. They are the largest class of DNA sequences in many genomes, accounting for up to 90% of some plant genomes and ~40-50% of mammalian genomes [3]. Transposons were first discovered by Barbara McClintock, who observed unusual colour patterns in maize, that she concluded were caused by mobile DNA [4,5]. Transposable elements can be divided into two fundamental classes based on their transposition mechanism: DNA transposons that rely on a "cut and paste" mechanism, or retrotransposons (LTR, SINE and LINE) that use a "copy and paste" mechanism via an RNA intermediate [6]. In spite of Susumu Ohno's labeling of repeats as "junk DNA" [7], transposable elements play an important role in creating genomic structural variation, regulating gene expression, creating mRNA alternative splicing and more [8-12].

Recent studies have shown that one TE can combine with another TE to create recombinant repeats [13-15]. Recombinant Repeats can be divided into two classes: simple and complex. Simple recombinant repeats are composed of 2 (bipartite) or 3 (tripartite) transposable elements, and complex recombinant repeats are created by fusion of more than three transposable elements [16-18]. Although recombinant repeats have until recently remained largely un-characterized, they serve as useful genetic markers for resolving phylogenetic trees containing highly similar species [19,20]. Simple recombinant repeats have also been used to explain how novel transposable elements can arise. There are two models used to explain novel SINEs and simple recombinant repeats formation; either via a template switching mechanism (TS) or a direct transposon-into-transposon (TinT) insertion

event [13,19,21-23]. Based on our analyses (S.L. Lim and D.L. Adelson unpublished), the TinT mechanism of simple recombinant repeat formation predominates.

Although simple recombinant repeats have been discovered and studied in mammals and plants [13,17,20,24], our knowledge of complex recombinant repeats is limited. There is no evidence for replication of complex recombinant repeats (CRR) based on RNA intermediates either *in-vitro* or *in-vivo*. They have only been described bioinformatic analyses of genomic sequence [14]. Characterisation of complex recombinant repeats with current TE annotation software is very difficult [25-27]. There is therefore a need for methods to identify and characterise complex recombinant repeats in order to describe their impact on complex genomes and their likely mode of origin.

In this report, we describe a pipeline that we have used to identify and classify complex recombinant repeats into families, from human (*Homo sapiens*), mouse (*Mus musculus*), cow (*Bos taurus*) and horse (*Equus callabus*). Our methods have allowed us to systematically study recombinant repeats, resulting in the discovery of complex recombinant repeat families, some of which are species-specific. We have also shown that most complex recombinant repeats are very likely generated via segmental duplication (SD) events. Finally, our comparative genomic analysis shows that the CRR family formation rate vary across primates, indicating the CRR families are useful for estimating SD formation in primate genomes.

Materials and Methods

Database

The data used in this research were downloaded from public databases. The genome assemblies of human (hg19), mouse (mm9), cow (BosTau 4.0) and horse

(EqCab 2.0) were downloaded from the UCSC Genome Browser (http://genome.ucsc.edu/) [28-31]. The mammalian haplotype sequences was discarded in this analysis. Species-specific repetitive element sequences (as of February 2011) were downloaded from Repbase (http://www.girinst.org/repbase/) [32].

Software for complex recombinant repeat identification and classification

The tools used in this pipeline, perl, krishna and PILER, RepeatMasker, LASTZ and BEDTools are open source software running as stand-alone applications in a Linux environment (Supplementary Information S1).

Complex Recombinant Repeat Identification Pipeline

The overview of the pipeline is shown in Supplementary Information S1. Complex recombinant repeat characteristics and how they were identified are shown in Figure 1. Mammalian genomes were masked with RepeatMasker using species-specific libraries from Repbase. We searched for interspersed type of complex recombinant repeats that contained sequences from four or more repeat types. Complex recombinant repeats that met the criteria were extracted for further classification.

Complex Recombinant Repeat Classification Pipeline

De novo repeat identification pipelines (krishna and PILER) [33] were used to classify the complex recombinant repeat families from each group as shown in Figure 2. First, krishna was used to align full-length CRR based on ≥ 70% sequence identity. If the CRR sequence was: a) overlapped with another CRR sequence by ≥ 1bp, and b) contained ‘gap’ or undetermined sequence ‘N’, the sequence was discarded. Subsequently, PILER was used to cluster aligned CRR into ‘families’. The families with ≥ 3 CRR copies were extracted for further analysis.

Identifying Human Complex Recombinant Repeats Generated Via Segmental Duplications

We used LASTZ to compare the sequence similarity in each CRR family. We downloaded the hg19 segmental duplication data from the Eichler lab (as of August 2011) (http://humanparalogy.gs.washington.edu/) [34,35] and used BEDTools to compare the human segmental duplication intervals with those of our human complex recombinant repeats.

Comparison Of Human Complex Recombinant Repeats With UCSC Refseq Genes

The mammalian RefSeq genes were downloaded from the UCSC genome repository and segmentally duplicated gene families from http://dgd.genouest.org/ [36]. We used BEDTools to intersect the recombinant repeat genomic intervals with the segmentally duplicated gene family intervals, and with RefSeq genes' 5'UTR, 3' UTR, Exonic and Intronic intervals in the relevant assembly.

Comparative Analysis Of Human Complex Recombinant Repeats With Other Primate Genomes

PILER was used to generate consensus sequences from each complex recombinant repeat family. We used BLAT (http://genome.ucsc.edu/) to find human complex recombinant repeats in marmoset (calJac3) (http://genome.wustl.edu/genomes/detail/callithrix-jacchus/ and https://www.hgsc.bcm.edu/content/marmoset-genome-project) , rhesus monkey (rheMac3) [37], orangutan (ponAbe2) [38] and chimpanzee (panTro4) [39].

Results

Characterisation of Complex Recombinant Repeats

We identified 1,218,613 complex recombinant repeats from four mammalian genomes (human, mouse, cow and horse), as described in the material and methods section (Figure 2). Of these only 1596 (Table S1) were present at a copy number of three or more and were subsequently classified into 389 families, leaving 1,217,019 recombinant repeat singletons that could not be clustered. Most families (235/60.4%) contained only 3 copies and families with 6 or more repeats numbered 52 (Figure 3). Most classified recombinant repeats were longer than 1kbp (1,287 copy numbers) and in cow and horse none was longer than 3kbp (Figure 4). It is worth noting that mouse differed from the other mammals in that it had the longest recombinant repeat, over 8kbp long, with 44 copies observed.

Complex recombinant repeat copy number and genome coverage varied significantly across the four mammal species used, and was highest in the mouse and human, compared to horse and cow (Table 1). Complex recombinant repeat copy numbers in human and mouse were not correlated with chromosome size and copy numbers in horse and cow were too low to address this question (Figure S1). Human recombinant repeat families were found either restricted to a single chromosome (intra-chromosomal) or on multiple chromosomes (inter-chromosomal). This was in contrast to other mammals, where families were primarily intra-chromosomal (Figure 5). The complex recombinant repeat family distributions in horse could not be classified as interchromosomal or intrachromosomal as they were almost exclusively found in unplaced contigs (ChrUn).

Composition and Origin of Complex Recombinant Repeats

Complex recombinant repeats were annotated using RepeatMasker, allowing us to determine repeat fragment type and length. The complexity of the annotation made it impractical to divide the recombinant repeats into types such as LINE-SINE

or LTR-LINE as we have done previously (S.L. Lim and D.L. Adelson Unpublished). In human, horse, and cow, non-LTR TE were a major component of complex recombinant repeats, followed by LTR and DNA TEs as shown in Table 1. However, in mouse complex recombinant repeats were primarily composed of LTR, followed by non-LTR and DNA TEs (Table 1). Further analysis showed that SINE did not proportionally contribute as much to complex recombinant repeats in mouse (3.81%) as in human (22.58%). In spite of the complexity of their annotation, we were able to partition complex recombinant repeats into 2 classes: a) "Lineage-specific" recombinant repeats (1,052 copies) that made up by clade-specific TEs (TE that appeared in specific species clades after mammalian divergence event), and b) "Mixed" recombinant repeats (543 copies) made up of a mixture of clade-specific and ancestral TEs (TE that appeared before mammalian divergence events, e.g. L2 and MIR) (Table 1). We did not find any complex recombinant repeats made up exclusively from ancestral TEs.

Complex Recombinant Repeat are generated by segmental duplication events

Our analyses suggested that most of the complex recombinant repeat families were distributed in chromosome-specific patterns, and our LASTZ analysis indicated that 1385 sequences (99.8%) in four mammalians shared ≥ 90% of sequence identity and sequence coverage, suggesting that they were created via duplication events. We therefore compared the genomic locations of human complex recombinant repeats with the hg19 segmental duplication map [34]. Our data showed that ~83.19% (787 copies) of the human complex recombinant repeats were co-located with segmental duplications, indicating that duplications drive the formation of complex recombinant repeat families (Table S2). We also investigated the complex recombinant repeats that did not overlap with human segmental

duplications, and found that all of these sequences were derived from: intra-chromosomal duplicated sequences found only on chromosome X (4 families, 12 copies); and recombinant repeat families that were located partially within reported segmental duplications (131 families, 147 copies). The comparison of human segmental duplicated gene coordinates [36] and CRR coordinates showed that 6 copies of CRRs were believed to be segmental duplicated (Table S1), but they were not exist in hg19 segmental duplication (Example shown in Figure S2). It is likely that all of our complex repeat families lie in segmental duplications, but that some of these duplication events have not yet been reported or annotated. In the mouse, complex recombinant repeat families were present only as intra-chromosomal elements, also indicating that they were most likely created via segmental duplication events. However, in the mouse mm9 assembly, segmental duplications are not annotated so we were unable to formally confirm this.

Complex Recombinant Repeat Exaptation Events.

It is well known that families of protein-coding genes can be created via segmental duplication. We therefore decided to identify exaptation events of complex recombinant repeats within RefSeq genes and specifically within segmentally duplicated protein-coding genes. We compared the complex recombinant repeat genomic intervals with mammalian UCSC RefSeq gene (http://genome.ucsc.edu/) genomic intervals and found 20.3% (324 copies) of recombinant repeats mapped to intronic regions, 1.32% (21 copies) exapted in 5' and 3' UTR regions and 0.69% (11 copies) exapted in protein coding exons (Table 1 and Table S2). However, only small sub-sequences from the complex recombinant repeats were exapted into UTR and exonic regions. With respect to human duplicated gene families we found exaptation events consistent with exaptation either before gene duplication or after

initial gene duplication, resulting in either fully familial (1 family) or partially familial exaptation (3 families) of complex repeat families. We found only partial familial exaptation of complex recombinant repeats in mouse (1 family). We were unable to find any exaptation events in horse and cow, as the CRR sequences were present in intronic regions only (Table 1).

Complex Recombinant Repeat Families In Primate Phylogeny

Primate genome architectures have undergone significant changes since the divergence of New and Old World Monkeys. Rates for substitution vary across primates, as do those for structural variation, segmental duplication and retrotransposon activity [40-47]. However, rates of complex recombinant repeat formation and turn over in primates are unknown [41,46]. We hypothesized that if a complex recombinant repeat was created before primate divergence, it should be found in all primate species. While if the complex repeat was created more recently, it would only appear in a group of closely related primates or a single species. We compared human complex recombinant repeat families with four primates: one new world monkey (marmoset) and three old world monkeys (Chimpanzee, Orangutan, Rhesus Monkey). We used human complex recombinant repeat family consensus sequences to find possible homologues in these primate genomes. Our data revealed that 82 recombinant repeat families remained conserved across the primates, 85 families appeared after marmoset divergence, 74 families appeared after rhesus monkey divergence, and 8 families appeared after orangutan divergence (Figure 6). We found that there were only 3 complex recombinant repeat families exclusive to humans and none of them appear to have been exapted by protein coding genes. This shows that the rate of CRR family formation has

decreased over time during the divergence of primates, with a currently much lower rate of CRR formation compared to 44mya.

4. Discussion

We have identified CRR in all mammalian genomes we have examined and found that they account for a large portion of those genomes (Table 1). However, only a small fraction of these can be classified into families of similar sequences. Many of these orphan CRR comprise more than three repetitive element sequences. For human, cow and horse the classified families are of one dominant type, made from LINE and SINE sequences, reflecting the dominant retrotransposon types in those genomes. However in the mouse, the dominant CRR family is comprised of LINE and LTR sequences. This probably reflects the greater abundance and activity of LTR repeats in the mouse genome [48-50].

The complex structure of the majority of CRR sequences, reflected in the existence of many singleton CRR sequences, may provide a clue as to how they are formed. For most CRR the individual repeat fragments that they consist of are of random size and orientation, and are often interspersed with short, non-repetitive sequences. Previous workers have described nested repetitive elements or composite elements resulting from the insertion of retrotransposon into other retrotransposon (Transposon into Transposon (TinT) [14,15,19]. This nesting or TinT behavior can only be reliably identified for a small fraction (1.25%, S.L. Lim and D.L. Adelson unpublished data) of the total interspersed repeat content of the human genome. Therefore a significant proportion of the human genome repetitive content consists of unique CRR sequences. If TinT can only account for a small fraction of this material we must consider alternative mechanisms for how these CRR are formed. It has been previously observed in plants that insertion of plastid genome

sequences into the nuclear genome, results in complex arrays of plastid, repetitive and other sequences that appear to be the result of double strand break repair [51]. It is therefore plausible that the singleton CRR sequences we have observed in mammals are primarily the result of double strand break repair events that create random patchworks of sequences that reflect the dominant proportion of repetitive element in the genome.

However, a small proportion of the CRR can be classified into families of similar sequences, and no mechanism has been proposed to explain how some CRR multiply to form families. For typical, single retrotransposon sequences increases in copy number are the result of retrotransposition, but this requires specific sequences such as an internal promoter and a 3' reverse transcriptase recognition site [16,52-54]. Because the CRR families consist of multiple fragments of retrotransposon in random orientations, it is unlikely that they can fulfill the requirements for SINE-like non-autonomous retrotransposition. In spite of these limitations, some CRR are present in multiple copies (Figure 3 and 4), and 99.8% of them (1385 copies) are nearly identical (≥ 90% similarity and coverage). If these sequences cannot duplicate through retrotransposition, they must be copied through another mechanism. We have observed that the majority of CRR families overlap with SD in human, mouse and cow, implicating SD as the mechanism for CRR family formation. However, in the horse most CRR are found in unplaced contigs, making it difficult to determine their location with respect to SD. In the cow, most CRR families are present in intrachromosomal SD and at much lower copy number than in human or mouse. We believe these observed differences in CRR location probably arise because of the difference between finished (human and mouse) vs draft (cow and horse) genome assemblies and are unlikely to reflect real species differences [44].

While our CRR families show significant overlap with SD, not all of our CRR family sequences overlap with SD as shown in Figure S2. A likely reason for this incomplete overlap of CRR families with SD lies in the methodology used to identify SD. The standard approach to finding SD relies on comprehensively masking repeats and then looking for highly conserved local sequence alignments greater than 1kbp in length [35,45]. Because repeats are excluded from the standard method for SD detection, SD comprised primarily of repeats will not be found. In contrast, our CRR families are based exclusively on the presence of repeats so can provide additional sequences to improve the existing SD annotation.

The rate of SD formation can be studied in primates, where draft genome assemblies are available for a number of relatively recently diverged species [45]. If CRR family formation depends on SD, the rate of family formation over time should provide an estimate of the rate of novel SD formation. Our estimates of CRR family formation (shown in Figure 6) indicate that the rate of CRR duplication has dropped after the divergence of New World and Old World monkeys and has continued to decrease until the present day. Previous work on the rate of SD formation in primates has yielded similar results in that the rate of SD formation is lower in recently diverged species [47]. Our results are comparable to these estimates and support the idea that CRR families can provide additional information to the inventory of SD carried out by standard means.

5. Conclusion

A significant proportion of mammalian genomes are made up of complex arrays of repetitive element fragments, as opposed to individual repetitive elements separated by non-repetitive sequence. When individual CRR are present in a region that undergoes SD, they become CRR families. These CRR families can provide

additional, unique information on the occurrence of SD that is not available through other means.

References

1. Walsh AM, Kortschak RD, Gardner MG, Bertozzi T, Adelson DL (2013) Widespread horizontal transfer of retrotransposons. Proc Natl Acad Sci U S A 110: 1012-1016.
2. Clark JB, Altheide TK, Schlosser MJ, Kidwell MG (1995) Molecular evolution of P transposable elements in the genus Drosophila. I. The saltans and willistoni species groups. Mol Biol Evol 12: 902-913.
3. Gogvadze E, Buzdin A (2009) Retroelements and their impact on genome evolution and functioning. Cellular and molecular life sciences : CMLS 66: 3727-3742.
4. McClintock B (1956) Controlling elements and the gene. Cold Spring Harbor symposia on quantitative biology 21: 197-216.
5. Mc CB (1950) The origin and behavior of mutable loci in maize. Proc Natl Acad Sci U S A 36: 344-355.
6. Jurka J, Kapitonov VV, Kohany O, Jurka MV (2007) Repetitive sequences in complex genomes: structure and evolution. Annual review of genomics and human genetics 8: 241-259.
7. Ohno S (1972) So much "junk" DNA in our genome. Brookhaven Symp Biol 23: 366-370.
8. Almeida LM, Amaral ME, Silva IT, Silva WA, Jr., Riggs PK, et al. (2008) Report of a chimeric origin of transposable elements in a bovine-coding gene. Genet Mol Res 7: 107-116.
9. Bowen NJ, Jordan IK (2007) Exaptation of protein coding sequences from transposable elements. Genome Dyn 3: 147-162.
10. Chen C, Ara T, Gautheret D (2009) Using Alu elements as polyadenylation sites: A case of retroposon exaptation. Molecular biology and evolution 26: 327-334.
11. Lee JY, Ji Z, Tian B (2008) Phylogenetic analysis of mRNA polyadenylation sites reveals a role of transposable elements in evolution of the 3'-end of genes. Nucleic acids research 36: 5581-5590.
12. Schumann GG, Gogvadze EV, Osanai-Futahashi M, Kuroki A, Munk C, et al. (2010) Unique functions of repetitive transcriptomes. International review of cell and molecular biology 285: 115-188.
13. Buzdin A, Ustyugova S, Gogvadze E, Vinogradova T, Lebedev Y, et al. (2002) A new family of chimeric retrotranscripts formed by a full copy of U6 small nuclear RNA fused to the 3' terminus of l1. Genomics 80: 402-406.
14. Giordano J, Ge Y, Gelfand Y, Abrusan G, Benson G, et al. (2007) Evolutionary history of mammalian transposons determined by genome-wide defragmentation. PLoS computational biology 3: e137.
15. Kriegs JO, Matzke A, Churakov G, Kuritzin A, Mayr G, et al. (2007) Waves of genomic hitchhikers shed light on the evolution of gamebirds (Aves: Galliformes). BMC evolutionary biology 7: 190.
16. Gilbert N, Lutz S, Morrish TA, Moran JV (2005) Multiple fates of L1 retrotransposition intermediates in cultured human cells. Molecular and cellular biology 25: 7780-7795.
17. Gogvadze E, Barbisan C, Lebrun MH, Buzdin A (2007) Tripartite chimeric pseudogene from the genome of rice blast fungus Magnaporthe grisea suggests double template

jumps during long interspersed nuclear element (LINE) reverse transcription. BMC genomics 8: 360.
18. Hasnaoui M, Doucet AJ, Meziane O, Gilbert N (2009) Ancient repeat sequence derived from U6 snRNA in primate genomes. Gene 448: 139-144.
19. Churakov G, Grundmann N, Kuritzin A, Brosius J, Makalowski W, et al. (2010) A novel web-based TinT application and the chronology of the Primate Alu retroposon activity. BMC evolutionary biology 10: 376.
20. Churakov G, Sadasivuni MK, Rosenbloom KR, Huchon D, Brosius J, et al. (2010) Rodent evolution: back to the root. Molecular biology and evolution 27: 1315-1326.
21. Kramerov DA, Vassetzky NS (2005) Short retroposons in eukaryotic genomes. International review of cytology 247: 165-221.
22. Ohshima K, Okada N (2005) SINEs and LINEs: symbionts of eukaryotic genomes with a common tail. Cytogenetic and genome research 110: 475-490.
23. Kramerov DA, Vassetzky NS (2011) Origin and evolution of SINEs in eukaryotic genomes. Heredity 107: 487-495.
24. Buzdin A, Gogvadze E, Lebrun MH (2007) Chimeric retrogenes suggest a role for the nucleolus in LINE amplification. FEBS letters 581: 2877-2882.
25. Lerat E (2010) Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. Heredity (Edinb) 104: 520-533.
26. Price AL, Jones NC, Pevzner PA (2005) De novo identification of repeat families in large genomes. Bioinformatics 21 Suppl 1: i351-358.
27. Saha S, Bridges S, Magbanua ZV, Peterson DG (2008) Empirical comparison of ab initio repeat finding programs. Nucleic Acids Res 36: 2284-2294.
28. Elsik CG, Tellam RL, Worley KC, Gibbs RA, Muzny DM, et al. (2009) The genome sequence of taurine cattle: a window to ruminant biology and evolution. Science 324: 522-528.
29. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921.
30. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520-562.
31. Wade CM, Giulotto E, Sigurdsson S, Zoli M, Gnerre S, et al. (2009) Genome sequence, comparative analysis, and population genetics of the domestic horse. Science 326: 865-867.
32. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, et al. (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenetic and genome research 110: 462-467.
33. Edgar RC, Myers EW (2005) PILER: identification and classification of genomic repeats. Bioinformatics 21 Suppl 1: i152-158.
34. She X, Jiang Z, Clark RA, Liu G, Cheng Z, et al. (2004) Shotgun sequence assembly and recent segmental duplications within the human genome. Nature 431: 927-930.
35. She XW, Jiang ZX, Clark RL, Liu G, Cheng Z, et al. (2004) Shotgun sequence assembly and recent segmental duplications within the human genome. Nature 431: 927-930.
36. Ouedraogo M, Bettembourg C, Bretaudeau A, Sallou O, Diot C, et al. (2012) The duplicated genes database: identification and functional annotation of co-localised duplicated genes across genomes. PLoS One 7: e50653.

37. Rhesus Macaque Genome S, Analysis C, Gibbs RA, Rogers J, Katze MG, et al. (2007) Evolutionary and biomedical insights from the rhesus macaque genome. Science 316: 222-234.
38. Locke DP, Hillier LW, Warren WC, Worley KC, Nazareth LV, et al. (2011) Comparative and demographic analysis of orang-utan genomes. Nature 469: 529-533.
39. Chimpanzee S, Analysis C (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437: 69-87.
40. Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE (2001) Segmental duplications: organization and impact within the current human genome project assembly. Genome Res 11: 1005-1017.
41. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, et al. (2002) Recent segmental duplications in the human genome. Science 297: 1003-1007.
42. Cheng Z, Ventura M, She X, Khaitovich P, Graves T, et al. (2005) A genome-wide comparison of recent chimpanzee and human segmental duplications. Nature 437: 88-93.
43. Perry GH, Tchinda J, McGrath SD, Zhang J, Picker SR, et al. (2006) Hotspots for copy number variation in chimpanzees and humans. Proc Natl Acad Sci U S A 103: 8006-8011.
44. Marques-Bonet T, Girirajan S, Eichler EE (2009) The origins and impact of primate segmental duplications. Trends Genet 25: 443-454.
45. Marques-Bonet T, Kidd JM, Ventura M, Graves TA, Cheng Z, et al. (2009) A burst of segmental duplications in the genome of the African great ape ancestor. Nature 457: 877-881.
46. Gazave E, Darre F, Morcillo-Suarez C, Petit-Marty N, Carreno A, et al. (2011) Copy number variation analysis in the great apes reveals species-specific patterns of structural variation. Genome Res 21: 1626-1639.
47. Sudmant PH, Huddleston J, Catacchio CR, Malig M, Hillier LW, et al. (2013) Evolution and diversity of copy number variation in the great ape lineage. Genome Res.
48. Baust C, Gagnier L, Baillie GJ, Harris MJ, Juriloff DM, et al. (2003) Structure and expression of mobile ETnII retroelements and their coding-competent MusD relatives in the mouse. J Virol 77: 11448-11458.
49. Maksakova IA, Romanish MT, Gagnier L, Dunn CA, van de Lagemaat LN, et al. (2006) Retroviral elements and their hosts: insertional mutagenesis in the mouse germ line. PLoS Genet 2: e2.
50. Rebollo R, Miceli-Royer K, Zhang Y, Farivar S, Gagnier L, et al. (2012) Epigenetic interplay between mouse endogenous retroviruses and host genes. Genome Biol 13: R89.
51. Lloyd AH, Timmis JN (2011) The origin and characterization of new nuclear genes originating from a cytoplasmic organellar genome. Mol Biol Evol 28: 2019-2028.
52. Cost GJ, Feng Q, Jacquier A, Boeke JD (2002) Human L1 element target-primed reverse transcription in vitro. The EMBO journal 21: 5899-5910.
53. Kazazian HH, Jr. (2004) Mobile elements: drivers of genome evolution. Science 303: 1626-1632.
54. Ostertag EM, Kazazian HH, Jr. (2001) Biology of mammalian L1 retrotransposons. Annual review of genetics 35: 501-538.
55. Hedges SB, Dudley J, Kumar S (2006) TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics 22: 2971-2972.

Figure Captions

Fig. 1. Overview of tandem (A) and vs interspersed (B) types of CRR..

Fig. 2. Overview of complex recombinant repeat identification and clustering process by Krishna and PILER.

Fig. 3. Overview of total complex recombinant repeat copy numbers in each family

Fig. 4. Overview of complex recombinant repeat length distributions. CRR lengths (in kbp) were binned and are displayed on the X axis, with copy number on the Y axis.

Fig. 5. Intra-chromosomal vs inter-chromosomal CRR family distribution in mammals. Green bar denotes unplaced contigs that cannot provide chromosomal location.

Fig. 6. Rate of appearance of recombinant repeat families in primates. Recombinant repeat families depicted on the primate phylogenetic tree [55]. The rate of novel CRR family formation was calculated based on the number of new CRR families divided by the time (mya) since the previous divergence event.

Table 1. The Overview classified complex recombinant repeats' statistical analysis in four different mammalian genomes

	Human (*H. sapiens*)	**Mouse (*M. musculus*)**	**Cow (*B. taurus*)**	**Horse (*E. caballus*)**
Total Complex Recombinant Repeat*	389,537	265,321	293,993	269,762
Total Complex Recombinant Repeats (base pairs)	804,966,916	494,851,557	677,661,544	614,201,737
Genome Coverage of Complex Recombinant Repeats (%)	26	18.64	25.72	25.95
Total Classified Complex Recombinant Repeats**	946	483	62	105
Total Classified Complex Recombinant Repeat Families	252	94	18	26
Total Classified Complex recombinant repeat (base pairs)	1,931,112	1,585,294	98,514	173,335
Genome Coverage of Classified Complex Recombinant Repeats (%)	0.06	0.06	0.004	0.007
Classified Complex Recombinant Repeat composition (%) LINE SINE LTR DNA Other	 42.75 22.58 16.77 7.03 10.87	 41.52 3.81 45.81 0.48 8.38	 42.27 22.19 17.84 1.54 16.16	 49.11 11.18 14.40 5.18 20.13

Complex Recombinant Repeat Content i) Lineage-specific ii) Mixed	515 431	440 43	41 21	57 48
Total Complex Recombinant Repeat Exaptation in Refgenes i) 5' UTR Exaptation ii) 3' UTR Exaptation iii) Exonic Exaptation	11 7 8	3 0 3	0 0 0	0 0 0

*Total Complex Recombinant Repeats are all the sequences found to contain ≥ 3 TE fragments.

**Classified Complex Recombinant Repeats sequences (≥ 3 copies) that could be clustered and classified into families based on their TE content.

Supporting Information Legends:

Supplementary Table S1: Complex recombinant repeat families in four mammals: human, mouse, cow and horse.

Supplementary Table S2: Complex recombinant repeats mapped into segmental duplication regions, and exapted into exonic regions or located in intronic regions of protein-coding genes.

Supplementary Information S1: The Detail Complex Recombinant Repeat Identification and Clustering Process.

Supplementary Figure S1: The Complex Recombinant Repeat Distributions in Human and Mouse Chromosome.

Supplementary Figure S2: The Comparison of hg19 Segmental duplication map, human segmental duplicated gene (TRIM49C), and CRR sequence (Family 26.7) coordinates presented in UCSC Genome Browser.

Figure 1

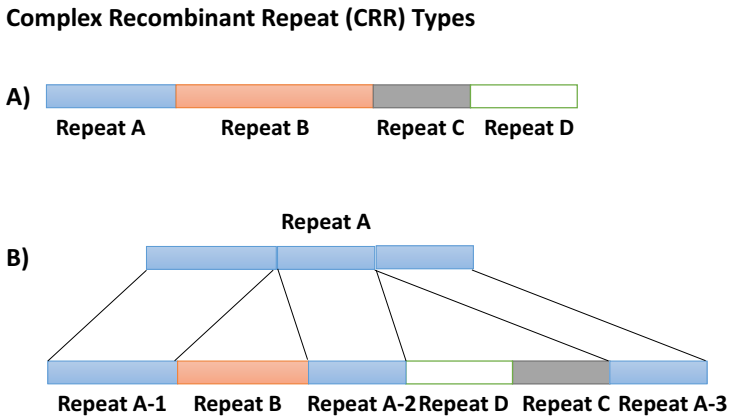


Figure 2:

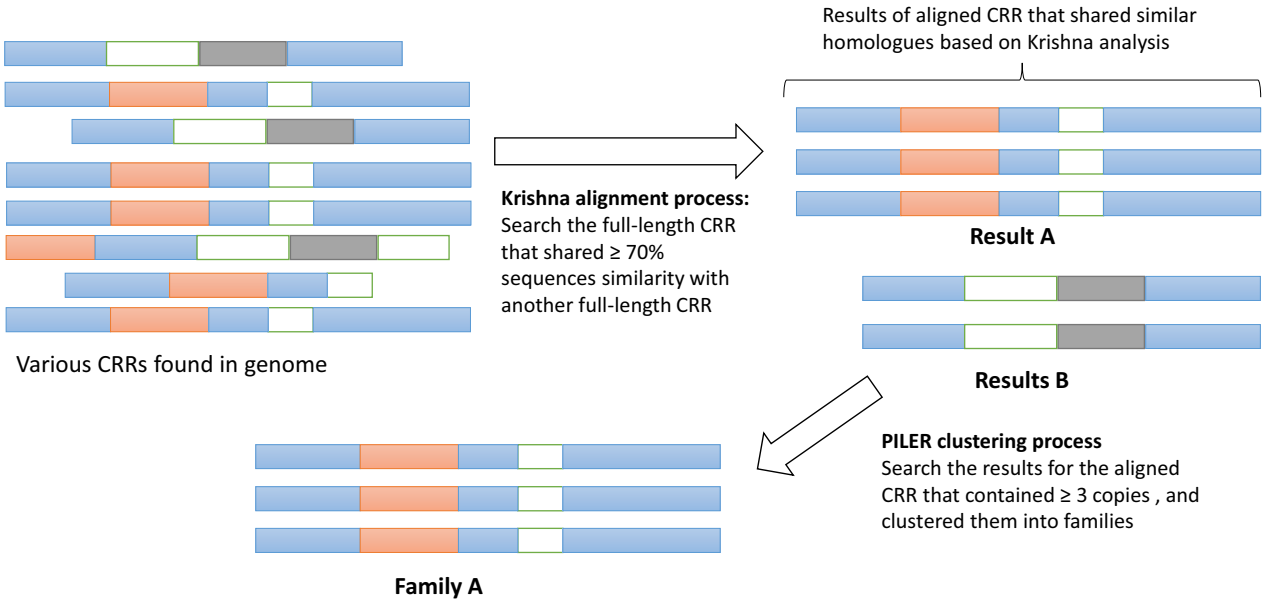


Figure 3:

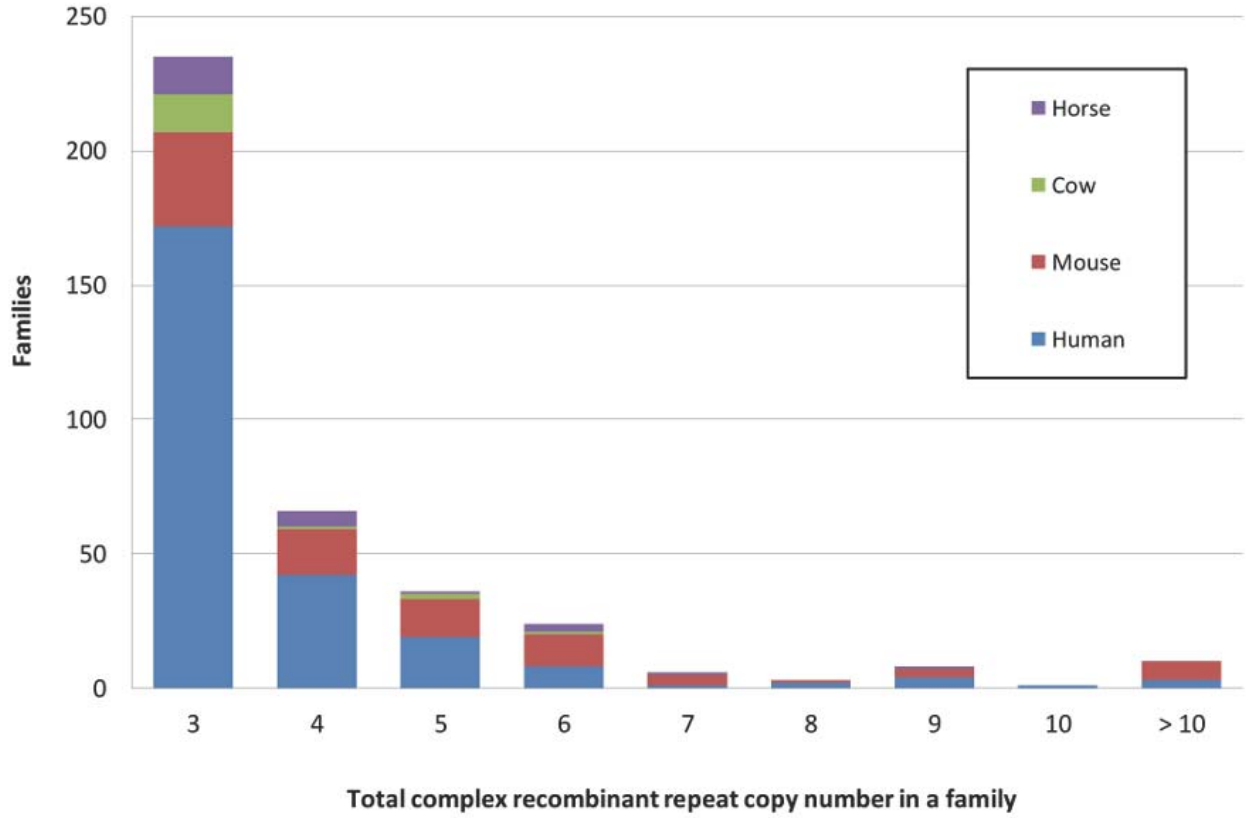

250
200
150
100
50
0
Families
Horse
Cow
Mouse
Human
3
4
5
6
7
8
9
10
> 10
Total complex recombinant repeat copy number in a family

Figure 4:

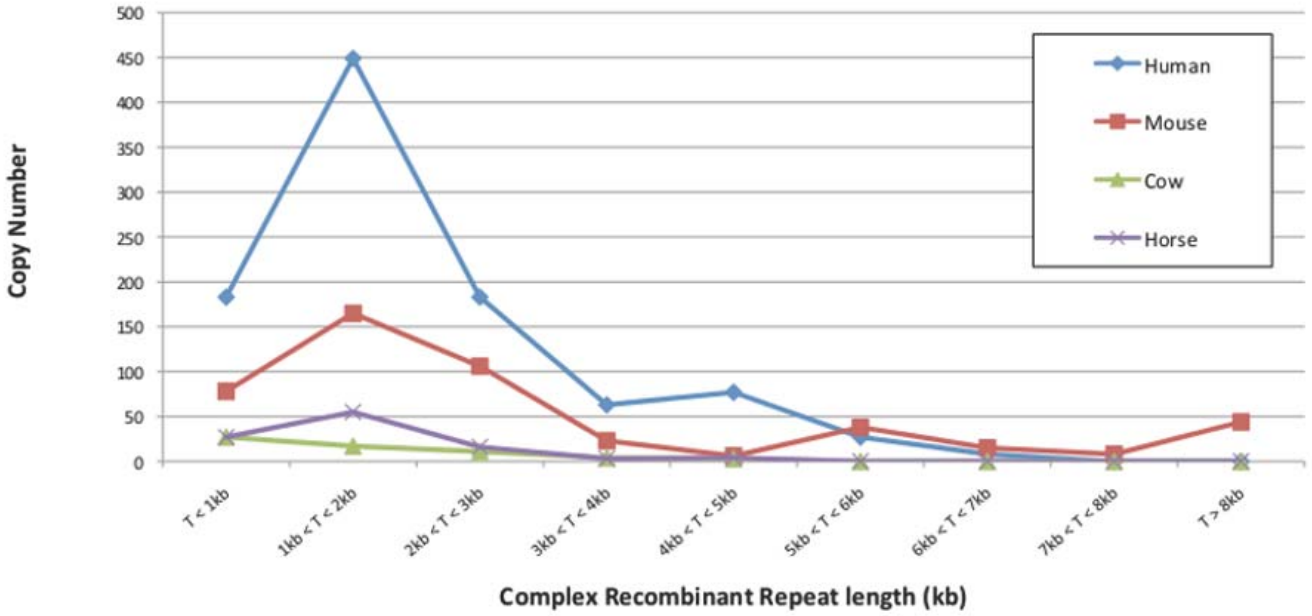

Copy Number
500
450
400
350
300
250
200
150
100
50
0
Human
Mouse
Cow
Horse
T < 1kb
1kb < T < 2kb
2kb < T < 3kb
3kb < T < 4kb
4kb < T < 5kb
5kb < T < 6kb
6kb < T < 7kb
7kb < T < 8kb
T > 8kb
Complex Recombinant Repeat length (kb)

Figure 5:

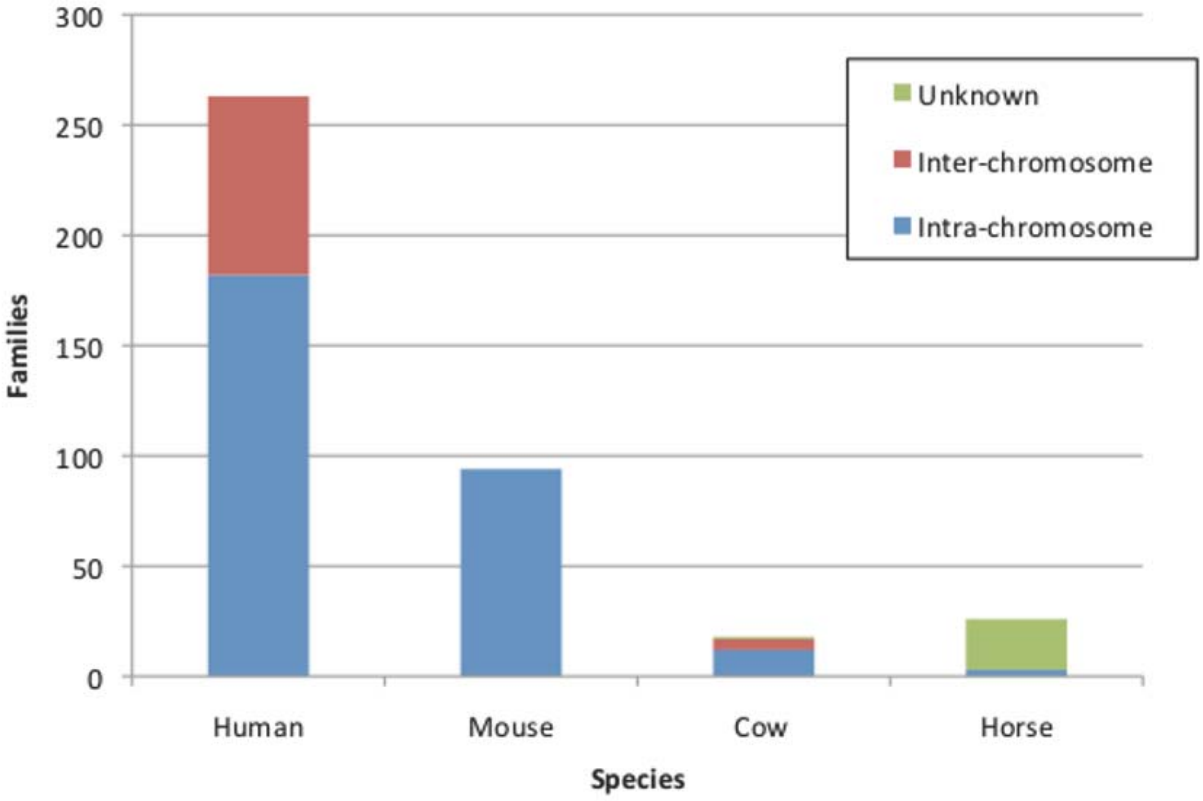
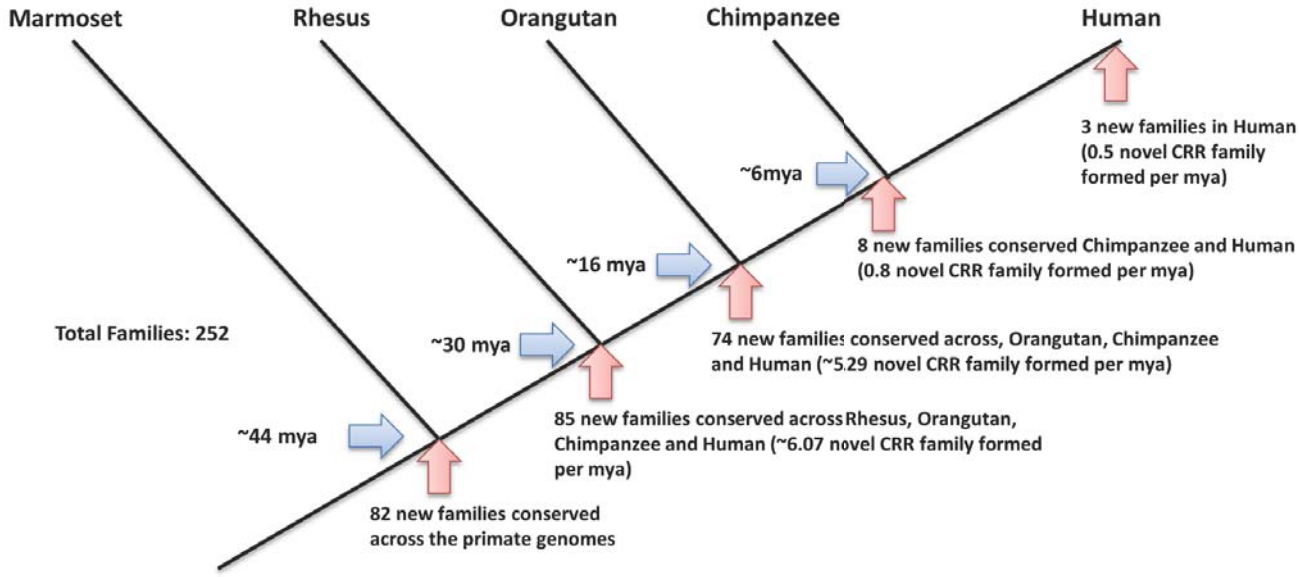

300
250
200
150
100
50
0
Families
Unknown
Inter-chromosome
Intra-chromosome
Human
Mouse
Cow
Horse
Species

Figure 6:


Marmoset
Rhesus
Orangutan
Chimpanzee
Human
3 new families in Human
(0.5 novel CRR family
formed per mya)
~6mya
8 new families conserved Chimpanzee and Human
(0.8 novel CRR family formed per mya)
~16 mya
74 new families conserved across, Orangutan, Chimpanzee
and Human (~5.29 novel CRR family formed per mya)
Total Families: 252
~30 mya
85 new families conserved across Rhesus, Orangutan,
Chimpanzee and Human (~6.07 novel CRR family formed
per mya)
~44 mya
82 new families conserved
across the primate genomes

Statement of Authorship

Title of Paper	
Publication Status	○ Published, ○ Accepted for Publication, ○ Submitted for Publication, ○ Publication style
Publication Details	

Author Contributions

By signing the Statement of Authorship, each author certifies that their stated contribution to the publication is accurate and that permission is granted for the publication to be included in the candidate's thesis.

Name of Principal Author (Candidate)			
Contribution to the Paper			
Signature		Date	

Name of Co-Author			
Contribution to the Paper			
Signature		Date	

Name of Co-Author			
Contribution to the Paper			
Signature		Date	

Name of Co-Author			
Contribution to the Paper			
Signature		Date	

Chapter 4

Discovery of A Novel LTR (LTR2i_SS) in *Sus scrofa*

Sim Lin Lim, R. Daniel Kortschak and David L. Adelson

School of Molecular and Biomedical Science
The University of Adelaide
North Terrace
Adelaide, 5005
South Australia
Australia

RESEARCH ARTICLE

Discovery Of A Novel LTR (LTR2i_SS) In *Sus Scrofa*

Sim L. Lim, R. Daniel Kortschak and David L. Adelson*

School of Molecular and Biomedical Science

The University of Adelaide

North Terrace

Adelaide, 5005

South Australia

Australia

***Corresponding Author: email: david.adelson@adelaide.edu.au,**

Tel: +61 (0)8 8313 7555,

Fax: +61 (0) 8313 4362

Abstract

LTR retrotransposons are transposable elements flanked with 5'/3' Long Terminal Repeats (LTRs). They have a similar structure to endogenous retroviruses (ERV) but they lack the envelope (*env*) gene making them non-infectious. LTRs are motif-rich sequences, and can act as bidirectional promoters or enhancers to regulate or inactivate genes by insertion. In this report, we identified a new chimeric LTR subfamily, LTR2i_SS, in the pig genome. This chimeric LTR family appears to be the ancestral form of the previously described LTR2_SS family. LTR2_SS appears to have deleted ~300bp of un-annotated, ancestral sequence from LTR2i_SS. We identified no functional provirus sequences for either of these LTR types. LTR2i_SS sequences have been exapted into the untranslated regions of two protein-coding gene mRNAs. Both of these genes lie within previously mapped pig quantitative trait loci (QTL).

Introduction

Eukaryotic genomes generally contain large numbers of transposable elements (TE), including variable numbers of Long terminal Repeat (LTR) retrotransposons. LTR retrotransposons contain flanking LTR sequences and their gene structures (*pol* and *gag*) share homology with endogenous retroviruses (ERV) (Garfinkel *et al.* 1985; Mellor *et al.* 1985; Adams *et al.* 1987). However, the LTR retrotransposons are not infectious since they lack the envelope gene (*env*). They replicate through a "Copy and Paste" RNA reverse-transcription process similar to yeast *Ty* elements (Mellor *et al.* 1985; Adams *et al.* 1987; Eichinger & Boeke 1988; Xu & Boeke 1990; Muller *et al.* 1991). During retrotranposition, they are also able to undergo intra-element homologous recombination via their LTRs, resulting in the excision of their structural genes (Belshaw *et al.* 2007), leaving a lone LTR at the

insertion site.

LTR retrotransposons are capable of integrating at almost any genomic location, potentially inactivating genes by physical disruption (Varmus 1982), and creating structural variation (Sverdlov 2000). They contain multiple motifs and transcription factor (TF) binding sites (van de Lagemaat *et al.* 2003; Wang *et al.* 2007; Bourque *et al.* 2008; Feschotte 2008) allowing them to exert transactivational regulatory effects or provide additional polyadenylation signals that may contribute to the formation of alternative transcripts (van de Lagemaat *et al.* 2003; Dunn *et al.* 2006; Huh *et al.* 2008; Faulkner *et al.* 2009).

The identification of novel LTR retrotransposons is challenging due to the complexity of LTR retrotransposon structures. Several tools have been developed to find novel LTR retrotransposons (McCarthy & McDonald 2003; Kalyanaraman & Aluru 2006; Sperber *et al.* 2007; Sperber *et al.* 2009). These tools are based on two broad approaches: *de novo* identification; and homology-based identification. *De novo* tools identify the repetitive structure of LTR retrotransposons based on sequence similarity within the target genome, clustering them into families of related elements (Rho *et al.* 2007; Steinbiss *et al.* 2009). Homology-based tools annotate potential LTR retrotransposons based on available consensus sequences from the Repbase Library (Jurka *et al.* 1996; Jurka *et al.* 2005).

In this report, we characterise a novel chimeric LTR family in pig (*Sus scrofa*) along with related proviral sequences. We have carried out a comparative analysis to determine the evolutionary relationship between this novel LTR and previously described LTR elements. Finally, we report exaptation events in protein-coding genes that are candidates for a number of pig production traits.

Materials and Methods

Database

The data used in this research were downloaded from public databases. The genome assembly of pig (susScr3) was obtained from the UCSC genome repository (http://genome.ucsc.edu/index.html) (Groenen *et al.* 2012). The Wuzhishan (minipig) genome project data, version 1.0 WGS assembly (Genbank number: AJKK00000000.1) was downloaded from NCBI Traces archive (http://www.ncbi.nlm.nih.gov/Traces/wgs/) (Fang *et al.* 2012).

Pig-specific repetitive element sequences (as of March 2013) were obtained from Repbase (http://www.girinst.org/repbase/) (Jurka *et al.* 2005). Quantitative Trait Loci (QTL) data were obtained from the Animal QTLdb (Hu *et al.* 2013).

Software For Complex Recombinant Repeat Identification And Classification

WU-BLAST, krishna (both used for local sequence alignment) and PILER (used for clustering of sequence families), RepeatMasker, Censor (both used to annotate families based on known repeat consensus sequences) and BEDTools (used to identify overlapping/intersecting sequence annotations ie genes and repeats) used in our pipeline are open source tools (Altschul *et al.* 1990; Jurka *et al.* 1996; Edgar & Myers 2005) (Supplementary Table 1). The NCBI BLAST tools used in this analysis are available in NCBI website (http://blast.ncbi.nlm.nih.gov/).

LTR2i_SS Family Identification Pipeline

The pig genome sequence was masked with RepeatMasker using pig-specific libraries from Repbase. Recombinant repeats (RR) made up of two or more repeat types were extracted and allocated to different groups based on the size, order and type of repeat fragments. A *de novo* repeat identification pipeline using krishna (http://godoc.org/code.google.com/p/biogo.examples/krishna) was used to align sequences from each group, based on sequence identity > 70% and PILER (Edgar

& Myers 2005) was used to cluster the alignments into families of three or more members.

Constructing LTR2i_SS And LTR2_SS Phylogenetic Tree

We used MUSCLE (Edgar 2004) to generate global alignments of LTR2i_SS sequences, followed by tree construction with FastTree (Price *et al.* 2009) using the maximum-likelihood method with the general-time reversible (GTR) model. LTR2_SS sequences were identified using RepeatMasker and were also included to determine their relationship with LTR2i_SS sequences.

Identifying Proviruses Flanked With LTR2i_SS Or LTR2_SS

Comparison of intact proviruses flanked with LTR2i_SS or LTR2_SS was performed by calculating the distance (bp) between pairs of LTR2i_SS/LTR2_SS sequences in each pig chromosome. If the distance was ≥4kbp but ≤10kb, we used RepeatMasker and Censor to annotate that genomic interval. If the interval was annotated as endogenous retrovirus or retrovirus, we used the NCBI BLASTX webtool (http://blast.ncbi.nlm.nih.gov/) to identify potential retroviral proteins in the UNIPROT/SWISSPROT database (Magrane & Consortium 2011). In order to compare the LTR2i_SS and LTR2_SS proviruses, we aligned them with MUMMER (Kurtz *et al.* 2004).

Motif Search On LTR2i_SS Sequences

It is known that LTR are enriched in functional motifs. To search for potential transcription factor sites in LTR2i_SS, we used MEME (Bailey *et al.* 2009) to predict 10 potential motif regions in LTR2i_SS sequences. We determined if the motif sequences from LTR2i_SS were present in the JASPAR CORE Vertebrate database (http://jaspar.genereg.net/) (Sandelin *et al.* 2004).

Overlap Of LTR2i_SS With Genes/mRNAs/Mapped Traits

We intersected LTR2i_SS intervals with susScr3 RefSeq gene 5' UTR, 3' UTR and Exon intervals using BEDTools to find potential exaptation events (http://code.google.com/p/bedtools/). We also used the LTR2i_SS consensus sequence as a query to search the NCBI reference RNA sequences (refseq_rna) and EST database with BLASTX (http://blast.ncbi.nlm.nih.gov/) in order to find any potential transcript variants containing LTR2i_SS sequences. QTL containing the genes with exapted LTR2i_SS were identified using Animal QTLdb.

Results and Discussion

A Novel Chimeric LTR Family In Pig

During a search for recombinant repeats in the susScr3 Duroc pig genome assembly (Groenen *et al.* 2012) (S.L. Lim, R. D. Kortchak, D.L Adelson, unpublished), we identified over one thousand copies of a novel chimeric LTR of 770bp average length (852bp length for consensus sequence) (See Table 1 and Supplementary Table 2). Approximately 531bp of this chimeric LTR concensus sequence could be annotated by RepeatMasker and Censor as LTR2_SS with a ~300bp un-annotated sequence insertion at position ~160 of the LTR2_SS consensus sequence (Figure 1). Based on BLAST similarity searches of all publicly available sequence data, the un-annotated sequence was only found in this chimeric LTR and could not be identified elsewhere in the pig genome or in any other NCBI database. We named this novel LTR variant 'LTR2i_SS', to denote it as an insertion variant of LTR2_SS.

LTR sequences are known to contain functional motifs such as transcription factor binding sites (van de Lagemaat *et al.* 2003; Wang *et al.* 2007) and the LTR2i_SS sequence is no exception. We identified ten motifs based on the full set

of LTR2i_SS sequences and five of these were present in the un-annotated sequence (see Supplementary Information 1).

There were approximately 50-fold more LTR2i_SS present in susScr3 compared to LTR2_SS (Table 1). This result is consistent with LTR2i_ss as the ancestral version of the LTR. We found that the other publicly available pig genome sequence, the minipig (Fang *et al.* 2012) contained the same ratio of LTR2i_SS to LTR2_SS (Table 1). We therefore inferred that both LTR2i_SS and LTR2_SS probably appeared prior to domestication.

Comparison Of LTR2i_SS And LTR2_SS Proviruses

Because the 50-fold difference in abundance of LTR2i_SS compared to LTR2_SS, indicated significant retrotranspositional activity for the LTR2i_SS, and because this requires the presence of intact proviruses, we decided to identify the potential proviral sequences for each LTR type. We found no potential proviral sequences in the Wuzhishan pig, but since this assembly contains only contigs and they were not assembled into known chromosomes, we could not identify provirus sequences split across contig boundaries. In the susScr3 assembly we found a total of six proviral sequences for both LTR2 types (Table 2). Four of proviruses contained pairs of the LTR2_SS sequences and two contained pairs of LTR2i_SS sequences. The proviral sequences could be classified into three types; two 8kbp LTR2_SS proviruses, two 6kbp LTR2_SS proviruses and two 6kbp LTR2i_SS proviruses. Comparison of these three types indicated that the 8kbp forms contained *Gag, Pro* and *Pol* genes whereas the 6kbp LTR2_SS proviruses contained *Gag* and *Pro* genes only and had a deletion of most of the *Pol* gene (Figure 2A). The 6kbp LTR2i_SS contained the *Pro* gene and most of the *Pol* gene, but had a partial deletion of the *Gag* gene and a large deletion in the non-coding region

between the *Pol* gene and the 3'LTR (Figures 2B, C). The only characteristic shared by all four 6kbp forms was an intact *Pro* gene. We could not identify a single functional polyprotein ORF in any of the 8kbp or 6kbp proviruses, indicating that none of these proviruses are capable of autonomous retrotransposition or LTR insertion. This implies that if any of these proviruses are currently being retrotransposed they are acting in a non-autonomous fashion and dependent on other closely related intact PERVs.

Three of the six provirus sequences have been previously described (Groenen *et al.* 2012) but three are novel (Table 2). Based on prior phylogenetic classification of ERV in the pig genome, all six of the provirus sequences identified belong to the PERV β 3 family (Groenen *et al.* 2012). This would suggest that the ancestral LTR2_SS provirus was of the 8kbp form. The over-representation of LTR2i_SS sequences suggests this was the ancestral form and the most parsimonious explanation for our observations is that the ancestral 8kbp form of the LTR2i_SS provirus is no longer present in the Duroc pig genome (susScr3 assembly). If the LTR2_SS form were ancestral we would expect those sequences to cluster into a single family, compared to the LTR2i_SS sequences. This is not supported by our phylogenetic analysis of these LTR sequences (see below, Figure 3). However, in the absence of intact autonomous provirus sequence of either variety, we cannot be confident of the ancestral relationship between LTR2i_SS and LTR2_SS PERVs.

Evolutionary Dynamics Of LTR2i_SS

The relationships between individual LTR2i_SS sequences, and LTR2_SS and LTR2i_SS sequences are best described using a phylogenetic tree (shown in Figure 3). In order to generate a tree, we used MUSCLE to align the LTR2i_SS

sequences and the LTR2_SS sequences and plotted a maximum-likelihood (General time reversible model) phylogenetic tree. The LTR2_SS sequences clustered into two sub-families, one of which exhibits long branch lengths. Inspection of the global alignment of all LTR2_SS sequences (Supplementary Information 3) revealed that there were many indels of 10 or more base pairs scattered over the LTRs and this probably accounts for the observed long branch lengths. The LTR2i_SS sequences are grouped into 7 or 8 distinct subfamilies with shorter branch lengths that are based on high support values (≥ 0.85) for the relevant tree nodes. The presence of two LTR2_SS sub-families suggests that they arose as the result of at least two independent deletion events of the un-annotated 300bp region of LTR2i_SS. This is a far likelier scenario than two independent insertion events of the same fragment into the same location and therefore supports our conclusion that LTR2i_SS sequences correspond to the ancestral form of the LTR. The observed diversity of the LTR2i_SS sequences is also consistent with an ancestral relationship to the LTR2_SS form.

The LTR associated with the two 6kbp LTR2i_SS-containing proviruses cluster together as a single subfamily. These two proviruses are located about 100kbp apart on SS7 indicating they may have resulted from a segmental duplication event. However, we found no known segmental duplication overlapping these two sequences (Groenen *et al.* 2012). The other 4 provirus sequences had LTR that clustered into the same LTR2_SS subfamily.

Potential Exaptation

We looked for potential exaptation events by identifying LTR2_SS and LTR2i_SS sequence overlaps with annotated genes or with known transcripts. We found no instance of possible exaptation of LTR2_SS, but two instances of possible

exaptation of LTR2i_SS in mRNA transcripts (Figure 4). Both potential exaptation events are into non-coding sequences, one is found at the 3' end of the 3' UTR of the Predicted Sus scrofa GTP-binding protein 10 (GTPBP10), transcript variant X2 (Accession number: XM_003482668) mRNA where ~550bp of LTR2i_SS has been inserted, including the full 300bp un-annotated sequence. The GTPBP10 gene is present on SS9 and lies within 26 previously mapped QTL for a variety of traits (Supplementary Table 3) most of which relate to carcass composition and/or growth rate. The other exaptation event was a 50bp insertion that did not include any of the un-annotated region into the 5' UTR of Sus scrofa sulfotransferase family 1E, estrogen-preferring member 1 (SULT1E1) mRNA (Accession number: NM_213992), located on SS8. We are confident this insertion is from LTR2i_SS as opposed to LTR2_SS based on the phylogenetic analysis of LTR2i_SS, with the most closely related LTR2i_SS sequence present in a sub-family that was well separated and did not cluster with either of the LTR2_SS subfamilies (Figure 3). *SULT1E1* is not present in the susScr3 genome assembly, but was previously published as part of QTL candidate gene localization (Kim *et al.* 2002). *SULT1E1* is a candidate gene for uterine capacity, but lies within a region that overlaps a total of 23 QTL, 5 of which are carcass traits and 2 of which are reproduction traits (Supplementary Table 3).

Conclusion

We have refined our understanding of the PERV β 3 family by describing a previously unknown LTR, LTR2i_SS. LTR2i_SS contains an extra 300bp of sequence compared to the previously characterized LTR2_SS. Based on copy number and phylogenetic analysis LTR2i_SS is most likely the ancestral form of LTR2_SS which has deleted 300bp of the ancestral sequence. LTR2i_SS has been

exapted into the non-coding regions of two protein-coding genes, both of which are present in previously mapped pig QTL.

References

Adams S.E., Mellor J., Gull K., Sim R.B., Tuite M.F., Kingsman S.M. & Kingsman A.J. (1987) The functions and relationships of Ty-VLP proteins in yeast reflect those of mammalian retroviral proteins. *Cell* **49**, 111-9.

Altschul S.F., Gish W., Miller W., Myers E.W. & Lipman D.J. (1990) Basic local alignment search tool. *J Mol Biol* **215**, 403-10.

Bailey T.L., Boden M., Buske F.A., Frith M., Grant C.E., Clementi L., Ren J., Li W.W. & Noble W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**, W202-8.

Belshaw R., Watson J., Katzourakis A., Howe A., Woolven-Allen J., Burt A. & Tristem M. (2007) Rate of recombinational deletion among human endogenous retroviruses. *J Virol* **81**, 9437-42.

Bourque G., Leong B., Vega V.B., Chen X., Lee Y.L., Srinivasan K.G., Chew J.L., Ruan Y., Wei C.L., Ng H.H. & Liu E.T. (2008) Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res* **18**, 1752-62.

Dunn C.A., Romanish M.T., Gutierrez L.E., van de Lagemaat L.N. & Mager D.L. (2006) Transcription of two human genes from a bidirectional endogenous retrovirus promoter. *Gene* **366**, 335-42.

Edgar R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-7.

Edgar R.C. & Myers E.W. (2005) PILER: identification and classification of genomic repeats. *Bioinformatics* **21 Suppl 1**, i152-8.

Eichinger D.J. & Boeke J.D. (1988) The DNA intermediate in yeast Ty1 element transposition copurifies with virus-like particles: cell-free Ty1 transposition. *Cell* **54**, 955-66.

Fang X., Mou Y., Huang Z., Li Y., Han L., Zhang Y., Feng Y., Chen Y., Jiang X., Zhao W., Sun X., Xiong Z., Yang L., Liu H., Fan D., Mao L., Ren L., Liu C., Wang J., Li K., Wang G., Yang S., Lai L., Zhang G., Li Y., Wang J., Bolund L., Yang H., Wang J., Feng S., Li S. & Du Y. (2012) The sequence and analysis of a Chinese pig genome. *Gigascience* **1**, 16.

Faulkner G.J., Kimura Y., Daub C.O., Wani S., Plessy C., Irvine K.M., Schroder K., Cloonan N., Steptoe A.L., Lassmann T., Waki K., Hornig N., Arakawa T., Takahashi H., Kawai J., Forrest A.R., Suzuki H., Hayashizaki Y., Hume D.A., Orlando V., Grimmond S.M. & Carninci P. (2009) The regulated retrotransposon transcriptome of mammalian cells. *Nat Genet* **41**, 563-71.

Feschotte C. (2008) Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* **9**, 397-405.

Garfinkel D.J., Boeke J.D. & Fink G.R. (1985) Ty element transposition: reverse transcriptase and virus-like particles. *Cell* **42**, 507-17.

Groenen M.A., Archibald A.L., Uenishi H., Tuggle C.K., Takeuchi Y., Rothschild M.F., Rogel-Gaillard C., Park C., Milan D., Megens H.J., Li S., Larkin D.M., Kim H., Frantz L.A., Caccamo M., Ahn H., Aken B.L., Anselmo A., Anthon C., Auvil L., Badaoui B., Beattie C.W., Bendixen C., Berman D., Blecha F., Blomberg J., Bolund L., Bosse M., Botti S., Bujie Z., Bystrom M., Capitanu B., Carvalho-Silva D., Chardon P., Chen C., Cheng R.,

Choi S.H., Chow W., Clark R.C., Clee C., Crooijmans R.P., Dawson H.D., Dehais P., De Sapio F., Dibbits B., Drou N., Du Z.Q., Eversole K., Fadista J., Fairley S., Faraut T., Faulkner G.J., Fowler K.E., Fredholm M., Fritz E., Gilbert J.G., Giuffra E., Gorodkin J., Griffin D.K., Harrow J.L., Hayward A., Howe K., Hu Z.L., Humphray S.J., Hunt T., Hornshoj H., Jeon J.T., Jern P., Jones M., Jurka J., Kanamori H., Kapetanovic R., Kim J., Kim J.H., Kim K.W., Kim T.H., Larson G., Lee K., Lee K.T., Leggett R., Lewin H.A., Li Y., Liu W., Loveland J.E., Lu Y., Lunney J.K., Ma J., Madsen O., Mann K., Matthews L., McLaren S., Morozumi T., Murtaugh M.P., Narayan J., Nguyen D.T., Ni P., Oh S.J., Onteru S., Panitz F., Park E.W., Park H.S., Pascal G., Paudel Y., Perez-Enciso M., Ramirez-Gonzalez R., Reecy J.M., Rodriguez-Zas S., Rohrer G.A., Rund L., Sang Y., Schachtschneider K., Schraiber J.G., Schwartz J., Scobie L., Scott C., Searle S., Servin B., Southey B.R., Sperber G., Stadler P., Sweedler J.V., Tafer H., Thomsen B., Wali R., Wang J., Wang J., White S., Xu X., Yerle M., Zhang G., Zhang J., Zhang J., Zhao S., Rogers J., Churcher C. & Schook L.B. (2012) Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* **491**, 393-8.

Hu Z.L., Park C.A., Wu X.L. & Reecy J.M. (2013) Animal QTLdb: an improved database tool for livestock animal QTL/association data dissemination in the post-genome era. *Nucleic Acids Res* **41**, D871-9.

Huh J.W., Kim D.S., Kang D.W., Ha H.S., Ahn K., Noh Y.N., Min D.S., Chang K.T. & Kim H.S. (2008) Transcriptional regulation of GSDML gene by antisense-oriented HERV-H LTR element. *Arch Virol* **153**, 1201-5.

Huson D.H. & Scornavacca C. (2012) Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Syst Biol* **61**, 1061-7.

Jurka J., Kapitonov V.V., Pavlicek A., Klonowski P., Kohany O. & Walichiewicz J. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110**, 462-7.

Jurka J., Klonowski P., Dagman V. & Pelton P. (1996) CENSOR--a program for identification and elimination of repetitive elements from DNA sequences. *Comput Chem* **20**, 119-21.

Kalyanaraman A. & Aluru S. (2006) Efficient algorithms and software for detection of full-length LTR retrotransposons. *J Bioinform Comput Biol* **4**, 197-216.

Kim J.G., Vallet J.L., Rohrer G.A. & Christenson R.K. (2002) Characterization of porcine uterine estrogen sulfotransferase. *Domest Anim Endocrinol* **23**, 493-506.

Kurtz S., Phillippy A., Delcher A.L., Smoot M., Shumway M., Antonescu C. & Salzberg S.L. (2004) Versatile and open software for comparing large genomes. *Genome Biol* **5**, R12.

Magrane M. & Consortium U. (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)* **2011**, bar009.

McCarthy E.M. & McDonald J.F. (2003) LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics* **19**, 362-7.

Mellor J., Malim M.H., Gull K., Tuite M.F., McCready S., Dibbayawan T., Kingsman S.M. & Kingsman A.J. (1985) Reverse transcriptase activity and Ty RNA are associated with virus-like particles in yeast. *Nature* **318**, 583-6.

Muller F., Laufer W., Pott U. & Ciriacy M. (1991) Characterization of products of TY1-mediated reverse transcription in Saccharomyces cerevisiae. *Mol Gen Genet* **226**, 145-53.

Price M.N., Dehal P.S. & Arkin A.P. (2009) FastTree: computing large minimum evolution

trees with profiles instead of a distance matrix. *Mol Biol Evol* **26**, 1641-50.
Rho M., Choi J.H., Kim S., Lynch M. & Tang H. (2007) De novo identification of LTR retrotransposons in eukaryotic genomes. *BMC Genomics* **8**, 90.
Sandelin A., Alkema W., Engstrom P., Wasserman W.W. & Lenhard B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* **32**, D91-4.
Sperber G., Lovgren A., Eriksson N.E., Benachenhou F. & Blomberg J. (2009) RetroTector online, a rational tool for analysis of retroviral elements in small and medium size vertebrate genomic sequences. *BMC Bioinformatics* **10 Suppl 6**, S4.
Sperber G.O., Airola T., Jern P. & Blomberg J. (2007) Automated recognition of retroviral sequences in genomic data--RetroTector. *Nucleic Acids Res* **35**, 4964-76.
Steinbiss S., Willhoeft U., Gremme G. & Kurtz S. (2009) Fine-grained annotation and classification of de novo predicted LTR retrotransposons. *Nucleic Acids Res* **37**, 7002-13.
Sverdlov E.D. (2000) Retroviruses and primate evolution. *Bioessays* **22**, 161-71.
van de Lagemaat L.N., Landry J.R., Mager D.L. & Medstrand P. (2003) Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends Genet* **19**, 530-6.
Varmus H.E. (1982) Form and function of retroviral proviruses. *Science* **216**, 812-20.
Wang T., Zeng J., Lowe C.B., Sellers R.G., Salama S.R., Yang M., Burgess S.M., Brachmann R.K. & Haussler D. (2007) Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proc Natl Acad Sci U S A* **104**, 18613-8.
Xu H. & Boeke J.D. (1990) Localization of sequences required in cis for yeast Ty1 element transposition near the long terminal repeats: analysis of mini-Ty1 elements. *Mol Cell Biol* **10**, 2695-702.

Figure Captions

Fig. 1. **Comparison of LTR2i_SS and LTR2_SS sequences.** The representation of the relationship between LTR2i_SS and LTR2_SS sequences is based on RepeatMasker annotation and BLAST alignments (Supplementary information 2) of LTR2i_SS and LTR2_SS consensus sequences. We can see that LTR2i_SS is essentially an LTR2_SS that contains additional un-annotated sequence.

Fig. 2. Comparison of 8kb LTR2 provirus (ss1976), 6kb LTR2 provirus (ssA_LTR2) and 6kb LTR2i provirus (ss908). 2A shows the *Pol* gene deletion in the 6kb LTR2 provirus, 2B shows the deletion of *Gag* along with sequences 3' of the *Pro* gene in the 6kb LTR2i provirus ss908 compared to ssA_LTR2. 2C shows the

deletion of the *Gag* gene and sequences 3' of the *Pol* gene in ss908 compared with ss1976. The sequence alignments between proviruses were plotted using MUMMER (Kurtz *et al.* 2004).

Fig. 3. Phylogenetic tree of LTR2i_SS and LTR2_SS sequences. The phylogenetic tree shows the LTR2_SS sequences (green) cluster into two subfamilies, one of which has longer branch lengths and includes four pairs of LTR2_SS associated with proviruses. Two pairs LTR2i_SS associated with proviruses (orange) are clustered into a single subfamily of the phylogenetic tree. The LTR2i_SS sequences most similar to those that exapted into GTP-binding protein-like (Gtpbp10-like) (red) and sulfotransferase family 1E (SULT1E1) (blue) mRNA are mapped to 2 separate subfamilies in the tree. The phylogenetic tree constructed with FastTree was visualized and annotated using Dendroscope (Huson & Scornavacca 2012).

Fig. 4. Potential LTR2i_SS exaptation in two mRNA sequences. The 5' region of LTR2i_SS_001116 (see Supplementary Table 2) including the un-annotated repetitive element (*I*) (~550bp) was exapted into the 3'UTR of the GTP-binding protein 10 (GTPBP10) transcript variant mRNA (XM_003482668). 50bp of the 3' region of LTR2i_SS_00500 (see Supplementary Table 2) was exapted into the 5'UTR of the sulfotransferase family 1E (SULT1E1) (NM_213992) mRNA.

Table 1. The distribution of LTR2_SS and LTR2i_SS sequences in Duroc (susScr3 assembly) and Wuzhishan (Minipig, AJKK00000000.1).

	Duroc (susScr3)	Wuzhishan (Minipig)
LTR2	22	17
LTR2i_SS	1194	986

Table 2. Identified LTR2_SS and LTR2i_SS proviruses in susScr3 genome assembly

Name	LTR type	Length	Chromosome	Start	End
ss1976	LTR2	8kb	X	53562664	53571113
ss2085	LTR2	8kb	X	84268150	84276679
ssA_LTR2	LTR2	6kb	11	33040648	33046767
ssB_LTR2	LTR2	6kb	1	260379969	260386212
ss908	LTR2i	6kb	7	22824165	22830303
ss_LTR2i	LTR2i	6kb	7	22733150	22739281

Supporting Information Legends:

Supplementary Info 1. Comparison of LTR2i_SS with LTR2_SS sequences with WU-BLAST.

Supplementary Info 2. LTR2i_SS motifs predicted by MEME. The predicted motifs were compared with JASPAR vertebrate database

Supplementary Info 3. Multiple Alignment of LTR2i_SS and LTR2_SS sequences by MUSCLE.

Table S1. The software used in pig LTR2i_SS identification

Table S2. The Genomic Location of LTR2i_SS in Duroc (susScr3.0 assembly)

Table S3. Potential LTR2i_SS exaptations (LTR2i_SS_001116 and LTR2i_SS_00500) that intersect with the pig QTL obtained from the Animal QTLdb.

Figure 1:

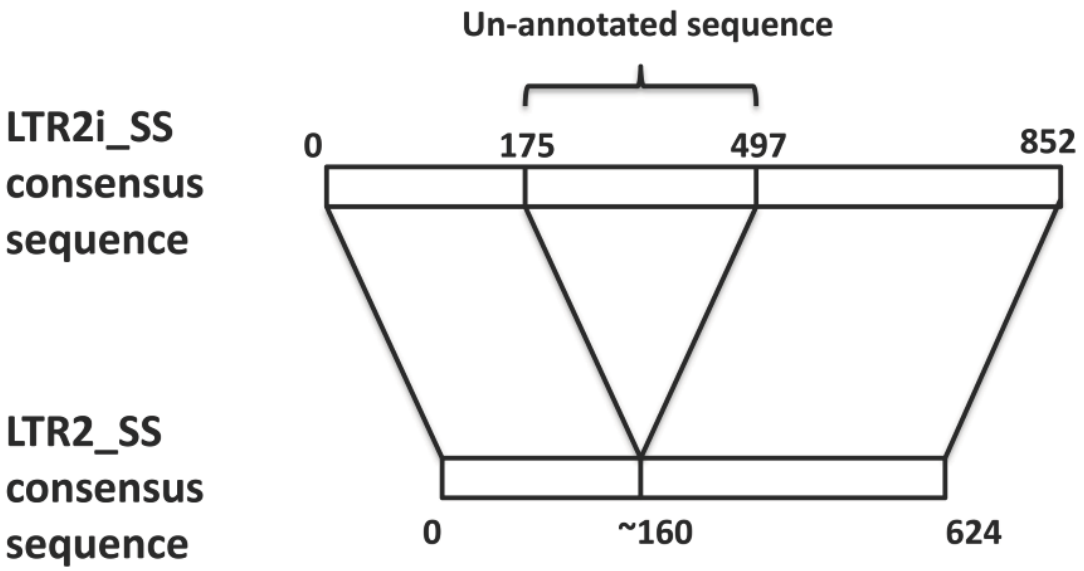

Un-annotated sequence
LTR2i_SS consensus sequence
0
175
497
852
LTR2_SS consensus sequence
0
~160
624

Figure 2:

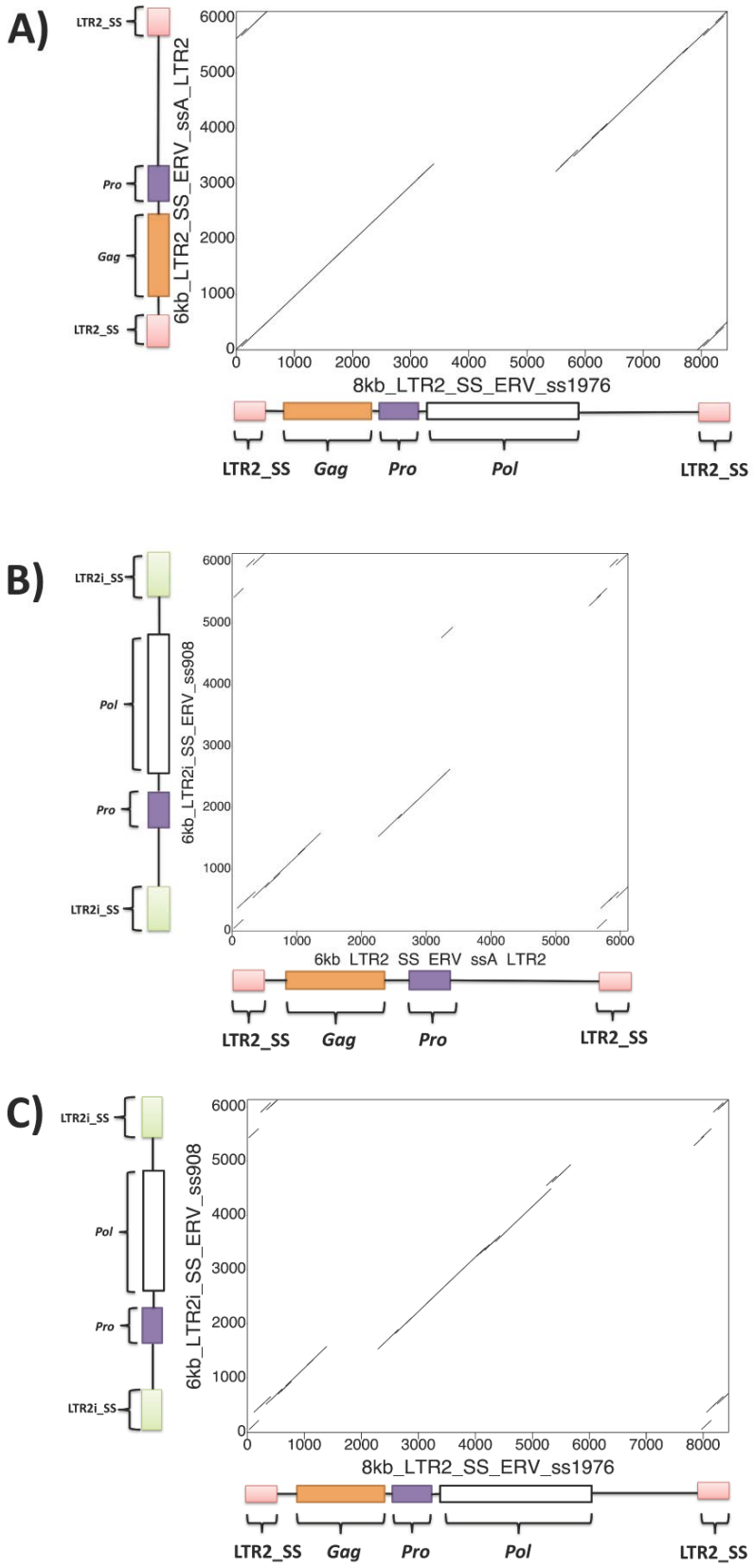

A)
LTR2_SS
Pro
Gag
LTR2_SS
6kb_LTR2_SS_ERV_ssA_LTR2
8kb_LTR2_SS_ERV_ss1976
LTR2_SS
Gag
Pro
Pol
LTR2_SS
B)
LTR2i_SS
Pol
Pro
LTR2i_SS
6kb_LTR2i_SS_ERV_ss908
6kb LTR2 SS ERV ssA LTR2
LTR2_SS
Gag
Pro
LTR2_SS
C)
LTR2i_SS
Pol
Pro
LTR2i_SS
6kb_LTR2i_SS_ERV_ss908
8kb_LTR2_SS_ERV_ss1976
LTR2_SS
Gag
Pro
Pol
LTR2_SS

Figure 3:

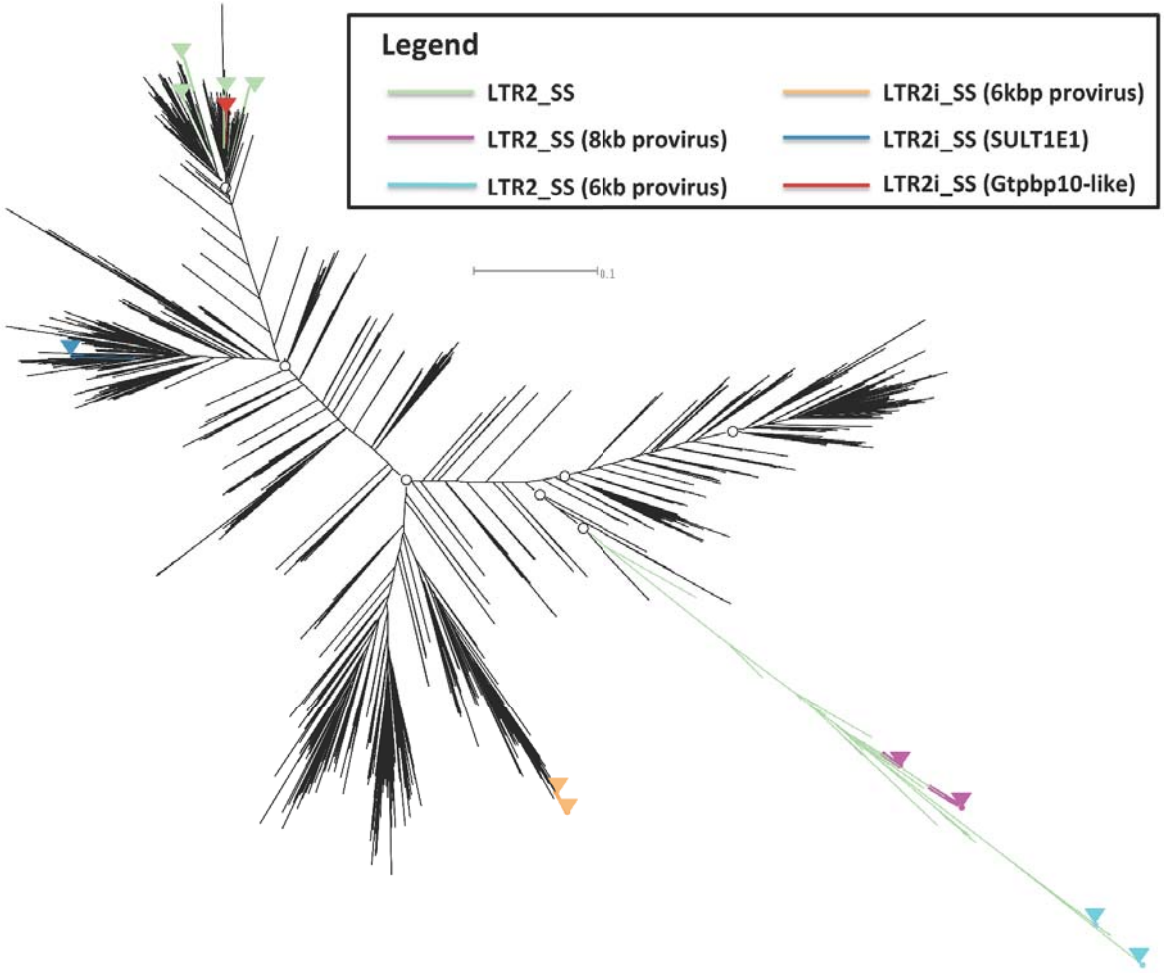
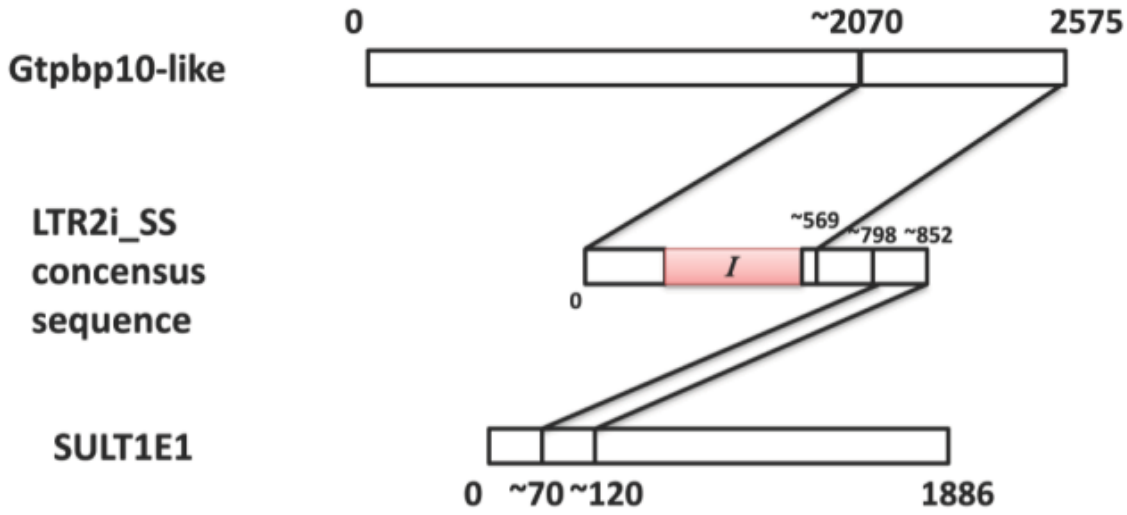

Legend
LTR2_SS
LTR2_SS (8kb provirus)
LTR2_SS (6kb provirus)
LTR2i_SS (6kbp provirus)
LTR2i_SS (SULT1E1)
LTR2i_SS (Gtpbp10-like)
0.1

Figure 4:


0
~2070
2575
Gtpbp10-like
LTR2i_SS
concensus
sequence
~569
~798 ~852
I
0
SULT1E1
0 ~70 ~120
1886

Chapter 5: Conclusions and Future Directions

Advances in genomic sequencing technologies have helped the research community to understand the distribution of transposable elements (TE) in eukaryotes. The availability of genome assemblies for many species has allowed us to identify specific TE involved in genome mutations and structural variations (SV). These resources also also help us to understand TE impacts on species divergence. Recombinant repeats (RR), the chimeric repeats composed of different repetitive and non-repetitive elements have become an active area of research in the genomic research community. Previous work has shown that RR are useful genetic markers for the reconstruction of phyogenetic trees, and understanding the history of TE activity on mammalian genomes. RR have served as important models to explain how novel TE families arise in complex genomes. However, our current understanding of RR distributions is limited because most RR are uncharacterized.

Prior to this thesis, a substantial amount of research had been done to understand different TE structures and their transposition mechanisms, and RR structures as discussed in chapter 1. We also discussed the five possible mechanisms that contribute to the novel TE and RR creations on mammalian genomes, and discussed the shortcomings of current repeat annotation pipelines in identifying RR.

In order to annotate the uncharacterized RR in genomes, we built a computational pipeline to identify simple recombinant repeats (SRR), bipartite (composed by two TE) and tripartite (composed by three TE) in four different mammals: human, mouse, cow and horse as described in chapter 2. Our pipeline helped us to identify more than 5,726 SRR families (36,234 copies) in mammals. Our

data provided an estimation of SRR family distributions in mammals (~0.245% coverage) that have not been previously reported. Our analysis showed that novel TE, especially SINE, are likely to be created via TinT events. We discovered that SRR have target site duplications (TSD) and polyA tails, indicating that they are capable of retrotransposition and integrate back to genome. However, further experimental work is required to validate SRR retrotranspositional ability.

We also carried out the identification of complex recombinant repeats (CRR) in mammals as described in chapter 3. As the name suggests, CRR are sequences that contained more than 3 TE. We used a *de novo* identification approach to find CRR families in the previously mentioned mammals. Our analysis showed that the CRR families have distinctive characteristics compared to SRR families. First, the CRR families (390) and copy numbers (1596) are significantly lower than the SRR families, and their distributions are intra-chromosomal rather than inter-chromosomal. We showed that CRR families were created via TinT events, but they were unlikely to expand through retrotransposition, indicating that the CRR families are duplicated by other mechanisms. However, our investigations showed that if individual CRR are present in a region that undergoes segmental duplications (SD), they replicate and become a CRR family. Therefore the CRR families can serve as markers to provide additional, or unique information on genomic regions that have undergone segmental duplication that is not available through other means.

In Chapter 4, we showed that our RR identification pipelines could incorrectly annotate ancestral repeats as RR if the repeats had not been described before. During the RR identification process in pig, we discovered a unique RR LTR family, LTR2i_SS. Initial analysis showed that it was a LTR2_SS repeat that contained

~300bp of additional, un-annotated sequence. The LTR2i_SS family is a pig-specific repeat and it is present at ~1000 copies in Duroc (susSc3 assembly) or Wuzhishan (minipig) pigs. We discovered two 6kbp β3 proviruses that were flanked by LTR2i_SS. These proviruses did not encode a functional polyprotein unit due to the partial deletion of the *gag* and 3' part of the *pol* genes, indicating they are unable to replicate via autonomous retrotransposition. The LTR2i_SS and LTR2_SS phylogenetic tree analysis revealed that LTR2i_SS is most likely the ancestral form of LTR2_SS. These observations showed that some of the RR identified in our pipelines are possibly ancestral forms of other TE that have not previously been described. This reinforces the need to annotate these RR cautiously in order to avoid mis-annotation in the future.

In conclusion, the future of RR research is likely to focus on improving the current RR identification pipeline, implementing new algorithms/software to identify the RR, and reducing the false-positive and false-negative of RR annotations. Once we are able to identify RR more effectively, the next challenge will be how to demonstrate that RR are expressed or retrotranspositionally active in mammals, and most importantly, to show how RR are created using *in vitro* or *in vivo* experiments. If we can solve these bioinformatic and experimental challenges, we will have a better understanding of RR, their biological impacts and how they contribute to novel TE families that arise in complex genomes.

Chapter 6: Supplementary Materials

Supplementary information of Chapter 2: Evolution of Novel Transposable Elements: Experimental Products of Recombinant Repeats?

Supplementary Information 1. The overview of software, custom PERL script and guide of simple recombinant repeat identification pipelines.

Content:

1. Method

1.1 Software Used

For repetitive elements annotation process, RepeatMasker version open-3.2.6 [1] was used. Custom script was written with PERL version 5.16.2 to identify and cluster recombinant repeat families (See Section 2) [2]. The recombinant repeat's location and information was stored in PostreSQL database [3]. The statistic overview and bar chart was plotted by Microsoft EXCEL [4], and the Boxplot was plotted by R [5]. The fastacmd version 2.2.19 was used to extract specific genomic sequences from genome assembly [6]. The TSDscan version 1.0 was used to search and align potential target site duplications in recombinant repeat flanking region [7]. BEDTools version 2.11.2 were used to manipulate genomic intervals [8].

Table 1. The software version details and downloads links.

Software	Version	Download Website
BEDTools	2.11.2	http://code.google.com/p/bedtools/
Excel for Mac	14.3.9	http://office.microsoft.com/en-au/excel/
R	3.0.2	http://www.r-project.org/
PERL	5.16.2	http://www.perl.org/
PostgreSQL	9.3.2	http://www.postgresql.org/
RepeatMasker	3.2.6	http://www.repeatmasker.org/
Fastacmd	2.2.19	http://www.ncbi.nlm.nih.gov/IEB/ToolBox/SDKDOCS/INDEX.HTML
TSDscan	1.0	http://www.intec-si.co.jp/technology/rna/

1.2 SRR identification Pipeline Overview

Species where full genome data was available were searched for simple recombinant repeats (SRR). Scripts and software settings shown in Section 2 were used to generate simple recombinant repeat families for each species. A flow chart showing the pipeline for the analysis is shown in Figure 1.

First, the RepeatMasker, shown in Section 2.1, was used to annotate the species-specific repetitive elements in different mammalian genomes. The species-specific repetitive element database was obtained from Repbase (http://www.girinst.org/repbase/).

Once the repeatmasker output results were generated, the rm2bed.pl script, shown in Section 2.2, was used to convert the output format into Bed format.

The STRR_discovery.pl and STRR_discovery_looping.pl scripts, as shown in Section 2.3, 2.4 and 2.5, were then used to search the Transposable Element (TE) sequences that have been inserted by another TE, or RR composed by 3 different TE (tripartite recombinant repeat).

The SBRR_discovery.pl and SBRR_discovery_looping.pl scripts, as shown in Section 2.3, 2.4 and 2.5, were then used to search two TE sequences that flanked with each other (Bipartite recombinant repeat. The bipartite was intersecting with tripartite sequence in previous analysis by BEDtools. If the bipartite was overlapped, the sequence was discarded.

Once those SRR sequences recovered, SRR_classification.pl script was used to cluster the SRR into families, as shown in Section. If there are ≥ 3 SRR sequences shared similar structure and content, they are classified into a single family, and the left-over sequences were discarded. The sequence information and location was stored in postgreSQL.

Figure 1. The flowchart of simple recombinant repeat family Identification


Get mammalian genome data
Annotate the Repetitive Elements in genome
RepeatMasker (Section 2.1)
Convert the RM output into BED format
rm2bed.pl (Section 2.2)
SBRR_discovery.pl, alternate_SBRR_discovery_loop.pl (Section 2.5, 2.6, 2.7)
Identify the Tripartite Recombinant Repeat in RM BED file
Identify the Bipartite Recombinant Repeat in RM BED file
STRR_discovery.pl, alternate_STRR_discovery.pl (Section 2.3, 2.4)
BEDTools (Section 2.8)
Intersecting the Bipartite genomic location with tripartite
Result: Bipartite recombinant repeats that did not overlapped with tripartite
Result: Genomic location of tripartite and bipartite recombinant repeat
Examine the TE contents in RR
If the RR contained TE derived from same family (DNA, LTR, LINE, SINE)
If the RR contained different types of TE
STRR_cluster.pl, SBRR_cluster.pl, clustering_looping.pl and copy number counting (GREP) command (Section 2.9, 2.10, 2.11, 2.12)
Cluster RR sequences into families
Discard
If the RR family contained < 3 sequences
Classified Bipartite and Tripartite RR families in mammalian genome

2. Script

2.1 RepeatMasker

Description

The genome repetitive elements can be annotated using RepeatMasker as shown below:

RepeatMasker -s -pa 6 -lib human_repetitive_element_library.fasta hg19_genome_assembly.fa

2.2 rm2bed.pl

Description

This script converts the RepeatMasker output into BED file format. The BED file contains: chromosome, sequence start location, sequence end location, matched repetitive element strand (+ for positive strand and C for negative strand), matched repetitive element types, matched repetitive element sequence start location and matched repetitive element sequence end location.

```
#!/usr/bin/perl

use strict;


my $f1 = $ARGV[0];

my @file;

my $chr;

my $seqstart;

my $seqend;

my $strand;

my $type;

my $alocation;
```

```
my $blocation;
my $clocation;
my @chr2;
my @seqstart2;
my @seqend2;
my @strand2;
my @type2;
my @alocation2;
my @blocation2;
my @clocation2;

open (FILE1, "<$f1")
   ||die "cannot open";

while ($_ = <FILE1>) {
   chomp;
   @file = split (' ', $_);
   $chr = $file[4];
   $seqstart = $file[5];
   $seqend = $file[6];
   $strand = $file[8];
   $type = $file[9];
   $alocation = $file[11];
   $blocation = $file[12];
   $clocation = $file[13];
   push(@chr2, $chr);
   push(@seqstart2, $seqstart);
   push(@seqend2, $seqend);
   push(@strand2, $strand);
```

```
   push(@type2, $type);
   push(@alocation2, $alocation);
   push(@blocation2, $blocation);
   push(@clocation2, $clocation);
}

my $number = @chr2;
my $i;

for ($i = 0; $i < $number; $i++) {
   if ($strand2[$i] =~ /\+/) {
   print "$chr2[$i]\t$seqstart2[$i]\t$seqend2[$i]\t$strand2[$i]\t$type2[$i]\t$alocation2[$i]\t$blocation2[$i]\n";
} else {

   print "$chr2[$i]\t$seqstart2[$i]\t$seqend2[$i]\t$strand2[$i]\t$type2[$i]\t$blocation2[$i]\t$clocation2[$i]\n";

}
}
```

2.3 STRR_discovery.pl

Description

This script searches potential tripartite recombinant repeats that created via insertion event. First We insert the TE (TE A) name that may inserted by another TE (TE B) name. It first examined the pair of TE A fragments are located in same chromosome, same strand, and contained TE B as an insertion. It then examined the pair of fragments were derived from a single TE or not based on the fragments' TE data (matched repetitive sequence start and sequence end location). If the pair of fragments was derived from the single TE, the script produced a new custom BED file.

```
#!/usr/bin/perl

use strict;

my $f1 = $ARGV[0];
my @file;
my $dna0 = 'TE A';
my $line0 = 'TE B';
my $chr;
my $seqstart;
my $seqend;
my $strand;
my $type;
my $alocation;
my $blocation;
my @chr2;
my @seqstart2;
my @seqend2;
my @strand2;
```

```
my @type2;
my @alocation2;
my @blocation2;




open (FILE1, "<$f1")
   ||die "cannot open";

while ($_ = <FILE1>) {
   chomp;
   @file = split (' ', $_);
   $chr = $file[0];
   $seqstart = $file[1];
   $seqend = $file[2];
   $strand = $file[3];
   $type = $file[4];
   $alocation = $file[5];
   $blocation = $file[6];

   push(@chr2, $chr);
   push(@seqstart2, $seqstart);
   push(@seqend2, $seqend);
   push(@strand2, $strand);
   push(@type2, $type);
   push(@alocation2, $alocation);
   push(@blocation2, $blocation);
```

```
}


my $number = @chr2;

my $i;

my $g;

my $h;

my $j;

my $k;

my $l;

my $m;


for ($i = 0; $i < $number; $i++) {

    $g = $i - 2;

    $h = $i - 1;

    $j = $i + 1;

    $k = $i + 2;

    $l = $i + 3;

    $m = $i + 4;

    if ($type2[$i] =~ /^$dna0$/ && $type2[$j] =~ /^$line0$/ && $type2[$k] =~
/^$dna0$/ && $strand2[$i] =~ $strand2[$k] && $blocation2[$i] > $alocation2[$k] &&
$seqstart2[$j] - $seqend2[$i] <= 10 && $seqstart2[$k] - $seqend2[$j] <= 10) {


    print
"$chr2[$i]\t$seqstart2[$i]\t$seqend2[$i]\t$strand2[$i]\t$type2[$i]\t$alocation2[$i]\t
$blocation2[$i]\t$seqstart2[$j]\t$seqend2[$j]\t$strand2[$j]\t$type2[$j]\t$alocation2[
$j]\t$blocation2[$j]\n$seqstart2[$k]\t$seqend2[$k]\t$strand2[$k]\t$type2[$k]\t$aloc
ation2[$k]\t$blocation2[$k]\n";


} elsif  ($type2[$i] =~ /^$dna0$/ && $type2[$j] =~ /^$line0$/ && $type2[$k] =~
/^$dna0$/ && $strand2[$i] =~ $strand2[$k] && $alocation2[$k] > $blocation2[$i] &&
$seqstart2[$j] - $seqend2[$i] <= 10 && $seqstart2[$k] - $seqend2[$j] <= 10) {
```

```
    print
"$chr2[$i]\t$seqstart2[$i]\t$seqend2[$i]\t$strand2[$i]\t$type2[$i]\t$alocation2[$i]\t
$blocation2[$i]\t$seqstart2[$j]\t$seqend2[$j]\t$strand2[$j]\t$type2[$j]\t$alocation2[
$j]\t$blocation2[$j]\n$seqstart2[$k]\t$seqend2[$k]\t$strand2[$k]\t$type2[$k]\t$aloc
ation2[$k]\t$blocation2[$k]\n";


}
```

2.4 alternate_STRR_discovery.pl

Description

This script searches 3 pieces of TE (TE A, TE B, and TE C) that could be potential tripartite recombinant repeat. It searched these pieces of TE that flanked each other in the RepeatMasker Outputfile. If it was found, the script produced the custom BED output file.

```
#!/usr/bin/perl

use strict;

my $f1 = $ARGV[0];
my @file;
my $dna0 = 'TE A';
my $line0 = 'TE B';
my $ltr0 = 'TE C';
my $chr;
my $seqstart;
my $seqend;
my $strand;
```

```
my $type;
my $alocation;
my $blocation;
my @chr2;
my @seqstart2;
my @seqend2;
my @strand2;
my @type2;
my @alocation2;
my @blocation2;




open (FILE1, "<$f1")
   ||die "cannot open";

while ($_ = <FILE1>) {
   chomp;
   @file = split (' ', $_);
   $chr = $file[0];
   $seqstart = $file[1];
   $seqend = $file[2];
   $strand = $file[3];
   $type = $file[4];
   $alocation = $file[5];
   $blocation = $file[6];

   push(@chr2, $chr);
```

```
    push(@seqstart2, $seqstart);
    push(@seqend2, $seqend);
    push(@strand2, $strand);
    push(@type2, $type);
    push(@alocation2, $alocation);
    push(@blocation2, $blocation);

}

my $number = @chr2;
my $i;
my $g;
my $h;
my $j;
my $k;
my $l;
my $m;

for ($i = 0; $i < $number; $i++) {
    $g = $i - 2;
    $h = $i - 1;
    $j = $i + 1;
    $k = $i + 2;
    $l = $i + 3;
    $m = $i + 4;
    if ($type2[$i] =~ /^$dna0$/ && $type2[$j] =~ /^$line0$/ && $type2[$k] =~
/^$ltr0$/ && $strand2[$i] =~ $strand2[$k] && $blocation2[$i] > $alocation2[$k] &&
$seqstart2[$j] - $seqend2[$i] <= 10 && $seqstart2[$k] - $seqend2[$j] <= 10) {
```

```
    print
"$chr2[$i]\t$seqstart2[$i]\t$seqend2[$i]\t$strand2[$i]\t$type2[$i]\t$alocation2[$i]\t
$blocation2[$i]\t$seqstart2[$j]\t$seqend2[$j]\t$strand2[$j]\t$type2[$j]\t$alocation2[
$j]\t$blocation2[$j]\n$seqstart2[$k]\t$seqend2[$k]\t$strand2[$k]\t$type2[$k]\t$aloc
ation2[$k]\t$blocation2[$k]\n";


} elsif  ($type2[$i] =~ /^$dna0$/ && $type2[$j] =~ /^$ltr0$/ && $type2[$k] =~
/^$line0$/ && $strand2[$i] =~ $strand2[$k] && $alocation2[$k] > $blocation2[$i] &&
$seqstart2[$j] - $seqend2[$i] <= 10 && $seqstart2[$k] - $seqend2[$j] <= 10) {


    print
"$chr2[$i]\t$seqstart2[$i]\t$seqend2[$i]\t$strand2[$i]\t$type2[$i]\t$alocation2[$i]\t
$blocation2[$i]\t$seqstart2[$j]\t$seqend2[$j]\t$strand2[$j]\t$type2[$j]\t$alocation2[
$j]\t$blocation2[$j]\n$seqstart2[$k]\t$seqend2[$k]\t$strand2[$k]\t$type2[$k]\t$aloc
ation2[$k]\t$blocation2[$k]\n";


} elsif  ($type2[$i] =~ /^$line0$/ && $type2[$j] =~ /^$ltr0$/ && $type2[$k] =~
/^$dna0$/ && $strand2[$i] =~ $strand2[$k] && $alocation2[$k] > $blocation2[$i] &&
$seqstart2[$j] - $seqend2[$i] <= 10 && $seqstart2[$k] - $seqend2[$j] <= 10) {


    print
"$chr2[$i]\t$seqstart2[$i]\t$seqend2[$i]\t$strand2[$i]\t$type2[$i]\t$alocation2[$i]\t
$blocation2[$i]\t$seqstart2[$j]\t$seqend2[$j]\t$strand2[$j]\t$type2[$j]\t$alocation2[
$j]\t$blocation2[$j]\n$seqstart2[$k]\t$seqend2[$k]\t$strand2[$k]\t$type2[$k]\t$aloc
ation2[$k]\t$blocation2[$k]\n";


} elsif  ($type2[$i] =~ /^$line0$/ && $type2[$j] =~ /^$dna0$/ && $type2[$k] =~
/^$ltr0$/ && $strand2[$i] =~ $strand2[$k] && $alocation2[$k] > $blocation2[$i] &&
$seqstart2[$j] - $seqend2[$i] <= 10 && $seqstart2[$k] - $seqend2[$j] <= 10) {


    print
"$chr2[$i]\t$seqstart2[$i]\t$seqend2[$i]\t$strand2[$i]\t$type2[$i]\t$alocation2[$i]\t
$blocation2[$i]\t$seqstart2[$j]\t$seqend2[$j]\t$strand2[$j]\t$type2[$j]\t$alocation2[
$j]\t$blocation2[$j]\n$seqstart2[$k]\t$seqend2[$k]\t$strand2[$k]\t$type2[$k]\t$aloc
ation2[$k]\t$blocation2[$k]\n";


} elsif  ($type2[$i] =~ /^$ltr0$/ && $type2[$j] =~ /^$dna0$/ && $type2[$k] =~
/^$ltr0$/ && $strand2[$i] =~ $strand2[$k] && $alocation2[$k] > $blocation2[$i] &&
$seqstart2[$j] - $seqend2[$i] <= 10 && $seqstart2[$k] - $seqend2[$j] <= 10) {
```

```
    print
"$chr2[$i]\t$seqstart2[$i]\t$seqend2[$i]\t$strand2[$i]\t$type2[$i]\t$alocation2[$i]\t
$blocation2[$i]\t$seqstart2[$j]\t$seqend2[$j]\t$strand2[$j]\t$type2[$j]\t$alocation2[
$j]\t$blocation2[$j]\n$seqstart2[$k]\t$seqend2[$k]\t$strand2[$k]\t$type2[$k]\t$aloc
ation2[$k]\t$blocation2[$k]\n";


} elsif  ($type2[$i] =~ /^$ltr0$/ && $type2[$j] =~ /^$line0$/ && $type2[$k] =~
/^$dna0$/ && $strand2[$i] =~ $strand2[$k] && $alocation2[$k] > $blocation2[$i] &&
$seqstart2[$j] - $seqend2[$i] <= 10 && $seqstart2[$k] - $seqend2[$j] <= 10) {


    print
"$chr2[$i]\t$seqstart2[$i]\t$seqend2[$i]\t$strand2[$i]\t$type2[$i]\t$alocation2[$i]\t
$blocation2[$i]\t$seqstart2[$j]\t$seqend2[$j]\t$strand2[$j]\t$type2[$j]\t$alocation2[
$j]\t$blocation2[$j]\n$seqstart2[$k]\t$seqend2[$k]\t$strand2[$k]\t$type2[$k]\t$aloc
ation2[$k]\t$blocation2[$k]\n";


}

}
```

2.5 SBRR_discovery.pl

Description

This script searches potential bipartite recombinant repeats that created via insertion event. First we insert the TE (TE A) name that may inserted by another TE (TE B) name. It first examined TE A that flanked with TE B in RM output. It then examined the bipartite RR 500bp flanking region did not contained any TE. If the bipartite was clear of other TE in 500bp region, the script produced a new custom BED file.

```
#!/usr/bin/perl


use strict;


my $f1 = $ARGV[0];
```

```
my @file;
my $dna0 = ‘TE A’;
my $line0 = ‘TE B’;
my $chr;
my $seqstart;
my $seqend;
my $strand;
my $type;
my $alocation;
my $blocation;
my @chr2;
my @seqstart2;
my @seqend2;
my @strand2;
my @type2;
my @alocation2;
my @blocation2;




open (FILE1, "<$f1")
   ||die "cannot open";

while ($_ = <FILE1>) {
   chomp;
   @file = split (' ', $_);
   $chr = $file[0];
   $seqstart = $file[1];
```

```
    $seqend = $file[2];
    $strand = $file[3];
    $type = $file[4];
    $alocation = $file[5];
    $blocation = $file[6];

    push(@chr2, $chr);
    push(@seqstart2, $seqstart);
    push(@seqend2, $seqend);
    push(@strand2, $strand);
    push(@type2, $type);
    push(@alocation2, $alocation);
    push(@blocation2, $blocation);

}

my $number = @chr2;
my $i;
my $g;
my $h;
my $j;
my $k;
my $l;
my $m;

for ($i = 0; $i < $number; $i++) {

    $h = $i - 1;
    $j = $i + 1;
```

```
    if ($type2[$i] =~ /^$dna0$/ && $type2[$j] =~ /^$line0$/ &&  $seqend2[$h] -
$seqstart2[$i] >= 500 && $seqstart[$k] - $seqend2[$j] >= 500 && $seqstart2[$j] -
$seqend2[$i] <=10) {


    print
"$chr2[$i]\t$seqstart2[$i]\t$seqend2[$i]\t$strand2[$i]\t$type2[$i]\t$alocation2[$i]\t
$blocation2[$i]\t$seqstart2[$j]\t$seqend2[$j]\t$strand2[$j]\t$type2[$j]\t$alocation2[
$j]\t$blocation2[$j]\n";


} elsif  ($type2[$i] =~ /^$line0$/ && $type2[$j] =~ /^$dna0$/ &&  $seqend2[$h] -
$seqstart2[$i] >= 500 && $seqstart[$k] - $seqend2[$j] >= 500) {


    print
"$chr2[$i]\t$seqstart2[$i]\t$seqend2[$i]\t$strand2[$i]\t$type2[$i]\t$alocation2[$i]\t
$blocation2[$i]\t$seqstart2[$j]\t$seqend2[$j]\t$strand2[$j]\t$type2[$j]\t$alocation2[
$j]\t$blocation2[$j]\n";

}

}
```

2.6 alternate_SBRR_matching.pl

Description

This script searches potential bipartite recombinant repeats that interrupted another TE through insertion in repeat-rich region. First we searched a pair of TE fragments that contain two TE sequences (TE A and TE B). If the TE A and B were flanked next to each other, the scripts identified it as bipartite and produce a new custom BED file.

```
#!/usr/bin/perl


use strict;
```

```
my $f1 = $ARGV[0];
my @file;
my $dna0 = ‘TE A’;
my $line0 = ‘TE B’;
my $chr;
my $seqstart;
my $seqend;
my $strand;
my $type;
my $alocation;
my $blocation;
my @chr2;
my @seqstart2;
my @seqend2;
my @strand2;
my @type2;
my @alocation2;
my @blocation2;




open (FILE1, "<$f1")
   ||die "cannot open";

while ($_ = <FILE1>) {
   chomp;
   @file = split (' ', $_);
   $chr = $file[0];
```

```
    $seqstart = $file[1];
    $seqend = $file[2];
    $strand = $file[3];
    $type = $file[4];
    $alocation = $file[5];
    $blocation = $file[6];

    push(@chr2, $chr);
    push(@seqstart2, $seqstart);
    push(@seqend2, $seqend);
    push(@strand2, $strand);
    push(@type2, $type);
    push(@alocation2, $alocation);
    push(@blocation2, $blocation);

}

my $number = @chr2;
my $i;
my $g;
my $h;
my $j;
my $k;
my $l;
my $m;

for ($i = 0; $i < $number; $i++) {

    $h = $i - 1;
```

```
    $j = $i + 1;

    $k = $i + 2;


    if ($type2[$i] =~ /^$dna0$/ && $type2[$j] =~ /^$line0$/ && $type2[h] =~ $type2[k]
&& $blocation2[$h] > $alocation2[$k] &&

 $seqstart2[$j] - $seqend2[$i] <=10 && $seqend2[$h] - $seqstart2[$i] <= 10 &&
$seqstart[$k] - $seqend2[$j] <= 10) {


    print
"$chr2[$i]\t$seqstart2[$i]\t$seqend2[$i]\t$strand2[$i]\t$type2[$i]\t$alocation2[$i]\t
$blocation2[$i]\t$seqstart2[$j]\t$seqend2[$j]\t$strand2[$j]\t$type2[$j]\t$alocation2[
$j]\t$blocation2[$j]\n";


} elsif  $type2[$i] =~ /^$dna0$/ && $type2[$j] =~ /^$line0$/ && $type2[h] =~
$type2[k] && $alocation2[$k] > $blocation2[$h] &&

 $seqstart2[$j] - $seqend2[$i] <=10 && $seqend2[$h] - $seqstart2[$i] <= 10 &&
$seqstart[$k] - $seqend2[$j] <= 10) {


    print
"$chr2[$i]\t$seqstart2[$i]\t$seqend2[$i]\t$strand2[$i]\t$type2[$i]\t$alocation2[$i]\t
$blocation2[$i]\t$seqstart2[$j]\t$seqend2[$j]\t$strand2[$j]\t$type2[$j]\t$alocation2[
$j]\t$blocation2[$j]\n";


}


    elsif ($type2[$i] =~ /^$line0$/ && $type2[$j] =~ /^$dna0$/ && $type2[h] =~
$type2[k] && $blocation2[$h] > $alocation2[$k] &&

 $seqstart2[$j] - $seqend2[$i] <=10 && $seqend2[$h] - $seqstart2[$i] <= 10 &&
$seqstart[$k] - $seqend2[$j] <= 10) {


    print
"$chr2[$i]\t$seqstart2[$i]\t$seqend2[$i]\t$strand2[$i]\t$type2[$i]\t$alocation2[$i]\t
$blocation2[$i]\t$seqstart2[$j]\t$seqend2[$j]\t$strand2[$j]\t$type2[$j]\t$alocation2[
$j]\t$blocation2[$j]\n";
```

```
} elsif  $type2[$i] =~ /^$line0$/ && $type2[$j] =~ /^$dna0$/ && $type2[h] =~
$type2[k] && $alocation2[$k] > $blocation2[$h] &&

 $seqstart2[$j] - $seqend2[$i] <= 10 && $seqend2[$h] - $seqstart2[$i] <= 10 &&
$seqstart[$k] - $seqend2[$j] <= 10) {



    print
"$chr2[$i]\t$seqstart2[$i]\t$seqend2[$i]\t$strand2[$i]\t$type2[$i]\t$alocation2[$i]\t
$blocation2[$i]\t$seqstart2[$j]\t$seqend2[$j]\t$strand2[$j]\t$type2[$j]\t$alocation2[
$j]\t$blocation2[$j]\n";



}

}
```

2.7 mapping_SBRR.pl

Description

This script searches potential bipartite recombinant in RM output file that did not detected by Section 2.5 and 2.6 analyses. First we used the bipartite from Section 2.5 and 2.5 as query to match the other undiscovered RR in the RM file. If the script found matches, it identified it as bipartite and produced a new custom BED file. The bipartite that overlapped with Section 2.5 and 2.6 results was discarded.

```
#!/usr/bin/perl


use strict;


my $f1 = $ARGV[0];

my $f2 = $ARGV[1];

my $f3 = $ARGV[2];

my @file;

my @file2;

my $chr;
```

```
my $seqstart;
my $seqend;
my $strand;
my $type;
my $alocation;
my $blocation;
my @chr2;
my @seqstart2;
my @seqend2;
my @strand2;
my @type2;
my @alocation2;
my @blocation2;
my @chr2;
my @seqstart2;
my @seqend2;
my @strand2;
my @type2;
my @alocation2;
my @blocation2;


my $chr10;
my $seqstart10;
my $seqend10;
my $strand10;
my $type10;
my $alocation10;
my $blocation10;
```

```
my $aseqstart10;
my $aseqend10;
my $astrand10;
my $atype10;
my $alocationa10;
my $blocationa10;
my @chr20;
my @seqstart20;
my @seqend20;
my @strand20;
my @type20;
my @alocation20;
my @blocation20;
my @aseqstart20;
my @aseqend20;
my @astrand20;
my @atype20;
my @alocationa20;
my @blocationa20;

open (FILE1, "<$f1")
   ||die "cannot open";

while ($_ = <FILE1>) {
   chomp;
   @file = split (' ', $_);
   $chr = $file[0];
   $seqstart = $file[1];
   $seqend = $file[2];
```

```
    $strand = $file[3];
    $type = $file[4];
    $alocation = $file[5];
    $blocation = $file[6];
    push(@chr2, $chr);
    push(@seqstart2, $seqstart);
    push(@seqend2, $seqend);
    push(@strand2, $strand);
    push(@type2, $type);
    push(@alocation2, $alocation);
    push(@blocation2, $blocation);
}

open (FILE2, "<$f2")
    ||die "cannot open";

while ($_ = <FILE2>) {
    chomp;
    @file2 = split (' ', $_);
    $chr10 = $file2[0];
    $seqstart10 = $file2[1];
    $seqend10 = $file2[2];
    $strand10 = $file2[3];
    $type10 = $file2[4];
    $alocation10 = $file2[5];
    $blocation10 = $file2[6];
    $aseqstart10 = $file2[7];
    $aseqend10 = $file2[8];
    $astrand10 = $file2[9];
```

```
    $atype10 = $file2[10];
    $alocationa10 = $file2[11];
    $blocationa10 = $file2[12];
    push(@chr20, $chr10);
    push(@seqstart20, $seqstart10);
    push(@seqend20, $seqend10);
    push(@strand20, $strand10);
    push(@type20, $type10);
    push(@alocation20, $alocation10);
    push(@blocation20, $blocation10);
    push(@aseqstart20, $aseqstart10);
    push(@aseqend20, $aseqend10);
    push(@astrand20, $astrand10);
    push(@atype20, $atype10);
    push(@alocationa20, $alocationa10);
    push(@blocationa20, $blocationa10);
}



my $number = @chr2;
my $i;
my $g;
my $h;
my $j;
my $k;
my $l;

for ($i = 0; $i < $number; $i++) {
```

```
    $g = $i - 2;

    $h = $i - 1;

    $j = $i + 1;

    $k = $i + 2;

    $l = $i + 3;


    if ($type2[$i] =~ /^$type20[0]$/ && $type2[$j] =~ /^$atype20[0]$/ &&
$blocation20[0] - $blocation2[$i] <= 10 && $blocation20[0] - $blocation2[$i] >= -10 &&
$alocationa20[0] - $alocation2[$j] >= -10 && $alocationa20[0] - $alocation2[$j] <= 10
&& $strand2[$i] =~ /$strand20[0]/ && $strand2[$j] =~ /$astrand20[0]/ &&
$type2[$h] !~ /^$type20[0]$/ && $type2[$g] !~ /^$type20[0]$/ && $type2[$k] !~
/^$type20[0]$/ && $type2[$l] !~ /^$type20[0]$/ && $type2[$h] !~ /^$atype20[0]$/
&& $type2[$g] !~ /^$atype20[0]$/ && $type2[$k] !~ /^$atype20[0]$/ &&
$type2[$l] !~ /^$atype20[0]$/) {


open (FILE3, ">>$f3")

   ||die "cannot open";

   print FILE3
"$chr2[$i]\t$seqstart2[$i]\t$seqend2[$i]\t$strand2[$i]\t$type2[$i]\t$alocation2[$i]\t
$blocation2[$i]\t$seqstart2[$j]\t$seqend2[$j]\t$strand2[$j]\t$type2[$j]\t$alocation2[
$j]\t$blocation2[$j]\n";


} elsif ($type2[$i] =~ /^$atype20[0]$/ && $type2[$j] =~ /^$type20[0]$/ &&
$blocation20[0] - $alocation2[$j] <= 10 && $blocation20[0] - $alocation2[$j] >= -10  &&
$alocationa20[0] - $blocation2[$i] >= -10 && $alocationa20[0] - $blocation2[$i] <= 10
&& $strand2[$j] !~ /$strand20[0]/ && $strand2[$i] !~ /$astrand20[0]/ &&
$type2[$h] !~ /^$type20[0]$/ && $type2[$g] !~ /^$type20[0]$/ && $type2[$k] !~
/^$type20[0]$/ && $type2[$l] !~ /^$type20[0]$/ && $type2[$h] !~ /^$atype20[0]$/
&& $type2[$g] !~ /^$atype20[0]$/ && $type2[$k] !~ /^$atype20[0]$/ &&
$type2[$l] !~ /^$atype20[0]$/) {


open (FILE3, ">>$f3")

   ||die "cannot open";

   print FILE3
"$chr2[$i]\t$seqstart2[$i]\t$seqend2[$i]\t$strand2[$i]\t$type2[$i]\t$alocation2[$i]\t
$blocation2[$i]\t$seqstart2[$j]\t$seqend2[$j]\t$strand2[$j]\t$type2[$j]\t$alocation2[
$j]\t$blocation2[$j]\n";
```

}

}

2.8 Intersecting Bipartite and tripartite RR results with BEDTools

Description

This script is the general process used to discard the bipartite recombinant (section 2.5, 2.6) repeats that might originate from tripartite recombinant repeats shown in Section 2.3, 2.4.

IntersectBed –a Bipartite.bed –b Tripartite.bed –u >> non-overlapped_Bipartite.bed

2.9 STRR_cluster.pl

Description

This script was used to cluster the tripartite recombinant repeat into families based on the RR structure. First, random single RR sequence (RR A) was picked from the tripartite custom BED file, and then it was compared with the rest of the tripartite sequence in the file. It had decent match, the sequences was extracted and cluster into new custom BED file. After that, a random RR sequence was picked again in the left-over sequences BED file and compared again for clustering process (The looping process was described in Section 2.10).

#!/usr/bin/perl

use strict;

```
my $f1 = $ARGV[0];
my $f2 = $ARGV[1];
my $f3 = $ARGV[2];
my $f4 = $ARGV[3];
my @file;
my $chr;
my $seqstart;
my $seqend;
my $strand;
my $type;
my $alocation;
my $blocation;
my $aseqstart;
my $aseqend;
my $astrand;
my $atype;
my $alocationa;
my $blocationa;
my $bseqstart;
my $bseqend;
my $bstrand;
my $btype;
my $alocationb;
my $blocationb;
my @chr2;
my @seqstart2;
my @seqend2;
my @strand2;
my @type2;
```

```
my @alocation2;
my @blocation2;
my @aseqstart2;
my @aseqend2;
my @astrand2;
my @atype2;
my @alocationa2;
my @blocationa2;
my @bseqstart2;
my @bseqend2;
my @bstrand2;
my @btype2;
my @alocationb2;
my @blocationb2;

my @file2;
my $chr10;
my $seqstart10;
my $seqend10;
my $strand10;
my $type10;
my $alocation10;
my $blocation10;
my $aseqstart10;
my $aseqend10;
my $astrand10;
my $atype10;
my $alocationa10;
my $blocationa10;
```

```
my $bseqstart10;
my $bseqend10;
my $bstrand10;
my $btype10;
my $alocationb10;
my $blocationb10;
my @chr20;
my @seqstart20;
my @seqend20;
my @strand20;
my @type20;
my @alocation20;
my @blocation20;
my @aseqstart20;
my @aseqend20;
my @astrand20;
my @atype20;
my @alocationa20;
my @blocationa20;
my @bseqstart20;
my @bseqend20;
my @bstrand20;
my @btype20;
my @alocationb20;
my @blocationb20;



#Comment: $f1 corresponding to the total tripartite recombinant repeat custom BED file
```

```
open (FILE1, "<$f1")
   ||die "cannot open";

while ($_ = <FILE1>) {
   chomp;
   @file = split (' ', $_);
   $chr = $file[0];
   $seqstart = $file[1];
   $seqend = $file[2];
   $strand = $file[3];
   $type = $file[4];
   $alocation = $file[5];
   $blocation = $file[6];
   $aseqstart = $file[7];
   $aseqend = $file[8];
   $astrand = $file[9];
   $atype = $file[10];
   $alocationa = $file[11];
   $blocationa = $file[12];
   $bseqstart = $file[13];
   $bseqend = $file[14];
   $bstrand = $file[15];
   $btype = $file[16];
   $alocationb = $file[17];
   $blocationb = $file[18];
   push(@chr2, $chr);
   push(@seqstart2, $seqstart);
   push(@seqend2, $seqend);
   push(@strand2, $strand);
```

```
    push(@type2, $type);
    push(@alocation2, $alocation);
    push(@blocation2, $blocation);
    push(@aseqstart2, $aseqstart);
    push(@aseqend2, $aseqend);
    push(@astrand2, $astrand);
    push(@atype2, $atype);
    push(@alocationa2, $alocationa);
    push(@blocationa2, $blocationa);
    push(@bseqstart2, $bseqstart);
    push(@bseqend2, $bseqend);
    push(@bstrand2, $bstrand);
    push(@btype2, $btype);
    push(@alocationb2, $alocationb);
    push(@blocationb2, $blocationb);
}

#Comment: $f2 corresponding to the randomly picked tripartite recombinant repeat
custom BED file used for clustering process
open (FILE2, "<$f2")
    ||die "cannot open";

while ($_ = <FILE2>) {
    chomp;
    @file2 = split (' ', $_);
    $chr10 = $file2[0];
    $seqstart10 = $file2[1];
    $seqend10 = $file2[2];
    $strand10 = $file2[3];
```

```
$type10 = $file2[4];
$alocation10 = $file2[5];
$blocation10 = $file2[6];
$aseqstart10 = $file2[7];
$aseqend10 = $file2[8];
$astrand10 = $file2[9];
$atype10 = $file2[10];
$alocationa10 = $file2[11];
$blocationa10 = $file2[12];
$bseqstart10 = $file2[13];
$bseqend10 = $file2[14];
$bstrand10 = $file2[15];
$btype10 = $file2[16];
$alocationb10 = $file2[17];
$blocationb10 = $file2[18];
push(@chr20, $chr10);
push(@seqstart20, $seqstart10);
push(@seqend20, $seqend10);
push(@strand20, $strand10);
push(@type20, $type10);
push(@alocation20, $alocation10);
push(@blocation20, $blocation10);
push(@aseqstart20, $aseqstart10);
push(@aseqend20, $aseqend10);
push(@astrand20, $astrand10);
push(@atype20, $atype10);
push(@alocationa20, $alocationa10);
push(@blocationa20, $blocationa10);
push(@bseqstart20, $bseqstart10);
```

```
   push(@bseqend20, $bseqend10);
   push(@bstrand20, $bstrand10);
   push(@btype20, $btype10);
   push(@alocationb20, $alocationb10);
   push(@blocationb20, $blocationb10);
}

unlink $f1;
unlink $f2;

my $number = @chr2;
my $i;

for ($i = 0; $i < $number; $i++) {
    if ($type2[$i] =~ /^$type20[0]$/ && $atype2[$i] =~ /^$atype20[0]$/ && $btype2[$i]
=~ /^$btype20[0]$/ && $blocation20[0] - $blocation2[$i] <= 10 && $blocation20[0] -
$blocation2[$i] >= -10 && $alocationb20[0] - $alocationb2[$i] <= 10 &&
$alocationb20[0] - $alocationb2[$i] >= -10 && $alocationa20[0] - $alocationa2[$i] >= -
10 && $alocationa20[0] - $alocationa2[$i] <= 10 && $blocationa20[0] -
$blocationa2[$i] >= -10 && $blocationa20[0] - $blocationa2[$i] <= 10) {

open (FILE3, ">>$f3")
   ||die "cannot open";
   print FILE3
"$chr2[$i]\t$seqstart2[$i]\t$seqend2[$i]\t$strand2[$i]\t$type2[$i]\t$alocation2[$i]\t
$blocation2[$i]\t$aseqstart2[$i]\t$aseqend2[$i]\t$astrand2[$i]\t$atype2[$i]\t$alocati
ona2[$i]\t$blocationa2[$i]\t$bseqstart2[$i]\t$bseqend2[$i]\t$bstrand2[$i]\t$btype2[
$i]\t$alocationb2[$i]\t$blocationb2[$i]\n";

} elsif ($type2[$i] =~ /^$btype20[0]$/ && $atype2[$i] =~ /^$atype20[0]$/ &&
$btype2[$i] =~ /^$type20[0]$/ && $blocation20[0] - $alocationb2[$i] <= 10 &&
$blocation20[0] - $alocationb2[$i] >= -10 && $alocationb20[0] - $blocation2[$i] <= 10
&& $alocationb20[0] - $blocation2[$i] >= -10 && $alocationa20[0] - $blocationa2[$i] >=
-10 && $alocationa20[0] - $blocationa2[$i] <= 10 && $blocationa20[0] -
$alocationa2[$i] >= -10 && $blocationa20[0] - $alocationa2[$i] <= 10) {
```

```
open (FILE3, ">>$f3")

   ||die "cannot open";

   print FILE3
"$chr2[$i]\t$seqstart2[$i]\t$seqend2[$i]\t$strand2[$i]\t$type2[$i]\t$alocation2[$i]\t
$blocation2[$i]\t$aseqstart2[$i]\t$aseqend2[$i]\t$astrand2[$i]\t$atype2[$i]\t$alocati
ona2[$i]\t$blocationa2[$i]\t$bseqstart2[$i]\t$bseqend2[$i]\t$bstrand2[$i]\t$btype2[
$i]\t$alocationb2[$i]\t$blocationb2[$i]\n";


} else {


open (FILE4, ">>$f4")

   ||die "cannot open";

   print FILE4
"$chr2[$i]\t$seqstart2[$i]\t$seqend2[$i]\t$strand2[$i]\t$type2[$i]\t$alocation2[$i]\t
$blocation2[$i]\t$aseqstart2[$i]\t$aseqend2[$i]\t$astrand2[$i]\t$atype2[$i]\t$alocati
ona2[$i]\t$blocationa2[$i]\t$bseqstart2[$i]\t$bseqend2[$i]\t$bstrand2[$i]\t$btype2[
$i]\t$alocationb2[$i]\t$blocationb2[$i]\n";


}

}
```

2.10 SBRR_cluster.pl

Description

This script was used to cluster the bipartite recombinant repeat into families based on the RR structure. First, random single RR sequence (RR A) was picked from the bipartite custom BED file, and then it was compared with the rest of the tripartite sequence in the file. It had decent match, the sequences was extracted and cluster into new custom BED file. After that, a random RR sequence was picked again in the left-over sequences BED file and compared again for clustering process (The looping process was described in Section 2.10).

```
#!/usr/bin/perl
```

```
use strict;

my $f1 = $ARGV[0];
my $f2 = $ARGV[1];
my $f3 = $ARGV[2];
my $f4 = $ARGV[3];
my @file;
my @file2;

my $chr;
my $seqstart;
my $seqend;
my $strand;
my $type;
my $alocation;
my $blocation;
my $aseqstart;
my $aseqend;
my $astrand;
my $atype;
my $alocationa;
my $blocationa;
my @chr2;
my @seqstart2;
my @seqend2;
my @strand2;
my @type2;
my @alocation2;
```

```
my @blocation2;
my @aseqstart2;
my @aseqend2;
my @astrand2;
my @atype2;
my @alocationa2;
my @blocationa2;


my $chr10;
my $seqstart10;
my $seqend10;
my $strand10;
my $type10;
my $alocation10;
my $blocation10;
my $aseqstart10;
my $aseqend10;
my $astrand10;
my $atype10;
my $alocationa10;
my $blocationa10;
my @chr20;
my @seqstart20;
my @seqend20;
my @strand20;
my @type20;
my @alocation20;
my @blocation20;
```

```
my @aseqstart20;
my @aseqend20;
my @astrand20;
my @atype20;
my @alocationa20;
my @blocationa20;



#Comment: $f1 corresponding to the total tripartite recombinant repeat custom BED file
open (FILE1, "<$f1")
   ||die "cannot open";


while ($_ = <FILE1>) {
   chomp;
   @file = split (' ', $_);
   $chr = $file[0];
   $seqstart = $file[1];
   $seqend = $file[2];
   $strand = $file[3];
   $type = $file[4];
   $alocation = $file[5];
   $blocation = $file[6];
   $aseqstart = $file[7];
   $aseqend = $file[8];
   $astrand = $file[9];
   $atype = $file[10];
   $alocationa = $file[11];
   $blocationa = $file[12];
```

```
    push(@chr2, $chr);
    push(@seqstart2, $seqstart);
    push(@seqend2, $seqend);
    push(@strand2, $strand);
    push(@type2, $type);
    push(@alocation2, $alocation);
    push(@blocation2, $blocation);
    push(@aseqstart2, $aseqstart);
    push(@aseqend2, $aseqend);
    push(@astrand2, $astrand);
    push(@atype2, $atype);
    push(@alocationa2, $alocationa);
    push(@blocationa2, $blocationa);
}



#Comment: $f2 corresponding to the randomly picked tripartite recombinant repeat
custom BED file used for clustering process
open (FILE2, "<$f2")
    ||die "cannot open";


while ($_ = <FILE2>) {
    chomp;
    @file2 = split (' ', $_);
    $chr10 = $file2[0];
    $seqstart10 = $file2[1];
    $seqend10 = $file2[2];
    $strand10 = $file2[3];
    $type10 = $file2[4];
```

```
    $alocation10 = $file2[5];
    $blocation10 = $file2[6];
    $aseqstart10 = $file2[7];
    $aseqend10 = $file2[8];
    $astrand10 = $file2[9];
    $atype10 = $file2[10];
    $alocationa10 = $file2[11];
    $blocationa10 = $file2[12];
    push(@chr20, $chr10);
    push(@seqstart20, $seqstart10);
    push(@seqend20, $seqend10);
    push(@strand20, $strand10);
    push(@type20, $type10);
    push(@alocation20, $alocation10);
    push(@blocation20, $blocation10);
    push(@aseqstart20, $aseqstart10);
    push(@aseqend20, $aseqend10);
    push(@astrand20, $astrand10);
    push(@atype20, $atype10);
    push(@alocationa20, $alocationa10);
    push(@blocationa20, $blocationa10);

}

unlink $f1;
unlink $f2;

my $number = @chr2;
my $i;
```

```
for ($i = 0; $i < $number; $i++) {

    if ($type2[$i] =~ /^$type20[0]$/ && $atype2[$i] =~ /^$atype20[0]$/ &&
$blocation20[0] - $blocation2[$i] <= 10 && $blocation20[0] - $blocation2[$i] >= -10 &&
$alocationa20[0] - $alocationa2[$i] >= -10 && $alocationa20[0] - $alocationa2[$i] <= 10
&& $strand2[$i] =~ /$strand20[0]/ && $astrand2[$i] =~ /$astrand20[$0]/) {



open (FILE3, ">>$f3")

   ||die "cannot open";

   print FILE3
"$chr2[$i]\t$seqstart2[$i]\t$seqend2[$i]\t$strand2[$i]\t$type2[$i]\t$alocation2[$i]\t
$blocation2[$i]\t$aseqstart2[$i]\t$aseqend2[$i]\t$astrand2[$i]\t$atype2[$i]\t$alocati
ona2[$i]\t$blocationa2[$i]\n";



} elsif ($type2[$i] =~ /^$atype20[0]$/ && $atype2[$i] =~ /^$type20[0]$/ &&
$blocation20[0] - $alocationa2[$i] <= 10 && $blocation20[0] - $alocationa2[$i] >= -10
&& $alocationa20[0] - $blocation2[$i] >= -10 && $alocationa20[0] - $blocation2[$i] <=
10 && $astrand2[$i] !~ /$strand20[0]/ && $strand2[$i] !~ /$astrand20[$0]/) {



open (FILE3, ">>$f3")

   ||die "cannot open";

   print FILE3
"$chr2[$i]\t$seqstart2[$i]\t$seqend2[$i]\t$strand2[$i]\t$type2[$i]\t$alocation2[$i]\t
$blocation2[$i]\t$aseqstart2[$i]\t$aseqend2[$i]\t$astrand2[$i]\t$atype2[$i]\t$alocati
ona2[$i]\t$blocationa2[$i]\n";



} else {



open (FILE4, ">>$f4")

   ||die "cannot open";

   print FILE4
"$chr2[$i]\t$seqstart2[$i]\t$seqend2[$i]\t$strand2[$i]\t$type2[$i]\t$alocation2[$i]\t
$blocation2[$i]\t$aseqstart2[$i]\t$aseqend2[$i]\t$astrand2[$i]\t$atype2[$i]\t$alocati
ona2[$i]\t$blocationa2[$i]\n";
```

}

}

2.11 clustering_looping.pl

Description

This script was used to start the looping process of clustering recombinant repeat into families. It produced a script to use for Section 2.9, 2.8 clustering process by using 'sh' command.

```
#!/usr/bin/perl

use strict;

my $f1 = $ARGV[0];
my @file;
my $chr;
my $seqstart;
my $seqend;
my $strand;
my $type;
my $alocation;
my $blocation;
my $aseqstart;
my $aseqend;
my $astrand;
my $atype;
my $alocationa;
my $blocationa;
my $bseqstart;
```

```
my $bseqend;
my $bstrand;
my $btype;
my $alocationb;
my $blocationb;

my @chr2;
my @seqstart2;
my @seqend2;
my @strand2;
my @type2;
my @alocation2;
my @blocation2;
my @aseqstart2;
my @aseqend2;
my @astrand2;
my @atype2;
my @alocationa2;
my @blocationa2;
my @bseqstart2;
my @bseqend2;
my @bstrand2;
my @btype2;
my @alocationb2;
my @blocationb2;



open (FILE1, "<$f1")
   ||die "cannot open";
```

```
while ($_ = <FILE1>) {
    chomp;
    @file = split (' ', $_);
    $chr = $file[0];
    $seqstart = $file[1];
    $seqend = $file[2];
    $strand = $file[3];
    $type = $file[4];
    $alocation = $file[5];
    $blocation = $file[6];
    $aseqstart = $file[7];
    $aseqend = $file[8];
    $astrand = $file[9];
    $atype = $file[10];
    $alocationa = $file[11];
    $blocationa = $file[12];
    $bseqstart = $file[13];
    $bseqend = $file[14];
    $bstrand = $file[15];
    $btype = $file[16];
    $alocationb = $file[17];
    $blocationb = $file[18];
    push(@chr2, $chr);
    push(@seqstart2, $seqstart);
    push(@seqend2, $seqend);
    push(@strand2, $strand);
    push(@type2, $type);
    push(@alocation2, $alocation);
```

```
    push(@blocation2, $blocation);
    push(@aseqstart2, $aseqstart);
    push(@aseqend2, $aseqend);
    push(@astrand2, $astrand);
    push(@atype2, $atype);
    push(@alocationa2, $alocationa);
    push(@blocationa2, $blocationa);
    push(@bseqstart2, $bseqstart);
    push(@bseqend2, $bseqend);
    push(@bstrand2, $bstrand);
    push(@btype2, $btype);
    push(@alocationb2, $alocationb);
    push(@blocationb2, $blocationb);
}

my $number = @chr2;
my $i;

for ($i = 0; $i < $number; $i++) {

    print "perl SBRR_cluster.pl combination_tripartite.out $i.testing.file.out
$i.tripartite.group.out combination_tripartite.out\n";

}
```

2.12 Copy number counting (GREP) command

Description

This script was used to calculate the copy number of sequences in each family custom BED file via GREP command. If the copy number is less than 3, the family was discarded.

Grep –c '.' Tripartite_family_A.bed >> Tripartite_family_copy_number.txt

References

1. Smit A, Hubley, R & Green, P. (1996-2010) RepeatMasker Open-3.0. http://wwwrepeatmaskerorg.

2. http://www.perl.org/.

3. http://www.postgresql.org/.

4. http://office.microsoft.com/en-au/excel/.

5. http://www.r-project.org/.

6. http://www.ncbi.nlm.nih.gov/IEB/ToolBox/SDKDOCS/INDEX.HTML.

7. Terai G, Yoshizawa A, Okida H, Asai K, Mituyama T (2010) Discovery of short pseudogenes derived from messenger RNAs. Nucleic acids research 38: 1163-1171.

8. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841-842.

Supplementary information of Chapter 3: Segmental Duplication Events Can Be Detected Using Repetitive Elements.

Supplementary Information 1: The Detail Complex Recombinant Repeat Identification and Clustering Process

Content

1. Method

1.1 Software Used

For repetitive elements annotation process, RepeatMasker version open-3.2.6 [1] was used. Krishna was used for sequence alignments and searches [2]. The sequence clustering process was done with PILER version 1.0 [3]. The fastacmd version 2.2.19 was used to extract specific genomic sequences from genome assembly [4]. The scripts were written in Perl (Perl version 5.16.2) [5]. BEDTools version 2.11.2 were used to manipulate genomic intervals [6]. Lastz The software is open source and they could be obtained from different websites as shown in Table 1 [7].

Table 1. The software version details and downloads links.

Software	Version	Download Website
BEDTools	2.11.2	http://code.google.com/p/bedtools/
Krishna	1.0	http://godoc.org/code.google.com/p/biogo.examples/krishna
PALS and PILER	1.0	http://www.drive5.com/piler/
PERL	5.16.2	http://www.perl.org/
LASTZ	1.02	www.bx.psu.edu/~rsharris/lastz/
RepeatMasker	3.2.6	http://www.repeatmasker.org/
Fastacmd	2.2.19	http://www.ncbi.nlm.nih.gov/IEB/ToolBox/SDKDOCS/INDEX.HTML

1.2 Pipeline Overview

Species where full genome data was available were searched for complex recombinant repeats. Scripts and software settings shown in Section 2 were used to generate complex recombinant repeat families for each species. A flow chart showing the pipeline for the analysis is shown in Figure 1.

First, the RepeatMasker, shown in Section 2.1, was used to annotate the species-specific repetitive elements in different mammalian genomes. The species-specific repetitive element database was obtained from Repbase (http://www.girinst.org/repbase/).

Once the repeatmasker output results were generated, the rm2bed.pl script, shown in Section 2.2, was used to convert the output format into Bed format. Grep unix command was used to extract specific TE classes from the BED file and generate into TE-specific BED file ('TE A' Bed file as an example).

The Fragment_recovery.pl and Fragment_recovery_looping scripts, as shown in Section 2.3, 2.4 and 2.5, were then used to search the TE A sequences that have been broken into 3~10 fragments due to possible insertion or deletion processes in the genome as shown in figure 2. The fragment recovery processes were repeated with other TE BED files.

Once those fragmented TE sequences recovered, the intact sequences were extracted from genome with fastacmd software, as shown in Section 2.6. These potential CRR fasta sequences were then compile into single fasta file.

The Krishna, shown in Section 2.7, was used to search the possible CRR sequence alignments. Piler, as shown in Section 2.8, was used to cluster the aligned sequences into families. Meanwhile, the RepeatMasker, as shown in Section 2.9, was used to annotate the repetitive elements in the clustered sequences.

Once the clustering process finished, the CRR family contents were examined by manually to remove false-positive result. If the CRR family contained > 3 repetitive elements and they were not derived from single type of TE class (i.e. LINE), it was confirmed as CRR family that exist in genome. The other families that did not match the criteria were discarded.

Figure 1. The flowchart of Complex recombinant repeat family Identification.

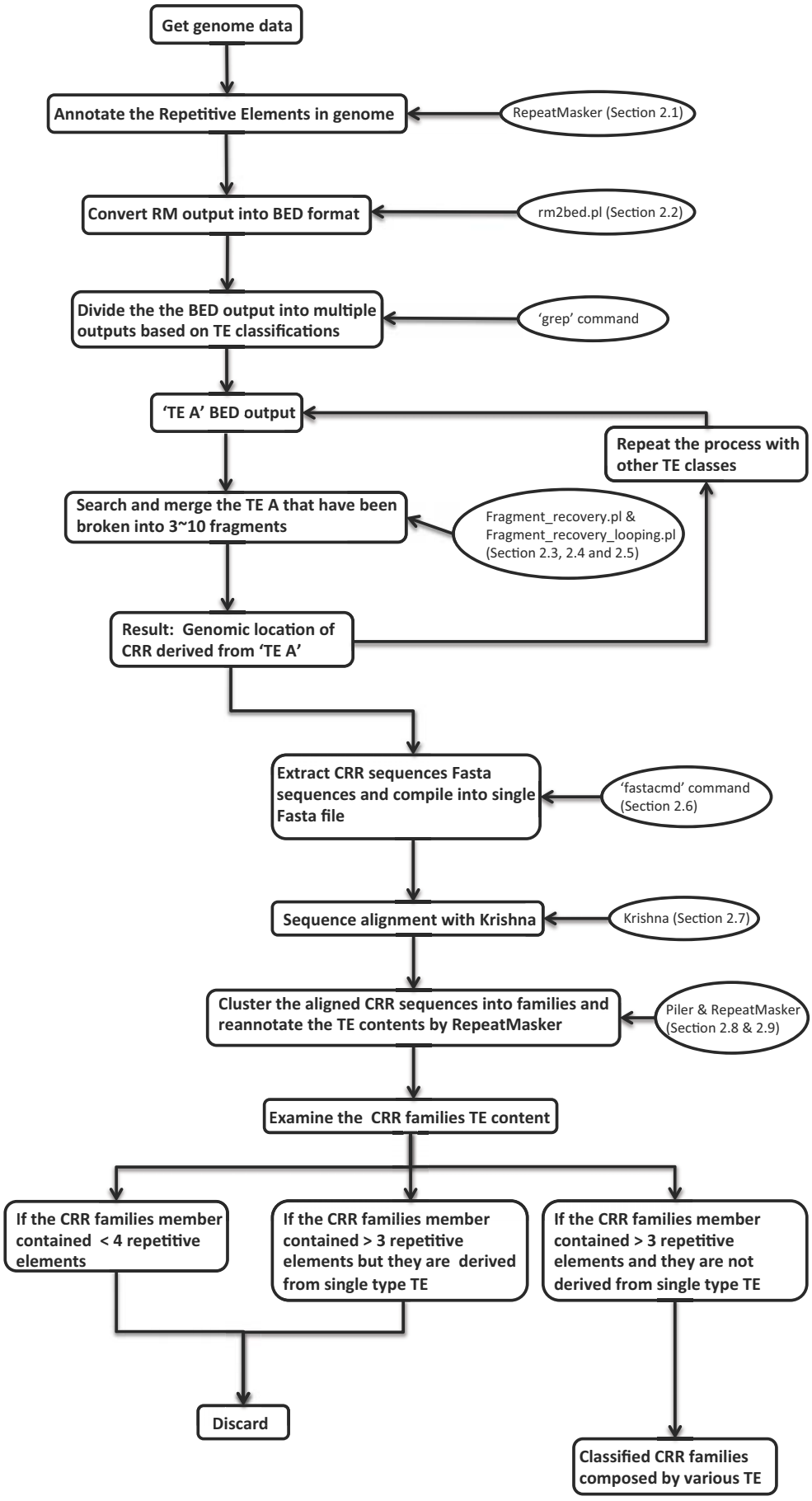
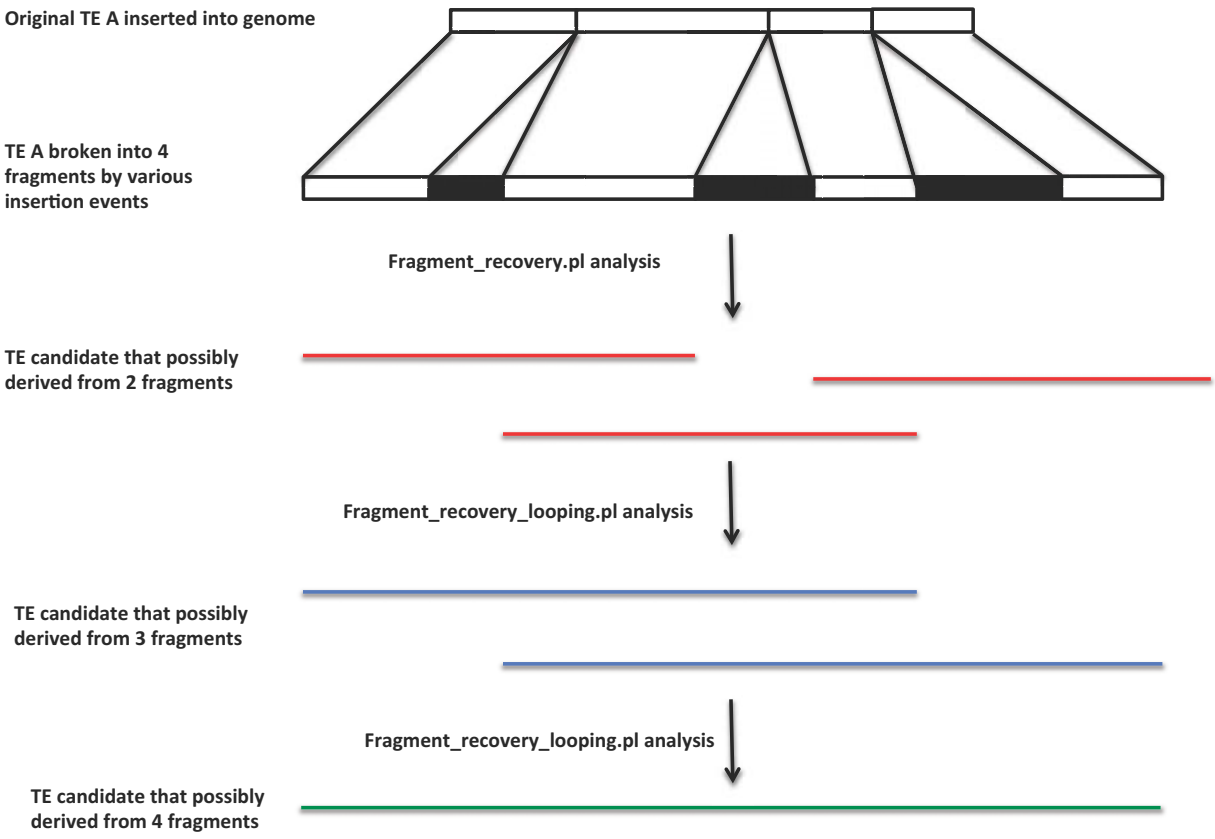

Get genome data
Annotate the Repetitive Elements in genome
RepeatMasker (Section 2.1)
Convert RM output into BED format
rm2bed.pl (Section 2.2)
Divide the the BED output into multiple outputs based on TE classifications
'grep' command
'TE A' BED output
Repeat the process with other TE classes
Search and merge the TE A that have been broken into 3~10 fragments
Fragment_recovery.pl & Fragment_recovery_looping.pl (Section 2.3, 2.4 and 2.5)
Result: Genomic location of CRR derived from 'TE A'
Extract CRR sequences Fasta sequences and compile into single Fasta file
'fastacmd' command (Section 2.6)
Sequence alignment with Krishna
Krishna (Section 2.7)
Cluster the aligned CRR sequences into families and reannotate the TE contents by RepeatMasker
Piler & RepeatMasker (Section 2.8 & 2.9)
Examine the CRR families TE content
If the CRR families member contained < 4 repetitive elements
If the CRR families member contained > 3 repetitive elements but they are derived from single type TE
If the CRR families member contained > 3 repetitive elements and they are not derived from single type TE
Discard
Classified CRR families composed by various TE

Figure 2. The overview of searching potential Complex recombinant repeats in the genome as shown in section 2.3, 2.4 and 2.5.



2. Script

2.1 RepeatMasker

Description

The genome repetitive elements can be annotated using RepeatMasker as shown below:

RepeatMasker -s -pa 6 -lib human_repetitive_element_library.fasta hg19_genome_assembly.fa

2.2 rm2bed.pl

Description

This script converts the RepeatMasker output into BED file format. The BED file contains: chromosome, sequence start location, sequence end location, matched repetitive element strand (+ for positive strand and C for negative strand), matched repetitive element types, matched repetitive element sequence start location and matched repetitive element sequence end location.

```
#!/usr/bin/perl

use strict;


my $f1 = $ARGV[0];

my @file;

my $chr;

my $seqstart;

my $seqend;

my $strand;

my $type;

my $alocation;
```

```
my $blocation;
my $clocation;
my @chr2;
my @seqstart2;
my @seqend2;
my @strand2;
my @type2;
my @alocation2;
my @blocation2;
my @clocation2;

open (FILE1, "<$f1")
   ||die "cannot open";

while ($_ = <FILE1>) {
   chomp;
   @file = split (' ', $_);
   $chr = $file[4];
   $seqstart = $file[5];
   $seqend = $file[6];
   $strand = $file[8];
   $type = $file[9];
   $alocation = $file[11];
   $blocation = $file[12];
   $clocation = $file[13];
   push(@chr2, $chr);
   push(@seqstart2, $seqstart);
   push(@seqend2, $seqend);
   push(@strand2, $strand);
```

```
    push(@type2, $type);
    push(@alocation2, $alocation);
    push(@blocation2, $blocation);
    push(@clocation2, $clocation);
}

my $number = @chr2;
my $i;

for ($i = 0; $i < $number; $i++) {
    if ($strand2[$i] =~ /\+/) {
    print "$chr2[$i]\t$seqstart2[$i]\t$seqend2[$i]\t$strand2[$i]\t$type2[$i]\t$alocation2[$i]\t$blocation2[$i]\n";
} else {

    print "$chr2[$i]\t$seqstart2[$i]\t$seqend2[$i]\t$strand2[$i]\t$type2[$i]\t$blocation2[$i]\t$clocation2[$i]\n";

}
}
```

2.3 Fragment_recovery.pl

Description

This script searches potential pieces of TE that have been broken into 2 fragments due to insertion or other events in the genome. It first examined the pair of fragments are located in same chromosome and same strand. It then examined the pair of fragments were derived from a single TE or not based on the fragments' TE data (matched repetitive sequence start and sequence end location). If the pair of fragments were derived from the single TE, the script will produced a new BED output and further processed by the Fragment_recovery_looping.pl.

```
#!/usr/bin/perl
use strict;

my $f1 = $ARGV[0];
my $f2 = $ARGV[1];
my $f3 = $ARGV[2];
my @file;
my $chr;
my $seqstart;
my $seqend;
my $strand;
my $type;
my $alocation;
my $blocation;
my $clocation;
my @chr2;
my @seqstart2;
my @seqend2;
```

```
my @strand2;
my @type2;
my @alocation2;
my @blocation2;
my @clocation2;

open (FILE1, "<$f1")
   ||die "cannot open";

while ($_ = <FILE1>) {
   chomp;
   @file = split (' ', $_);
   $chr = $file[0];
   $seqstart = $file[1];
   $seqend = $file[2];
   $strand = $file[3];
   $type = $file[4];
   $alocation = $file[5];
   $blocation = $file[6];
   push(@chr2, $chr);
   push(@seqstart2, $seqstart);
   push(@seqend2, $seqend);
   push(@strand2, $strand);
   push(@type2, $type);
   push(@alocation2, $alocation);
   push(@blocation2, $blocation);

}
```

```
my $number = @chr2;

my $i;

my $j;

my $k;


for ($i = 1; $i < $number; $i++) {

$k = $i - 1;


   if ($chr2[$i] =~ $chr2[$k] && $strand2[$i] =~ /P/ && $strand2[$i] =~
/^$strand2[$k]$/ && $blocation2[$i] > $blocation2[$k] && $blocation2[$i] >
$alocation2[$i] && $blocation2[$k] > $alocation2[$k] && $alocation2[$i] -
$blocation2[$k] >= -100 && $alocation2[$i] - $blocation2[$k] <= 3000 && $seqstart2[$i]
- $seqend2[$k] > 0 && $seqstart2[$i] - $seqend2[$k] < 20000) {


open (FILE2, ">>$f2")

   ||die "cannot open";

   print FILE2
"$chr2[$i]\t$seqstart2[$k]\t$seqend2[$i]\t$strand2[$i]\t$type2[$i]\t$alocation2[$k]\
t$blocation2[$i]\n";


} elsif ($chr2[$i] =~ $chr2[$k] && $strand2[$i] =~ /C/ && $strand2[$i] =~
/^$strand2[$k]$/ && $alocation2[$k] > $alocation2[$i] && $alocation2[$i] >
$blocation2[$i] && $alocation2[$k] > $blocation2[$k] && $blocation2[$k] -
$alocation2[$i] >= -100 && $blocation2[$k] - $alocation2[$i] <= 3000 && $seqstart2[$i]
- $seqend2[$k] > 0 && $seqstart2[$i] - $seqend2[$k] < 20000) {


open (FILE2, ">>$f2")

   ||die "cannot open";

   print FILE2
"$chr2[$i]\t$seqstart2[$k]\t$seqend2[$i]\t$strand2[$i]\t$type2[$i]\t$alocation2[$k]\
t$blocation2[$i]\n";


}

}
```

2.4 Fragment_recovery_looping.pl

Description

This script searches potential pieces of TE that have been broken into 3 fragments based on the BED output generated by Fragment_recovery_looping.pl. It first examined the pair of fragments are located in same chromosome and same strand. It then examined the pair of fragments were derived from 3 TE or not based on the fragments' TE data (matched repetitive sequence start and sequence end location). If the pair of fragments were derived from the single TE, the script will produced a new BED output. This script could be used to detect the TE that broken into 4~10 fragments based on the script described in 2.4.

```
#!/usr/bin/perl
use strict;

my $f1 = $ARGV[0];
my $f2 = $ARGV[1];
my $f3 = $ARGV[2];
my $f4 = $ARGV[3];
my @file;
my $chr;
my $seqstart;
my $seqend;
my $strand;
my $type;
my $alocation;
my $blocation;
my $clocation;
my @chr2;
my @seqstart2;
```

```
my @seqend2;
my @strand2;
my @type2;
my @alocation2;
my @blocation2;
my @clocation2;

open (FILE1, "<$f1")
   ||die "cannot open";

while ($_ = <FILE1>) {
   chomp;
   @file = split (' ', $_);
   $chr = $file[0];
   $seqstart = $file[1];
   $seqend = $file[2];
   $strand = $file[3];
   $type = $file[4];
   $alocation = $file[5];
   $blocation = $file[6];
   push(@chr2, $chr);
   push(@seqstart2, $seqstart);
   push(@seqend2, $seqend);
   push(@strand2, $strand);
   push(@type2, $type);
   push(@alocation2, $alocation);
   push(@blocation2, $blocation);

}
```

```
my $number = @chr2;

my $i;

my $j;

my $k;


   if ($seqstart2[1] - $seqend2[0] >= 20000) {


open (FILE3, ">>$f3")

   ||die "cannot open";

   print FILE3
"$chr2[0]\t$seqstart2[0]\t$seqend2[0]\t$strand2[0]\t$type2[0]\t$alocation2[0]\t$bl
ocation2[0]\n";

}


for ($i = 1; $i < $number; $i++) {

$k = $i - 1;

$j = $i + 1;


   if ($chr2[$i] =~ $chr2[$k] && $strand2[$i] =~ /P/ && $strand2[$i] =~
/^$strand2[$k]$/ && $blocation2[$i] > $blocation2[$k] && $blocation2[$i] >
$alocation2[$i] && $blocation2[$k] > $alocation2[$k] && $alocation2[$i] -
$alocation2[$k] >= -100 && $alocation2[$i] - $alocation2[$k] <= 6000 && $seqstart2[$i]
- $seqend2[$k] < 20000) {


open (FILE2, ">>$f2")

   ||die "cannot open";

   print FILE2
"$chr2[$i]\t$seqstart2[$k]\t$seqend2[$i]\t$strand2[$i]\t$type2[$i]\t$alocation2[$k]\
t$blocation2[$i]\n";
```

```
} elsif ($chr2[$i] =~ $chr2[$k] && $strand2[$i] =~ /C/ && $strand2[$i] =~
/^$strand2[$k]$/ && $alocation2[$k] > $alocation2[$i] && $alocation2[$i] >
$blocation2[$i] && $alocation2[$k] > $blocation2[$k] && $alocation2[$k] -
$alocation2[$i] >= -100 && $alocation2[$k] - $alocation2[$i] <= 6000 && $seqstart2[$i]
- $seqend2[$k] < 20000) {




open (FILE2, ">>$f2")

   ||die "cannot open";

   print FILE2
"$chr2[$i]\t$seqstart2[$k]\t$seqend2[$i]\t$strand2[$i]\t$type2[$i]\t$alocation2[$k]\
t$blocation2[$i]\n";



}

}
```

2.5 Looping script

Description

This script is the general process used to run the previously mention script (2.2 and 2.3) that used to search complex TE (potential complex recombinant repeat) that had been broken into 3~10 fragments.

```
#Comment: The initial script is used to search TE broken into 2 pieces

perl DNA_repair.pl TE_A_data.bed 2_TE_fragment_merge_result.out


#Comment: This script is used to search TE broken into 3 pieces based on previous
output

perl DNA_repair_level_2.pl 2_TE_fragment_merge_result.out
3_TE_fragment_merge_result_for_further_identification.out
confirmed_TE_A_created_by_3_fragments.bed


#Comment: This script is used to search TE broken into 4 pieces based on previous
output
```

perl DNA_repair_level_2.pl 3_TE_fragment_merge_result.out 4_TE_fragment_merge_result_for_further_identification.out confirmed_TE_A_created_by_4_fragments.bed

#Comment: This script is used to search TE broken into 5 pieces based on previous output

perl DNA_repair_level_2.pl 4_TE_fragment_merge_result.out 5_TE_fragment_merge_result_for_further_identification.out confirmed_TE_A_created_by_5_fragments.bed

#Comment: This script is used to search TE broken into 6 pieces based on previous output

perl DNA_repair_level_2.pl 5_TE_fragment_merge_result.out 6_TE_fragment_merge_result_for_further_identification.out confirmed_TE_A_created_by_6_fragments.bed

#Comment: This script is used to search TE broken into 7 pieces based on previous output

perl DNA_repair_level_2.pl 6_TE_fragment_merge_result.out 7_TE_fragment_merge_result_for_further_identification.out confirmed_TE_A_created_by_7_fragments.bed

#Comment: This script is used to search TE broken into 8 pieces based on previous output

perl DNA_repair_level_2.pl 7_TE_fragment_merge_result.out 8_TE_fragment_merge_result_for_further_identification.out confirmed_TE_A_created_by_8_fragments.bed

#Comment: This script is used to search TE broken into 9 pieces based on previous output

perl DNA_repair_level_2.pl 8_TE_fragment_merge_result.out 9_TE_fragment_merge_result_for_further_identification.out confirmed_TE_A_created_by_9_fragments.bed

#Comment: This script is used to search TE broken into 10 pieces based on previous output

perl DNA_repair_level_2.pl 9_TE_fragment_merge_result.out
10_TE_fragment_merge_result_for_further_identification.out
confirmed_TE_A_created_by_10_fragments.bed

2.6 fastacmd

Description

The potential CRR FASTA sequences were extracted from the relevant genome assembly with fastacmd based on the BED output that generated from previous analysis as shown below:

fastacmd -p F -i hg19_genome_assembly -s 'chr1' -L 10000,12000 >> CRR_sequences.fa

2.7 Krishna

Description

The CRR fast file was undergo sequence alignment with Krishna as shown below. The sequence identity/similarity between two sequence alignment was set to > 70%.

krishna -self=true -filtid=0.70 -query CRR_sequences.fa -target=CRR sequences.fa -output krishna_CRR_sequences.gff -log=true

2.8 PILER and RepeatMasker

Description

The aligned CRR sequences (obtained from krishna) was undergo clustering process with PILER as shown below. The family size (-famsize) is set to 3 sequences (minimum size to form low copy number repeat family) in order to avoid false positive result.

piler -trs krishna_CRR_sequences.gff -out trs_krishna_CRR_sequences.gff -famsize 3

2.9 Repeat Annotation with PILER

Description

The CRR family sequences obtained from precious PILER analysis (as shown in Section 2.8) were further annotated by RepeatMasker output file previously done in Section X.

piler -annot trs_krishna_CRR_sequences.gff -rep human_repeatmasker_output.gff -out RM_trs_krishna_CRR_sequences.gff

References

1. Smit A, Hubley, R & Green, P. (1996-2010) RepeatMasker Open-3.0. http://wwwrepeatmaskerorg.

2. Kortschak D (2010) Krishna. http://godocorg/codegooglecom/p/biogoexamples/krishna.

3. Edgar RC, Myers EW (2005) PILER: identification and classification of genomic repeats. Bioinformatics 21 Suppl 1: i152-158.

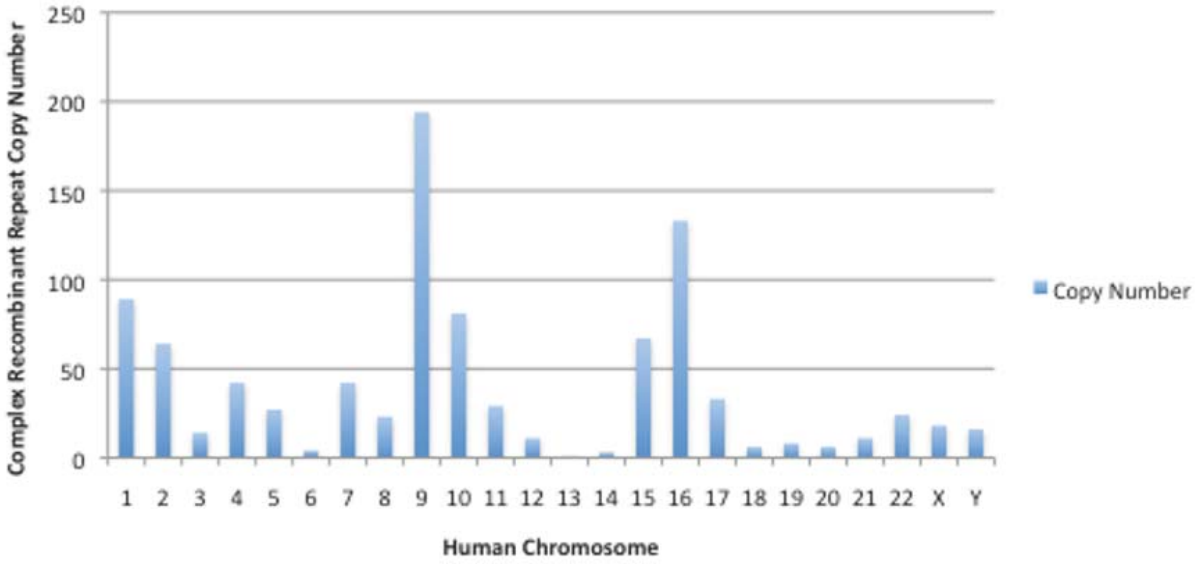
4. http://www.ncbi.nlm.nih.gov/IEB/ToolBox/SDKDOCS/INDEX.HTML.

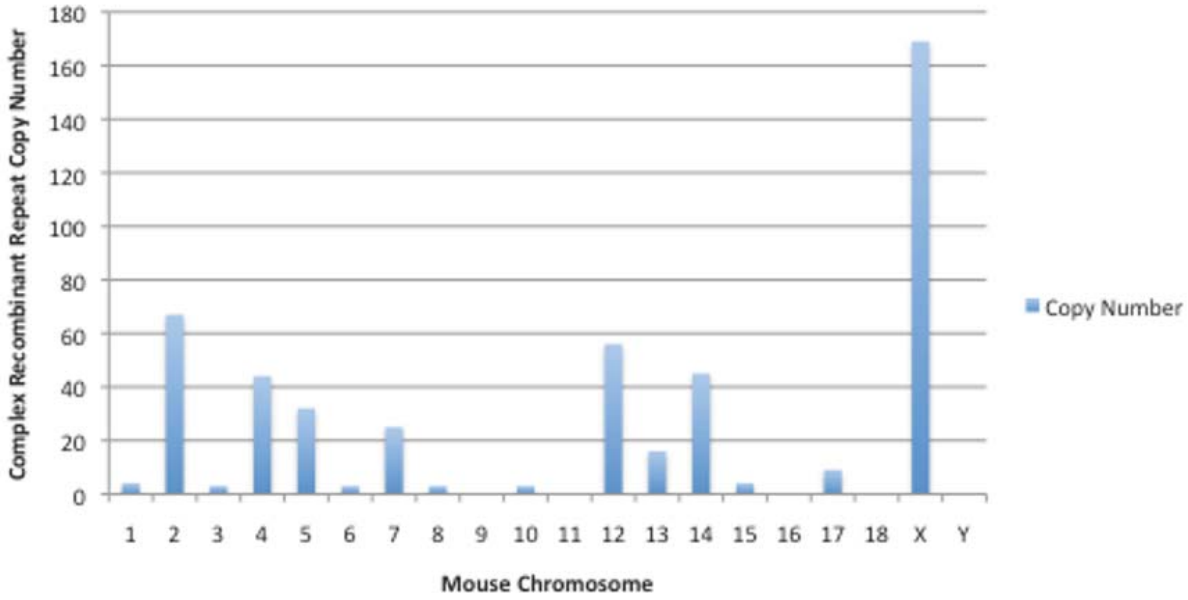
5. http://www.perl.org/.

6. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841-842.

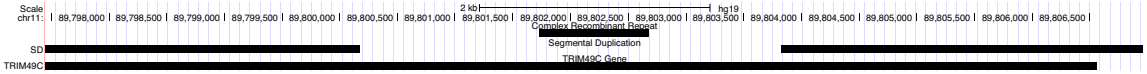
7. Harris RS (2007) Improved pairwise alignment of genomic DNA. PhD Thesis, The Pennsylvania State University.

Supplementary Figure S1. The Complex Recombinant Repeat Distributions in Human and Mouse Chromosome.





Supplementary Figure S2. The Comparison of hg19 Segmental duplication map, human segmental duplicated gene (TRIM49C), and CRR sequence (Family 26.7) coordinates presented in UCSC Genome Browser.



Supplementary information of Chapter 4: Discovery of Chimeric LTR, LTR2i_SS in *Sus scrofa*

Table S1. The software used in pig LTR2i_SS identification.

Software	Website/Link
CENSOR	http://www.girinst.org/censor/
RepeatMasker	http://www.repeatmasker.org/
PERL	http://www.perl.org/
BEDTools	http://code.google.com/p/bedtools/
PostgreSQL	http://www.postgresql.org/
PILER	http://www.drive5.com/piler/
WU-BLAST	http://blast.wustl.edu/
krishna	http://godoc.org/code.google.com/p/biogo.examples/krishna