Landscape of Insertion Polymorphisms in the Human Genome

Masahiro Onozawa, Liat Goldberg, and Peter D. Aplan*

Genetics Branch, National Cancer Institute, National Institutes of Health, Bethesda, Maryland

*Corresponding author: E-mail: aplanp@mail.nih.gov.

Accepted: February 26, 2015

Data deposition: This project has been deposited at dbVar under the accession nstd105.

Abstract

Nucleotide substitutions, small (<50 bp) insertions or deletions (indels), and large (>50 bp) deletions are well-known causes of genetic variation within the human genome. We recently reported a previously unrecognized form of polymorphic insertions, termed templated sequence insertion polymorphism (TSIP), in which the inserted sequence was templated from a distant genomic region, and was inserted in the genome through reverse transcription of an RNA intermediate. TSIPs can be grouped into two classes based on nucleotide sequence features at the insertion junctions; class 1 TSIPs show target site duplication, polyadenylation, and preference for insertion at a 5′-TTTT/A-3′ sequence, suggesting a LINE-1 based insertion mechanism, whereas class 2 TSIPs show features consistent with repair of a DNA double strand break by nonhomologous end joining. To gain a more complete picture of TSIPs throughout the human population, we evaluated whole-genome sequence from 52 individuals, and identified 171 TSIPs. Most individuals had 25–30 TSIPs, and common (present in >20% of individuals) TSIPs were found in individuals throughout the world, whereas rare TSIPs tended to cluster in specific geographic regions. The number of rare TSIPs was greater than the number of common TSIPs, suggesting that TSIP generation is an ongoing process. Intriguingly, mitochondrial sequences were a frequent template for class 2 insertions, used more commonly than any nuclear chromosome. Similar to single nucleotide polymorphisms and indels, we suspect that these TSIPs may be important for the generation of human diversity and genetic diseases, and can be useful in tracking historical migration of populations.

Key words: templated sequence insertion polymorphisms (TSIPs), mitochondria, polymorphism, human migration, DNA repair, LINE-1 retrotransposon.

Introduction

A large number of polymorphisms in human genomes have been identified and classified as single nucleotide polymorphisms (SNPs), small (<50 bp) insertions or deletions, referred to collectively as short indels (Montgomery et al. 2013), and large deletions (>50 bp; Beck et al. 2010; Huang et al. 2010; Iskow et al. 2010; 1000 Genomes Project Consortium et al. 2012; Montgomery et al. 2013). Analysis of genomes from 1,092 individuals identified 38 million SNPs, 1.4 million short indel polymorphisms, and 14,000 larger (>50 bp) deletion polymorphisms (1000 Genomes Project Consortium et al. 2012). In addition, a large number of polymorphic LINE-1 or Alu elements have been catalogued (Beck et al. 2010; Huang et al. 2010; Iskow et al. 2010), and retroposed, processed gene transcript polymorphisms similar to processed pseudogenes have been predicted by whole-genome sequence analysis (Ewing et al. 2013). However, the experimental designs used by these investigators were designed to specifically identify LINE-1 (Huang et al. 2010; Iskow et al. 2010) or processed pseudogene (Ewing et al. 2013) insertions.

We recently discovered that experimentally induced DNA double strand breaks (DSBs) can be repaired by insertion of sequences derived from distant regions of the genome, termed templated sequence insertions, or TSIs (Onozawa et al. 2014). Moreover, analysis of whole-genome sequence files from two myeloma cell lines indicated that TSIs could be detected as a pair of structural variations (SVs; fig. 1*A*). We identified TSIs from SV data and verified each TSI by Sanger sequencing (fig. 1*B* and *C*; Onozawa et al. 2014). Identical TSIs were found in several cell lines, suggesting that these TSIs were polymorphic within the human genome (fig. 1*C*). We designated this form of polymorphism as a “TSIP (templated sequence insertion polymorphism).” In this manuscript, we study a large database of whole-genome sequence from healthy individuals of varied ethnic background to identify a nonbiased landscape of TSIPs.

Materials and Methods

TSIP Identification from Publicly Available Data

We studied a publicly available catalog of whole-genome sequence from 52 individuals (http://www.completegenomics.com/public-data/, last accessed March 17, 2015; Drmanac et al. 2010). Whole-genome sequence data were generated on 52 healthy normal individuals from 12 distinct ethnic backgrounds: 7 YRI; 5 ASW; 4 LWK; 4 MKK; 4 CEU; 4 CEPH; 4 TSI; 4 CHB; 4 JPT; 4 GIH; 4 MXL; 2 PUR. These samples are further categorized into four super regional groups: AFR (YRI, ASW, LWK, MKK); EUR (CEU; CEPH; TSI); ASN (CHB, JPT, GIH); and AMR (MXL; PUR) (supplementary table S1, Supplementary Material online).

Paired end reads (which consisted of two DNA sequences flanking a nonsequenced internal region), that mapped far apart, in the wrong orientation, or to different chromosomes were flagged as "discordant." Clusters of multiple discordant pair reads that had similar positioning and orientation were assembled as an SV, which could represent a deletion, inversion, insertion, or translocation. Each individual had between 3,938 and 8,243 SVs (mean 5,451 per individual). We screened the combined data of 52 normal individuals contained in a "SV baseline genome set" file (B37baseline-junctions.tsv. available at www.completegenomics.com/sequence-data/download-data/), which contained a total of 39,595 independent SVs from the 52 individuals. We identified pairs of reciprocal fusion sequences that might represent interchromosomal TSIs by applying the following criteria. First, both fusion junctions had to be located within 50 kb of one another. Second, the strand polarity had to align such that an insertion was feasible. Third, both ends of the SV needed to be aligned to a single genomic loci; SV that showed multiple or imperfect alignments (<95% sequence identity) were excluded as misaligned events. Finally, SV files (allJunctionsBeta file) from specific individuals were examined to identify individuals with the candidate TSI. We excluded intrachromosomal SVs because these may include *cis* genomic events including inversions, duplications, large deletions, distal duplications, or combinations thereof. These TSIP variants have been submitted to dbVar (accession number nstd105).

Verification of TSIP Sequences

Nine genomic DNA samples (NA18501 (YRI), NA19834 (ASW), NA21732 (MKK), NA06985 (CEU), NA20509 (TSI), NA18555 (HCB), NA18558 (HCB), NA18947 (JPT), HG00732 (PUR)) were purchased from Coriell institute (Camden, NJ). Primers flanking the predicted insertions fragment were generated and the potential insertions were validated by polymerase chain reaction (PCR) amplification. PCR amplifications were performed using PCR SuperMix High Fidelity enzyme and buffers (Invitrogen, Carlsbad, CA) for short (<1.5 kb) fragments or Long Amp Taq (New England Biolabs, Ipswich, MA) for longer fragments (>1.5 kb). Nucleotide sequences were determined by Sanger sequencing and capillary electrophoresis at the NCI DNA core facility. Inserted sequences were aligned to the reference genome (GRCh37/hg19) using BLAT (Kent et al. 2002) to identify the origin of the inserted fragment. Inserted fragments were designated as genic (exonic or intronic sequence) or retrotransposon sequence by using UCSC Genes, ENCODE/GENCODE version 17, and Repeat Masker.

Results and Discussion

In order to generate a detailed landscape of interchromosomal TSIPs in human genomes, we studied a publicly available catalog of whole-genome sequence from 52 individuals (http://www.completegenomics.com/public-data/; Drmanac et al. 2010). Whole-genome sequence data was generated on 52 healthy normal individuals from 12 distinct ethnic backgrounds: 7 YRI; 5 ASW; 4 LWK; 4 MKK; 4 CEU; 4 CEPH; 4 TSI; 4 CHB; 4 JPT; 4 GIH; 4 MXL; 2 PUR. These samples are further categorized into four super regional groups: AFR (YRI, ASW, LWK, MKK); EUR (CEU; CEPH; TSI); and ASN (CHB, JPT, GIH); AMR (MXL; PUR) (supplementary table S1, Supplementary Material online).

In total, we identified 171 potential TSIPs present in one or more individuals (supplementary tables S2 and S3, Supplementary Material online). Each TSIP contained a unique inserted donor sequence and a unique acceptor site. The prevalence of each TSIP varied from 2% (1/52) to 98% (51/52) (fig. 2*A*). "Common" TSIPs were defined as those present in at least 20% of the individuals sampled; 25.7% (45/171) of the TSIPs were common, whereas 74.3% (127/171) were rare (fig. 2*A* and supplementary table S3, Supplementary Material online). Sixty-three of 171 TSIPs were found in a single individual, and could represent de novo germline mutations. It is noteworthy that three TSIPs (chr22 into 6, chr8 into 11, chr12 into 15) were present at a frequency of close to 100% (96.2–98.1%), suggesting that the reference genome (GRCh37/hg19) may be based on an uncommon variant that lacks the TSIPs at these insertion sites (supplementary table S3, Supplementary Material online).

Not surprisingly, all of the common TSIPs were found in individuals from more than one regional group (fig. 2*B*); the most recent common ancestor for these common TSIPs is predicted to have resided in Africa. The number of unique TSIPs in AFR was disproportionally large, and most (45/64) of the rare TSIPs were present in only a single regional group (fig. 2*B*), as might be expected for polymorphisms that originated more recently in evolutionary time. Each individual had a unique set of TSIPs (supplementary table S3, Supplementary Material online). AFR individuals had significantly more TSIPs compared with individuals from other regions, whereas there was no significant difference in TSIPs per individual among individuals from EUR, ASN, and AMR

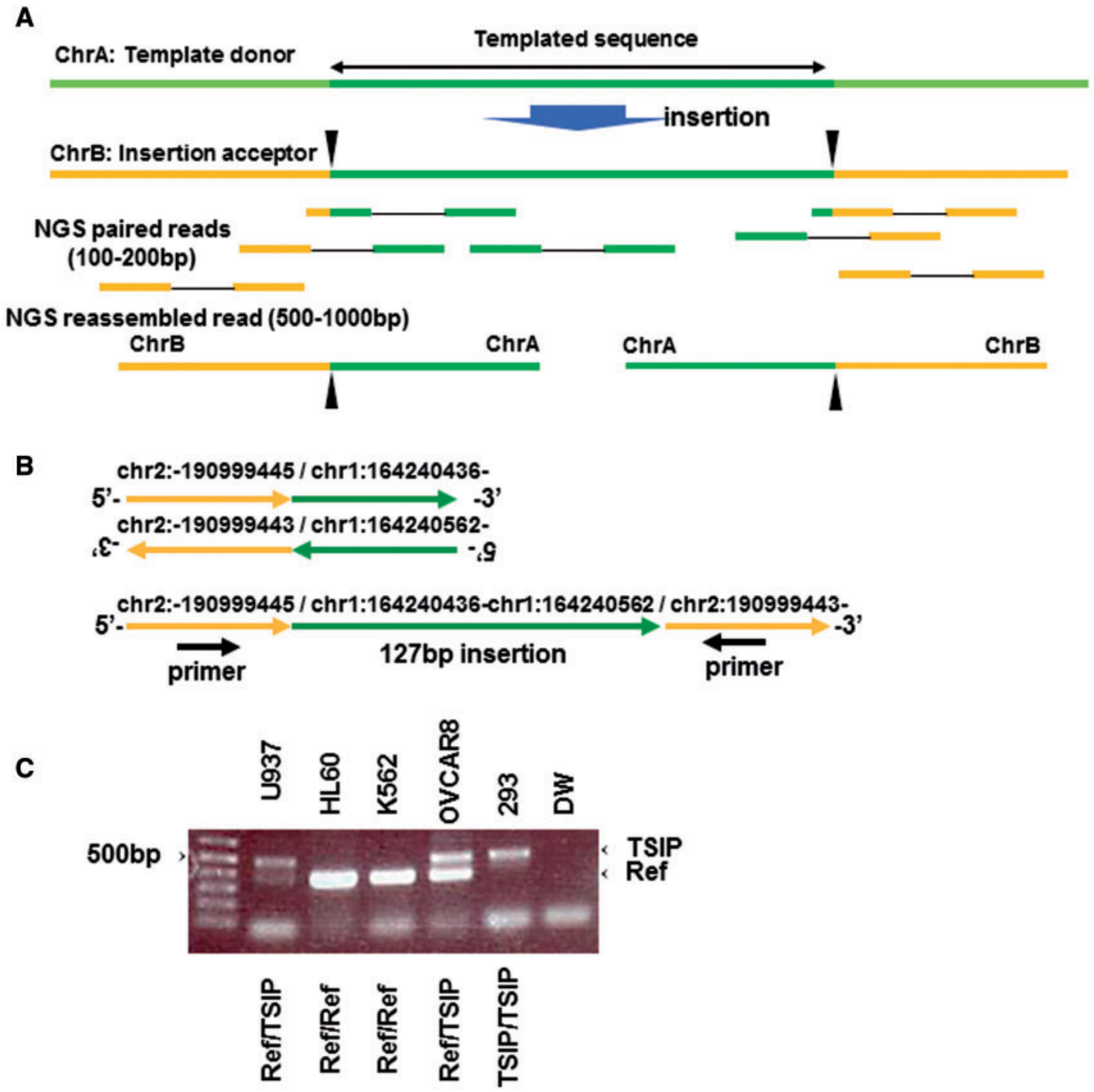

FIG. 1.—Identification and validation of TSIPs. (*A*) A TSI produces a pair of closely linked SVs. (*B*) Example of a pair of SVs that could be caused by insertion of 127 bp of chr 1 sequence into chr 2. PCR primers flanking the putative insertion are indicated. (*C*) The putative TSI shown in (*B*) was verified by PCR amplification and sequencing from several human cell lines. An identical TSIP was found in several cell lines demonstrating that the TSI is polymorphic in the human genome. Ref: reference sequence.

(fig. 2*C*). The increased TSIP diversity found in AFR individuals is consistent with findings from mitochondrial (Cann et al. 1987) and Y-chromosome (Hammer 1995; Underhill et al. 2000) polymorphisms and supports the accepted notion that Homo sapiens originated in Africa.

We analyzed the origin of the donor templated loci and acceptor insertion loci. Somewhat surprisingly, almost 20% of the TSIP donor sequences identified through analysis of the 52 individuals were derived from mitochondrial sequences (fig. 3*A*). Conversely, there were no TSIPs inserted into mitochondria (fig. 3*A*). We previously reported that experimentally induced TSIs, produced by repair of a DNA DSB, were derived from all 22 autosomes and 2 sex chromosomes, with no clear preference for any chromosome. However, none of the experimentally induced TSIs were templated from mitochondrial sequence (Onozawa et al. 2014). Excluding mitochondrial sequence insertions, 67% of the TSIP donor sequences were derived from genic regions, and 58% of the TSIP acceptor sites were genic regions (fig. 3*B*). Compared with the human genome composition, genic (exon + intron) regions were overrepresented for both TSIP donor and acceptor sites (fig. 3*B*).

To validate the predicted TSIPs, we obtained genomic DNA from 9 of the 52 genomes that were sequenced (NA18501 (YRI), NA19834 (ASW), NA21732 (MKK), NA06985 (CEU), NA20509 (TSI), NA18555 (HCB), NA18558 (HCB), NA18947 (JPT), HG00732 (PUR)). These nine samples had a total of 89

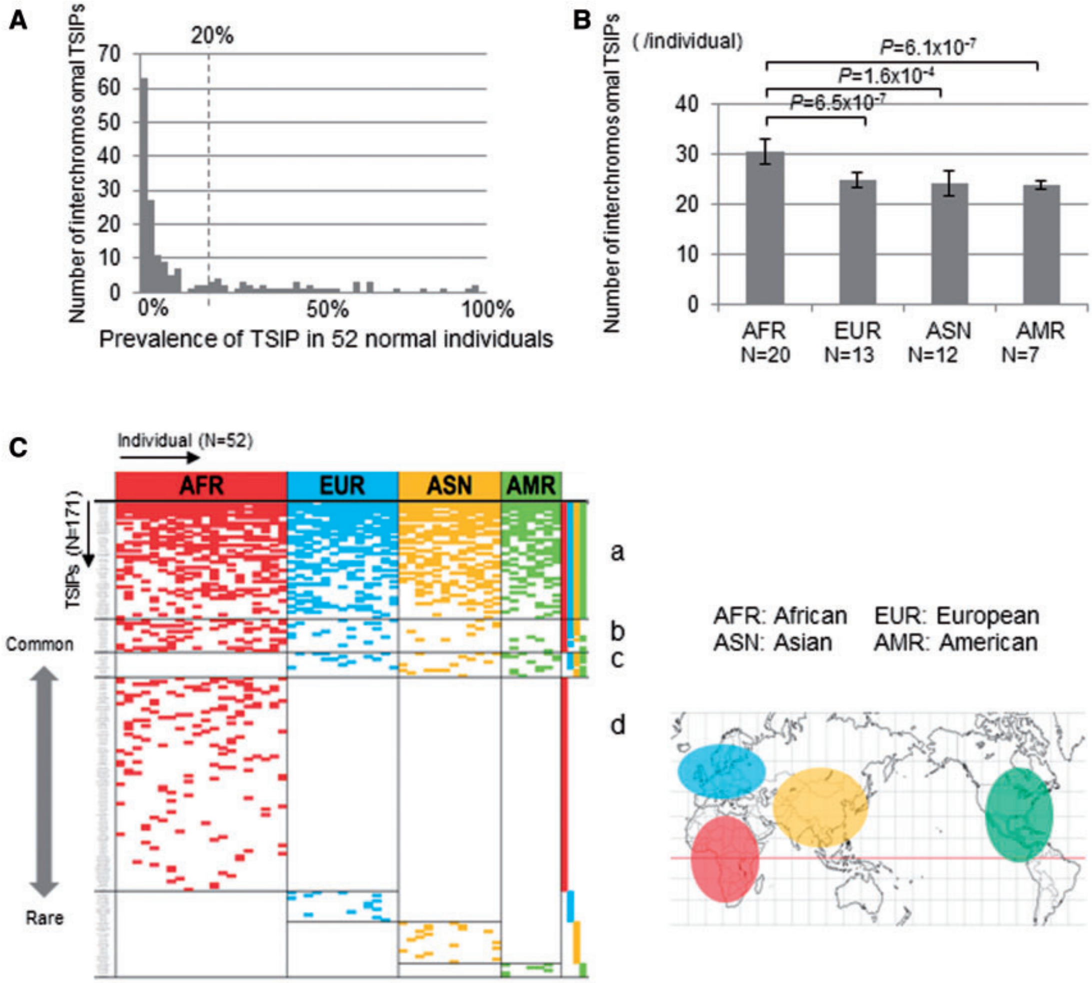

Fig. 2.—TSIPs identified in normal individuals. (*A*) The prevalence of each TSIP in the 52 genomes studied was determined. 127 of the 171 (73.4%) TSIPs were present in less than 20% of the individuals and were designated "rare," the remainder were designated common. (*B*) Each column represents a single individual (N = 52), grouped by region. Each row represents an independent TSIP (N = 171). TSIP data were sorted by 1) geographic distribution and 2) prevalence (common to rare). Geographic distribution of the TSIPs was sorted into four groups: a, present in all four regions; b, present in AFR+1–2 other regions; c, not present in AFR but present in 2–3 other region; d, present only in 1 region. (*C*) Each individual had a unique number and combination of TSIPs. AFR individuals had a significantly larger number of TSIPs (27–38, mean 30.5) compared with other regional super groups (EUR [19–28, mean 24.5], ASN [19–28, mean 24.3], AMR [22–26, mean 24]). Statistical analysis utilized the Students' *t*-test.

different predicted TSIPs, and we designed primers flanking the predicted insertions for each of these 89 TSIPs. Using PCR and Sanger sequencing of the PCR products, we successfully validated 69 (77.5%) of the 89 predicted TSIPs (supplementary text S1 and table S4, Supplementary Material online). Of the 20 samples where we were unable to amplify a predicted TSIP, four cases did not amplify either the TSIP or the reference allele, and five cases had a predicted TSIP PCR that was greater than 1 kb. In the remaining 11 cases, the predicted TSIP may have been a false prediction or, alternatively, the inserted sequence may have contained difficult to amplify sequences. Although most of the inserted fragments were derived from a single genomic donor site, two TSIPs (Chr2+10 into chr 20, two discrete ChrM sequences into 5p13.3) contained two donor fragments from distinct genomic or mitochondrial regions inserted into a single acceptor site (supplementary text S1 and table S4, Supplementary Material online). Three TSIPs (Chr5 into 4, Chr22 into 1, ChrX into 11) were inserted into coding exons (SCD5, C1orf194, DCDC5, respectively) and generated frameshift mutations, showing that TSIPs have the potential for functional consequences (supplementary text S1 and table S4, Supplementary Material online).

We have previously classified TSIPs into class 1 or class 2 events based on nucleotide sequence features at the insertion junction (fig. 4*A*). Class 1 TSIPs display all of the hallmarks of a retrotransposon-induced event. The cleavage is typically at a

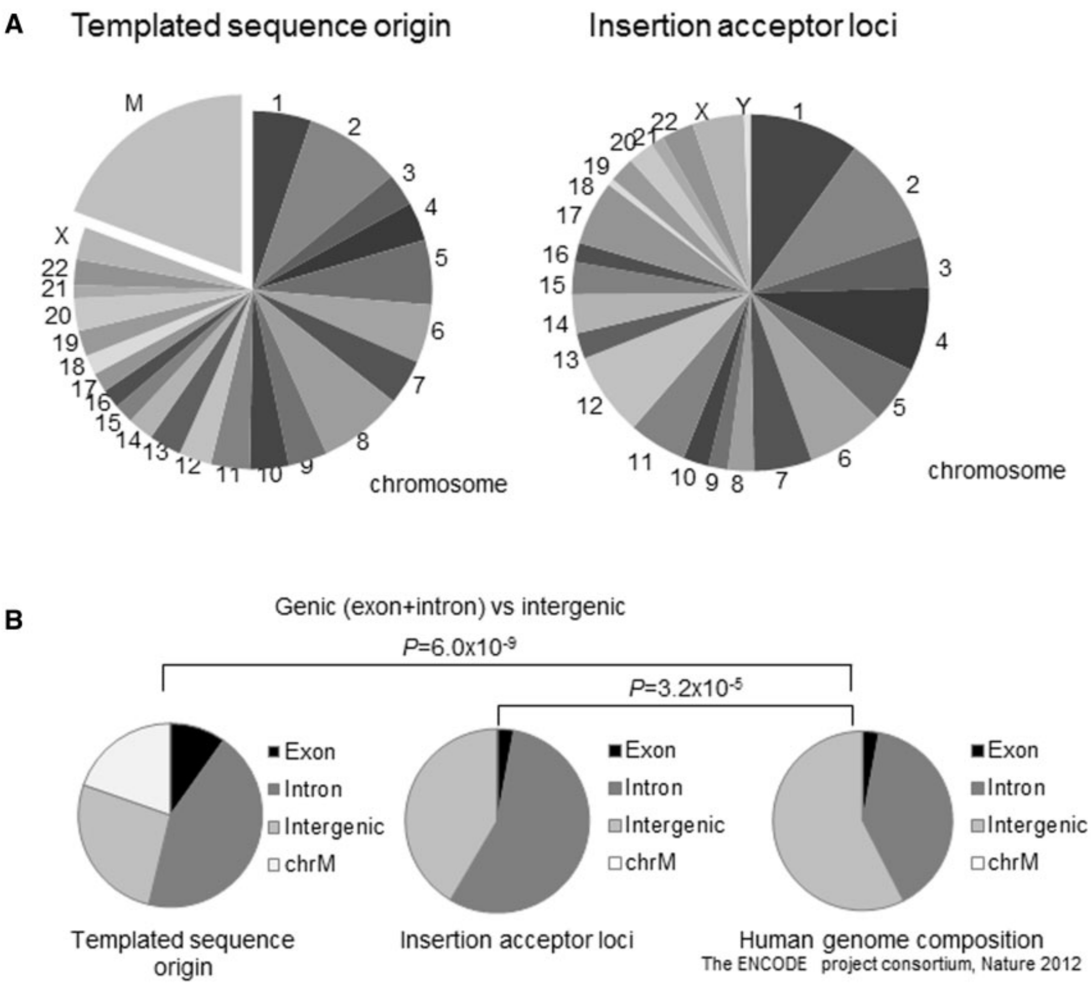

Fig. 3.—Characteristics of the template donor and acceptor sites. (*A*) Mitochondrial sequence (chrM) donors are over-represented as 19% of the TSIPs. Conversely, no chrM serves as an insertion acceptor site. (*B*) Genic regions (exon + intron) were significantly overrepresented as TSIP origins ($P = 6.0 \times 10^{-9}$) and insertion acceptor loci ($P = 3.2 \times 10^{-5}$) compared with human genome composition (χ^2 test). Insertion of a contiguous exon–intron fragment was considered to be an exon insertion.

preferred 5′-TTTT/A-3′ LINE-1 integration site. Target site duplications (TSDs) flank the inserted sequence, which is accompanied by a nontemplated polyA tract directed by a polyadenylation signal (5′-AATAAA-3′) typically located within 20 bp of the polyA tract, leading to the suspicion that these insertions were templated from RNA species and integrated by LINE-1 integrase and reverse transcriptase activity. Class 1 TSIPs included LINE-1/SINE sequence insertions as well as insertions of processed pseudogenes (fig. 4*B*). However, nearly half of the class 1 TSIs were derived from intronic or intergenic regions, and were apparently mobilized "in *trans*," without active LINE-1 sequences nearby the donor sequence. Similar to the LINE-1 and pseudogene sequences, these intronic or intergenic sequences were polyadenylated and preceded by a consensus polyadenylation signal (supplementary text S1 and table S4, Supplementary Material online); and 5/6 of the inserted intronic sequences were derived from the noncoding strand. We suspect that these intronic and intergenic class 1 TSIPs are derived from nonannotated RNA transcripts.

Class 2 TSIPs represent a novel form of insertion polymorphism which do not show TSD, polyA tracks, or a preferred integration site (fig. 4*A* and *B* and supplementary text S1, Supplementary Material online). Instead, they show features typically associated with NHEJ, such as short deletions at the insertion site, short track microhomology, and the addition of nontemplated nucleotides at the insertion junctions. We suspect that class 2 TSIPs were produced by repair of a DNA DSB created by physiologic or environmental DNA damage (ROS, ionizing radiation, genotoxins, etc) via a "patch" of sequence that was derived from a distant region of the genome (Onozawa et al. 2014). The class 2 TSIPs were similar to the

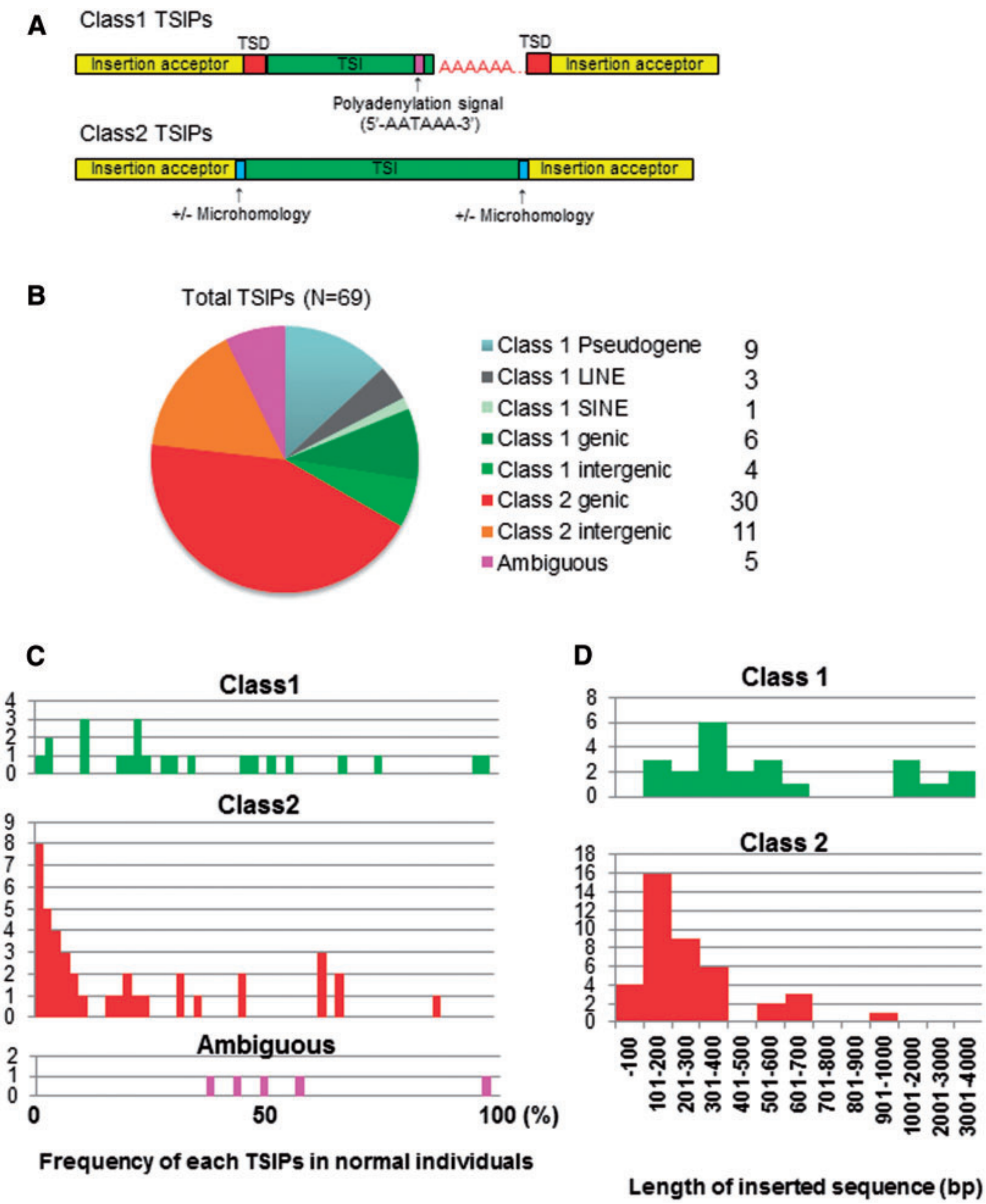

Fig. 4.—Nucleotide sequence features of verified TSIPs. (*A*) Class 1 TSIPs have a TSD of 5–20 bp, a nontemplated polyA tract, and a polyadenylation signal (5′-AATAAA-3′). Class 2 TSIPs lack a TSD and polyA tract, and display variable-sized deletions at the insertion acceptor site as well as microhomology, and, less commonly, nontemplated nucleotide addition at the junction site. TSIPs had to show a clear TSD and polyA tract to be designated "class 1." TSIPs with only a TSD or polyA, were classified as "Ambiguous." (*B*) 69 TSIPs were sequence verified, including 23 class1, 41 class2, and 5 ambiguous TSIPs. The 23 Class 1 TSIPs could be further divided based on the origin of the inserted sequence (9 pseudogene, 3 3′ LINE-1 transductions, 1 SINE, 6 polyadenylated intronic sequence, and 4 polyadenylated intergenic sequence). (*C*) Class 2 TSIPs are more commonly rare TSIPs (χ^2 $P = 0.018$). (*D*) Lengths of the inserted templated sequence are shown (Class 1 + 2, $N = 64$). Class 2 TSIs were shorter (45–988 bp, median: 204 bp) than class 1 TSIs (107–3,771 bp, median: 417 bp) (*t*-test, $P = 0.00016$).

TSIs produced by creating a DNA DSB using the I-SceI enzyme (Onozawa et al. 2014), in terms of mean size of insertion and the NHEJ features discussed above. Several lines of evidence indicated that these experimentally produced patches used to repair the I-SceI induced DNA DSB were derived from reverse transcription of RNA, as opposed to deletion of genomic DNA from the donor site (Onozawa et al. 2014). We showed that the experimentally induced TSIs could be templated from RNA

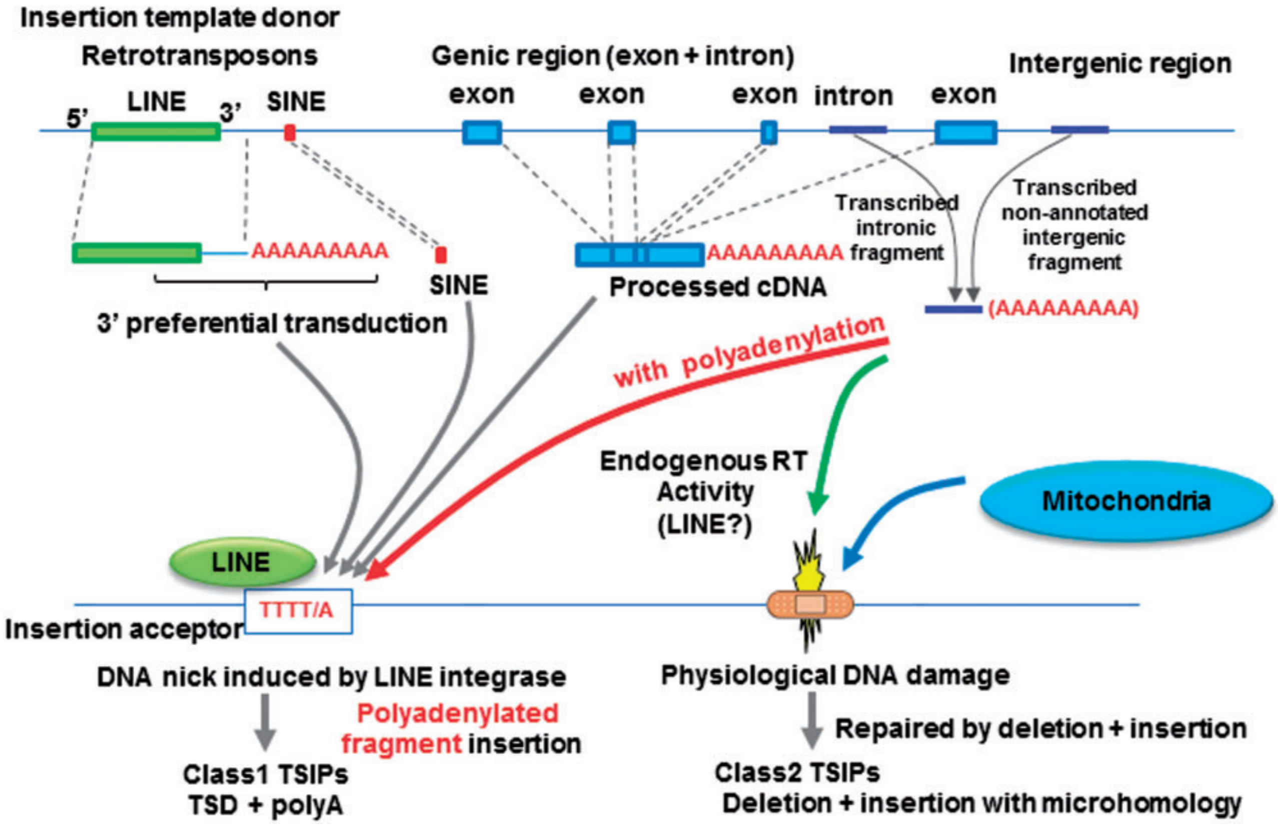
Insertion template donor
Retrotransposons
LINE
5'
3'
SINE
AAAAAAAAA
SINE
3' preferential transduction
Genic region (exon + intron)
exon
exon
exon
intron
exon
Intergenic region
Processed cDNA
AAAAAAAAA
Transcribed intronic fragment
Transcribed non-annotated intergenic fragment
(AAAAAAAAA)
with polyadenylation
Endogenous RT Activity (LINE?)
Mitochondria
LINE
TTTT/A
Insertion acceptor
DNA nick induced by LINE integrase
Polyadenylated fragment insertion
Class1 TSIPs
TSD + polyA
Physiological DNA damage
Repaired by deletion + insertion
Class2 TSIPs
Deletion + insertion with microhomology

Fig. 5.—Landscape of TSIPs in Human genome. LINE-1 mediated integration of LINE-1/SINE sequences, LINE-1 plus 3′-transduced sequences, and processed cDNA insertions are known to create insertion polymorphisms. Polyadenylated intronic or intergenic fragments can also be acted upon in *trans* by LINE-1 ORF2 and integrate at the site of a nick created by LINE-1 ORF2, resulting in a class 1 TSIP. Reverse transcription of exonic or nonannotated transcripts can be used to patch and repair a DNA DSB. Alternatively, a DNA DSB can be repaired by fragments of mitochondrial DNA or reverse-transcribed RNA. Both forms of repair would generate a class 2 TSIP.

that was cotransfected with an I-SceI expression vector, and that the frequency of TSIs was reduced more than 3-fold by treatment with reverse transcriptase inhibitors. In addition, the TSI sequences were not excised from the "donor" site (Onozawa et al. 2014), and a copy number gain of the inserted sequence could be detected in the genome of the subclone containing the TSI (Onozawa et al. 2014). Thus, although we cannot exclude the possibility that some of the class 2 TSIPs were produced from genomic DNA, by analogy to the experimentally induced TSIs, we suggest that a majority of the class 2 TSIPs were produced via reverse transcription of an RNA intermediate.

Interestingly, a large number of class 2 TSIPs were caused by insertions of sequences derived from mitochondrial DNA. The human genome contains numerous mitochondrial pseudogenes, equal to approximately 35 mitochondrial genomes (Thomas et al. 1996; Yuan et al. 1999; Willett-Brozick et al. 2001; Tourmen et al. 2002). Therefore, it is possible that one or more of the TSIPs we designated as insertions of mitochondrial origin instead originated from a nuclear pseudogene; this mechanism has been previously suggested as one of several models that could explain the insertion of mitochondrial sequence at the site of a balanced translocation breakpoint (Willett-Brozick et al. 2001). However, we regard this possibility to be an unlikely mechanism for generation of mitochondrial TSIPs, as the inserted sequences in question matched mitochondrial sequence far better than any nuclear sequence. For instance, the insertion of ChrM (161 bp, *MTND5* exon) into chr4p15.1 is a 99.4% match for the reference mitochondrial genome (one mismatch), whereas the best match in the nuclear genome (chr 5:134261491–134261651) has ten mismatches. Although it remains possible that the chromosome 5 sequence was inserted into chr4, and then underwent specific mutations that coincidentally matched the mitochondrial sequence, this sequence of events would seem to be improbable, as previously discussed (Willett-Brozick et al. 2001).

As discussed above, the TSI events at experimentally induced DNA DSB closely resembled class 2 TSIPs, in terms of microdeletions at the DNA DSB site, as well as microhomology and nontemplated nucleotide addition (Varga and Aplan

2005; Cheng et al. 2010; Onozawa et al. 2014). However, in those experiments, none of 180 independent TSI events used to repair an experimentally induced DNA DSB were derived from mitochondrial DNA, in stark contrast with 33/171 (19.3%) of the TSIPs identified in our current study (χ^2 test, $P = 4.72 \times 10^{-9}$). In addition, we searched publicly available data from two reports of matched tumor and normal whole-genome sequence, and identified no instances of mtDNA fusions among a total of 620 verified somatic SVs (Welch et al. 2012; Shern et al. 2014). Therefore we suspect that the polymorphic mitochondrial sequence insertions are reproduction-specific events that take place in germ cells or early stage embryos, but do not occur, or occur only rarely, in somatic cells. Intriguingly, ubiquitination and subsequent destruction of sperm mitochondria shortly after fertilization is a well-known phenomenon (Sutovsky et al. 2000). Therefore, it is conceivable that fragmented paternal mitochondrial DNA can be inserted into nuclear genome, exclusively in fertilized embryos (Woischnik and Moraes 2002). As discussed above, mitochondrial pseudogenes have previously been identified, but the mechanism(s) leading to these insertions have not been determined (Thomas et al. 1996; Tourmen et al. 2002). We speculate that fragmented paternal mitochondria DNA or reverse transcribed RNA could be used as a patch to repair a DNA-DSB in the fertilized egg before initiating cell division, thus leading to a TSIP containing mitochondrial sequence.

There is potential for this mechanism of DNA DSB repair to cause genetic disease. A case of Pallister–Hall syndrome was reported in which 72-bp of mitochondrial sequence was inserted into and disrupted the *GLI3* gene (Turner et al. 2003). The conception of this patient was temporally and geographically associated with high-level radioactive contamination following the Chernobyl accident (Turner et al. 2003). Although speculative, it is conceivable that a DNA DSB in the germ cell, caused by ionizing radiation, was repaired by a TSI derived from mitochondrial DNA in this case.

Rare TSIPs, which are likely to be more recent evolutionary events, tended to be class 2 events (fig. 4*C*), consistent with the notion that LINE-1 activity has decreased in the human lineage over the past 25 million years (Lander et al. 2001; Zhang et al. 2003; Khan et al. 2006). In addition, the lengths of the inserted class 2 TSIPs were smaller (45–988 bp, median: 204 bp) than class 1 TSIPs (107–3,771 bp, median: 417 bp) (*t*-test, $P = 0.00016$) (fig. 4*D*).

Conclusion

Each TSIP that we have identified is unique, composed of unique donor and acceptor sequences, as well as unique sequence features at the insertion junctions. We propose two distinct mechanisms to account for the generation of these TSIPs, namely LINE-1 mediated insertions (class 1) and DNA DSB repair (class 2) (fig. 5). As opposed to SNPs, which can arise as recurrent, independent events, the likelihood that one of these TSIPs, would occur as an independent, recurrent event would seem to be vanishingly small, and could be used to trace founder populations. These TSIPs provide unique polymorphic markers, similar to SNPs and variable tandem repeats, and may contribute to genetic disease, especially if the TSIP disrupts a coding exon.

Supplementary Material

Supplementary tables S1–S4 and text S1 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This work was supported by the intramural research program of the NCI, NIH. M.O. was supported by the Japan Society for the Promotion of Science (JSPS), Postdoctoral Fellowships for Research Abroad program. The authors thank Michael Kuehl, Javed Khan, Joshua Waterfall, and Paul Meltzer for insightful discussions.

Literature Cited

1000 Genomes Project Consortium, et al. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65.

Beck CR, et al. 2010. LINE-1 retrotransposition activity in human genomes. Cell 141:1159–1170.

Cann RL, Stoneking M, Wilson AC. 1987. Mitochondrial DNA and human evolution. Nature 325:31–36.

Cheng Y, et al. 2010. Efficient repair of DNA double-strand breaks in malignant cells with structural instability. Mutat Res. 683: 115–122.

Drmanac R, et al. 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science 327:78–81.

Ewing AD, et al. 2013. Retrotransposition of gene transcripts leads to structural variation in mammalian genomes. Genome Biol. 14:R22.

Hammer MF. 1995. A recent common ancestry for human Y chromosomes. Nature 378:376–378.

Huang CR, et al. 2010. Mobile interspersed repeats are major structural variants in the human genome. Cell 141:1171–1182.

Iskow RC, et al. 2010. Natural mutagenesis of human genomes by endogenous retrotransposons. Cell 141:1253–1261.

Kent WJ, et al. 2002. The human genome browser at UCSC. Genome Res. 12:996–1006.

Khan H, Smit A, Boissinot S. 2006. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. Genome Res. 16:78–87.

Lander ES, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.

Montgomery SB, et al. 2013. The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. Genome Res. 23:749–761.

Onozawa M, et al. 2014. Repair of DNA double-strand breaks by templated nucleotide sequence insertions derived from distant regions of the genome. Proc Natl Acad Sci U S A. 111:7729–7734.

Shern JF, et al. 2014. Comprehensive genomic analysis of rhabdomyosarcoma reveals a landscape of alterations affecting a common genetic axis in fusion-positive and fusion-negative tumors. Cancer Discov. 4: 216–231.

Sutovsky P, et al. 2000. Ubiquitinated sperm mitochondria, selective proteolysis, and the regulation of mitochondrial inheritance in mammalian embryos. Biol Reprod. 63:582–590.

Thomas R, Zischler H, Paabo S, Stoneking M. 1996. Novel mitochondrial DNA insertion polymorphism and its usefulness for human population studies. Hum Biol. 68:847–854.

Tourmen Y, et al. 2002. Structure and chromosomal distribution of human mitochondrial pseudogenes. Genomics 80:71–77.

Turner C, et al. 2003. Human genetic disease caused by de novo mitochondrial-nuclear DNA transfer. Hum Genet. 112:303–309.

Underhill PA, et al. 2000. Y chromosome sequence variation and the history of human populations. Nat Genet. 26:358–361.

Varga T, Aplan PD. 2005. Chromosomal aberrations induced by double strand DNA breaks. DNA Repair 4:1038–1046.

Welch JS, et al. 2012. The origin and evolution of mutations in acute myeloid leukemia. Cell 150:264–278.

Willett-Brozick JE, Savul SA, Richey LE, Baysal BE. 2001. Germ line insertion of mtDNA at the breakpoint junction of a reciprocal constitutional translocation. Hum Genet. 109:216–223.

Woischnik M, Moraes CT. 2002. Pattern of organization of human mitochondrial pseudogenes in the nuclear genome. Genome Res. 12: 885–893.

Yuan JD, Shi JX, Meng GX, An LG, Hu GX. 1999. Nuclear pseudogenes of mitochondrial DNA as a variable part of the human genome. Cell Res. 9:281–290.

Zhang Z, Harrison PM, Liu Y, Gerstein M. 2003. Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. Genome Res. 13:2541–2558.

Associate editor: Hidemi Watanabe