Human Endogenous Retrovirus (HERV) Insertional Polymorphisms

by

Scott Matthew Bradley Golem

A Thesis

submitted to the Department of Biological Sciences

in partial fulfillment of the requirements

for the degree of

Master of Science

June, 2013

Brock University

St. Catharines, Ontario

Abstract

Human endogenous retroviruses (HERVs) are the result of ancient germ cell infections of human germ cells by exogenous retroviruses. HERVs belong to the long terminal repeat (LTR) group of retrotransposons that comprise ~8% of the human genome. The majority of the HERVs documented have been truncated and/or incurred lethal mutations and no longer encode functional genes; however a very small number of HERVs seem to maintain functional in making new copies by retrotranspositon as suggested by the identification of a handful of polymorphic HERV insertions in human populations. The objectives of this study were to identify novel insertion of HERVs via analysis of personal genomic data and survey the polymorphism levels of new and known HERV insertions in the human genome. Specifically, this study involves the experimental validation of polymorphic HERV insertion candidates predicted by personal genome-based computation prediction and survey the polymorphism level within the human population based on a set of 30 diverse human DNA samples. Based on computational analysis of a limited number of personal genome sequences, PCR genotyping aided in the identification of 15 dimorphic, 2 trimorphic and 5 fixed full-length HERV-K insertions not previously investigated. These results suggest that the proliferation rate of HERVKs, perhaps also other ERVs, in the human genome may be much higher than we previously appreciated and the recently inserted HERVs exhibit a high level of instability. Throughout this study we have observed the frequent presence of additional forms of genotypes for these HERV insertions, and we propose for the first time the establishment of new genotype reporting nomenclature to reflect all possible combinations of the pre-integration site, solo-LTR and full-length HERV alleles.

Acknowledgements

I am heartily thankful to my supervisor, Ping Liang, whose encouragement, guidance and support from the initial to the final level enabled me to develop an understanding of the subject. It is a pleasure to thank my committee members who have made this thesis possible Mike Bidochka and Jeffrey Atkinson. I would also like to acknowledge the encouragement of my colleagues in the lab, Amanda Bering, Daniel Tang, Musa Ahmed, as well as Xuemei Luo for her contributions to the data set used in this project. I would like to thank my friends and family for their continued support in all my endeavours. Finally I would like to thank my beautiful wife Monika Ovsonka for all of her support throughout the whole process, and my daughter Adele for always making me smile when things got tough.

Table of Contents

List of Tables

List of Figures

List of abbreviations

BLAST-like alignment tool (BLAT)

Copy number variation (CNV)

Database of retrotransposon insertion polymorphisms (dbRIP)

Deoxyribonucleic acid (DNA)

Endogenous retrovirus (ERV)

Full-length (FL)

Human endogenous retrovirus (HERV)

Human MMTV-like (HML)

Long terminal repeat (LTR)

Mouse mammary tumor virus (MMTV)

Next generation sequencing (NGS)

Open reading frame (ORF)

Polymerase chain reaction (PCR)

Pre-integration site (PI)

Retrotransposon element (RE)

Ribonucleic acid (RNA)

Single nucleotide polymorphism (SNP)

Solo-long terminal repeat (sLTR)

Target site duplication (TSD)

Transposable element (TE)

Transposable insertion polymorphisms in the human reference genome but absent in the test genome (TIPs_IN)

Transposable insertion polymorphisms absent from the human reference genome but present in the test genome (TIPs_OUT)

Introduction

Overview of transposable elements

Transposable elements (TE) (often referred to as "jumping genes") are sequences of DNA that are able to mediate their movement throughout the genome either via retrotransposition or by splicing itself and moving to a different location within the genome (Cordaux 2009). Since Barbara McClintock first described TEs in 1953 while studying the mosaic colouration of maize, it has become well established that such elements are universal (McClintock 1953, Goodier 2008, Medstrand 2002). Through the analysis of the many completed genome sequences, TEs are estimated to account for a major proportion of the plant and animal genomes, with approximately 50% for the human genome (Lander 2001). With such a high number of insertions and the ongoing activity for some of these members, they are thought to contribute significantly to the inter- and intra-species genetic variation.

There are two major classes which TEs can be separated into based on their method of transposition within the genome: DNA transposons and retrotransposons (Cordaux 2009). DNA transposons are able to move and insert themselves into new genomic sites in a "cut and paste" fashion while remaining as DNA (Stoye 2001). DNA transposons were active during early primate evolution (~37 million years ago), but are currently thought to be immobile in the human genome (Griffiths 2001). They are estimated to account for ~3 percent of the human genome (Cordaux 2009). Conversely, retrotransposons duplicate through a "copy and paste" technique using RNA intermediates which are reverse transcribed before inserting into a new genomic location. Retrotransposons are currently estimated to account for ~40 percent of the human genome (Cordaux 2009). The class of retrotransposons is further divided into two sub-groups which are identified by the presence or absence of long terminal repeats (LTRs) (Bannert 2006).

Non-LTR retrotransposons represent the majority of TEs present in the human genome, collectively accounting for approximately one-third (Cordaux 2009). The main elements composing the sub-group of non-LTR retrotransposons are: long interspersed element 1 (LINE-1) and Alu. Each of these has been indisputably shown as currently active within humans; with more than 80 reported cases of *de novo* insertions causing genetic disorders. The current estimated retrotransposition rates are as follows: Alu approximately one in twenty live births, LINE-1 approximately one in 200 live births (Konkel 2010), serving probably only as very rough guidelines.

The LTR retrotransposons sub-category is composed of endogenous retroviruses (ERVs), which are estimated to account for approximately 8 percent of the human genome (Goodier 2008). Some of these elements have integrated exclusively within the human genome and thus also called human endogenous retroviruses (HERVs). Many of the HERVs are thought to have inserted into the human genome roughly 25-30 million years ago (Cordaux 2009, Macfarlane 2004). Unlike other mammals such as mice, cats and sheep, there are currently no active infectious ERVs reported in humans (Bannert 2006). Although their activity is currently thought to be very limited, the youngest elements HERV-K(HML-2) have shown evidence of persisting activity (Belshaw 2005, Turner 2001, Hughes 2004).

Origin and classification of ERVS and HERVS

Among RNA viruses, retroviruses are unique in their ability to integrate DNA copies of their genome into the genome of the infected cell (Johnson, Coffin 1999). Each retrovirus genome is composed of two copies of positive single-strand RNAs containing mainly four genes in the order 5'-*gag-pro-pol-env*-3'. *Gag* encodes the matrix and capsid proteins, *pro* encodes the protease, *pol* encodes the reverse transcriptase and integrase, and *env* encodes the surface envelope proteins (Figure 1) (Bannert 2006). Very few exogenous viruses (such as HIV) possess additional non-structural accessory genes that facilitate their replication or impair host defences. It should be noted that these accessory genes are rare among endogenous viruses, with the exception of HERV-K (see ERV Classification below) (Kurth 2010). After infection, a cellular tRNA molecule is used as a primer by the co-packaged reverse transcriptase in order to reverse transcribe the RNA into the ensuing double-stranded cDNA and virion protein (Sverdlov 2000, Bannert 2006). Next the integrase mediates the insertion of the DNA genome into the host chromosomal DNA at a seemingly randomized location, while generating a duplication of short sequence at the genomic integration site, flanking each LTR (Bannert 2006). Generally these insertions occur in somatic cells, and passed onto all progeny cells (Bannert 2006). If this integration takes place within a germ line cell, it will give rise to an endogenous retrovirus (ERV) in the genome of a new birth that is derived from the gamete carrying this allele (Macfarlane 2004). When this occurs, the insertion is passed vertically from the infected host to their offspring according to Mendelian laws (Bannert 2006). It is important to note that there are no reports of the eradication of an ERV from an infected host (Johnson, Coffin 1999). This has resulted in the chromosomes of mammals and most other vertebrates to be interlaced with ERV sequences, some considered ancient by the identical site of integration present in more than one species; whereas other have been acquired in more evolutionarily recent times as being specific to one species or even some individuals within the species (Tristem 2000).

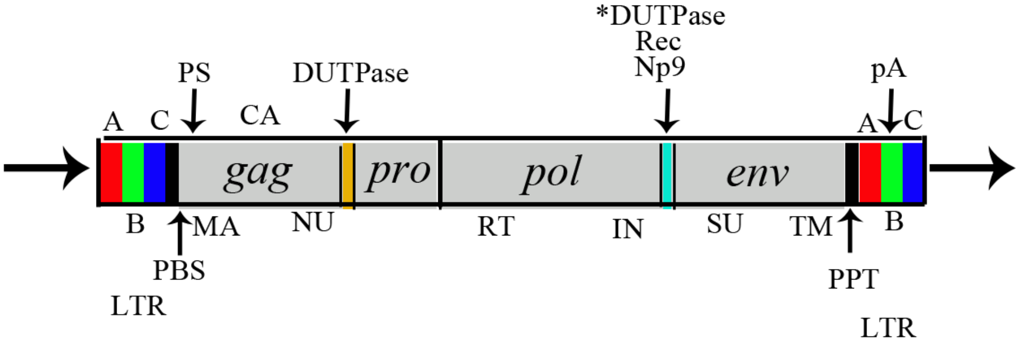

Figure 1: Generalized organization of an integrated HERV (provirus).

The viral sequence is flanked by a short duplication of host DNA produced during the integration process. The long terminal repeats (LTRs) at the 5′ and 3′ ends are composed of the A, B, and C regions. Transcription starts in the B region of the 5′-LTR and the polyadenylation signal (pA) resides at the end of the B region in the 3′-LTR. The primer binding site (PBS), commonly used to classify HERVs, is located in front of the *gag* reading frame. The *gag* portion encodes the structural portion of the viral matrix (MA), capsid (CA), and nucleocapsid (NU). It is processed by the product of the subsequent protease (*pro*). The reverse transcriptase (RT) and integrase (IN) are part of the polymerase (*pol*). The envelope protein (*env*) consists of the surface (SU) and transmembrane (TM) units and is translated from a spliced transcript. Complex retroviruses encode accessory proteins (e.g., Rec, Np9) at the *pol/env* junction. The location of a dUTPase encoded by several retroviruses, the position of the packaging signal (PS), and the polypurine tract (PPT) are also shown (* indicates another site of dUTPase in some retroviruses) [adapted from (Bannert 2006)].

ERV classification has become extremely complex as a result of different research teams providing arbitrary nomenclatures using a variety of classification criteria such as specificity of tRNA primer binding site, morphological type and copy numbers (Blomberg, Benachenhou et al. 2009). Many of these classifications are directly related to the small number of detected HERV sequences as well as the limited knowledge and methods used to study them at the time (Blomberg, Benachenhou et al. 2009). This ambiguity has led to imprecise naming, as well as the subsequent overlapping of different family classifications as outlined by Blomberg et al. 2009. Current retrovirus taxonomy divides the status of a family into one of the seven genera: alpha-, beta-, gamma-, delta-, epsilon-, lenti- and spuma-retroviruses (Bannert 2006). ERVs are currently loosely grouped into three classes due to their phylogenetic relatedness to exogenous viruses (Bannert 2006). Class I is composed of viruses that cluster with the gamma- and epsilon-genera, Class II ERVs are related to beta-retroviruses or very distantly to delta- and lenti-viruses, while Class III is comprised of those most similar to spuma-viruses (Bannert 2006). Although many ERVs belong to Class II, it is important to note that there are none that are known to be closely related to lenti-viruses (e.g. HIV) (Bannert 2006). Given that a traditional criteria used to name and classify a new human ERV was the sequence of 18 nucleotides constituting the primer binding site used to initiate reverse transcription, most HERVs are further organised by adding the one-letter code of the amino acid specificity of the most likely tRNA as a suffix to the acronym HERV (Bannert 2006). Therefore a provirus using a lysine tRNA would be classified as HERV-K, whereas HERV-W would use tryptophan (Moyes 2007). All of the Class II HERVs contain a lysine tRNA primer binding site, leading to the alternative nomenclature of simply HERV-K (Nelson, Carnegie et al. 2003). Although mutational events have rendered most HERVs replication defective following integration, the HERV-Ks are thought to be the most active class of HERV, as they have retained the ability to encode a functional retroviral protein (Macfarlane 2004). The HERV-K clade of beta retrovirus-like endogenous retroviruses currently contains a total of ten groups ranging from HML-1 to HML-10 (Subramanian, Wildschutte et al. 2011). They are most closely related to the mouse mammary tumor virus (MMTV), which is a causative agent for breast cancer in mice, leading to their acronym generated from human MMTV-like (Nelson, Carnegie et al. 2003). Genome-wide studies have shown that the most recently active retroviruses belong to the HML-2 group, which has been estimated to include ~60 proviruses and over 2500 solo-LTRs in the human genome (Subramanian, Wildschutte et al. 2011). The HML-2 group is further classified into type 1 or type 2 based on the presence or absence of a 292bp deletion at the *pol-env* junction respectively

(Subramanian, Wildschutte et al. 2011). HML-2 elements are distinguished from their progenitor HERV-K(OLD) by their 96bp deletion in *gag* which has not disrupted the open reading frame, as well as a 8 to 23bp deletion found within their LTRs (Macfarlane 2004). It is estimated that the HML-2 group integrated into the germ line roughly 28 million years ago, before the evolutionary divergence of lower Old World primates and hominoids (Macfarlane 2004). The HML-2 group is unique among all ERVs being the only group that includes human-specific proviruses, of which 11 are known to be insertionally polymorphic within the human population (Barbulescu, Turner et al. 1999, Turner 2001, Costas 2001, Hughes 2004, Belshaw 2005). The insertion rate of the HML-2 group appears to have been fairly constant since the *Homo-Pan* divergence, lending to the evidence that replication competent HML-2 viruses may still exist within the human population (Subramanian, Wildschutte et al. 2011). Although there has been no evidence of an infectious HERV in humans to date, other vertebrates have been found to contain replication-competent ERVs (Subramanian, Wildschutte et al. 2011).

Polymorphism levels of HERVs in humans

Upon integration, the ERV may retain the potential to be both vertically (parent to offspring) and horizontally (re-infection) transmitted. Both the length of this stage and the proviral frequency reached in the host population are mainly determined by the effects that the integration has on the fitness of the host (Bannert 2006). If the integration is neutral or in some way beneficial to the host, the allelic frequency is more likely to increase. Conversely, the integration will not reach high allelic frequencies or remain in the host population for long if it causes strong detrimental or pathogenic effects (Bannert 2006). For a detrimental proviral insertion to become fixed within a host population it must be preceded by either partial or complete inactivation of the insertion (Bannert 2006). Unless selective pressure ensures the retention of some functionality during evolution, the provirus will eventually be subjected to random modifications within the host genome resulting in the loss of expression and the capacity to proliferate (Bannert 2006). These integrated proviruses then become fossilised in the host genome at which point further decay will eventually render the retroviral sequence barely recognizable. It should be noted that the likelihood that two independent integrations will occur at the same chromosomal position is essentially negligible (Stoye 2001). An increase in the provirus allele may also occur as a result of genetic drift, where a population bottle neck may rapidly alter the allelic frequencies by possibly sparing more individuals with a specific integration than those without (Bannert 2006). This can also result from a founder effect, where a single individual or small group of provirus carriers create a population burst. Eventually, an advantageous or neutral integration may become fixed within the genome of a species causing the loss of the insert-free allele (Bannert 2006).

The activity of an ERV can be abolished through a variety of reversible and irreversible mechanisms. The most drastic inactivation is caused by a homologous recombination between the two LTRs, resulting in the deletion of all viral sequences but a single chimeric LTR, termed a solo-LTR (Medstrand 2002, Vitte 2003). Compared to their full-length ERV counterparts, it is estimated that solo-LTRs are approximately ten-fold more abundant within the human genome (Stoye 2001). During host replication, mutations, deletions and recombination can also lead to the inactivation of their transcription regulatory elements and loss of protein function via non-synonymous or non-sense mutations (Belshaw 2005). Hypermethylation of ERV promoters can also quickly silence provirus expression (Moyes 2007).

Currently there are only 11 HERVs that have been found to be insertionally polymorphic within the human population (Barbulescu, Turner et al. 1999, Turner 2001, Costas 2001, Hughes

2004, Belshaw 2005). Barbulescu et al. (1999) were the first group to identify a polymorphic HERV. In their study they were able to identify five loci where the full-length provirus and sLTR were found among the human samples but absent in the *P.pygmaeus, G. gorilla*, *P. paniscus* and *P. troglodytes* samples tested. A study done by Costas (2001) found one polymorphic sLTR present among the human individuals tested. In contrast Belshaw et al. (2005) were able to identify one polymorphic loci where both the sLTR and full-length provirus alleles were present in the individuals tested. Arguably the most monumental finding with regards to HERV insertional polymorphism came from Turner et al. (2001) when they found two full-length proviruses to be polymorphic among a subset of human individuals they tested. These loci were named HERV-K113 and HERV-K115 and it was found that the provirus allele frequencies were 0.19 (9/48) and 0.4 (2/46) respectively, indicating that the insertion occurred fairly recently (Turner 2001). They estimated these two HERVs integrated into the host genome within the last 1 million years (Turner 2001). Interestingly, both HERV-K113 and HERV-K115 were found to have full-length open reading frames (ORF) for all viral proteins. The HERV-K113 element appears to be capable of coding for all structural, regulatory and enzymatic proteins, as there are no detrimental mutations in the full-length provirus sequence (Turner 2001). HERV-K115 has obtained a 1bp deletion located 92bp upstream from the stop codon of the *gag* ORF, causing a frame shift that is likely to result in the inability to translate the *pro* and *pol* ORFs (Turner 2001). Given this, it has been proposed that HERV-K113 represents the best candidate of a provirus that is still active today in humans (Turner 2001).

Functional importance of HERVS in the genome

With few exceptions, retrotransposon insertions are neutral or in some instance detrimental to the host, with the latter likely to be eliminated as a result of negative selection and therefore unlikely to reach a high allele frequency within the population. As HERVs litter our genome, their evolutionary conservation in the host genomes indicates that there must be some beneficial functions HERVs provide to the host. It has been found that HERV-K (HML-2) LTRs contribute to the expression of nearby genes by acting as active promoters for host non-repetitive DNA transcription *in vivo* (Buzdin, Kovalskaya-Alexandrova et al. 2006). Retrotransposon-mediated sequence transduction and gene duplication have also been found to have led to both the creation of novel genes and aiding in the diversity of multi-gene families such as MHC- or T-Cell receptor genes (Brandt, Schrauth et al. 2005, Xing, Wang et al. 2006, Agrawal, Eastman et al. 1998, Doxiadis, De Groot et al. 2008). The reverse transcriptase has been shown to repair chromosomal breaks, and it has been suggested that telomerase is derived from the TE-coded reverse transcriptase (Teng, Kim et al. 1996, Eickbush 1997). HERV LTRs have also been found to contain binding sites for the p53 regulator, accounting for over 30% of these binding sites genome wide (Wang, Zeng et al. 2007). Consequently these HERV LTRs are suspected to contribute to the anti-oncogenic function of the stress-responsive p53 pleiotropic regulator (Wang, Zeng et al. 2007). It has also been found *in vitro* that cellular resistance to infection by exogenous retroviruses can be conferred by the HERV-W envelope glycoproteins (Ponferrada, Mauck et al. 2003). Another study found that HERV-Es are activated in some renal cancer cells, providing target antigens that were recognizable by cytotoxic T-cells after allogeneic hematopoetic stem cell transplantation (Takahashi, Harashima et al. 2008). This resulted in the regression of the tumor, providing evidence that humans are apparently not immunologically tolerant of HERVs (Takahashi, Harashima et al. 2008). The HERV-W and HERV-FRD envelope proteins (syncitin-1 and syncytin-2 respectively) have been detected in the placenta, and are thought to mediate the cell-cell fusion of cytotrophoblasts to syncytiotrophoblasts resulting in the physiological morphogenesis of the placenta (Kurth 2010). The previously demonstrated immunosuppressive property of retroviral ENV proteins has also been demonstrated from syncytin-2, which may be instrumental in fetal-maternal tolerance (Mangeney, Renard et al. 2007). The immunosuppressive and fusogenic endogenous retrovirus proteins have also been detected in sheep and mice, leading to the evidence of positive selection over millions of years (Kurth 2010).

Most, if not all HERV insertions that reach a high allelic frequency in the human population have acquired knockout mutations, deletions or undergone recombination events rendering them inactive (Bannert 2006). Therefore it is not unexpected that the limited investigations have identified the majority of these insertions to be neutral or defective. It is likely that any direct disease causing insertions may be too rare to allow the recognition of any infectious and replication competent (Bannert 2006). A variety of studies have indicated that HERVK (HML-2) expression is up-regulated in tissues associated with a variety of diseases including melanomas, germ cell tumors, breast cancer, ovarian cancer, leukemias/lymphomas, schizophrenia and rheumatoid arthritis, although the functional consequences of this expression remain unknown (Frank, Verbeke et al. 2008, Herbst, Sauter et al. 1998, Büscher, Hahn et al. 2006, Hu, Hornung et al. 2006, Iwabuchi, Kakihara et al. 2004, Dickerson, Rubalcaba et al. 2008, Sicat, Sutkowski et al. 2005, Subramanian, Wildschutte et al. 2011). Despite research in these areas, there has been no clear data on which specific loci are being transcribed, nor the reasons for their activation (Subramanian, Wildschutte et al. 2011). It has been proposed that the HERV immunosuppressive ENV proteins may indirectly facilitate tumor development through inhibition of an immune response, which may explain why high expression levels have been found in the diseases listed above (Nelson, Carnegie et al. 2003).

Methods for detection of HERV polymorphisms

There have been a variety of approaches used for the detection of ERV diversity. One of the earliest approaches was to stimulate ERV replication cells derived from one species and co-cultivate them with appropriate indicator cells from a different species, which isolated replication competent endogenous viruses (Gifford, Tristem 2003). Another approach used to approximate the distribution diversity of ERVs was through the use of low and high-stringency hybridization with retrovirus-derived probes (Gifford, Tristem 2003). They have also been detected using synthetic primer binding site (PBS) probes (Gifford, Tristem 2003). This technique uses the PBS probe hybridization to detect ERV-containing clones in genomic libraries based on bacterial or phage PI artificial chromosomes (BAC and PAC libraries respectively), followed by sequencing of the positive clones. Although this method is very time-consuming, it is able to provide the complete sequence of an ERV insertion (Gifford, Tristem 2003). A more efficient way of studying ERV diversity is to use primers flanking the insertion site, followed by PCR to amplify the novel ERVs from host genomic DNA (Gifford, Tristem 2003). Although this method does not provide the complete sequence of the ERV insertion, it is a very useful tool for providing sufficient data for phylogenetic analysis among the samples tested. PCR validation is now considered the gold standard for HERV polymorphism detection, as it allows the determination of the presence or absence of the insertion by comparing the sizes of the amplified products.

Now with the availability of the complete genome sequences, computational methods are used to investigate this diversity by comparing different genomes. These methods provide the opportunity to investigate the diversity of the ERV sequences as well as the distribution throughout the genome. Computational algorithms such as RepeatMasker have allowed the automatic annotation of ERV insertions, providing a fast and extremely efficient basis for the preliminary analysis of the structural variations found within the assembled genomes. Although there are a variety of computational approaches for detecting structural variations, they are limited in detecting mobile element insertional polymorphisms, as they can only compare assembled genomes. All of the recent genome projects (see below) have been generating genome sequences using next generation sequencing technologies, which result in a vast quantity of unassembled DNA sequence data, making this information useless to the algorithms requiring an assembled genome for comparison. VariationHunter is the only tool that has been developed so far to identify mobile element insertional polymorphisms by comparing the test genome and the human reference genome sequence using the unassembled next generation sequence data (Hormozdiari, Hajirasouliha et al. 2010).

Evolution of sequencing technologies

Biological sciences have been fundamentally transformed by the ability to rapidly determine nucleic acid sequences (Korlach, Bjornson et al. 2010). It has created a landslide of information that has revolutionised the way we think about scientific approaches, and stimulated an immense number of scientific advances. For over two decades the Sanger sequencing method (Sanger 1988) has been responsible for a variety of fundamental accomplishments, one of the most monumental being the completion of the first finished-grade human reference genome sequence (Lander 2001, Venter 2001, Collins, Lander et al. 2004). What was once accomplished over years with a high financial burden using the Sanger method, can now be accomplished in weeks for magnitudes lower in price using the next generation sequencing (NGS) technologies (Bentley 2009).

No matter which method is used, all NGS technologies follow the same three basic phases consisting of sample preparation, physical sequencing and re-assembly (Schadt 2010).

The genomic library is created by randomly fragmenting the template DNA into roughly 1kb long pieces. The fragments are then spatially separated and immobilized to allow the parallel sequencing of thousands to billions of sequencing reactions (Metzker 2010). The 4 main NGS technologies that currently dominate the commercial market are Roche/454, Illumina (Solexa), ABI (SOLid) and Ion Proton (Ion Torrent). Roche/454 is generally the method of choice for applications where long read lengths are critical such as *de novo* sequencing and metagenomics. The Illumina/Solexa sequencing platform is best used for re-sequencing applications.

The ABI SOLid sequencing platform is one of the most reliable technologies for identifying true single nucleotide polymorphisms (SNPs) (Shendure 2008). The Ion Proton is best for smaller runs, and sequencing can occur in real time. The read lengths and data outputs are summarized in

Table 1. The main bottlenecks that NGS face are found within the computational resources needed for assembly, annotation and analysis of sequence reads. NGS technologies currently have read lengths far smaller than the smallest genomes (Miller 2010). Shorter read lengths deliver less information per read, requiring higher coverage to satisfy minimum overlap criteria for assembly (Scholz, Lo et al. 2012). Assembly software is challenged by genomic regions with shared perfect repeats, which can be indistinguishable when the repeats are longer than the read lengths (de Magalhães, Finch et al. 2010). This challenge becomes amplified even further by the raw accuracy of these reads being inferior to Sanger sequencing (Scholz, Lo et al. 2012). In order to combat this issue, assembly software must tolerate imperfect sequence alignments to avoid overlooking library overlaps, which leads to false positives especially with polymorphic repeats (Miller 2010).

Table 1: Comparison of next generation sequencing technologies. Adapted from (Scholz, Lo et al. 2012)

Technology	Average Read Length	Output (Mb)	Run Time
Roche/454	400bp-700bp	500-900	10-20 hours
Illumina/Solexa	35bp-150bp	400,000-600,000	8-14 hours
ABI-SOLiD	35bp-60bp	71,000-155,000	8-12 hours
Ion Proton	100-200bp	10-1000	3 hours

Human genomes

The advancement in sequencing technologies has allowed genome-wide association studies to identify genetic variants and provide insight into those that are associated with human disorders. It has also allowed researchers to focus on the development and validation of prognostic and predictive markers to work towards the goal of personalized medicine (Ziogas 2009). The only finished-grade human reference genome sequence (NCBI build 36) was published in 2004 (Metzker 2010). This genome is estimated to be composed of 99.99% European origin and still contains ~210 gaps (Snyder 2010). With sequencing technologies creating fairly short read lengths, most genome sequencing projects rely on re-sequencing (assembly based on comparison to the reference genome) as opposed to *de novo* sequencing. Currently there are several thousand genome sequences that have been reported, providing valuable data for genome variant studies by comparing personal genomes to the reference genome (Snyder 2010). The genome sequence of J. Craig Venter in 2007 compared to the reference genome identified 3.2 million single-nucleotide variants (SNVs) and 900,000 structural variants (SV). The diploid genome of James D. Watson became the first whole genome to be sequenced using next generation sequencing technologies (Metzker 2010). Comparison of the

James D. Watson genome with the reference identified 3.3 million SNVs (Wheeler, Srinivasan et al. 2008). The 1000 Genomes Project (http://www.1000genomes.org/) began in January 2008, with the goal of sequencing the complete genomes of 1000 individuals from around the world using the NGS technologies. This project aims to provide the genome data needed to discover and characterize the variants and polymorphisms with a frequency of at least 1% in the human genome by comparing each of the 5 major population groups that include: ancestry from Europe, East Asia, South Asia, West Africa and the Americas. The project has completed the pilot phase which included a component known as the Trio Project, where whole-genome shotgun sequencing at a high coverage rate (averaging 42x) of two families (one Yoruba from Ibaden, Nigeria; one of European ancestry in Utah) that each included the two parents and one daughter (Altshuler, Lander et al. 2010). The sequence read data generated using mainly the Illumina/Solexa and Roche/454 platforms. All of these sequence reads have been made available to the scientific community through their website. This is the data that the computational prediction of this study is based on.

Objectives

Using the personal genome sequence data generated from the next-generation sequencing platforms to study transposable element-derived structural variations represents an emerging and very promising direction for genetic and genomics research. Current data on polymorphic HERVs is extremely small. However, we believe that HERVs remain a certain level of transposition activity in the human genome and their level of polymorphism may be much higher than currently known. The main objectives of this study are to explore different approaches to discovering novel polymorphic HERV insertions aided by computational prediction from the analysis of available personal genome sequence data and to provide further characterization of the newly identified HERV insertions in sequence and allele frequency. These will allow us to assess the efficiency and accuracy of each approach, lending insight into the most efficient future discovery method and providing a more accurate assessment of the activity and polymorphism level of HERV insertions in the human genome, as well as their potential impact on genome function and evolution.

Materials and Methods

The prediction methods chosen to be evaluated involve three distinct approaches. Two of the methods involve computational prediction to assemble, align and compare the genomes of 6 individuals to the human reference genome. The third method involves screening the full-length HERV sequences identified within the human reference genome via PCR genotyping to examine if any insertional variation can be detected at that locus.

Generation of TIPs_IN candidate list

The TIPs_IN list of candidate polymorphic HERV loci represent insertions that are present in the human reference genome but absent in one or more of the donor human genomes. This list was generated by Dr. Ping Liang via computational comparative genomics analysis, using the personal genome sequence data generated by the 1000 Genomes Project, based on the 6 individuals from the two trio families, for which deep sequencing data were generated. These families are from Yoruba in Ibaden, Nigeria and from European ancestry located in Utah, USA, with each consisting of the mother, father and a daughter. Paired-end reads represent the two short reads of the two ends of a genomic fragment with a known estimated size (herein referred to as the library size). The paired-end reads for the 6 individuals from the Utah and Nigerian families were used as the test genome data for predicting the TIPs_IN candidates using paired-end mapping (PEM). These paired-end reads were generated using the Illumina Solexa platform, with read lengths averaging 250bp at a standard deviation of 100bp. Below is a brief description of the computational algorithms and procedures used.

Alignment data of the pair reads for each genome was downloaded from the NCBI short read trace data web site at ftp://ftp.ncbi.nih.gov/1000genomes/ftp/data/. Paired-end reads showing a mapping distance (span size in the human reference genome) considerably larger than the sequencing library size were selected for genotyping. Their location in the human reference genome was then compared to the locations of HERV sites that have already been annotated in that genome. The presence of multiple pair reads with their two read mates located outside of a HERV insertion indicated a possible absence of the HERV insertion in the examined genome (Figure 2). The predicted candidates were further processed for selection by requiring the absence of a recently duplicated region without the insertion in the human genome and the absence of the HERV insertion in the closely related chimpanzee genome. This was done based on the assumption that a polymorphic HERV insertion in human genome must have originated from a human-specific insertional event and as such should be absent in the genome of an outgroup species. This final candidate list was subjected to experimental verification by PCR genotyping.

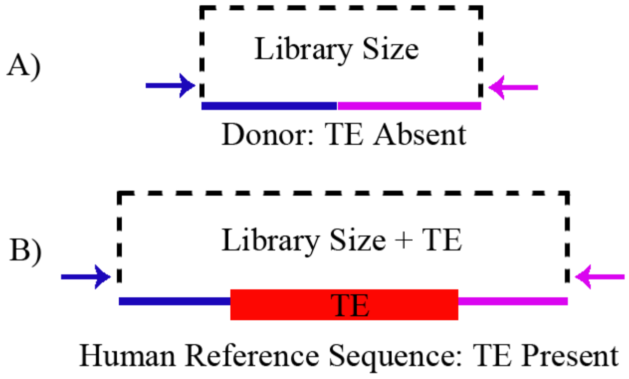

Figure 2: Schematic representation of TIPs_IN candidate generation.

A) Represents paired-end reads originated from the donor genome at the same location, with a size smaller than what to be expected within the human reference genome. B) Represents the expected distance of the same read pairs within the human reference.

Processing & Filtering TIPs_IN Candidate List Data

The candidate HERV lists produced by computational analysis was narrowed down by a number of specific criteria. The HERV-K candidates classified as members of the LTR5_Hs solo-LTR subfamily were the focus of this study due to the increased probability of activity (Taruscio & Mantovani, 1998). If the size of a candidate was less than 500 base pairs, it was assumed to be a fragment and not considered. Using the UCSC Genome Browser (http://genome.ucsc.edu/), the human genome was compared with those of two outgroups, in this case *Pan troglodytes* (chimpanzee) and *Macaca mulatta* (rhesus monkey), to ensure that the HERV insertion was not present in either of the genomes. If either outgroup species carried the insertion, then it was not considered a HERV, as it was inserted into an ancestral genome before the divergence of humans and chimpanzee. It was also essential to make certain that the regions flanking the candidate were unique, that is, they did not contain repeating elements (such as LINEs, SINEs, SNPs, etc.) so that primers could be designed specifically for that region of the genome. Additional criteria used include a minimal 5 pairs of reads supporting the same polymorphic pattern and the absence of the candidate insertion in one of the six genomes for TIPs_IN. All of the selected candidates are found in

Table 2.

Generation, processing, and filtering of Full-length TIPs_IN Candidate List

The full-length TIPs_IN candidate list represents HERV loci that are found within the human reference genome as a full-length insertion with a sequence similar to HERV-K113. The initial candidate list was established by comparing the sequence identity of the HERV-K113 *gag* and *env* sequences published by (Belshaw 2005a, 12507) using the BLAT tool found at (http://dbrip.brocku.ca/cgi-bin/hgBlat?command=start). The initial candidate list was composed of entries that were obtained using BLAT against the GRCh37/hg19 assembly of the human genome sequence, and found to have a sequence identity of 95% or greater.

The initial full-length TIPs_IN candidate list was compared to the known polymorphic HERVs that had been previously published to keep those not studied before. The list was further narrowed using the UCSC Genome Browser (http://genome.ucsc.edu/) to ensure their absence in the outgroup genomes and for primer availability as described in the previous section. All of the selected candidates are found in Table 3.

Generation and Filtering of TIPs_Out Candidate List

The TIPs_Out list of candidate polymorphic HERV loci represent insertions that were absent in the human reference genome but present in one or more of the donor human genomes. This candidate list was generated by Xuemei Luo in Dr. Liang's lab via computational comparative genomics analysis using the same donor genome data set as the TIPs_IN predictions. Below is a brief description of the computational algorithms and procedures used.

The paired-end mapping sequences were aligned using the MAQ software. This software labels concordant (mapped to the same chromosome) reads with a flag of 18 (MF18) and those mapped to two different chromosomes with a flag of 32 (MF32 reads). Based on the alignments of paired-end reads to the human reference genome, the MF32 reads provided the signatures for possible TIPs_Out candidates, by indicating an insertion is found within that region (Figure 3). These reads were then further required to have one of the reads mapped to non-repetitive region position and the other mapped to a HERV sequence. The location for the transposons in the human reference genome was based on the RepeatMasker annotation obtained from the UCSC Genome Browser (http://genome.ucsc.edu). Candidate reads were then clustered based on their positions on the reference genome where each cluster represented a candidate HERV insertion locus. The insertion genotype was then determined by comparing the ratio of concordant MF18 reads to the MF32 reads. A genotype of "+/+" required most paired-end reads in the region flanking the insertion to display an MF32 flag and very few or no concordant pairs (MF18); while the genotype for "+/-" show roughly half MF18 and MF32 reads. Due to the alignment of paired-end reads, the size of the novel insertion was roughly predicted by adding the flanking region to the size of the LTR sequence the read was mapped to, but is not a reliable source of information.

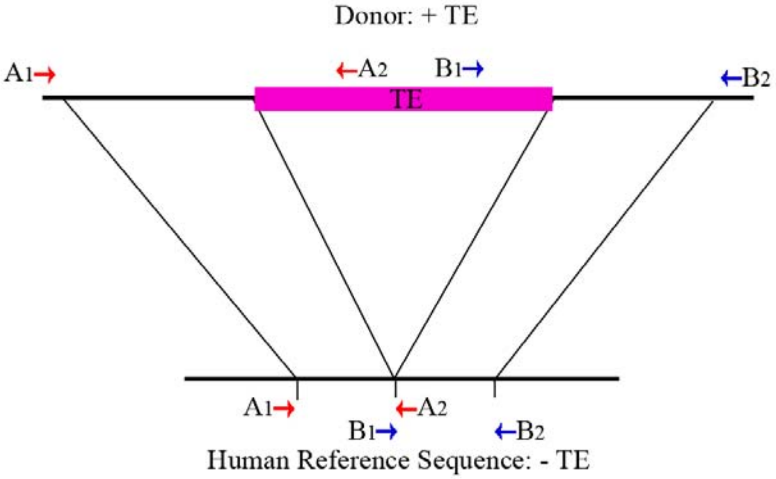

Figure 3: Schematic representation of TIPs_OUT prediction.

A reads has one read (A1) in the 5' flanking region of the predicted TIP_OUT insertion and the other read (A2) mapping into a TE of the same subfamily elsewhere in the genome. B reads (B2) has one read in the 3' flanking region of the predicted insertion and other read (B1), mapped into the same TE elsewhere in the genome as A2.

The initial TIPs_OUT candidate list was narrowed by a number of very strict criteria while examining the predicted genotype data to minimize the possible false positives. First, all of the entries that gave an uncertain genotype prediction for all 6 individual test samples were excluded, as the data associated with these entries was very poor. The entries which gave an MF32/MF18 value of 0/0 for each of the 6 individual test samples were excluded, as these were likely to represent false positive insertions. If the predicted genotype for all of the 6 individual test samples was predicted to be +/+ for an insertion, these were excluded because these would represent an insertion found in every individual, and therefore less likely to be polymorphic. If the size of the candidate was less than 500 base pairs, it was assumed to be a fragment, and was excluded. The candidate list was further narrowed to those with between 10-50 reads, as this represented the most likely number based on the sequencing coverage depth. In order to narrow the remaining list, the list was filtered for candidates that were of high interest. This included TIPs that were specific to either the Nigerian or Utah populations, as these would represent a polymorphic status, followed by checking primer availability in the flanking regions. All of the selected candidates are listed in Table 4.

Genome Position

All of the TIPs_IN and TIPs_OUT candidate lists were generated from the March 2006 NCBI36/hg18 assembly of the genome, whereas the full-length TIPS_IN were generated from the February 2009 GRCh37/hg19 assembly of the genome. In order to make comparisons and avoid overlap between these lists, the genome position annotations were converted from one version to another using the liftOver tool found on the UCSC genome browser website (http://genome.ucsc.edu/cgi-bin/hgLiftOver).

DNA Test Samples

The computational analysis of TIPs was done using the data from the six test DNA samples used in the first pilot of the 1000 Genomes Project, and the DNA from these individuals were purchased from the Coriell Biorepository. The Coriell sample IDs for the trio family from Yoruba in Ibaden, Nigeria were NA19238 (mother), NA19239 (father) and NA19240 (daughter) and the Utah trio family were NA12892 (mother), NA12891 (father) and NA12878 (daughter) (Coriell 2011). These samples were used for genotyping using PCR to validate the computational prediction.

To perform an extended allele frequency of those polymorphic insertions verified using the 6 trio samples, a 24-sample panel from Coriell's Polymorphism Discovery Resource [M24PDR] was also used. The following samples come with no specific ethnicity data associated with each DNA sample, but are intended to cover a wide variety of population groups: NA15029, NA15036, NA15215, NA15223, NA15245, NA15224, NA15236, NA15510, NA15213, NA15221, NA15227, NA15385, NA15590, NA15038, NA15056, NA15072, NA15144, NA15216, NA15226, NA15242, NA15268, NA15324, NA15386 and NA15594 (Coriell 2011). In addition to these samples, for some loci with low allele frequency, two additional Coriell sample sets were used: the HD11 panel of Africans North of the Sahara samples: 17380, 17379, 17382, 17378, 17384, 17381, 17383 and the HD12 panel of Africans South of the Sahara samples: 17348, 17341, 17344, 17342, 17347, 17346, 17343, 17345, 17349 (Coriell 2011). Since there is no available sequencing data associated with these samples, they were genotyped based solely on PCR results. Each of the stock DNA samples obtained from

Coriell were stored at -80°C, and 50ng/µl working solutions were created for each DNA sample by suspending in TE buffer solution (Tris 1M, EDTA 0.5M at pH 8.0) and stored at -20°C.

Primer Design

Primers were designed for the final polymorphic HERV insertion candidates using UCSC Genome Browser (http://genome.ucsc.edu/) and Primer3 online software (http://frodo.wi.mit.edu/primer3/). The UCSC Genome Browser was used to obtain the DNA sequence from the human reference genome in the locations of the predicted insertions. For the TIPs_IN candidates, the primers were designed in the regions flanking the insertion that was present in the human reference genome, while trying to avoid being placed within another repetitive element (Figure 4A). The TIPs_OUT candidates do not have an insertion within the human reference genome; therefore these primers were designed in the regions that flanked the predicted chromosomal position acquired from the computational analysis (Figure 4B). The DNA sequences obtained for each candidate location and surrounding flanking regions was used in Primer3 software to create primers using the following general primer picking conditions: Primer size minimum 18, optimal 20, maximum 27 bases, Primer melting temperature minimum 57.0°C, optimal 60.0°C, maximum 63.0°C, and GC content between 40 and 70 percent. The designed primers were tested using the UCSC Genome Browser's In-Silico PCR program to ensure that one and only product at the expected locus and size is predicted in the human reference genome. The information provided by the In-Silico PCR was used to determine the size of the alleles with and without the insertion.

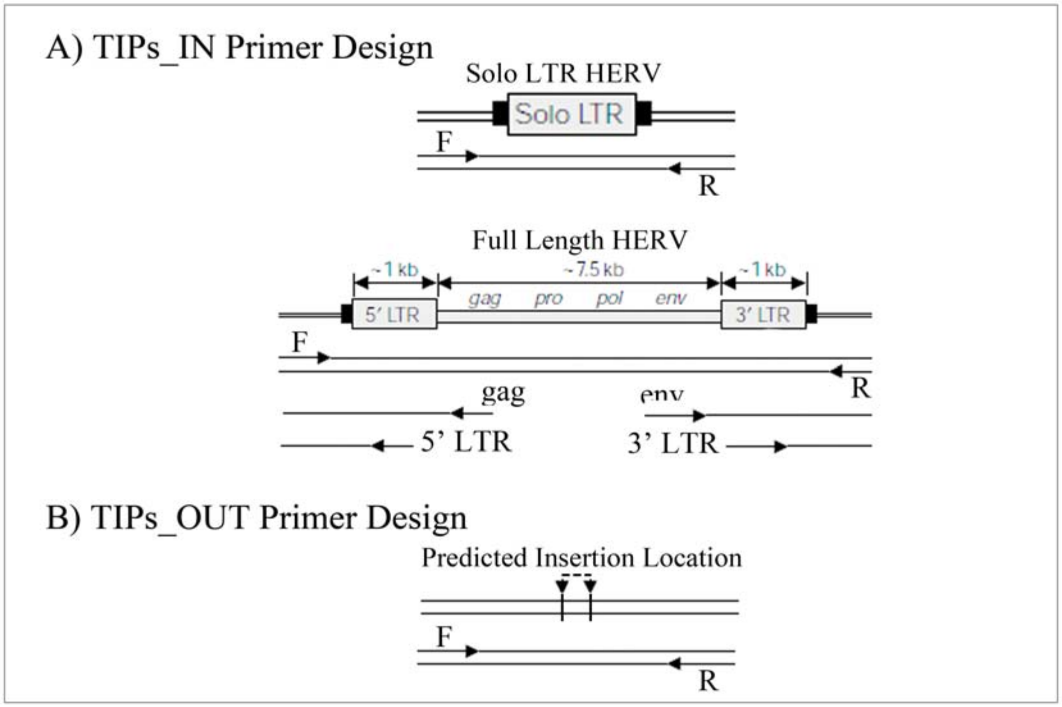

Figure 4: TIPs_IN and TIPs_OUT primer design strategy.

A) TIPs_IN forward and reverse primers are designed to flank the solo-LTR or the full-length HERV depending on the status within the human reference genome sequence. Internal primers designed to be located in either the *gag*, *env*, 5' LTR or 3' LTR were used in conjunction with either the forward or reverse primers. B) TIPs_OUT forward and reverse primers are designed to flank the computationally predicted insertion location within the human reference genome sequence.

It is difficult to genotype a full-length HERV (~9.5kb) using regular PCR due to the large product sizes. To overcome this issue, "universal" primers were designed using the full-length HERV-K113 sequence that was published by (Belshaw 2005). To aid in identifying full-length HERV insertions, primers were manually designed for amplifying the internal regions, specifically on the – strand within the *gag* region, and on the + strand within the *env* region of the HERV (see Table 5). These primers are used to help determine the orientation of TIPs_OUT candidates, as well as confirming the presence of an LTR by using each of the "universal" primers is in conjunction with the forward and/or reverse primer designed for each candidate locus. Some of the forward and reverse primers designed for the full-length HERVs found within the literature were also ordered, along with previously designed internal and LTR primers. All of the primers were ordered from AlphaDNA (www.alphadna.com). Each of the stock primers were stored at -20°C, and 10μM working solutions were suspended in TE buffer solution (Tris 1M, EDTA 0.5M at pH 8.0) and stored at -20°C.

Table 2: TIPs_IN candidate locations for which primers were designed

Location (chr: s-e)	TE Size (bp)	Size +TE (bp)	Size -TE (bp)	Primer Sequences (5´-3´)	T_M(F/R) (°C)
chr3:14107685-14108653	969	1636	667	F: aaagggcatggagaaatgtg R: cccacctaggctctgacttc	59.9/58.9
chr4:120483136-120484102	967	1608	641	F: gaggtgtgcaagggacattt R: catccttcaaggccagaaaa	60.0/60.2
chr7:157722243-157723211	969	1415	446	F: tgctcattcagaagccacac R: aacgagaagccagcatcagt	60.0/60.0
chr8:18695738-18696706	969	1495	526	F: ctgcaggacgatgagaggat R: tatcatgccctgtggtctga	60.4/60.1
chr8:37170043-37171011	969	1487	518	F: ctgggagagatggcagagag R: gcagtgagatgtggctttga	60.1/60.0
chr11:71155928-71156598	671	846	175	F: ctggttcttcagagccacct R: cgactttgccttgaactgtg	59.4/59.5
chr12:54013481-54014450	970	1449	479	F: ttcagtccctagaggtactatgctc R: ggtttccagatcttaccagca	59.4/59.2

Table 3: Full-length TIPs_IN candidate locations for which primers were designed

Location (chr: s-e) Hg19	TE Size (bp)	Size +TE (bp)	Size -TE (bp)	Primer Sequences (5´-3´)	T_M(F/R) (°C)
chr3:185281305-185288547	9181	9273	92	F:CATCCCTTCCATGCCTTAG R:GGGATTATGAGACAGGTACATG	58.1/56.0
chr3:101411706-101418889	9124	10233	1109	F: TCTCTGCAGGCTTGCAATC R: CCCACCCCAGATCCAAGTAC	60.2/61.5
chr3:125610106-125617634	9126	9702	576	F: CTTACCAATGTGCCCACGTAC R: AGAGGCAGAATGATATGGTGGT	60.3/59.9
chr8:140473118-140475236	3090	3517	427	F: TCCCACTGCCAAGAAGACC R: TCCCCCATCTTGCCTAGC	61.2/61.1
chr10:6867110-6874635	9463	9737	274	F:GAGTTGGAGTGAGGAAATCAGTTC R: GCATTACCTGCAGATACTCGTG	60.5/59.8
chr10:101581556-101587716	7054	7231	177	F: CAGGTAGTAGCGTGGAGAAAAC R: CTTCACCCTCCATTCCAGG	58.1/60.4
chr11:101566762-101574290	9466	9620	154	F: AAACACTTCCATGCTCAGAAAG R: CCATCCCTGGCAAAATGAC	58.5/61.3
chr12:58722211-58729730	9457	11432	1975	F: TGTTGGGGCTGAGGACAG R: CTACAGCTGCCCCATGATTAC	60.8/59.6
chr21:19933917-19940998	8305	8778	473	F: CTGAACATGAATTCTTTGCAAG R: CTTGCAAAGAATTCATGTTCAG	57.6/57.6

Table 4: TIPs_OUT candidate locations for which primers were designed

Location (chr: s-e) Hg18	Size -TE (bp)	Primer Sequences (5´-3´)	T_M(F/R) (°C)
chr4:9590458-9590900	500	F:ACCCTCCAGCTCCAGTGC R:GGGCATCTTTTCAAGACGTT	61.4/59.2
chr6:161190594-161191142	933	F:ACCATGAAGCCAGAGAGAAAAT R:GTCGCCTTTCCTTGGTCTC	59.2/59.8
chr13:89540897-89541416	650	F:TTGATAAAATTTGGACAAGAAGTCTC R:TGCATCATAATTGAATGCAAAA	59.2/59.1
chr14:89981676-89981754	837	F:CCCTATGGATACGACCATCAC R:GCACCTGCTCTTTCTCTTCC	59.1/59.2
chr15:26103457-26104092	969	F:GATCTTACCAGAACAAAACCCAT R:CGCTTTGGATTGCTAGTGTG	58.4/59.5
chr20:12350168-12350569	518	F:GCCCTGTTGTAATAGGCATGA R:GCAAAGGAATTTGAGCCAAG	60.0/59.8
chr21:44526904-44527067	215	F:CACGGCATGGTGGAAAGT R:GATGCCTTAGCCCAGAGATG	60.5/59.8
chr5:74942501-74942793	589	F:GCTTTGAATACCCTCCCCAAT R:TGCACAGTGCTAAGGTTTGG	61.4/59.9
chr6:32565513-32565534	369	F:AACAGAGAATGCCGTCAAATG R:TTGGGTTCACTTTATCCACCA	60.1/60.2
chr8:26596800-26597248	1015	F:CATGGTGGGAATTTATCAACG R:CCATCTGGGAAGTGTGGAG	60.1/59.0
chr9:33120476-33120527	170	F:GCTGGTGAGCTAAGGTCAGG R:CCACTCCCATTTGGCTTATG	60.0/60.3
chr9:71604561-71604867	1520	F:CCCAAGGCAGAAAGTCTTAAG R:GGCTCTGGCTCCAATTACAC	58.1/59.7
chr11:124879106-124879109	257	F:AAGGAAAACTGAGGACTGGTG R:CCCAAAAAGCAGCAGTTTGTA	58.3/60.3
chr12:11502680-11503061	759	F:AAGGGTGGGGGAATACGTC R:CCTCACACATTTGCTTCTGC	61.0/59.4
chr14:50903129-50903366	849	F:TCCTACTCTTGGGAGGCTCA R:AGCAAGGCACCAGGACTTAG	59.9/59.5
chr21:15966603-15966908	997	F:CAGCTGCTGGGTGCTGAG R:TCAGCAGAACAATGAGTACAAGG	62.0/59.4
chr1:110110564-110110730	916	F:GGGCATGTCCTTGAAATTGT R:CTCTTTCTTTTCCCCACAGG	59.8/58.8
chr16:79673433-79673677	638	F:TGGAGCTTTGCATTGTTCTG R:AATAACGCAAGCCAGCAGAG	60.0/60.5
chr16:79687236-79687414	775	F:CCCTAGGGCAAAGGCTACTC R:CATGTGGAAAGGAACCCAGT	60.2/59.8
chr1:16943334-16943482	342	F:GCTGGGATTATAGGCACACG R:AAATTGTTCAAAAGCATCAAAGA	60.5/58
chr9:32849689-	456	F:TGTGTTTGTTTTGCGCATTT	

32849840		R:TGAAAGGTGCATGCTCAGTC	
chr19:22205964-22206428	649	F:CGACACAAAGGAAGACACAGAG R:GACGGTTTTTGACTTAAGATAGAGC	59.9/59
chr1:225025914-225026184	600	F:TCAAACTCTAGCTCACATGTCCT R:GAAAACAATGGAGGGTGAGG	58.1/59.4
chr1:223336301-223336671	450	F: TTTCCCTTGATGTTCTTCCA R: CATTACCCTTCCATGAGAATCA	58.1/58.9
chr2:24987735-24988289	850	F: ACAGGCTCCGAGGGAAGATA R: TGTTACAGTTTAGTGCCTTCTGG	61.1/58.5
chr11:99317377-99317663	629	F: TACATGCATTCCCAGGGTTT R: TGACATGATTTTGCCTGACTCT	60.2/59.6
chr15:24169171-24169651	574	F: AGGCTGCTCAAGGCTACAGA R: GATTCAGGCTGTTTCGTGTG	60.3/59.3
chr17:10388873-10389199	487	F:TGCCACAAGTAGTTTAGATTGGTC R:TGAAGGAGAAGGTCCAGGAA	59.6/59.8
chr17:4959277-4959821	670	F: GCCAGTGAGCCTCTGACTTT R: CTCGAGGACCGCCTCAGT	59.6/61.6
chr17:24467429-24467979	651	F: GCCTCCACATTCCCTGAGTA R: ACTTCACTCTGAGGCGGTGA	60.1/61.0
chr17:32220207-32220618	1477	F: GACTGACTGTGCCCTTGGAT R: TGGAAAATTCAAGCAATATGGA	60.1/59.4

Table 5: Universal internal HERV primers designed

Primer Name	HERV Location	Primer	Distance from LTR (bp)	Primer Sequence (5’-3’)	T_M (F/R) (°C)
HERV_5LTR_1	5’LTR	-	5' LTR 922/3' LTR 66	GTGGGACGAGAGATTTGGAA	60
HERV_5LTR_2	5’LTR	-	5' LTR 776/ 3' LTR 212	TTCTCAAAGAGGGGGATGTG	60
HERV_5LTR_3	5’LTR	-	5' LTR 845/ 3' LTR 143	GCGTTCAGCATATGGAGGAT	60.1
HERV_3LTR_1	3’ LTR	+	5' LTR 776/3' LTR 212	CACATCCCCCTCTTTGAGAA	60
HERV_3LTR_2	3’ LTR	+	5' LTR 770/3' LTR 238	TCCCCACAATTGTCTTGTGA	59.9
HERV_3LTR_3	3’ LTR	+	5' LTR 550/ 3' LTR 431	CCCGATTGTATGCTCCATCT	59.9
HERV_Gag_1	*gag*	-	from 5'	TTTGCCAGAATCTCCCAATC	60

			2175		
HERV_Gag_2	*gag*	-	from 5' 2248	TCGGACCTGTTCTTGTACCC	60
HERV_Gag_3	*gag*	-	from 5' 2572	CTCAGGATTGGCGTTTTCAT	60.1
HERV_Env_1	*env*	+	from 3' 1208	AAATTTGGTGCCAGGAACTG	60
HERV_Env_2	*env*	+	from 3' 1206	ATTTGGTGCCAGGAACTGAG	60.1
HERV_Env_3	*env*	+	from 3' 1575	TGCTGTAGCAGGAGTTGCAT	59.6
Barbulescu et al. 1999 LTR Primers					
LTR+_M1	5' LTR	+	5' LTR 20 /3' LTR 968	TGTGGGGAAAAGCAAGAGAG	60.4
LTR+_M2	5'LTR	+	5' LTR 446 /3' LTR 542	CTGTGCTGAGGAGGATTAGT	54.5
LTR+_M3	3'LTR	+	5' LTR 849 /3' LTR 139	TCCATATGCTGAACGCTGGT	61.6
LTR+_M7F	5'LTR	+	5' LTR 342 /3' LTR 646	AAGCCAGGTATTGTCCAAGG	59.1
LTR+_M8F	3'LTR	+	5' LTR 821 /3' LTR 167	TAAGGGAACTCAGAGGCTGG	59.4
LTR-_M4	3'LTR	-	5' LTR 940	GTGGGTGTTTCTCGTAAGGT	56.6
LTR-_M5		-	not found in K113 sequence	GGACAGGCAGGAGACAGATG	60.8
LTR-_M6		-	not found in K113 sequence	CTGAGTTGACACAGCACACG	59
LTR-_M7R	5' LTR	-	5' LTR 342 bp	CCTTGGACAATACCTGGCTT	59.1
LTR-_M8R	3'LTR	-	5' LTR 821 bp	CCAGCCTCTGAGTTCCCTTA	59.4
Full-length HERV from Barbulescu et al. 1999					
HERV-K103_MB-2_F			1508 total	GATTTCGAGCCACCTCTGAAG	61.3
HERV-K103_MB-3_R			1508 total	CTCAGAAACAGGCTTAAGACG	56.9
HERV-K109_MB-23_F			9794 total	GTCCTTTAATGTCTCCCCTC	55.2
HERV-K109_MB-37_R			9794 total	CAGATGAGATGTCAAGCAAGGT	59.4

Full-length ERV from Belshaw et al. 2005					
s859c12_F				TAGGCTTGAGGTATAAGTCAC	51.4
s859c12_R				TTGGTTTCCAGATCTTACCAGC	60.5

PCR Genotyping

Genotyping and validation of the selected candidates was performed using PCR. Each 25µl PCR reaction was setup using Invitrogen's AccuPrime™ Taq DNA Polymerase System kit reagents. Each reaction was carried out in a sterile thin walled 0.25 ml PCR tube and composed of 2.5 µl 10x AccuPrime™ PCR Buffer II, 2.5 µl of each primer (10µM), 50 ng of the DNA sample of choice, 0.5 µl AccuPrime™ *Taq* DNA Polymerase, and 16.0 µl of autoclaved distilled water. PCR reactions were run on Eppendorf Mastercycler or Mastercycler gradient (Eppendorf; Mississauga, Ontario) with the following protocol:

Lid Temperature 95°C
94°C 2 minutes
*94°C 30 seconds *35 cycles
*54°C-60°C 30 seconds (depending on T_M of primers)
*68°C 1 minute 30 seconds – 4 minutes (depending on expected product size)
68°C 10 minutes
Hold 4°C upon completion

PCR products were loaded on a 1%-2% agarose gel (depending on expected product size) stained with ethidium bromide or RedSafe™ Nucleic Acid Staining Solution (iNtron Biotechnology, Toronto, Ontario). Each agarose gel was then subjected to electrophoresis (45 minutes-90 minutes at a voltage of 75V-110V, depending on the size of the gel) in 1x TAE buffer and visualized by exposing to UV light using a BioRad Gel Doc 1000 with a camera. Each PCR amplification product was compared against the Invitrogen 100 bp DNA Ladder (Cat

15628-050), Invitrogen 1 Kb Plus DNA Ladder (Cat #10787-018) or Norgen Biotek LowRanger 100 bp DNA Ladder (Cat# 11500) depending on the expected product size.

Possible Genotypes for Each Individual DNA Sample

Traditionally throughout the literature there are three possible genotypes used to identify a polymorphic HERV insertion (Moyes 2007). Throughout this study more allele combinations were observed throughout the screening of all 30 individual DNA samples. This has led to the proposal of using six possible allele combinations to describe the genotype of HERV insertions (Table 6). The first combination is the Pre-integration site on both alleles with no HERV insertion present on wither allele (abbreviated as PI/PI). The second possible combination is a solo-LTR being present on one allele, and the pre-integration site with no HERV insertion on the other allele (abbreviated PI/sLTR). The third combination is the presence of a full-length HERV insertion on one allele and the pre-integration site with no HERV insertion on the other allele (abbreviated (PI/FL). The fourth possible combination is the presence of a solo-LTR on both alleles (abbreviated sLTR/sLTR). The fifth possible combination is the presence of a solo-LTR on one allele and a full-length HERV insertion on the other allele (abbreviated sLTR/FL). The sixth and final possible combination is the presence of a full-length HERV insertion on both alleles (abbreviated FL/FL).

Table 6: Possible allele combinations and their allele genotype abbreviation

Possible Allele Combinations	**Allele Genotype Abbreviation**
Both Pre-integration Alleles	PI/PI
Pre-Integration and Solo-LTR Allele	PI/sLTR
Pre-integration and Full-length Allele	PI/FL
Both Alleles Solo-LTR	sLTR/sLTR
Solo-LTR and Full-length Allele	sLTR/FL
Both Full-length Alleles	FL/FL

Given that there are six genotype combinations that are possible for each locus, these are likely to result in a variation of genotypes throughout multiple individuals as a result of Mendelian inheritance. This has led to three possible categories for which to classify the overall genotype status for each HERV locus that has been tested (Table 7). The first category is dimorphic, which results from at least two individuals with a different genotype at a given locus. This is further broken down into three sub-categories. The first sub-category is a dimorphic solo-LTR, which is used to describe the presence of only the pre-integration allele and solo-LTR genotype combinations occurring within all of the tested individuals at that locus. This can occur with one individual having both pre-integration alleles and another individual having the pre-integration allele and the solo-LTR allele or both solo-LTR alleles. The second sub-category is dimorphic Solo-LTR and Full-length, which is used to describe the presence of only the solo-LTR allele and the full-length allele genotype combinations occurring within all individuals at that locus. This can occur when one individual has both solo-LTR alleles and another individual has both full-length alleles. This can also occur when one individual has one solo-LTR allele and one full-length allele while another individual has either both full-length alleles or both solo-LTR alleles. The third sub-category is dimorphic full-length, which occurs when only the pre-integration allele and the full-length allele are present in all of the individuals at that locus. This can occur when one individual has both pre-integration alleles and another individual has both full-length alleles. This can also occur when one individual has one pre-integration allele and one full-length allele while another individual has both pre-integration alleles or both full-length alleles.

The second category is trimorphic, which is used to describe the presence of the pre-integration allele, solo-LTR allele and full-length HERV insertion allele within at least two individuals at that locus. This can occur in four different combinations. The first combination occurs when one individual having both pre-integration alleles while another individual has one solo-LTR allele and one full-length allele. The second combination occurs when one individual has one pre-integration allele and one solo-LTR allele while the other individual has one pre-integration allele and one full-length allele. The third and fourth combinations occur when one individual has one pre-integration allele and one solo-LTR allele while the other individual has both full-length alleles or both solo-LTR alleles.

The third category is fixed, which is used to describe either a solo-LTR allele or full-length HERV insertion allele that is homozygous in all individuals within a locus. They are categorized as fixed because with a homozygous genotype in all individuals this insertion pattern will not change as a result of Mendelian inheritance, but can only change through unrelated insertions/deletions within that locus.

Table 7: HERV locus classification categories, sub-categories and their corresponding genotypes

HERV Locus Classification Category	Locus Classification Sub-Category	Genotypes Required in at Least One Individual at that Locus
Dimorphic	Dimorphic Solo-LTR	PI/PI and PI/SLTR
		PI/PI and sLTR/sLTR
	Dimorphic Solo-LTR and Full-length	sLTR/sLTR and FL/FL
		sLTR/sLTR and sLTR/FL
		sLTR/FL and FL/FL
	Dimorphic Full-length	PI/PI and PI/FL
		PI/PI and FL/FL
Trimorphic		PI/PI and sLTR/FL
		PI/sLTR and PI/FL
		PI/sLTR and FL/FL
		sLTR/sLTR and PI/FL
Fixed	Fixed Solo-LTR	All show sLTR/sLTR
	Fixed Full-length	All show FL/FL

TIPs_IN PCR Genotyping Strategy

The TIPs_IN candidate list is composed of computationally predicted solo-LTRs that are present within the human reference genome sequence, but predicted to be absent from one or more of the six test individuals. Therefore for each candidate to be considered polymorphic, the solo-LTR must be absent in one or more individuals. To test for these polymorphisms, first the F+R primer was run using the DNA sample with the best prediction data showing the absence of the solo-LTR insertion. If this sample showed the absence of the solo-LTR insertion, the F+R was run on the panel of 6 test sample individuals, followed by the panel of 24 anonymous individuals. Although this data shows the presence and absence of a solo-LTR, it does not give any information with regards to the presence of a possible full-length HERV insertion. To test for a full-length insertion, each candidate was first checked in the UCSC Genome Browser to determine the orientation of the solo-LTR insertion. With this information the "universal" internal primers were tested along with the appropriate flanking primers for each candidate insertion's orientation. By overlaying all of these images, it is possible to compare each individual and identify the pre-integration band, solo-LTR band, and full-length bands to determine a complete genotype. In the cases where the internal primers showed non-specific amplification with the flanking primers, the "universal" 5' or 3'LTR primers were run with the corresponding flanking primer, to see if any samples that showed the absence of a solo-LTR band during initial screening, resulted in amplification of the flanking and LTR primer combination. These cases are listed as potentially full-length insertions, as they have not been fully validated with the internal primers.

Given that the full-length TIPs_IN candidates are present within the human reference genome sequence, in order to be polymorphic, these candidates must be found as either a solo-

LTR or absent from an individual. Because they were found within the reference genome this allowed access to a variety of information using the UCSC genome browser, such as: HERV orientation, size of insertion, In-Silico testing of both the flanking primers, as well as the flanking primers combined with the “universal” internal primers. This information was obtained for all of the full-length TIPs_IN candidates prior to PCR genotyping. Unlike the other TIPs_IN candidates, these insertions were not computationally predicted to be polymorphic, so there were no DNA samples for which the insertion was predicted to be absent. Therefore to test these candidates, a 50 ng/µl mixture of the 24 anonymous DNA samples was created for preliminary testing of these candidates. To test for the absence of the full-length insertion, the F+R primers were run to check for amplification of either a pre-integration band or a solo-LTR. If either of these two bands were present, these candidates were run on both the panel of 24 anonymous individuals, followed by the panel of 6 individuals. Although screening the candidates with the mixed DNA proved effective in some candidates, in cases where the polymorphism is present in very few individuals, amplification of these bands were either very faint or not visible. The flanking primers were also run with the internal primers selected from the In-Silico testing to determine if the full-length insertion was absent from any of the test individuals. By overlaying these gel images, it is possible to identify a complete genotype for each individual.

TIPs_OUT PCR Genotyping Strategy

Unlike the TIPs_IN candidates, the TIPs_OUT candidates are absent from the human reference genome; and to be polymorphic must appear as either a solo-LTR or a full-length insertion in one or more of the test individuals. The only information available before screening the TIPs_OUT candidates was the computationally predicted location of each insertion within the human reference genome sequence. Therefore the first step in testing each of the candidates was to run a PCR (using the DNA sample with the best prediction data), which consisted of 3 separate reactions, F+R primers, F+5' LTR primers, and F+3'LTR primers. By running this combination of primers, it was possible to identify whether or not a solo-LTR existed in this sample (from the F+R), and determine the orientation of the HERV insertion (by comparing the F+5'LTR and F+3'LTR). If the F+R showed a solo-LTR, this was run on the panel of 6 individuals, followed by the panel of 24 anonymous individuals. Using the candidate's insertion orientation, the "universal" internal primers were tested along with the appropriate flanking primers on the DNA sample with the best prediction data. Although this worked in some cases, the universal primers will not work if only the solo-LTR is present in the sample. Therefore the "universal" internal primers were also tested with the mixed DNA to identify a working combination. If a combination of internal primers worked, they were run on the 6 panel of test sample individuals, followed by the 24 panel of anonymous individuals. In the cases where a suitable internal primer could not be found, the F+5' or 3' LTR primer combination used in the initial screening was run on the panel of 6 test sample individuals, followed by the panel of 24 anonymous individuals. By overlaying these gel images, it was possible to identify any individuals that did not have a solo-LTR band, but the LTR primer worked, thus indicating a possible full-length insertion. These specific individuals were then used to test the flanking primers combined with the "universal" internal primers to find a suitable combination for testing on the panel of 6 test sample individuals, and the panel of 24 anonymous individuals. By overlaying these gel images, it is possible to identify a complete genotype for each individual.

PCR and Sample Preparation for DNA Sequencing

The samples selected for sequencing were those that represented a novel pre-integration site or insertion sequence as a way of further validation and characterization. Samples from candidates of interest (trimorphic status etc.) were also sequenced to allow for comparisons of the LTR sequences between multiple individuals. Each sample sent for sequencing was first amplified in a 50μl PCR reaction as described in the PCR genotyping section above. All of the selected PCR products were purified using the Norgen Biotek PCR purification kit or Gel purification kit following the manufacturer's instructions (Norgen Biotek Corp; St. Catharines, Ontario), followed by DNA concentration determination, and if necessary dilution to obtain a concentration between 30 and 100 ng/μl before sending for sequencing at The Centre for Applied Genomics (TCAG) in the Hospital for Sick Children in Toronto, Ontario. All sequencing was done on the Applied Biosystems 3760XL DNA Analyzer.

PCR Genotype Data Analysis

PCR results for all of the candidates were examined for evidence of insertional polymorphisms as demonstrated by the variation in the size of the product in at least one test sample. Each of these candidates were used to calculate and compare the expected and observed allele frequencies as follows:

$$\text{Observed Allele Frequency} = \frac{\text{\# of the allele}}{\text{Total \# of alleles in population}}$$

Using the Hardy-Weinberg equilibrium ($p^2 + 2pq + q^2 = 1$), the genotype frequencies can be calculated, where:
p = allele frequency of PI present = # of alleles/total # of alleles in population
q = allele frequency of sLTR present = # of alleles/total # of alleles in population

In the case of 3 allele combinations being present, this must be expanded to:
$p^2 + q^2 + r^2 + 2pq + 2pr + 2qr = 1$
r = allele frequency of FL present = # of alleles/total # of alleles in population

The expected frequencies are calculated as:
Expected p = allele frequency of PI present x total individuals genotyped
Expected q = allele frequency of sLTR present x total individuals genotyped
Expected p = allele frequency of FL present x total individuals genotyped

Estimation of proviral ages by LTR sequence comparisons

Given the abundance of ERVs in primate genomes, they are ideal candidates for exploitation as phylogenetic markers (Johnson, Coffin 1999). The relative age of ERVs can be estimated by comparing the sequences of the two LTRs. Due to the mechanisms behind reverse transcription, the two LTRs are identical at the time a provirus forms. Any mutations that accumulate over evolutionary time will be unique to one of the two LTRs, thus allowing the estimation of the proviral age (Turner 2001). The current accepted rate of mutation was developed by dividing the number of substitutions per site between the human and chimpanzee ERV sequences, by the age of the most common ancestor of humans and chimpanzees (~4.5 million years ago). This gives an estimated mutation rate of $2.3x10^{-9}$ to $5.0x10^{-9}$ substitution per site per year (Johnson, Coffin 1999). Given that the average HERV-K LTR is 970 bp long, this gives an estimate of approximately one difference per LTR every 200,000-450,000 years (Turner 2001). Due to the imprecise estimates of divergence dates found throughout the literature, these calculations can only provide a rough estimate of absolute time, but they are still very useful for comparing the relative ages and rates of evolution of different HERV loci (Johnson, Coffin 1999). In the scenario where only a solo-LTR allele is present, this type of analysis cannot be completed as the second LTR is not available. Instead, it is possible to compare the solo-LTRs of multiple individuals at the same locus to estimate the time that those individuals diverged from one another. This can also give a very rough indication of how long ago the solo-LTR was formed.

Results

Verifying TIPS_IN solo-LTR Candidates using Trio Samples and Anonymous 24 Individual Samples

From the list of computationally generated TIPs_IN candidates for HERV insertions, 7 loci were chosen to be verified and genotyped by PCR based on the selection criteria outlined in the Materials and Methods section (

Table 2). Each candidate was tested on the two trio family samples from which the insertion prediction was made. Each of these candidates exists as a solo-LTR within the human reference genome, and therefore a polymorphic status is obtained by verifying the absence of this insertion within at least one haploid test genome. In addition to validating the absence of an insertion, this PCR assay allowed the determination of the genotype for each individual sample.

All of the 7 candidate loci tested were verified to be polymorphic using the 6 trio samples, demonstrating a 100% accuracy of the computational prediction. These loci were subjected to further analysis using a panel of 24 anonymous diverse human DNA samples for surveying their frequency in the human populations. Since these loci were presented as solo-LTRs in the reference genome, we also checked to see whether a full-length LTR was present in any of the samples used in this study. For this purpose, we used the "universal" primers for LTR internal viral regions designed based on the internal region of the full-length HERVK, that are close to the LTR sequences, i.e. the *gag* and *env* genes, in combination with the primers designed in the region flanking the insertions. These primer combinations between the internal primers and the flanking primers were first tested on individuals that were homozygous with the PI alleles during the initial screening, since an apparent "homozygous" PI genotype can be obtained for samples carrying a PI/FL genotype due to the failure of amplifying the full-length LTR. The same is true for samples showing an apparent homozygous sLTR genotype. The combinations of internal primers with flanking primers only amplified in one locus, chr12:54013481-54014450, indicating the presence of the FL allele. By overlapping the images produced from the F+R, F+HERV_GAG_1 and R+HERV_ENV_3 primers, the full genotype of each individual locus can be determined. From the panel of 6 test individuals (Figure 5A), NA19238 was heterozygous for the sLTR and FL alleles. NA19239 remained homozygous with the PI alleles.

NA12892 was found to be homozygous with the FL alleles. The remaining individuals were all found to be heterozygous with the PI and FL. When the same primer combinations were used to screen the 24 sample panel (Figure 5B), 4 individuals (NA15233, NA15245, NA15224, NA15226) were shown to be heterozygous for the sLTR and FL, 3 individuals (NA15236, NA15385, NA15038) were heterozygous for PI and sLTR, 3 individuals (NA15510, NA15227, NA15242) were homozygous with the FL, one individual (NA15386) that was homozygous for the sLTR. The remaining 8 samples were all found to be heterozygous with the PI and FL. Based on the co-presence of the PI, sLTR and FL alleles among the individuals tested, this insertion is shown to be a novel case of trimorphic HERV-K.

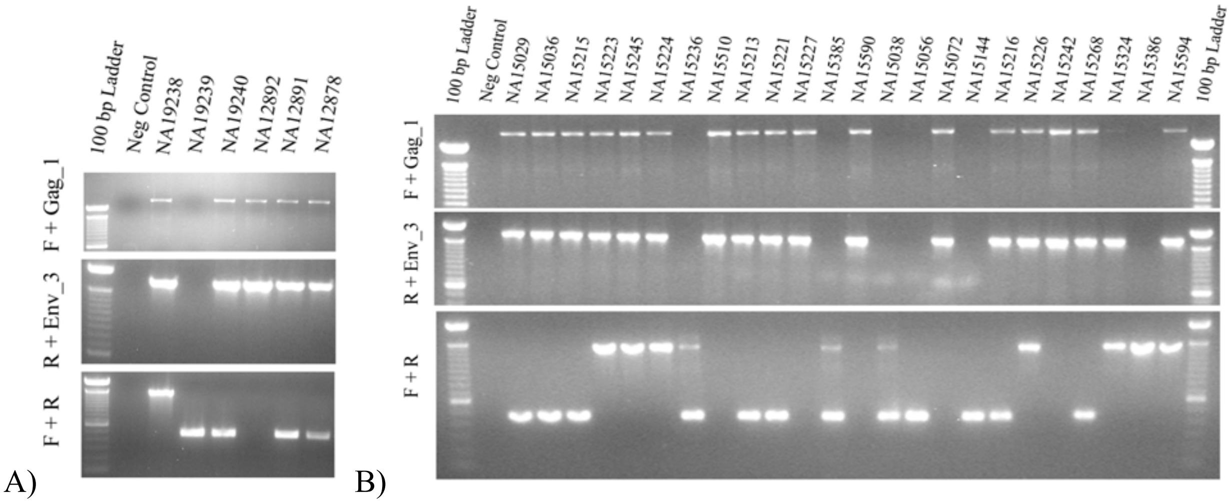

Figure 5: PCR genotype results for the combination of the F+R, F+GAG_1 and R+ENV_3 primers on candidate chr12:54013481-54014450.

A) Result for the 6 trio individuals. B) Results for the 24 human diversity panel. The sizes of the pre-integration band and sLTR alleles are 479 bp and 1449 bp respectively.

Candidate locus chr11:71155928-71156598 was shown to be heterozygous with the PI and sLTR alleles in all 30 of the individuals tested (Figure 6). In theory this is not impossible, but it is highly improbable based on Hardy-Weinberg equilibrium that a heterozygous allele combination will become fixed within a population, as there should be some individuals homozygous for the insertion as well as some homozygous insertion-free. Although the sample size tested is extremely small compared to the entire human population, the individuals tested are supposed to represent the major human ethnic populations. Therefore more individuals must be tested before this candidate locus can be classified as a fixed PI/sLTR insertion. Based on the presence of the PI and sLTR alleles among the individuals tested, this insertion has been categorised as dimorphic PI/sLTR. This may also be a result of PCR contamination.

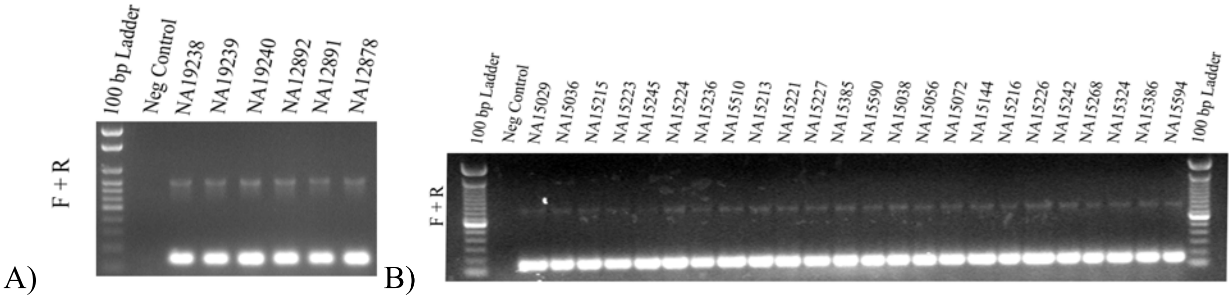

Figure 6: PCR genotype results for the combination of the F+R primers on candidate chr11:71155928-71156598.

A) Result for the 6 trio individuals. B) Results for the 24 human diversity panel. The sizes of the pre-integration and sLTR alleles are 175 bp and 897 bp, respectively.

The other 5 candidate loci only showed the presence of different combinations of the PI and sLTR alleles throughout the samples tested. All of these were all classified as dimorphic PI/sLTR. An example of this genotype status is illustrated with chr3:14107685-14108653 in Figure 7, whereas genotyping images for the remaining loci can be found in the

AGRAWAL, A., EASTMAN, Q.M. and SCHATZ, D.G., 1998. Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. *Nature,* **394**(6695), pp. 744-751.

ALTSHULER, D.M., LANDER, E.S., AMBROGIO, L., BLOOM, T., CIBULSKIS, K., FENNELL, T.J., GABRIEL, S.B., JAFFE, D.B., SHEFLER, E. and SOUGNEZ, C.L., 2010. A map of human genome variation from population scale sequencing.

BANNERT, N., 2006. The evolutionary dynamics of human endogenous retroviral families. *Annual review of genomics and human genetics,* **7**(1), pp. 149.

BARBULESCU, M., TURNER, G., SEAMAN, M.I., DEINARD, A.S., KIDD, K.K. and LENZ, J., 1999. Many human endogenous retrovirus K (HERV-K) proviruses are unique to humans. *Current biology,* **9**(16), pp. 861-S1.

BELSHAW, R., 2005. Genomewide screening reveals high levels of insertional polymorphism in the human endogenous retrovirus family HERV-K (HML2): implications for present-day activity. *Journal of virology,* **79**(19), pp. 12507.

BELSHAW, R., 2005. High copy number in human endogenous retrovirus families is associated with copying mechanisms in addition to reinfection. *Molecular biology and evolution,* **22**(4), pp. 814.

BENTLEY, G., 2009. High- resolution, high- throughput HLA genotyping by next-generation sequencing. *Tissue antigens,* **74**(5), pp. 393.

BLOMBERG, J., BENACHENHOU, F., BLIKSTAD, V., SPERBER, G. and MAYER, J., 2009. Classification and nomenclature of endogenous retroviral sequences (ERVs): problems and recommendations. *Gene,* **448**(2), pp. 115-123.

BRADY, T., LEE, Y.N., RONEN, K., MALANI, N., BERRY, C.C., BIENIASZ, P.D. and BUSHMAN, F.D., 2009. Integration target site selection by a resurrected human endogenous retrovirus. *Genes & development,* **23**(5), pp. 633-642.

BRANDT, J., SCHRAUTH, S., VEITH, A.M., FROSCHAUER, A., HANEKE, T., SCHULTHEIS, C., GESSLER, M., LEIMEISTER, C. and VOLFF, J.N., 2005. Transposable elements as a source of genetic innovation: expression and evolution of a family of retrotransposon-derived neogenes in mammals. *Gene,* **345**(1), pp. 101-111.

BÜSCHER, K., HAHN, S., HOFMANN, M., TREFZER, U., ÖZEL, M., STERRY, W., LÖWER, J., LÖWER, R., KURTH, R. and DENNER, J., 2006. Expression of the human endogenous retrovirus-K transmembrane envelope, Rec and Np9 proteins in melanomas and melanoma cell lines. *Melanoma research,* **16**(3), pp. 223-234.

BUZDIN, A., KOVALSKAYA-ALEXANDROVA, E., GOGVADZE, E. and SVERDLOV, E., 2006. At least 50% of human-specific HERV-K (HML-2) long terminal repeats serve in vivo as active promoters for host nonrepetitive DNA transcription. *Journal of virology,* **80**(21), pp. 10752-10762.

BUZDIN, A., LEBEDEV, Y.B. and SVERDLOV, E., 2003. Human-specific HERV-K intron LTRs have nonaccidental opposite orientation relative to the direction of gene transcription and might be involved in the antisense regulation of gene expression. *Russian Journal of Bioorganic Chemistry,* **29**(1), pp. 91-93.

COLLINS, F., LANDER, E., ROGERS, J., WATERSTON, R. and CONSO, I., 2004. Finishing the euchromatic sequence of the human genome. *Nature,* **431**(7011), pp. 931-945.

CORDAUX, R., 2009. The impact of retrotransposons on human genome evolution. *Nature reviews.Genetics,* **10**(10), pp. 691.

CORIELL, 2011-last update, Human Population Collections [Homepage of Coriell Institute], [Online]. Available: http://www.ccr.coriell.org/Sections/BrowseCatalog/Populations.aspx?PgId=42011].

COSTAS, J., 2001. Evolutionary dynamics of the human endogenous retrovirus family HERV-K inferred from full-length proviral genomes. *Journal of Molecular Evolution,* **53**(3), pp. 237-243.

DANGEL, A.W., BAKER, B.J., MENDOZA, A.R. and YU, C.Y., 1995. Complement component C4 gene intron 9 as a phylogenetic marker for primates: long terminal repeats of the endogenous retrovirus ERV-K (C4) are a molecular clock of evolution. *Immunogenetics,* **42**(1), pp. 41-52.

DE MAGALHÃES, J.P., FINCH, C.E. and JANSSENS, G., 2010. Next-generation sequencing in aging research: emerging applications, problems, pitfalls and possible solutions. *Ageing research reviews,* **9**(3), pp. 315.

DICKERSON, F., RUBALCABA, E., VISCIDI, R., YANG, S., STALLINGS, C., SULLENS, A., ORIGONI, A., LEISTER, F. and YOLKEN, R., 2008. Polymorphisms in human endogenous retrovirus K-18 and risk of type 2 diabetes in individuals with schizophrenia. *Schizophrenia research,* **104**(1), pp. 121-126.

DOXIADIS, G.G.M., DE GROOT, N. and BONTROP, R.E., 2008. Impact of endogenous intronic retroviruses on major histocompatibility complex class II diversity and stability. *Journal of virology,* **82**(13), pp. 6667-6677.

EICKBUSH, T.H., 1997. Telomerase and retrotransposons: which came first? *Science,* **277**(5328), pp. 911-912.

FRANK, O., VERBEKE, C., SCHWARZ, N., MAYER, J., FABARIUS, A., HEHLMANN, R., LEIB-MÖSCH, C. and SEIFARTH, W., 2008. Variable transcriptional activity of endogenous retroviruses in human breast cancer. *Journal of virology,* **82**(4), pp. 1808-1818.

GIFFORD, R. and TRISTEM, M., 2003. The evolution, distribution and diversity of endogenous retroviruses. *Virus genes,* **26**(3), pp. 291-315.

GOODIER, J.L., 2008. Retrotransposons revisited: the restraint and rehabilitation of parasites. *Cell,* **135**(1), pp. 23.

GRIFFITHS, D.J., 2001. Endogenous retroviruses in the human genome sequence. *Genome Biol,* **2**(6), pp. 1017.1-1017.5.

HERBST, H., SAUTER, M., KÜHLER-OBBARIUS, C., LÖNING, T. and MUELLER-LANTZSCH, N., 1998. Human endogenous retrovirus (HERV)-K transcripts in germ cell and trophoblastic tumours. *Apmis,* **106**(1-6), pp. 216-220.

HORMOZDIARI, F., HAJIRASOULIHA, I., DAO, P., HACH, F., YORUKOGLU, D., ALKAN, C., EICHLER, E.E. and SAHINALP, S.C., 2010. Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics,* **26**(12), pp. i350-i357.

HORMOZDIARI, F., ALKAN, C., VENTURA, M., HAJIRASOULIHA, I., MALIG, M., HACH, F., YORUKOGLU, D., DAO, P., BAKHSHI, M. and SAHINALP, S.C., 2011. Alu repeat discovery and characterization within human genomes. *Genome research,* **21**(6), pp. 840-849.

HU, L., HORNUNG, D., KUREK, R., ÖSTMAN, H., BLOMBERG, J. and BERGQVIST, A., 2006. Expression of human endogenous gammaretroviral sequences in endometriosis and ovarian cancer. *AIDS Research & Human Retroviruses,* **22**(6), pp. 551-557.

HUGHES, J.F., 2004. Human endogenous retrovirus K solo-LTR formation and insertional polymorphisms: implications for human and viral evolution. *Proceedings of the National Academy of Sciences of the United States of America,* **101**(6), pp. 1668.

ILLARIONOVA, A., VINOGRADOVA, T. and SVERDLOV, E., 2007. Only those genes of the KIAA1245 gene subfamily that contain HERV (K) LTRs in their introns are transcriptionally active. *Virology,* **358**(1), pp. 39-47.

IWABUCHI, H., KAKIHARA, T., KOBAYASHI, T., IMAI, C., TANAKA, A., UCHIYAMA, M. and FUKUDA, T., 2004. A gene homologous to human endogenous retrovirus overexpressed in childhood acute lymphoblastic leukemia. *Leukemia & lymphoma,* **45**(11), pp. 2303-2306.

JHA, A.R., NIXON, D.F., ROSENBERG, M.G., MARTIN, J.N., DEEKS, S.G., HUDSON, R.R., GARRISON, K.E. and PILLAI, S.K., 2011. Human endogenous retrovirus K106 (HERV-K106) was infectious after the emergence of anatomically modern humans. *PloS one,* **6**(5), pp. e20234.

JHA, A.R., PILLAI, S.K., YORK, V.A., SHARP, E.R., STORM, E.C., WACHTER, D.J., MARTIN, J.N., DEEKS, S.G., ROSENBERG, M.G. and NIXON, D.F., 2009. Cross-sectional dating of novel haplotypes of HERV-K 113 and HERV-K 115 indicate these proviruses originated in Africa before Homo sapiens. *Molecular biology and evolution,* **26**(11), pp. 2617-2626.

JOHNSON, W.E. and COFFIN, J.M., 1999. Constructing primate phylogenies from ancient retrovirus sequences. *Proceedings of the National Academy of Sciences,* **96**(18), pp. 10254-10260.

KOBAYASHI, S., GOTO-YAMAMOTO, N. and HIROCHIKA, H., 2004. Retrotransposon-induced mutations in grape skin color. *Science,* **304**(5673), pp. 982-982.

KONKEL, M.K., 2010. A mobile threat to genome stability: the impact of non-LTR retrotransposons upon the human genome. *Seminars in cancer biology,* **20**(4), pp. 211.

KORLACH, J., BJORNSON, K.P., CHAUDHURI, B.P., CICERO, R.L., FLUSBERG, B.A., GRAY, J.J., HOLDEN, D., SAXENA, R., WEGENER, J. and TURNER, S.W., 2010. Real-Time DNA Sequencing from Single Polymerase Molecules. In: NILS G. WALTER, ed, *Methods in Enzymology.* Academic Press, pp. 431-455.

KUMAR, S. and SUBRAMANIAN, S., 2002. Mutation rates in mammalian genomes. *Proceedings of the National Academy of Sciences,* **99**(2), pp. 803-808.

KURTH, R., 2010. Beneficial and detrimental effects of human endogenous retroviruses. *International journal of cancer,* **126**(2), pp. 306.

LANDER, E.S., 2001. Initial sequencing and analysis of the human genome. *Nature,* **409**(6822), pp. 860.

LEBEDEV, Y.B., BELONOVITCH, O.S., ZYBROVA, N.V., KHIL, P.P., KURDYUKOV, S.G., VINOGRADOVA, T.V., HUNSMANN, G. and SVERDLOV, E.D., 2000. Differences in HERV-K LTR insertions in orthologous loci of humans and great apes. *Gene,* **247**(1), pp. 265-277.

LIFE TECHNOLOGIES, 2010-last update, **AccuPrime™ GC-Rich DNA Polymerase**. Available: http://products.invitrogen.com/ivgn/product/12337024 [01/05, 2013].

LIFE TECHNOLOGIES, 2010-last update, AccuPrime™*Pfx* DNA Polymerase. Available: http://products.invitrogen.com/ivgn/product/12344024 [01/05, 2013].

LIJAVETZKY, D., RUIZ-GARCÍA, L., CABEZAS, J.A., DE ANDRÉS, M.T., BRAVO, G., IBÁÑEZ, A., CARREÑO, J., CABELLO, F., IBÁÑEZ, J. and MARTÍNEZ-ZAPATER, J.M., 2006. Molecular genetics of berry colour variation in table grape. *Molecular Genetics and Genomics,* **276**(5), pp. 427-435.

MACFARLANE, C., 2004. Allelic variation of HERV-K (HML-2) endogenous retroviral elements in human populations. *Journal of Molecular Evolution,* **59**(5), pp. 642.

MACK, M., BENDER, K. and SCHNEIDER, P.M., 2004. Detection of retroviral antisense transcripts and promoter activity of the HERV-K (C4) insertion in the MHC class III region. *Immunogenetics,* **56**(5), pp. 321-332.

MANGENEY, M., RENARD, M., SCHLECHT-LOUF, G., BOUALLAGA, I., HEIDMANN, O., LETZELTER, C., RICHAUD, A., DUCOS, B. and HEIDMANN, T., 2007. Placental syncytins: Genetic disjunction between the fusogenic and immunosuppressive activity of retroviral envelope proteins. *Proceedings of the National Academy of Sciences,* **104**(51), pp. 20534-20539.

MCCLINTOCK, B., 1953. Induction of instability at selected loci in maize. *Genetics,* **38**(6), pp. 579.

MEDSTRAND, P., 2002. Retroelement distributions in the human genome: variations associated with age and proximity to genes. *Genome research,* **12**(10), pp. 1483.

METZKER, M.L., 2010. Sequencing technologies-the next generation. *Nature reviews.Genetics,* **11**(1), pp. 31.

MILLER, J.R., 2010. Assembly algorithms for next-generation sequencing data. *Genomics,* **95**(6), pp. 315.

MOYES, D., 2007. Insertional polymorphisms: a new lease of life for endogenous retroviruses in human disease. *Trends in genetics,* **23**(7), pp. 326.

NELSON, P.N., CARNEGIE, P., MARTIN, J., EJTEHADI, H.D., HOOLEY, P., RODEN, D., ROWLAND-JONES, S., WARREN, P., ASTLEY, J. and MURRAY, P.G., 2003. Demystified... human endogenous retroviruses. *Molecular Pathology,* **56**(1), pp. 11-18.

PANARO, M.A., CALVELLO, R., LISI, S., SACCIA, M., MITOLO, C.I. and CIANCIULLI, A., 2009. Viral sequence integration into introns of chemokine receptor genes. *Immunopharmacology and immunotoxicology,* **31**(4), pp. 589-594.

PONFERRADA, V., MAUCK, B. and WOOLEY, D.P., 2003. The envelope glycoprotein of human endogenous retrovirus HERV-W induces cellular resistance to spleen necrosis virus. *Archives of Virology,* **148**(4), pp. 659-675.

SANGER, F., 1988. Sequences, Sequences, and Sequences. *Annual Review of Biochemistry,* **57**(1), pp. 1-29.

SCHADT, E.E., 2010. A Window into Third Generation Sequencing. *Human molecular genetics,* **19**(r2), pp. R227.

SCHOLZ, M.B., LO, C. and CHAIN, P.S., 2012. Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. *Current opinion in biotechnology,* **23**(1), pp. 9-15.

SHENDURE, J., 2008. Next-generation DNA sequencing. *Nature biotechnology,* **26**(10), pp. 1135.

SICAT, J., SUTKOWSKI, N. and HUBER, B.T., 2005. Expression of human endogenous retrovirus HERV-K18 superantigen is elevated in juvenile rheumatoid arthritis. *The Journal of rheumatology,* **32**(9), pp. 1821-1831.

SNYDER, M., 2010. Personal genome sequencing: current approaches and challenges. *Genes development,* **24**(5), pp. 423.

STEWART, C., KURAL, D., STRÖMBERG, M.P., WALKER, J.A., KONKEL, M.K., STÜTZ, A.M., URBAN, A.E., GRUBERT, F., LAM, H.Y. and LEE, W., 2011. A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS genetics,* **7**(8), pp. e1002236.

STOYE, J.P., 2001. Endogenous retroviruses: Still active after all these years? *Current biology,* **11**(22), pp. R914.

SUBRAMANIAN, R.P., WILDSCHUTTE, J.H., RUSSO, C. and COFFIN, J.M., 2011. Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. *Retrovirology,* **8**(90doi), pp. 4690-4698.

SVERDLOV, E.D., 2000. Retroviruses and primate evolution. *BioEssays,* **22**(2), pp. 161.

TAKAHASHI, Y., HARASHIMA, N., KAJIGAYA, S., YOKOYAMA, H., CHERKASOVA, E., MCCOY, J.P., HANADA, K., MENA, O., KURLANDER, R. and ABDUL, T., 2008. Regression of human kidney cancer following allogeneic stem cell transplantation is associated with recognition of an HERV-E antigen by T cells. *The Journal of clinical investigation,* **118**(3), pp. 1099.

TARUSCIO, D., FLORIDIA, G., ZORAQI, G.K., MANTOVANI, A. and FALBO, V., 2002. Organization and integration sites in the human genome of endogenous retroviral sequences belonging to HERV-E family. *Mammalian genome,* **13**(4), pp. 216-222.

TENG, S.C., KIM, B. and GABRIEL, A., 1996. Retrotransposon reverse-transcriptase-mediated repair of chromosomal breaks. *Nature,* **383**(6601), pp. 641-644.

TRISTEM, M., 2000. Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database. *Journal of virology,* **74**(8), pp. 3715.

TURNER, G., 2001. Insertional polymorphisms of full-length endogenous retroviruses in humans. *Current biology,* **11**(19), pp. 1531.

VENTER, J.C., 2001. The sequence of the human genome. *Science,* **291**(5507), pp. 1304.

VITTE, C., 2003. Formation of solo-LTRs through unequal homologous recombination counterbalances amplifications of LTR retrotransposons in rice Oryza sativa L. *Molecular biology and evolution,* **20**(4), pp. 528.

WANG, J., SONG, L., GONDER, M.K., AZRAK, S., RAY, D.A., BATZER, M.A., TISHKOFF, S.A. and LIANG, P., 2006. Whole genome computational comparative genomics: A fruitful approach for ascertaining< i> Alu insertion polymorphisms. *Gene,* **365,** pp. 11-20.

WANG, T., ZENG, J., LOWE, C.B., SELLERS, R.G., SALAMA, S.R., YANG, M., BURGESS, S.M., BRACHMANN, R.K. and HAUSSLER, D., 2007. Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proceedings of the National Academy of Sciences,* **104**(47), pp. 18613-18618.

WHEELER, D.A., SRINIVASAN, M., EGHOLM, M., SHEN, Y., CHEN, L., MCGUIRE, A., HE, W., CHEN, Y., MAKHIJANI, V. and ROTH, G.T., 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature,* **452**(7189), pp. 872-876.

XING, J., WANG, H., BELANCIO, V.P., CORDAUX, R., DEININGER, P.L. and BATZER, M.A., 2006. Emergence of primate genes by retrotransposon-mediated sequence transduction. *Proceedings of the National Academy of Sciences,* **103**(47), pp. 17608-17613.

ZIOGAS, D., 2009. Genetics and personal genomics for personalized breast cancer surgery: progress and challenges in research and clinical practice. *Annals of Surgical Oncology,* **16**(7), pp. 1771.

Appendix 1 – TIPs_IN Positive Results and all results summarized in

Table 8: Summary of TIPs_IN candidate loci PCR genotyping. With the 24 sample panel, only one test sample (NA15224) did not amplify either the PI or sLTR. To ensure this was not a result of a failed PCR reaction, this sample was run a second time, with the same result. Since neither band amplified, this sample was further tested with all of the internal primers to determine if a full-length insertion was present. All of the internal primers did not amplify with the proper size, or as a single band (non-specific amplification). Therefore this sample is likely

not amplifying the PI or sLTR due to either a mutation in the primer binding site or a possible deletion of this region within this individual genome.

In summary, of the 7 candidates, 6 were classified as dimorphic PI/sLTR after genotyping, due to the presence of both the PI and sLTR among the individuals tested, while locus, chr12:54013481-54014450 was shown to be trimorphic based on the presence of all three possible alleles (PI, sLTR and FL). The full-length HERV allele identified here adds to a very short list of full-length HERV-K sequences outside the human reference genome sequences. Further, our results suggest that for a HERV insertion that is shown in the reference genome as a solo-LTR and dimorphic in some sample, it is very likely a full-length version can still exist in the human population.

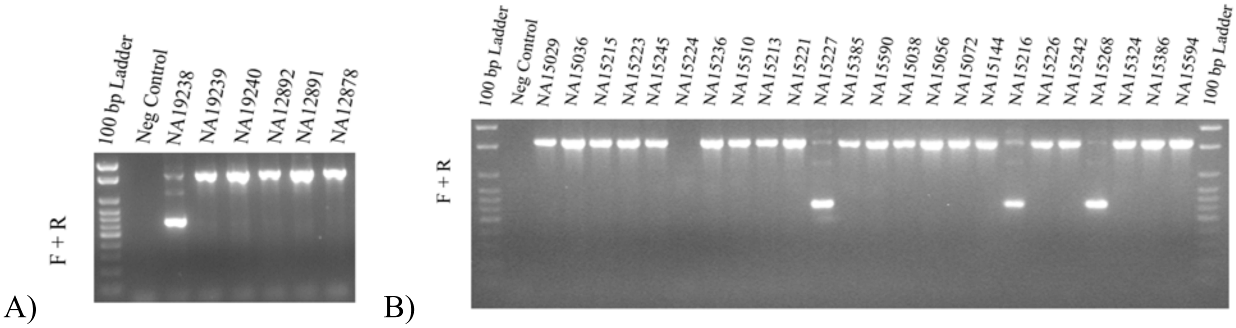

Figure 7: PCR genotype results for the combination of the F+R primers on candidate chr. 3:14107685-14108653.

A) Result for the 6 trio individuals. B) Results for the 24 human diversity panel. The size of the pre-integration band is 667 bp, and the size with the insertion is 1636 bp.

Table 8: Summary of TIPs_IN candidate loci PCR genotyping

			Number of Individuals with Each Genotype (Observed/Expected)						
Candidate Position	Insertion Status	Sample Size	PI/PI	PI/ sLTR	PI/ FL	sLTR/ sLTR	sLTR/ FL	FL/ FL	Observed AF PI/sLTR /FL
chr3:14107685-14108653	Dimorphic PI/sLTR	30	0/0.1	4/3.9	0	25/25.9	0	0	0.07/0.93/0
chr4:120483136-120484102	Dimorphic PI/sLTR	30	5/10.1	25/14.6	0	0/5.3	0	0	0.58/0.42/0
chr7:157722243-157723211	Dimorphic PI/sLTR	30	18/19.2	12/9.6	0	0/1.2	0	0	0.8/0.2/0
chr8:18695738-18696706	Dimorphic PI/sLTR	30	17/15.98716.0	10/11.8	0	3/2.2	0	0	0.73/0.27/0
chr8:37170043-37171011	Dimorphic PI/sLTR	30	0/6.6	28/14.9	0	2/8.4	0	0	0.47/0.53
chr11:71155928-71156598	Dimorphic PI/sLTR	30	0/7.5	30/15	0	0/7.5	0	0	0.5/0.5/0
chr12:54013481-54014450	Trimorphic	30	3/3.1	3/3.8	10/9.2	1/1.2	7/5.8	6/6.9	0.32/0.2/0.48

Verifying Full-Length TIPS_IN HERV-K Candidates using the Trio Samples and Anonymous 24 Individual Samples

There were a total of eight FL TIPs_IN candidates chosen to be genotyped by PCR based on the selection criteria outlined in the Materials and Methods section (Table 3). Each of these candidates exists as a full-length insertion within the human reference genome, and therefore a polymorphic status can be confirmed by verifying either the presence of sLTR or the absence of this insertion within at least one haploid test genome using a strategy similar to the ones used in genotyping TIPs_IN solo-LTRs. As polymorphic status of these loci were not predicted in any particular individual, each candidate locus was tested on both the 6 trio-individuals and the 24 sample panel. Of the eight candidate loci tested, five were found to only contain the full-length insertion, as no sLTR or PI could be amplified in any of the test individuals, thus are categorised as Fixed FL-HERV-K. The genotype images for the five candidates that appeared to be fixed FL insertions can be found in the

Appendix 2 – FL TIPs_IN Positive Results.

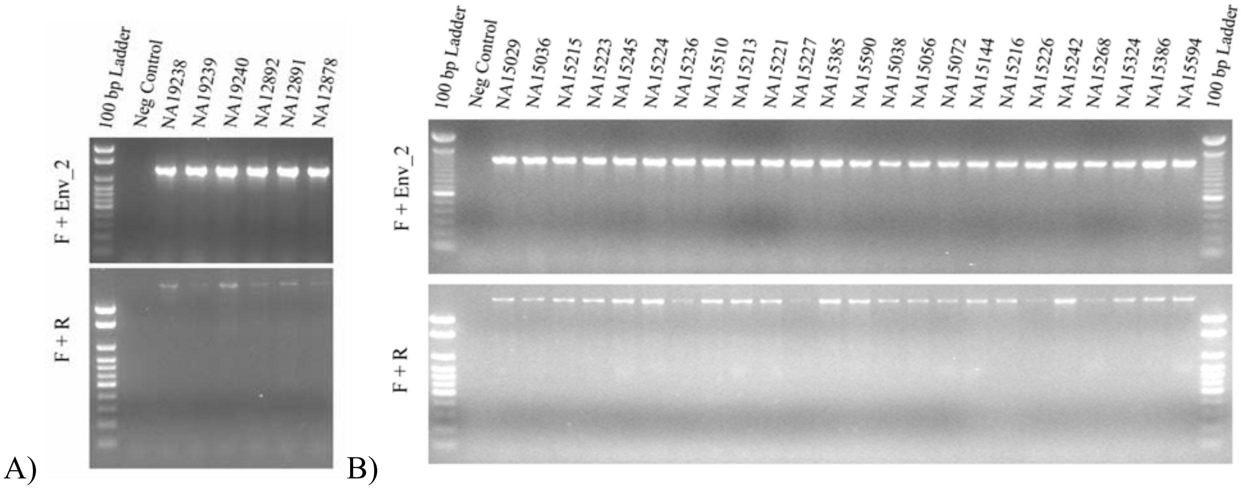

Figure 8: PCR genotype results for the combination of the F+R, F+GAG_1 and R+ENV_2 primers on candidate chr8:140473118-140475236.

A) Result for the 6 trio individuals. B) Results for the 24 human diversity panel. The sizes of the FL fragment band and the F+ENV_2 band are 3517 bp and 1230 bp, respectively.

Two of the eight candidates tested were found to be dimorphic. For candidate chr3:185281305-185288547, four of the 30 individuals tested (NA19238, NA15215, NA15056, NA15268) were found to be heterozygous with the sLTR and FL, with the rest being homozygous for the FL allele. Based on the presence of the sLTR and FL alleles among the individuals tested, this insertion has been categorised as dimorphic sLTR/FL. The genotype image for this candidate can be found in

Appendix **2** – FL TIPs_IN Positive **Results**. Candidate chr3:125610106-125617634 was found to be heterozygous with the PI and FL alleles in all 30 of the individuals tested. Again this is not impossible, but it is very improbable that this genotype would be found in the entire population as in the case of chr11:71155928-71156598 described earlier. This again may indicate contamination of the PCR.

Interestingly, locus chr11:101566762-101574290 was shown to be trimorphic by having the FL-LTR, sLTR and PI allele detected in the 30 samples (Figure 9). This becomes the second trimorphic HERV-K insertion identified in this study, increasing the total number of trimorphic HERV-K insertions ever identified to three (Belshaw 2005). This also indicates that for any full-length HERV-K to be documented in the human reference genome, there is a chance to find a dimorphism of PI and FL and even trimorphism.

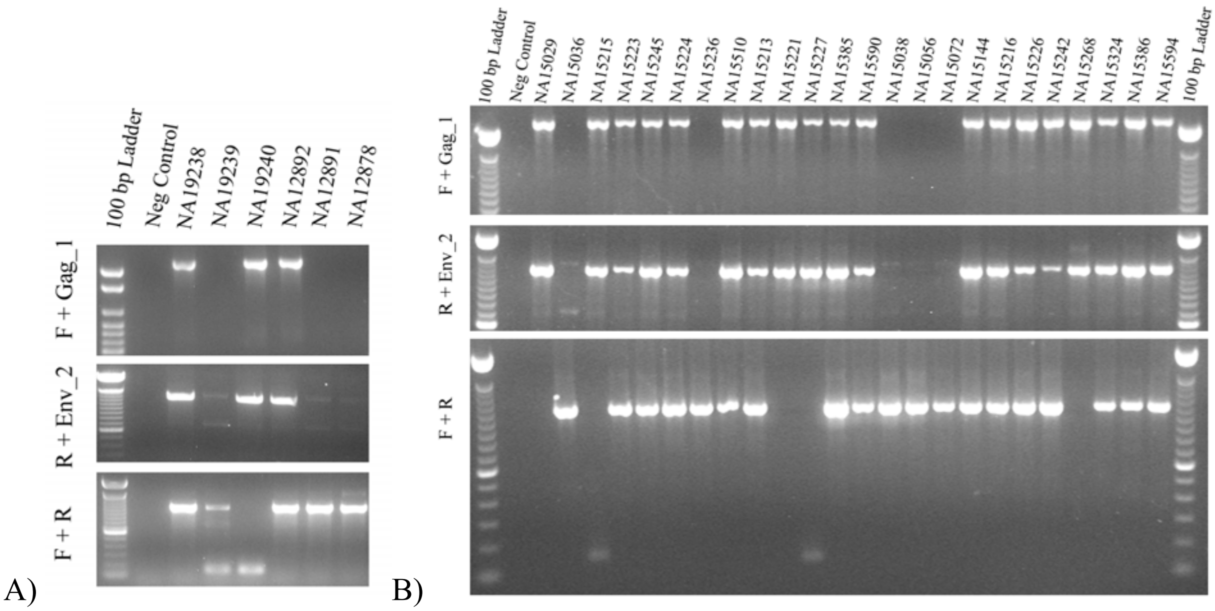

Figure 9: PCR genotype results for the combination of the F+R, F+GAG_1 and R+ENV_2 primers on candidate chr11:101566762-101574290.

A) Result for the 6 trio individuals. B) Results for the 24 human diversity panel. The sizes of the pre-integration band and F+GAG_1 band are 2249 bp and 1286 bp, respectively

In summary, a polymorphic status was identified for 3 of the 8 full-length HERVK loci tested, with one being dimorphic with the PI and FL alleles, another being dimorphic with the sLTR and FL alleles, and a third being a trimorphic case (Table 9: Summary of Full-length TIPs_IN candidate loci genotyping). For the 5 loci (chr3:10, chr8, chr10, chr21) that appear to show a fixed FL insertion, it may suggest that these insertions occurred relatively early and have certain advantage by maintaining in the full-length status, thus have had the chance to spread over all human populations an extremely high frequency. However, a PI allele may be later found when more diverse samples, especially samples from old population, are analyzed. In the same time, a sLTR allele of this insertion may be generated in future generations via recombination.

Table 9: Summary of Full-length TIPs_IN candidate loci genotyping

		Number of Individuals with Each Genotype (Observed/Expected)						
Candidate Position	Insertion Status	PI/PI	PI/sLTR	PI/FL	sLTR/sLTR	sLTR/FL	FL/FL	Observed AF PI/sLTR/FL
Chr3:185281305-185288547	Dimorphic sLTR/FL	0/0	0	0	0/0.1	4/3.9	26/25.9	0/0.07/0.93
Chr3:101411706-101418889	Fixed FL	0	0	0	0	0	30	0/0 /1
Chr3:125610106-125617634	Dimorphic PI/FL	0/7.5	0	30/15	0	0	0/7.5	0.5/0/0.5
Chr8:140473118-140475236	Fixed FL Fragment	0	0	0	0	0	30	0/0/1
Chr10:6867110-6874635	Fixed FL	0	0	0	0	0	30	0/0/1
Chr11:101566762-101574290	Trimorphic	0/0.1	1/2.2	3/1.7	7/8.1	16/12.8	3/5.0	0.07/0.52/0.41
Chr12:58722211-58729730	Fixed FL	0	0	0	0	0	30	0/0/1
Chr21:19933917-19940998	Fixed FL	0	0	0	0	0	30	0/0 /1

*All candidates that were homozygous for all individuals tested were excluded from the expected allele frequency (AF) calculations as there was not enough data available to calculate this.

Verifying TIPS_OUT candidates using the trio samples and anonymous 24 individual samples

In addition to the above polymorphic HERV-K insertions identified based on those present in the human reference genome sequences, we also explored the use of the newly available personal genome sequence data to identify novel HERV-K insertions not present in the reference genome. For this, 29 candidate loci were selected for genotyping from a larger list of computationally predicted TIPs_OUT HERV insertions, which was based on the analysis of the personal genome data of the 6 trio samples from the 1000 Genome Project, using criteria outlined in the Materials and Methods section (Table 4). Each candidate was first tested on the two trio-family samples from which the insertion prediction was based on. Each of these candidates exists as a pre-integration site within the human reference genome. Therefore a polymorphic status is obtained by verifying the presence of either a solo-LTR or full-length insertion.

Using a PCR design similar to the screening of the TIPs_IN, of the 29 candidates tested, 6 were found to be dimorphic PI/sLTR. Of these six loci, chr4:9590458-9590900 exhibited a very low insertion rate, with only one (NA19238) individual of the 30 tested having a heterozygous PI/sLTR insertion and all remaining 29 samples being homozygous for the PI allele (Figure 10). Although none of the internal universal primers were able to amplify the FL allele, the possibility that any of these might actually be PI/FL cannot be excluded. This candidate was further tested on the HD 11 panel of Africans North of the Sahara and the HD 12 panel of Africans South of the Sahara. Of these 23 individuals tested, only one other (NA17348) also displayed a heterozygous PI/sLTR insertion, while all remaining samples shown as homozygous for the PI alleles. This extremely low frequency of this sLTR allele suggests that this insertion is likely to be fairly recent and limited to certain Nigerian/African populations. However, a much larger sample size is needed to confirm this. Due to its extreme low allele frequency and presumably very recent insertion, it is reasonable to expect the likelihood of identifying the presence of a full-length allele by screening more individuals in the related populations

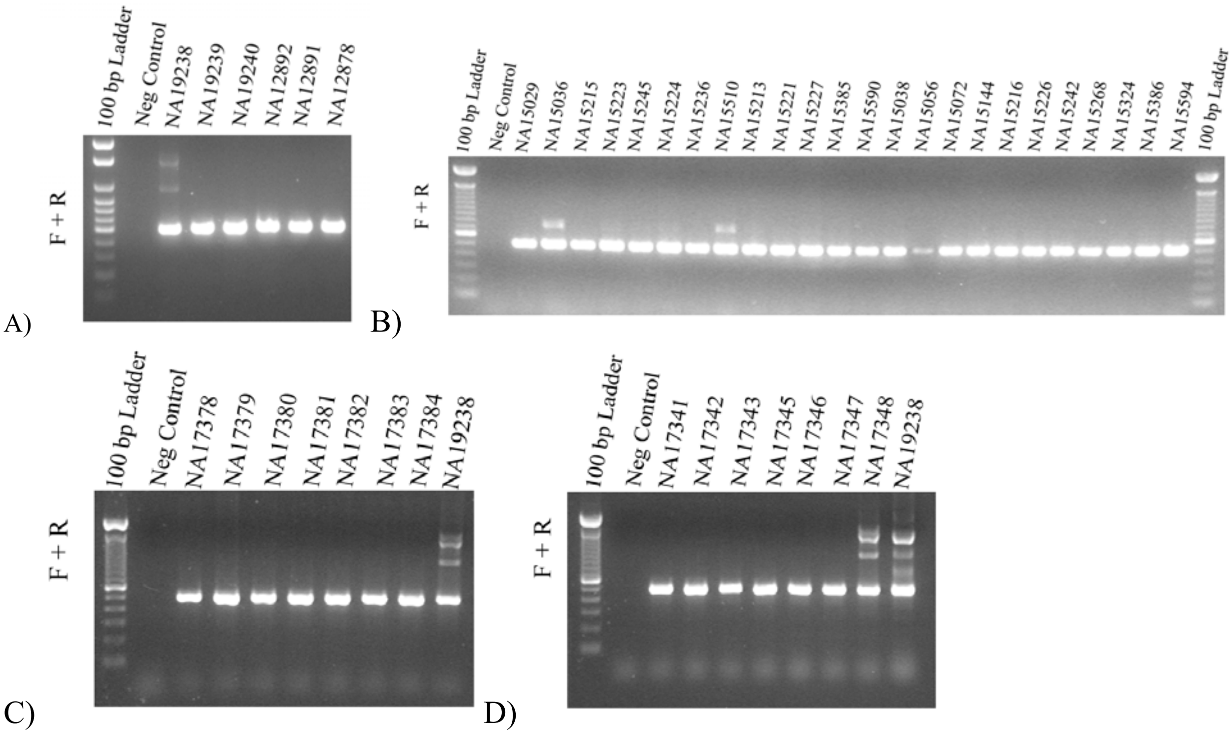

Figure 10: PCR genotype results for the combination of the F+R primers on candidate chr4:9590458-9590900.

A) Results for the 6-sample panel. B) Results for on the 24-sample panel. C) Results for HD 11 panel of Africans North of the Sahara. D) Results for the HD 12 panel of Africans South of the Sahara. The size of the PI band is 500 bp, and the sLTR band is roughly 1500 bp.

The genotype data of the other five candidates which were also found to be dimorphic PI/sLTR are provided in

Appendix 3 – TIPs_OUT Positive Results and summarized in Table 10. Among these, locus chr6:161190594-161191142 was shown to have a high frequency with 21 of the 30 individuals being homozygous with the sLTR, and 8 being heterozygous (NA12892, NA12891, NA15223, NA15224, NA15385, NA15590, NA15216, NA15242) with the PI/sLTR, and only one being homozygous for the PI (NA12878). By allele frequency, this is followed by locus chr13:89540897-89541416 for which 23 of the 30 individuals were homozygous with the PI alleles, and 7 (NA19238, NA19239, NA19240, NA15221, NA15144, NA15216, NA15268) were heterozygous with the PI/sLTR, and by locus chr20:12350168-12350569, chr21:15966603-15966908 and chr17:4959277-4959821.

To search for the presence of a full-length allele for each locus in samples with an apparent homozygous PI or sLTR genotype, which include the previous 6 loci and remaining loci for which no sLTR was detected, universal internal primers were used as described earlier. One difference in this case is that due to their absence in the reference genome, the orientation of the insertions was uncertain unlike in the case of TIPs_IN insertions, because this was not included as an output of the prediction data. Therefore, two possible combinations between the internal primers and the flanking primers were needed to be tested. One candidate, chr19:22205964-22206428 was showed to be dimorphic with the PI/FL alleles (Figure 11). In this case, the F+R combination of primers showed that all 6 individuals were homozygous with the PI, but the F+3LTR_1 amplified a product with the correct size for the presence of a full-length insertion in NA19239, NA19240 and NA12892, leading to their genotype of heterozygous PI/FL, and indicating that this insertion was in – strand orientation. These primers were then run on the 24 panel anonymous individuals where 7 individuals (NA15029, NA15223, NA15224, NA15213, NA15221, NA15590, NA15268) were homozygous with the FL, 6 individuals

(NA15036, NA15215, NA15242, NA15385, NA15038) were homozygous with the PI, and the remaining 11 individuals were heterozygous with the PI and FL. Based on the presence of only the FL alleles among the individuals tested, this insertion has been categorised as dimorphic PI/FL, and it represents a novel full-length HERV insertion outside of the reference genome.

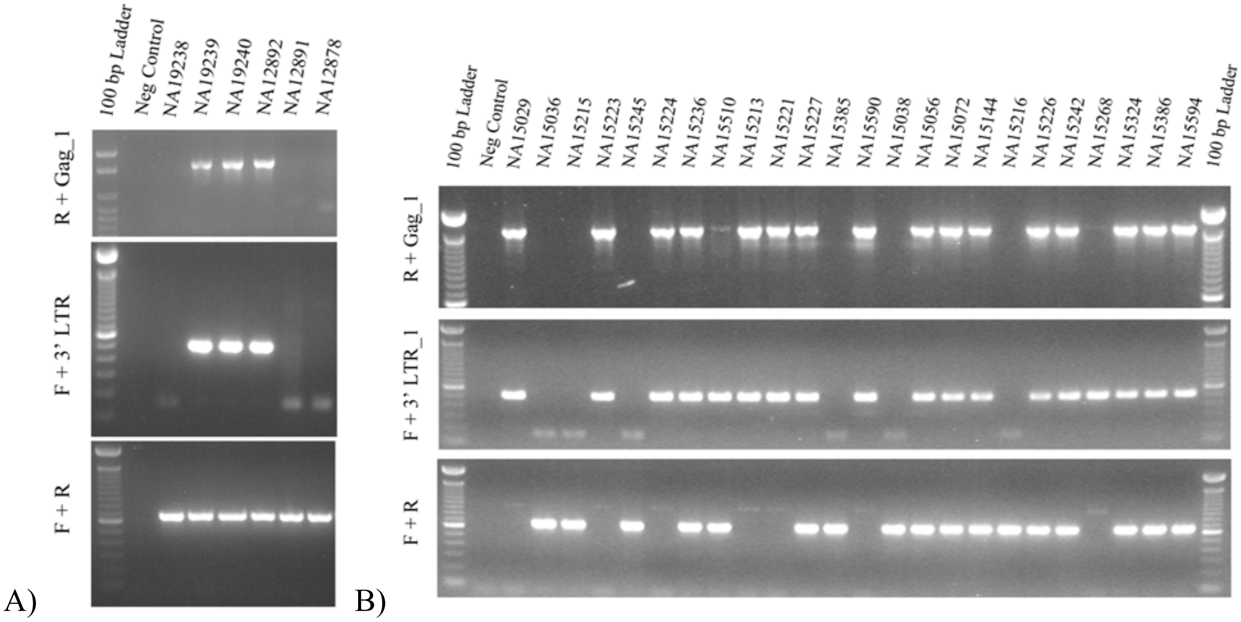

Figure 11: PCR genotype results for the combination of the F+R primers on candidate chr19:22205964-22206428.

A) Result for the 6 trio individuals. B) Results for the 24 human diversity panel. The size of the PI band is 646 bp, the LTR band is ~550 bp and the GAG band is ~1850 bp.

Two candidates, chr6:32565513-32565534 and chr11:124879106-124879109, were found to have a polymorphic PI allele that only amplified in some the 30 individuals tested, whereas the rest of the individuals showed no amplification with F+R, as well as combination of internal primers with flanking primers. The PCR results of chr6:32565513-32565534 is shown in Figure 12. To ensure the absence of amplification was not due to operation-error resulted PCR failure, PCR was repeated and the result ended up being the same. A likely explanation for this scenario is that a mutation/deletion may have occurred in the primer sites and prevents PCR from working in some individuals. One way to determine whether a full-length insertion is present as homozygous FL genotype is to run a long-range PCR with F+R primers. This was performed by a colleague and no amplification was observed.

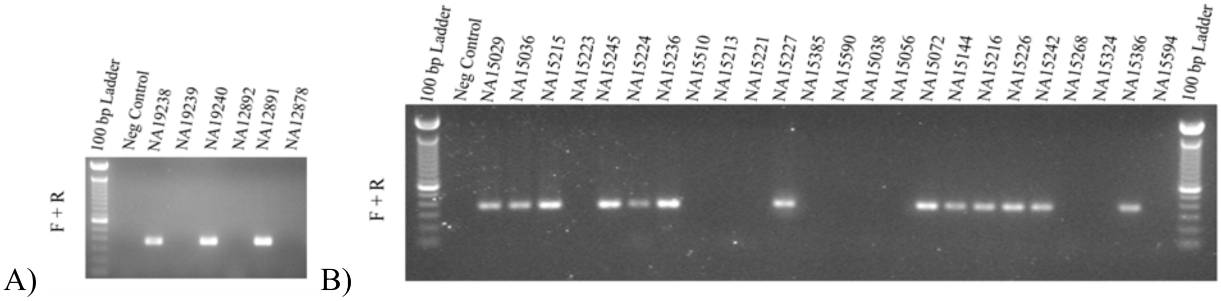

Figure 12: PCR genotype results for the combination of the F+R primers on candidate chr6:32565513-32565534.

A) Result for the 6 trio individuals. B) Results for the 24 human diversity panel. The size of the PI band is 369 bp.

Table 10: Summary of positive TIPs_OUT candidate loci PCR genotyping

			Number of Individuals with Each Genotype (Observed/Expected)						
Candidate Position	Insertion Status	Sample Size	PI/PI	PI/sLTR	PI/FL	sLTR/sLTR	sLTR/FL	FL/FL	Observed Allele Frequency PI/sLTR/FL
chr4:9590458-9590900	Dimorphic PI/sLTR	44	42/ 42.3	2/1.7	0	0/0.01	0	0	0.98/0.02/0
chr6:161190594-161191142	Dimorphic PI/sLTR	30	1/ 0.9	8/ 8.5	0	21/20.7	0	0	0.17/0.83/0
chr13:89540897-89541416	Dimorphic PI/sLTR	30	23/ 23.2	7/6.3	0	0/0.4	0	0	0.88/0.12/0
chr20:12350168-12350569	Dimorphic PI/sLTR	30	27/ 27.1	3/2.9	0	0/0.08	0	0	0.95/0.05/0
chr6:32565513-32565534	Polymorphic PI	30	16	0	0	0	0	0	0.53/0/0
chr11:124879106-124879109	Polymorphic PI	30	17	0	0	0	0	0	0.57/0/0
chr21:15966603-15966908	Dimorphic PI/sLTR	30	24/ 24.3	6/5.4	0	0/0.3	0	0	0.9/0.1/0
chr19:22205964-22206428	Dimorphic PI/FL	30	9/ 8.4	0	14/ 14.9	0	0	7/6.6	0.53/0/0.47

chr17:4959277-4959821	Dimorphic PI/sLTR	30	28/ 25.9	2/3.9	0	0/0.1	0	0	0.93/0.07/0
							Insertion Prediction Accuracy		24.1%

*All candidates that were homozygous for all individuals tested as well as those with a polymorphic pre-integration site were excluded from the expected allele frequency (AF) calculations as there was not enough data available to calculate this. This table excludes all of the candidates that were not shown to be polymorphic.

Based on these PCR result, a polymorphic status was verified for 7 of the 29 loci tested with their genotype results summarized in Table 10. Despite the presence of the sLTR, it should be noted that for individuals representing a genotype of "PI/PI" the possibility of the genotype actually representing "PI/sLTR" cannot be excluded due to the preferential amplification of the smaller product. Also in the instances where the only polymorphic genotype was shown as PI/sLTR or those with a polymorphic PI, the possibility of those with the genotype "PI/PI" or no amplification at all actually representing the genotype "PI/FL" cannot be excluded as the "universal" internal primers did not always amplify with one clear product. In these instances, due to the non-specific amplification, these primer combinations were not used to determine a FL insertion as it became too subjective to classify which product band reflected the FL insertion. Therefore these candidates should be considered for further screening in the future.

Combined Results of Polymorphism Survey from this Study

Overall in this study we were able to identify a total of 17 novel polymorphic insertions (

Table 11). Of these confirmed candidates, 12 were classified dimorphic PI/sLTR. One was classified dimorphic sLTR/FL, and two were dimorphic PI/FL. Interestingly 2 candidates were found to exhibit a trimorphic status. An additional 4 candidates were found to be a fixed FL insertion. Overall this study indicates that the level of HERV polymorphism is much higher than previously demonstrated in the published literature (Barbulescu, Turner et al. 1999, Turner 2001, Costas 2001, Hughes 2004, Belshaw 2005).

Table 11: Total number of candidates found within this study and their corresponding insertion status

Insertion Status	Number of Candidates Confirmed In This Study
Dimorphic PI/sLTR	12
Dimorphic sLTR/FL	1
Dimorphic PI/FL	2
Trimorphic	2
Fixed FL	4
Total Confirmed Polymorphic Loci	17

Sequence Data Analysis

A total of seventeen loci were chosen to be sequenced based on the presence of one or two alleles outside the human reference genome sequence, thus representing novel sequences. Such alleles can be the pre-integration allele for insertion present in the reference genome, the sLTR allele for a FL in the reference genome or the FL allele for insertion reported as sLTR in the reference genome. Out of those sequenced, only twelve were of high enough quality to properly assemble and analyse. The sequence data was used to construct a complete sequence of the insertion along with the flanking sequences and target site duplications (TSDs) by comparing the sequences of the pre-integration allele and the insertional allele. The target-site duplication, as a hallmark of retrotransposition, represents the additional sequence rearrangements caused by a HERV-K insertion. Also importantly the availability of the detailed sequence allows the definitive confirmation of the computational prediction and PCR genotyping result. An example of a novel sequence is provided in Figure 13, while the remaining sequences have been submitted to GenBank (http://www.ncbi.nlm.nih.gov/genbank/); accession numbers can be found in Table 12.

```
>NA_19238|locus: hg18:chr6:161190427-161191359, insertion site:
chr6:161190896, classification: LTR5_Hs
ACAGTGAACACAGAGAAAAGATGATGGTACACCCAGAGAAAAGATGATGGTACGCCgcctgtgg
ggaaaagcaagagagatcagattgttactgtgtctgtgtagaaagaagtagacataggagactc
cattttgttctgtactaagaaaaattcttctgccttgagattctgttaatctataacctcaccc
ccaaccccgtgctctctgaaacatgtgctgtgtcaaaatcagagttaaakggattaagggcggt
gcaagatgtgctttgttaaacagatgcttgaaggcagcatgctccttaagagtcatcaccactc
cctaatctcaagtacccagggacacaaaaactgcggaaggccgcagggacctctgcctaggaaa
gccaggtattgtccaaggtttctccccatgtgatagtctgaaatatggcctcctgggaagggaa
agacctgaccgtcccccagcccgacatccgtaaagggtctgtgctgaggaggattagtgaaaga
ggaaggaatgcctcttgcagttgagacaagaggaagtcatctgtctcctgcccgtccctgggca
atggaatgtctcggtataaaacccgattgtatgctccatctactgagatagggaaaagccacct
tagggctggaggtgggacctgcgggcagcaatactgctttgtaaagcattgagatgtttatgtg
tatgcatatctaaaagcacagcacttaatcctttacattgtctatgatgcaaagacctttgttc
acgtgtttgtctgctgaccctctccccacaattgtcttgtgaccctgacacatcccccttyttcg
agaaacacccacagatgatcaataaatactaagggaactcagaggttggcgggatcctccatat
gctgaacgctggttcccccggtccccttatttctttctctatactttgtctytgtgtctttttc
ttttccaaatctytcgtccccccttacgagaaacacccacaggtgtggaggggcaacccacccc
tacaTACGCCCAGAGAAAAGACGACATTGCACCCAGAGAAAAGATGACAGTGAACCTAGA
```

Figure 13: Detailed sequence and annotation of solo-LTR insertion at locus chr.6:161190427-161191359 subjected to DNA sequencing.

The sequence is presented in fasta format. The UPPER CASE sequence indicates the flanking/pre-integration sequence, while the lower case represents the solo-LTR insertion sequence and the TSDs are underlined. The information about the sample ID, original locus ID based on genomic location of the region represented by the flanking sequence in the hg18 reference genome, the exact insertion site, and the HERV subfamily designation are provided in the sequence description line.

In all cases, the sequencing result matches with the predicted loci positions, and validated the genotype result. This is based on the match of sequences flanking the involved HERV insertion within the human reference genome. The availability of the pre-integration and insertion allele enables very accurate identification of TSD sequences.

Table 12: Type of novel sequences obtained for each candidate, and their target site duplication.

Candidate	Type of TIP	Sequence Obtained	DNA Sample	Target Site Duplication (TSD)	GenBank Accesion #
chr3:14107685	TIPs_IN (sLTR)	Pre-integration Site	NA19238	AGAAGA	Pending
chr3:125610106	FL_TIPs_IN	Pre-integration Site	NA19238	TGTGG	Pending
chr6:161190594	TIPs_OUT	Solo-LTR	NA19238	TACGCC	Pending
chr7:15772224	TIPs_IN (sLTR)	Pre-integration Site	NA19238	CACTCTGC	Pending
chr8:18695738	TIPs_IN (sLTR)	Pre-integration Site	NA19238	ATGAAGC	Pending
chr8:37169884	TIPs_IN	Pre-integration Site	NA19240	GATTTT	Pending
chr11:71155870	TIPs_IN	Pre-integration Site	NA19238	CCCTC	Pending
chr11:101566762	FL_TIPs_IN	Solo-LTR	NA19238	ATTGTGT	Pending
chr11:101566762	FL_TIPs_IN	Pre-integration Site	NA19240	ATTGTGT	Pending
chr12:54013481	TIPs_IN (sLTR)	Pre-integration Site	NA19239	TAAAA	Pending
chr13:89540897	TIPs_OUT	Solo-LTR	NA19238	CTACTT	Pending
chr20:12402168	TIPs_OUT	Solo-LTR	NA19238	AAGTGG	Pending

As the two LTRs flanking the proviral insertion upon integration are identical initially, the sequence divergence found between the 5' and 3'LTRs of HERVs can serve as a molecular clock for estimating the ages of the insertion (Dangel, Baker et al. 1995). Using the LTR sequence divergence, the age of the full-length HERV proviral sequences found within the Hg19 human reference genome were estimated by comparing the sequence divergence between the 5’ and 3’ LTR (

Table 13). The upper bound was generated using the inferred evolutionary rate specific to HERV LTR of 1.3x10^{-9} mutations/site/year, resulting in a divergence rate of 0.13% per million years (My) as determined by Lebedev et al. 2000. This rate was generated by comparing the LTR sequence divergence of orthologous ERVs in different species and factoring in the time passed since these species diverged (Lebedev, Belonovitch et al. 2000). The lower bound was generated using the inferred mammalian genome rate of 2.2x10^{-9} mutations/site/year, resulting in a divergence rate of 0.22% per My as determined by Kumar and Subramanian 2002. This boundary was chosen as this rate has been reported as relatively invariant within and between primate genomes, thus representing the most conservative age estimation (Kumar, Subramanian 2002). The age estimates placed the overall insertion time of these HERV proviruses between 1.4-23 million years. It is highly unlikely that the estimates above 5 million years are accurate, as the human and chimpanzee lineages are thought to have diverged between 4-6 my ago (Buzdin, Lebedev et al. 2003). The remaining estimates ranging from 1.4-3.18 million years, agree with previous studies done with HERV-K (Dangel, Baker et al. 1995, Jha, Pillai et al. 2009, Subramanian, Wildschutte et al. 2011), furthering the evidence that these are the youngest subgroup of HERVs.

Table 13: Age estimation of full-length HERV insertions by comparing the sequence divergence between the 5' and 3' LTRs.

Candidate Position Hg19	#SNPs between 5' and 3' LTR	# Transitions /Transversions	% Sequence Variation	Lower Boundary (Kumar and Subramanian 2002) My	Upper Boundary (Lebedev 2000) My
Chr3:185281305-185288547	3	3/0	0.31	1.409	2.384
Chr3:101411706-101418889	15	9/6	2.77	7.036	11.91
Chr3:125610106-125617634	16	14/2	2.05	9.312	15.759
Chr10:6867110-6874635	29	24/5	3.3	13.6	23.02
Chr11:101566762-101574290	4	4/0	0.41	1.878	3.179
Chr12:58722211-58729730	4	4/0	0.41	1.878	3.179

*Million years (My)

Given a solo-LTR does not have a second LTR to compare to, their ages have mainly been overlooked in all of the previous studies. Subramanian et al. 2011 however, have proposed estimating the age of a solo-LTR by comparing it to the subgroup consensus sequence generated from the alignment of all known sLTRs in the human reference genome. This estimate was normalized to an average of $3.4x10^{-9}$ mutations/site/year, resulting in a sequence divergence of 0.34% per million years (Subramanian, Wildschutte et al. 2011). Using this estimate as an upper bound, the age of all polymorphic sLTRs identified in this study (with quality sequence data) were estimated (see Table 14). The lower bound of 0.22% per My as determined by Kumar and Subramanian (2002) was used to reflect a relatively conservative mutation rate. All of the solo-LTR sequences were determined to be from the LTR5_Hs subgroup, and this consensus sequence was used for all of these sequence comparisons. Only the SNPs and 1bp insertion/deletions were factored into the sequence divergence, as larger insertions/deletions do not accurately reflect a point mutation, and can severely skew the age estimates. The age estimates for the polymorphic insertions range from 1.8 million years to 7.9 million years. These estimates coincide with the estimated time of divergence between the human and chimpanzee lineages, and agrees with other published sLTR estimates (Buzdin, Lebedev et al. 2003, Subramanian, Wildschutte et al. 2011). This reveals that all of these insertions are very young with regards to evolutionary timeframes, furthering the evidence that the HERV-K subgroup has been active most recently.

Table 14: Age estimation of HERV-K solo-LTR by comparing the sequence divergence between polymorphic loci identified and the LTR5_Hs consensus sequence.

Candidate Position Hg18	#SNPs compared to LTR_Hs consensus sequence	# Transitions /Transversions	% Sequence Variation	Lower Boundary (Subramanian 2011) My	Upper Boundary (Kumar and Subramanian 2002) My
Chr3:14107685-14108653	12	8/4	1.24	3.646	5.635
Chr4:120483136-120484102	9	8/1	1.14	3.342	5.165
Chr7:157722243-157723211	8	7/1	0.83	2.431	3.757
Chr8:18695738-18696706	10	8/2	1.03	3.038	4.696
Chr8:37170043-37171011	6	5/1	0.62	1.823	2.817
Chr11:71155928-71156598	10	10/0	1.47	4.338	6.704
Chr12:54013481-54014450	12	11/1	1.34	3.946	6.098
Chr6:161190594-161191142	17	10/7	1.76	5.165	7.983

Functional Impact of polymorphic HERVs Based on Gene Context

To predict the functional impact each of the verified polymorphic loci was compared to the UCSC Genome Browser to determine the gene context of these insertions. If an insertion was within 2kb of any genes, it is considered as genic, otherwise as intergenic. Of the 17 polymorphic candidates identified, 12 were found to be intergenic. The low number of HERV insertions within or near gene regions is expected, as these are more likely to reduce fitness and become excised or inactivated (Kurth 2010). Intergenic insertions are less likely to have a definitive functional impact on genes due to their distance from the gene. However, they may also exert impact on genes via interfering distant gene regulatory sites or epigenetic regulation. Among the 5 insertions located within gene regions, 4 are in the intron region and 1 in the down stream region. Interestingly, all of these insertions were found to be oriented anti-sense compared to the gene transcription, which provides support to the theory that HERV integrations within the intron are more likely to be found in the anti-sense direction due to selective pressure (Illarionova, Vinogradova et al. 2007, Doxiadis, De Groot et al. 2008).

Table 15: Gene context for all of the insertionally polymorphic HERVs identified in this study

Candidate Position HG18	Insertion Status	Gene Context
Chr3:14107685-14108653	Dimorphic PI/sLTR	Intergenic
Chr4:120483136-120484102	Dimorphic PI/sLTR	Intergenic
Chr7:157722243-157723211	Dimorphic PI/sLTR	intron:PTPRN2:NM:002847
Chr8:18695738-18696706	Dimorphic PI/sLTR	intron:PSD3:NM:206909
Chr8:37170043-37171011	Dimorphic PI/sLTR	Intergenic
Chr11:71155928-71156598	Dimorphic PI/sLTR	Intergenic
Chr12:54013481-54014450	Trimorphic	~790bp downstream of OR6C3
Chr3:185281305-185288547	Dimorphic sLTR/FL	Intergenic
Chr3:125610106-125617634	Dimorphic PI/FL	Intergenic
Chr11:101566762-101574290	Trimorphic	Intergenic
Chr4:9590458-9590900	Dimorphic PI/sLTR	intron:SLC2A9:NM:001001290
Chr6:161190594-161191142	Dimorphic PI/sLTR	Intergenic
Chr13:89540897-89541416	Dimorphic PI/sLTR	Intergenic
Chr20:12350168-12350569	Dimorphic PI/sLTR	Intergenic
Chr21:15966603-15966908	Dimorphic PI/sLTR	Intergenic
Chr19:22205964-22206428	Dimorphic PI/FL	Intergenic
Chr17:4959277-4959821	Dimorphic PI/sLTR	intron:ZNF232: NM:014519.2

Discussion

HERV-K shows evidence of being active since the divergence of humans and chimpanzees, but to date there have been no replication competent endogenous HERVs identified (Brady, Lee et al. 2009). This had led to the widespread belief that HERVs have very limited to no current activity level within the human genome. The availability of multiple human genome sequences has led us to examine the current activity level of HERVs. The identification of HERVs has been previously been achieved by detecting ERV containing clones in BAC libraries using high-stringency hybridization with retrovirus-derived probes or synthetic primer binding sites; followed by PCR amplification of the novel ERV sequences from host genomic DNA (Gifford, Tristem 2003). These methods were extremely time consuming and plagued by limitations without having the full human genome sequence. Since the first publication of the human reference genome (Lander 2001, Venter 2001), and the increase in the more recently published personal genome data (Altshuler, Lander et al. 2010), the methods for identifying retrotransposon and HERV insertions have changed considerably and become much more efficient. Computational comparative genomics has become the method of choice for identifying retrotransposon insertions by comparing individual genomes to the human reference genome (Wang, Song et al. 2006, Stewart, Kural et al. 2011, Hormozdiari, Alkan et al. 2011). This method allows a much more efficient way (with both cost and time) to identify potential insertion candidates, but is limited by the quality and sequence read lengths of the genome sequence data generated by the next generation sequencing technologies. Given these limitations and the relative infancy of these prediction algorithms, this method of prediction is bound to generate false positive results, and therefore must be validated experimentally.

Evaluating approaches for discovering novel polymorphic HERV insertions

In this study, we explored three different approaches to identify novel polymorphic HERV-K insertions. Each of the three prediction methods used to identify polymorphic loci in this study proved fruitful, although with varying degrees of success. Using a broad panel of individual DNA samples, the FL TIPs_IN method allowed the survey of FL HERV-K insertions found within the human reference genome that shared a 95% or higher sequence similarity to the HERV-K113 sequence. This survey revealed that three of the eight loci tested are insertionally polymorphic, with the PI or sLTR alleles found in at least one test individual.

The other two methods relied entirely on the computational comparison of individual test genomes to the human reference genome. All seven of the TIPs_IN candidates were shown to be polymorphic, as the sLTR found within the human reference genome was proven to be absent in at least one haploid test genome (

Table 8). Given that the insertion was present in each candidate initially, this leads to the assumption that the full-length HERV-K retroviral element parent copy was active during early human migration before the divergence of different ethnic groups. When comparing all of the predicted genotypes for each donor genome with those obtained experimentally (

Table 8), it was found that only 28.6% were correct. This indicates that although this method did produce 100% accuracy for predicting a polymorphic locus at each candidate position, it is highly inaccurate for predicting the genotype of an individual sample.

In contrast, only 7 of the 29 candidates predicted using the TIPs_OUT method were verified to be polymorphic; as the sLTR or FL alleles were found in at least one of the haploid test genomes whereas only the pre-integration site was present within the human reference genome (Table 10). Six of the seven polymorphic candidates were classified as dimorphic PI/sLTR as these two alleles were present among the individuals tested. Only one of the 7 verified polymorphic candidates was found to be heterozygous with the PI/FL alleles. There were two candidates (chr6:32565513-32565534, chr11:124879106-124879109) that did not show any sLTR or FL alleles, but the pre-integration site appeared polymorphic among the 30 individuals tested, as it was absent in 14 and 13 individuals respectively. Due to the absence of the PI, the most likely explanation for this observed phenomenon is that these individuals have a mutation within the primer annealing site. This would render the primers incapable of amplifying this target region. Another explanation could be the possibility of a FL allele existing within the target region, which a normal PCR is unable to amplify due to the large size (~10kb). There were no “universal” primers that amplified without non-specific amplification, so there was no evidence provided from this study supporting any FL alleles. The PCR verification of the TIPs_OUT candidate list demonstrates 24.1% accuracy for the computational prediction of the polymorphic loci candidates that were selected (ie. with 75.9% false positives) (Table 10). The actual false positive rate from the computationally generated results is likely to be much higher than this, as the tested candidate list was heavily screened during the selection process and not randomly chosen. Therefore this accuracy rate can only reflect the candidates chosen, and should not be applied to the entire list that was initially generated.   Although the genotype prediction was 100% correct for 3 of the candidates tested in each of the 6 individuals from the trio samples, the combined prediction accuracy of the computational algorithm was only 52.4% correct. This number can be misleading, as one candidate’s genotype was not predicted correctly in any of the 6 individuals from the trio samples (0%), and two other candidate’s genotypes were only predicted correctly in 1 of the 6 samples (16.7%). Therefore the prediction accuracy overall appears to be either very accurate, or very inaccurate for the candidates that were verified. Although this method generated a very high rate of false positives, some of these results may in fact represent a positive result that was not observed due to the limitations of the PCR verification for FL alleles. One of these limitations is the PCR reagents themselves, which are limited to a maximum amplification size of ~5kb (Life Technologies 2010), which is only half of the FL HERV insertion length. In these instances the use of a long range PCR reagent kit may allow the amplification of the FL allele, as they can amplify fragments up to ~12kb (Life Technologies 2010). This method however would be limited by the presence of a pre-integration site allele or sLTR allele, causing the PCR to preferentially amplify the smaller product, leading to the potential of the FL allele to amplify with far fewer copies, and may appear faint or not at all when visualized on the agarose gel after electrophoresis. This is also the case in some of the individuals that were found to be homozygous with the PI alleles during a normal PCR reaction. The sLTR may actually be present, but given that the PI will preferentially amplify, the sLTR may not be visible on the agarose gel, resulting in a false indication of the genotype. In instances where the sLTR appeared to amplify very faint, each of the samples were re-tested in order to try and limit the occurrence of this phenomenon as much as possible throughout this study. The second limitation was the use of the "universal" internal and LTR primers. All of these primers were designed using an alignment of the known FL HERV-K insertions that are present in the human reference genome, and were designed within the most conserved areas. The computational prediction did not provide the predicted orientation of the insertion, resulting in this information to be gathered by testing combinations of the flanking primers with the "universal" LTR primers. In some instances the LTR primers did amplify, but as the size of the product can only be estimated, the possible correct size became a large range, with no guarantee that the product was actually amplifying from the target site. In the instances where the combination of a flanking primer and a "universal" LTR primer appeared to be within the correct size range, that individual sample was tested using all of the "universal" internal primers (in the *gag* and *env* genes) that were available using the orientation information provided from the LTR primer tests. If any of the primer annealing sites had any mutations, these primers may not amplify the target site. The fact that these "universal" primers are designed separately from the flanking primers also does not guarantee the proper efficiency, and these primers may not work as well together as those generated in pairs. These scenarios can lead to either no amplification or non-specific amplification of the target site. Given all of these complications, it is possible that the rate of FL alleles in the candidate list may actually be higher than was able to be verified in this study. The third possible limitation could be attributed to the primer design using the human reference genome. Since the TIPs_OUT candidates reflect insertions that are absent in the human reference genome but present in one of the haploid test genomes; more insertions are likely to be found in the Nigerian population, as the human reference genome is derived from mostly Caucasian DNA. It is also possible that the genomic sequence from the Nigerian trio samples have more sequence divergence in the primer annealing sites than are found within the human reference genome. This would lead to a higher rate of PCR failure in these individuals or the anonymous individuals which are closely related to this population group.

The major reason that there are so many false positives with the TIPs_OUT prediction is related to the method of prediction using paired-end reads. All of the current next generation sequencing technologies are limited by short read lengths and the accuracy of base calls, making the assembly of the whole genome from this raw data much more difficult. The largest sequence library size that are available from the 1000 Genome Project are only 250bp (with a standard deviation of 100bp) which is still very far from the 10kb size of a FL HERV insertion, and the ~970bp sLTR. If the read lengths could be increased to create a library size large enough to span the entire size of the insertion, the accuracy of the computational would be increased monumentally.

Although the TIPs_IN methods allowed the identification of a greater number of polymorphic loci, these methods are limited to the number of insertions found within the human reference genome. Therefore although these methods have produced the largest number of results, there are only a finite number of polymorphic loci that can be discovered using these two methods. In contrast, the TIPs_OUT method produced a much smaller number of true polymorphic loci, but this method is theoretically capable of detecting an infinite number of polymorphic loci, as it is only limited by the number of genomes available to compare to the human reference genome. Therefore as next generation sequencing technologies improve read lengths; the prediction accuracy of this method is bound to improve, and should be considered the method of choice for future discovery as it represents an unlimited number of potential polymorphic loci.

Redefining current classification nomenclature for documenting HERV insertion polymorphism

Overall, a regular RE insertion is limited to two possible combinations including the presence of the RE allele, or the absence of the RE allele. This has provided the basis of RE insertion polymorphism nomenclature, in which the presence of both the presence and absence of the RE insertion (PI) leads to a dimorphic classification of that RE insertion. Although HERVs are REs, the LTR class for which they belong increases the variability found within this group of polymorphism. Traditionally there have been three possible genotypes used to identify a polymorphic HERV insertion, which were originally proposed by Moyes et al. (2007). These include the absence of the HERV allele (PI), the presence of a FL HERV allele, or the sLTR allele which occurs as a result of the homologous recombination of the FL provirus LTRs (Moyes 2007). Throughout this study we have observed the frequent presence of additional forms of genotypes for these HERV insertions, and we propose for the first time the use of 6 types of genotypes to reflect all possible combinations of the PI, sLTR and FL alleles. As shown in Table 6, the six genotypes of a HERV insertion locus include PI/PI, PI/FL, FL/FL, PI/sLTR, sLTR/sLTR, sLTR/FL. In the context of polymorphism for the entire human population, any locus showing a genotype of PI/FL and/or PI/sLTR are considered to be polymorphic by the standard insertion polymorphism criterion, which applies to all non-LTR TE insertions. In addition, we argue that a locus showing a genotype of sLTR/FL, which may be in co-presence with sLTR/sLTR or FL/FL in the populations, should also be considered polymorphic by the definition of sequence polymorphism, since it means that there are clearly two different alleles for the same locus with one being sLTR and the other being FL, with these two combinations differing in sequence length by several kilo-bases, but both exist within the human population (Table 7).

Polymorphic loci sequence analysis

Since all of the polymorphic candidates were found by comparing to the human reference genome, it was important to obtain the DNA sequences for all novel alleles. For the TIPs_IN candidates these are represented by the PI and FL alleles, whereas for the FL_TIPs_IN they are represented by the PI and sLTR. The TIPs_OUT novel sequences are represented by the sLTR and FL alleles. These sequences provide the ultimate validation of the computational prediction of insertion polymorphism with complete or partial sequence, as well as the exact location of the insertion within the human reference genome. In all cases examined, the novel PI allele sequences matched those of the sequences flanking the HERV insertions found within the human reference genome. This indicates that these are true polymorphic insertions, and are not the result of the HERV proviral loss due to a recombination event involving a non-orthologous locus containing a sequence similar to that flanking the provirus. If this were the case the sequences flanking the provirus would likely have been mutated, resulting in a sequence difference between the flanking sequence and the PI sequence obtained (Turner 2001). All of the PI sequences were compared to the HERV insertion allele sequences and the target site duplications were noted in the PI sequence, also providing the exact location of each insertion within the human genome sequence (Table 12).

Using the candidates from the FL_TIPs_IN, representing the FL insertions found within the human reference genome, the 5' and 3' LTR sequence data was exploited in an attempt to estimate the age of these insertions (

Table 13). Given that the sequence of the 5' and 3' LTR that flank the proviral genes are identical upon integration as a result of retrotransposition, each sequence will evolve independently. Therefore the divergence between the two LTR sequences can serve as a molecular clock of the integration time by comparing to the proposed rates of sequence mutations. For this study an upper bound was generated using the inferred evolutionary rate specific to HERV LTR of $1.3x10^{-9}$ mutations/site/year, resulting in a divergence rate of 0.13% per million years (My) as determined by Lebedev et al. (2000). The lower bound was generated using the inferred mammalian genome rate of $2.2x10^{-9}$ mutations/site/year, resulting in a divergence rate of 0.22% per My as determined by Kumar and Subramanian 2002, as this represented a more conservative estimate. Using these methods the youngest FL insertion tested was found to be chr3:185281305-185288547 with an age of 1.4-2.4 million years, whereas the oldest insertion was chr10:6867110-6874635 with an age of 13.6-23.02 million years (

Table 13). The older the insertion is, the more likely those alleles will become fixed within the population, as it is more likely to incur mutations which preventing the homologous recombination between the 5' and 3' LTRs. It is important to note that the age estimates for chr3:101411706-101418889, chr3:125610106-125617634, chr10:6867110-6874635 are severely skewed by the high number of transitions. It has been proposed that transitions occur roughly 5-10 fold more often than transversions, and therefore these candidates may not follow the proposed molecular clock trends (Johnson, Coffin 1999). When comparing the estimated age to the polymorphic status of these insertions, the oldest insertions (ranging from 9.3-23.02 million years) are all fixed FL within the samples tested, with the exception of chr12:58722211-58729730 which deviates from this pattern. This candidate was found to be fixed FL within all the samples tested, yet it appears to be only 1.87-3.17 million years. Jha et al. (2011) also found discordance with insertion age and fixation, suggesting the probability of an ERV fixation may be inversely correlated with local chromosomal recombination rate and local gene density. If this is the case, then a low rate of recombination in the region surrounding this insertion may have led to the acceleration of its fixation. It is also possible that areas with a high rate of recombination will decelerate the fixation of the insertions that are older (Jha, Nixon et al. 2011). Aside from this exception, the younger insertions (ranging from 1.4-3.2 million years) have not yet become fixed within the population.

Similar age estimates were applied to the polymorphic solo-LTRs found throughout this study. Unlike the FL estimates that rely on divergence of the two LTRs, the solo-LTR can only be compared to the LTR subgroup consensus; which was the LTR5_Hs consensus sequence (Subramanian, Wildschutte et al. 2011). The upper bound of 0.34% per million years, or 3.4×10^{-9} mutations/site/years as proposed by Subramanian et al. (2011) was chosen as it is the only sLTR sequence divergence rate found within the literature. As a lower bound the 0.22% per My as determined by Kumar and Subramanian 2002 was chosen as it reflected a more conservative estimate. Using these age estimates, the youngest sLTR insertion was found to be chr8:37170043-37171011 with an age of 1.8-2.8 My, whereas the oldest insertion chr6:161190594-161191142 was found to be 5.2-7.9 My old. When comparing the sLTR allele frequencies in association with their estimated ages, there does not appear to be any direct trends, which may be a result of the relatively young ages of these insertions with regard to evolutionary timeframes. The age estimates for the FL and sLTR loci do however coincide with previous studies, furthering the evidence that the LTR5_Hs group have been continuously integrating into the germline since the divergence of humans and chimpanzees roughly 4-6 my ago (Subramanian, Wildschutte et al. 2011, Buzdin, Lebedev et al. 2003). It is important to note that all of the age estimates based on molecular clock calibrations are subject to a wide margin of error, as they are based on imprecise estimates of divergence dates (Johnson, Coffin 1999). Therefore these age estimates should only be used to provide rough estimates of absolute time, but are more useful for comparing relative evolution rates and ages of different HERV loci (Johnson, Coffin 1999). The sequence differences of the same allele among different individuals can also be used as an indication of their age, but this was beyond the scope of this study.

Functional Impact

HERVs have played an important role in primate evolution as they are known to create genomic rearrangements through several recombinational processes such as the generation of solo-LTRs, gene conversion events, excision of sequences located between two homologous proviruses and the recombination between LTRs of allelic proviruses (Doxiadis, De Groot et al. 2008). It has also been shown that full–length provirus and solo-LTR insertions can disrupt gene function or regulation of host genes to a different degree between the two types of LTR alleles (Doxiadis, De Groot et al. 2008). An example is the FL LTR retrotransposon insertion located upstream of the *VvmybA1*-coding sequence in *V. vinifera* grapes, which alters the gene expression resulting in the loss of red pigmentation, resulting in a white-skinned grape (Kobayashi, Goto-Yamamoto et al. 2004). In the instances where this insertion is found as a sLTR due to a post-insertion homologous recombination event, there is partial recovery of the gene expression, thus partial recovery of grape colour, resulting in two spontaneous colour variants of grapes (Ralli Seedless and Super Red) (Lijavetzky, Ruiz-García et al. 2006). Therefore it is possible for both the FL and separate homologous recombination events forming the sLTR alleles to have distinct impacts on the same gene.

HERV-K and their LTRs have been found to function *in vivo* as enhancers, promoters, transcription terminators and the origin of splice sites (Dangel, Baker et al. 1995, Buzdin, Lebedev et al. 2003, Taruscio, Floridia et al. 2002, Doxiadis, De Groot et al. 2008, Illarionova, Vinogradova et al. 2007, Panaro, Calvello et al. 2009). Therefore it is important to identify the proximity of these insertions to their surrounding genes. To investigate if any of these polymorphic insertions have any functional impact within the genome, each locus was analyzed using the UCSC Genome Browser to determine the gene context of these insertions (

Table 15). Out of the 16 candidates examined, 12 (71%) were found to be intergenic, as they were located at least 2kb or further from any surrounding genes. This result coincides with the supporting data that HERVs are less common in introns or in close proximity to genes than to intergenic regions (Kurth 2010). These are more likely to reduce fitness and become excised or inactivated, leading to a disproportionate accumulation of HERV sequences in gene-sparse regions (Buzdin, Lebedev et al. 2003, Doxiadis, De Groot et al. 2008, Kurth 2010). As such, these insertions are likely to have little to no impact on the expression of any of the surrounding genes, but it is important to note that a lot of regulatory elements are far away from the gene, so these may still be able to impact the regulation of surrounding genes. Of the remaining candidates, four (24%) were found within introns, and one (1%) was found ~790bp downstream of the nearest gene. Although there are mixed reviews on the impact of HERV insertions found within introns, the impact is suspected to be most closely related to the orientation of the insertion and the transcription direction of the gene (Dangel, Baker et al. 1995, Buzdin, Lebedev et al. 2003, Taruscio, Floridia et al. 2002, Doxiadis, De Groot et al. 2008, Illarionova, Vinogradova et al. 2007, Panaro, Calvello et al. 2009). Observing the insertions found within the introns revealed that the LTR orientation is opposite of the genes' transcription direction in all four loci. This supports the current data that HERV integrations into the introns are more likely found to be anti-sense to the direction of genes' transcription, aiding to the theory that there is strong selection against sense-directed integrations (Illarionova, Vinogradova et al. 2007, Doxiadis, De Groot et al. 2008). It has been proposed that the reason for this trend is due to the negative influence of sense-oriented HERVs on correct splicing of the targeted genes and post-transcriptional gene regulation due to RNA interference (Buzdin, Lebedev et al. 2003, Doxiadis, De Groot et al. 2008). This can be a consequence of the formation of double-stranded RNA between mRNA and the anti-sense transcript; resulting in the degradation of all mRNAs containing sites homologous to the double-stranded fragment (Buzdin, Lebedev et al. 2003). Anti-sense orientations have been found to down-regulate splicing activity, suggesting that splicing/exonization by anti-sense HERVs may be suppressed due to hybridizations with sense-oriented mRNA (Doxiadis, De Groot et al. 2008). It has also been speculated that an anti-sense transcript could be generated in the case of reverse orientation of the proviral sequence during host transcription; which may serve as a defence against other exogenous retroviruses of homologous sequence (Mack, Bender et al. 2004). This would allow protection by blocking the translation of newly expressed retroviral genes already in the genome, or prevent the initial integration of the viral DNA (Mack, Bender et al. 2004). As such, these may provide a selective advantage for the host, leading to the maintenance of the insertions in these loci (Mack, Bender et al. 2004, Doxiadis, De Groot et al. 2008).

Sense-oriented LTR insertions with powerful transcriptional termination signals in the gene intron can inactivate the gene by causing an early transcription termination (Buzdin, Lebedev et al. 2003). These types of insertions are likely to be deleterious mainly in monogenic systems, as there is an underrepresentation of these integrations in the human genome (Doxiadis, De Groot et al. 2008). Despite these observations, there have been instances where sense-oriented HERV insertions within the introns may have had a positive effect, contributing to the plasticity and diversity of the primate genomes in particular with multi-gene families, with the majority of this activity being attributed to the splice sites within the LTR and solo-LTRs (Doxiadis, De Groot et al. 2008). Therefore these studies indicate that the presence or absence of the insertions can have a strong influence on the particular hosts' genes, and these genes should be the subject of future studies.

Summary and Conclusions

Overall in this study we were able to identify a total of 17 novel polymorphic insertions using a broad sample of 30 individuals covering the major ethnic population groups by exploring three different strategies (

Table 11). This represents the largest discovery of polymorphic HERVs ever found, increasing the known polymorphic loci by over 150%. Of these confirmed candidates, 12 were classified dimorphic PI/sLTR. One was classified dimorphic sLTR/FL, and one was dimorphic PI/FL. Interestingly 2 candidates were found to exhibit a trimorphic status. An additional 4 candidates were found to be a fixed FL insertion. Overall this study indicates that the level of HERV polymorphism is much higher than previously demonstrated in the published literature. All of the prediction methods used throughout this study resulted in the identification of polymorphic HERV loci. While the TIPs_In prediction methods can only identify a finite number of polymorphic insertions, the TIPs_OUT prediction method represents a theoretically unlimited number of polymorphic candidate loci. The insertion age estimates of the loci tested place their integration time in the germline within the 1-6 million years, adding further evidence that HERV-K HML-2 represents the youngest HERV family in the human genome; thus HERV-K have infected humans in recent evolutionary times. There were four polymorphic loci that have inserted with the intron of surrounding genes, but as they are anti-sense orientations compared to the gene transcription direction, their impact can only be speculated. Given that these insertions are polymorphic among the individuals tested, future studies should be directed at identifying the functional impact that these insertions may have on these genes. Future studies should also be directed towards obtaining the sequence data associated with the novel insertions identified in this study that this study was unable to obtain, most notably those that are full-length insertions.

HERV insertional polymorphism identification will greatly benefit from the influx of personal genome data that is currently being obtained through next generation sequencing technologies. As these become available, the TIPs_OUT prediction method is likely to prove the most suitable way to identify polymorphic insertions. Thus this method is suspected to demonstrate the most useful for identifying HERV insertions associated with human diseases, especially those that are autoimmune in nature. As current PCR genotyping is limited by the "universal" internal primers needed to amplify the full-length HERV insertions, a more efficient method of detection is needed improve the identification of these insertions predicted using the TIPs_OUT algorithms. Although the panel of 30 individuals is capable of providing information on the distribution of polymorphic HERV alleles, this sample size is extremely small compared to the current human population and a larger sample size should be used in future studies in order to be able to draw more accurate conclusions. Although this study has greatly increased the current data on polymorphic HERVs, it is still extremely small, and the identification of new polymorphic loci will help provide more insight to both their current activity levels as well as any functional impact the may incur on the host. Therefore future research will ultimately help us gain more information and insights into how our genome works and evolves.

Literature Cited

AGRAWAL, A., EASTMAN, Q.M. and SCHATZ, D.G., 1998. Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. *Nature,* **394**(6695), pp. 744-751.

ALTSHULER, D.M., LANDER, E.S., AMBROGIO, L., BLOOM, T., CIBULSKIS, K., FENNELL, T.J., GABRIEL, S.B., JAFFE, D.B., SHEFLER, E. and SOUGNEZ, C.L., 2010. A map of human genome variation from population scale sequencing.

BANNERT, N., 2006. The evolutionary dynamics of human endogenous retroviral families. *Annual review of genomics and human genetics,* **7**(1), pp. 149.

BARBULESCU, M., TURNER, G., SEAMAN, M.I., DEINARD, A.S., KIDD, K.K. and LENZ, J., 1999. Many human endogenous retrovirus K (HERV-K) proviruses are unique to humans. *Current biology,* **9**(16), pp. 861-S1.

BELSHAW, R., 2005. Genomewide screening reveals high levels of insertional polymorphism in the human endogenous retrovirus family HERV-K (HML2): implications for present-day activity. *Journal of virology,* **79**(19), pp. 12507.

BELSHAW, R., 2005. High copy number in human endogenous retrovirus families is associated with copying mechanisms in addition to reinfection. *Molecular biology and evolution,* **22**(4), pp. 814.

BENTLEY, G., 2009. High- resolution, high- throughput HLA genotyping by next-generation sequencing. *Tissue antigens,* **74**(5), pp. 393.

BLOMBERG, J., BENACHENHOU, F., BLIKSTAD, V., SPERBER, G. and MAYER, J., 2009. Classification and nomenclature of endogenous retroviral sequences (ERVs): problems and recommendations. *Gene,* **448**(2), pp. 115-123.

BRADY, T., LEE, Y.N., RONEN, K., MALANI, N., BERRY, C.C., BIENIASZ, P.D. and BUSHMAN, F.D., 2009. Integration target site selection by a resurrected human endogenous retrovirus. *Genes & development,* **23**(5), pp. 633-642.

BRANDT, J., SCHRAUTH, S., VEITH, A.M., FROSCHAUER, A., HANEKE, T., SCHULTHEIS, C., GESSLER, M., LEIMEISTER, C. and VOLFF, J.N., 2005. Transposable elements as a source of genetic innovation: expression and evolution of a family of retrotransposon-derived neogenes in mammals. *Gene,* **345**(1), pp. 101-111.

BÜSCHER, K., HAHN, S., HOFMANN, M., TREFZER, U., ÖZEL, M., STERRY, W., LÖWER, J., LÖWER, R., KURTH, R. and DENNER, J., 2006. Expression of the human endogenous retrovirus-K transmembrane envelope, Rec and Np9 proteins in melanomas and melanoma cell lines. *Melanoma research,* **16**(3), pp. 223-234.

BUZDIN, A., KOVALSKAYA-ALEXANDROVA, E., GOGVADZE, E. and SVERDLOV, E., 2006. At least 50% of human-specific HERV-K (HML-2) long terminal repeats serve in vivo as active promoters for host nonrepetitive DNA transcription. *Journal of virology,* **80**(21), pp. 10752-10762.

BUZDIN, A., LEBEDEV, Y.B. and SVERDLOV, E., 2003. Human-specific HERV-K intron LTRs have nonaccidental opposite orientation relative to the direction of gene transcription and might be involved in the antisense regulation of gene expression. *Russian Journal of Bioorganic Chemistry,* **29**(1), pp. 91-93.

COLLINS, F., LANDER, E., ROGERS, J., WATERSTON, R. and CONSO, I., 2004. Finishing the euchromatic sequence of the human genome. *Nature,* **431**(7011), pp. 931-945.

CORDAUX, R., 2009. The impact of retrotransposons on human genome evolution. *Nature reviews.Genetics,* **10**(10), pp. 691.

CORIELL, 2011-last update, Human Population Collections [Homepage of Coriell Institute], [Online]. Available: http://www.ccr.coriell.org/Sections/BrowseCatalog/Populations.aspx?PgId=42011].

COSTAS, J., 2001. Evolutionary dynamics of the human endogenous retrovirus family HERV-K inferred from full-length proviral genomes. *Journal of Molecular Evolution,* **53**(3), pp. 237-243.

DANGEL, A.W., BAKER, B.J., MENDOZA, A.R. and YU, C.Y., 1995. Complement component C4 gene intron 9 as a phylogenetic marker for primates: long terminal repeats of the endogenous retrovirus ERV-K (C4) are a molecular clock of evolution. *Immunogenetics,* **42**(1), pp. 41-52.

DE MAGALHÃES, J.P., FINCH, C.E. and JANSSENS, G., 2010. Next-generation sequencing in aging research: emerging applications, problems, pitfalls and possible solutions. *Ageing research reviews,* **9**(3), pp. 315.

DICKERSON, F., RUBALCABA, E., VISCIDI, R., YANG, S., STALLINGS, C., SULLENS, A., ORIGONI, A., LEISTER, F. and YOLKEN, R., 2008. Polymorphisms in human endogenous retrovirus K-18 and risk of type 2 diabetes in individuals with schizophrenia. *Schizophrenia research,* **104**(1), pp. 121-126.

DOXIADIS, G.G.M., DE GROOT, N. and BONTROP, R.E., 2008. Impact of endogenous intronic retroviruses on major histocompatibility complex class II diversity and stability. *Journal of virology,* **82**(13), pp. 6667-6677.

EICKBUSH, T.H., 1997. Telomerase and retrotransposons: which came first? *Science,* **277**(5328), pp. 911-912.

FRANK, O., VERBEKE, C., SCHWARZ, N., MAYER, J., FABARIUS, A., HEHLMANN, R., LEIB-MÖSCH, C. and SEIFARTH, W., 2008. Variable transcriptional activity of endogenous retroviruses in human breast cancer. *Journal of virology,* **82**(4), pp. 1808-1818.

GIFFORD, R. and TRISTEM, M., 2003. The evolution, distribution and diversity of endogenous retroviruses. *Virus genes,* **26**(3), pp. 291-315.

GOODIER, J.L., 2008. Retrotransposons revisited: the restraint and rehabilitation of parasites. *Cell,* **135**(1), pp. 23.

GRIFFITHS, D.J., 2001. Endogenous retroviruses in the human genome sequence. *Genome Biol,* **2**(6), pp. 1017.1-1017.5.

HERBST, H., SAUTER, M., KÜHLER-OBBARIUS, C., LÖNING, T. and MUELLER-LANTZSCH, N., 1998. Human endogenous retrovirus (HERV)-K transcripts in germ cell and trophoblastic tumours. *Apmis,* **106**(1-6), pp. 216-220.

HORMOZDIARI, F., HAJIRASOULIHA, I., DAO, P., HACH, F., YORUKOGLU, D., ALKAN, C., EICHLER, E.E. and SAHINALP, S.C., 2010. Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics,* **26**(12), pp. i350-i357.

HORMOZDIARI, F., ALKAN, C., VENTURA, M., HAJIRASOULIHA, I., MALIG, M., HACH, F., YORUKOGLU, D., DAO, P., BAKHSHI, M. and SAHINALP, S.C., 2011. Alu repeat discovery and characterization within human genomes. *Genome research,* **21**(6), pp. 840-849.

HU, L., HORNUNG, D., KUREK, R., ÖSTMAN, H., BLOMBERG, J. and BERGQVIST, A., 2006. Expression of human endogenous gammaretroviral sequences in endometriosis and ovarian cancer. *AIDS Research & Human Retroviruses,* **22**(6), pp. 551-557.

HUGHES, J.F., 2004. Human endogenous retrovirus K solo-LTR formation and insertional polymorphisms: implications for human and viral evolution. *Proceedings of the National Academy of Sciences of the United States of America,* **101**(6), pp. 1668.

ILLARIONOVA, A., VINOGRADOVA, T. and SVERDLOV, E., 2007. Only those genes of the KIAA1245 gene subfamily that contain HERV (K) LTRs in their introns are transcriptionally active. *Virology,* **358**(1), pp. 39-47.

IWABUCHI, H., KAKIHARA, T., KOBAYASHI, T., IMAI, C., TANAKA, A., UCHIYAMA, M. and FUKUDA, T., 2004. A gene homologous to human endogenous retrovirus overexpressed in childhood acute lymphoblastic leukemia. *Leukemia & lymphoma,* **45**(11), pp. 2303-2306.

JHA, A.R., NIXON, D.F., ROSENBERG, M.G., MARTIN, J.N., DEEKS, S.G., HUDSON, R.R., GARRISON, K.E. and PILLAI, S.K., 2011. Human endogenous retrovirus K106 (HERV-

K106) was infectious after the emergence of anatomically modern humans. *PloS one,* **6**(5), pp. e20234.

JHA, A.R., PILLAI, S.K., YORK, V.A., SHARP, E.R., STORM, E.C., WACHTER, D.J., MARTIN, J.N., DEEKS, S.G., ROSENBERG, M.G. and NIXON, D.F., 2009. Cross-sectional dating of novel haplotypes of HERV-K 113 and HERV-K 115 indicate these proviruses originated in Africa before Homo sapiens. *Molecular biology and evolution,* **26**(11), pp. 2617-2626.

JOHNSON, W.E. and COFFIN, J.M., 1999. Constructing primate phylogenies from ancient retrovirus sequences. *Proceedings of the National Academy of Sciences,* **96**(18), pp. 10254-10260.

KOBAYASHI, S., GOTO-YAMAMOTO, N. and HIROCHIKA, H., 2004. Retrotransposon-induced mutations in grape skin color. *Science,* **304**(5673), pp. 982-982.

KONKEL, M.K., 2010. A mobile threat to genome stability: the impact of non-LTR retrotransposons upon the human genome. *Seminars in cancer biology,* **20**(4), pp. 211.

KORLACH, J., BJORNSON, K.P., CHAUDHURI, B.P., CICERO, R.L., FLUSBERG, B.A., GRAY, J.J., HOLDEN, D., SAXENA, R., WEGENER, J. and TURNER, S.W., 2010. Real-Time DNA Sequencing from Single Polymerase Molecules. In: NILS G. WALTER, ed, *Methods in Enzymology.* Academic Press, pp. 431-455.

KUMAR, S. and SUBRAMANIAN, S., 2002. Mutation rates in mammalian genomes. *Proceedings of the National Academy of Sciences,* **99**(2), pp. 803-808.

KURTH, R., 2010. Beneficial and detrimental effects of human endogenous retroviruses. *International journal of cancer,* **126**(2), pp. 306.

LANDER, E.S., 2001. Initial sequencing and analysis of the human genome. *Nature,* **409**(6822), pp. 860.

LEBEDEV, Y.B., BELONOVITCH, O.S., ZYBROVA, N.V., KHIL, P.P., KURDYUKOV, S.G., VINOGRADOVA, T.V., HUNSMANN, G. and SVERDLOV, E.D., 2000. Differences in HERV-K LTR insertions in orthologous loci of humans and great apes. *Gene,* **247**(1), pp. 265-277.

LIFE TECHNOLOGIES, 2010-last update, **AccuPrime™ GC-Rich DNA Polymerase**. Available: http://products.invitrogen.com/ivgn/product/12337024 [01/05, 2013].

LIFE TECHNOLOGIES, 2010-last update, AccuPrime™*Pfx* DNA Polymerase. Available: http://products.invitrogen.com/ivgn/product/12344024 [01/05, 2013].

LIJAVETZKY, D., RUIZ-GARCÍA, L., CABEZAS, J.A., DE ANDRÉS, M.T., BRAVO, G., IBÁÑEZ, A., CARREÑO, J., CABELLO, F., IBÁÑEZ, J. and MARTÍNEZ-ZAPATER, J.M.,

2006. Molecular genetics of berry colour variation in table grape. *Molecular Genetics and Genomics,* **276**(5), pp. 427-435.

MACFARLANE, C., 2004. Allelic variation of HERV-K (HML-2) endogenous retroviral elements in human populations. *Journal of Molecular Evolution,* **59**(5), pp. 642.

MACK, M., BENDER, K. and SCHNEIDER, P.M., 2004. Detection of retroviral antisense transcripts and promoter activity of the HERV-K (C4) insertion in the MHC class III region. *Immunogenetics,* **56**(5), pp. 321-332.

MANGENEY, M., RENARD, M., SCHLECHT-LOUF, G., BOUALLAGA, I., HEIDMANN, O., LETZELTER, C., RICHAUD, A., DUCOS, B. and HEIDMANN, T., 2007. Placental syncytins: Genetic disjunction between the fusogenic and immunosuppressive activity of retroviral envelope proteins. *Proceedings of the National Academy of Sciences,* **104**(51), pp. 20534-20539.

MCCLINTOCK, B., 1953. Induction of instability at selected loci in maize. *Genetics,* **38**(6), pp. 579.

MEDSTRAND, P., 2002. Retroelement distributions in the human genome: variations associated with age and proximity to genes. *Genome research,* **12**(10), pp. 1483.

METZKER, M.L., 2010. Sequencing technologies-the next generation. *Nature reviews.Genetics,* **11**(1), pp. 31.

MILLER, J.R., 2010. Assembly algorithms for next-generation sequencing data. *Genomics,* **95**(6), pp. 315.

MOYES, D., 2007. Insertional polymorphisms: a new lease of life for endogenous retroviruses in human disease. *Trends in genetics,* **23**(7), pp. 326.

NELSON, P.N., CARNEGIE, P., MARTIN, J., EJTEHADI, H.D., HOOLEY, P., RODEN, D., ROWLAND-JONES, S., WARREN, P., ASTLEY, J. and MURRAY, P.G., 2003. Demystified... human endogenous retroviruses. *Molecular Pathology,* **56**(1), pp. 11-18.

PANARO, M.A., CALVELLO, R., LISI, S., SACCIA, M., MITOLO, C.I. and CIANCIULLI, A., 2009. Viral sequence integration into introns of chemokine receptor genes. *Immunopharmacology and immunotoxicology,* **31**(4), pp. 589-594.

PONFERRADA, V., MAUCK, B. and WOOLEY, D.P., 2003. The envelope glycoprotein of human endogenous retrovirus HERV-W induces cellular resistance to spleen necrosis virus. *Archives of Virology,* **148**(4), pp. 659-675.

SANGER, F., 1988. Sequences, Sequences, and Sequences. *Annual Review of Biochemistry,* **57**(1), pp. 1-29.

SCHADT, E.E., 2010. A Window into Third Generation Sequencing. *Human molecular genetics,* **19**(r2), pp. R227.

SCHOLZ, M.B., LO, C. and CHAIN, P.S., 2012. Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. *Current opinion in biotechnology,* **23**(1), pp. 9-15.

SHENDURE, J., 2008. Next-generation DNA sequencing. *Nature biotechnology,* **26**(10), pp. 1135.

SICAT, J., SUTKOWSKI, N. and HUBER, B.T., 2005. Expression of human endogenous retrovirus HERV-K18 superantigen is elevated in juvenile rheumatoid arthritis. *The Journal of rheumatology,* **32**(9), pp. 1821-1831.

SNYDER, M., 2010. Personal genome sequencing: current approaches and challenges. *Genes development,* **24**(5), pp. 423.

STEWART, C., KURAL, D., STRÖMBERG, M.P., WALKER, J.A., KONKEL, M.K., STÜTZ, A.M., URBAN, A.E., GRUBERT, F., LAM, H.Y. and LEE, W., 2011. A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS genetics,* **7**(8), pp. e1002236.

STOYE, J.P., 2001. Endogenous retroviruses: Still active after all these years? *Current biology,* **11**(22), pp. R914.

SUBRAMANIAN, R.P., WILDSCHUTTE, J.H., RUSSO, C. and COFFIN, J.M., 2011. Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. *Retrovirology,* **8**(90doi), pp. 4690-4698.

SVERDLOV, E.D., 2000. Retroviruses and primate evolution. *BioEssays,* **22**(2), pp. 161.

TAKAHASHI, Y., HARASHIMA, N., KAJIGAYA, S., YOKOYAMA, H., CHERKASOVA, E., MCCOY, J.P., HANADA, K., MENA, O., KURLANDER, R. and ABDUL, T., 2008. Regression of human kidney cancer following allogeneic stem cell transplantation is associated with recognition of an HERV-E antigen by T cells. *The Journal of clinical investigation,* **118**(3), pp. 1099.

TARUSCIO, D., FLORIDIA, G., ZORAQI, G.K., MANTOVANI, A. and FALBO, V., 2002. Organization and integration sites in the human genome of endogenous retroviral sequences belonging to HERV-E family. *Mammalian genome,* **13**(4), pp. 216-222.

TENG, S.C., KIM, B. and GABRIEL, A., 1996. Retrotransposon reverse-transcriptase-mediated repair of chromosomal breaks. *Nature,* **383**(6601), pp. 641-644.

TRISTEM, M., 2000. Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database. *Journal of virology,* **74**(8), pp. 3715.

TURNER, G., 2001. Insertional polymorphisms of full-length endogenous retroviruses in humans. *Current biology,* **11**(19), pp. 1531.

VENTER, J.C., 2001. The sequence of the human genome. *Science,* **291**(5507), pp. 1304.

VITTE, C., 2003. Formation of solo-LTRs through unequal homologous recombination counterbalances amplifications of LTR retrotransposons in rice Oryza sativa L. *Molecular biology and evolution,* **20**(4), pp. 528.

WANG, J., SONG, L., GONDER, M.K., AZRAK, S., RAY, D.A., BATZER, M.A., TISHKOFF, S.A. and LIANG, P., 2006. Whole genome computational comparative genomics: A fruitful approach for ascertaining< i> Alu insertion polymorphisms. *Gene,* **365**, pp. 11-20.

WANG, T., ZENG, J., LOWE, C.B., SELLERS, R.G., SALAMA, S.R., YANG, M., BURGESS, S.M., BRACHMANN, R.K. and HAUSSLER, D., 2007. Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proceedings of the National Academy of Sciences,* **104**(47), pp. 18613-18618.

WHEELER, D.A., SRINIVASAN, M., EGHOLM, M., SHEN, Y., CHEN, L., MCGUIRE, A., HE, W., CHEN, Y., MAKHIJANI, V. and ROTH, G.T., 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature,* **452**(7189), pp. 872-876.

XING, J., WANG, H., BELANCIO, V.P., CORDAUX, R., DEININGER, P.L. and BATZER, M.A., 2006. Emergence of primate genes by retrotransposon-mediated sequence transduction. *Proceedings of the National Academy of Sciences,* **103**(47), pp. 17608-17613.

ZIOGAS, D., 2009. Genetics and personal genomics for personalized breast cancer surgery: progress and challenges in research and clinical practice. *Annals of Surgical Oncology,* **16**(7), pp. 1771.

Appendix 1 – TIPs_IN Positive Results

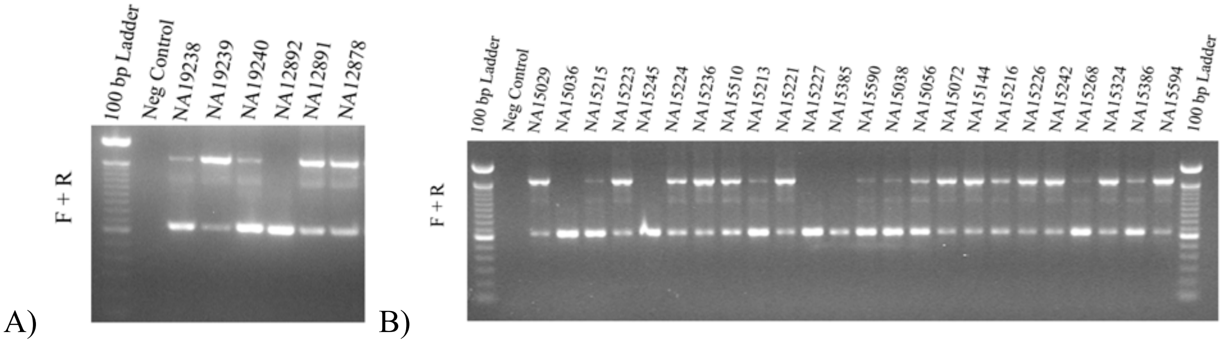

Figure 14: PCR genotype results for the F+R primers on candidate chr4:120483136-120484102.

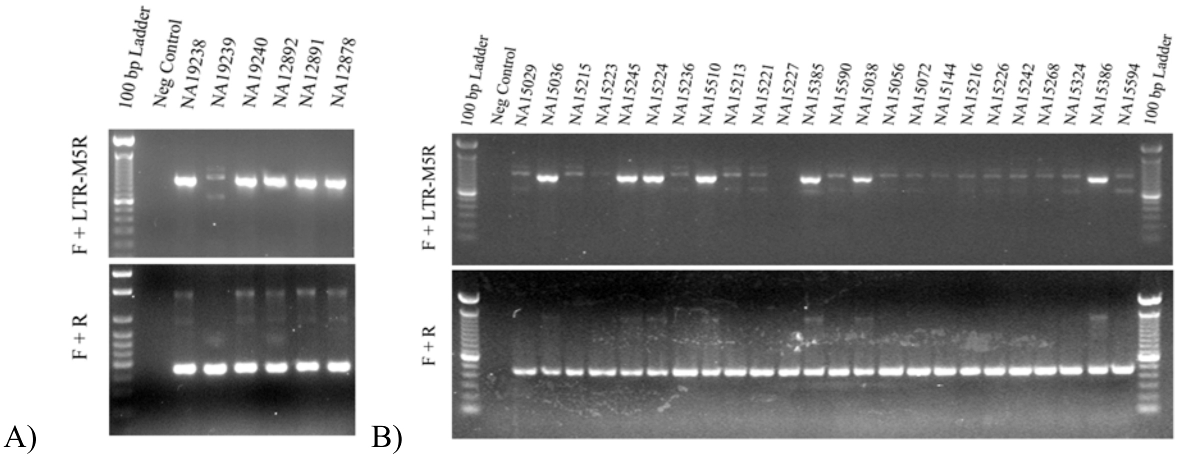

Figure 15: PCR genotype results for the F+R and F+LTR-M5R primers on candidate chr7:157722243-157723211.

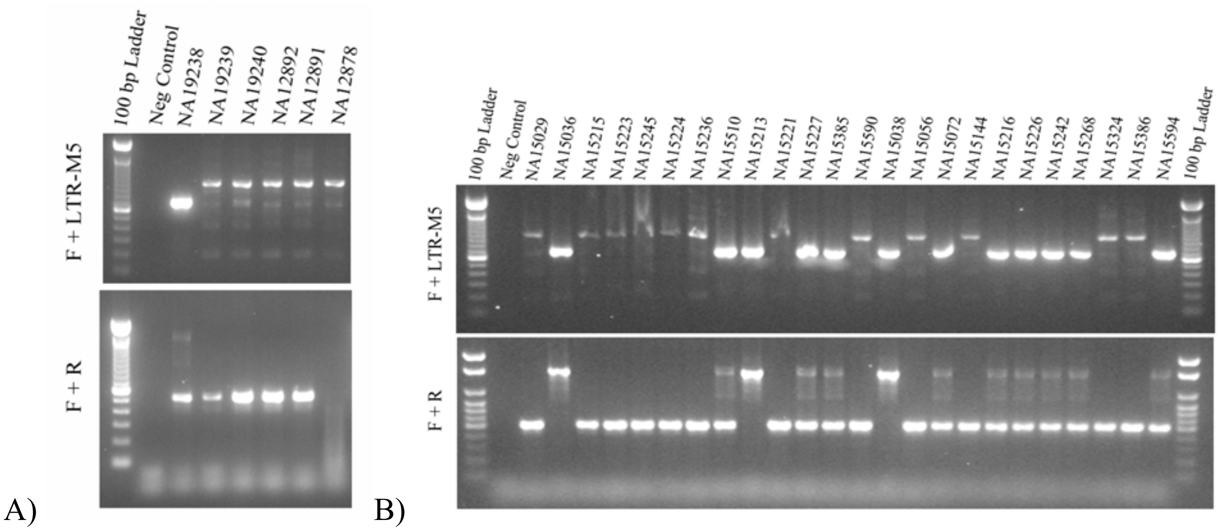

Figure 16: PCR genotype results for the F+R and F+LTR-M5R primers on candidate chr8:18695738-18696706.

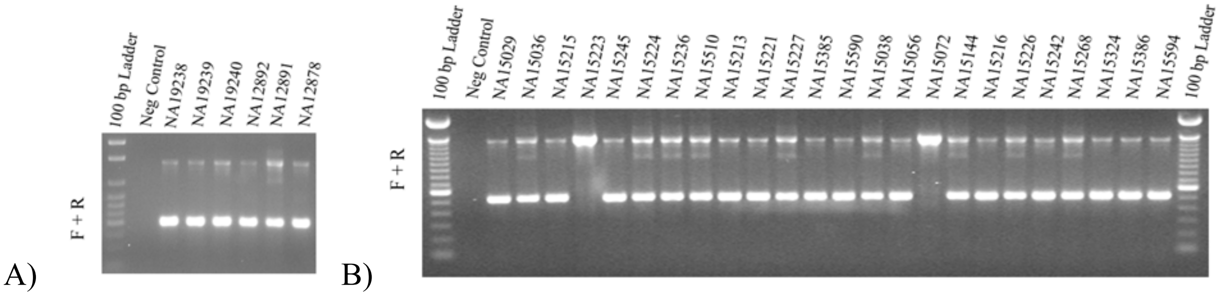

Figure 17: PCR genotype results for the F+R primers on candidate chr8:37170043-37171011.

Appendix 2 – FL TIPs_IN Positive Results

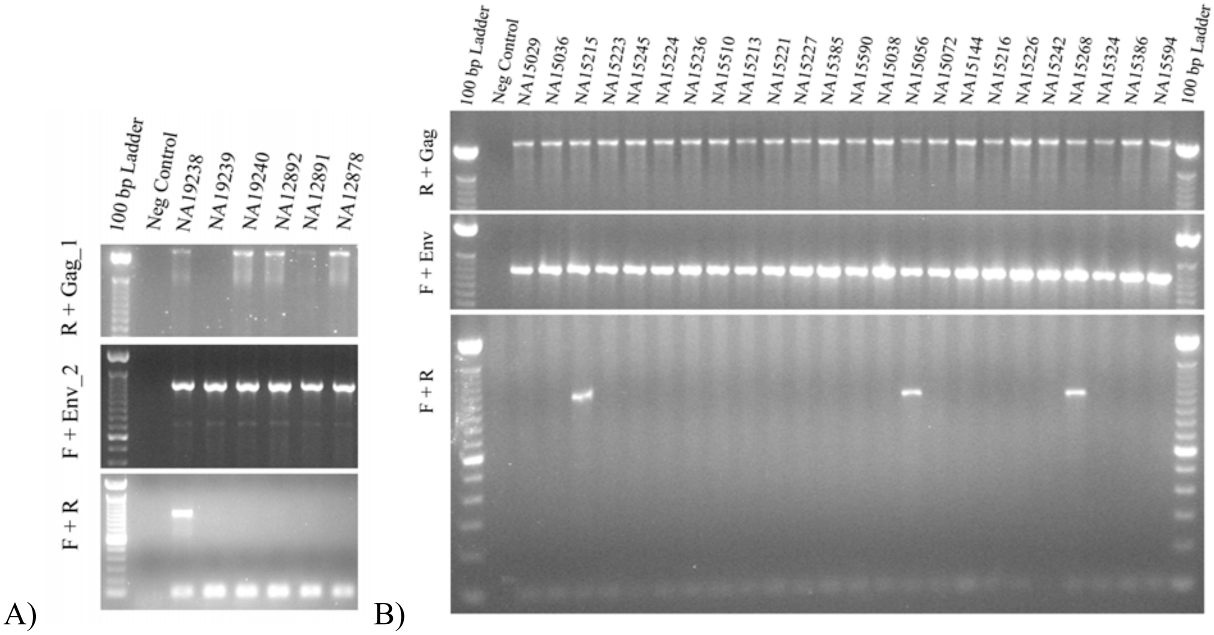

Figure 18: PCR genotype results for the F+R, F+ENV_2 and R+GAG_1 primers on candidate chr3:185281305-185288547.

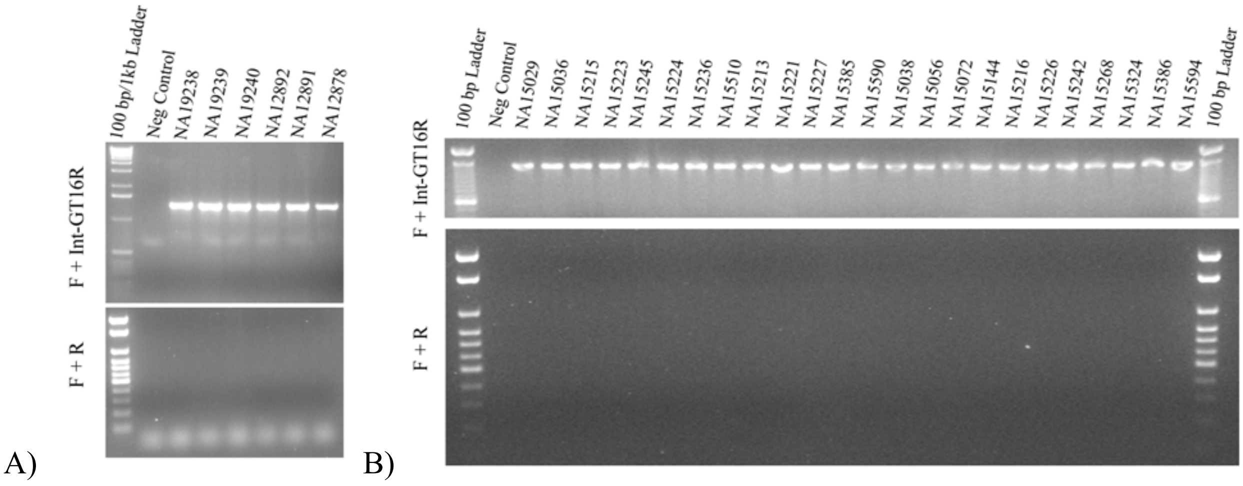

Figure 19: PCR genotype results for the F+R, F+Int-GT16R primers on candidate chr3:101411706-101418889.

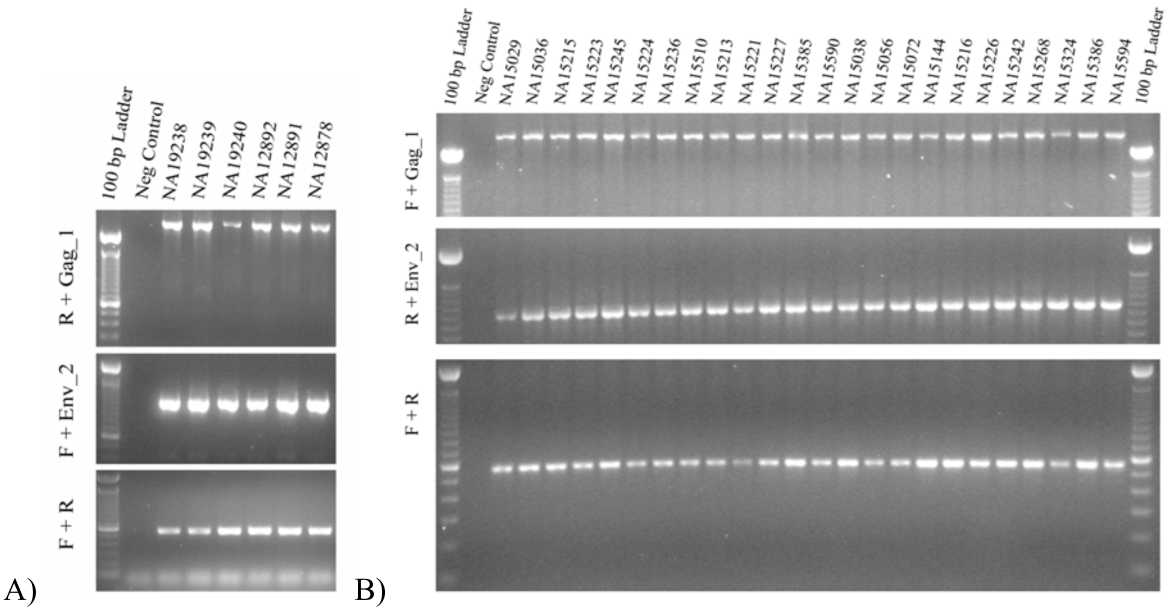

Figure 20: PCR genotype results for the F+R, F+GAG_1 and R+ENV_2 primers on candidate chr3:125610106-125617634.

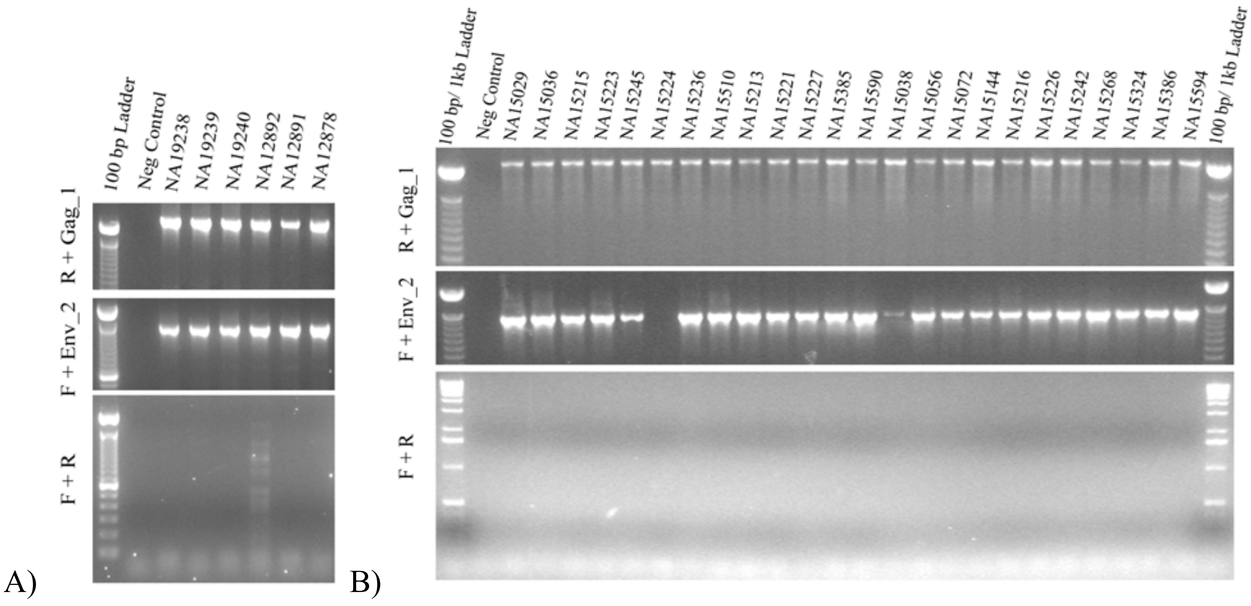

Figure 21: PCR genotype results for the combination of the F+R, F+ ENV_2 and R+GAG_1 primers on candidate chr10:6867110-6874635.

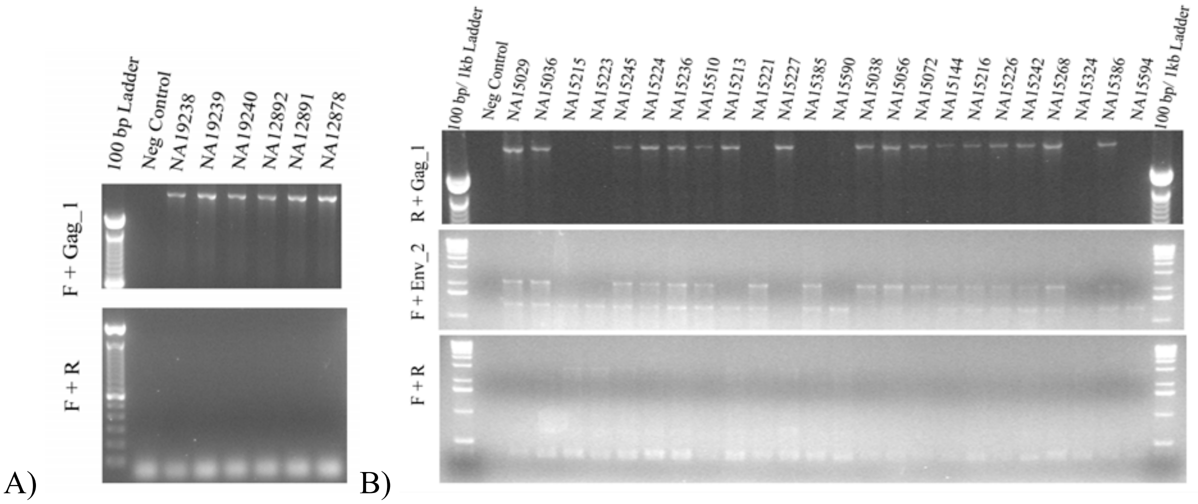

Figure 22: PCR genotype results for the combination of the F+R, F+ENV_2 and R+GAG_1 primers on candidate chr12:58722211-58729730.

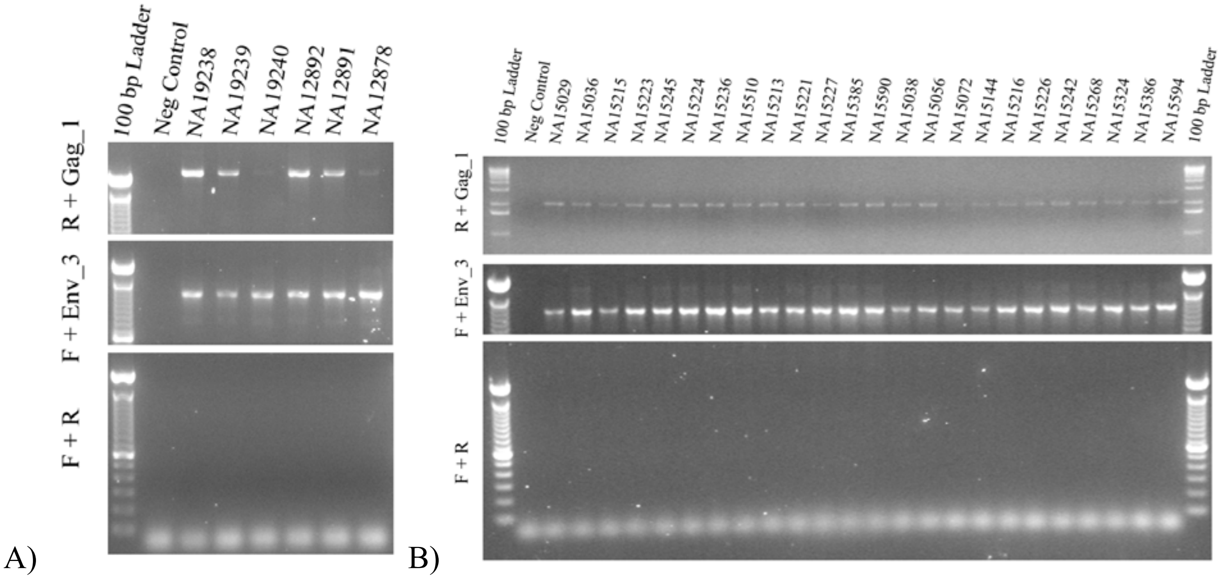

Figure 23: PCR genotype results for the combination of the F+R, F+ENV_3 and R+GAG_1 and primers on candidate chr21:19933917-19940998.

Appendix 3 – TIPs_OUT Positive Results

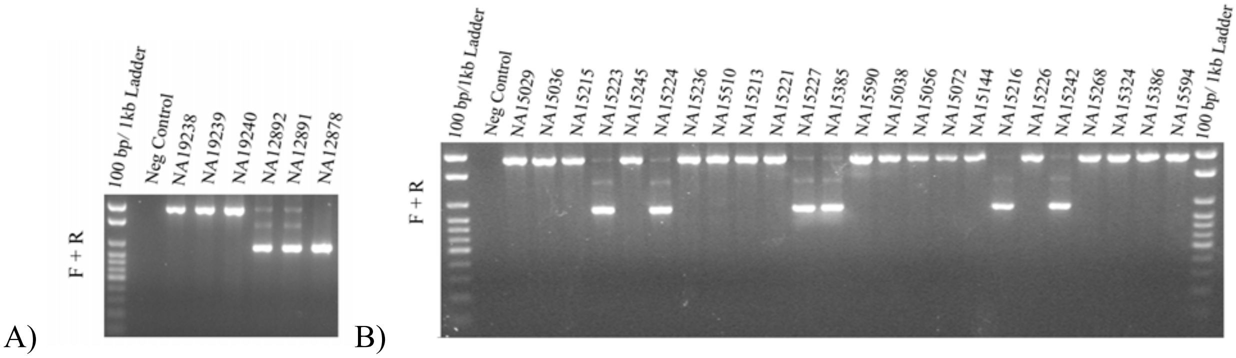

Figure 24: PCR genotype results for the combination of the F+R primers on candidate chr6:161190594-161191142.

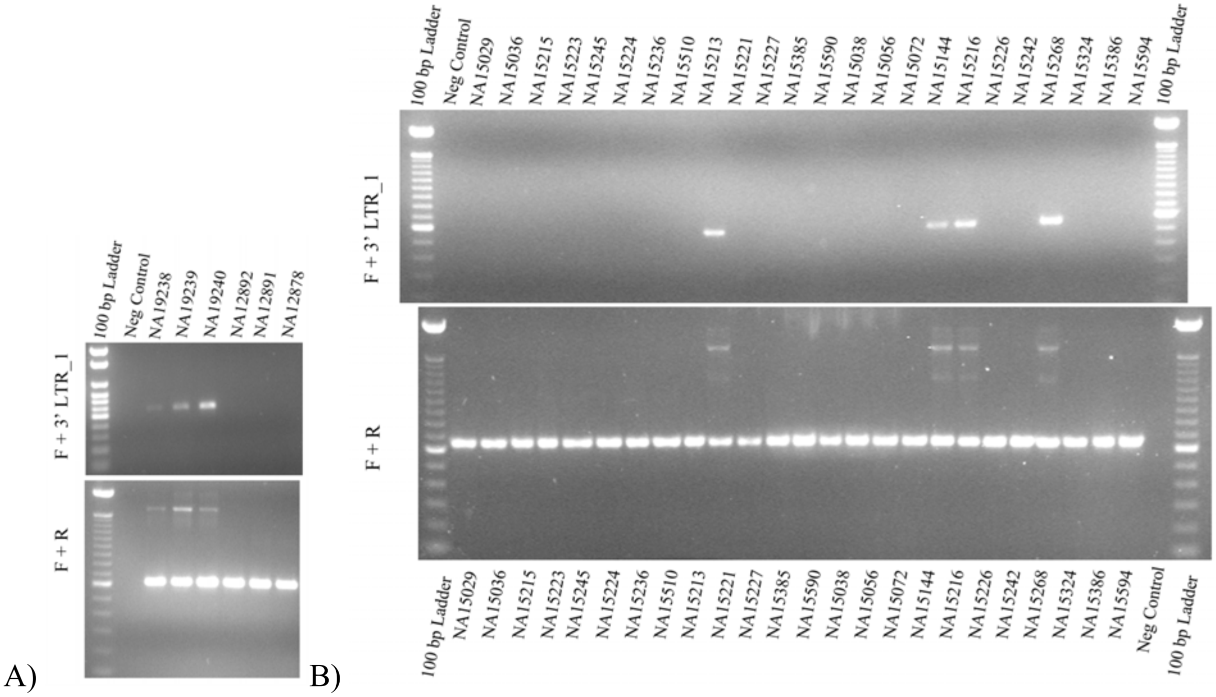

Figure 25: PCR genotype results for the combination of the F+R and F+3_LTR_1 primers on candidate chr13:89540897-89541416.

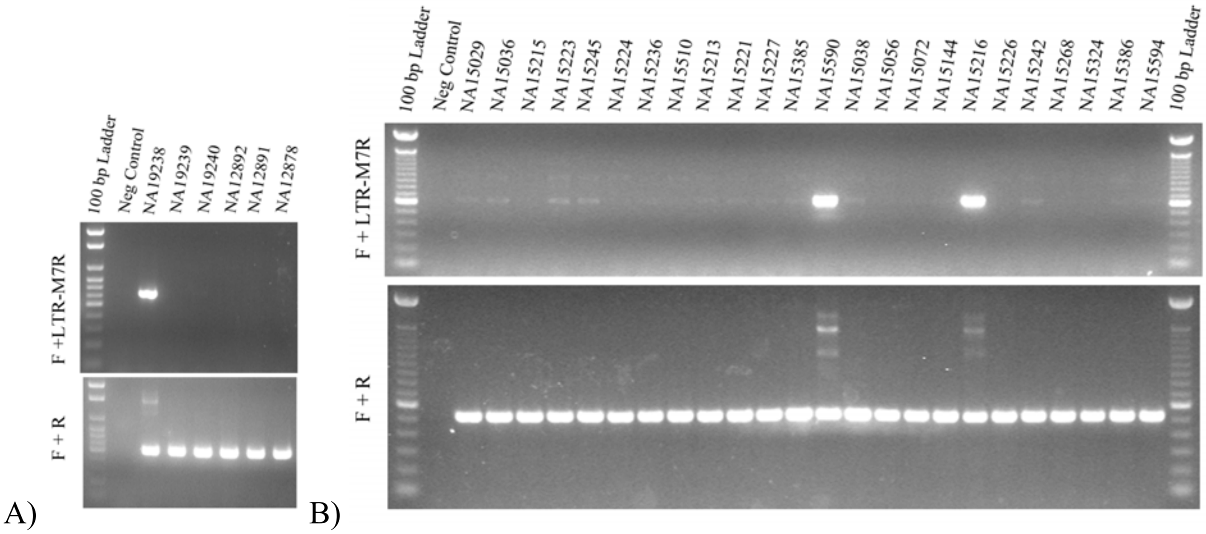

Figure 26: PCR genotype results for the combination of the F+R and F+LTR-M7R primers on candidate chr20:12350168-12350569.

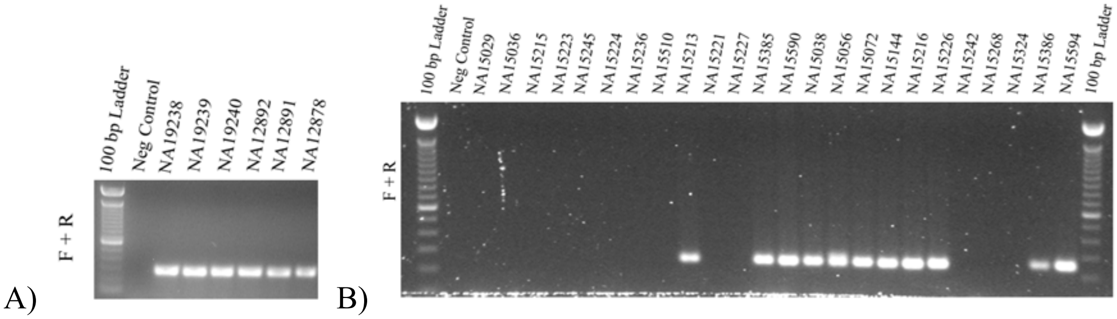

Figure 27: PCR genotype results for the combination of the F+R primers on candidate chr11:124879106-124879109.

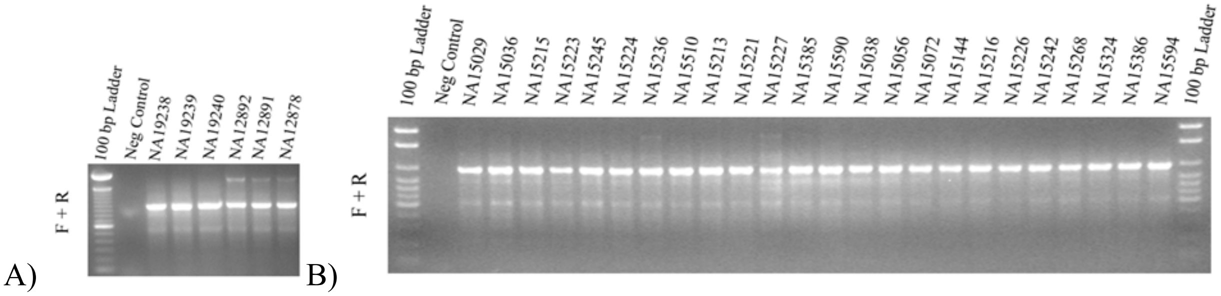

Figure 28: PCR genotype results for the combination of the F+R primers on candidate chr21:15966603-15966908.

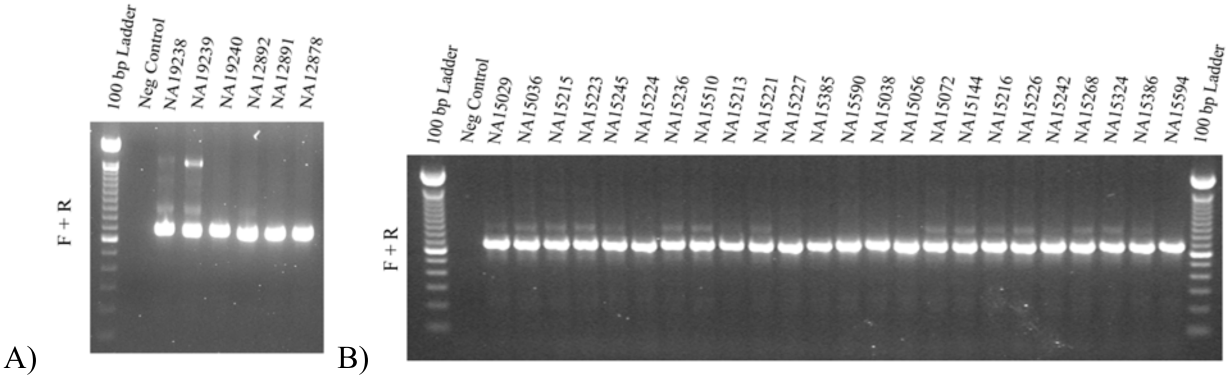

Figure 29: PCR genotype results for the combination of the F+R primers on candidate chr17:4959277-4959821.