SUPPLEMENTAL INFORMATION INVENTORY

SUPPLEMENTAL FIGURES

Figure S1, related to Figure 1
Figure S2, related to Figure 3
Figure S3, related to Figure 4
Figure S4, related to Figure 5
Figure S5, related to Figure 6
Figure S6, related to Figure 7
Table S1, related to Figure 3
Table S2, related to Figure 5
Table S3, related to Figure 6
Table S4, related to Figure 7

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

SUPPLEMENTAL REFERENCES

Supplement to
Huang et al. 2010; "TIP-chip reveals that interspersed repeats are major structural variants in the human genome"

Supplemental Figures

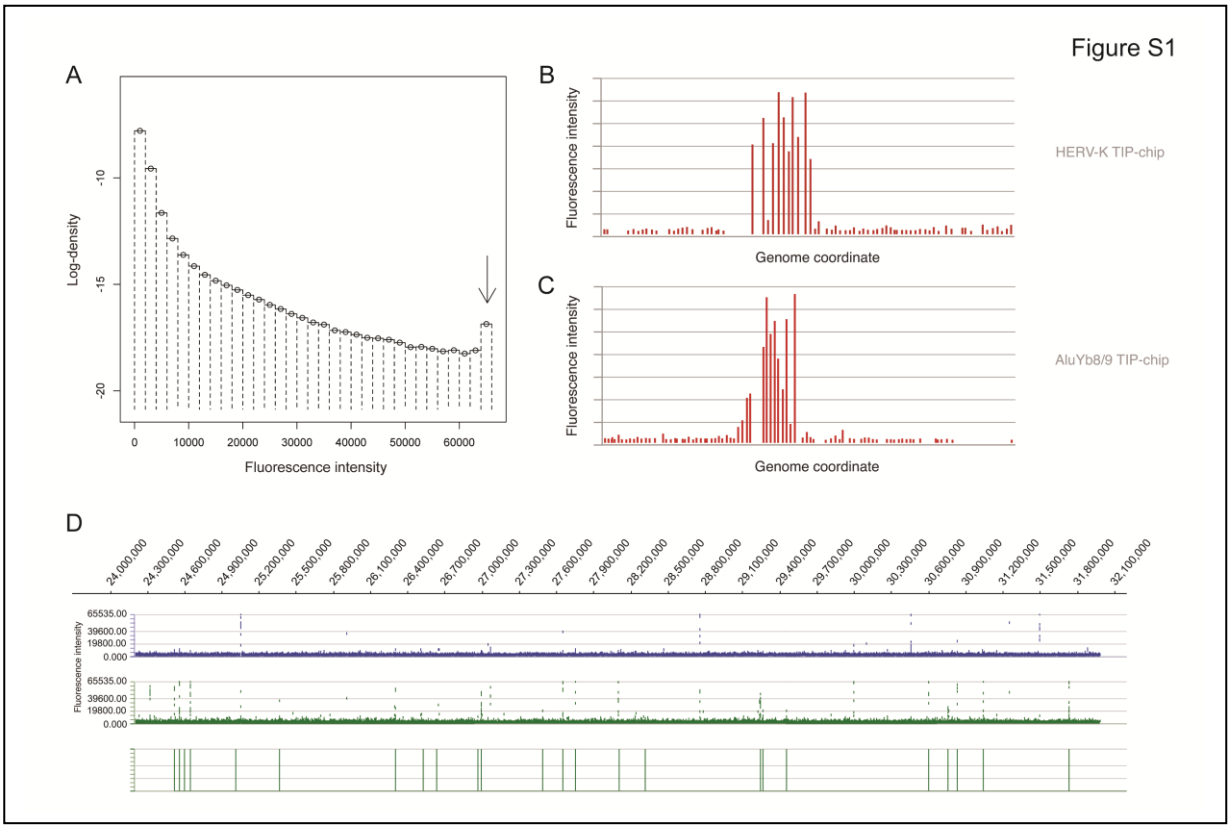

Figure S1, related to Figure 1: A. Log scale histogram of fluorescence intensity on TIP-chip. Log-density values are plotted against fluorescence intensities. The histogram is biphasic. The arrow shows high intensity values that comprise TIP-chip peaks. **B.** Raw intensity data of one reference HERV-K LTR insertion. X axis indicates genomic coordinate. Probe fluorescence intensity is shown on Y axis. Each bar represents one array probe. **C.** Raw intensity data of one reference *Alu*Yb8/9 insertion. Axes are as in **B**. **D.** Detection of L1(Ta) and *Alu*Ya5/8 and Yb8/9. The X axis displayed on top is the genomic coordinate. The Y axis for the top two tracks shows probe fluorescence intensity values. Each dot in the top two tracks represents one array probe. The top track displays raw intensity data from a L1(Ta) TIP-Chip. The middle track displays raw intensity data from an *Alu* TIP-Chip. The bottom tracks marks the location of *Alu* reference insertions. Multiple high intensity probes compose each peak. A fraction of the peaks on the *Alu*Yb TIP-Chip line up with positions of the reference insertions. The remaining peaks denote candidate novel insertions.

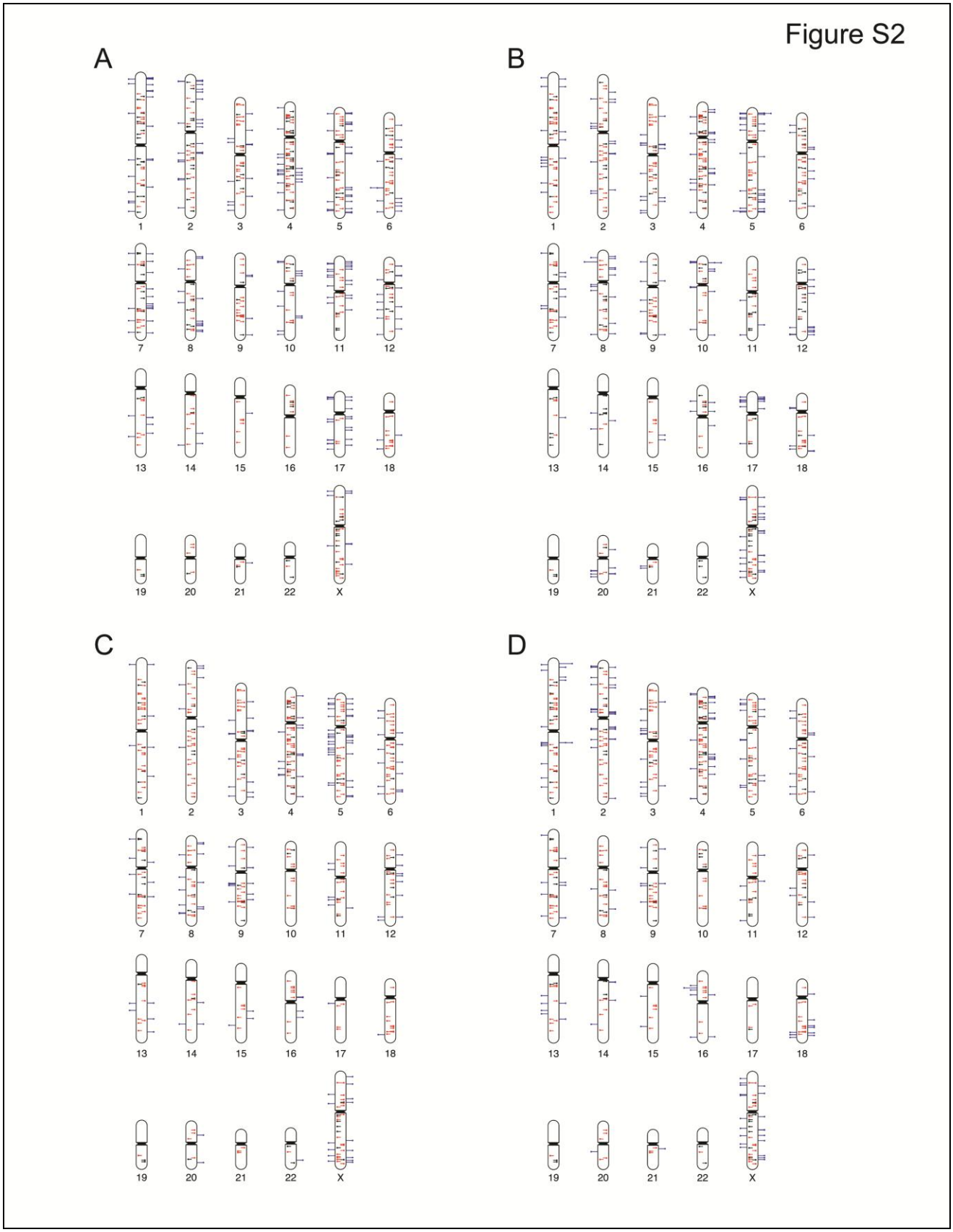

Figure S2, related to Figure 3: Whole genome L1(Ta) profiles of 4 individuals.

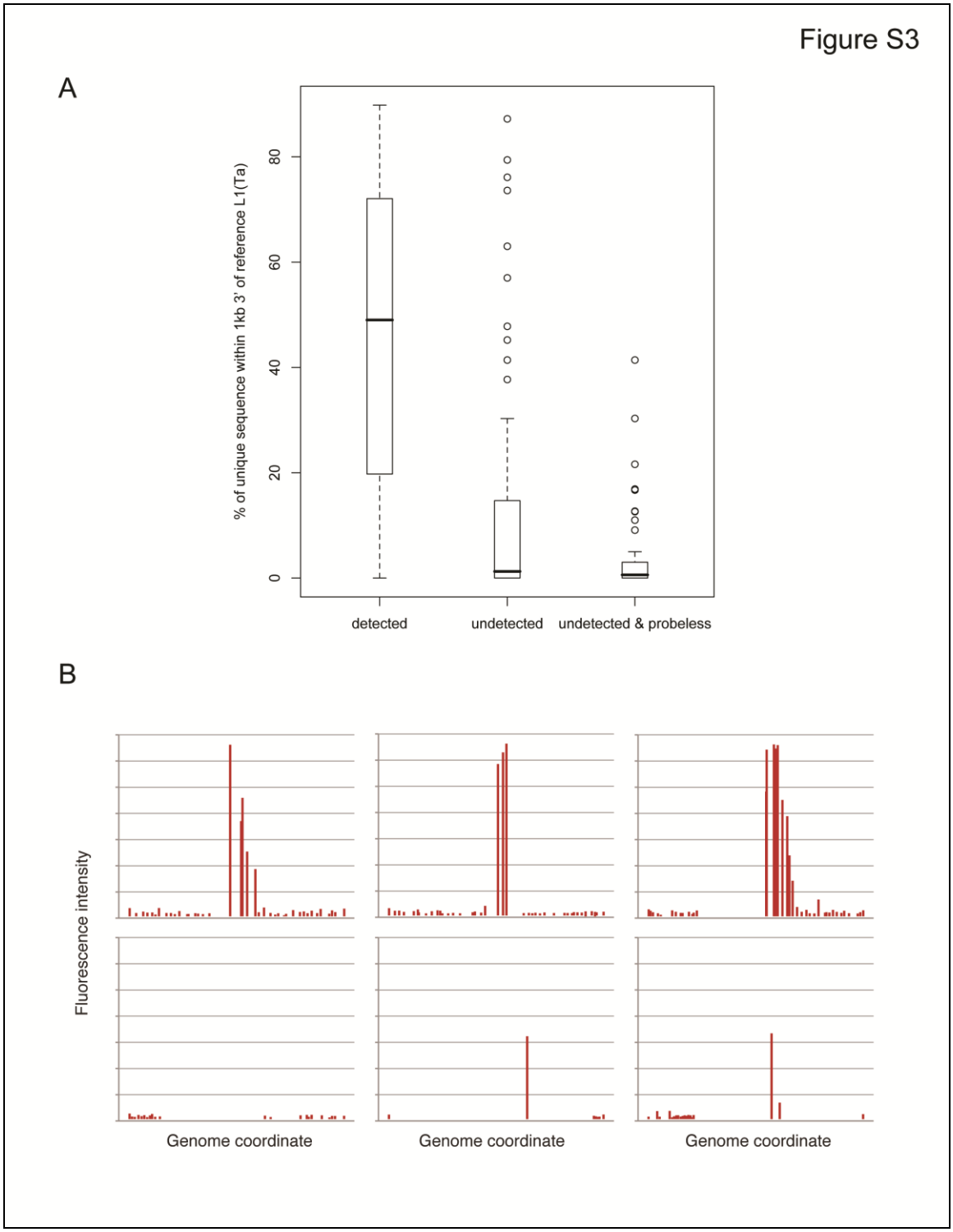

Figure S3, related to Figrure 4: A. Unique sequences adjacent to reference L1(Ta)s. One kilobase regions 3' of reference L1(Ta)s were analyzed for repetitive sequence composition by Repeatmasker; the unique, non-repetitive percentage is shown on the Y-axis. These data are plotted for reference L1(Ta)s 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.

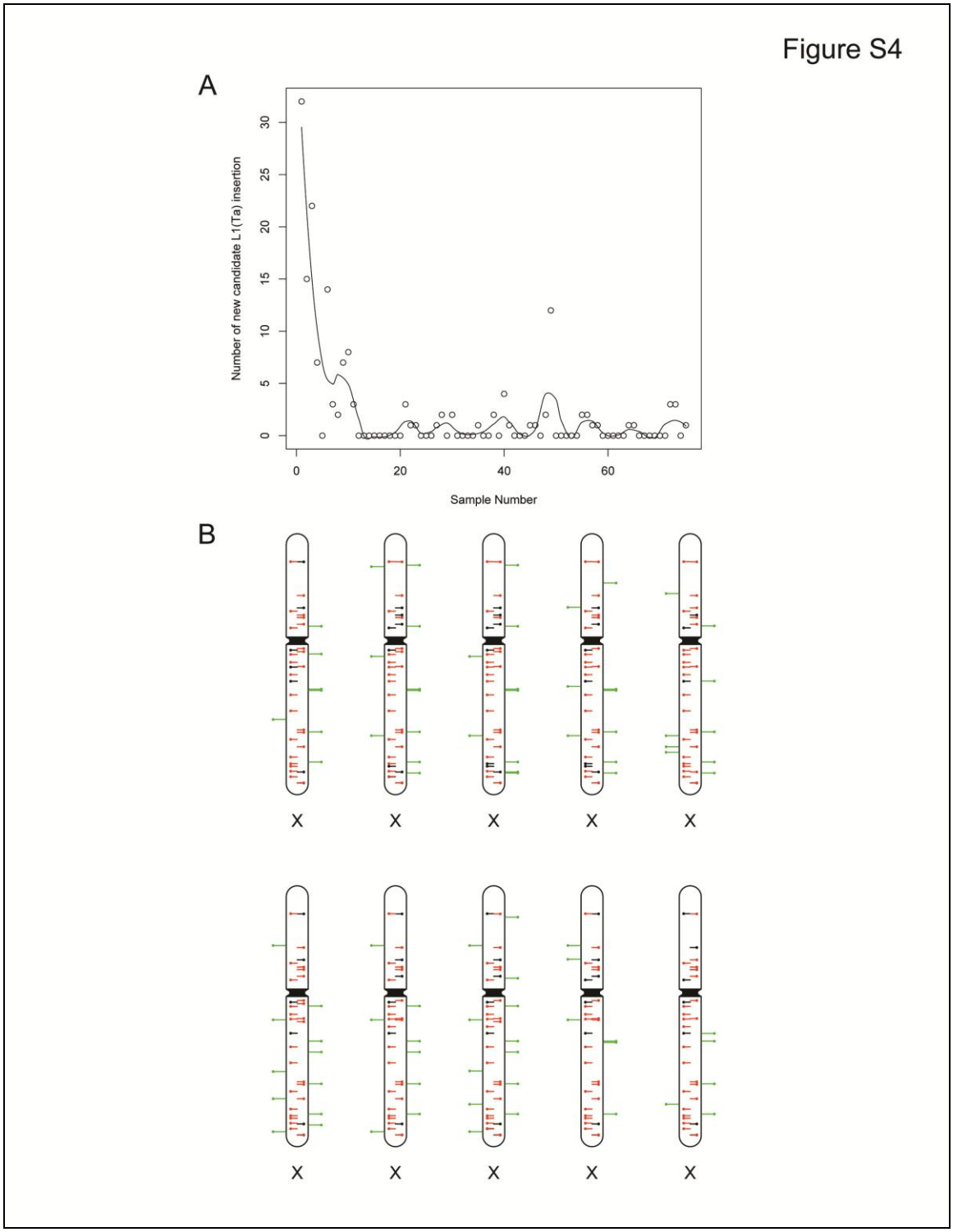

Figure S4, related to Figure 5: Novel X chromosome insertions A. Discovery rate of candidate L1(Ta)s on X chromosome. Seventy-five unrelated male samples are included in this analysis. Samples are arranged sequentially in the order profiled by TIP-chip (X-axis). The number of new (i.e., not detected 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.

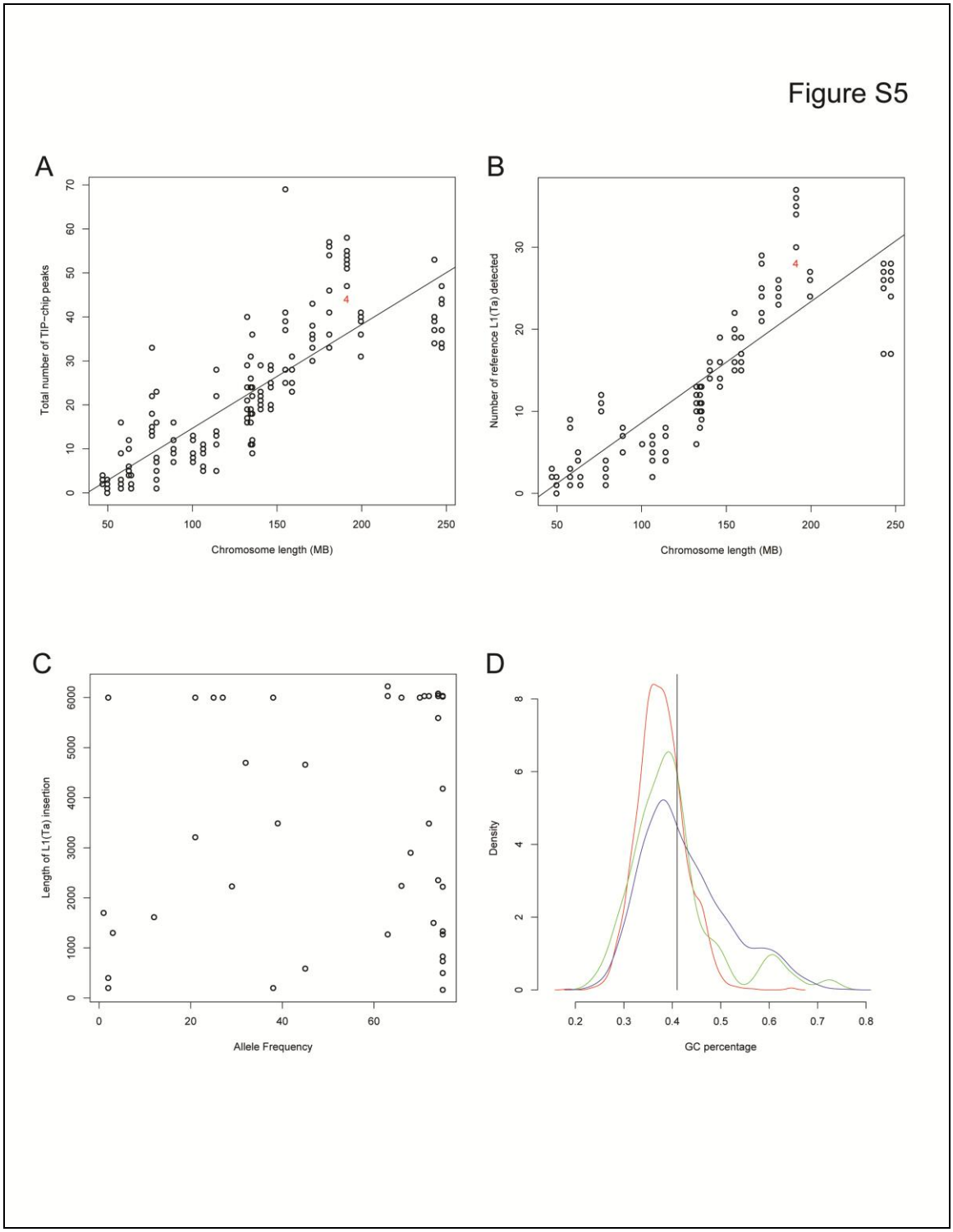

Figure S5, related to Figure 6: L1(Ta) genome wide analyses per megabase (MB) DNA in different samples (n=10) over different chromosomes. A. L1(Ta)s per megabase (MB) DNA in various samples (n=10) on different chromosomes. The position on the X-axis shows chromosome length in MB, and chromosome numbers are indicated. The Y-axis displays total number of TIP-chip peaks detected on each chromosome. The fitted line depicts the relationship between the number of peaks and chromosome length. **B.** The same display for L1(Ta)s as **A.** in the reference genome (hs_ref). Chromosome 4 is an outlier, with significantly more reference peaks (F-ratio: 119.5 degrees of freedom=213, $p<0.001$) than other chromosomes. **C.** Length of L1(Ta) verses allelic frequencies. The length of each L1(Ta) insertion is plotted against its allele frequency for 49 insertions; 75 unrelated male

samples were used for allele frequency determination. No statistically significance correlation was observed between retrotransposon length and allele frequency. **D**. GC content of 1000bp flanking L1(Ta) insertions. Probability density function of GC percentage in the 1kb flanking each L1(Ta) insertion site is plotted here (500bp 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 non-reference L1(Ta) insertions; blue curve, non-reference 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.

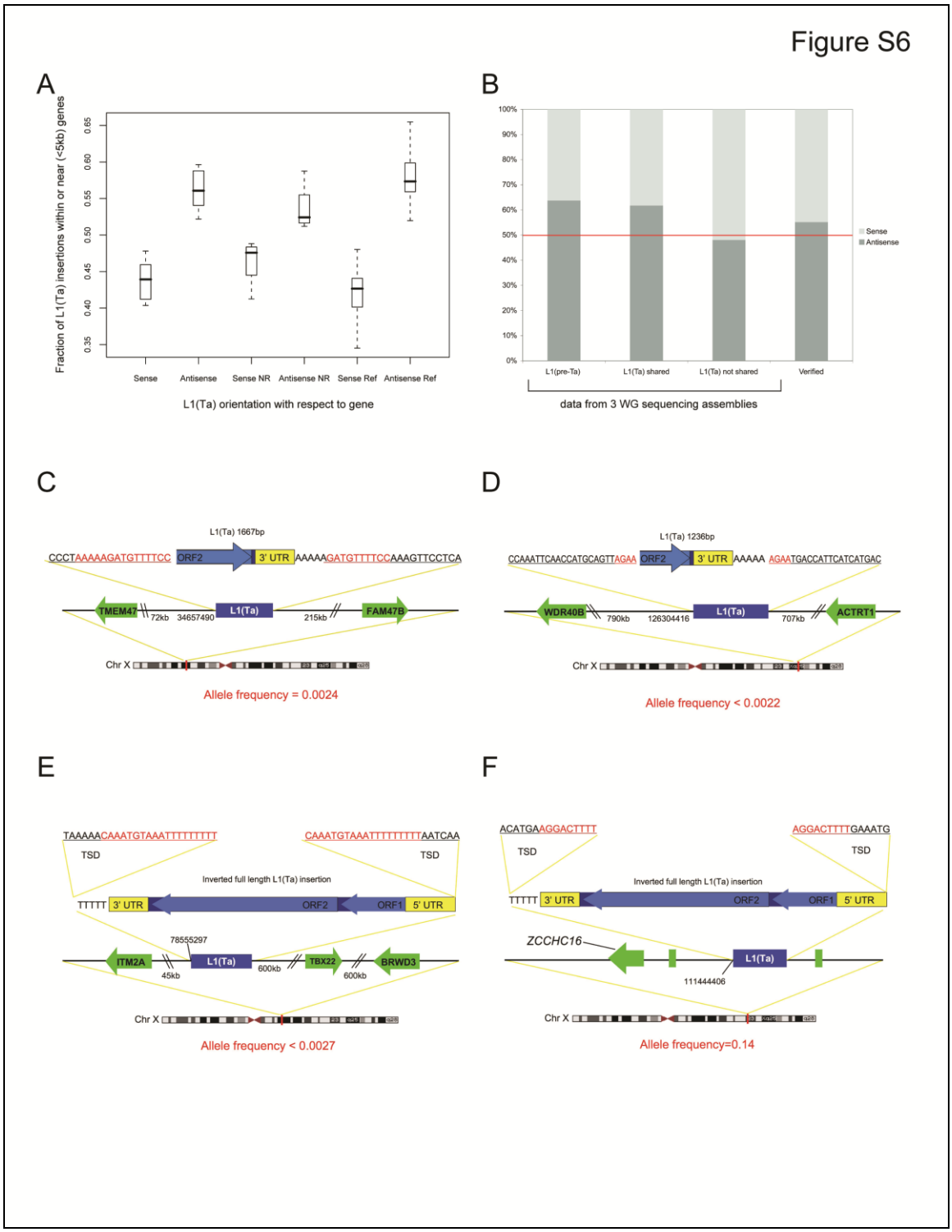

Figure S6, related to Figure 7: Insertions in genes and clinical samples. A. Box plot of the fraction of L1(Ta) insertions, within or near (<5kb) genes, oriented in the sense or antisense direction relative to the host gene is plotted here. The data are further divided into all candidate novel (left pair), non-reference (NR, middle pair) or reference (right pair) L1(Ta) insertions found in TIP-chip profile of 10 different samples. **B**. The percentage of L1(Ta) insertions, within or near (<5kb) genes, oriented in the sense or antisense direction relative to the host gene is plotted. Darker shading shows percentage of antisense insertions and lighter shade marks that of sense insertions. The 50% expected for no orientation bias is denoted by the red line. Data displayed represent four different categories: “L1(pre-Ta)”, “L1(Ta) shared”, “L1(Ta) not shared” and “Verified”. The first three bars are data generated from the three publicly available whole genome sequencing assemblies. Insertions present in all three

assemblies are classified as “shared”, while those present in one or two assemblies are marked as “not shared”. “Shared” insertions tend to have a higher allele frequency, representing older insertion events as a group, whereas “not shared” insertions will be more polymorphic and younger in age. Similarly, L1(pre-Ta) is a older group of L1s, expected to have experienced more selection time. The right most bar represents novel, verified L1(Ta) insertions from TIP-chip. Along with the “L1(Ta) not shared” group, these are likely the youngest subset, and the relatively limited exposure to selection is reflected in less orientation bias. **C-F.** Four infrequent L1(Ta) insertions identified in X-linked intellectual disability patients. See **Figure 7** legend.

Supplemental Tables

Tables S1 and S2 are more than two pages long and are therefore provided as Excel files

Table S1, related to Figure 3. The Tab "Hu_Ref NCBI_Assembly" represents L1(Ta) insertions already in Hu_Ref, some of which are newly reported here (the ones without a "Y" in column D.) The Tab "Hu_Ref TIP-Chip_Seq Ver" indicates sequence verified insertions not in the sequence assembly. The Tab "HuRef TIP-Chip PCR Ver" represents 25 insertions newly found by TIP-chip in Hu_Ref DNA, and the primers used for spanning PCR confirmation. Coordinates given are of the first nonA residue following the 3' A tail of each sequence. The Tab "Sample 1 TIP-ChIP 3'" lists all verified novel insertions discovered by TIP-chip and validated by 3' junction PCRs with the primer indicated. The Tab "Sample 1 TIP-chip spanning" indicates all verified novel insertions discovered by TIP-chip and validated by spanning PCRs with the primer pairs indicated. The Tab "Other Seq verified insertions" lists all sequence verified insertions discovered in this study.

Table S2, related to Figure 5. This table lists the allele frequencies for the insertions found on the X chromosomes of 75 males.

Population	Sample size	XMR32				XMR42			
		+/+	+/-	-/-	Allele fr.	+/+	+/-	-/-	Allele fr.
All	96	1	0	0	0.01	1	0	0	0.01
African Americans	14	0	0	0	0.00	0	0	0	0.00
Biaka	4	0	0	0	0.00	0	0	0	0.00
Caucasian	26	1	0	0	0.04	1	0	0	0.04
Chinese	4	0	0	0	0.00	0	0	0	0.00
Druze	5	0	0	0	0.00	0	0	0	0.00
Indo Pakistani	3	0	0	0	0.00	0	0	0	0.00
Mbuti	5	0	0	0	0.00	0	0	0	0.00
Mexican	17	0	0	0	0.00	0	0	0	0.00
Puerto Rican	8	0	0	0	0.00	0	0	0	0.00
Krasnodar	3	0	0	0	0.00	0	0	0	0.00
South American	7	0	0	0	0.00	0	0	0	0.00

Population	Sample size	CADM2				CTNND2				C4				C8				NRCAM				BC048982				MAMDC2				F8			
		+/+	+/-	-/-	Allele fr.	+/+	+/-	-/-	Allele fr.	+/+	+/-	-/-	Allele fr.	+/+	+/-	-/-	Allele fr.	+/+	+/-	-/-	Allele fr.	+/+	+/-	-/-	Allele fr.	+/+	+/-	-/-	Allele fr.	+/+	+/-	-/-	Allele fr.
North America	75.2 ± 1.4	11	22	41	0.29	0	76	0	0.51	1	18	56	0.13	15	28	32	0.39	42	31	3	0.76	19	28	29	0.44	0	9	67	0.06	4	63	6	0.47
African Americans	13,8 ± 0,2	0	4	9	0.14	0	14	0	0.50	0	5	9	0.18	3	4	7	0.36	8	6	0	0.79	5	4	5	0.50	0	0	14	0.00	0	13	0	0.46
Caucasian	36.7 ± 0.5	9	10	17	0.38	0	37	0	0.50	0	8	28	0.11	9	14	14	0.44	21	16	0	0.79	10	16	11	0.49	0	6	31	0.08	3	27	6	0.45
Mexican	16.8 ± 0.4	2	7	8	0.33	0	17	0	0.51	0	5	12	0.15	3	6	8	0.36	11	4	2	0.77	2	4	11	0.24	0	1	16	0.03	0	16	0	0.48
Puerto Rican	7.9 ± 0.3	0	1	7	0.06	0	8	0	0.51	1	0	7	0.13	0	4	3	0.25	2	5	1	0.57	2	4	2	0.51	0	2	6	0.13	1	7	0	0.57
South America	14.5 ± 1.6	1	2	7	0.14	0	15	0	0.52	0	2	13	0.07	4	7	4	0.52	9	4	2	0.76	1	5	9	0.24	0	1	14	0.03	2	11	2	0.52
South American	7 ± 0.0	0	2	5	0.14	0	7	0	0.50	0	2	5	0.14	0	4	3	0.29	4	2	1	0.71	0	2	5	0.14	0	1	6	0.07	0	6	1	0.43
S. Am. Indian	7.5 ± 1.6	1	0	2	0.13	0	8	0	0.50	0	0	8	0.00	4	3	1	0.69	5	2	1	0.75	1	3	4	0.31	0	0	8	0.00	2	5	1	0.56
Europe	19.8 ± 0.6	3	4	12	0.25	0	19	1	0.48	0	4	16	0.10	6	4	10	0.40	10	6	3	0.66	4	9	7	0.43	0	5	15	0.13	1	9	10	0.28
French	5.9 ± 0.3	3	1	2	0.59	0	6	0	0.51	0	1	5	0.08	2	1	3	0.42	3	1	1	0.59	1	2	3	0.34	0	2	4	0.17	0	3	3	0.25
Italian	10.9 ± 0.3	0	2	8	0.09	0	10	1	0.46	0	3	8	0.14	4	3	4	0.50	7	2	2	0.73	3	6	2	0.55	0	3	8	0.14	1	4	6	0.28
Russian (Krasnodar)	3 ± 0.0	0	1	2	0.17	0	3	0	0.50	0	0	3	0.00	0	0	3	0.00	0	3	0	0.50	0	1	2	0.17	0	0	3	0.00	0	2	1	0.33
Africa	12.9 ± 0.3	0	4	9	0.15	1	12	0	0.54	0	1	12	0.04	1	6	5	0.31	6	3	4	0.58	5	7	1	0.65	0	0	13	0.00	0	9	4	0.35
Black African	4.0 ± 0.0	0	0	4	0.00	1	3	0	0.63	0	0	4	0.00	0	1	3	0.13	1	1	2	0.38	2	1	1	0.63	0	0	4	0.00	0	0	4	0.00
Biaka	3.9 ± 0.3	0	0	4	0.00	0	4	0	0.50	0	1	3	0.13	1	2	0	0.50	1	1	2	0.38	3	1	0	0.88	0	0	4	0.00	0	4	0	0.50
Mbuti	5 ± 0.0	0	4	1	0.40	0	5	0	0.50	0	0	5	0.00	0	3	2	0.30	4	1	0	0.90	0	5	0	0.50	0	0	5	0.00	0	5	0	0.50
Middle East	10 ± 0.0	0	4	6	0.20	0	10	0	0.50	1	1	8	0.15	1	4	5	0.30	5	3	2	0.65	1	6	3	0.40	0	2	8	0.10	1	4	5	0.30
Ashkenazi Jewish	5 ± 0.0	0	2	3	0.20	0	5	0	0.50	1	0	4	0.20	1	2	2	0.40	0	3	2	0.30	0	4	1	0.40	0	2	3	0.20	1	1	3	0.30
Druze	5 ± 0.0	0	2	3	0.20	0	5	0	0.50	0	1	4	0.10	0	2	3	0.20	5	0	0	1.00	1	2	2	0.40	0	0	5	0.00	0	3	2	0.30
Asia/Pacific	67.4 ± 5.3	16	9	34	0.30	1	68	1	0.52	2	13	52	0.13	16	25	25	0.42	49	12	7	0.82	14	33	23	0.45	7	16	44	0.22	9	41	20	0.44
Oriental	12 ± 0.0	4	3	5	0.46	0	11	1	0.46	1	2	9	0.17	2	5	5	0.38	8	4	0	0.83	1	6	5	0.33	0	6	6	0.25	4	5	3	0.54
Aboriginal	13.8 ± 0.6	2	1	11	0.18	0	14	0	0.51	0	5	7	0.18	6	4	4	0.58	7	1	6	0.54	5	4	5	0.51	3	2	9	0.29	3	7	4	0.47
Chinese	4.7± 0.5	1	1	2	0.30	0	5	0	0.50	0	1	4	0.10	0	3	1	0.30	2	2	1	0.60	0	1	4	0.10	0	1	4	0.10	1	4	0	0.60
Indo Pakistani	3.0 ± 0.0	1	1	1	0.50	0	3	0	0.50	0	0	3	0.00	1	1	1	0.50	2	1	0	0.83	0	3	0	0.50	0	0	3	0.00	0	3	0	0.50
Japanese	17.6 ± 2.5	2	1	8	0.14	1	18	0	0.56	0	3	15	0.08	5	4	7	0.39	18	0	0	1.00	5	9	5	0.53	2	5	12	0.25	0	11	8	0.31
Pacific	3.7 ± 0.9	4	0	0	1.00	0	4	0	0.50	1	1	2	0.38	1	3	0	0.63	3	1	0	0.88	0	2	2	0.25	1	0	0	0.25	0	3	1	0.38
Thai	9.7±0.5	0	2	7	0.10	0	10	0	0.52	0	0	10	0.00	1	4	5	0.31	6	3	0	0.77	2	8	0	0.62	0	0	10	0.00	1	5	4	0.36
Southest Asian	2.9 ± 0.3	2	0	0	0.67	0	3	0	0.50	0	1	2	0.17	0	1	2	0.17	3	0	0	1.00	1	0	2	0.33	1	2	0	0.67	0	3	0	0.50

Table S3, related to Figure 6. Upper portion: frequency distribution of rare alleles found in XMR samples. Lower portion: frequency distributions of 8 insertion alleles discovered in whole genome TIP-chip.

Coordinate	Length	Strand	TSD Length	Allele Freq	Location	Gene
34657490	1667	+	15	0.0024	intergenic	--
126304416	1236	+	4	<0.0018	intergenic	--
78555457	6022	-	19	<0.0021	conserved intergenic	--
85568983	368	-	Δ4	<0.0025	2nd intron	*DACH2*
17517693	206	-	15	0.0138	1st intron	*NHS*
111444405	6022	-	9	0.14	2nd intron	*ZCCHC16*

Table S4, related to Figure 7. Rare alleles found in XMR samples.

Supplemental Experimental Procedures

DNA samples

Genomic DNAs from 69 deidentified X-linked intellectual disability samples were derived from lymphoblastoid cell lines by standard phenol:chloroform extraction and ethanol precipitation. Patients were chosen based on a family history suggesting X-linked inheritance (two or more affected males in the same or ≥ two generations; no male-to-male transmission; females not affected or only mildly) or because of the presence of distinct features characteristic of known X-linked intellectual disability syndromes. The samples were collected at the Greenwood Genetic Center following an informed IRB consent protocol and from the HPA (Health Protection Agency, also known as European Collection of Cell Cultures (ECACC). Salibury, UK) culture collections. Genomic DNAs from X-linked dilated cardiomyopathy or Becker muscular dystrophy cases were isolated from PBLs of consented patients undergoing screening for dystrophin gene mutations (Flanigan et al. 2009). For whole genome studies, samples included PBLs from consented volunteers, and 5 twin pairs discordant for diabetes or myocardial infarction. The human diversity panels used to assess allelic frequencies were purchased from Coriell Medical Research Institute (Camden, NJ; catalog number SNP500V) and Health Protection Agency Culture Collections, Salisbury, UK (EDP-1, catalog number 07020701).

Verification PCRs

PCRs to verify LINE insertions were designed to recover the 3' end of a LINE and unique DNA at the insertion site. Site specific primers were chosen using batch Primer3 (Whitehead Institute for Biomedical Research; Cambridge, MA) (Rozen and Skaletsky, 2000) paired with one of two LINE specific primers, GAGATATACCTAATGCTAGATGACAC or AGATATACCTAATGCTAGATGACACA. Verification PCRs spanning entire LINE insertions were performed using LATaq (Takara Bio) or SuperMix HF (Invitrogen; Carlsbad, CA) as described by the manufacturer. Primer sequences used for verification PCRs are included in the **Table**.

Statistical Calculations

All two-group comparisons are performed with a t-test to calculate p-values. To assess correlation between allele frequencies and L1(Ta) lengths (**Figure S5C**) and between L1(Ta) counts and chromosome lengths (**Figure S5A,B**), linear regression was performed. The p-value for the coefficient is provided. To describe deviations from Hardy-Weinberg equilibrium predicted genotype distributions (**Figure 6**), chi-square tests were used. To test whether chromosome 4 is a statistically significant outlier for L1(Ta) insertions (**Figure S5,A,B**), two different models were used for regression. In one model, all data were used for linear regression, while in the other model, data from chromosome 4 were fitted separately. An F-test was performed on the residual sum of square (RSS) of the two different models to obtain the p-value.

Supplemental References

Flanigan, K.M., Dunn, D.M., von Niederhausern, A., Soltanzadeh, P., Gappmaier, E., Howard, M.T., Sampson, J.B., Mendell, J.R., Wall, C., King, W.M.*, et al.* (2009). Mutational spectrum of DMD mutations in dystrophinopathy patients: application of modern diagnostic techniques to a large cohort. Hum Mutat *30*, 1657-1666.

Rozen, S., and Skaletsky, H. (2000). Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol *132*, 365-386.