Mobile Interspersed Repeats Are Major Structural Variants in the Human Genome

Cheng Ran Lisa Huang, Anna M. Schneider, Yunqi Lu, Tejasvi Niranjan,Peilin Shen, Matoya A. Robinson, Jared P. Steranka, David Valle, Curt I. Civin, Tao Wang, Sarah J. Wheelan, Hongkai Ji, Jef D. Boeke and Kathleen H. Burns

Cell, Volume 171, Issue 7, 1171-82, June 25 2010

Triinu Kõressaar,
Seminar in bioinformatics

TARTU 2010

Genome wide studies

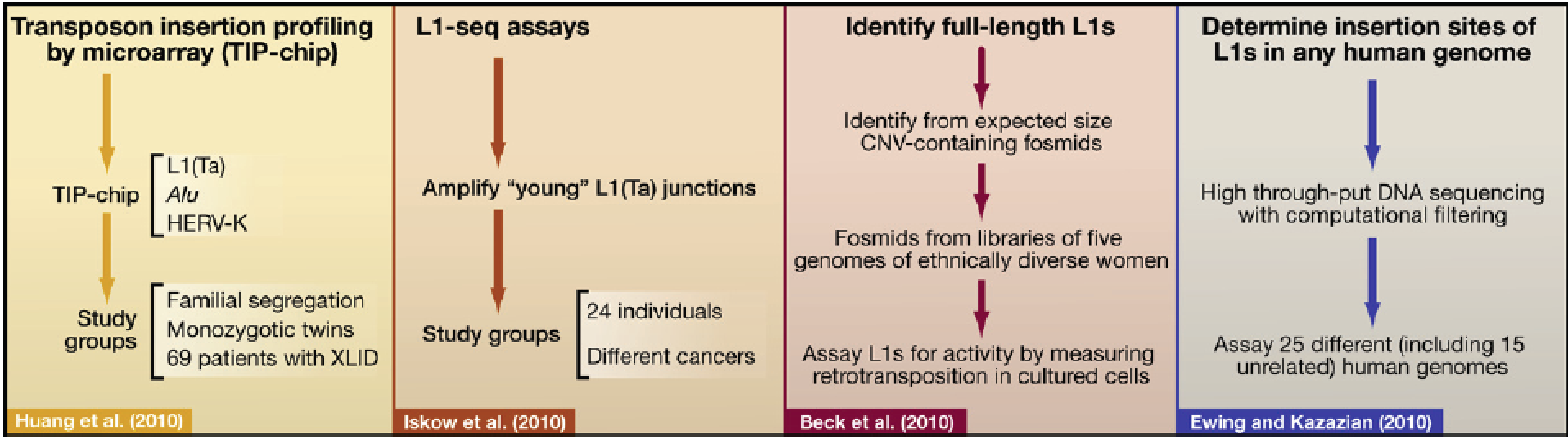
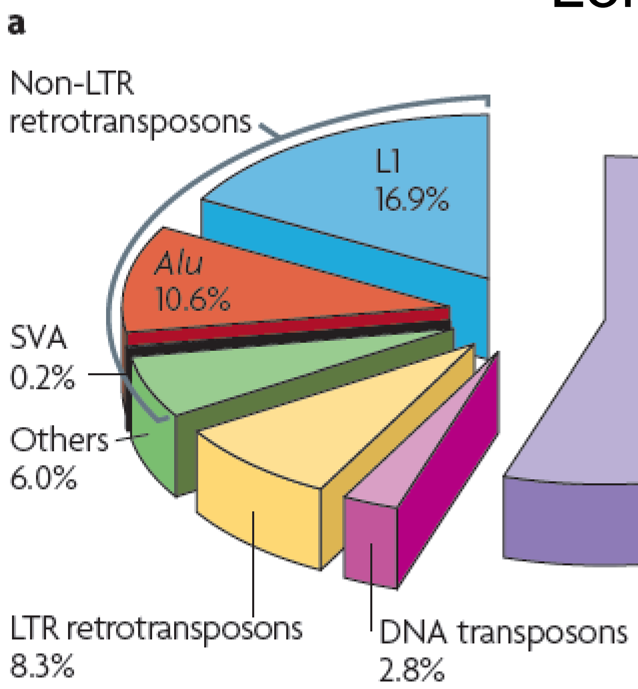

Transposon insertion profiling by microarray (TIP-chip)
TIP-chip
L1(Ta)
Alu
HERV-K
Study groups
Familial segregation
Monozygotic twins
69 patients with XLID
Huang et al. (2010)
L1-seq assays
Amplify "young" L1(Ta) junctions
Study groups
24 individuals
Different cancers
Iskow et al. (2010)
Identify full-length L1s
Identify from expected size CNV-containing fosmids
Fosmids from libraries of five genomes of ethnically diverse women
Assay L1s for activity by measuring retrotransposition in cultured cells
Beck et al. (2010)
Determine insertion sites of L1s in any human genome
High through-put DNA sequencing with computational filtering
Assay 25 different (including 15 unrelated) human genomes
Ewing and Kazazian (2010)

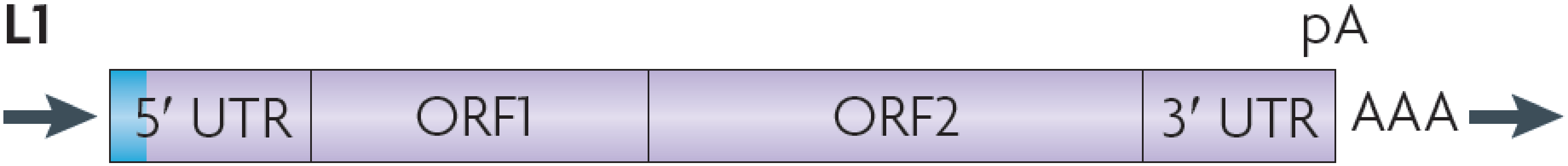
Figure 1. Identifying Repetitive Sequences and Structural Variations in the Human Genome
Four different studies use diverse genome-wide assays of personal genomes with or without next-generation massively parallel DNA sequencing to identify repetitive sequences and structural variations in the human genome (Beck et al., 2010; Huang et al., 2010; Iskow et al., 2010; Ewing and Kazazian, 2010). The L1(Ta) repeat sequences represent relatively young retrotransposition events of LINE elements. *Alu* is the most frequent repetitive sequence class in the human genome, originally identified more than 30 years ago by reassociation techniques. All four studies show that LINE and Alu elements contribute to structural variations in the human genome. XLID, X-linked intellectual disability.

1. Huang, C. R. L. *et al*. Mobile interspersed repeats are major structural variants in the human genome. Cell 141, 1171–1182 (2010)
2. Iskow, R. C. *et al*. Natural mutagenesis of human genomes by endogenous retrotransposons. Cell 141, 1253–1261 (2010)
3. Beck, C. R. *et al*. LINE-1 retrotransposon activity in human genomes. Cell 141, 1159–1170 (2010)
4. Ewing, A. D. & Kazazian, H. H. Jr. High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. Genome Res. 20 May 2010 (doi: 10.1101/gr.106419.110)

Long Interspersed element 1 (LINE-1, L1)



LTR – Long Terminal Repeats,
SVA – Short Interspersed Element region + Variable Number of Tandem Repeats region + *Alu-like* region



~6kb

5’ UTR – RNAPII promoter
ORF1 – RNA binding protein
ORF2 – protein with endonuclease and reverse-transcriptase activity

Cordaux and Batzer, 2010 (Nature Reviews)

Retrotransposones

Long interspersed element-1s (LINE-1s or L1s) are the most abundant class of retrotransposon in the human genome

A major component of interindividual variation

Might contribute to disease and complex traits

There is a great deal more recent or current retrotransposon activity than anticipated

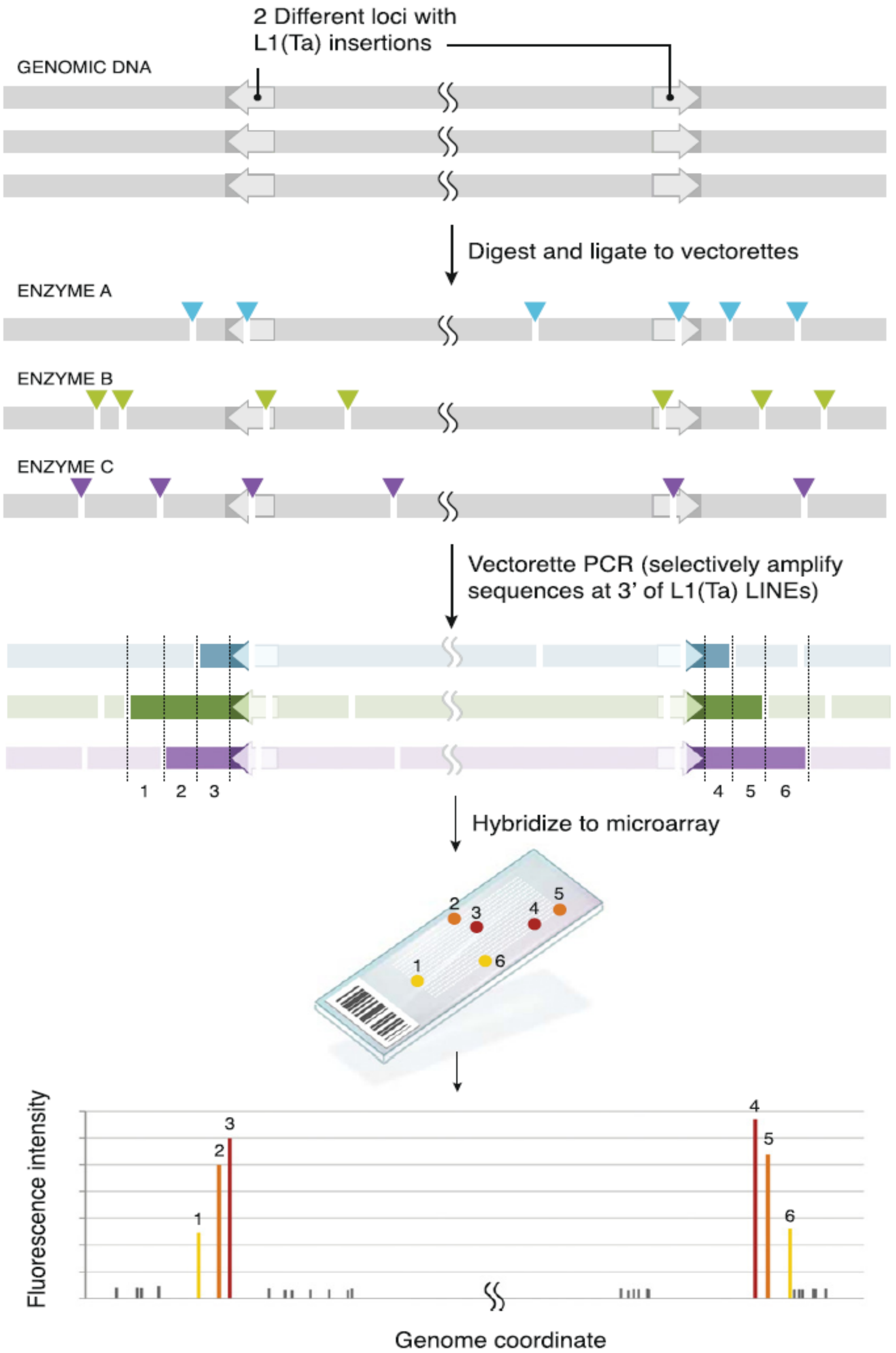


Figure 1. Transposon Insertion Profiling Chip Method

Human genomic DNA contains numerous L1(Ta) insertions (arrows 5′ → 3′); minus (left) and plus strand (right) insertion are illustrated here. Multiple copies of genomic DNA are digested in parallel with different REs (colored arrows, sites; each color is a different RE), and vectorette linkers (data not shown) are ligated to fragments. Vectorette PCR then specifically amplifies 3′ L1(Ta) sequence and unique genomic sequence 3′ of the L1(Ta) insertions (resulting amplicons are denoted by colored fragments). The cuts create a series of variable-length PCR templates for each L1(Ta) insertion. Genomic DNA fragments lacking L1(Ta) insertions are not amplified. Amplicons are labeled and hybridized to genomic tiling microarrays, generating peaks of signal intensity at probes (1–6) corresponding to genomic locations immediately adjacent to L1(Ta) insertions. For each peak, probes closest to L1(Ta) have highest fluorescence intensity with a gradient of diminishing signal seen downstream of the insertion because proximal probes are represented in more PCR products and shorter PCR products including them amplify more efficiently. Thus, slope of the signal gradient (±) opposes insertion orientation.
See also Figure S1.

The youngest L1 family – transcribed L1, subset a L1(Ta)

Peak recognition (1/2)

By HMM based L1 Signal Analysis software (Huang et al, in preparation)

Quality control measurement - insertions in the hs_ref genome

Peaks are ranked by the sum of posterior probability of each probe being in a peak

Peaks were removed

(i) after the i-th number of reference peaks in the ranked list

(ii) if the region showed ‘noisy’ background (variance = j)

(iii) if the peak was made up of less than k number of consecutive probes (allowing 1 failed probe within the peak interval), and

(iv) if local background intensity (defined by a 40 probe window flanking the peak) was above threshold m.

Peak recognition (2/2)

Peaks were reranked based on maximum probe intensity

Peaks below the last reference peak are deleted

Cutoff values for each variable (4) were imposed to target a total peak number closest to the expected number of L1(Ta) insertion positions per diploid human, $N_e = 515$, while removing the fewest reference L1(Ta) peaks.

Reference L1(Ta)s that did not make the cutoff (on average < 12% per sample) are retained in the final list

1. Average allele frequency $\overline{F}_i$ in single individual

Average allele frequency for L1(Ta) is assumed to be constant and invariant among genomes

Allele frequencies of 161 candidate novel L1(Ta) insertions found by chromosome X TIP-chip were defined based on 75 male samples profiled (number of TIP-chip peaks found at that genomic location divided by 75)

The average allele frequency for each individuals is determined by averinging the allele frequencies for each insertion on their X-chromosome (the mean of $\overline{F}_i$ is **0.75**)

2. Expected number of different L1(Ta) alleles in diploid individual (N_e)

Total L1(Ta) number does not vary significantly between individuals

In three sequenced haploid genome assemblies, the L1(Ta) counts are 413, 363, 460 (hs_ref, hs_alt_HuRef, hs_alt_celera)

Average of these values (412) is used as estimate of L1(Ta) insertions per haploid genome

The number of expected homozygous insertions is 412x0.75 = 309

The expected number of distinct L1(Ta) alleles per diploid human genome (N_e) is 412x2-309 = **515**

L1(Ta) discovery and inheritance patterns on the X chromosome

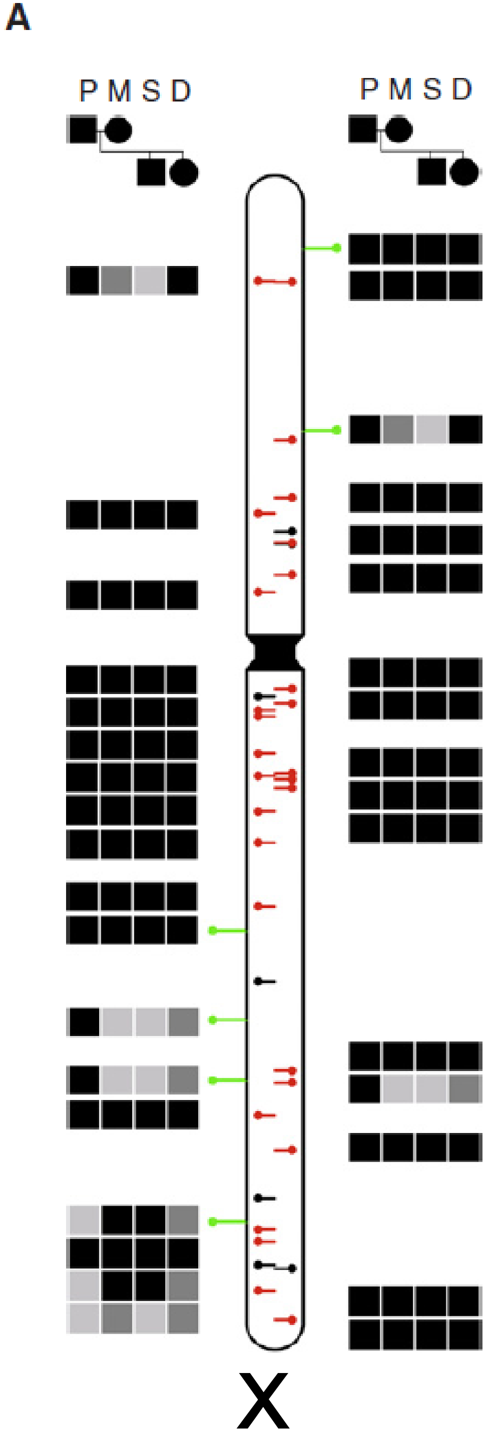


Figure 2. Inheritance Pattern of X Chromosome L1s

(A) L1(Ta) insertion profiles were generated for a family by TIP-chip using X chromosome microarrays. Presence (filled squares) or absence (empty squares) of peaks is indicated in paternal (P), maternal (M), son (S), and daughter (D) samples. Black or gray filled squares indicate an L1(Ta) detected at a specific site, as opposed to no fill; gray indicates inferred heterozygosity. Lollipops on the ideogram correspond to insertion coordinates. Black lines in center mark L1(Ta) incorporated in hs_ref NCBI Build 36.1. These are overlaid with red where observed. Green lines are PCR-verified novel insertions. Side represents insertion orientation (left = plus strand). In this family, 6 L1(Ta)s are paternal, nonmaternal; 4 are maternal, nonpaternal; and 4 additional maternal L1(Ta)s were not passed to her son, indicating maternal heterozygosity. Thus at least 33.33% of insertions found are polymorphic in this family.

38 known L1(Ta)
28 L1(Ta)s found by TIP-chip correspond to reference L1(Ta)s
84% (of 28) right orientation identified
6 previously unknown L1(Ta)s
34 L1(Ta)s seen in the family, 13 are polymorphic

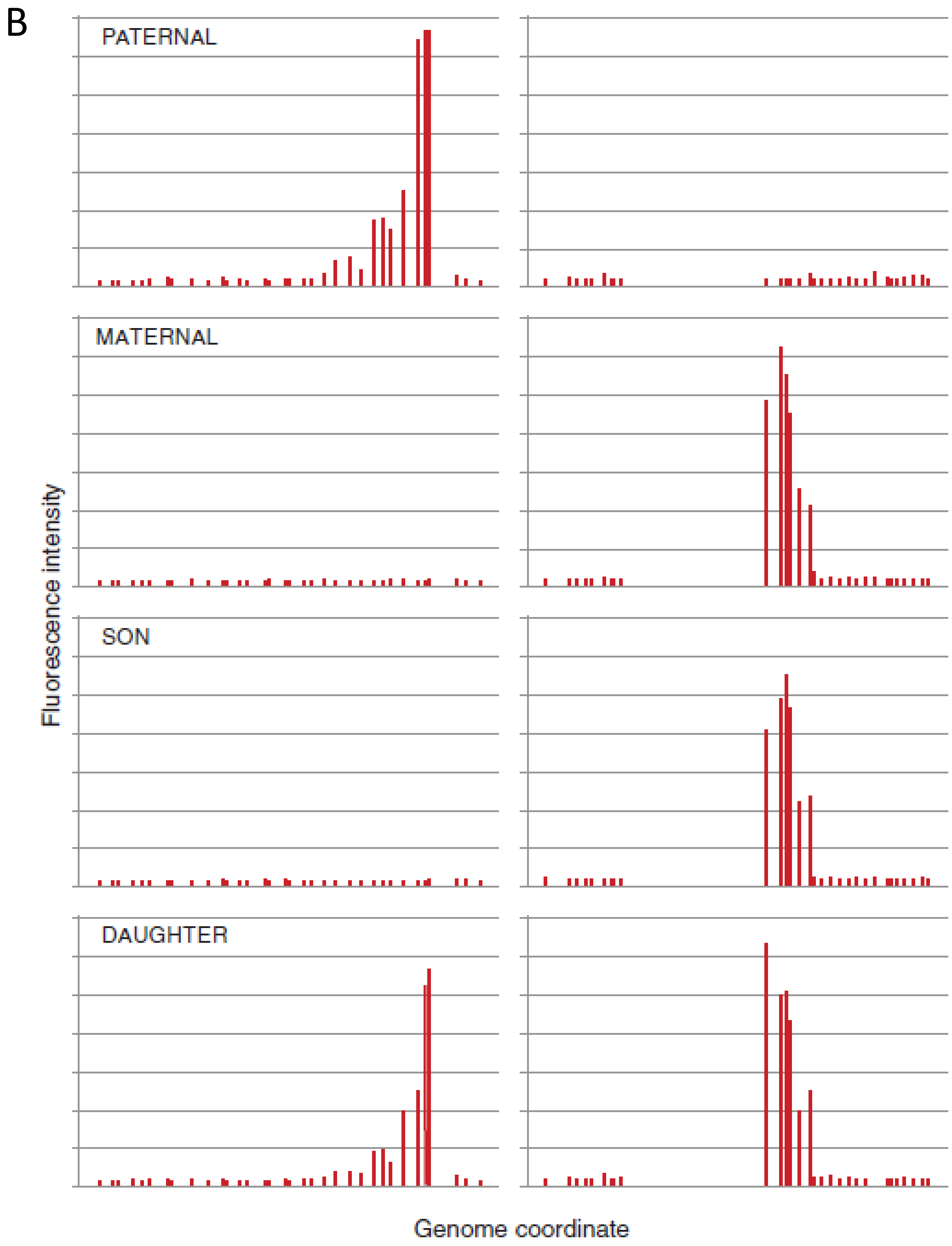


Figure 2. Inheritance Pattern of X Chromosome L1s

(B) Raw intensity data of two representative reference L1(Ta) insertions (one in each orientation) across four family members. x axis indicates genomic coordinate.
Probe fluorescence intensity is shown on y axis. Each bar represents one array probe.

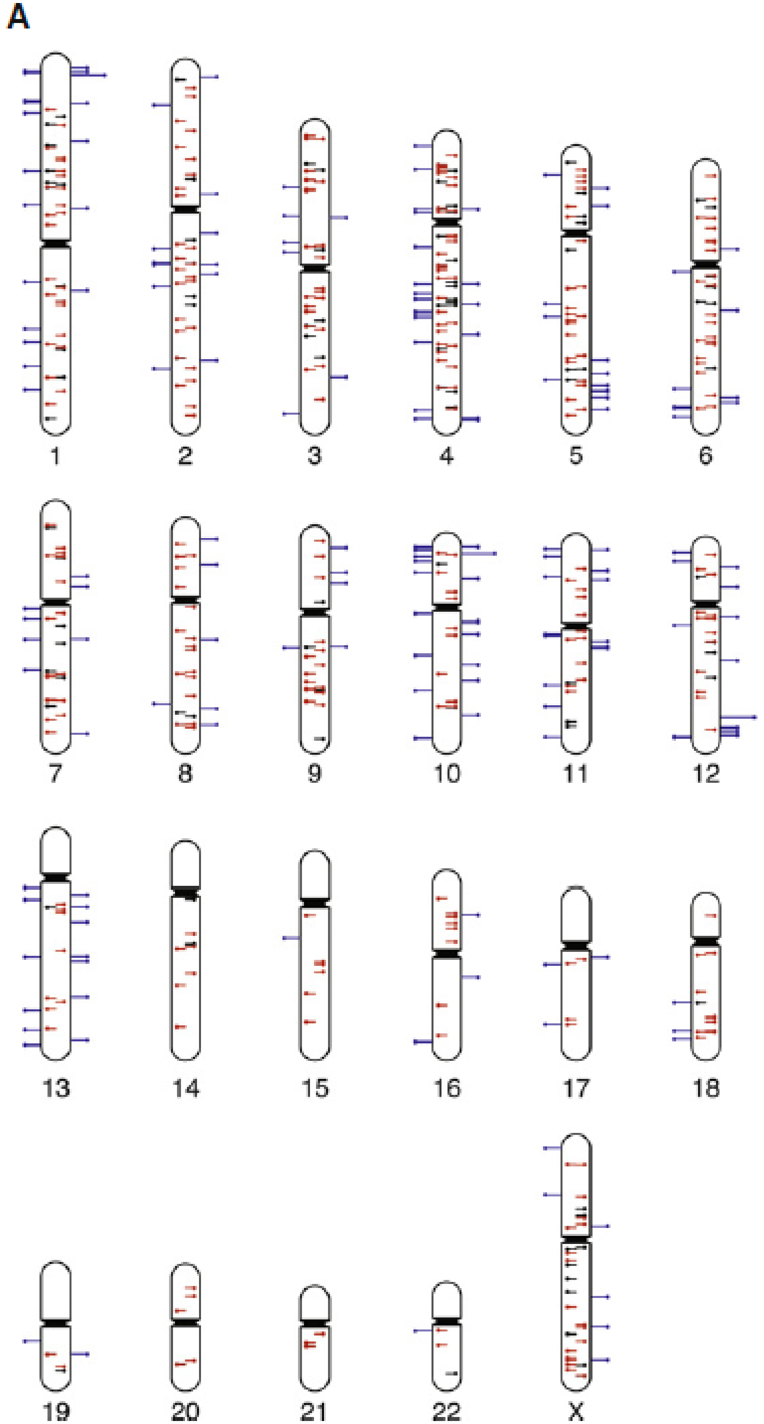


Figure 3. Genome-wide Mapping of L1(Ta) Insertions in an Individual

(A) Ideogram illustrates TIP-chip peaks in an individual; 514 peaks are included after imposing the cutoff (Experimental Procedures). Marks show predicted positions of L1(Ta) insertions on the plus (left side) and minus strands. Central lines similarly illustrate position and orientation of L1(Ta)s in the human reference sequence (hs_ref NCBI Build 36.1). These are color coded to indicate those identified by TIP-chip in this individual (red, n = 323) and those not seen in this sample (black). Blue lines on the outside of the chromosome correspond to nonreference insertions (n = 191). In addition to reference L1(Ta)s, 52 were considered true positives because they correspond to insertions included in dbRIP (n = 25) or were described by human sequencing projects (n = 24), as well as 3 by Beck and Moran (Beck et al., 2010). As described further in the text, additional TIP-chip peaks were verified by PCR and sequencing.

+56 true L1 insertions verified by PCR and sequencing

Overall assay positive predicitve value 84%

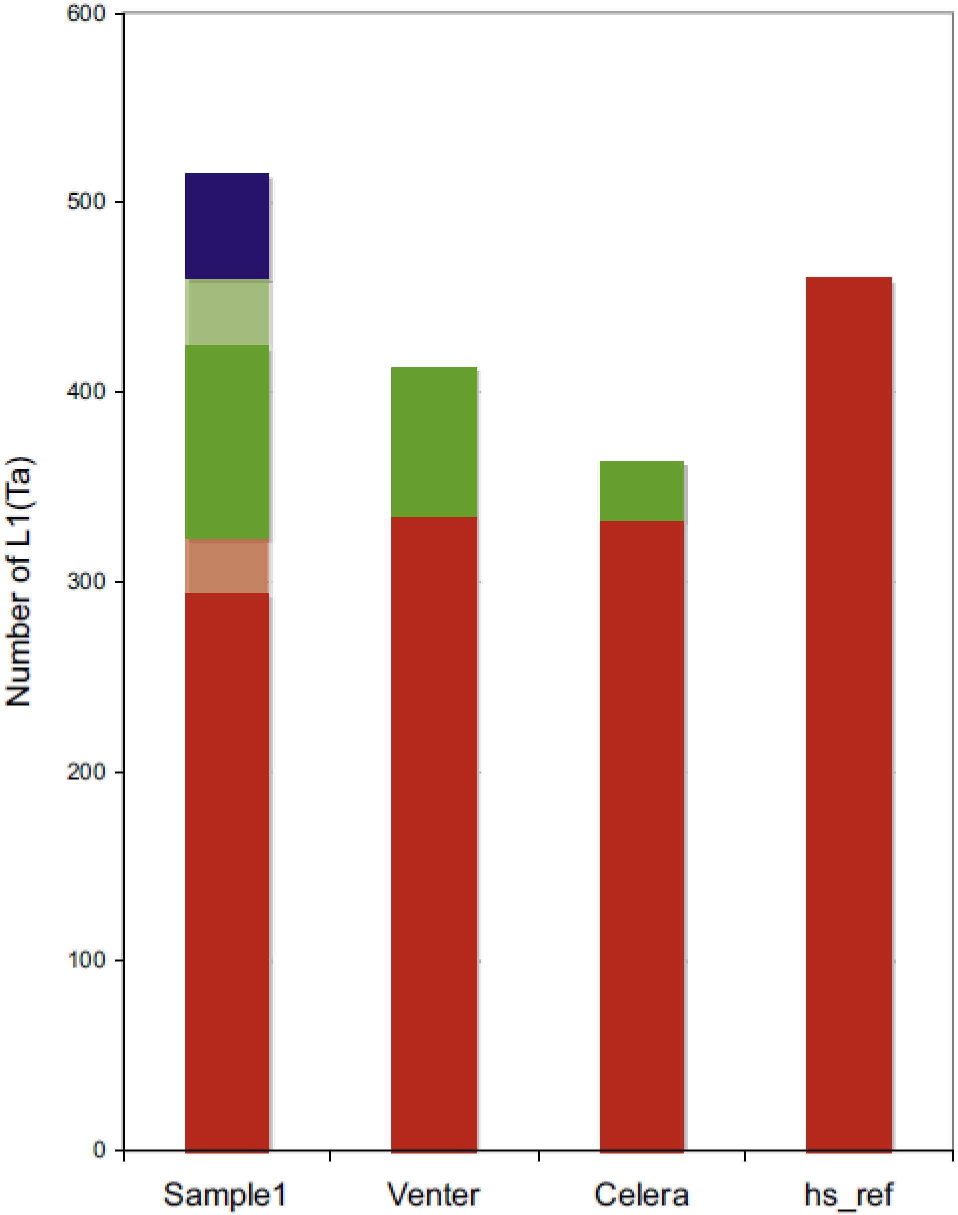


Figure 3. Genome-wide Mapping of L1(Ta) Insertions in an Individual

(B) TIP-chip and whole-genome sequencing in identifying L1(Ta) insertions. The y axis shows the L1(Ta) count in each sample. Sample1 was profiled by TIP-chip, whereas the other three samples are from different whole-genome sequencing approaches. Insertions present in hs_ref are displayed in red. Verified nonreference L1(Ta) insertions are shown in green. Lighter shades of red reflect reference insertions that were not retained after the imposed cutoff, while that of green reflects 30 PCR verified insertions that might not become sequence verified. Candidate novel L1(Ta) insertions identified by TIP-chip after the cutoff and awaiting further verification, are marked in blue. The ability of TIP-chip to identify L1(Ta) insertions is comparable to whole-genome sequencing.

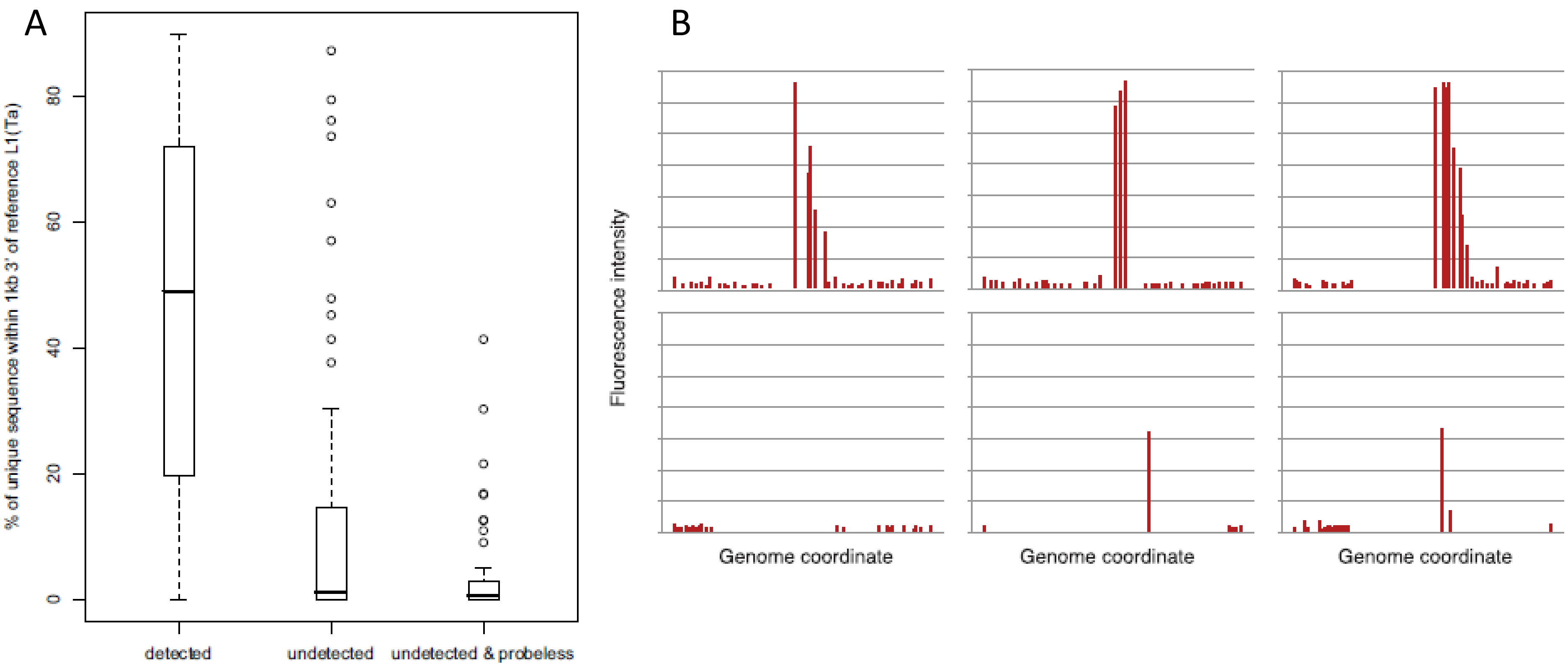


Figure S3. Related to Figure 4

(A) Unique sequences adjacent to reference L1(Ta)s. One kilobase regions 3′ of reference L1(Ta)s were analyzed for repetitive sequence composition by Repeatmasker; the unique, nonrepetitive percentage is shown on the y axis. These data are plotted for reference L1(Ta)s detected by TIP-chip, for those undetected by TIP-chip in 15 unrelated samples, for those undetected and in the 'probe poor' category (see Results section).

(B) Comparison of X chromosome array and whole-genome array. x axis indicates genomic coordinate. Probe fluorescence intensity is shown on y axis. Each bar represents one array probe. The first row is raw intensity data from the X chromosome array. The second row is raw intensity data from the corresponding section of the whole-genome array. These are examples of three reference L1(Ta) insertions detected on the X platform but not on the whole-genome array, making them false negatives. These panels demonstrate that missing probes in the whole-genome array at the area where peak forms on the X chromosome array data accounts for the reason for the false negative and suggests that improvement can be easily made by revising the array design.

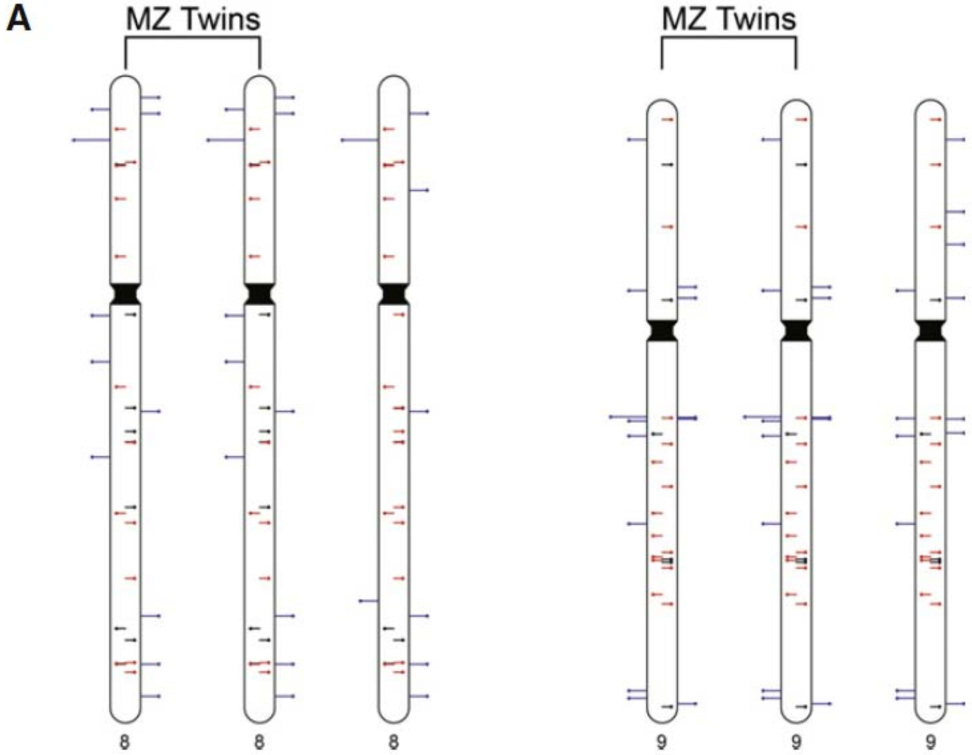


Figure 4. **High Reproducibility of Whole-Genome TIP-chip**

(A) Ideogram illustrating TIP-chip peaks on chromosomes 8 and 9 in a monozygotic twin pair and an unrelated individual. Marks on chromosomes show predicted positions of L1(Ta) insertions on the plus (left side) and minus strands. Central lines similarly illustrate position and orientation of L1(Ta)s in hs_ref. These are color-coded to indicate L1(Ta)s identified by TIP-chip in these individuals (red) and those not seen in this sample (black). Blue lines on the outside of the chromosome correspond to candidate nonreference L1(Ta)s. When our automated peak identification program is complemented by visual inspection of the raw data, twins have identical peak patterns while displaying many polymorphisms as compared to the unrelated individual (right most).

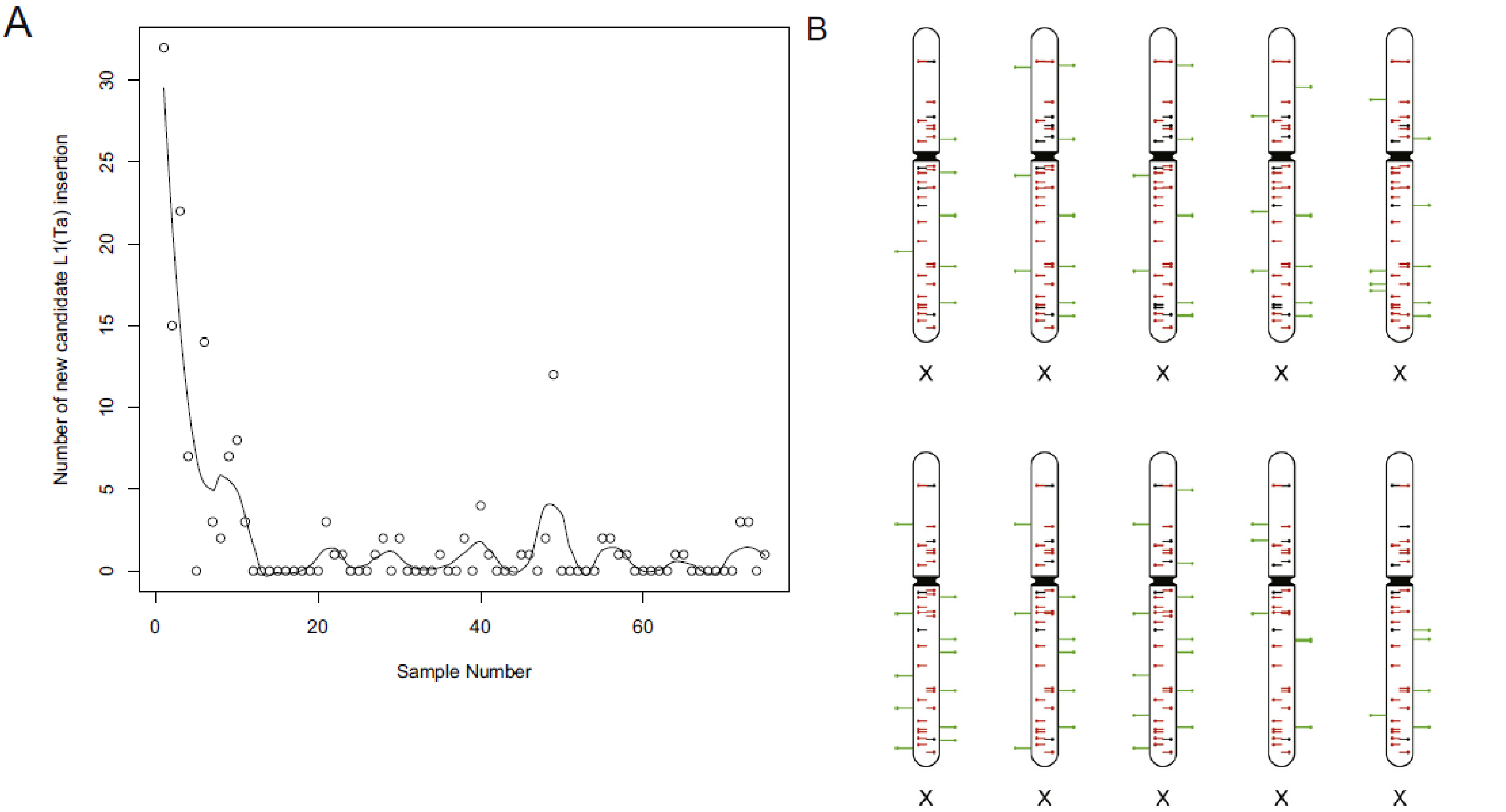


Figure S4. Novel X Chromosome Insertions, Related to Figure 5

(A) Discovery rate of candidate L1(Ta)s on X chromosome. Seventy-five unrelated male samples are included in this analysis. Samples are arranged sequentially in the order profiled by TIP-chip (x axis). The number of new (i.e., not detected in preceding samples) candidate L1(Ta) insertions found in each sample is plotted on the y axis.

(B) Ideograms of ten X chromosomes for which novel insertions were validated. See Figure 4A legend. This shows the high degree of variation in L1(Ta) insertion profile in different individuals.

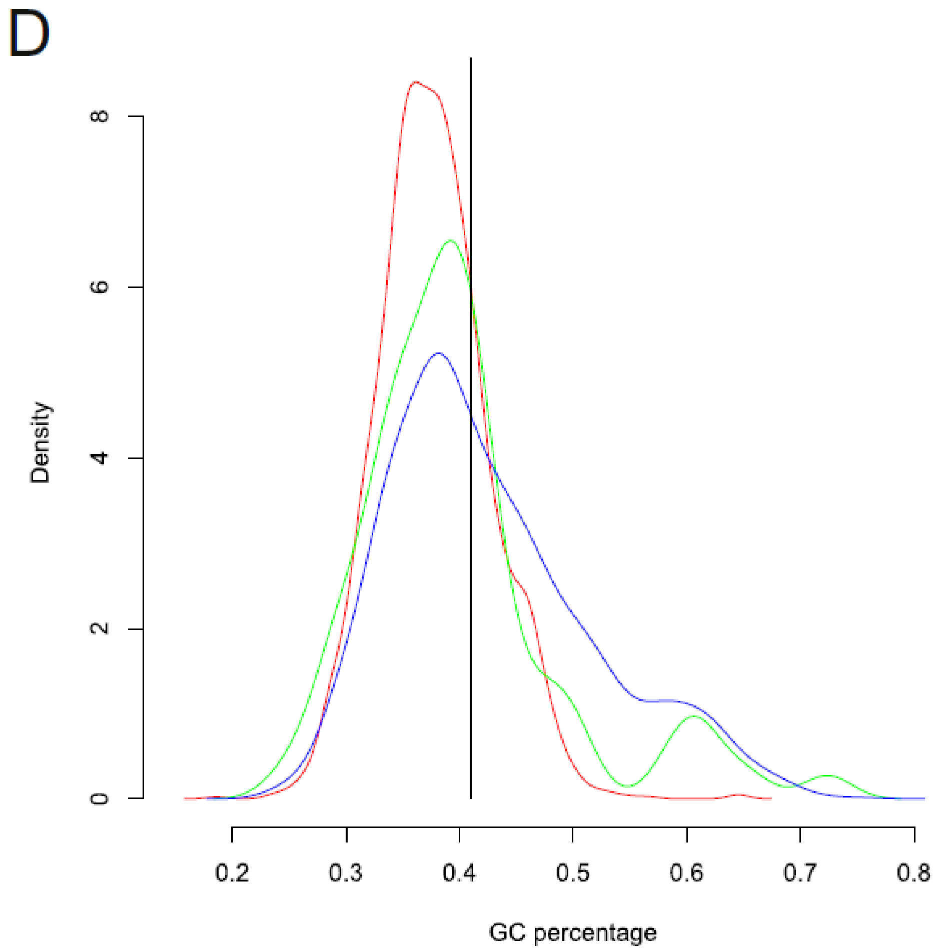


Figure S5. L1(Ta) Genome-wide Analyses per Megabase (MB) DNA in Different Samples (n = 10) over Different Chromosomes

(D) GC content of 1000bp flanking L1(Ta) insertions. Probability density function of GC percentage in the 1 kb flanking each L1(Ta) insertion site is plotted here (500 bp 5′ and 500bp 3′ of the insertion site). The area under each curve is equal. (Red curve, insertion sites of reference L1(Ta); green curve, PCR verified nonreference L1(Ta) insertions; blue curve, nonreference L1(Ta) maximal probe peak positions on TIP-chip.) Note that sequenced insertions have higher (base-pair) precision whereas PCR-verified and TIP-chip mapped insertions are less precise. Both green and blue curves behave similarly to the reference insertion curve, showing preferential accumulation of L1(Ta)s in AT-rich regions. The average GC content in the human genome is 41%, denoted by the vertical line.

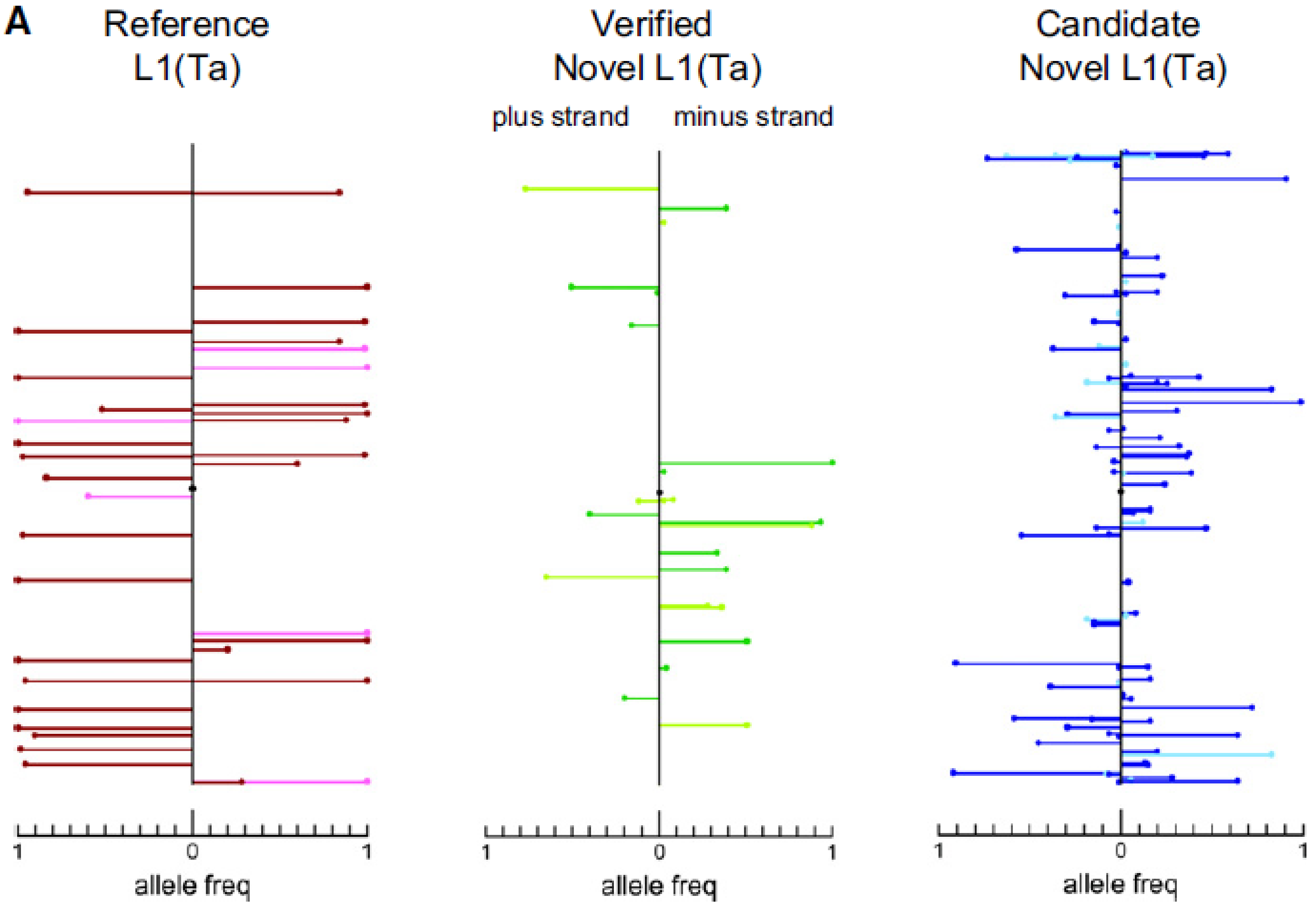


Figure 5. Polymorphism of X chromosome L1(Ta)s (A) Each mark represents a L1(Ta) insertion. y axis denotes position along the X chromosome and the x axis reflects allele frequencies for L1(Ta) insertions on the plus (left) and minus strands (i.e., % of males with respective insertion). In total, 75 unrelated clinical male samples collected in the United States were included in this analysis; samples were not selected based on ethnic background. As a generalization, L1(Ta)s included in hs_ref (reference L1(Ta)s, red; leftmost panel) had higher allele frequencies (0.896 ± 0.202) than novel L1(Ta)s identified (0.263 ± 0.266, green and blue for PCR verified and not yet verified, respectively, see Table S2). No significant difference in allele frequencies were observed comparing intergenic L1(Ta)s (darker hue) with intronic/intragenic insertions (lighter hue).

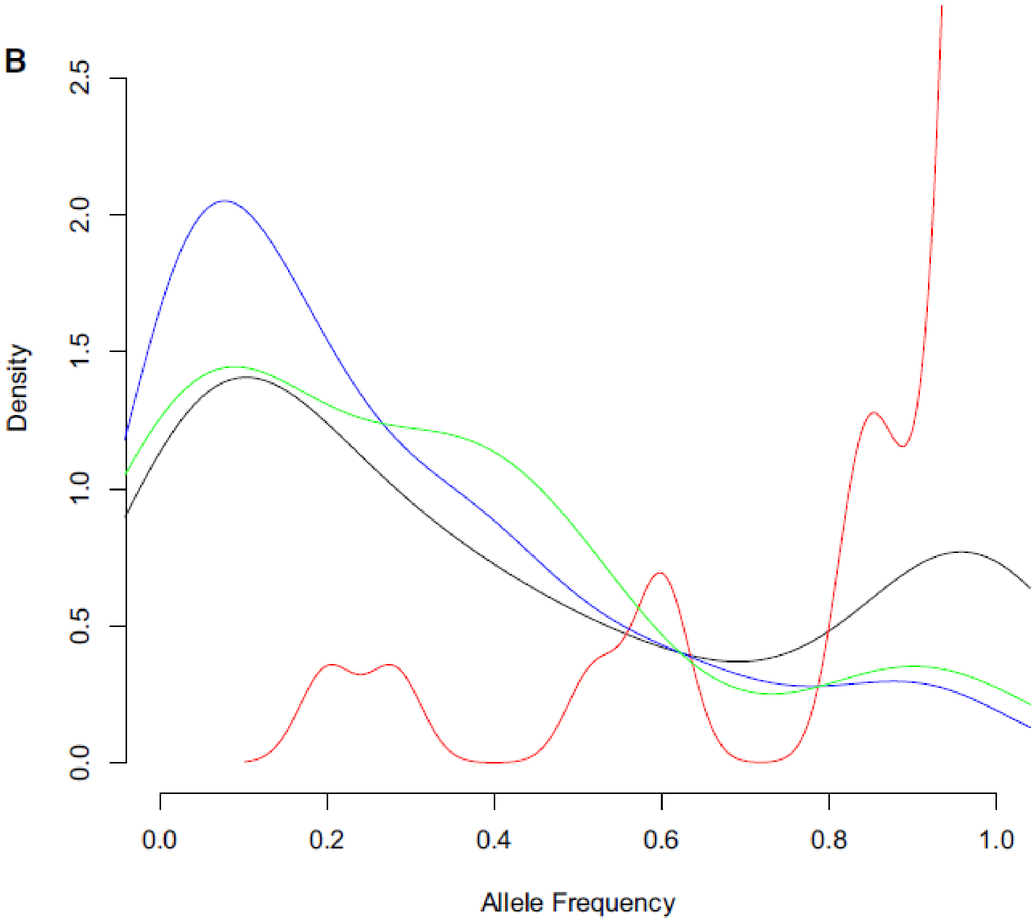


Figure 5. Polymorphism of X chromosome L1(Ta)s
(B) Probability density function of allele frequencies of L1(Ta) insertions on the X chromosome. The area under each curve equals one. The x axis denotes the allele frequency ranging from 0 to 1 (present in all samples tested). Allele frequencies are calculated using X chromosome TIP-chip profiles of 75 unrelated males. The red curve shows the probability density function for insertions in hs_ref. The green curve depicts verified insertions. The blue curve displays TIP-chip peaks not yet verified. Black indicates the combined total of all three classes described above.

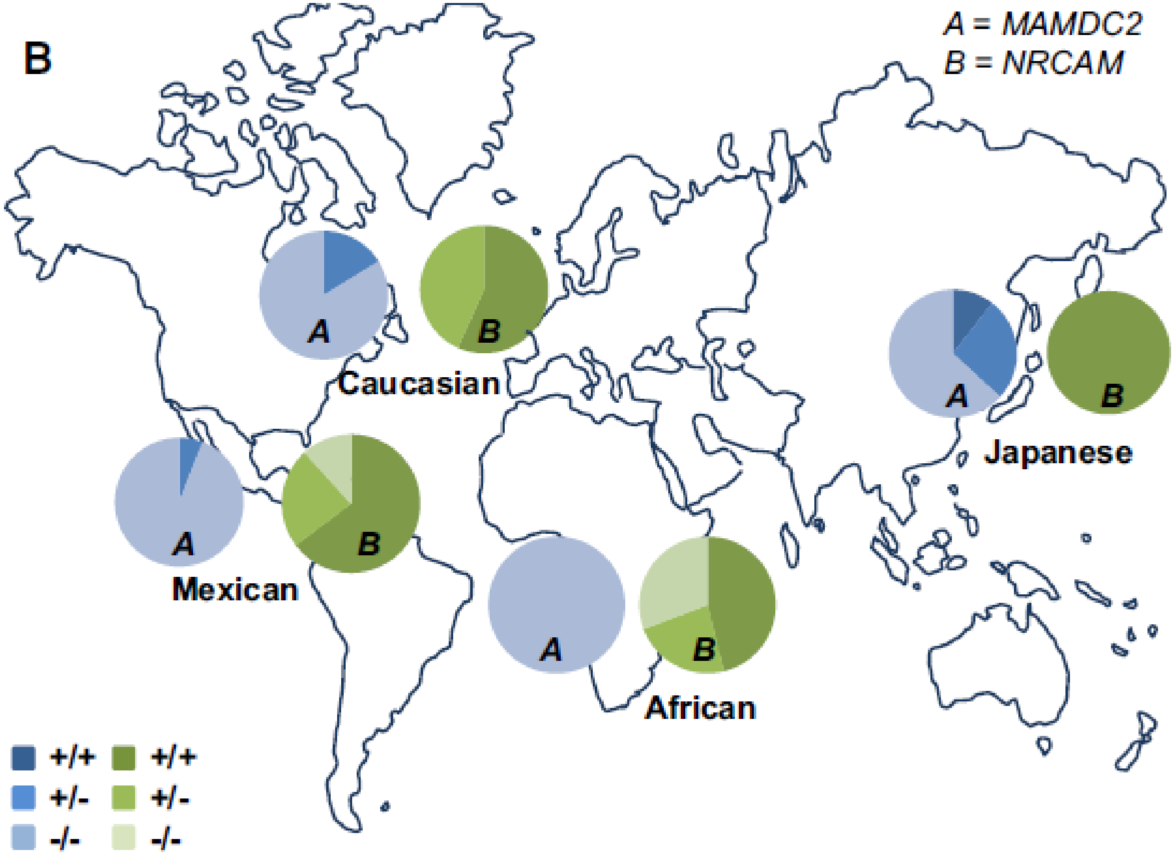


Figure 6. Polymorphism of L1(Ta)s

(B) Pie charts indicate genotype distribution for two representative nonreference L1(Ta)s (not included in hs_ref) identified by TIP-chip studies of an individual (see Figure 3) across two human ethnic diversity panels. The total sample size of both diversity panels is 198 people. The Caucasian, Mexican and Japanese sample groups were represented most highly (n = 37, 17 and 18 respectively) and were used for Hardy-Weinberg calculations. For Locus A (*MAMDC2*) the allele frequencies for each population, as well as the chi square values for the biggest population groups are as follows: Caucasians (0.08; $\chi^2 = 0.29$); Mexican (0.03; $\chi^2 = 0.02$); Japanese (0.25; $\chi^2 = 1.41$); African (0.00, n = 13). For Locus B (*NRCAM*) the allele frequencies for each population, as well as the chi-square values for the biggest population groups are as follows: Caucasians (0.79; $\chi^2 = 2.82$); Mexican (0.77; $\chi^2 = 2.04$); Japanese (1.00); African (0.58, n = 13).

Conclusions

1. The reference genome assembly with respect to inserted sequence variations (ISV) is incomplete
2. The small quantity and low allele frequency of many novel L1(Ta)s suggest thst they remain highly active in modern humans

3. TIP-chip represents the high-throughput method for mapping retroelement insertions
4. TIP-chip enables to discover novel L1(Ta) insertions
5. TIP-chip is fast and cost effective
6. TIP-chip detects many types of ISVs
7. With TIP-chip, insertions in repetitive regions are difficult or impossible to map