euL1db: the European database of L1HS retrotransposon insertions in humans

Ashfaq A. Mir, Claude Philippe and Gaël Cristofari

SUPPLEMENTARY DATA

Supplementary Data online contains a “Supplementary Methods” section, 2 Supplementary Figures and 1 Supplementary Table.

Supplementary Methods

Window size selection to build MRIP records. Since each study use a distinct set of methods to detect L1 insertions and specific algorithms to define the junctions between an L1 copy and its flanking sequence, identical L1 insertions might be reported with slightly different coordinates in different datasets. For this reason, each MRIP was built by merging SRIPs that fall into a 200-bp sliding window. To choose the size of this window, we evaluated the distribution of the SRIP-to-SRIP distances in euL1db. We first extracted non-redundant SRIPs from each Study. Practically, if a Study includes multiple SRIPs with the same exact genomic coordinates, we kept only one of them. This step prevents Studies with a large number of samples to be overrepresented in the distribution. Then, we used Bedtools (1) to form clusters of SRIPs falling in a 1 kb-window (bedtools cluster with option –d 1000). Clusters with a single SRIP were excluded. Finally, we calculated the distance between each pair of non-redundant SRIP pair falling in a given cluster (using bedtools closest with option -d). Strand orientation was not taken into account since strand information is missing for some SRIPs. The distribution of the SRIP-to-SRIP distances is graphed in Supplementary Figure 2 and indicates that 95% of the SRIP are less than 200-bp distant (median: 7bp; 75% quartile: 27 bp; 95% percentile: 214 bp). We choose the 200 bp value to keep acceptable insertion site accuracy and to avoid splitting a unique biological event into artificially distinct records.

Data curation and inclusion. Data included in euL1db were often directly extracted from the main or supplementary data of the original publications. When positions were reported in hg18 reference genome coordinates, we translated them into GRCh37/hg19 coordinates using the liftOver tool of the UCSC Genome Browser website (available at: https://genome.ucsc.edu/cgi-bin/hgLiftOver) (2). Only L1HS insertions were kept. Putative insertions that were tested by PCR and/or Sanger sequencing but failed these additional validations were excluded. Some studies required additional processing, which are detailed below. The sources of high-throughput data included in euL1db at the time of writing are summarized in Supplementary Table 1.

The original Beck *et al.* 2010 publication (3) only reports a short DNA sequence corresponding to the preintegration site, and its chromosome number and cytogenetic band. We remapped the preintegration site sequence to reference genome hg19 with BWA (4) to obtain the precise genomic coordinates. For sequences with multiple possible positions (MAPQ=0), we selected the position consistent with the reported cytogenetic band.

The original Iskow *et al.* 2010 publication (5) provides the DNA sequence downstream of each putative L1 insertions obtained by 454 sequencing (in opposite orientation relative to L1) and the genomic coordinates of the BLAT best hit after mapping them on the hg18 reference genome. To obtain the strand information of these putative insertions, we remapped the published DNA sequences (1389 in total) to hg19 using BWA. In a first step, we used bwa mem (bwa mem with options -t4 -M). We recovered 980 uniquely mapped positions consistent with the original BLAT analysis; 285 unmapped sequences; and 124 sequences mapping to multiple positions or corresponding to chimeric sequence. The latter were discarded and not included in euL1db. In a second step, unmapped sequences from the first step were mapped again using an algorithm with increased sensitivity for short sequences (bwa aln with options -l12 -o2 / bwa samse with –n 10 option), allowing us to recover an additional set of 116 uniquely mapping sequences consistent with the original BLAT coordinates. In total, 1095 high-confidence insertions out of 1389 sequences from the 454 Iskow experiments have been included in euL1db. The published table reporting the results of ABI Sanger sequencing experiments includes the coordinates of the flanking region sequenced but not the DNA sequence itself, nor its orientation. We used the middle of this segment as a coordinate for the insertion point and we could not deduce strand information.

The Baillie *et al.* publication (6) contains two distinct sets of data obtained by RC-seq: (i) germline polymorphic retrotransposon insertions discovered in a pool of genomic DNA isolated from the blood of several individuals; and (ii) putative somatic and germline retrotransposon insertions discovered in genomic DNA isolated from different brain regions of three individuals. Because many of the putative somatic insertions are only supported by a single sequencing read, we only kept in euL1db somatic insertions that have been validated by PCR and/or Sanger sequencing. In contrast, germline insertions were more robustly identified (several sequencing reads in multiple independent libraries) and were kept, whether further tested by PCR or not. Only L1HS elements were included (L1-Ta and L1-Pre-Ta).

Similarly the Solyom *et al.* article (7) reports insertions found in multiple colon cancer samples or their matched normal tissues. This was achieved by a combination of L1-seq and RC-seq approaches. For the L1-seq approach (their Sup. Tab. 1A and 1B), we kept as germline insertions, those found in both normal and tumor tissues and, as somatic insertions, those found only in one tissue and validated by PCR and/or Sanger sequencing (their Sup. Tab. 3). For the RC-seq data (their Sup. Table 2), we only kept those tagged as "high confidence" and we further added/removed insertions validated/invalidated by PCR and/or Sanger sequencing data (their Sup. Tab. 3). Here as well, only L1HS elements were included.

Some studies have searched L1 insertions in pooled samples (obtained from different individuals, cell lines, etc…) as a facilitated mean to find common inherited L1 polymorphisms. In these cases, insertions cannot be linked to samples/individuals. Such samples were recorded in euL1db but tagged as “Mix” in their sample relationship field.

Supplementary References

1. Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics,* **26**, 841-842.

2. Karolchik, D., Barber, G.P., Casper, J., Clawson, H., Cline, M.S., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., *et al.* (2014) The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res,* **42**, D764-D770.

3. 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.

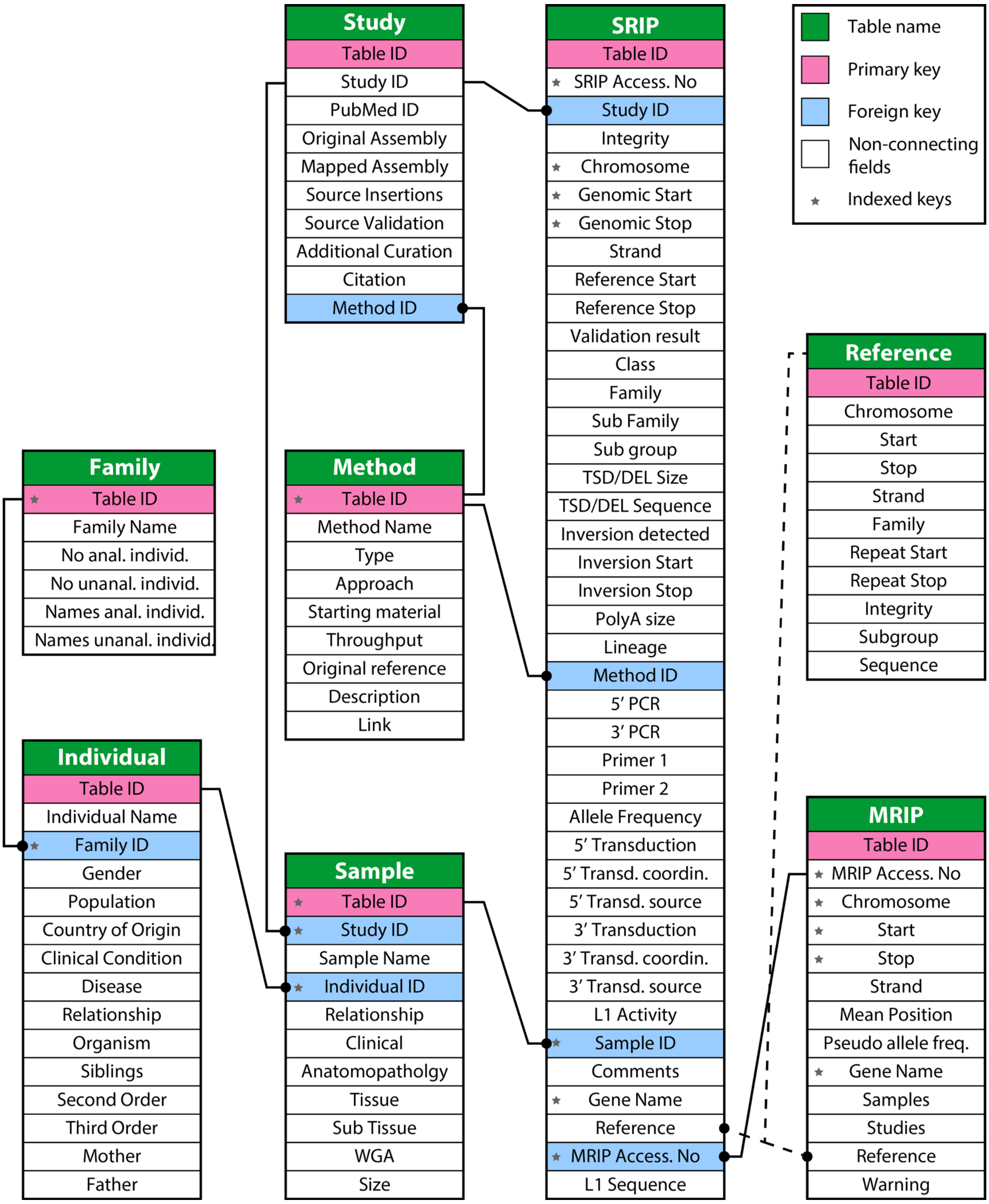
4. Li, H. and Durbin, R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics,* **26**, 589-595.

5. 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.

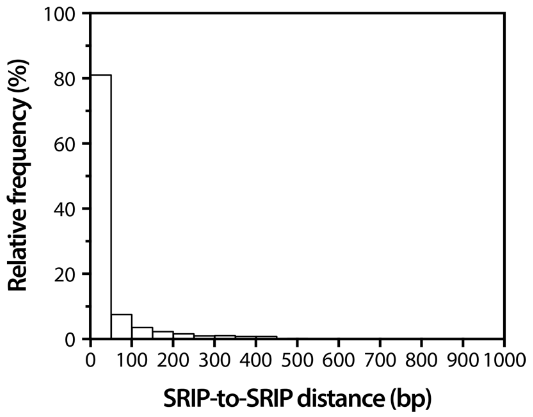
6. Baillie, J.K., Barnett, M.W., Upton, K.R., Gerhardt, D.J., Richmond, T.A., De Sapio, F., Brennan, P.M., Rizzu, P., Smith, S., *et al.* (2011) Somatic retrotransposition alters the genetic landscape of the human brain. *Nature,* **479**, 534-537.

7. Solyom, S., Ewing, A.D., Rahrmann, E.P., Doucet, T., Nelson, H.H., Burns, M.B., Harris, R.S., Sigmon, D.F., Casella, A., *et al.* (2012) Extensive somatic L1 retrotransposition in colorectal tumors. *Genome Res,* **22**, 2328-2338.

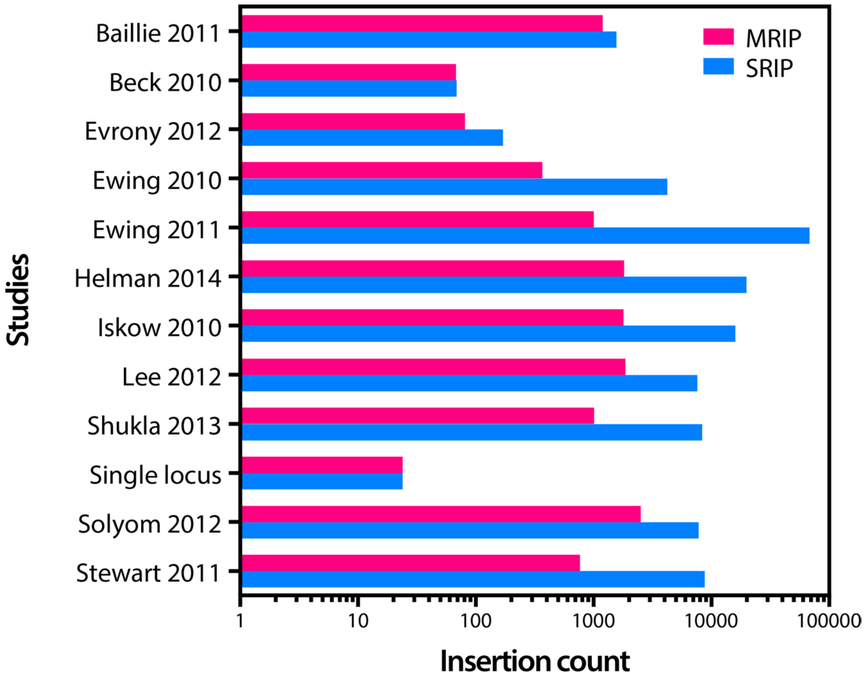
Supplementary Figures



Supplementary Figure 1. euL1db table structure. Data have been organized in 8 Meta tables (names in green) tailored to allow efficient data mining and reduce processing time by built-in functions. Meta data tables distribute the information into chunks of systematic and interconnected data records. Primary keys are shown in pink, foreign keys are shown in blue and are linked by plain lines to their original table. Indexed keys are marked with a star. Fields related to a single table are presented with a white background and carry information specific to each stored data type. Meta tables form a light layer of information network for zooming down to a particular detail without processing all of the data. The "reference" fields of the SRIP and MRIP tables are filled non-dynamically (dashed line) with the "Reference" table, which contains the reference L1HS insertions. This data structure allows maximum amount of information to be stored without compromising processing time and memory utilization by interrogating external programs.



Supplementary Figure 2. Distribution of SRIP-to-SRIP distances in euL1db. Since 95% of SRIP are less than 200 bp distant from each others, we choose this window size to merge SRIP into MRIP (see Fig. 1B). The detailed procedure is explained in Supplementary Methods.



Supplementary Figure 3. Count and origin of L1HS insertions included in euL1db.

Supplementary Table

Study.ID	PMID	Source of insertions data	Source of validation data	Coordinates liftover	Additional processing
Baillie2011	22037309	Tables S4, S5	Tables S6, S7	-	See text
Beck2010	20602998	Table S2	Table S2	hg18 to hg19	See text
Evrony2012	23101622	Table S3	Table S3	-	-
Ewing2010	20488934	Table S1	Table S1	hg18 to hg19	-
Ewing2011	20980553	Table S4	-	hg18 to hg19	-
Helman2014	24823667	Tables 2, 3, 5*	Table 1	hg18 to hg19 **	
Iskow2010	20603005	Tables S1, S2	Tables S1, S2	hg18 to hg19	See text
Lee2012	22745252	Tables S2, S6, S8	Tables S3, S4, S7	hg18 to hg19	-
Shukla2013	23540693	Table S3	Tables S5, S6	-	-
Solyom2012	22968929	Tables 1, S1A, S1B, S2	Table S3	-	See text
Stewart2011	21876680	Table S1	Table S1	hg18 to hg19	DEL excluded§

Supplementary Table 1. Source data for high- and medium-troughput L1 mapping studies.
*, Table numbers are conflicting in the Supplementary Data of this article. **, Only for 3 tissues (COAD, READ, LAML). Other samples were already mapped to the hg19 assembly. §, In this study, mobile element insertion (MEI) annotated as DEL are insertions present in the reference genome, but absent from a particular sample.