Molecules and Cells

Novel Discovery of LINE-1 in a Korean Individual by a Target Enrichment Method

Wonseok Shin[1], Seyoung Mun[1], Junse Kim[1], Wooseok Lee[1], Dong-Guk Park[2], Seungkyu Choi[3], Tae Yoon Lee[4], Seunghee Cha[5], and Kyudong Han[1,*]

[1]Department of Nanobiomedical Science & BK21 PLUS NBM Global Research Center for Regenerative Medicine, Dankook University, Cheonan 31116, Korea, [2]Department of Surgery, [3]Department of Pathology, Dankook University College of Medicine, Cheonan 31116, Korea, [4]Department of Technology Education and Department of Biomedical Engineering, Chungnam National University, Daejeon 34134, Korea, [5]Department of Oral and Maxillofacial Diagnostic Sciences, University of Florida College of Dentistry, Gainesville, FL 32610, USA
*Correspondence: jim97@dankook.ac.kr
http://dx.doi.org/10.14348/molcells.2018.0351
www.molcells.org

Long interspersed element-1 (LINE-1 or L1) is an autonomous retrotransposon, which is capable of inserting into a new region of genome. Previous studies have reported that these elements lead to genomic variations and altered functions by affecting gene expression and genetic networks. Mounting evidence strongly indicates that genetic diseases or various cancers can occur as a result of retrotransposition events that involve L1s. Therefore, the development of methodologies to study the structural variations and interpersonal insertion polymorphisms by L1 element-associated changes in an individual genome is invaluable. In this study, we applied a systematic approach to identify human-specific L1s (i.e., L1Hs) through the bioinformatics analysis of high-throughput next-generation sequencing data. We identified 525 candidates that could be inferred to carry non-reference L1Hs in a Korean individual genome (KPGP9). Among them, we randomly selected 40 candidates and validated that approximately 92.5% of non-reference L1Hs were inserted into a KPGP9 genome. In addition, unlike conventional methods, our relatively simple and expedited approach was highly reproducible in confirming the L1 insertions. Taken together, our findings strongly support that the identification of non-reference L1Hs by our novel target enrichment method demonstrates its future application to genomic variation studies on the risk of cancer and genetic disorders.

Keywords: L1Hs, long interspersed elements-1, non-reference L1 screening, target enrichment system

INTRODUCTION

Almost half of the human genome is derived from transposable elements (TEs) that are divided into two classes, DNA transposons and retrotransposons (Cordaux and Batzer, 2009). Retrotransposons consist of long interspersed elements (LINEs), short interspersed elements (SINEs), and endogenous retroviruses (ERVs) (Cordaux and Batzer, 2009). Retrotransposons have the ability to generate genomic variations because they can be inserted into another genomic location through RNA intermediates. These intermediates are reverse transcribed and mobilize TEs (O'Donnell and Burns, 2010). Previous studies focusing on TEs provide important insights into understanding of human genome evolution and diversity (Ayarpadikannan and Kim, 2014; Beck et al., 2011; Park et al., 2015; Schrader and Schmitz, 2018; Sotero-Caio et al., 2017). Among retrotransposons, LINE-1s (L1s) constitute a common large family of retrotransposons composed of approximately 17% of the human genome, which have resided in the mammalian genomes for over 150 million years (Lutz et al., 2003; Seleme et al., 2006). Full-

Received 17 August 2018; revised 10 October 2018; accepted 26 October 2018; published online 6 December, 2018

eISSN: 0219-1032

length L1 is around 6 kb and consists of a 5' untranslated region (UTR), two open reading frames (ORF1 and ORF2) and a 3' UTR with a poly A terminal sequence. The ORF1 encodes a RNA-binding protein that functions as a nucleic acid chaperone. The ORF2 encodes endonuclease (EN) and reverse transcriptase (RT) that are crucial protein-complexes for self-mobilization (Lutz et al., 2003; Philippe et al., 2016; Seleme et al., 2006). The L1 is reverse transcribed and integrated into the genome through a mechanism known as target-primed reverse transcription (TPRT) (Philippe et al., 2016). During this process, the L1 endonuclease generates a single stranded nick in genomic DNA to expose a 3'-OH, which is used as a primer for reverse transcription of L1 RNA by the L1 reverse transcriptase (Brouha et al., 2003). The newly integrated L1 is flanked on both sides by identical direct repeats (7 to 20 bp) called "target-site duplications" (TSDs) flanking newly integrated elements (Fanning and Singer, 1987). These L1 proteins can also act in trans, to generate mobilization of noncoding retroelements such as *Alu*, SVA elements, and processed pseudogenes (Dewannieux et al., 2003). L1s exist in > 500,000 copies in the human genome and most of these copies are inactive due to accumulation of mutations. However, a few human-specific L1s (including L1Hs) are capable of retrotransposition, which contribute to our genetic diversity (Bennett et al., 2008; Philippe et al., 2016). It is estimated that 80 to 100 copies are able to retrotranspose themselves into the human genome, and that only a small number of highly active L1Hs are classified as "hot" L1s (Beck et al., 2010). These L1Hs can be subdivide into pre-Ta and Ta (for transcribed subset a) subfamilies (Myers et al., 2002), and Ta family has differentiated into Ta-0 and Ta-1, each of which branched additional subsets. In addition, Ta elements contain ACA/G in their 3' UTR as the shared sequences variants (SSV), whereas old elements generally have GAG/A at these positions (Ovchinnikov et al., 2002). These active retrotransposons are still capable of being transcribed into an RNA that is reverse-transcribed and integrated into another genomic region through TPRT (Boissinot et al., 2004). Thus, the newly identified L1 insertion polymorphisms are one of the important sources for studying human population genetics and understanding genomic diversity because they can be used as genetic biomarkers (Wang et al., 2017).

Recently, next-generation sequencing (NGS)-based applications such as whole-genome sequencing, transcriptome sequencing, exome sequencing, and microRNA profiling has the tremendous impact on genome research (Kenny et al., 2011; Valencia et al., 2012). Nevertheless, the identification of non-reference retrotransposons in specific region is not suitable by using these NGS techniques due to the limitations of resequencing. For example, when mapping on the human reference genome, resequencing the data of additional reads could be discarded by the bioinformatics algorithm. In addition, due to the repetitive sequence characteristics of the TEs included in the short reads, it is mapped to a paralogous region other than its original position (Ewing, 2015). Furthermore, the quality of sequencing reads is degraded due to the poly A tail sequence on 3' UTR of L1, and sequencing data cannot be used effectively. Therefore, it was challenging to detect the newly inserted TEs and its polymorphisms using the NGS method (Collier et al., 2005; Iskow et al., 2010) even though there are several methods based on TE-priming system, such as ATLAS and SIMPLE, reported that are currently used for identifying non-reference L1Hs insertions from human individuals (Badge et al., 2003; Boissinot et al., 2000; Konkel et al., 2007; Streva et al., 2015).

Here, we introduce an improved method of L1Hs target sequencing library construction and bioinformatics analysis for detecting non-reference L1Hs insertions in the human genome through the target enrichment system. Using our system, we discovered 525 non-reference L1Hs insertion candidates from a Korean individual (called KPGP9). Furthermore, to confirm the sensitivity and specificity of our method, we randomly selected 40 out of the non-reference L1Hs insertions and performed experimental validation by PCR amplification and Sanger sequencing with the KPGP9 sample. We propose that L1Hs target enrichment system could be useful to explore the dynamics of L1Hs retrotransposition in human populations.

MATERIALS AND METHODS

Probe design

To design a probe specific for non-reference L1Hs insertions, we compared L1Hs consensus sequence with other L1 subfamilies (L1PA2~L1PA17). Their consensus sequences were collected from the Genetic Information Research Institute (GIRI) repbase browser (http://www.girinst.org/repbase/update/search.php). To distinguish between L1Hs and other L1 families, we targeted to ACA at positions 5930-5932 of L1Hs family. Ultimately, the forty nucleotide long probe sequences (5'-AGGGATAGCATTGGGAGATATACCTAATGCTAGATGACAC-3') were designed to be specific for L1Hs.

Library construction

Fragmentation

Sample donors in this study signed a written informed consent to participate, and the Genome Research Foundation (IRB-20101202-001 for KPGP9) provided an approval for this study. The first stage in a standard genomic DNA library preparation is DNA fragmentation by sonication. The genomic DNA extracted from KPGP9 (a Korean individual) was used for this study. The genomic DNA (500 ng) was sheared using a Covaris S2 sonicator (Covaris, USA) to achieve typical size range of 400 to 700 bp and a target peak around 550 bp (settings: Duty Cycle, 10%; Intensity, 2; Peak Incident Power, 175; cycles per burst, 200; DNA treatment time, 45 s; water bath temperature, 4℃). The fragmented DNAs were quantified by Colibri Microvolume Spectrometer (Titertek-Berthold, Germany) and approximately 500 ng of the sample was run on a 2% TAE agarose gel along with the 1 KB Plus DNA Ladder (BioFACT, Korea) to verify the average fragment size of 500 bp.

DNA end repair and adaptor ligation

The Ovation Target Enrichment System kit (NuGen, USA) was used to construct Illumina libraries with the fragmented

DNAs for NGS on Illumina sequencing by synthesis platform. In brief, the following steps of DNA end repair, Illumina sequencing-based adaptor ligation and purification of the ligated gDNA were performed as described in the Ovation Target Enrichment System kit protocol. To repair both ends of fragmented DNAs, the End repair Master Mix (NuGen) was added to the fragmented DNA solution and incubated at 25℃ for 30 min, followed by incubation at 70℃ for 10 min. Next, the Ligation Adapter Master Mix (Nugen, USA) was added to end-repaired samples in order to ligate an adaptor and this mixture was incubated at 25℃ for 30 min, followed by 70℃ for 10 min. Immediately, adaptor dimers and unused reagents in the reaction were removed from the enriched libraries by using 0.8 volume per sample of the Agencourt RNAClean XP Beads (Beckman Coulter, USA). After ligation purification, the quality of each library was assessed by using a Bioanalyzer high Sensitivity DNA chip (Agilent Technologies, USA) (Supplementary Fig. S1).

Target enrichment

The probe hybridization and extension step were performed by a thermal cycler (Bio-Rad Laboratories, USA) with the following conditions: 95℃ for 5 min; 200 cycles of 80℃ for 10 s, decrease 0.1℃ each cycle; 60℃ for 16 h. After the hybridization, we immediately mixed the Extension Enzyme (Nugen) into the sample for hybridization-based target elongation with the following conditions on the thermocycler: elongation at 72℃ for 10 min and stabilization at 4℃. The increased DNA from target sequence purification was immediately carried out following the manufacturer's instructions. After the hybridization-based target elongation, this library pool was amplified using the Library Amplification Master Mix (NuGen) with following program: one cycle of 37℃ for 10 min; an initial denaturation step of 3 min at 95℃, followed by 15 cycles of PCR at 30 s of denaturation at 95℃, 15 s at the annealing temperature 62℃, and 20 s of extension at 72℃, followed by a final extension step of 3 min at 72℃, followed by a hold at 4℃. The amplified product was then cleaned using Agencourt RNAClean XP beads following the manufacturer's protocol (NuGen). A library of genomic DNA fragments amplified using the Illumina's primer set is called a target-enrichment library. The library had been sequenced by a NGS equipment of Illumina platform.

Next-generation sequencing

Cluster generation and sequencing were carried out on the amplified samples using Illumina HiSeq 2500 System (Illumina, USA) in 100 bp paired-end format according to the Illumina Paired-End Sequencing Platform Library protocol. The L1Hs-targeted sequencing was performed on a HiSeq equipment at the NGS and analysis facility, Theragen Etex Bio Institute (http://www.theragenetex.com/bio/).

Data analyses

HiSeq sequencing short reads were analyzed by custom pipelines to preprocess the raw data, which were contained with trimming the adaptor sequences, quality control in base calling, and verification of read mapping scores. The statistics analysis workflow was as follows.

Quality control and statistics processing

Prior to trimming sequencing reads, the reads were tag-sorted and quality control using the FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc), which is known as quality control application for sequencing data. In addition, various statistics such as total reads, total sequences, GC content, and sequence duplication levels from raw data were obtained by using the FastQC program.

Trimming and alignment

Prior to alignment between paired-end reads and reference genome sequences, the adaptor sequences at the 5' end of reads, poly A tail, ambiguous long N bases, and low complexity sequences were trimmed and removed using Cutadapt 1.1 version (https://cutadapt.readthedocs.io/en/stable). In addition, some reads that included the L1Hs sequences with the probe sequence from adaptor-trimmed reads were removed. As a result, reads having only the flanking sequences of L1Hs were obtained. In the fasta format file processed by the trimming step, this read sequences were aligned to the human reference genome (GRCh38/hg38; December 2013 freeze) with Bowtie2 (http://bowtie-bio.sourceforge.net) aligner. Furthermore, PCR-duplicated reads were removed from generated BAM alignment files after mapping using the Picard's MarkDuplicates (http://broadinstitute.github.io/picard) tool to eliminate potential PCR bias. Our peak-calling analysis is similar to the ChIP-seq analysis. In the ChIP-seq analysis, the peak-calling finds statistically enrichment compared to noise or background. Likewise, we applied the peak-calling strategy to find L1Hs-target enrichment regions in the mapped files (BAM alignment files). Peaks were annotated using both Hypergeometric Optimization of Motif Enrichment program (HOMER) v4.10 (http://homer.ucsd.edu/homer) and Model-based Analysis of ChIP-Seq (MACS) v2.1 (https://pypi.python. org/pypi/MACS2). These programs for peak-calling were performed by the parameter of default values and each calling algorithm. To determine the relative overlap between peaks in HOMER and MACS, we merged the two sets of peak calling outputs. Human genome reference sequences and their repeat annotation were obtained from the UCSC genome data (http://hgdownload.soe.ucsc.edu/downloads. html).

Bioinformatics analysis

The IGV program (https://software.broadinstitute.org/software/igv/home) was utilized to examine if the non-reference L1Hs candidates extracted by the computational approach are different from the existing L1 position. To examine the GC content of the flanking sequences of the non-reference L1Hs insertions, we extracted 20 kb of flanking sequence upstream and downstream of each insertion site using the UCSC Genome Browser (https://genome.ucsc.edu/index.html). The percentage of GC nucleotides in the flanking sequence was then calculated using the Bedtools v2.27.0 (http://bedtools.readthedocs.io/en/latest/) with regions of non-reference L1Hs insertion. For the gene density analysis, we counted the number of genes within a 2 Mb window of flanking sequence centered on each non-reference L1Hs using the National Center for Biotechnology Information

Map Viewer utility (http://www.ncbi.nlm.nih.gov/projects/mapview/map_search.cgi?taxid=9606).

To investigate the genomic location of target enrichment regions, we manually inspected the non-reference L1Hs candidate loci and annotated genes related to their insert sites using the UCSC Genome Browser.

Validation of non-reference L1Hs insertions

Primer design

The UCSC genome browser gateway (http://genome-asia.ucsc.edu/cgi-bin/hgGateway) was used to acquire the flanking sequence of target regions both upstream and downstream in the human reference genome. Oligonucleotide primer pairs for the PCR amplification of each locus were designed using the Primer3 (http://bioinfo.ut.ee/pri-mer3-0.4.0/primer3) and Oligo Analysis Tools (http://www.oper-on.com/tools/oligo-analysis-tool.aspx) programs. In addition, *in silico* PCR was conducted for each locus to estimate the expected PCR product size and the optimal annealing temperature, and to further identify that primer pairs only amplified a single locus.

PCR amplification and DNA sequence analysis

PCR was performed using two different human genomic DNA samples (KPGP9 and NA10851 (Coriell Cell Repository, USA)) as templates. PCR amplification of each locus was performed in 20 ul reactions containing 20 ng of template DNA, 10 ul of 2X EF-Taq Premix4 (BioFACT), 10 nM of each oligo nucleotide primers, and nuclease-free water. Each PCR was subjected to initial denaturation step of 5 min at 95℃, followed by 35 cycles of 30 s at 95℃, 40 s of annealing at optimal annealing temperature, and a long extension step at 68℃ for 7 min, followed by a final extension step at 68℃ for 10 min. The PCR products were run on a 1% agarose gel electrophoresis with EcoDye (BioFACT) and visualized using Gel Doc (Bio-Rad, Germany). The PCR product was purified by the PCR purification kit (FAVORGEN, Taiwan) and some products were cloned with the Dr. TA TOPO cloning kit (Doctor Protein, Korea) according to the manual instructions. Target clone were confirmed by colony PCR, following plasmid

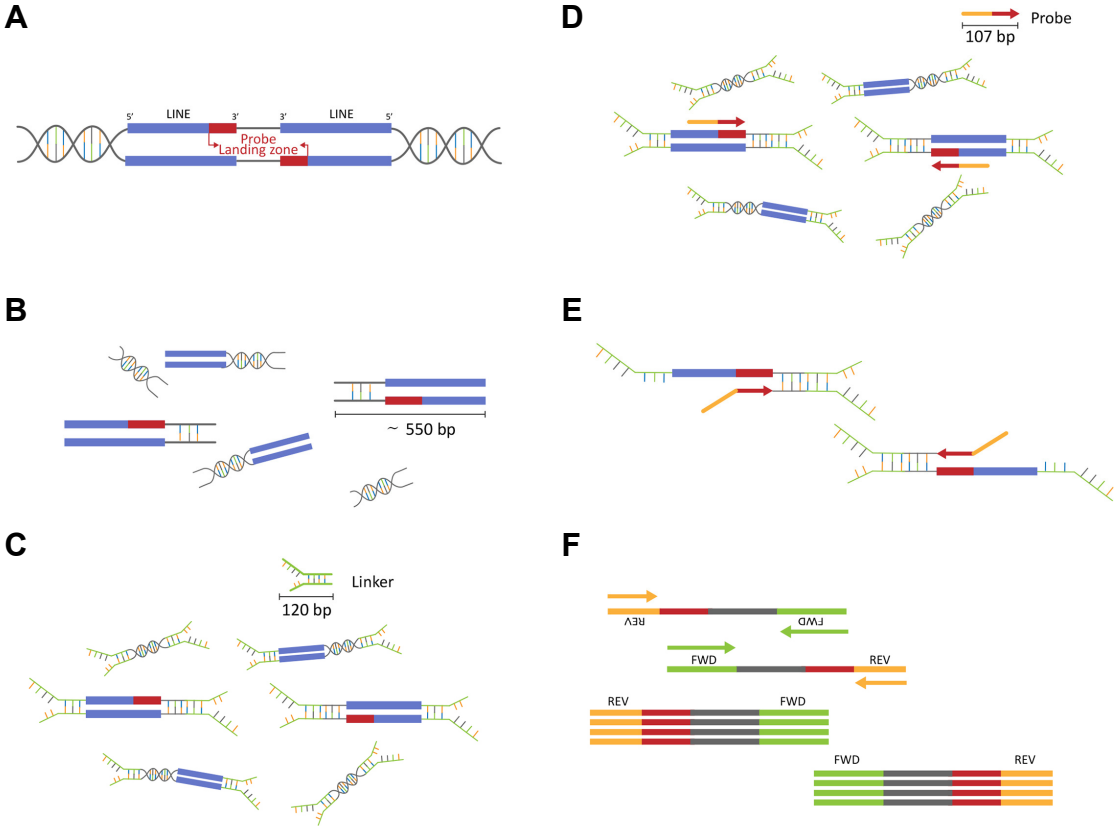

Fig. 1. The workflow of L1Hs-targeted enrichment library preparation. (A) Double-strand genomic DNA (blue) is extracted from a Korean individual genome (KPGP9). Red boxes indicate the regions where the probe binds to the 3' UTR of L1Hs elements. (B) Genomic DNA is fragmented by aquatic ultrasonic wave of the Covaris S2 system. Sheared DNAs have an average size of 550 bp, which is suitable for HiSeq sequencing. (C) The Illumina's adaptor (green) is ligated at both ends of the fragmented DNAs. (D) The adaptor-ligated DNAs is hybridized with the L1Hs-targeted probe (red and orange). Only the presence of the L1Hs 3' UTR allows the sequence-specific binding of the probe. (E) Targeted DNA fragments are selectively elongated from the probe-binding strands. (F) Because the probe sequence attached to the L1Hs 3' UTR and the Illumina's adaptor sequences at both ends are known, targeted DNAs are enriched by PCR with the primer set. After library construction, the final product is confirmed using the Agilent Bioanalyzer High Sensitivity chip assay.

preparation with the Exprep™ plasmid SV kit (GeneAll Biotechnology, Korea) and sequenced using chain-termination sequencing on an ABI 3500 Genetic Analyzer (Applied Biosystems, USA).

RESULTS AND DISCUSSION

Description of the L1Hs-targeted enrichment method

Previous studies on non-reference TEs have suggested that the amplification system was problematic because randomly fragmented DNA in different sizes was inserted in the library construction and the length of a producible read was insufficient to cover either the L1Hs-targeted region or the genomic region (Van den Broeck et al., 1998; Witherspoon et al., 2010). To overcome those issues, we used 100 bp paired-end sequencing using the Illumina platform and high-throughput targeted sequencing methods to identify L1Hs elements in the human genome. In addition, we added a fragment size selection step to selectively enrich the L1Hs-targeted reads. Our method is a developed system to detect L1Hs elements based on targeted high-throughput sequencing using a L1Hs target-specific probe (Ewing and Kazazian, 2010) and sample-specific indexing (Smith et al., 2009).

The workflow of the method for the L1Hs-targeted library is illustrated in Fig. 1 with additional explanations, and the detailed protocol is outlined in the Materials and Methods section. The extracted genomic DNAs from the KPGP9 (Korean individual) had been fragmented by using physical sonication with the Covaris system. During this time, genomic DNAs can be randomly generated to form either blunt- or sticky-end. To extract DNAs of suitable size (about 550 bp) for the Illumina HiSeq platform, we performed beads size selection on fragmented DNAs. AMPure XP Beads (Beckman Coulter, A63881) can select the desired size of DNAs based on its concentration. The selected DNAs in shapes were performed to repair for blunt ends, and adenosines was provided to each 3' ends for TA ligation. Adaptors comprised of 3' thymine oligonucleotide are compatible with the Illumina sequencing platform and ligated to the sheared DNAs.

As shown in Fig. 2, the L1Hs-specific targeted probe was designed in the specific region of L1Hs 3' UTR and this probe underwent ligation with the Illumina adaptor sequence. To select fragment DNAs that contain an L1Hs region, we perform *in situ* DNA-DNA hybridization with a target-probe that anneals to a specific-region found only in L1Hs elements and flanking genomic sequence. When the L1Hs-targeted probe is hybridized with the L1Hs sequence, the 3' end is extended to the opposite adaptor sequence. To enrich the target fragments containing L1Hs regions, PCR amplification was conducted with a primer pair, Illumina adaptors-ligated with targeted L1Hs-specific probe. Subsequently, final libraries included 3' terminus of L1Hs element and its unique flanking sequence. The enriched final library was subjected to qualitative analysis through a BioAnalyzer 2100 instrument (Agilent Technologies). In addition, qRT-PCR was performed using the Kapa Library Quantification Kit (KAPA Biosystems, KR0405) to conduct accurate quantitative analysis. The final libraries were sequenced by using Illumina HiSeq 2500 system though Rapid SBS Kit v2. In the HiSeq system, the library on the Flow Cell is amplified through the cluster generation. Sequencing starts and reads from the first sequencing primer (Read 1) at the adaptor sequence in the flanking sequence direction of L1Hs downstream. After the template switching step on the sequencing process, the second sequencing primer (Read 2) started and read from the probe region of L1Hs 3' UTR.

The obtained raw sequencing data could be classified by each sample through "Demultiplexing step", was generated by using the FastQC program. As shown in Fig. 3, there are bioinformatics analysis flowchart for the detection of non-reference L1Hs insertions. The Illumina's adaptor sequence with L1Hs-targed sequence is not the L1Hs-targed sequence and should be removed using the Cutadapt1.1 version bioinformatics tool. The sequence data from which the L1Hs-targeted probe sequences had been removed were mapped to the human reference genome using Bowtie aligner program. The overlapped target regions were selected as the L1Hs insertion loci using two programs, HOMER and

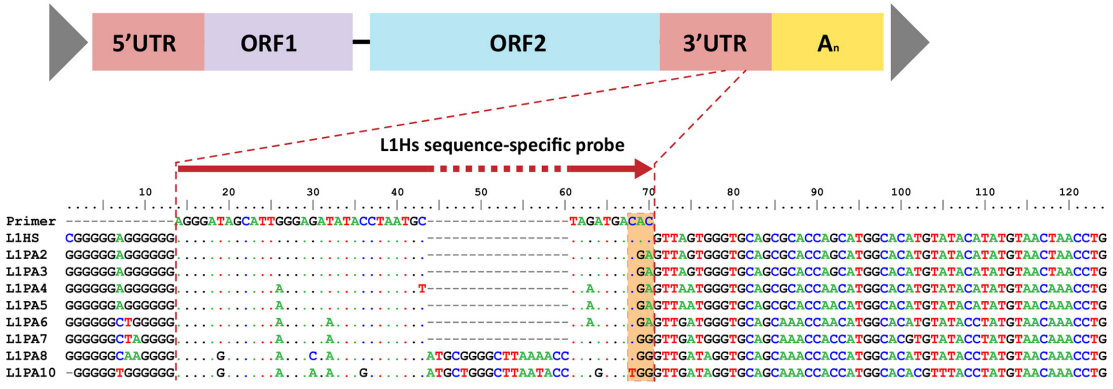

Fig. 2. The design of the probe specific for the L1Hs-target sequence. Using Clustal W Multiple alignment on BioEdit v.7.2.5, we aligned the L1 subfamilies (L1Hs and L1PA2 to L1PA10) based on their 3' UTR region (Thompson, J. D., 1994). To design a probe with high specificity, by using the human genome database and the Repeat Masker web-based tool, we collected more than 30 L1 sequences for each subfamily and designed the target-probe common to the L1Hs element, but in different sequence position for another subfamily.

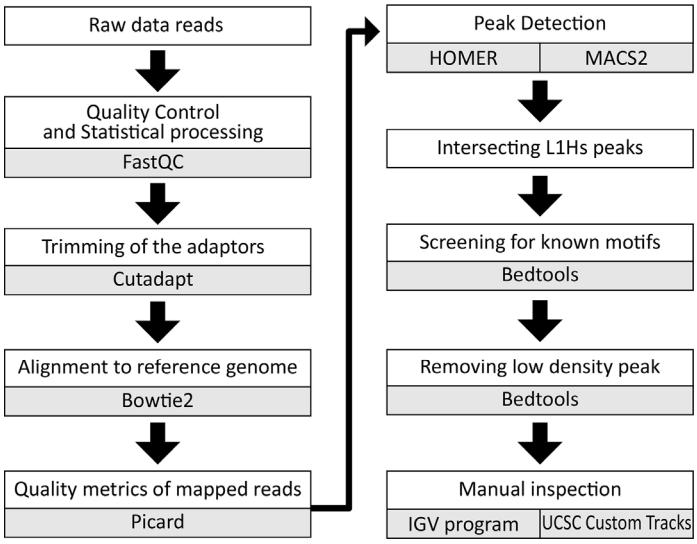

Fig. 3. NGS Data analysis. This schematic diagram describes the process of our computational approach. Raw data was obtained by paired-end sequencing on the Illumina HiSeq2500 system.

MACS2, which were developed to display the target region showing a significant degree of mapping within the aligned genomic sequence. There is no gold-standard for peak-calling step even though several developed approaches are used (Strino and Lappe, 2016). Because of the algorithm difference in the programs, false-positive results could be increased or the peaks that should actually be called is not selected. Nevertheless, it is known that the use of multiple replicate tools in the peak-calling programs has a high recall rate (Steinhauser et al., 2016).

Identification of non-reference L1Hs insertions

As for a sequencing result, we obtained a total of approximately 30 million reads with 3.05 Gb using paired-end sequencing on the HiSeq2500 system. Each manufactured sequencing read contained partial L1Hs-targeted sequence (Read 2) and its flanking sequence (Read 1), respectively. However, in the case of Read 2 from beginning of the L1Hs 3' UTR region, the poly A tail region is continuously recorded in the target library, and the quality score deteriorated from the signal detector of the HiSeq system. Furthermore, it was difficult to use because Read 2 containing poly A tail sequence was mapped to a number of paralogous region. Therefore, in bioinformatics analysis, the "Read 1" that is considered to contain the flanking sequence of L1Hs was used to identify the predicted region of non-reference L1Hs insertion (Table 1). The qualified reads after filtering out low-quality reads and trimming adaptors for Illumina platform were implicated in alignment to the human reference genome (Hg38). After mapping trimmed reads to human reference genome, we performed peak calling of mapped reads using two peak caller programs, HOMER and MACS, by which we find 2900 overlapped peak-callings (Lun and Smyth, 2016). Considering the normalized tag counts (number of tags found at the peak) and fold change (FC) value on the HOMER program, the low density peak and the annotated region (L1 subfamilies) were removed using the Bedtools command. The peak-callings with the same position as L1s present in the human reference genome and with small number of mapping reads (i.e, normalized tag counts < 4) were eliminated. As a result, a total of 525 non-reference L1Hs insertion candidates were selected (Table 2). In addition, we manually inspected the all loci by using the

Table 1. Summary of the high-throughput sequencing data

Classification	Paired-end	Read 1[†]
Total reads	30,228,074	15,114,037
Total bases	3.05 Gb	1.53 Gb
GC contents (%)	1,306,397,565 (42.79%)	647,427,902 (60.93%)
N zero reads (%)	30,127,532 (99.67%)	15,086,912 (99.82%)
Q30 bases	1,977,974,878 (64.79%)	1,268,378,122 (83.09%)

[†]Read1 was the sequence that we used for the computational analysis.

Table 2. Summary of non-reference L1Hs elements in the KPGP9 genome

Classification	No. of loci
L1Hs-targeted sequencing result	2,900
Excepted reference L1 subfamilies, segmental duplication regions, and duplicated peak calling	2,375
Computationally predicted non-reference L1Hs	525
Intergenic regions	261
Intronic regions	247(40)[†]
Exonic regions	17(1)[†]
Validation regions	40

[†]The number in parentheses indicates the number of predicted genes.

IGV program and the Custom Track with these positions on the UCSC browser (Supplementary Table S1).

Based on our data analysis, although we have found 525 non-reference L1Hs regions with 1.53 Gb data of single read, we expect to find more accurate and more L1Hs if data production is increased. However, as data production increases, the frequency of false positive is expected to increase. On the other hand, because our probe sequence is located at the 3' UTR of L1Hs element, we could identify non-reference L1Hs integrated by a typical TPRT mechanism rather than targeting the 5' UTR. Thus, this method is one way to find non-reference L1Hs elements that are mostly truncated in the 5' end.

Characterization of non-reference L1Hs insertions

Through the manual inspection of 525 non-reference L1Hs insertions, we examined chromosomal distribution and genomic contribution. Chromosome 4 included the higher number of L1Hs insertions while we have not detected non-reference L1Hs insertions at chromosome 21 (Supplementary Fig. S2). When compared to the L1 distribution of the human genome, the non-reference L1Hs elements are distributed according to the chromosome size and corresponds to the results of previous study (Sellis et al., 2007). Among the total of 525 L1Hs insertions, 261 L1Hs elements were located in intergenic regions and 247 L1Hs elements were located in the intronic regions of 215 genes. Interestingly, 17 out of 525 L1Hs insertions were detected at exonic regions (5 L1Hs in 3' UTR, 11 L1Hs in coding exonic regions, and 1 L1Hs in small nucleolar RNA host gene; Table 2). Interestingly, three of the L1Hs elements in the coding exonic regions were associated with the zinc finger protein. Moreover, we investigated the overall non-reference L1Hs regions inserted into the intron and exon related regions of genes except for the regions inserted into the predicted genes (Supplementary Table S2). The composition of L1 sequence of 223 genes, excluding 41 regions of predicted genes, was examined in 264 intronic and exonic regions. As shown in Fig. 4, the average L1 composition of the non-reference L1Hs-inserted genes is 12.8%, which is higher than that of the human genes (7.3%). The typical L1 insertion donates an endonuclease cleavage site to the genome and is known to have a chance of inducing a novel L1 insertion because its target site is duplicated (Tripathi et al., 2015). Furthermore, we investigated the gene density of the genomic regions flanking each non-reference L1Hs element by extracting 2 Mb of flanking genomics sequences (±1 Mb in either direction), and counting the number of known or predicted human RefSeq genes (Supplementary Table S1). The gene density of non-reference L1Hs regions averaged about 18.4 genes per Mb, which is substantially higher than the about 10 genes per Mb average reported for the human genome (Lander et al., 2001). In addition, we investigated the GC content of all non-reference L1Hs regions. The GC content was calculated for the 20 kb of flanking genomic sequence on each side of each locus. The GC content of these flanking regions averaged 41.15%. This is similar to the human reference genomic average GC content of 41% (Lander et al., 2001). Our result indicates that the non-reference L1Hs is randomly integrated regardless of the GC content of the target region. In the previous study, it was reported that there is not a significant amount of the GC content in the newly inserted L1 region (Ovchinnikov et al., 2001). Thus, our finding is consistent with this assertion.

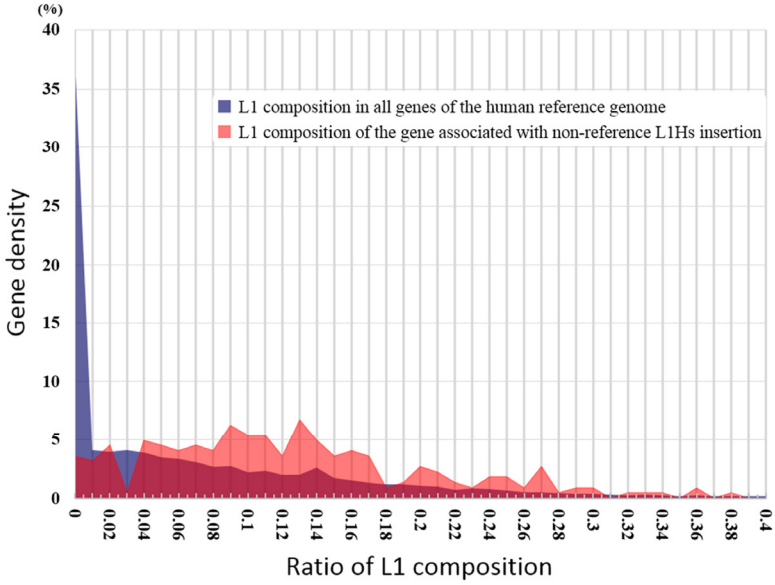

Fig. 4. Comparison of the L1 composition on the genes of non-reference L1Hs insertion and on the human genes. The blue-filled distribution is the L1 composition in all genes of the human reference genome. The red-filled distribution is the L1 composition of the gene associated with non-reference L1Hs insertion. The numbers on the Y axis and the X axis indicate the percentage of genes and the ratio of L1 composition, respectively.

Validation of non-reference L1Hs insertions and identification of false positives

Among the 525 non-reference L1Hs insertion candidates, 40 L1Hs insertion loci were randomly selected and subjected to PCR verification (Supplementary Table S3). The PCR result showed that 37 out of the 40 insertions (92.5%) were authentic (Supplementary Table S4). The three false positive could be a result of either from computational algorithm bias or from non-specific binding of probes during sequencing library construction. We also examined the orientation and homozygous/heterozygous genotype of the L1Hs insertions in the Korean genome. Ten out of the 37 authentic non-reference L1Hs insertions located in the sense orientation while the rest resided in the antisense orientation. In addition, we found that 14 insertions are homozygous and 23 insertions are heterozygous at the respective locus, indicating that approximately 62% of the L1Hs insertions are heterozygous in the Korean genome. We examined if any of the heterozygous insertion is a full-length element (> 6 kb), which means that it could still retain the ability to retrotranspose in the genome. Our result showed that two insertions are full-length.

CONCLUSION

Most of the L1s in the human genome are inactive. However, approximately 100 L1 copies still remain active in the human genome. By retrotransposition, they could influence on changing genomic structure, amplifying and/or disrupting gene expression, and leading genomic variations. Here, we aimed to identify non-reference L1Hs that still are retrotranspositionally competent in the human genome. Thus, we introduced the new L1 targeted-enrichment method based on 3' UTR sequence to discover the typical L1Hs insertion and screened the non-reference L1Hs insertion loci from a Korean individual (KPGP9) by using Illumina platform. In conclusion, we propose that this fast and cost-effective method for L1Hs screening is useful to further investigate somatic mutations induced by TEs in cancer and genetic diseases.

Note: Supplementary information is available on the Molecules and Cells website (www.molcells.org).

ACKNOWLEDGMENTS

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2017R1D1A1A02019421).

REFERENCES

Ayarpadikannan, S., and Kim, H.S. (2014). The impact of transposable elements in genome evolution and genetic instability and their implications in various diseases. Genomics Inform. *12*, 98-104.

Badge, R.M., Alisch, R.S., and Moran, J.V. (2003). ATLAS: a system to selectively identify human-specific L1 insertions. Am. J. Hum. Genet. *72*, 823-838.

Beck, C.R., Collier, P., Macfarlane, C., Malig, M., Kidd, J.M., Eichler, E.E., Badge, R.M., and Moran, J.V. (2010). LINE-1 retrotransposition activity in human genomes. Cell *141*, 1159-1170.

Beck, C.R., Garcia-Perez, J.L., Badge, R.M., and Moran, J.V. (2011). LINE-1 elements in structural variation and disease. Annu. Rev. Genomics. Hum. Genet. *12*, 187-215.

Bennett, E.A., Keller, H., Mills, R.E., Schmidt, S., Moran, J.V., Weichenrieder, O., and Devine, S.E. (2008). Active Alu retrotransposons in the human genome. Genome. Res. *18*, 1875-1883.

Boissinot, S., Chevret, P., and Furano, A.V. (2000). L1 (LINE-1) retrotransposon evolution and amplification in recent human history. Mol. Biol. Evol. *17*, 915-928.

Boissinot, S., Entezam, A., Young, L., Munson, P.J., and Furano, A.V. (2004). The insertional history of an active family of L1 retrotransposons in humans. Genome Res. *14*, 1221-1231.

Brouha, B., Schustak, J., Badge, R.M., Lutz-Prigge, S., Farley, A.H., Moran, J.V., and Kazazian, H.H., Jr. (2003). Hot L1s account for the bulk of retrotransposition in the human population. Proc. Natl. Acad. Sci. USA *100*, 5280-5285.

Collier, L.S., Carlson, C.M., Ravimohan, S., Dupuy, A.J., and Largaespada, D.A. (2005). Cancer gene discovery in solid tumours using transposon-based somatic mutagenesis in the mouse. Nature *436*, 272-276.

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

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

Ewing, A.D. (2015). Transposable element detection from whole genome sequence data. Mob. DNA *6*, 24.

Ewing, A.D., and Kazazian, H.H., Jr. (2010). High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. Genome Res. *20*, 1262-1270.

Fanning, T., and Singer, M. (1987). The LINE-1 DNA sequences in four mammalian orders predict proteins that conserve homologies to retrovirus proteins. Nucleic Acids Res. *15*, 2251-2260.

Iskow, R.C., McCabe, M.T., Mills, R.E., Torene, S., Pittard, W.S., Neuwald, A.F., Van Meir, E.G., Vertino, P.M., and Devine, S.E. (2010). Natural mutagenesis of human genomes by endogenous retrotransposons. Cell *141*, 1253-1261.

Kenny, E.M., Cormican, P., Gilks, W.P., Gates, A.S., O'Dushlaine, C.T., Pinto, C., Corvin, A.P., Gill, M., and Morris, D.W. (2011). Multiplex target enrichment using DNA indexing for ultra-high throughput SNP detection. DNA Res. *18*, 31-38.

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

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Lun, A.T., and Smyth, G.K. (2016). csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. Nucleic Acids Res. *44*, e45.

Lutz, S.M., Vincent, B.J., Kazazian, H.H., Jr., Batzer, M.A., and Moran, J.V. (2003). Allelic heterogeneity in LINE-1 retrotransposition activity. Am. J. Hum. Genet. *73*, 1431-1437.

Myers, J.S., Vincent, B.J., Udall, H., Watkins, W.S., Morrish, T.A., Kilroy, G.E., Swergold, G.D., Henke, J., Henke, L., Moran, J.V., et al. (2002). A comprehensive analysis of recently integrated human Ta L1 elements. Am. J. Hum. Genet. *71*, 312-326.

O'Donnell, K.A., and Burns, K.H. (2010). Mobilizing diversity: transposable element insertions in genetic variation and disease. Mob. DNA *1*, 21.

Ovchinnikov, I., Rubin, A., and Swergold, G.D. (2002). Tracing the LINEs of human evolution. Proc. Natl. Acad. Sci. USA *99*, 10522-10527.

Ovchinnikov, I., Troxel, A.B., and Swergold, G.D. (2001). Genomic characterization of recent human LINE-1 insertions: evidence supporting random insertion. Genome Res. *11*, 2050-2058.

Park, S.J., Kim, Y.H., Lee, S.R., Choe, S.H., Kim, M.J., Kim, S.U., Kim, J.S., Sim, B.W., Song, B.S., Jeong, K.J., et al. (2015). Gain of a new exon by a lineage-specific Alu element-Integration event in the BCS1L gene during primate evolution. Mol. Cells. *38*, 950-958.

Philippe, C., Vargas-Landin, D.B., Doucet, A.J., van Essen, D., Vera-Otarola, J., Kuciak, M., Corbin, A., Nigumann, P., and Cristofari, G. (2016). Activation of individual L1 retrotransposon instances is restricted to cell-type dependent permissive loci. Elife. *5*.

Schrader, L., and Schmitz, J. (2018). The impact of transposable elements in adaptive evolution. Mol. Ecol. [In press] Available at: https://doi.org/10.1111/mec.14794.

Seleme, M.C., Vetter, M.R., Cordaux, R., Bastone, L., Batzer, M.A., and Kazazian, H.H., Jr. (2006). Extensive individual variation in L1 retrotransposition capability contributes to human genetic diversity. Proc. Natl. Acad. Sci. USA *103*, 6611-6616.

Sellis, D., Provata, A., and Almirantis, Y. (2007). Alu and LINE1 distributions in the human chromosomes: evidence of global genomic organization expressed in the form of power laws. Mol. Biol. Evol. *24*, 2385-2399.

Smith, A.M., Heisler, L.E., Mellor, J., Kaper, F., Thompson, M.J., Chee, M., Roth, F.P., Giaever, G., and Nislow, C. (2009). Quantitative phenotyping via deep barcode sequencing. Genome Res. *19*, 1836-1842.

Sotero-Caio, C.G., Platt, R.N., 2nd, Suh, A., and Ray, D.A. (2017). Evolution and diversity of transposable elements in vertebrate genomes. Genome Biol. Evol. *9*, 161-177.

Steinhauser, S., Kurzawa, N., Eils, R., and Herrmann, C. (2016). A comprehensive comparison of tools for differential ChIP-seq analysis. Brief Bioinform. *17*, 953-966.

Streva, V.A., Jordan, V.E., Linker, S., Hedges, D.J., Batzer, M.A., and Deininger, P.L. (2015). Sequencing, identification and mapping of primed L1 elements (SIMPLE) reveals significant variation in full length L1 elements between individuals. BMC Genomics. *16*, 220.

Strino, F., and Lappe, M. (2016). Identifying peaks in *-seq data using shape information. BMC Bioinformatics *17 Suppl 5*, 206.

Tripathi, S., Pohl, M.O., Zhou, Y., Rodriguez-Frandsen, A., Wang, G., Stein, D.A., Moulton, H.M., DeJesus, P., Che, J., Mulder, L.C., et al. (2015). Meta- and orthogonal integration of influenza "OMICs" data defines a role for UBR4 in virus budding. Cell Host Microbe. *18*, 723-735.

Valencia, C.A., Rhodenizer, D., Bhide, S., Chin, E., Littlejohn, M.R., Keong, L.M., Rutkowski, A., Bonnemann, C., and Hegde, M. (2012). Assessment of target enrichment platforms using massively parallel sequencing for the mutation detection for congenital muscular dystrophy. J. Mol. Diagn. *14*, 233-246.

Van den Broeck, D., Maes, T., Sauer, M., Zethof, J., De Keukeleire, P., D'Hauw, M., Van Montagu, M., and Gerats, T. (1998). Transposon Display identifies individual transposable elements in high copy number lines. Plant J. *13*, 121-129.

Wang, L., Norris, E.T., and Jordan, I.K. (2017). Human retrotransposon insertion polymorphisms are associated with health and disease via gene regulatory phenotypes. Front Microbiol. *8*, 1418.

Witherspoon, D.J., Xing, J., Zhang, Y., Watkins, W.S., Batzer, M.A., and Jorde, L.B. (2010). Mobile element scanning (ME-Scan) by targeted high-throughput sequencing. BMC Genomics. *11*, 410.