Spontaneous retrotranspositions in normal tissues are rare and associated with cell-type-specific differentiation

Xiao Dong[1]*, Lei Zhang[1]*, Kristina Brazhnik[1]*, Moonsook Lee[1], Xiaoxiao Hao[1], Alexander Y. Maslov[1], Zhengdong Zhang[1], Tao Wang[2], Jan Vijg[1,3]

[1]Department of Genetics, Albert Einstein College of Medicine, Bronx, NY 10461, USA.
[2]Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY 10461, USA.
[3]Center for Single-Cell Omics in Aging and Disease, School of Public Health, Shanghai Jiao Tong University School of Medicine, Shanghai, 200025, China

*These authors contributed equally to this work.
Corresponding to X.D. (biosinodx@gmail.com) and J.V. (jan.vijg@einstein.yu.edu)

Activation of retrotransposons and their insertions into new genomic locations, i.e., retrotranspositions (RTs), have been identified in about 50% of tumors. However, the landscape of RTs in different, normal somatic cell types in humans remains largely unknown. Using single-cell whole-genome sequencing we identified 528 RT events, including LINE-1 (L1), and Alu, in 164 single cells and clones of fibroblasts, neurons, B lymphocytes, hepatocytes and liver stem cells, of 29 healthy human subjects aged from 0 to 106 years. The frequency of RTs was found to vary from <1 on average per cell in primary fibroblasts to 7.8 per cell in hepatocytes. Somewhat surprisingly, RT frequency does not increase with age, which is in contrast to other types of spontaneous mutation. RTs were found significantly more likely to insert in or close to target genes of the Polycomb Repressive Complex 2 (PRC2), which represses most of the genes encoding developmental regulators through H3K27me3 histone modification in embryonic stem cells. Indeed, when directly comparing RT frequency between differentiated liver hepatocytes with liver stem cells, the latter were almost devoid of RTs. These results indicate that spontaneous RTs are associated with cellular differentiation and occur, possibly, as a consequence of the transient chromatin transition of differentiation-specific genes from a transcriptionally repressed to activated state during the differentiation process.

Retrotransposons are widespread repetitive elements in the genome. They are usually classified into LTR (Long Terminal Repeat) retrotransposons and non-LTR retrotransposons. The latter type is the most abundant and include Long Interspersed Nuclear Element (LINE)-1 (L1), Alu and SVA elements. Together these account for more than one-third of the human genome[1]. Only L1 elements are autonomous retrotransposons and about 100 of them have been demonstrated as still active in the human genome[2-4]. Most are inactivated by truncations, rearrangements and other mutations. L1 elements can be activated, transcribed and, after reverse transcription, re-integrated in the genome[2,5]. While normally repressed, possibly through epigenetic mechanisms[6], retrotranspositions (RTs) have been reported in both tumors and in the germline[4,7-9].

In normal somatic cells, derepression of retrotransposons has been documented during early embryonic development, neurological disorders, and replicative senescence[10-13]. Activation of retrotransposons has been implicated in the loss of genome integrity during aging of somatic cells[14]. However, the landscape of RTs in normal human cells is almost completely unknown apart from some conflicting reports on human neurons[15-18]. Indeed, quantitative information about RTs cannot be obtained by studying bulk tissues and requires advanced, well-validated single-cell genomics technology[19,20].

To identify RTs in normal somatic cells, we used whole-genome sequencing data of 152 single cells and 12 single-cell derived clones, of 56 B lymphocytes, 28 fibroblasts, 36 neurons, 36 hepatocytes and 8 liver stem cells, of 29 healthy subjects aged between 0 year and 106 years. Apart from our own data this also included single-cell whole genome sequences generated by others using different single-cell amplification methods (Table S1)[20-22]. Whole-genome sequencing of bulk DNA of the same subjects were also analyzed to filter out germline polymorphisms. On average, 30.3±8.4x depth of single cells and clones, and 29.4±7.2x depth of bulk DNA were obtained after alignment, covering on average 87.3±9.2% and 91.9±0.3% of the genome of single cells/clones, and bulk DNA separately (Table S2). Somatic RTs were then identified comparing the alignment of single cells and clones to their corresponding bulk DNAs using TraFiC[8,9]. We validated our variant calling by PCR analysis of 10 randomly chosen RTs (Supplementary Information; Table S3). In addition, to ensure that none of the identified RTs were artifacts of the amplification we directly compared RT frequencies obtained after whole-genome amplification of single human fibroblasts with those found in unamplified DNA from clones derived from single cells in the same population of cells (Fig. S1). The results obtained show very similar results between single-cell and unamplified clone analysis, indicating that the RTs identified were bona fide insertions and not artifacts of the amplification procedure or variant calling.

Across all samples we identified 528 RT events. Consistent with previous studies in tumors[7-9], most of the RTs were L1 insertions (77.5%), with Alu as the second most frequent type (14.2%) (Fig. 1a, Fig. S2a,b, Table S4). RT frequency appeared to be highly cell-type specific, with almost no insertions found in fibroblasts, 2-3 per cell in B lymphocytes and neurons and more than 7 in hepatocytes (Fig. 1b, Fig. S3); about 50% to 90% of the cells carried at least one RT depending on the cell-type (Fig. 1c, Fig. S4a,b). The RT frequency observed is in agreement with a previous report of almost no RTs in fibroblast clones [23]. The frequency observed in neurons is largely in agreement with ref [16], which reported <1 RT per neuron, but not in agreement with ref [15], which reported 13.7 RTs per neurons.

For the most frequent L1 family, germline RT occurs from a small number of polymorphic, "hot" elements[4]. We identified three hot L1 source elements, which were also found to be cell-type specific (Fig. 1d-g, Table S5): 15q24.3 is the source of 19 events in 8 hepatocytes and 11 B lymphocytes of 14 human subjects; 5q23.1 is the source of 17 events in 5 neurons of 2 subjects; and 12q14.3 is the source of 10 events in 3 B lymphocytes of 2 subjects. Although similar numbers of hot L1 sources were reported in cancers[9], there is virtually no overlap with the L1 sources in normal cells. Indeed,

Xp22.2-Xp22.13, a hot L1 source found in cancers, accounted for only one single event in a B lymphocyte. This cell-type specificity in source elements may be a result of potential differences in transcriptional activation of L1 sources between tumors and normal cells.

Thus far, only expression but not re-integration of retrotransposons has been studied as a function of age. The wide age-coverage of the samples of three cell types, i.e., B lymphocytes, hepatocytes and neurons, allowed us for the first time to test if RT insertions accumulate during aging. To our surprise, unlike other types of mutations [24-26], no increase in RT frequency was observed in any of the three cell types, B lymphocytes (aged between 0 and 106 years), hepatocytes (5 months - 77 years), and neurons (15 - 42 years) during aging (Fig. 2a, Fig. S5). This is in spite of the demonstrated increase in expression of retrotransposons during aging[27-29]. This finding suggests that RT activation may be a transient event occurring only once during a cell's lifetime after which repression is again implemented. To test this, we studied possible preferences in genomic location of the RTs identified.

RT distribution across the genome was analyzed for all RTs collectively. The results showed that insertion sites are significantly depleted from the most obvious functional sequences. That is, the RT frequencies in gene exons, 5' and 3' UTRs are 62%, 48%, and 49%, respectively, significantly less than what would be expected by chance alone (P=0.004, 0.022 and 0.006, permutation test; Fig. 2b, Fig. S6a-c). RT insertion sites are also depleted from CpG islands and their flanking island shores (38% less than the frequency expected by chance alone; P=0.041), which is consistent with the finding that insertions have preferentially been found at AT regions in tumors[7].

We then tested the target regions of 161 transcription factors (TFs), identified from multiple cell types by ENCODE[30], for enrichment or depletion of RTs. For most of the TFs, their target regions were neither depleted nor enriched for RTs (Table S6). However, the target regions of seven and three TFs were found depleted or enriched for RT insertions, respectively (fold change≥2 and FDR-adjusted P<0.05 for multiple testing correction; Fig. 2c, Fig. S7). Interestingly, the most significant association was observed as an enrichment of RTs at the target regions of SUZ12, i.e., 4.8-fold higher frequency than expected by chance alone (P<0.0005 permutation test, FDR-adjusted P=0.0345). SUZ12 is a major component of the Polycomb Repressive Complex (PRC2), which functions in embryonic stem (ES) cells to repress expression of developmental genes[31]. The majority of genes encoding developmental regulators exhibit extended regions of PRC2 binding in humans [32].

We then validated the enrichment at PRC2 target regions with an independent dataset of ES cells[32], in which target genes of PRC2 were defined as the intersection of three target gene sets of SUZ12, EED and H3K27me3 separately. We found that RT insertion sites are significantly enriched in any of the three gene sets as well as their intersection, i.e., the PRC2 target genes, (3.0-fold higher than expected by chance alone, P=0.0015 permutation test; Fig. 2d, Fig. S8a-c, Table S7).

Based on the observed enrichment of RT insertions at PRC2 and H3K2me3 target genes, we hypothesized that these events are in some way associated with the temporal regulation of chromatin during differentiation from a stem or progenitor cell to a fully mature cell type. To test this, we made use of the available whole-genome sequences of both differentiated, mature liver hepatocytes and liver stem cells. The identity of the latter was verified using a set of stem-cell-specific and epithelial-progenitor-cell-specific cell surface markers (EpCAM, Lgr5, CD90, CD29, CD105, and CD73; Supplementary Information). If RT would be a transient event occurring only once in a lifetime, we would expect no RTs in the stem cells. Indeed, we found that although hepatocytes have the highest RT frequency (7.83±11.51 RTs per cell, avg.±s.d.) among the four cell types analyzed, liver stem cells have almost no RTs, i.e., 0.25±0.48 per cell with 25% of cells carrying at least one insertion (P=0.00037, Fig. 3a,b, Fig. S9). This result confirmed that RTs are associated, at least in human liver, with cellular differentiation.

The cell type-specific differences observed could well be related to the state of differentiation and/or the origin of the cell type. For example, fibroblasts are the least specialized cells in the connective-tissue family [33], which may be related to their very low frequency of RTs. The RTs during differentiation may have both beneficial and damaging effects. As documented, expression of certain L1 elements is probably required during early embryonic development[34,35]. However, L1 insertions are stochastic events and of low abundance, i.e., no more than a few per cell. Therefore, it seems highly unlikely that they have any consistent regulatory effect on development. Instead, they are probably mere by-products of essential biological processes.

In summary, using single-cell whole-genome sequencing, we demonstrate that RTs in normal somatic cells of healthy human subjects occur at a low frequency, which is cell type-specific. Multiple lines of evidence indicate that rather than gradually increasing with age, for example, as a consequence of an age-related activation [27], spontaneous RTs in normal cells occur mainly during cellular differentiation, probably as by-products of the dramatic chromatin alterations associated with terminal differentiation. While this suggests that RTs do not significantly contribute to the normal aging process due to their low frequency and lack of further accumulation, RTs may well play a role in age-related diseases, most notably cancer [7-9,18].

Acknowledgements This study was supported by NIH grants P01 AG017242 (J.V.), K99 AG056656 (X.D.), P01 AG047200 (J.V.), U01 ES029519 (J.V.) and the Paul F. Glenn Center for the Biology of Human Aging.

Author Contributions J.V. and X.D. conceived this study and designed the experiments. X.D., X.H., T.W. and Z.Z. analyzed the data. L.Z., K.B., M.L. and A.Y.M. performed the experiments. X.D. and J.V. wrote the manuscript.

Competing interests: X.D., L.Z., M.L., A.Y.M., J.V. are co-founders of SingulOmics Corp. The other authors declare no competing interests

References

1 Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921, doi:10.1038/35057062 (2001).
2 Sassaman, D. M. *et al.* Many human L1 elements are capable of retrotransposition. *Nature genetics* **16**, 37-43, doi:10.1038/ng0597-37 (1997).
3 Beck, C. R. *et al.* LINE-1 retrotransposition activity in human genomes. *Cell* **141**, 1159-1170, doi:10.1016/j.cell.2010.05.021 (2010).
4 Brouha, B. *et al.* Hot L1s account for the bulk of retrotransposition in the human population. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 5280-5285, doi:10.1073/pnas.0831042100 (2003).
5 Boeke, J. D., Garfinkel, D. J., Styles, C. A. & Fink, G. R. Ty elements transpose through an RNA intermediate. *Cell* **40**, 491-500 (1985).
6 Slotkin, R. K. & Martienssen, R. Transposable elements and the epigenetic regulation of the genome. *Nature reviews. Genetics* **8**, 272-285, doi:10.1038/nrg2072 (2007).
7 Lee, E. *et al.* Landscape of somatic retrotransposition in human cancers. *Science (New York, N.Y.)* **337**, 967-971, doi:10.1126/science.1222077 (2012).
8 Tubio, J. M. C. *et al.* Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science (New York, N.Y.)* **345**, 1251343, doi:10.1126/science.1251343 (2014).
9 Rodriguez-Martin, B. *et al.* Pan-cancer analysis of whole genomes reveals driver rearrangements promoted by LINE-1 retrotransposition in human tumours. *bioRxiv*, doi:10.1101/179705 (2017).
10 Molaro, A. *et al.* Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. *Cell* **146**, 1029-1041, doi:10.1016/j.cell.2011.08.016 (2011).
11 Shpyleva, S., Melnyk, S., Pavliv, O., Pogribny, I. & Jill James, S. Overexpression of LINE-1 Retrotransposons in Autism Brain. *Molecular neurobiology* **55**, 1740-1749, doi:10.1007/s12035-017-0421-x (2018).
12 Guo, C. *et al.* Tau Activates Transposable Elements in Alzheimer's Disease. *Cell reports* **23**, 2874-2880, doi:10.1016/j.celrep.2018.05.004 (2018).
13 De Cecco, M. *et al.* Genomes of replicatively senescent cells undergo global epigenetic changes leading to gene silencing and activation of transposable elements. *Aging cell* **12**, 247-256, doi:10.1111/acel.12047 (2013).
14 Gorbunova, V., Boeke, J. D., Helfand, S. L. & Sedivy, J. M. Human Genomics. Sleeping dogs of the genome. *Science (New York, N.Y.)* **346**, 1187-1188, doi:10.1126/science.aaa3177 (2014).
15 Upton, K. R. *et al.* Ubiquitous L1 mosaicism in hippocampal neurons. *Cell* **161**, 228-239, doi:10.1016/j.cell.2015.03.026 (2015).
16 Evrony, G. D. *et al.* Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* **151**, 483-496, doi:10.1016/j.cell.2012.09.035 (2012).
17 Evrony, G. D., Lee, E., Park, P. J. & Walsh, C. A. Resolving rates of mutation in the brain using single-neuron genomics. *eLife* **5**, doi:10.7554/eLife.12966 (2016).
18 Faulkner, G. J. & Billon, V. L1 retrotransposition in the soma: a field jumping ahead. *Mobile DNA* **9**, 22, doi:10.1186/s13100-018-0128-1 (2018).

19 Gundry, M. & Vijg, J. Direct mutation analysis by high-throughput sequencing: from germline to low-abundant, somatic variants. *Mutation research* **729**, 1-15, doi:10.1016/j.mrfmmm.2011.10.001 (2012).

20 Dong, X. *et al.* Accurate identification of single-nucleotide variants in whole-genome-amplified single cells. *Nature methods* **14**, 491-493, doi:10.1038/nmeth.4227 (2017).

21 Lodato, M. A. *et al.* Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science (New York, N.Y.)* **350**, 94-98, doi:10.1126/science.aab1785 (2015).

22 Chen, C. *et al.* Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI). *Science (New York, N.Y.)* **356**, 189-194, doi:10.1126/science.aak9787 (2017).

23 Saini, N. *et al.* The Impact of Environmental and Endogenous Damage on Somatic Mutation Load in Human Skin Fibroblasts. *PLoS genetics* **12**, e1006385, doi:10.1371/journal.pgen.1006385 (2016).

24 Zhang, L. & Vijg, J. Somatic Mutagenesis in Mammals and Its Implications for Human Disease and Aging. *Annual review of genetics* **52**, 397-419, doi:10.1146/annurev-genet-120417-031501 (2018).

25 Dolle, M. E. *et al.* Rapid accumulation of genome rearrangements in liver but not in brain of old mice. *Nature genetics* **17**, 431-434, doi:10.1038/ng1297-431 (1997).

26 Dolle, M. E., Snyder, W. K., Gossen, J. A., Lohman, P. H. & Vijg, J. Distinct spectra of somatic mutations accumulated with age in mouse heart and small intestine. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 8403-8408 (2000).

27 De Cecco, M. *et al.* Transposable elements become active and mobile in the genomes of aging mammalian somatic tissues. *Aging* **5**, 867-883, doi:10.18632/aging.100621 (2013).

28 Chen, H., Zheng, X., Xiao, D. & Zheng, Y. Age-associated de-repression of retrotransposons in the Drosophila fat body, its potential cause and consequence. *Aging cell* **15**, 542-552, doi:10.1111/acel.12465 (2016).

29 Elsner, D., Meusemann, K. & Korb, J. Longevity and transposon defense, the case of termite reproductives. *Proceedings of the National Academy of Sciences of the United States of America* **115**, 5504-5509, doi:10.1073/pnas.1804046115 (2018).

30 Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91-100, doi:10.1038/nature11245 (2012).

31 Margueron, R. & Reinberg, D. The Polycomb complex PRC2 and its mark in life. *Nature* **469**, 343-349, doi:10.1038/nature09784 (2011).

32 Lee, T. I. *et al.* Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* **125**, 301-313, doi:10.1016/j.cell.2006.02.043 (2006).

33 Alberts, B., Johnson, A., Lewis, J. & al., e. Ch. Fibroblasts and Their Transformations: The Connective-Tissue Cell Family., (Garland Science, 2002).

34 Beraldi, R., Pittoggi, C., Sciamanna, I., Mattei, E. & Spadafora, C. Expression of LINE-1 retroposons is essential for murine preimplantation development. *Molecular reproduction and development* **73**, 279-287, doi:10.1002/mrd.20423 (2006).

35 Jachowicz, J. W. *et al.* LINE-1 activation after fertilization regulates global chromatin accessibility in the early mouse embryo. *Nature genetics* **49**, 1502-1510, doi:10.1038/ng.3945 (2017).

Figure legends

Fig. 1. RT spectrum, frequency and distribution.
(**a**) Spectrum of RTs identified across all cells collectively. (**b**) Frequencies of RTs in each cell type. A dot presents a cell; bars present the average with its standard deviation (s.d.). *P* values were determined for the difference in total RTs between cell types using two tailed Student's t-test. (**c**) Fractions of cells with different numbers of RTs. (**d-g**) Genome distribution of RTs in each cell type. RTs of all cells per cell type were plotted together in each figure. Within each circus plot, links in the inner layer present directions from L1 sources to their insertion sites. Only sources of those with transductions can be determined by aligning non-repetitive sequences transduced with RT to the reference genome (Supplementary Information), and were plotted. The middle layer of a circus plot presents insertion sites of all RTs. The external layer presents cytobands of chromosomes.

Fig. 2. Age-related frequency and genomic-feature distribution of RTs.
(**a**) Frequency of RTs during aging. Linear regressions were performed for each cell type separately. (**b**) Depletion of RTs in genomic features. (**c**) Enrichment of RTs in target regions of SUZ12. (**d**) Enrichment of RTs in target genes of PRC2. (**b-d**) A red dot presents the number of RTs observed. A violin plot (with a box plot inside) presents a distribution (with median and quantiles) of expected numbers of RTs by chance alone (2,000 times of random sampling, Supplementary Information). *P* values were estimated as Monte Carlo *P* values based on the random sampling.

Fig. 3. RT frequency in liver stem cells and hepatocytes.
(**a**) Frequencies of RTs in liver stem cells and terminally differentiated hepatocytes. A dot presents a cell; bars present the average and s.d.. *P* values were determined for the difference of total RTs using two tailed Student's t-test. (**b**) Fractions of cells with different number of RTs.

Fig. S1. Comparison of RT frequency observed in single cells and single-cell derived clones.
Clones and SCMDA-amplified single cells were obtained from the same population of human fibroblasts[20]; in the paper about LIANTI amplification, single cells of another human fibroblast population were used and amplified using multiple different amplification protocols[22]. No significant difference was observed in number of RTs per cell between the three datasets.

Fig. S2. Spectra of RTs in different cell types.

(**a**) Spectrum of RTs in all cells collectively. L1-td0, td1, and td2 represent solo-L1 events, partnered transductions and orphan transductions, respectively (Supplementary Information). (**b**) Spectra of RTs in different cell types.

Fig. S3. Average frequencies of RTs in each cell type.
L1-td0, td1, td2 represent solo-L1 events, partnered transductions and orphan transductions, respectively (Supplementary Information).

Fig. S4. Fractions of cells with different numbers of RTs.
(**a**) L1. (**b**) Alu.

Fig. S5. Frequency of L1 insertions during aging.
Linear regressions were performed for each cell type separately.

Fig. S6. Depletion of RTs in genomic features per cell type.
A red dot represents the number of RTs observed. A violin plot (with a box plot inside) presents the distribution (with median and quantiles) of expected numbers of RTs by chance alone (2,000 times of random sampling, Supplementary Information). *P* values were estimated as Monte Carlo *P* values based on the random sampling.

Fig. S7. Enrichment of RTs in target regions of TFs.
A red dot represents the number of RTs observed. A violin plot (with a box plot inside) presents the distribution (with median and quantiles) of expected numbers of RTs by chance alone (2,000 times of random sampling, Supplementary Information). *P* values were estimated as Monte Carlo *P* values based on the random sampling.

Fig. S8. Enrichment of RTs in target genes of PRC2.
A red dot represents the number of RTs observed. A violin plot (with a box plot inside) presents the distribution (with median and quantiles) of expected numbers of RTs by chance alone (2,000 times of random sampling, Supplementary Information). *P* values were estimated as Monte Carlo *P* values based on the random sampling.

Fig. S9. Average frequencies of RTs in liver stem cells and hepatocytes.
L1-td0, td1, and td2 represent solo-L1 events, partnered transductions and orphan transductions, respectively (Supplementary Information).

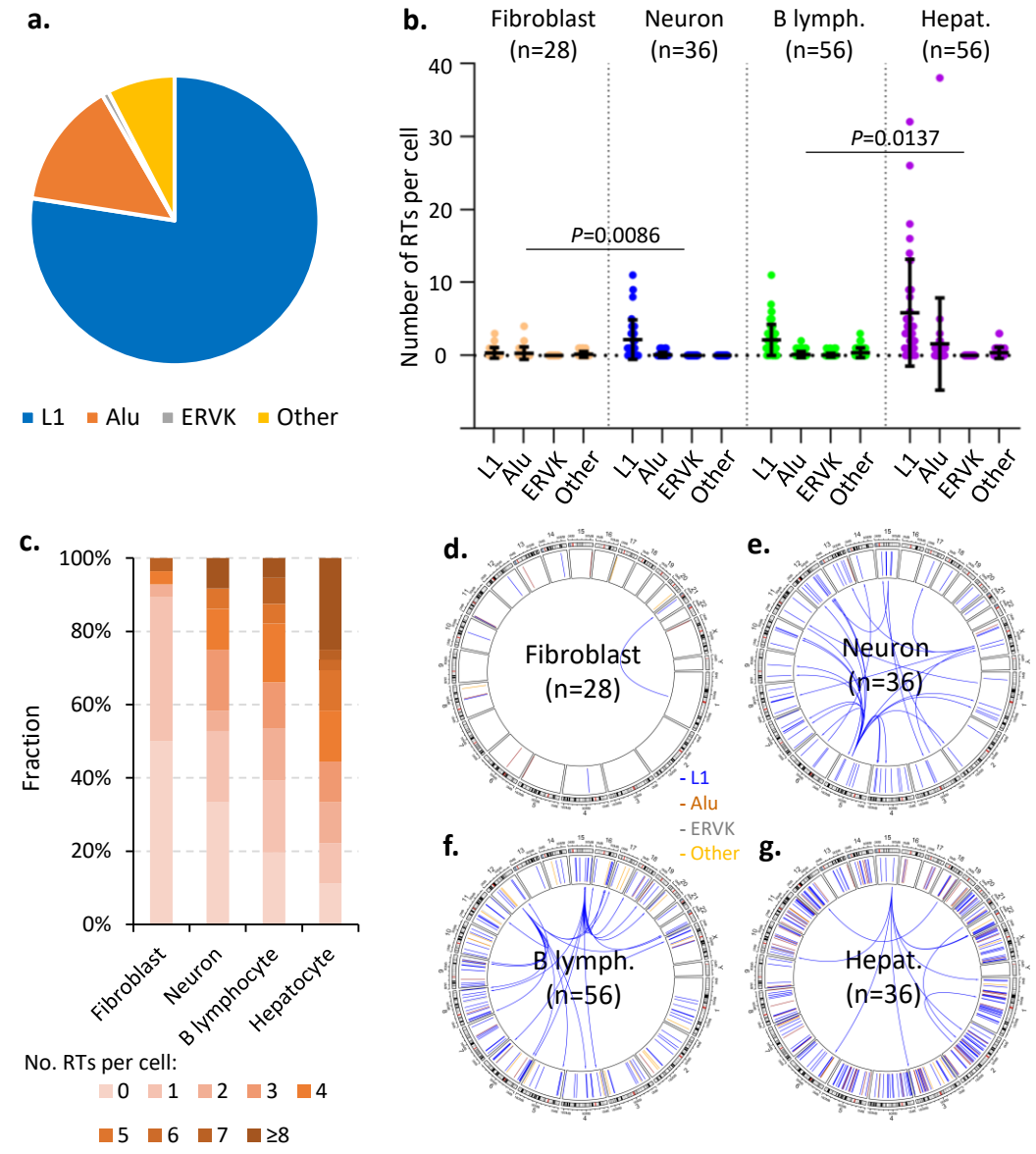
a.
L1
Alu
ERVK
Other
b.
Fibroblast (n=28)
Neuron (n=36)
B lymph. (n=56)
Hepat. (n=56)
Number of RTs per cell
P=0.0086
P=0.0137
c.
Fraction
Fibroblast
Neuron
B lymphocyte
Hepatocyte
No. RTs per cell:
d.
Fibroblast (n=28)
e.
Neuron (n=36)
- L1
- Alu
- ERVK
- Other
f.
B lymph. (n=56)
g.
Hepat. (n=36)

Fig. 1.

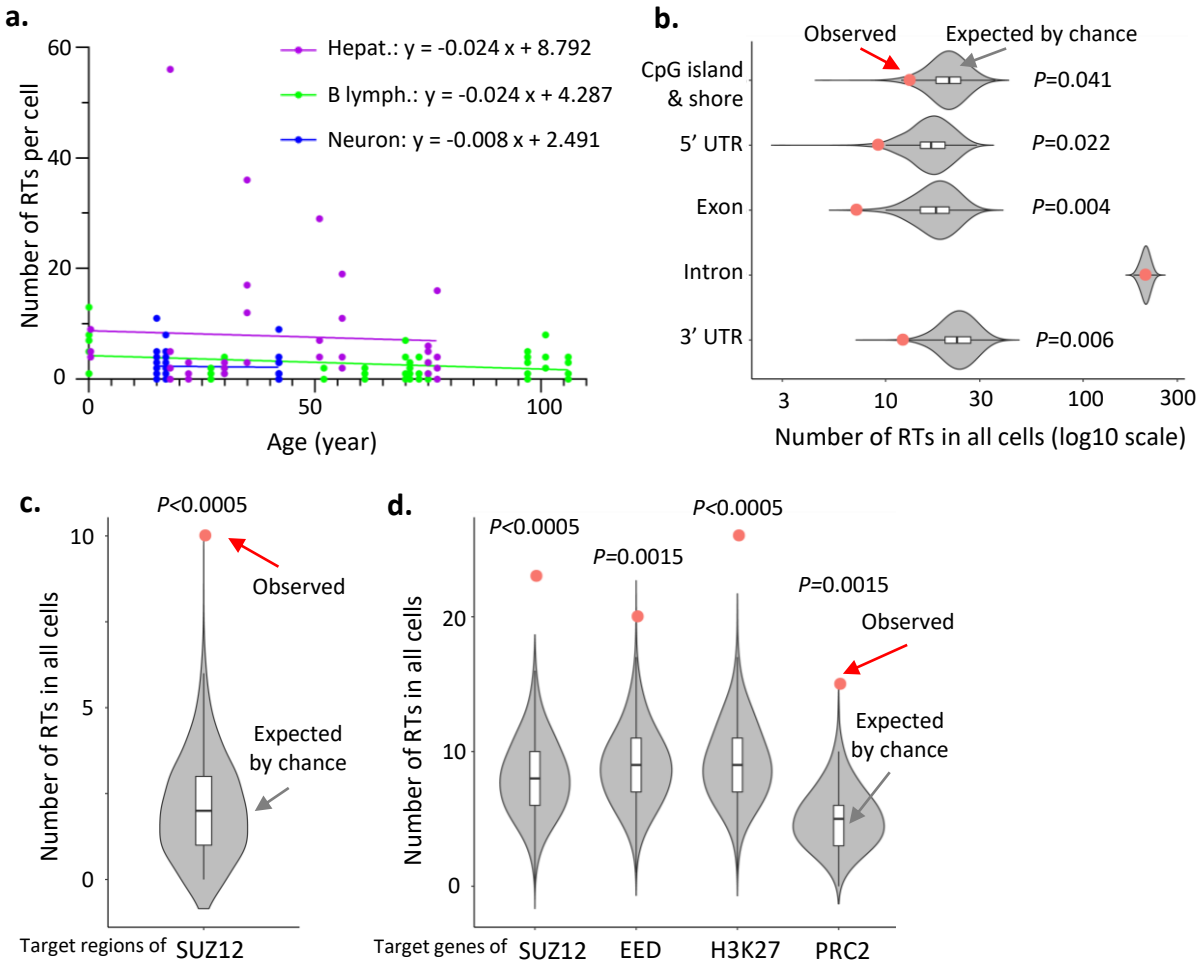
a.
Number of RTs per cell
Age (year)
Hepat.: y = -0.024 x + 8.792
B lymph.: y = -0.024 x + 4.287
Neuron: y = -0.008 x + 2.491
b.
Observed
Expected by chance
CpG island & shore
5' UTR
Exon
Intron
3' UTR
P=0.041
P=0.022
P=0.004
P=0.006
Number of RTs in all cells (log10 scale)
c.
P<0.0005
Observed
Expected by chance
Number of RTs in all cells
Target regions of SUZ12
d.
P<0.0005
P=0.0015
P=0.0015
P<0.0005
Observed
Expected by chance
Number of RTs in all cells
Target genes of SUZ12 EED H3K27 PRC2

Fig. 2.

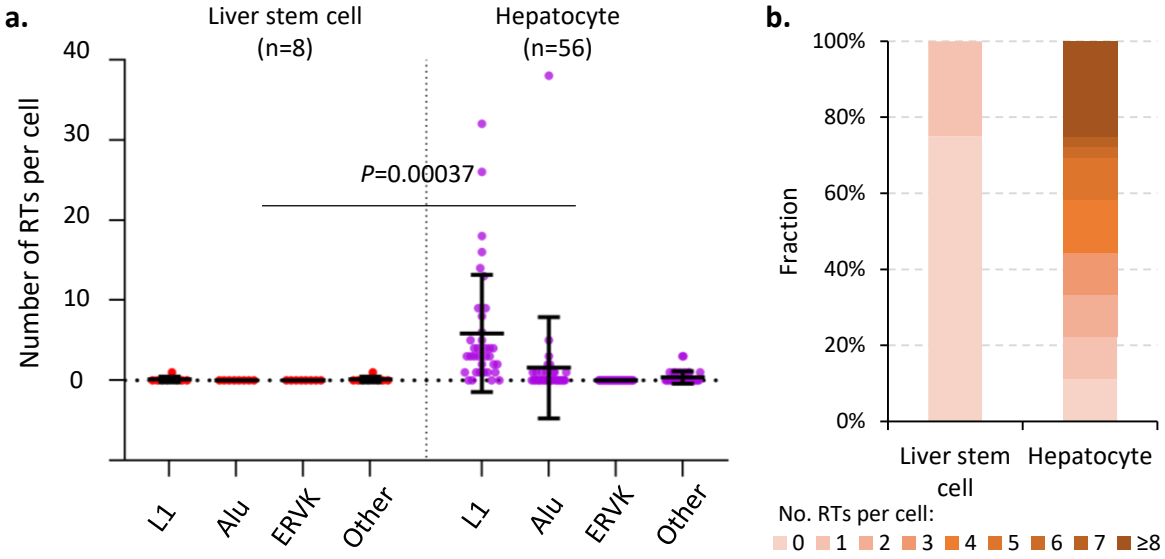
a.
Liver stem cell
(n=8)
Hepatocyte
(n=56)
Number of RTs per cell
40
30
20
10
0
P=0.00037
L1
Alu
ERVK
Other
L1
Alu
ERVK
Other
b.
100%
80%
60%
40%
20%
0%
Fraction
Liver stem cell
Hepatocyte
No. RTs per cell:
0 1 2 3 4 5 6 7 ≥8

Fig. 3.