[Frontiers in Bioscience E4, 1542-1555, January 1, 2012]

Database documentation of retrotransposon insertion polymorphisms

Ping Liang[1], Wanxiangfu Tang[1]

[1]*Department of Biological Sciences, Brock University, 500 Glenridge Ave, St. Catharines, Ontario, Canada L2S 3A1*

TABLE OF CONTENTS

1. ABSTRACT

Retrotransposons constitute more than 40% of the human genome with L1, Alu, SVA, and HERVs known to remain active in transposition. Retrotransposition contribute to genetic diversity in the form of retrotransposon insertion polymorphism (RIP) that is defined as the presence or absence of a retrotransposon insertion among human populations at a specific genomic location. So far close to 5000 cases of RIPs have been identified with more than 50 cases associated with disease. A large number of new RIPs are being and to be identified from newly available personal genomes data, making RIPs an important source of genetic variations/mutations that deserve proper documentation. In this review, we discuss the special characteristics of RIPs and the challenges in their compiling and annotating, and we examine the current status of database documentation of RIPs and describe in details the design, data schema, and utilities of dbRIP, which is currently the only database dedicated to the documentation of retrotransposon insertion polymorphism. Some future perspectives and outstanding issues associated with documentation of RIPs are also presented.

2. INTRODUCTION—RETROTRANSPOSONS AND THEIR CLASSIFICATION

Retrotransposon elements (REs) are a group of transposable elements (TEs) that propagate themselves into different places in the genome via an intermediate process of reverse transcription. In a sense, REs proliferate in the genome in a copy-and-paste fashion. In the human genomes, as in most other mammalian and plant genomes, REs exist in millions of copies and all together they constitute more than 40% of the human genome. REs have played very important roles in shaping the evolution of human and other primate genomes. They impact the functions of genes and the genome via a variety of mechanisms, which include, but not limited to, generation of insertion mutations and genomic instability, creation of new genes or gene isoforms, and alteration of gene expression regulation and epigenetic regulation (1-13). The major types of retrotransposons in the human genome include the LTR retrotransposons, i.e., the Endogenous Retrovirus (ERVs), that are characterized by the presence of the two long-terminal repeats (LTRs), and the non-LTR retrotransposons LINE1 (L1) and Alus. L1s, Alus, and

ERVs comprise approximately 500,000, 1,000,000 and 300,000 copies and constitute 17%, 11% and 8.5% of the human genome, respectively (14, 15). These REs represent the groups that were once very successful during the evolution of the mammals and primates and to certain degrees have remained active in the current human genomes. Also worthy of mention are the SVAs, a type of non-LTR retrotransponsons, which in sequence represent chimeras of SINE, VNTR and Alu-like regions. SVAs are very young and highly active, despite their small population size of a few thousand copies (16, 17).

Each of these major types of retrotransposons can be divided into subfamilies of more closely related elements based on a set of diagnostic nucleotide sequences. For example, Alus can be divided into more than 200 subfamilies (18), whereas SVAs have only 6 subfamilies (16). The formation of subfamilies, which often exist in a hierarchical structure, reflects the evolutionary dynamics of retrotransposon amplification in a sequential or a linear accumulation fashion. In others words, as explained by the "master gene model", a limited number of "master" copies that are competent for retrotransposition in the genome are responsible for the generation of most new copies, with all progeny copies from each "master" copy forming a subfamily at a variable size depending on the activity level of the "master" copy and the length of its existence. New "master" copies may emerge from a subfamily and form a new cluster within that parent subfamily. For example, AluYb9 originated from AluYb8 by carrying one extra diagnostic nucleotide variation, where as AluYb8 came from Yb5, which belong to the larger AluYb subfamily, which in turn is part of a relatively older AluY family. The proliferation rate of retrotransposons during the evolution of mammals and primates has not been constant with significant differences seen among the major types of retrotransposons and among subfamilies of the same RE type. For example, the ERVs were more active during the early evolution of primates but have become much less active at least in the human genome (19-21). Furthermore, different REs in many cases have showed quite different activities among closely related species after their divergence, as well demonstrated by the dramatically different profiles of Alu elements among different primate species (22, 23). For instance, AluYb and AluYa subfamilies have been highly active in the human genomes, but not in the chimpanzee genome (23, 24), whereas the activity of AluYc5 subfamily is relatively small in the human genome but is much larger in the chimpanzee genome, apparently as the most active Alu subfamily in this genome (unpublished data).

As a result of this past and ongoing proliferation activity of retrotransposons that fluctuate both vertically and horizontally (cross-species), a significant amount of genome diversity has been generated between human and other closely related primate species and among different human populations, as well as individuals within populations, leading to the generation of retrotransposon insertions that are species-, population- and family lineage-specific. For example, there are ~2000 L1, ~7000 Alus, and ~1000 SVA insertions that are only found in the human genome, and together these insertions contributed more than 8 Mb of nucleotide sequences to the human genome, a major factor leading to the increase in genome size in humans in comparison with chimpanzees (25-32). There are also close to 17,000 copies of ERVs that are specific to human genomes (human endogenous retrovirus or HERVs), and they make up more than 22 Mb of sequences. In this case, the generation of the HERVs may be a combined effect of the proliferation from existing ERVs and newly domesticated virus and their proliferation in the human genome (19, 33).

Different from other types of regular sequence variation, Retrotransposon Insertion Polymorphism (RIP) refers to the presence or absence of a retrotransposon insertion at a specific genomic location in populations of a given species. Due to their significant impact on genes and genome as a whole, and despite the relatively little attention received so far, these RIPs constitute a very important source of all human genetic polymorphisms that together with other types of genetic variations are responsible for the full spectrum of the vivid phenotypic differences observed among human individuals, such as the physical appearance and susceptibility to diseases. In this review, we examine the history of research related to identification of RIPs in humans and the associated methodologies, and we discuss the special characteristics of retrotransposon insertion polymorphisms and the challenges in compiling the data. We also examining the current status of database documentation of RIPs and describe in details the design, data schema, and utilities of dbRIP, the current only database specially designed for the documentation of RIPs.

3. COMPUTATIONAL DOCUMENTATION OF RETROTRANSPOSON INSERTION POLYMORPHISM

3.1. Methods for identification and ascertainment of retrotransposon insertion polymorphism

The currently known polymorphic retrotransposon insertions were identified using a number of approaches and methodologies all within the last two decades. Earlier studies using genomic library screening with probes/primers specific for young Alu elements contributed to the discovery of a small number of RIPs (34-37). A recent study employing the library screening approach combined with high throughput pair-end Sanger sequencing successfully identified 198 L1 insertions, as well as 1 HERV-K insertion, not present in the reference genome from the analysis of 17 genomes (38). Most of the disease-related retrotransposon insertions were discovered from the mutational screening of candidate genes, using methods including Southern blot, DNA sequencing, etc. (e.g. 39, 40 and reviews 41, 42).

The task of finding RIPs among millions of copies that are highly similar in sequence in a genome is essentially like "finding a needle in a hay stack". For this reason, no large-scale comprehensive study was possible until the human genome sequences became available (14, 43). The use of the human genome sequences for identification of RIPs was first explored by Batzer's group.

In this approach, Alu elements belonging to young subfamilies were identified by computational sequence analysis based on the level of sequence divergence among family members, and polymerase chain reactions (PCRs) using primers designed in regions flanking the insertion were used to ascertain the polymorphism status of these candidates by screening DNA samples from diverse human populations. The first study using such a strategy identified 106 polymorphic Alu insertions out of 475 Ya5 and Yb8 insertions (44). Subsequently, this method was extensively used to analyze almost all Y subfamilies including Ya, Yb, Yc, Yd, Yg and Yi, Ye, and multiple AluY subfamily members on the X chromosome (23, 45-51). Together these studies are responsible for the identification of over 400 polymorphic Alu insertions. While successful, the use of this strategy was limited to REs that are covered in the public version of the human genome sequence and the selection of candidates was biased towards certain relatively small and young subfamilies for which the numbers of candidates are manageable for PCR assays.

To identify RIPs that are absent in the reference genome, genomic DNA sequences from more human individuals representing different populations are needed. With the genomic sequences becoming available for multiple human individuals, computational comparative genomic approaches were developed to more effectively identify RIPs. The first attempt with this type of strategy used partial human trace genomic sequences representing 36 diverse humans to compare with the reference genome and identified over 600 Alu, L1 and SVA insertion polymorphism (52). The first study comparing two genome sequences was performed by our group, in which we took advantage of the availability of the public and Celera versions of human genome sequences that roughly represent two different individual genomes despite the mixing nature of DNA used for genome sequencing. In that study, we identified more than 800 new Alu insertion polymorphisms, the largest set of polymorphic Alu insertions identified by a single study at that time (26). Among these Alu RIPs, more than one third were insertions outside the public versions of the genome sequence. Subsequently, the same approach was used for identification of ~150 polymorphic L1 insertions (27). Somewhat surprising was the fact that there is very little overlap among lists of the RIPs identified from the above three large-scale computational studies. This is likely because each method used a different genomic sequence source and identified an incomplete list among a large number of possible RIPs, and it served as the first strong hint that the actual level of RIPs may be much higher than what we could have expected from the limited number of RIPs previously identified. More recent studies using the diploid Venter genome in comparison with the reference genome revealed more polymorphic insertions of Alus, L1s and SVA and confirmed our previous speculation, as well as demonstrated the usefulness of diploid genome sequences for identification of new RIPs (29, 53).

With the advent of next generation sequencing (NGS) technologies and their applications in sequencing a large number of personal genomes, a few more approaches have been developed identification of de novo retrotransposon insertions in the genomes of individuals in question. One of the strategies that have been tested in a few laboratories is the use of NGS to selectively sequence the junction areas between RE insertions and their flanking genomic sequences. For example, using this approach Witherspoon *et al.* has identified a large number of novel Alu RIPs from several Japanese individuals (54). Similarly, Ewing & Kazazian devised a NGS sequencing approach for L1, and by surveying 26 individuals, they identified 367 L1s not present in the reference genome, majority of which are novel polymorphic L1s (55). Further more, the availability of personal genome sequences in large number, such as those that have and being generated by the 1000 genome projects, permits identification of novel RIPs via computational comparative genomic analysis (56, 57). An unprecedented larger number of novel RIPs from the known families of active retrotransposons, Alu, L1, and SVA, are being identified, among which include those that are specific to populations or groups of populations (58-60). In addition to the sequencing approaching, a microarray-based method has also been explored for identifying polymorphic L1 insertions (61). We can expect the discovery of many more novel RIPs from the analysis of a sufficiently large number of individual genomes representing diverse populations, particularly the ancient or highly isolated populations, such as the Bushmen and Neanderthal genomes (62-64). In addition to novel RIPs identified mostly as population and individual from these analyses, we can also expect a certain number of RE insertions present in the reference genome to be recognized as RIPs, particular from the analysis of the ancient populations.

The current gold standard for ascertaining a RIP is PCR, in which a pair of PCR primers are designed in the flanking regions of the insertion, such that the presence and absence of the RE insertion will lead to differences in PCR product size, i.e., the size for the insertion positive allele is larger than that of the product for the insertion negative allele roughly by the size of the insertion. This strategy works well when the RE insertion is relative small, i.e. below 2kb, and it can distinguish among the three genotypes of a RE insertion, "+/+", "+/-", and "-/-", as having one large product, one large and one small product, and one small product, respectively (see illustration in Figure 1A). When the insertion size is large, such as insertions of full length L1s and HERVs, which can be as long as 10 kb, it becomes difficult to obtain a product for the insertion positive allele even with a long range PCR. In this case, as used for genotyping HERV RIPs by Belshaw *et al* (21), a better strategy is to design two additional primers inside the RE, which are oriented outwards, such that in the presence of the RE insertion these two primers will work with the two primers in the flanking region to generate two shorter products, while for the insertion negative allele there would be only one product to be generated from the two primers in the flanking regions. Due to the variability of subgroup sequences, these primers generally require to be designed for each specific subgroup of RE, unless a "universal" primer can be found based on a region highly conserved among the larger RE family. With

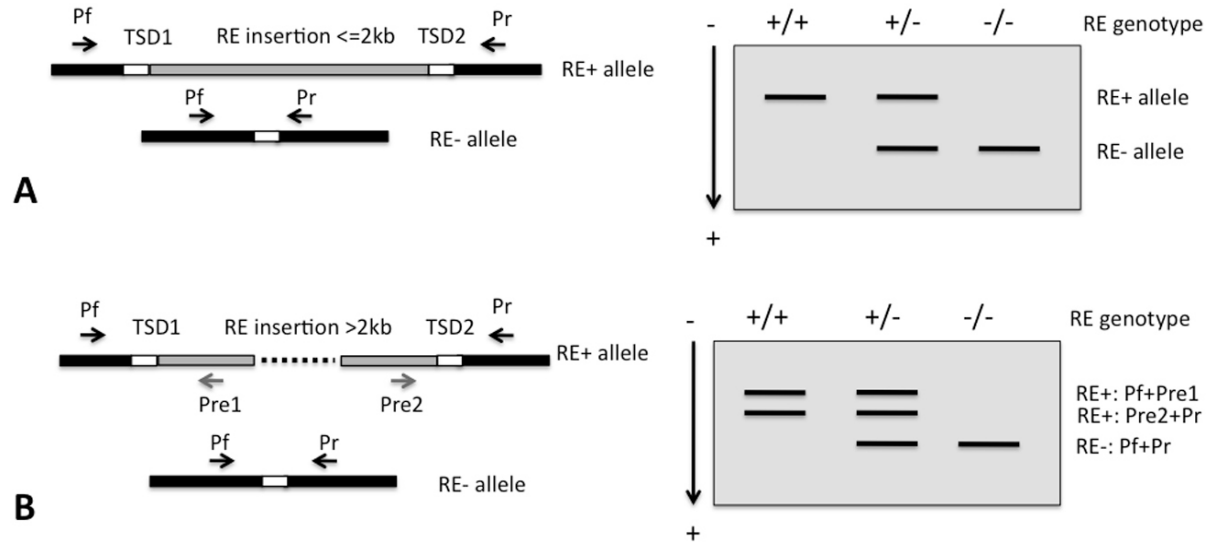


Figure 1. A schematic representation of RIP ascertainment by PCR. The left panels illustrate the design of the PCR primers, while the right panels illustrate the patterns of PCR products on agarose gels. Panel A is for ascertaining short RE insertions (e.g. <=2kb), while panel B is for ascertaining long RE insertions, such as full-length L1 and HERV insertions. The sizes of the three PCR products relative to one another in panel B may be different from locus to locus depending the specific location of the primers relative to the insertion site. Arrows labeled as "Pf" and "Pr" indicate the locations of the forward and reverse primers designed in the genome regions flanking the RE insertion, respectively, while those labeled as "Pre1" and "Pre2" indicate the primers designed with the RE insertion and are oriented outwards.

these two sets of primers, 2, 3, and 1 product(s) are expected for samples with a genotype of "+/+", "+/-", and "-/-", respectively (see illustration in Figure 1B). Therefore, a PCR assay can provide complete and accurate genotyping of the RIP by distinguishing between the three possible genotypes. In addition, PCR assay can also provides DNA for sequencing to obtain the sequence of the RE insertion, which may be highly desired for RIPs outside the reference sequence, for which the RE insertion and target site duplication (TSD) sequences are usually not available.

Among the known RIPs, most were identified by computational analysis of genomic sequences, among which usually only a small portion was subjected to PCR verification, more as a way for assessing the accuracy of the methods than an attempt for validation. In these studies, it is often not feasible to experimentally verify all candidate RIPs due to their large numbers and the prohibitive cost of validation by PCR. However, to make the RIP data as reliable and usable genetic variation data for the genetic community, it is important that we can experimentally validate all RIPs identified computationally or using any other methods that do not provide complete sequences for the insertions and the associated sequences, such as TSDs.

3.2. Special characteristics and issues associated with RIPs.

In comparison with other types of genetic variations, such as such as single nucleotide polymorphisms (SNPs), indels, genomic rearrangements, and copy number variations (CNVs), RIPs have several unique features. They require special handling in characterization, compiling, and data display as discussed below.

First, a RIP represents an evolutionary event that has a definitive ancestral status, which is always the absence of the insertion. In other words, the pre-integration sequence is always the ancestral form. Since the chance for two individuals to have the same retrotransposon insertion as a result of two independent insertion events is almost zero and there is no known mechanism that specifically removes a retrotransposon insertion, the only reason for two individuals to share a retrotransposon insertion is their shared ancestry. For this very reason, RIPs are considered to be homoplasy-free, and it is this characteristic that makes RIPs a very useful type of genetic markers in population studies, particularly in resolving the ancestral relationship (65, 66). Also for these reasons, all RIPs are always true insertion polymorphisms and they should not be called "deletions" even when absent in the reference sequences. But it does present a challenge using the current reference-based nomenclature for documenting polymorphisms (67, 68) (http://www.hgvs.org/mutnomen/recs.html).

Second, the final outcome of a retrotransposon insertion event often carries complex sequence rearrangements beyond the insertion of the RE sequence. These rearrangements include the more common generation of TSDs at variable lengths, ranging from a few bps to a few hundred bps, or the less common deletion of flanking region at the integration site. Furthermore, in some cases, more often seen in association with L1 and SVA insertions, extra sequences flanking the parent copy of the retrotransposons can be carried to the progeny copy via 5' or 3' transduction (2, 4-6, 8, 9, 13, 69, 70). The sizes of these transduced sequences range from a few to a few hundred bps, and can lead to exon shuffling if coding

sequences are included in the transduced sequences. Accurate identification of the transduced sequence associated with a RE insertion is important for assessing the impact of the RIP. Again, using the current nomenclature for polymorphism/mutations presents difficulties for documenting these extra sequence rearrangements.

Third, insertions caused by certain types of retrotransposons, such as LTR (e.g. HERVs), can generate post-insertion secondary changes within the RE insertion, such as the homologous sequence-mediated recombination between the two LTRs, leading to the deletion of the internal ERV sequence and generation of a solo-LTR (19, 71, 72). As a result, different forms of the insertion sequences, despite their origin from the same insertion event, may co-exist at the same site among the populations. In the case of a HERV insertion, at least three forms of sequences at the site, i.e., the pre-integration sequence, insertion containing the full HERV sequence, and the insertion containing only the solo-LTR, can exist as shown by Belshaw *et al* (21). Therefore, the genotype data in this case has to be dealt and presented differently.

Lastly, RE insertions, including the associated TSDs, are also sources of SNPs and microsatellite variations (73). The two copies of TSDs can be subjected to random mutations and become different from each other, while the poly-dA tracks carried by Alus, L1s and SVAs are a major source of microsatellite DNA subjected to a high level of sequence variations. These SNP variations carried by a RE sequence that is polymorphic itself by way of presence or absence adds an additional dimension to the genetic diversity, and it is a challenge to report and document them.

For all these reasons, it is much more challenging to document these retrotransposon insertion polymorphisms than other types of sequence polymorphisms. Therefore, RIPs warrant to be treated as a special type of genetic variations.

3.3. Database documentation of retrotransposon insertion polymorphism in dbRIP

3.3.1. Current RIP database documentation status

Due to the large number of known RIPs and the many more expected to be identified, it is essential that these data are compiled in a way that is accurate and easy to access. Accuracy here refers not only to the reliability of the data in all components of the information accurately, including the sequence of the insertion, location, and classification, but also the completeness of the data. For example, just knowing the presence of the insertion at a specific location does not provide sufficient information about the potential impact of the insertion, and it is important also to know the exact sequence of the insertion and the TSDs or deletion of the flanking sequence and/or the presence of 5' or 3' transduced sequence. Other types of information, such as the source of the polymorphism (i.e. the specific population or individual showing the presence or absence of the insertion), the ascertaining/genotyping method, the insertion allele frequency in the examined populations, the phenotype association, etc. are also very useful. The sample source is very important for future study of rare RIPs. Currently, the RIP data can be found in a few databases which include the dbSNP at NCBI (http://www.ncbi.nlm.nih.gov/projects/SNP/), database of retrotransposon insertion polymorphisms (dbRIP) at Brock University (http://dbrip.org) and database of genome variants (DGV) at the Centre of Applied Genomics (http://projects.tcag.ca/variation/) (25, 74). Some mouse RIP data were included in MouseIndelDB (http://variation.osu.edu/) (75). In all these databases other than dbRIP, it was not straightforward if possible at all, to find specifically the entries related to retrotransposon insertion polymorphism, nor do they allow query by disease, gene context-based location or RE class. For example, unless the dbSNP IDs for polymorphic retrotransposon insertions, such as those by Bennett *et al* (52), are known at the time of query, it is very hard to find these data from dbSNP, and there is no indication of the insertion sequence and the TSDs, neither is the classification of the RE provided due to the different purpose of the database. Among all these databases, only dbRIP was designed specifically to accommodate the special needs of retrotransposon insertion polymorphism data (25). dbRIP has been recognized as an important reference resource for the research community as demonstrated by the large number of citations it receives since its relative short inception (29, 53, 54, 58, 75-80). We describe in the subsequent sections in detail about the design, database schema, and utilities, and future development of dbRIP.

3.3.2. Overview of dbRIP

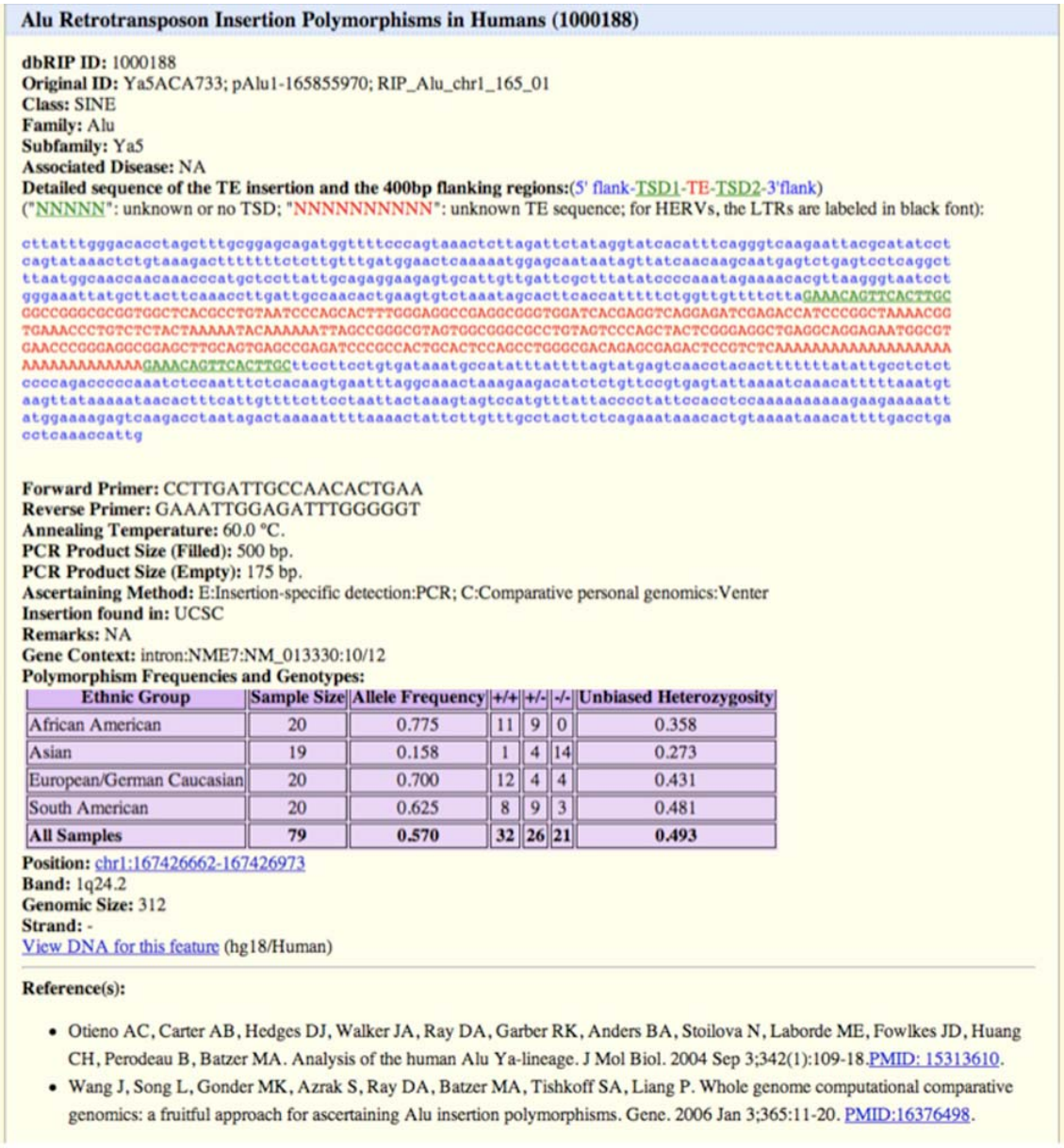
In designing dbRIP, instead of having it as a standalone database like many other biological databases, we decided to have it integrated with a genome browser, and among the existing genome browsers, we choose to go with the UCSC Genome Browser for its easy-to-use interface and comprehensive coverage of functional genome data. This integration made dbRIP very user friendly, and more importantly, it allows the RIP data to be viewed in context of genome sequences, gene and many other related genomic and functional genomic data that are made available via the UCSC Genome Browser (81, 82).
In the first full release of dbRIP data in June of 2006, there were 2095 non-redundant entries from a total 2897 reported cases, including 1625, 407 and 63 cases of Alus, L1s and SVAs, respectively (25), and we have recently extended the coverage to include RIPs derived from HERVs. As of writing, dbRIP covers a total of 2,771 non-redundant RIP entries, including 2086, 598, 77, 10 cases of Alus, L1s, SVAs, and HERVs, respectively (Table 1; Tang *et al*, manuscript in submission). These RIP data were collected from over 70 publications and were curated manually to characterize all properties associated with a RIP.

One of the unique features of dbRIP, which is very important for retrotransposon insertion polymorphism, is that we provide the detailed sequence information associated with a RIP by distinguishing the sequences of the RE insertion, the target site duplications (TSD) and the flanking regions (Figure 2). For LTR retrotransposons,

Table 1. A summary statistics of RIPs documented in dbRIP release 2.0 for hg18

RIP Class	# of loci (unique/total)	# of loci outside hg18	# of loci with genotype	Gene context (D/P/E/I/IG*)	Disease- related loci
Alu	2086/2708	858	526	6/7/24/765/1284	33
L1	598/800	299	123	2/2/10/183/401	15
SVA	77/87	18	31	1/1/4/29/42	3
HERV	10/10	2	6	0/0/5/5	/
Total	2771/3605	1177	686	9/10/38/982/1732	51

*D: downstream up to 500bp, P: promoter up to 1kb, E: exon, I: intron, IG: intergenic region



Alu Retrotransposon Insertion Polymorphisms in Humans (1000188)

dbRIP ID: 1000188
Original ID: Ya5ACA733; pAlu1-165855970; RIP_Alu_chr1_165_01
Class: SINE
Family: Alu
Subfamily: Ya5
Associated Disease: NA
Detailed sequence of the TE insertion and the 400bp flanking regions:(5' flank-TSD1-TE-TSD2-3'flank)
("NNNNN": unknown or no TSD; "NNNNNNNNNN": unknown TE sequence; for HERVs, the LTRs are labeled in black font):

cttatttgggacacctagctttgcggagcagatggttttcccagtaaactcttagattctataggtatcacatttcagggtcaagaattacgcatatcct
cagtataaactctgtaaagacttttttctcttgtttgatggaactcaaaaatggagcaataatagttatcaacaagcaatgagtctgagtcctcaggct
ttaatggcaaccaacaaacccatgctcctttattgcagaggaagagtgcattgttgattcgctttatatcccaaatagaaaacacgttaagggtaatcct
gggaaattatgcttacttcaaaccttgattgccaacactgaagtgtctaaatagcacttcaccatttttctggttgttttcttaGAAACAGTTCACTTGC
GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGTGGATCACGAGGTCAGGAGATCGAGACCATCCCGGCTAAAACGG
TGAAACCCTGTCTCTACTAAAAATACAAAAAATTAGCCGGGCGTAGTGGCGGGCGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATGGCGT
GAACCCGGGAGGCGGAGCTTGCAGTGAGCCGAGATCCCGCCACTGCACTCCAGCCTGGGCGACAGAGCGAGACTCCGTCTCAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAGAAACAGTTCACTTGCttccttcctgtgataaatgccatatttattttagtatgagtcaacctacacttttttatattgcctctct
ccccagacccccaaatctccaatttctcacaagtgaatttaggcaaactaaagaagacatctctgttccgtgagtattaaaatcaaacatttttaaatgt
aagttataaaaataacactttcattgttttcttcctaattactaaagtagtccatgtttattaccccctattccacctccaaaaaaaaaagaagaaaaatt
atggaaaagagtcaagacctaatagactaaaaattttaaaactattcttgtttgcctacttctcagaaataaacactgtaaaataaacattttgacctga
cctcaaaccattg

Forward Primer: CCTTGATTGCCAACACTGAA
Reverse Primer: GAAATTGGAGATTTGGGGGT
Annealing Temperature: 60.0 °C.
PCR Product Size (Filled): 500 bp.
PCR Product Size (Empty): 175 bp.
Ascertaining Method: E:Insertion-specific detection:PCR; C:Comparative personal genomics:Venter
Insertion found in: UCSC
Remarks: NA
Gene Context: intron:NME7:NM_013330:10/12
Polymorphism Frequencies and Genotypes:

Ethnic Group	Sample Size	Allele Frequency	+/+	+/-	-/-	Unbiased Heterozygosity
African American	20	0.775	11	9	0	0.358
Asian	19	0.158	1	4	14	0.273
European/German Caucasian	20	0.700	12	4	4	0.431
South American	20	0.625	8	9	3	0.481
All Samples	**79**	**0.570**	**32**	**26**	**21**	**0.493**

Position: chr1:167426662-167426973
Band: 1q24.2
Genomic Size: 312
Strand: -
View DNA for this feature (hg18/Human)

Reference(s):

- Otieno AC, Carter AB, Hedges DJ, Walker JA, Ray DA, Garber RK, Anders BA, Stoilova N, Laborde ME, Fowlkes JD, Huang CH, Perodeau B, Batzer MA. Analysis of the human Alu Ya-lineage. J Mol Biol. 2004 Sep 3;342(1):109-18.PMID: 15313610.
- Wang J, Song L, Gonder MK, Azrak S, Ray DA, Batzer MA, Tishkoff SA, Liang P. Whole genome computational comparative genomics: a fruitful approach for ascertaining Alu insertion polymorphisms. Gene. 2006 Jan 3;365:11-20. PMID:16376498.

Figure 2. An example of a dbRIP record detailed information page. A screen shot of the RIP data page showing all information associated with dbRIP_Alu_00188.

which are characterized by the presence of two long-terminal repeats (LTRs) at the ends, we also define and label the LTR sequences differently from the internal endogenous retrovirus related sequence, which is useful for assessing the impact of a HERV RIP. Other informational items we currently collect and provide for RIPs includes identifications (including the original ID(s)), RE classification, known disease associations, available PCR conditions (primers, TM, and allele sizes) used for genotyping, methods of ascertainment, source of the RIPs,

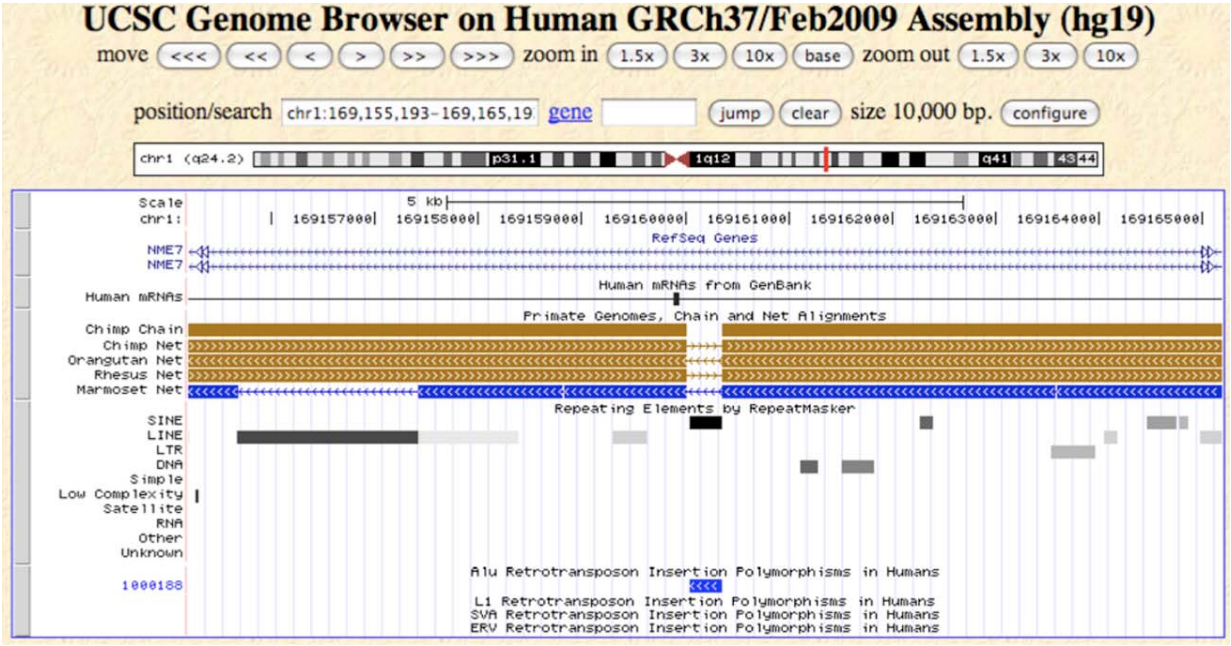


Figure 3. A screen shot of the UCSC Genome Browser showing a dbRIP entry shown. An Alu RIP, dbRIP 1000188, as shown in **figure 2** was displayed in the UCSC Genome Browser (the blue bar at the bottom) along with other data tracks. This RIP is located in the intron of gene NME7 and represents a RE insertion present in the reference genome (see the RepeatMasker track). Comparative genomic data track indicate this RE insertion is absent in non-human primate genomes (see the tracks above the RepeatMasker

genomic location in gene context, available genotyping data in details, and the original reference(s) reporting each RIP (Figure 2).

3.3.3. dbRIP utilities

A key function of a database is to allow querying its data by all properties associated with the data. The current major utilities of dbRIP include a data-search interface and a position-mapping tool. In dbRIP, the RIP data can be searched using the standard utilities provided by the UCSC genome browser. For example, one can use the genome position, gene name, and blat to find all RIPs located in the specified regions or associated with the specified genes or query sequence. Advanced users can also use the Table Browser utility to perform more sophisticated queries (83). To further facilitate the search of dbRIP data, a "SearchdbRIP" tool was developed for querying the RIP data, and it allows querying of RIP data based on one or more of RIP properties, which may not be possible or easy to perform using the UCSC Genome Browser utilities. The SearchdbRIP utilities are divided into two sections. The first part allows quick search by RIP IDs (can be either the dbRIP ID or names used in the original study), and chromosome coordinates. The second part provides advanced search by one or more of the RIP data properties including the chromosome, location in gene context, source of RIP, RE subfamily, population, allele frequency range, disease association, and author name (Figure 4). For location in gene context, we break the genome into 5 categorical regions, including exon, promoter (1kb upstream transcription start site), intron, downstream (500 bp downstream the end of the gene), and inter-genic regions in priority order from high to low to handle situations where two or more categories can be assigned to the same location. For exons, we further divide into 5'-UTR, CDS, 3'-UTR and non-coding RNA. Using this search parameter, one can query all RIPs falling into a specific gene location category, for instance, all RIPs in coding regions, as a way to study the functional impact of RIPs. The currently available categories for sources of the RIPs include "reference (e.g. hg18, hg19)", "Venter" and "others". This allows users to collect RIPs from a specific genome or data source. It is our plan to expand this list in the near future to cover other important sources, such as Watson, 1000 Genome Project, Bushmen, Neanderthal, etc. In short, SearchdbRIP allows users to collect a specific set of RIPs by using one of more of the search parameters. The output of SearchdbRIP is a detailed list of matched entries. For each matched entry, two hyper links are provided: the link to "Detailed" provides access to the detailed record page of the RIP as shown in Figure 2, and the link to "Browser" brings users to the genome browser at a default window size of the RIP insertion size plus 5kb on each side as shown in Figure 3.

The tool, PositionMapping, is designed for users to determine in a batch style among a list of newly identified candidate RIP entries which represent known RIPs (i.e. those in dbRIP) and which represent novel RIPs (i.e. not in dbRIP). The utility compares the positions of the user's input list with the coordinates of all dbRIP entries and identifies overlapping entries. It should be a useful tool for researchers who identify large lists of RIPs and want to compare with the data in dbRIP. As output, the utility

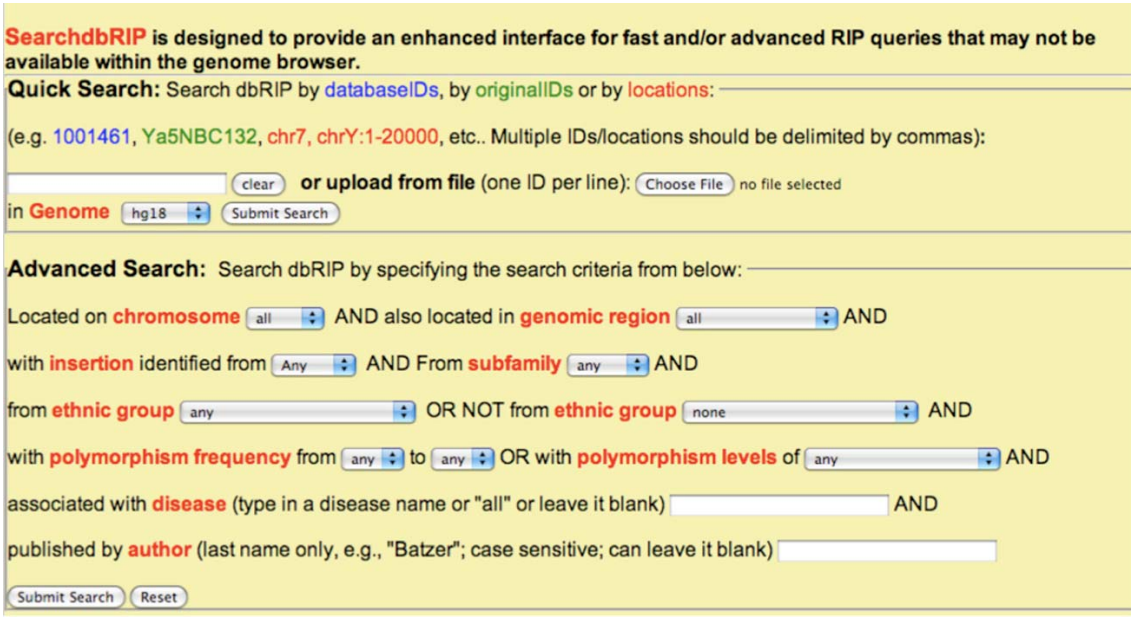


Figure 4. A screen shot of the SearchdbRIP interface showing all available search parameters.

provides a summary of the mapping result and generates a list of input IDs that are novel RIPs and a list of IDs that overlap with the data in dbRIP.

3.3.4. dbRIP record tracking and data releases

In designing the first version of dbRIP, we used an ID system that reveals the RE type, the chromosome, the position in million base pair of a RIP and the number of RIPs in the same position designation. For example, "RIP_Alu_chr7_003_01" indicates the RIP as the first Alu record located within the 3 million bp of chromosome 7. The intention was to provide as much information as possible via the ID about the RIP. However, this causes a problem of consistency when migrating to a new genome version due to the changes of the chromosome coordinates, making it either necessary to change the IDs with every genome migration or possibly rendering the ID meaningless. To avoid this problem, starting from release 2 of dbRIP data, we changed it to a 7-digit numerical ID system, similar to the on used in OMIM database (http://www.ncbi.nlm.nih.gov/omim). In this ID system, the first digit is used to indicate the major type of retrotransposons (1xxxxxx, 2xxxxxx, 3xxxxxx, 4xxxxxxx for Alu, L1, SVA and HERV, respectively) and the rest of 6 digits are used to indicate the sequential order of RIPs of this type deposited into dbRIP. For example, dbRIP 1000001 is the first Alu RIP, while 2000100 is the 100th L1 RIP. The remaining 5 digits (i.e., 5-9) for the 1st position in the ID are reserved for new types of RIPs and for accommodating existing types that exceed 1 million in number. Therefore, the system allows a maximal of 3 million RIP entries for one type of RE and a total of 10 million entries for all REs. The new ID system provides stability and consistency not affected by migrating to newer genome versions and it allows referencing a RIP record via a permanent identification.

To better track the changes of the dbRIP data, we developed a data-version system using release numbers. We assign a sequential data release version number, such as release 1 and 2, for each major update and use the decimal number after the major release number to label each minor update. For instance, release 2.1 will be used for the first minor update for release 2 data (either addition of a small number of RIPs and/or modifications to the existing data). The data release version is an indication of status of RIPs data (e.g. the total number of RIPs) and is not necessarily directly tied to a genome version, since one data release may be provided simultaneously for more than one genome version. For example, release 2.0 is made available for both hg18 and hg19. For the last version of each major data release, we provide a summary statistics table (see Table 1 for example).

3.3.5. Future development of dbRIP

We are committed to maintain and update dbRIP for the community as a free research resource. Future maintenance of dbRIP will focus on: 1) timely collection of newly published RIP data, 2) accommodation of new RIP data identified using new methodologies that may not complete information of a RIP, 3) support for newer versions of the reference human genomes and 4) possible expansion to other model organisms, such as mouse. Due to the expected availability of personal genome sequences in a large number and development of new strategies for experimental identification of novel RIPs, a large number of novel RIPs have been and will be identified via computational comparative genomics, microarray or next-generation sequencing approaches (54, 55, 58-61). Timely updates to accommodate these new data will be in high demand. For utility, we plan to develop an interface for users to submit new RIP data to facilitate the deposition of data into dbRIP. In addition, it may be useful to implement

a mechanism to inform interested users about the database updates, e.g. an email alerting service.

Future improvements for dbRIP need to address a few outstanding issues, including 1) how to document rare case of RIPs with deletions in flank regions; 2) how to report and describe the transduced sequences associated with RIPs; 3) development of controlled vocabularies and/or nomenclatures for describing RIP properties, such as the population name, methods used to identify or ascertain RIPs, genotype data and allele frequency, disease association, etc. To make RIP data accessible to a broader user community, we are working with teams of dbVAR at NCBI (http://www.ncbi.nlm.nih.gov/dbvar) and Database of Genome Variants (http://projects.tcag.ca/variation/) and to exchange data with these databases that have been covering or intent to cover RIP data. We will also continue to work with the UCSC Genome Browser team to make the dbRIP data track available for all human genome versions on their genome browser server (http://genome.ucsc.edu). To make dbRIP a more valuable resource for the community, we welcome suggestions and contributions from users regarding future data updating and improvement of the interface and utilities.

4. SUMMARY AND CONCLUDING REMARKS

Over the past decade, with the availability of human reference genome sequences and that of other primate genomes, we have obtained a panoramic view for retrotransposons, the major class of mobile elements in the genomes. We begin to learn more about their evolutionary history, proliferation dynamics during evolution and to appreciate their important impact on genome evolution, and gene function and genome diversity. The process of retrotransposition not only allows the propagation of these retrotransposons, many to a great success, to achieve their impact in the genome, but also serves as a major mechanism responsible for generating inter- and intra-species genome diversity. The retrotransposon insertion polymorphisms represent an important source of genetic polymorphisms, not only because the ever-increasing number of loci and the large amount of sequences involved, but also because of their significant and complex impact on genome structure and gene function. Their highly complex characteristics make them distinct from other types of genetic variations and require them to be handled in differently in curation and documentation. As of writing, close to 5000 entries of RIPs, amounting to over 25 million base pairs of sequences, have been identified, mostly from genome-wide surveys and comparative genomic analysis. Now with the advent of the newer generations of genome sequencing technologies and ever increasing number of personal genomes, we are provided with an exiting new opportunity to obtain a more complete picture for the level of genetic polymorphisms contributed by these retrotransposons, the related mechanisms, and functional impact. From these analyses, we expect to identify a large the number of new RIPs, speculatively triple of what we have known so far, reaching to those that represent very rare *de novo* insertion events from the known active retrotransposons and those by non-canonical mechanisms, perhaps also those from mobile elements not currently known to be active.

Complete and accurate documentation of such special complex types of genetic variations in an integrated and intuitive manner is essential for fully realizing the benefits of these research data. In the mean time, this also imposes many challenges and brings the need of developing new standards and nomenclature by the research community. dbRIP is currently the only database specially designed to accommodate the documentation of RIPs. Such databases that target at specific data types and user communities provide many benefits over general databases, mainly due to their high quality data curation and clearly defined data applications. Future improvements with community support are needed to meet the needs and challenges. Last but not least, as a community of the researchers on mobile elements, we need to advocate more about the importance of retrotransposon insertion polymorphisms for genetics and human population study and solicit more funding to support related research, particularly the large scale of validation and genotyping of RIPs, as well as the database documentation.

5. ACKNOWLEDGEMENT

We thank Scott Golem and the two anonymous reviewers for their critical review of the manuscript and suggestions. This work is in part supported by grants from the Canada Research Chair program, Canadian Foundation of Innovation (CFI), Ontario Ministry of Research & Innovation (OMRI), Brock University, and Natural Sciences and Engineering Research Council (NSERC) to PL, and was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET, http://www.sharcnet.ca) and Compute/Calcul Canada (https://computecanada.org/).

6. REFERENCES

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

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

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

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

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

6. K Han, J Lee, TJ Meyer, J Wang, SK Sen, D Srikanta, P Liang, MA Batzer: Alu recombination-mediated structural deletions in the chimpanzee genome. *PLoS Genet* 3, 1939-1949 (2007)

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

8. JD Boeke, OK Pickeral: Retroshuffling the genomic deck. *Nature* 398, 108-9, 111 (1999)

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

10. DV Babushok, EM Ostertag, HH Kazazian Jr: Current topics in genome evolution: molecular mechanisms of new gene formation. *Cell Mol Life Sci* 64, 542-554 (2007)

11. HH Kazazian Jr: Mobile elements: drivers of genome evolution. *Science* 303, 1626-1632 (2004)

12. PL Deininger, JV Moran, MA Batzer, HH Kazazian Jr: Mobile elements and mammalian genome evolution. *Curr Opin Genet Dev* 13, 651-658 (2003)

13. SK Sen, K Han, J Wang, J Lee, H Wang, PA Callinan, M Dyer, R Cordaux, P Liang, MA Batzer: Human genomic deletions mediated by recombination between Alu elements. *Am J Hum Genet* 79, 41-53 (2006)

14. ES Lander, LM Linton, B Birren, C Nusbaum, MC Zody, J Baldwin, K Devon, K Dewar, M Doyle, W FitzHugh, R Funke, D Gage, K Harris, A Heaford, J Howland, L Kann, J Lehoczky, R LeVine, P McEwan, K McKernan, J Meldrim, JP Mesirov, C Miranda, W Morris, J Naylor, C Raymond, M Rosetti, R Santos, A Sheridan, C Sougnez, N Stange-Thomann, N Stojanovic, A Subramanian, D Wyman, J Rogers, J Sulston, R Ainscough, S Beck, D Bentley, J Burton, C Clee, N Carter, A Coulson, R Deadman, P Deloukas, A Dunham, I Dunham, R Durbin, L French, D Grafham, S Gregory, T Hubbard, S Humphray, A Hunt, M Jones, C Lloyd, A McMurray, L Matthews, S Mercer, S Milne, JC Mullikin, A Mungall, R Plumb, M Ross, R Shownkeen, S Sims, RH Waterston, RK Wilson, LW Hillier, JD McPherson, MA Marra, ER Mardis, LA Fulton, AT Chinwalla, KH Pepin, WR Gish, SL Chissoe, MC Wendl, KD Delehaunty, TL Miner, A Delehaunty, JB Kramer, LL Cook, RS Fulton, DL Johnson, PJ Minx, SW Clifton, T Hawkins, E Branscomb, P Predki, P Richardson, S Wenning, T Slezak, N Doggett, JF Cheng, A Olsen, S Lucas, C Elkin, E Uberbacher, M Frazier, RA Gibbs, DM Muzny, SE Scherer, JB Bouck, EJ Sodergren, KC Worley, CM Rives, JH Gorrell, ML Metzker, SL Naylor, RS Kucherlapati, DL Nelson, GM Weinstock, Y Sakaki, A Fujiyama, M Hattori, T Yada, A Toyoda, T Itoh, C Kawagoe, H Watanabe, Y Totoki, T Taylor, J Weissenbach, R Heilig, W Saurin, F Artiguenave, P Brottier, T Bruls, E Pelletier, C Robert, P Wincker, DR Smith, L Doucette-Stamm, M Rubenfield, K Weinstock, HM Lee, J Dubois, A Rosenthal, M Platzer, G Nyakatura, S Taudien, A Rump, H Yang, J Yu, J Wang, G Huang, J Gu, L Hood, L Rowen, A Madan, S Qin, RW Davis, NA Federspiel, AP Abola, MJ Proctor, RM Myers, J Schmutz, M Dickson, J Grimwood, DR Cox, MV Olson, R Kaul, C Raymond, N Shimizu, K Kawasaki, S Minoshima, GA Evans, M Athanasiou, R Schultz, BA Roe, F Chen, H Pan, J Ramser, H Lehrach, R Reinhardt, WR McCombie, M de la Bastide, N Dedhia, H Blocker, K Hornischer, G Nordsiek, R Agarwala, L Aravind, JA Bailey, A Bateman, S Batzoglou, E Birney, P Bork, DG Brown, CB Burge, L Cerutti, HC Chen, D Church, M Clamp, RR Copley, T Doerks, SR Eddy, EE Eichler, TS Furey, J Galagan, JG Gilbert, C Harmon, Y Hayashizaki, D Haussler, H Hermjakob, K Hokamp, W Jang, LS Johnson, TA Jones, S Kasif, A Kaspryzk, S Kennedy, WJ Kent, P Kitts, EV Koonin, I Korf, D Kulp, D Lancet, TM Lowe, A McLysaght, T Mikkelsen, JV Moran, N Mulder, VJ Pollara, CP Ponting, G Schuler, J Schultz, G Slater, AF Smit, E Stupka, J Szustakowski, D Thierry-Mieg, J Thierry-Mieg, L Wagner, J Wallis, R Wheeler, A Williams, YI Wolf, KH Wolfe, SP Yang, RF Yeh, F Collins, MS Guyer, J Peterson, A Felsenfeld, KA Wetterstrand, A Patrinos, MJ Morgan, P de Jong, JJ Catanese, K Osoegawa, H Shizuya, S Choi, YJ Chen, International Human Genome Sequencing Consortium: Initial sequencing and analysis of the human genome. *Nature* 409, 860-921 (2001)

15. PL Deininger, MA Batzer: Mammalian retroelements. *Genome Res* 12, 1455-1465 (2002)

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

17. RE Mills, EA Bennett, RC Iskow, SE Devine: Which transposable elements are active in the human genome? *Trends Genet* 23, 183-191 (2007)

18. AL Price, E Eskin, PA Pevzner: Whole-genome analysis of Alu repeat elements reveals complex evolutionary history. *Genome Res* 14, 2245-2252 (2004)

19. N Bannert, R Kurth: Retroelements and the human genome: new perspectives on an old relation. *Proc Natl Acad Sci U S A* 101 Suppl 2, 14572-14579 (2004)

20. N Bannert, R Kurth: The evolutionary dynamics of human endogenous retroviral families. *Annu Rev Genomics Hum Genet* 7, 149-173 (2006)

21. R Belshaw, AL Dawson, J Woolven-Allen, J Redding, A Burt, M Tristem: Genomewide screening reveals high levels of insertional polymorphism in the human endogenous retrovirus family HERV-K(HML2): implications for present-day activity. *J Virol* 79, 12507-12514 (2005)

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

23. AB Carter, AH Salem, DJ Hedges, CN Keegan, B Kimball, JA Walker, WS Watkins, LB Jorde, MA Batzer:

Genome-wide analysis of the human Alu Yb-lineage. *Hum Genomics* 1, 167-178 (2004)

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

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

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

27. MK Konkel, J Wang, P Liang, MA Batzer: Identification and characterization of novel polymorphic LINE-1 insertions through comparison of two human genome sequence assemblies. *Gene* 390, 28-38 (2007)

28. J Wang, W Wang, R Li, Y Li, G Tian, L Goodman, W Fan, J Zhang, J Li, J Zhang, Y Guo, B Feng, H Li, Y Lu, X Fang, H Liang, Z Du, D Li, Y Zhao, Y Hu, Z Yang, H Zheng, I Hellmann, M Inouye, J Pool, X Yi, J Zhao, J Duan, Y Zhou, J Qin, L Ma, G Li, Z Yang, G Zhang, B Yang, C Yu, F Liang, W Li, S Li, D Li, P Ni, J Ruan, Q Li, H Zhu, D Liu, Z Lu, N Li, G Guo, J Zhang, J Ye, L Fang, Q Hao, Q Chen, Y Liang, Y Su, A San, C Ping, S Yang, F Chen, L Li, K Zhou, H Zheng, Y Ren, L Yang, Y Gao, G Yang, Z Li, X Feng, K Kristiansen, GK Wong, R Nielsen, R Durbin, L Bolund, X Zhang, S Li, H Yang, J Wang: The diploid genome sequence of an Asian individual. *Nature* 456, 60-65 (2008)

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

30. J Lee, R Cordaux, K Han, J Wang, DJ Hedges, P Liang, MA Batzer: Different evolutionary fates of recently integrated human and chimpanzee LINE-1 retrotransposons. *Gene* 390, 18-27 (2007)

31. MC Seleme, MR Vetter, R Cordaux, L Bastone, MA Batzer, HH Kazazian Jr: Extensive individual variation in L1 retrotransposition capability contributes to human genetic diversity. *Proc Natl Acad Sci U S A* 103, 6611-6616 (2006)

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

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

34. SS Arcot, TH Shaikh, J Kim, L Bennett, M Alegria-Hartman, DO Nelson, PL Deininger, MA Batzer: Sequence diversity and chromosomal distribution of "young" Alu repeats. *Gene* 163, 273-278 (1995)

35. MA Batzer, SS Arcot, JW Phinney, M Alegria-Hartman, DH Kass, SM Milligan, C Kimpton, P Gill, M Hochmeister, PA Ioannou, RJ Herrera, DA Boudreau, WD Scheer, BJ Keats, PL Deininger, M Stoneking: Genetic variation of recent Alu insertions in human populations. *J Mol Evol* 42, 22-29 (1996)

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

37. MA Batzer, VA Gudi, JC Mena, DW Foltz, RJ Herrera, PL Deininger: Amplification dynamics of human-specific (HS) Alu family members. *Nucleic Acids Res* 19, 3619-3623 (1991)

38. JM Kidd, T Graves, TL Newman, R Fulton, HS Hayden, M Malig, J Kallicki, R Kaul, RK Wilson, EE Eichler: A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell* 143, 837-847 (2010)

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

40. Y Miki, I Nishisho, A Horii, Y Miyoshi, J Utsunomiya, KW Kinzler, B Vogelstein, Y Nakamura: Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. *Cancer Res* 52, 643-645 (1992)

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

42. R Druker, E Whitelaw: Retrotransposon-derived elements in the mammalian genome: a potential source of disease. *J Inherit Metab Dis* 27, 319-330 (2004)

43. JC Venter, MD Adams, EW Myers, PW Li, RJ Mural, GG Sutton, HO Smith, M Yandell, CA Evans, RA Holt, JD Gocayne, P Amanatides, RM Ballew, DH Huson, JR Wortman, Q Zhang, CD Kodira, XH Zheng, L Chen, M Skupski, G Subramanian, PD Thomas, J Zhang, GL Gabor Miklos, C Nelson, S Broder, AG Clark, J Nadeau, VA McKusick, N Zinder, AJ Levine, RJ Roberts, M Simon, C Slayman, M Hunkapiller, R Bolanos, A Delcher, I Dew, D Fasulo, M Flanigan, L Florea, A Halpern, S Hannenhalli, S Kravitz, S Levy, C Mobarry, K Reinert, K Remington, J Abu-Threideh, E Beasley, K Biddick, V Bonazzi, R Brandon, M Cargill, I Chandramouliswaran, R Charlab, K Chaturvedi, Z Deng, V Di Francesco, P Dunn, K Eilbeck, C

Evangelista, AE Gabrielian, W Gan, W Ge, F Gong, Z Gu, P Guan, TJ Heiman, ME Higgins, RR Ji, Z Ke, KA Ketchum, Z Lai, Y Lei, Z Li, J Li, Y Liang, X Lin, F Lu, GV Merkulov, N Milshina, HM Moore, AK Naik, VA Narayan, B Neelam, D Nusskern, DB Rusch, S Salzberg, W Shao, B Shue, J Sun, Z Wang, A Wang, X Wang, J Wang, M Wei, R Wides, C Xiao, C Yan, A Yao, J Ye, M Zhan, W Zhang, H Zhang, Q Zhao, L Zheng, F Zhong, W Zhong, S Zhu, S Zhao, D Gilbert, S Baumhueter, G Spier, C Carter, A Cravchik, T Woodage, F Ali, H An, A Awe, D Baldwin, H Baden, M Barnstead, I Barrow, K Beeson, D Busam, A Carver, A Center, ML Cheng, L Curry, S Danaher, L Davenport, R Desilets, S Dietz, K Dodson, L Doup, S Ferriera, N Garg, A Gluecksmann, B Hart, J Haynes, C Haynes, C Heiner, S Hladun, D Hostin, J Houck, T Howland, C Ibegwam, J Johnson, F Kalush, L Kline, S Koduru, A Love, F Mann, D May, S McCawley, T McIntosh, I McMullen, M Moy, L Moy, B Murphy, K Nelson, C Pfannkoch, E Pratts, V Puri, H Qureshi, M Reardon, R Rodriguez, YH Rogers, D Romblad, B Ruhfel, R Scott, C Sitter, M Smallwood, E Stewart, R Strong, E Suh, R Thomas, NN Tint, S Tse, C Vech, G Wang, J Wetter, S Williams, M Williams, S Windsor, E Winn-Deen, K Wolfe, J Zaveri, K Zaveri, JF Abril, R Guigo, MJ Campbell, KV Sjolander, B Karlak, A Kejariwal, H Mi, B Lazareva, T Hatton, A Narechania, K Diemer, A Muruganujan, N Guo, S Sato, V Bafna, S Istrail, R Lippert, R Schwartz, B Walenz, S Yooseph, D Allen, A Basu, J Baxendale, L Blick, M Caminha, J Carnes-Stine, P Caulk, YH Chiang, M Coyne, C Dahlke, A Mays, M Dombroski, M Donnelly, D Ely, S Esparham, C Fosler, H Gire, S Glanowski, K Glasser, A Glodek, M Gorokhov, K Graham, B Gropman, M Harris, J Heil, S Henderson, J Hoover, D Jennings, C Jordan, J Jordan, J Kasha, L Kagan, C Kraft, A Levitsky, M Lewis, X Liu, J Lopez, D Ma, W Majoros, J McDaniel, S Murphy, M Newman, T Nguyen, N Nguyen, M Nodell, S Pan, J Peck, M Peterson, W Rowe, R Sanders, J Scott, M Simpson, T Smith, A Sprague, T Stockwell, R Turner, E Venter, M Wang, M Wen, D Wu, M Wu, A Xia, A Zandieh, X Zhu: The sequence of the human genome. *Science* 291, 1304-1351 (2001)

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

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

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

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

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

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

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

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

52. EA Bennett, LE Coleman, C Tsui, WS Pittard, SE Devine: Natural genetic variation caused by transposable elements in humans. *Genetics* 168, 933-951 (2004)

53. S Levy, G Sutton, PC Ng, L Feuk, AL Halpern, BP Walenz, N Axelrod, J Huang, EF Kirkness, G Denisov, Y Lin, JR MacDonald, AW Pang, M Shago, TB Stockwell, A Tsiamouri, V Bafna, V Bansal, SA Kravitz, DA Busam, KY Beeson, TC McIntosh, KA Remington, JF Abril, J Gill, J Borman, YH Rogers, ME Frazier, SW Scherer, RL Strausberg, JC Venter: The diploid genome sequence of an individual human. *PLoS Biol* 5, e254 (2007)

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

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

56. J Kaiser: DNA sequencing. A plan to capture human diversity in 1000 genomes. *Science* 319, 395 (2008)

57. N Siva: 1000 Genomes project. *Nat Biotechnol* 26, 256 (2008)

58. AD Ewing, HH Kazazian: Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans. *Genome Res* (2010)

59. 1000 Genomes Project Consortium, RM Durbin, GR Abecasis, DL Altshuler, A Auton, LD Brooks, RM Durbin, RA Gibbs, ME Hurles, GA McVean: A map of human genome variation from population-scale sequencing. *Nature* 467, 1061-1073 (2010)

60. F Hormozdiari, C Alkan, M Ventura, I Hajirasouliha, M Malig, F Hach, D Yorukoglu, P Dao, M Bakhshi, SC

Sahinalp, EE Eichler: Alu repeat discovery and characterization within human genomes. *Genome Res* (2010)

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

62. SC Schuster, W Miller, A Ratan, LP Tomsho, B Giardine, LR Kasson, RS Harris, DC Petersen, F Zhao, J Qi, C Alkan, JM Kidd, Y Sun, DI Drautz, P Bouffard, DM Muzny, JG Reid, LV Nazareth, Q Wang, R Burhans, C Riemer, NE Wittekindt, P Moorjani, EA Tindall, CG Danko, WS Teo, AM Buboltz, Z Zhang, Q Ma, A Oosthuysen, AW Steenkamp, H Oostuisen, P Venter, J Gajewski, Y Zhang, BF Pugh, KD Makova, A Nekrutenko, ER Mardis, N Patterson, TH Pringle, F Chiaromonte, JC Mullikin, EE Eichler, RC Hardison, RA Gibbs, TT Harkins, VM Hayes: Complete Khoisan and Bantu genomes from southern Africa. *Nature* 463, 943-947 (2010)

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

64. JP Noonan, G Coop, S Kudaravalli, D Smith, J Krause, J Alessi, F Chen, D Platt, S Paabo, JK Pritchard, EM Rubin: Sequencing and analysis of Neanderthal genomic DNA. *Science* 314, 1113-1118 (2006)

65. DJ Witherspoon, EE Marchani, WS Watkins, CT Ostler, SP Wooding, BA Anders, JD Fowlkes, S Boissinot, AV Furano, DA Ray, AR Rogers, MA Batzer, LB Jorde: Human population genetic structure and diversity inferred from polymorphic L1(LINE-1) and Alu insertions. *Hum Hered* 62, 30-46 (2006)

66. NT Perna, MA Batzer, PL Deininger, M Stoneking: Alu insertion polymorphism: a new type of marker for human population studies. *Hum Biol* 64, 641-648 (1992)

67. JT den Dunnen, SE Antonarakis: Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. *Hum Mutat* 15, 7-12 (2000)

68. JT den Dunnen, SE Antonarakis: Nomenclature for the description of human sequence variations. *Hum Genet* 109, 121-124 (2001)

69. OK Pickeral, W Makalowski, MS Boguski, JD Boeke: Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. *Genome Res* 10, 411-415 (2000)

70. F Charbonnier, S Baert-Desurmont, P Liang, F Di Fiore, C Martin, S Frerot, S Olschwang, Q Wang, MP Buisine, B Gilbert, M Nilbert, A Lindblom, T Frebourg: The 5' region of the MSH2 gene involved in hereditary non-polyposis colorectal cancer contains a high density of recombinogenic sequences. *Hum Mutat* 26, 255-261 (2005)

71. JF Hughes, JM Coffin: Human endogenous retrovirus K solo-LTR formation and insertional polymorphisms: implications for human and viral evolution. *Proc Natl Acad Sci U S A* 101, 1668-1672 (2004)

72. AR Jha, SK Pillai, VA York, ER Sharp, EC Storm, DJ Wachter, JN Martin, SG Deeks, MG Rosenberg, DF Nixon, KE Garrison: Cross-sectional dating of novel haplotypes of HERV-K 113 and HERV-K 115 indicate these proviruses originated in Africa before Homo sapiens. *Mol Biol Evol* 26, 2617-2626 (2009)

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

74. AJ Iafrate, L Feuk, MN Rivera, ML Listewnik, PK Donahoe, Y Qi, SW Scherer, C Lee: Detection of large-scale variation in the human genome. *Nat Genet* 36, 949-951 (2004)

75. K Akagi, RM Stephens, J Li, E Evdokimov, MR Kuehn, N Volfovsky, DE Symer: MouseIndelDB: a database integrating genomic indel polymorphisms that distinguish mouse strains. *Nucleic Acids Res* 38, D600-6 (2010)

76. JS Mattick, RJ Taft, GJ Faulkner: A global view of genomic information--moving beyond the gene and the master regulator. *Trends Genet* 26, 21-28 (2010)

77. SH Rangwala, L Zhang, HH Kazazian Jr: Many LINE1 elements contribute to the transcriptome of human somatic cells. *Genome Biol* 10, R100 (2009)

78. R Khaja, J Zhang, JR MacDonald, Y He, AM Joseph-George, J Wei, MA Rafiq, C Qian, M Shago, L Pantano, H Aburatani, K Jones, R Redon, M Hurles, L Armengol, X Estivill, RJ Mural, C Lee, SW Scherer, L Feuk: Genome assembly comparison identifies structural variants in the human genome. *Nat Genet* 38, 1413-1418 (2006)

79. JM Chen, C Ferec, DN Cooper: Mechanism of Alu integration into the human genome. *Genomic Med* 1, 9-17 (2007)

80. JT Simpson, K Wong, SD Jackman, JE Schein, SJ Jones, I Birol: ABySS: a parallel assembler for short read sequence data. *Genome Res* 19, 1117-1123 (2009)

81. D Karolchik, G Bejerano, AS Hinrichs, RM Kuhn, W Miller, KR Rosenbloom, AS Zweig, D Haussler, WJ Kent: Comparative genomic analysis using the UCSC genome browser. *Methods Mol Biol* 395, 17-34 (2007)

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

83. D Karolchik, AS Hinrichs, TS Furey, KM Roskin, CW Sugnet, D Haussler, WJ Kent: The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* 32, D493-6 (2004)

Abbreviations: RE: retrotransposon element, RIPs: retrotransposon insertion polymorphisms, dbRIP: database of retrotransposon insertion polymorphism; TSD: target site duplication

Key Words: Retrotransposon, Mobile elements, DNA transposition, Database, Computational comparative genomics, Polymorphism, dbRIP, Human, Review

Send correspondence to: Ping Liang, Department of Biological Sciences, Brock University, 500 Glenridge Avenue, St. Catharines, Ontario, Canada L2S 3A1, Tel: 905-688-5550 Ext 5922, Fax: 905-688-1855, E-mail: pliang@brocku.ca

http://www.bioscience.org/current/vol4E.htm