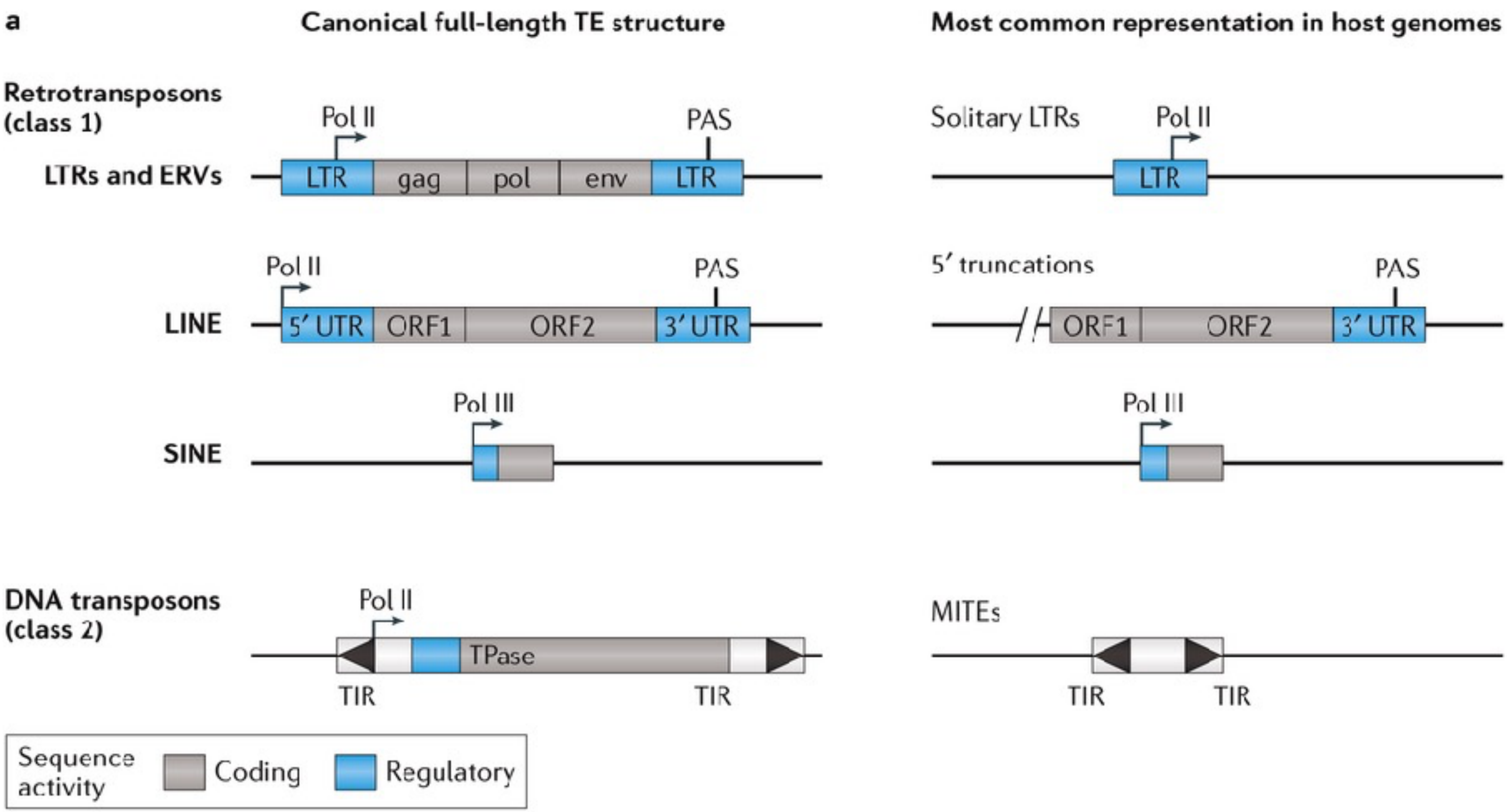
Transcription of L1 elements in the human somatic tissue

Group Meeting 2017

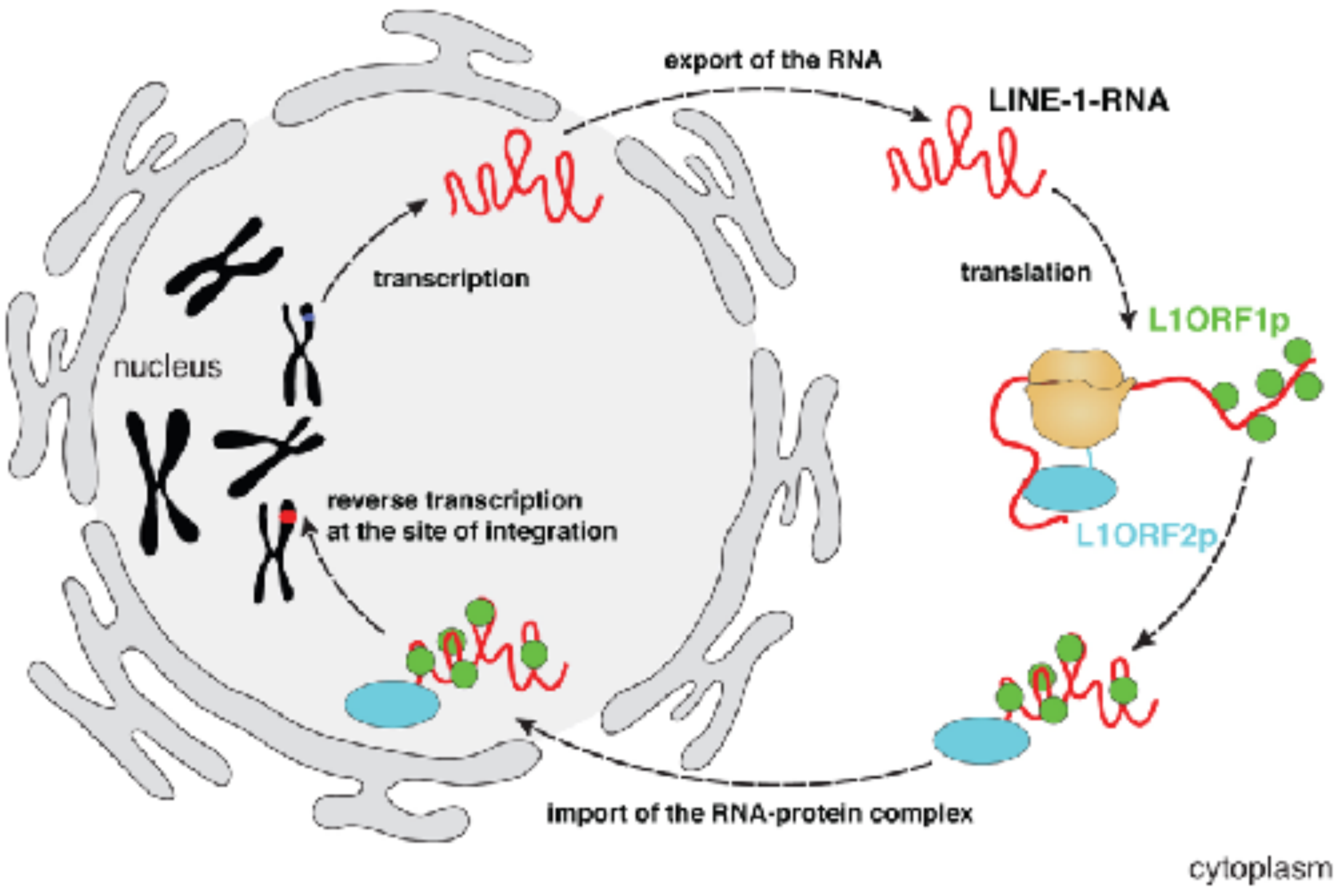
Fabio Navarro

What and how they are usually find in the genome



Chuong, E. B., Elde, N. C., & Feschotte, C. (2016). Regulatory activities of transposable elements: from conflicts to benefits. *Nature Reviews. Genetics*. http://doi.org/10.1038/nrg.2016.139

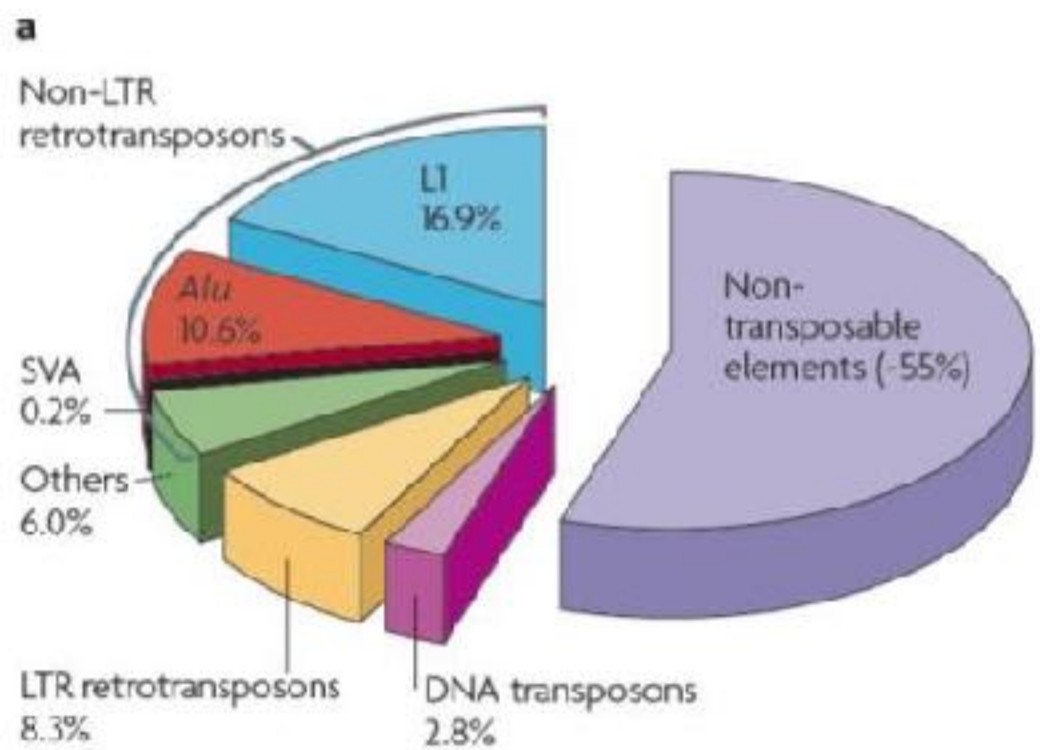
L1 Life cycle



- Preferentially retrotranspose the mRNA used during translation (cis-preference)
- Copy and paste mechanism

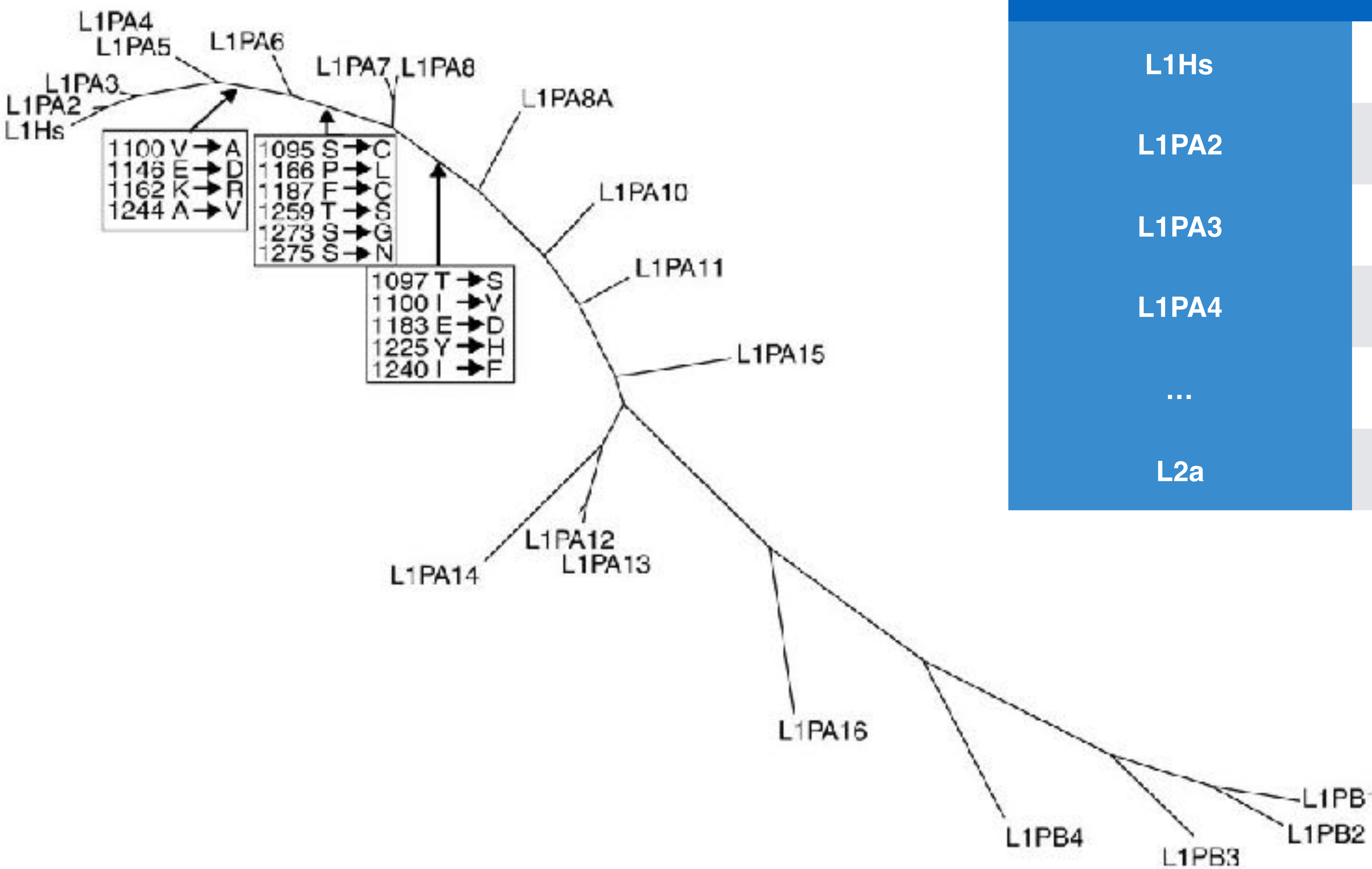
Non-LTR retrotransposons encode noncanonical RRM domains in their first open reading frame. E. Khazina, et al. PNAS, 2009.

TEs in the human genome



Cordaux, R., & Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nature Reviews. Genetics*, *10*(10), 691–703. http://doi.org/10.1038/nrg2640

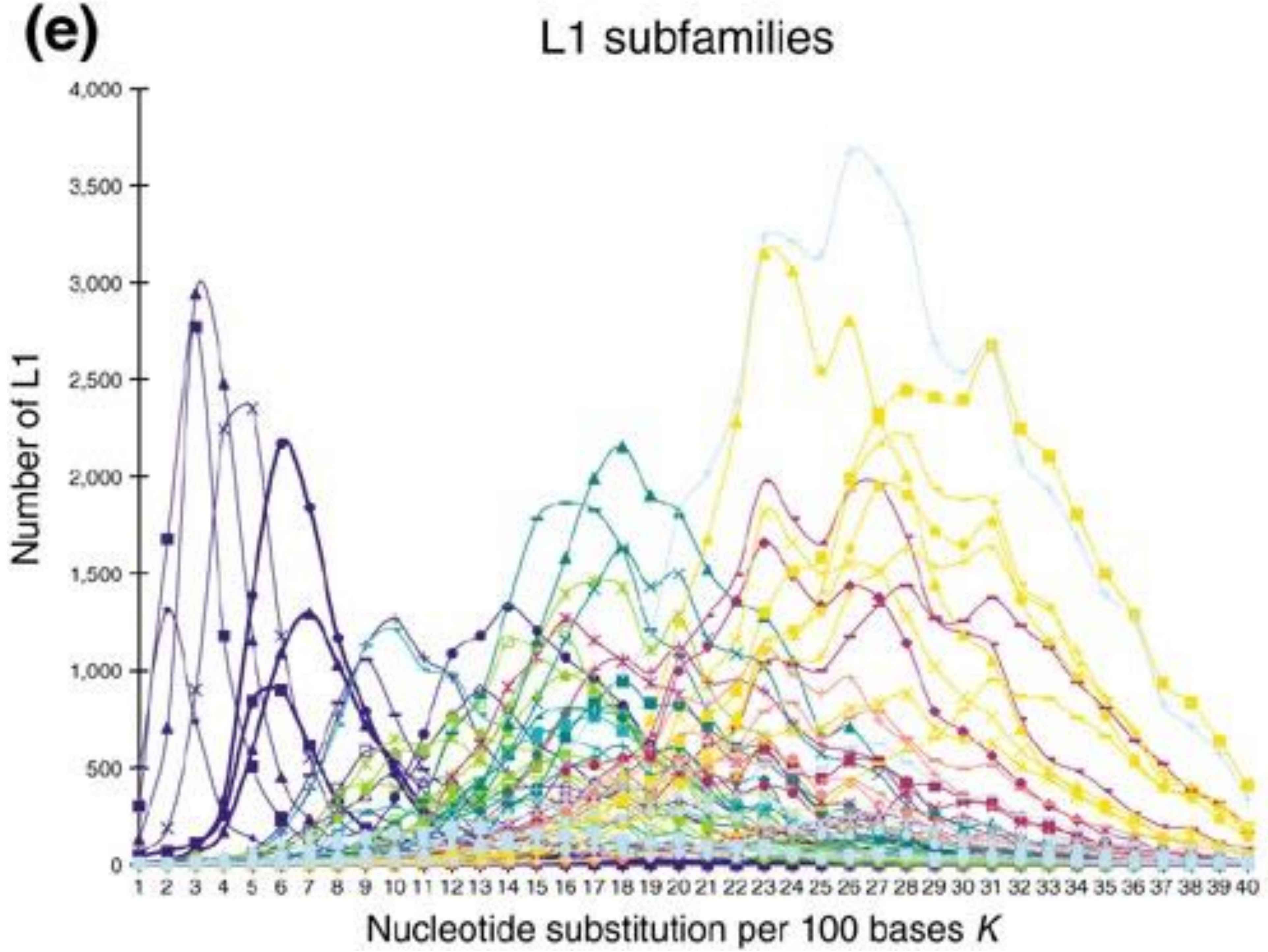
L1 Subfamilies



	Instances in HG38
L1Hs	1,686
L1PA2	5,113
L1PA3	11089
L1PA4	12272
…	…
L2a	174,058

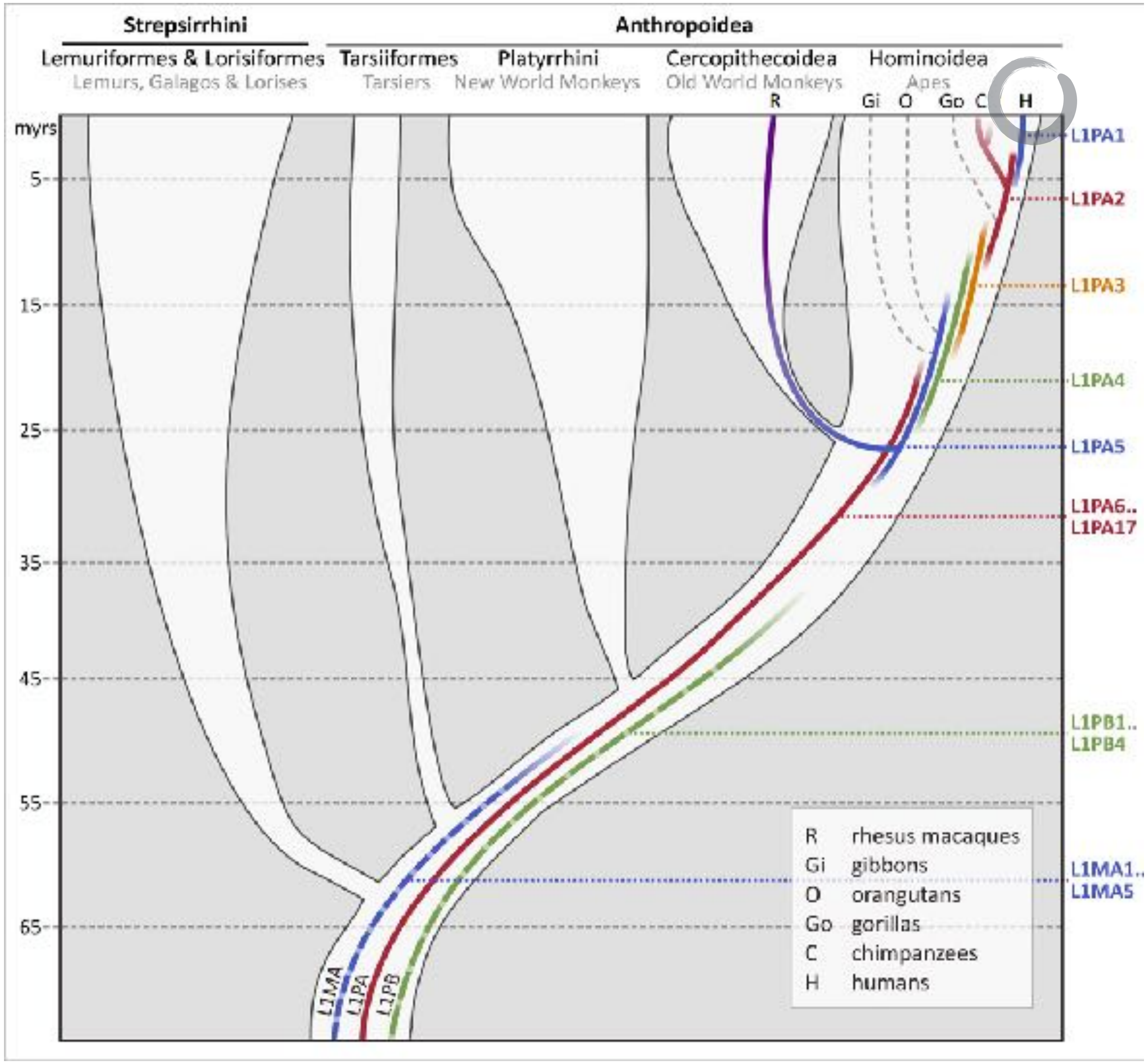
Ohshima, K., Hattori, M., Yada, T., Gojobori, T., Sakaki, Y., & Okada, N. (2003). Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. *Genome Biology*, *4*(11), R74. http://doi.org/10.1186/gb-2003-4-11-r74

L1 Subfamilies



Ohshima, K., Hattori, M., Yada, T., Gojobori, T., Sakaki, Y., & Okada, N. (2003). Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. *Genome Biology*, *4*(11), R74. http://doi.org/10.1186/gb-2003-4-11-r74

L1 Subfamilies

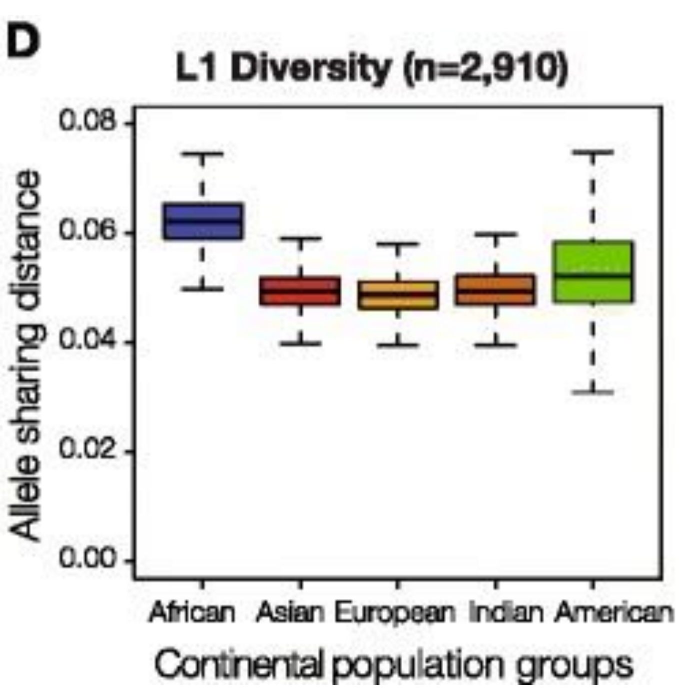


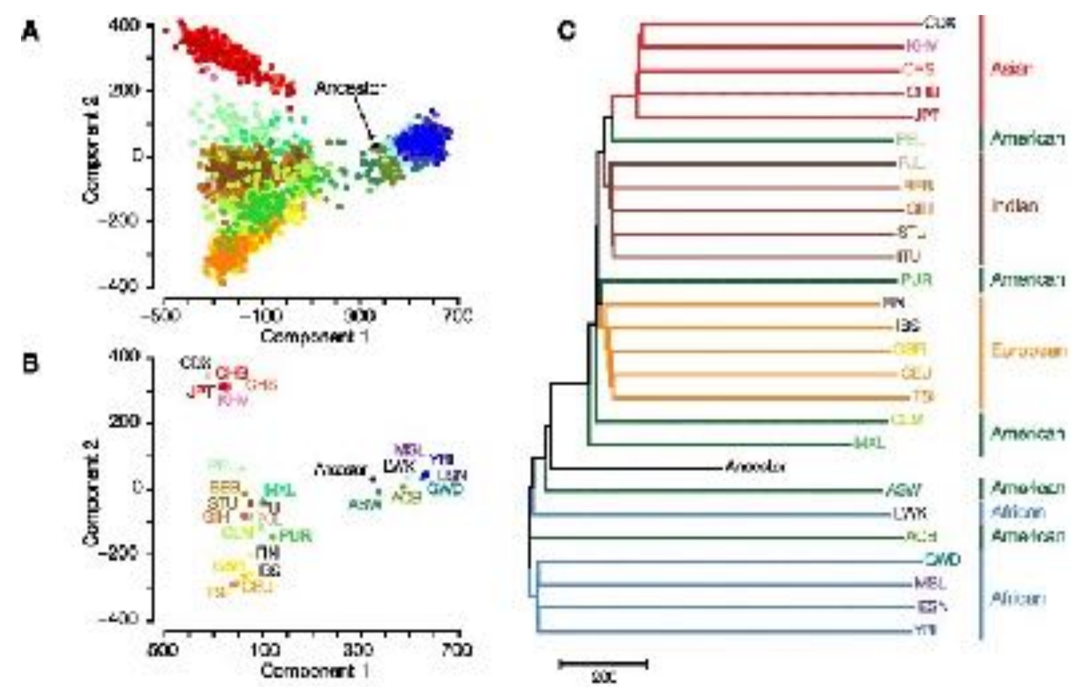
Konkel, M. K., Walker, J. A., & Batzer, M. A. (2011). LINEs and SINEs of primate evolution. *Evolutionary Anthropology: Issues, News, and Reviews*, *19*(6), 236–249. http://doi.org/10.1002/evan.20283

L1Hs is active in germiline

(and mostly L1Hs)

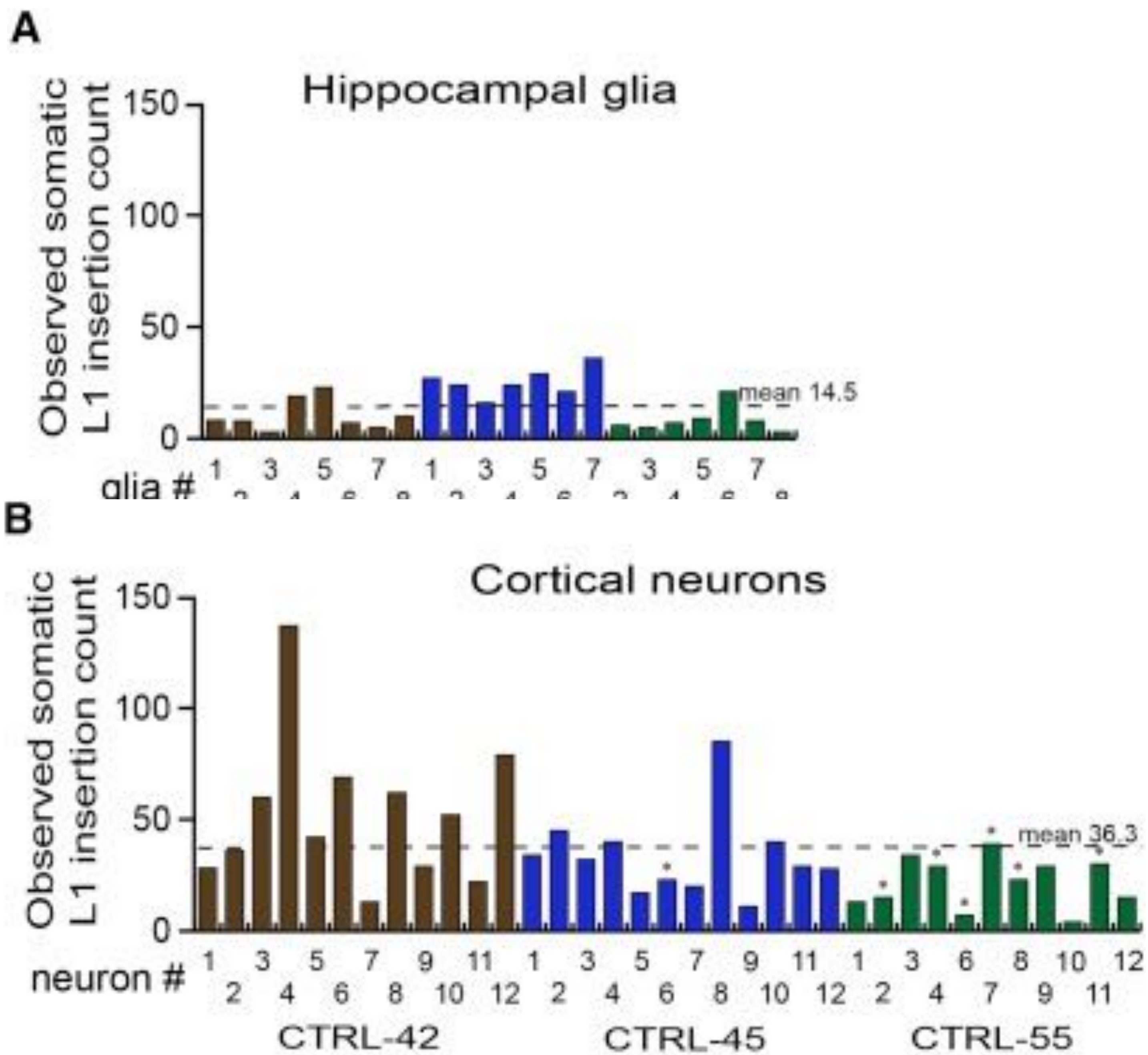
- dbRIP - Database of Transposable Elements presence/absence polymorphism: **> 90% of polymorphic sites are L1Hs**

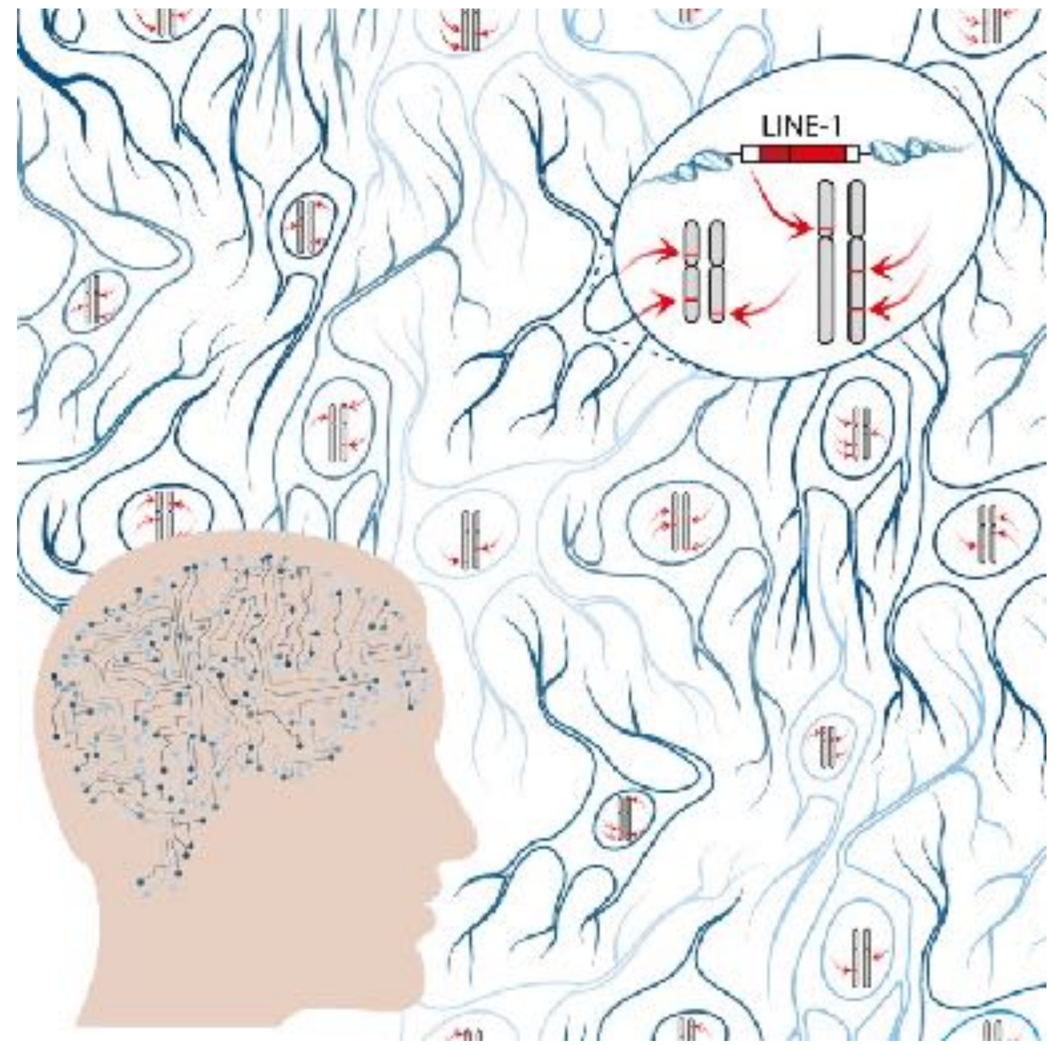




Wang, J., Song, L., Grover, D., Azrak, S., Batzer, M. A., & Liang, P. (2006). dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. *Human Mutation*, *27*(4), 323–329. http://doi.org/10.1002/humu.20307
Rishishwar, L., Tellez Villa, C. E., & Jordan, I. K. (2015). Transposable element polymorphisms recapitulate human evolution. *Mobile DNA*, *6*(1), 21. http://doi.org/10.1186/s13100-015-0052-6

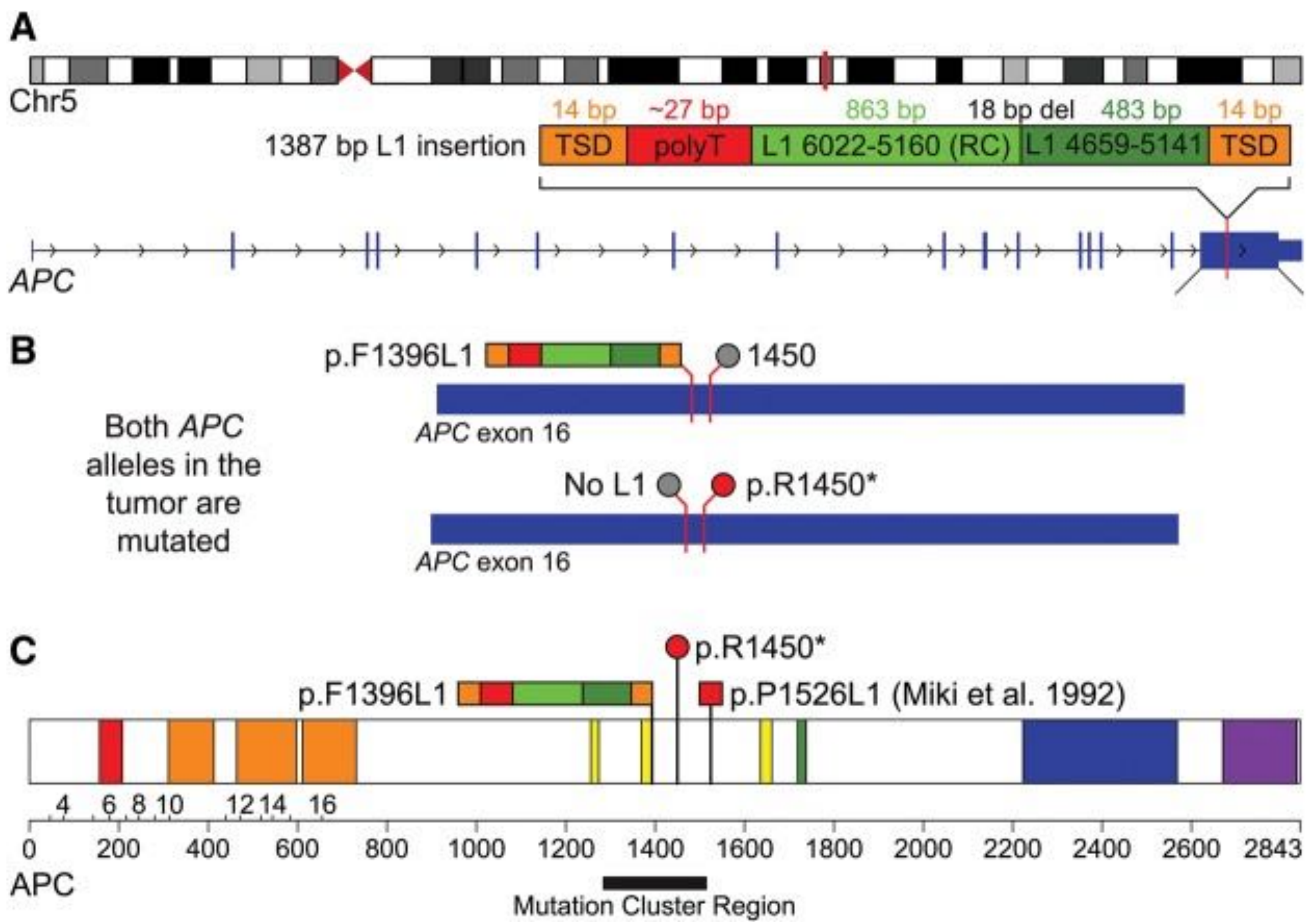
L1 as a **somatic** mutagenic factor





Chuong, E. B., Elde, N. C., & Feschotte, C. (2016). Regulatory activities of transposable elements: from conflicts to benefits. *Nature Reviews. Genetics*. http://doi.org/10.1038/nrg.2016.139

Tumorigenic L1



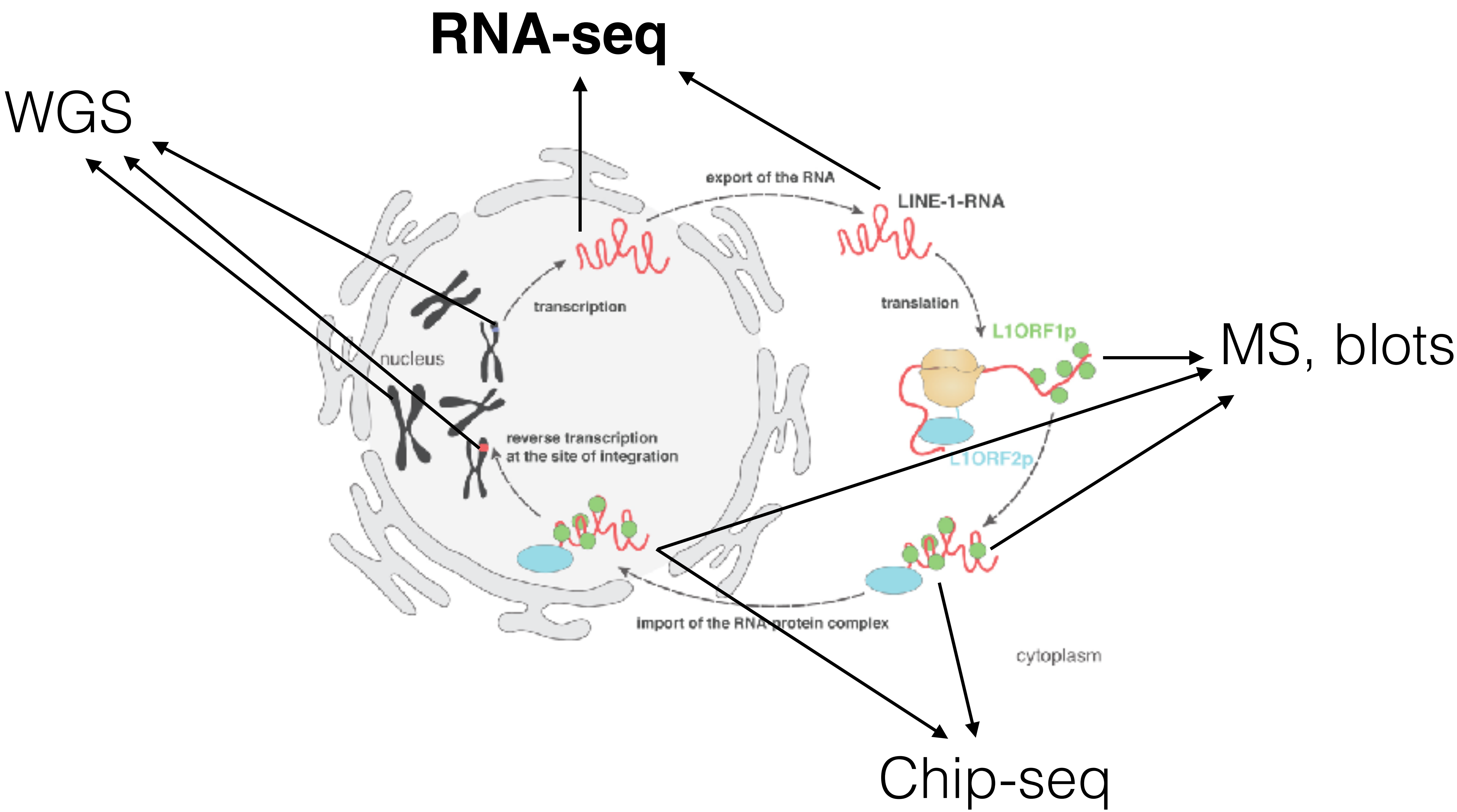
Scott, E. C., Gardner, E. J., Masood, A., Chuang, N. T., Vertino, P. M., & Devine, S. E. (2016). A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. *Genome Research*, *26*(6), 745–755. http://doi.org/10.1101/gr.201814.115

L1 is associated to other genetic diseases

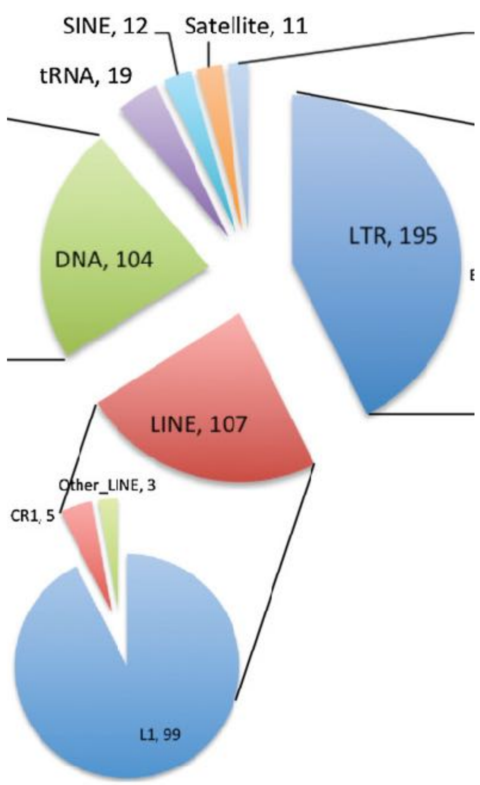
Disease	Gene disrupted by L1
Familial Retinoblastoma	RB1
β-thalassemia	HBB
(Fukuyama-type congenital) Muscular dystrophy	FKTN
Hemophilia A	FVIII
Hemophilia B	FIX
Cancer, cancer, cancer…	…

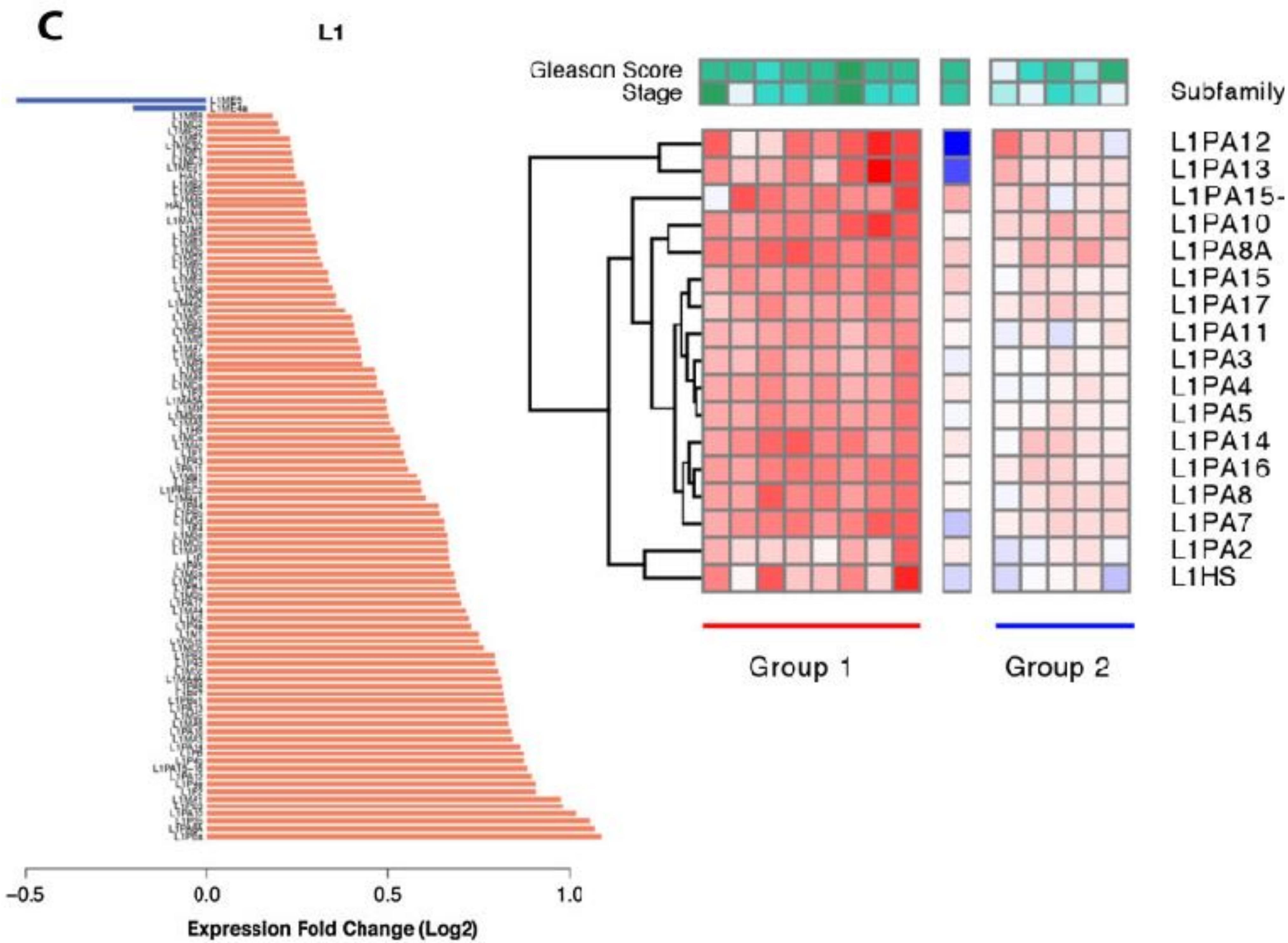
Hancks, D. C., & Kazazian, H. H. (2016). Roles for retrotransposon insertions in human disease. *Mobile DNA*, *7*(1), 9. http://doi.org/10.1186/s13100-016-0065-9

Detecting the activity of L1



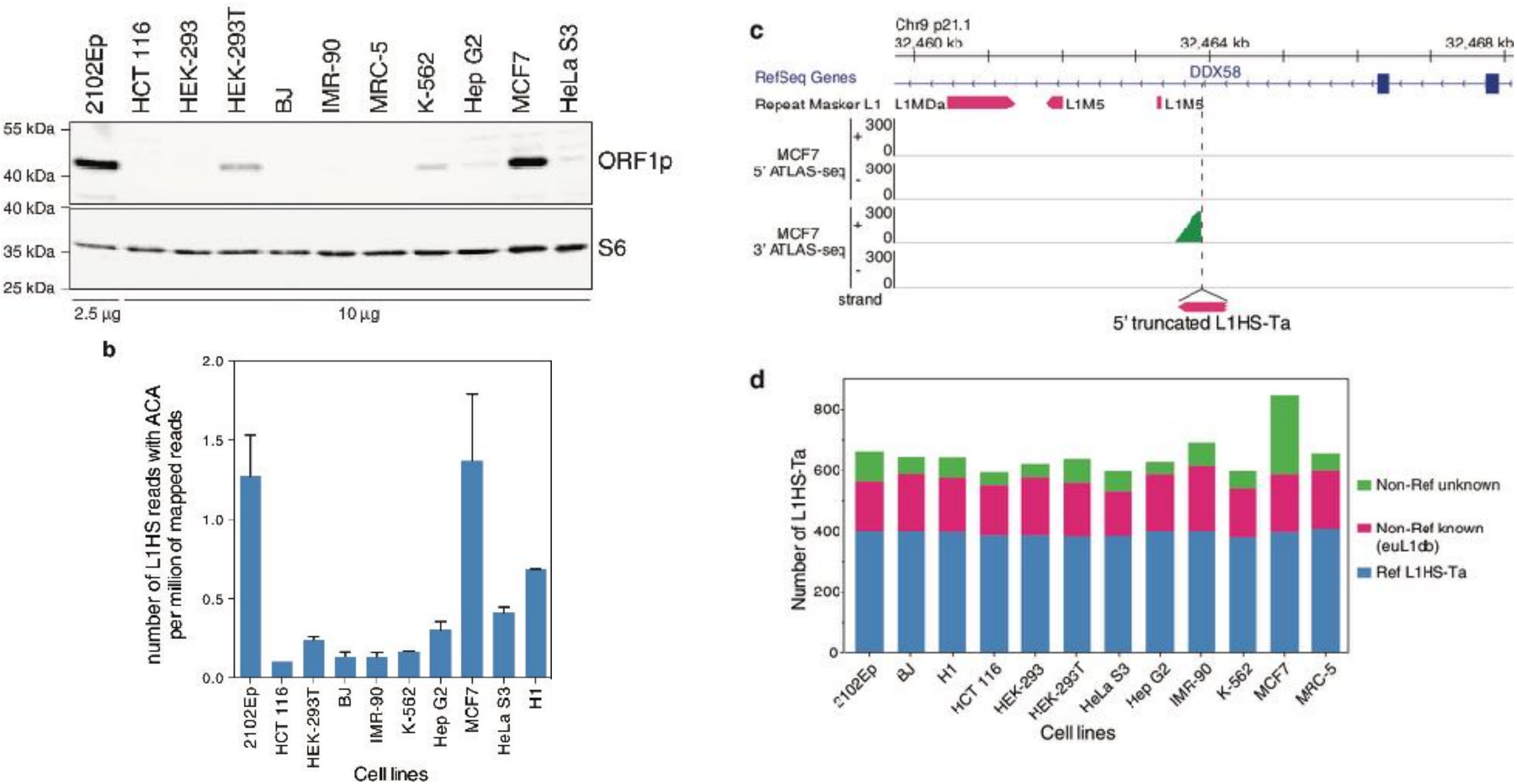
Transcriptional landscape of repetitive elements in normal and cancer human





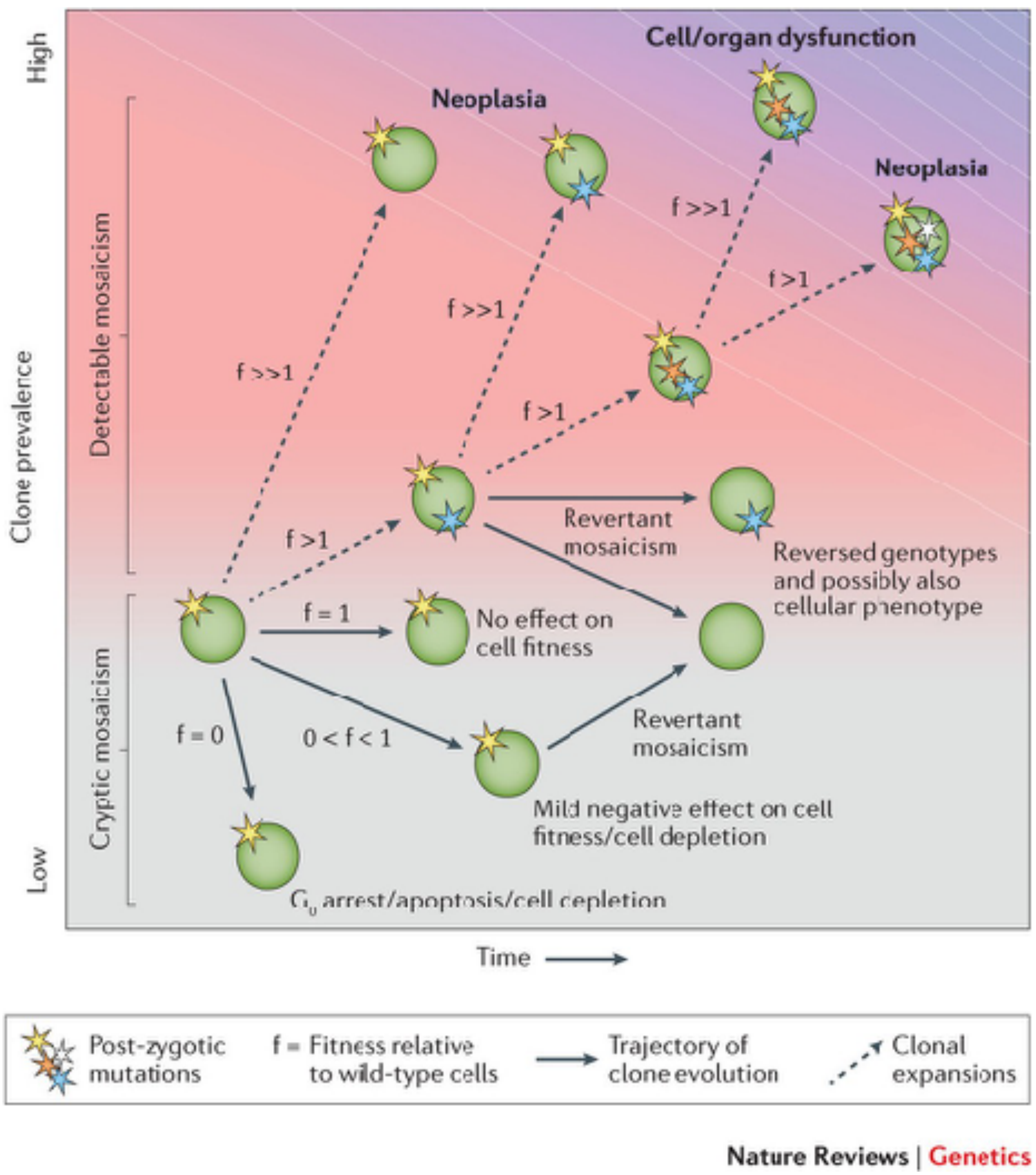
Criscione, S. W., Zhang, Y., Thompson, W., Sedivy, J. M., & Neretti, N. (2014). Transcriptional landscape of repetitive elements in normal and cancer human cells. *BMC Genomics*, *15*(1), 583–17. http://doi.org/10.1186/1471-2164-15-583

Activation of individual L1 retrotransposon instances is restricted to cell-type dependent permissive loci



Philippe, C., Vargas-Landin, D. B., Doucet, A. J., van Essen, D., Vera-Otarola, J., Kuciak, M., et al. (2016). Activation of individual L1 retrotransposon instances is restricted to cell-type dependent permissive loci. *eLife*, *5*, 166. http://doi.org/10.7554/eLife.13926

Somatic Variation

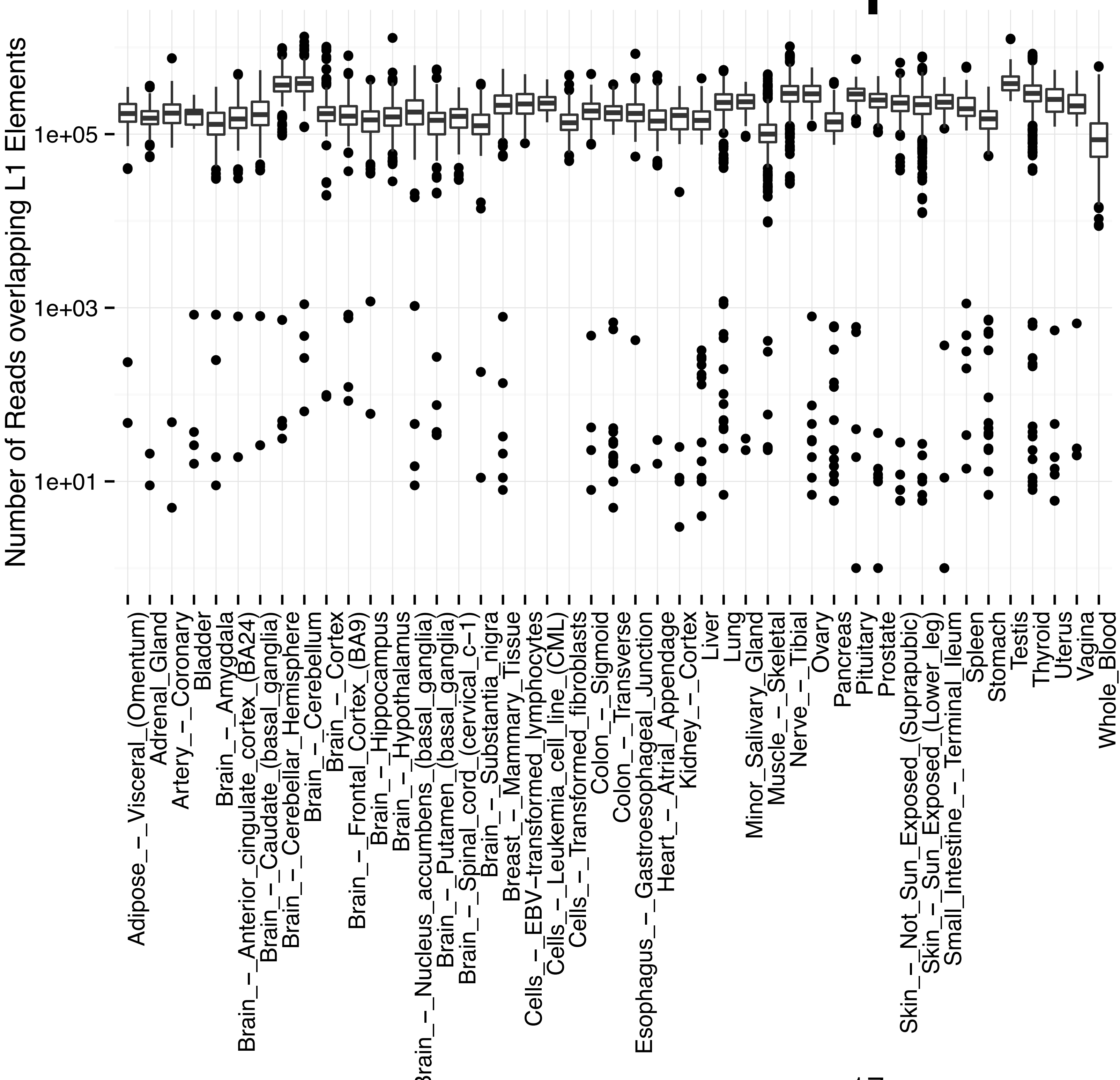


- The total number of mutations that can be expected to arise in the soma as a consequence of mitotic divisions is a function of two basic parameters: the number of cell divisions that occurred after conception and the mutation rate per cell division.
- Long interspersed nuclear element 1 retrotransposition have been shown to cause DNA copy-number alterations during embryogenesis, in neural precursors and in the adult brain.
- Alu element retrotransposition has been detected in human embryonic stem cells, as well as in the brain and myocardium.

Forsberg, L. A., Gisselsson, D., & Dumanski, J. P. (2016). Mosaicism in health and disease - clones picking up speed. *Nature Reviews. Genetics*. http://doi.org/10.1038/nrg.2016.145

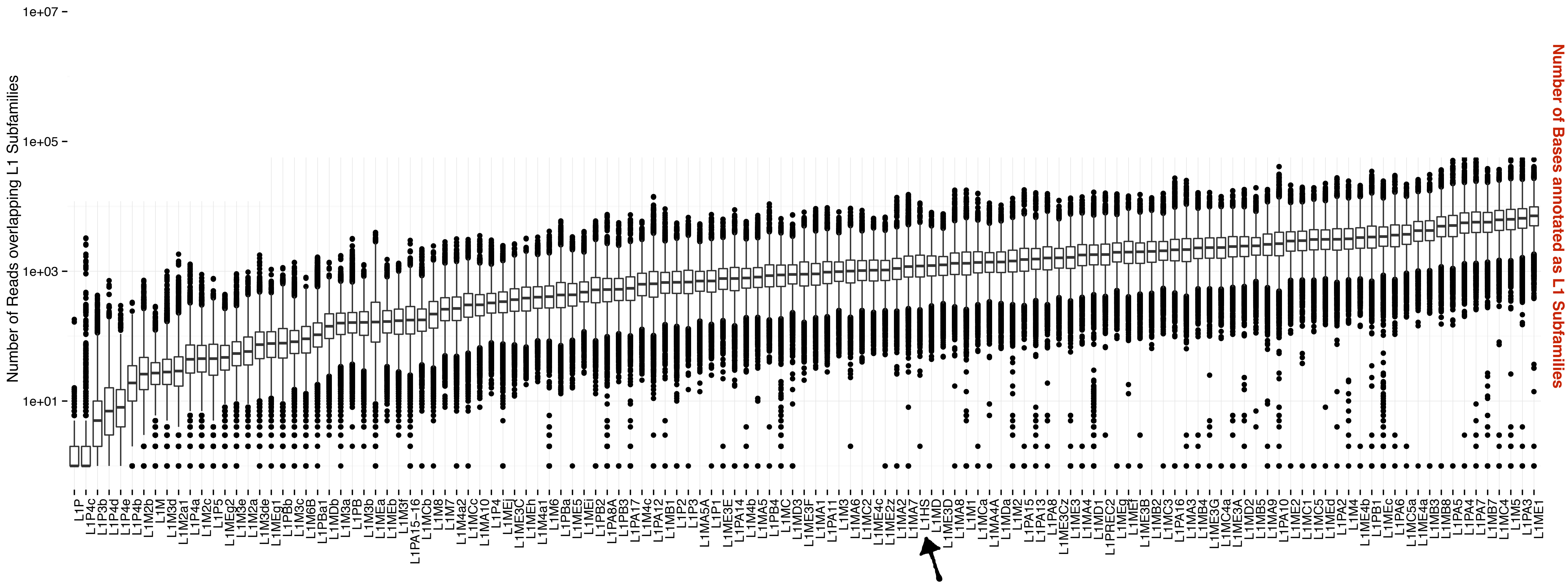
The method: TeXP

RNA-Seq and L1s



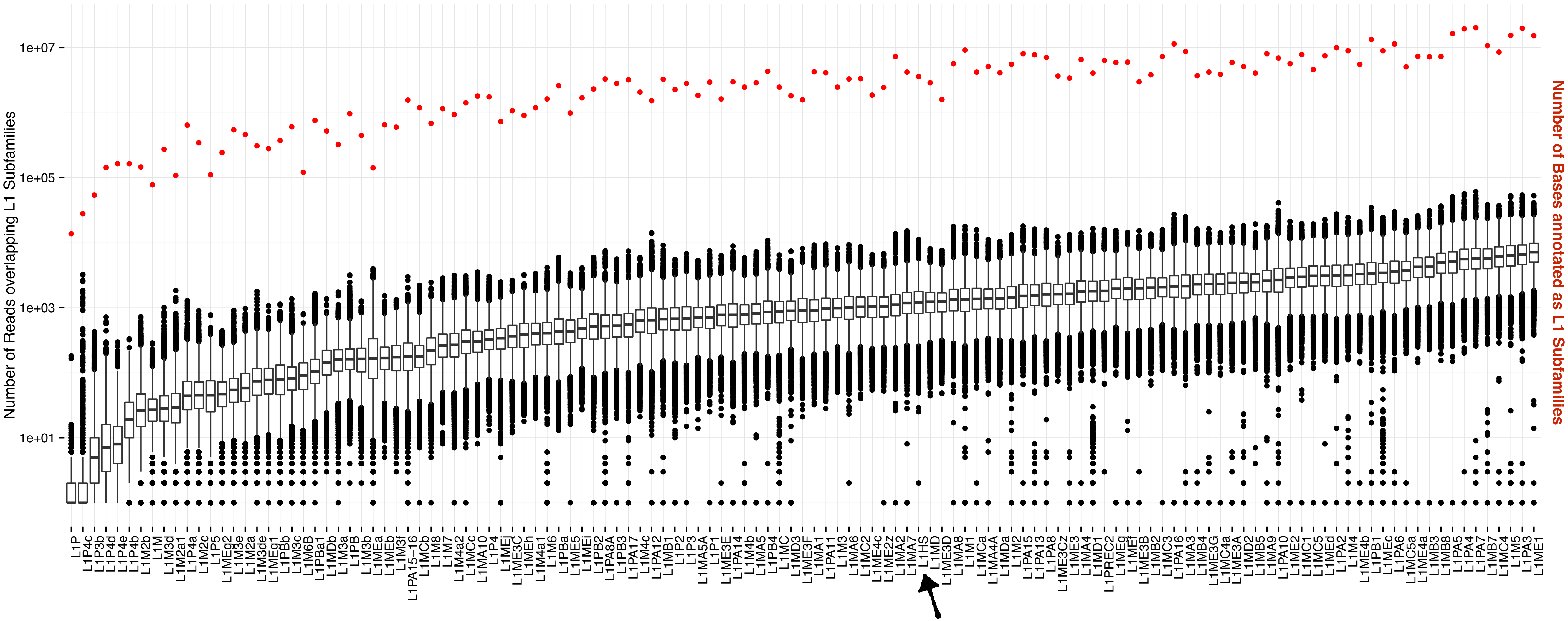
- Every RNA sequencing experiment has on average 200 thousand L1 reads.
- Or.. on average 0.5% of the reads in a RNA-seq reads map to L1 instances.
- With Cerebellum, Testis and a few other tissues showing higher levels of L1 levels.

Number per L1 subfamily



Number per L1 subfamily

(genome-transcriptome correlation)



What is a potential source of this signal?

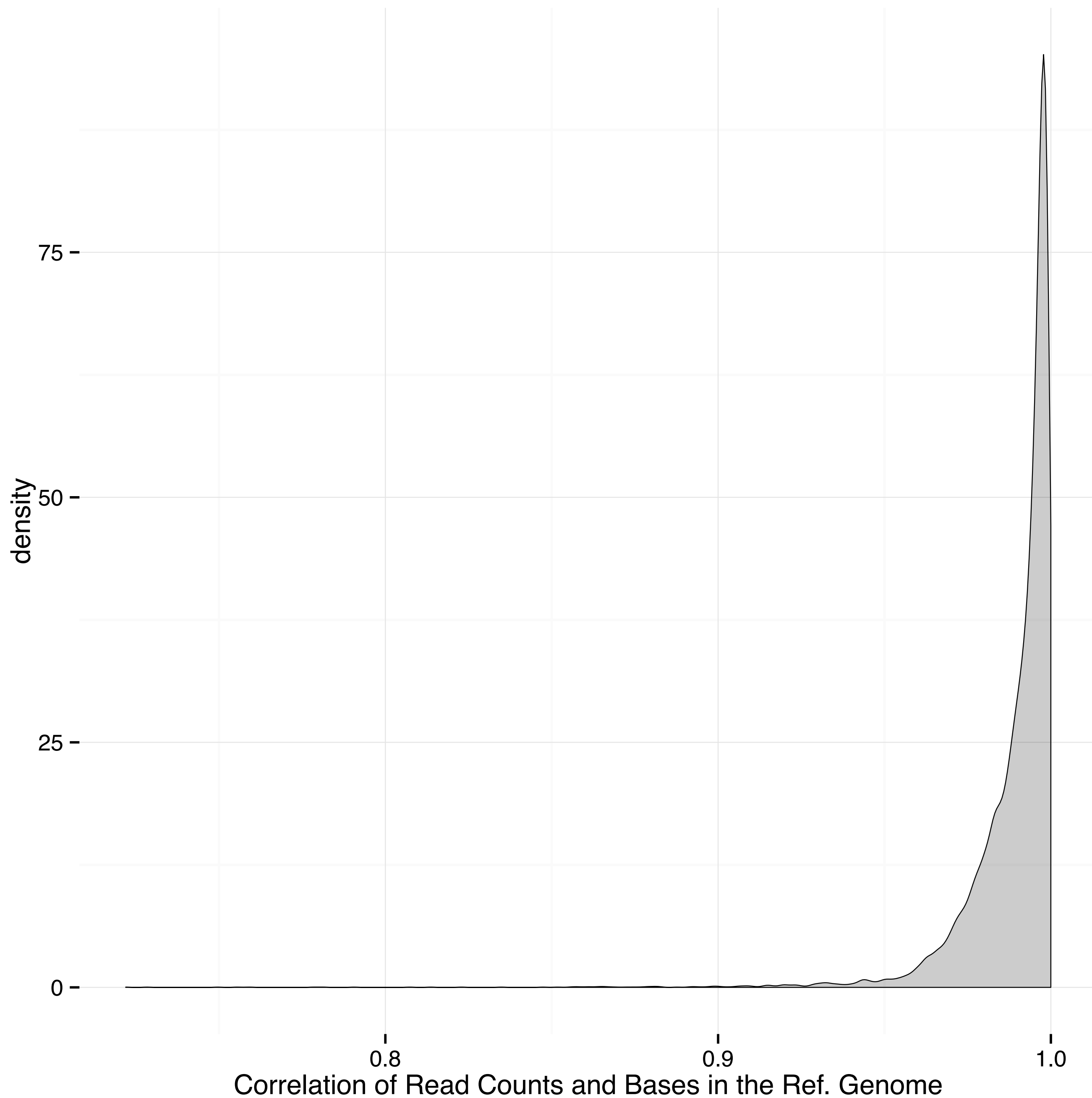
Pervasive transcription?

Pervasive transcription

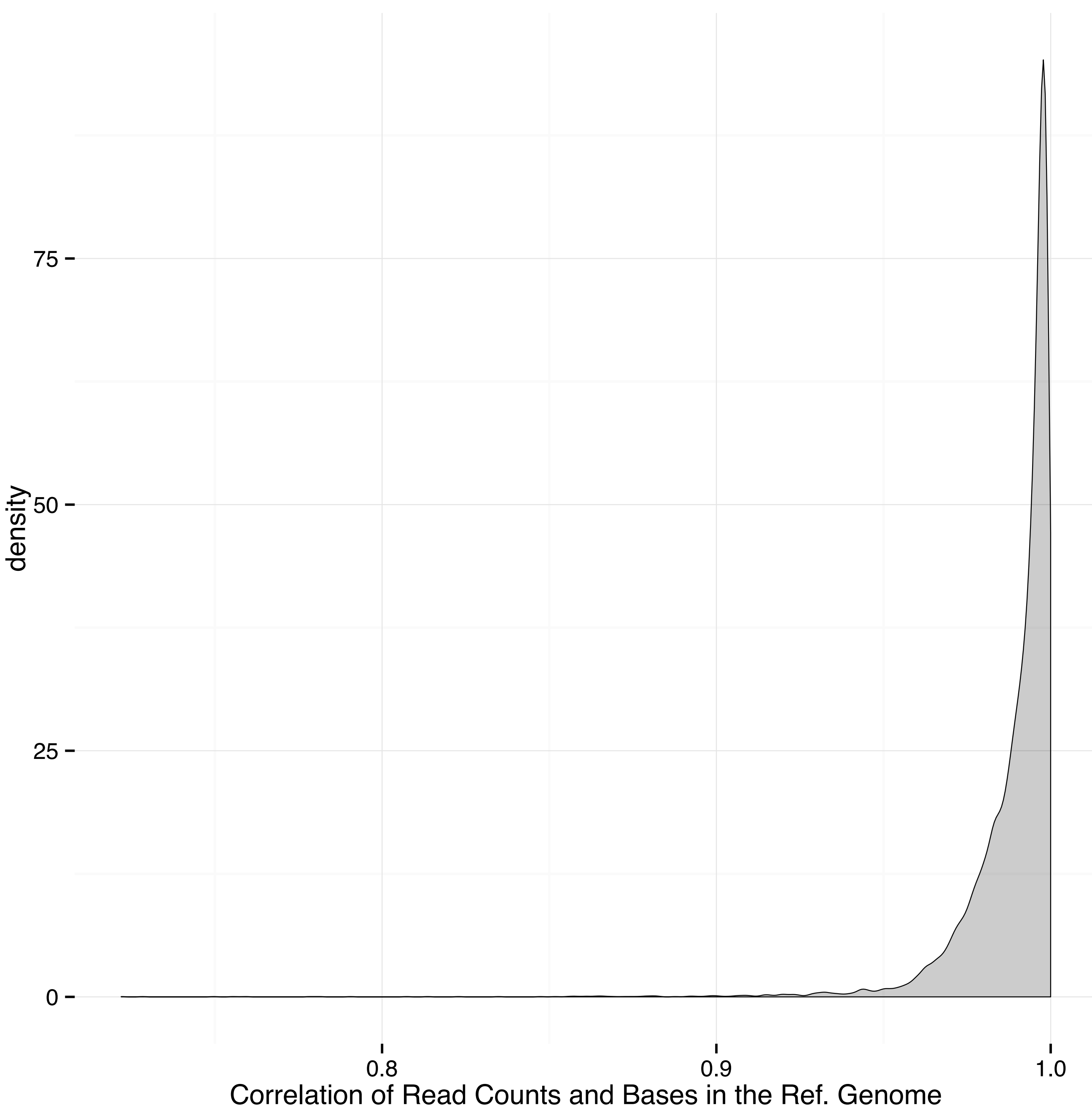
- The phenomena known as pervasive transcription is defined as the transcription of regions well beyond the boundaries of known genes.

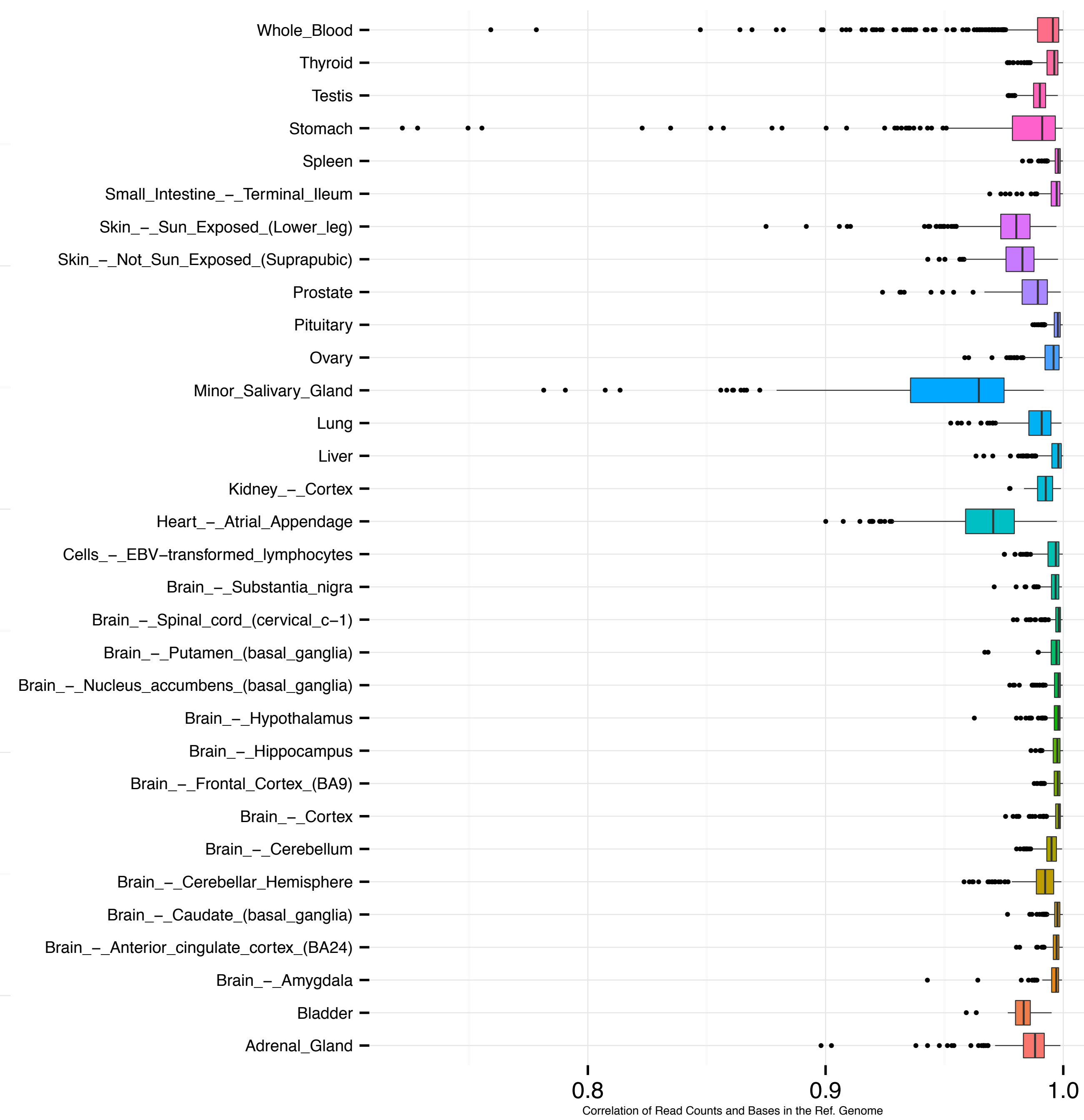
- Pervasive transcription does not affect the transcription level quantification of the transcription level of protein coding genes since they protein coding genes are present either as a single copy or low copy numbers in the genome. On the other hand, the transcription level quantification of L1 transposable elements, including L1 elements, transcription level is specially affected by pervasive transcription due to its multi-copy nature.

Most of RNA-seq samples have high genome-transcriptome correlation



Most of GTEx samples have high genome-transcriptome correlation





Model

- Read counts in L1 is a combination of **Pervasive transcription signal** and:
  - **L1Hs** autonomous transcription signal
  - L1PA2, L1PA3, etc. autonomous transcription signals

(BUT, in theory, older L1 subfamilies are not expected to be active (they are > 8My old and degraded) - plus, we have no evidence of recent retrotransposition of their transcripts.)

Model

$$N_i = t * (P_j * S_{i,j})$$

- N_i = Number of reads overlapping subfamily i;
- P_j = Signal Proportion of subfamily or pervasive transcription j;
- $S_{i,j}$ = Proportion of signal j mapped to subfamily i;
- t = Total number of reads overlapping all i subfamilies;
- The vector P is the hidden variable

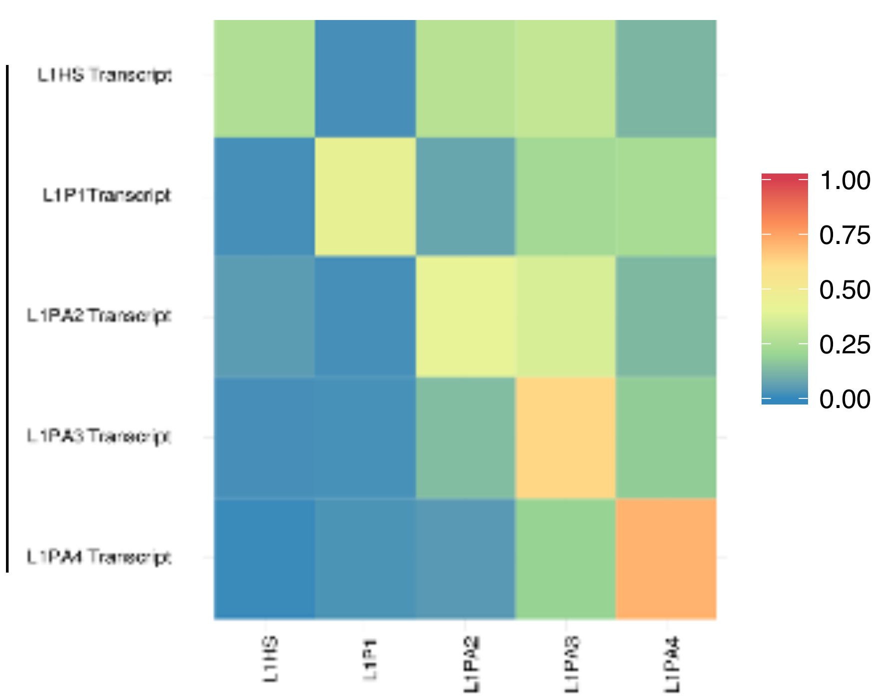
Signature matrix (mappability fingerprint)

1. Proportion of bases annotated as each subfamily is assumed as the **Pervasive Transcription** signal.

(1)

2. Based on simulations of reads originating from putative subfamily mature transcript, subfamily signal is defined by the Proportion of reads mapped to each subfamily.

(2)



On the L1 transcripts simulation (2)

1. Select putative full-length L1 transcripts;

2. Simulate reads of N base pairs and 0.1% error rate;

3. Align to the reference genome and;

4. Count the number of reads overlapping L1 subfamilies

   ps. randomly picking one of the best alignments (counting the alignment multiple times yielded similar results).

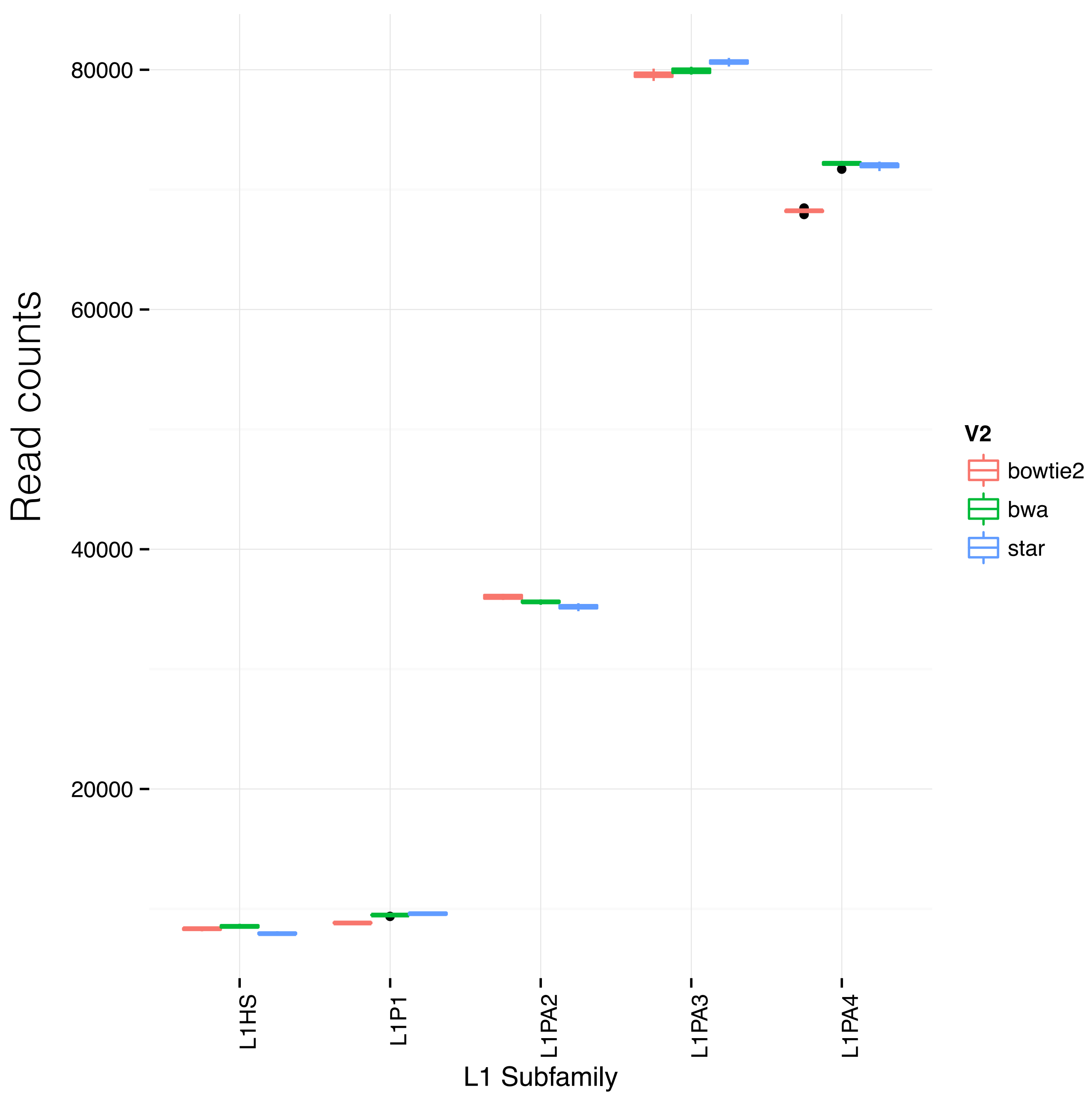
Model

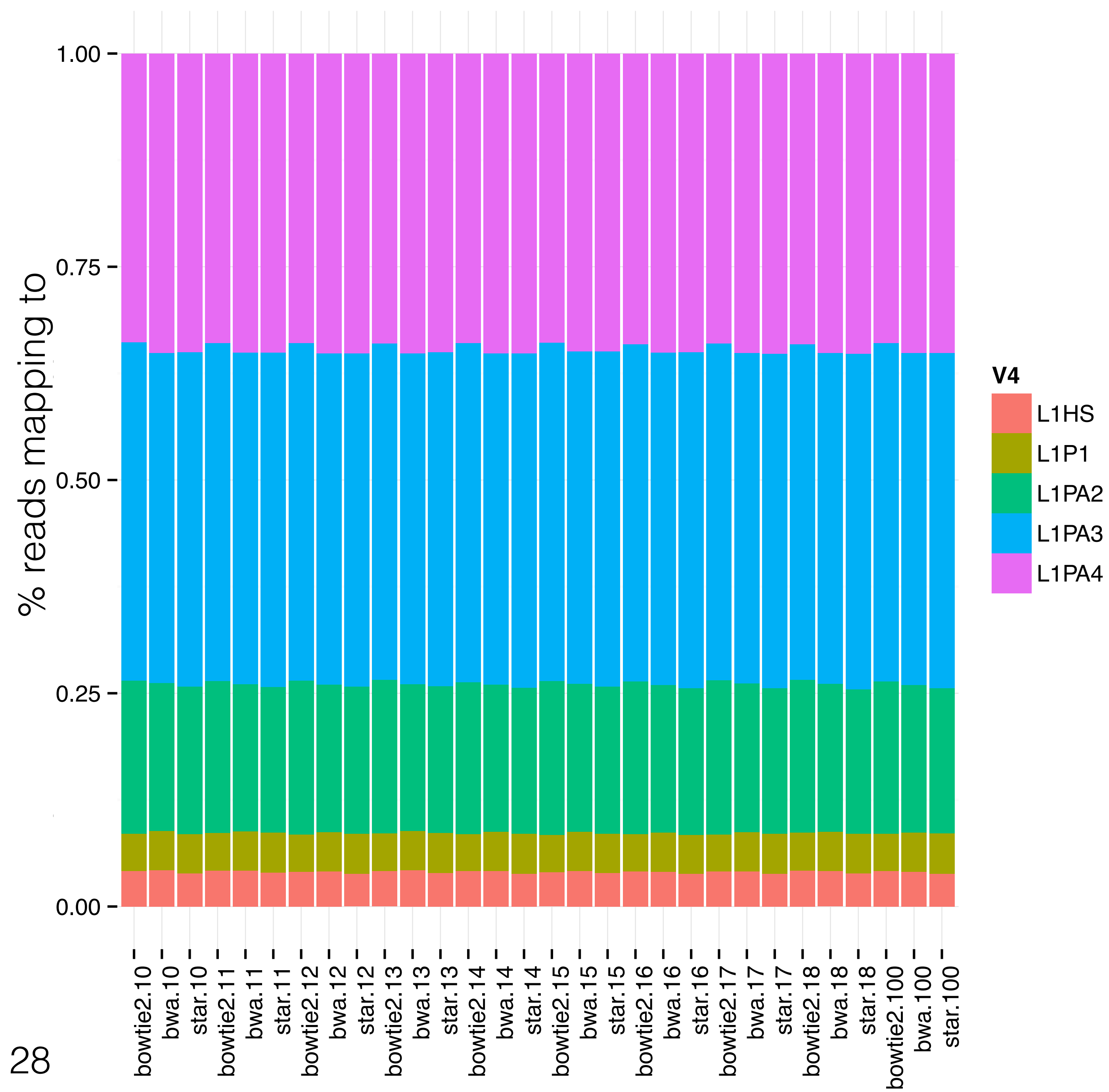
Ni=t*(Pj*Si,j)

- Now this becomes a simple regression problem:
  - Least Squares with Equalities and Inequalities (lsei)
  - Mixed Membership (mixedMem - Erosheva et al (2004))
  - LASSO (penalized)

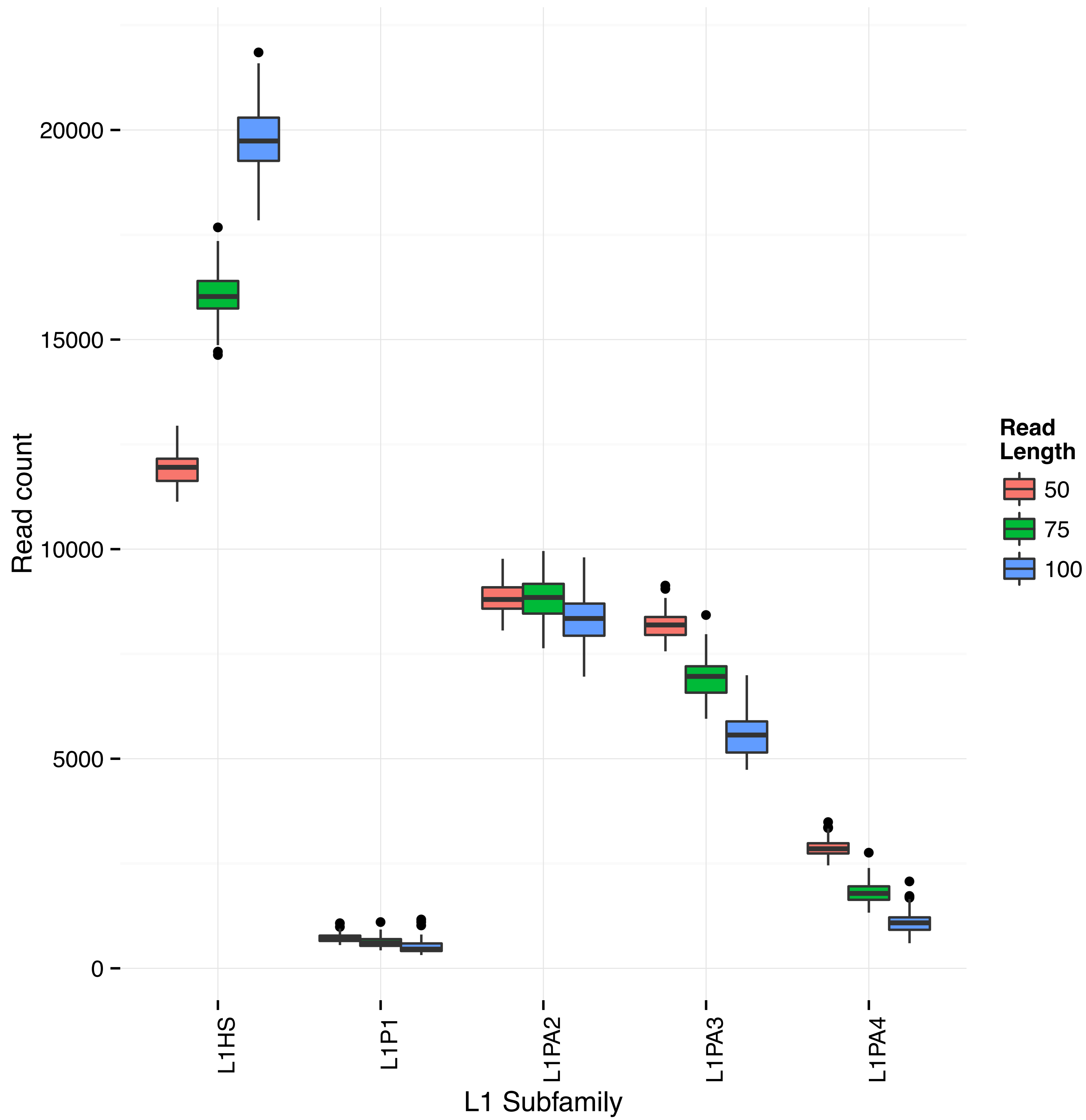
LASSO regression end up being the best method due to the expected sparsity.

Aligner assessment





Read length assessment



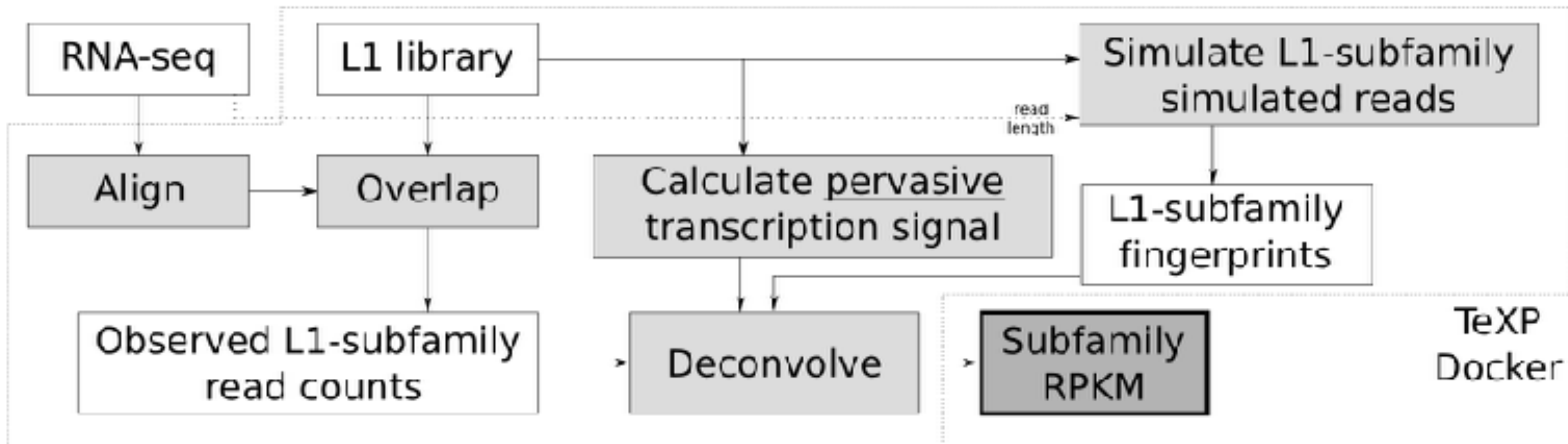
TeXP

- Is a tool to simulate reads compatible to a RNA-seq experiment and calculate the mappability fingerprint for L1 elements;

- It maps RNA-seq reads to a reference genome and uniformly quantify the L1 subfamily read counts;

- Finally, TeXP estimates the rate of pervasive transcription and autonomous transcription of L1 subfamilies;

TeXP

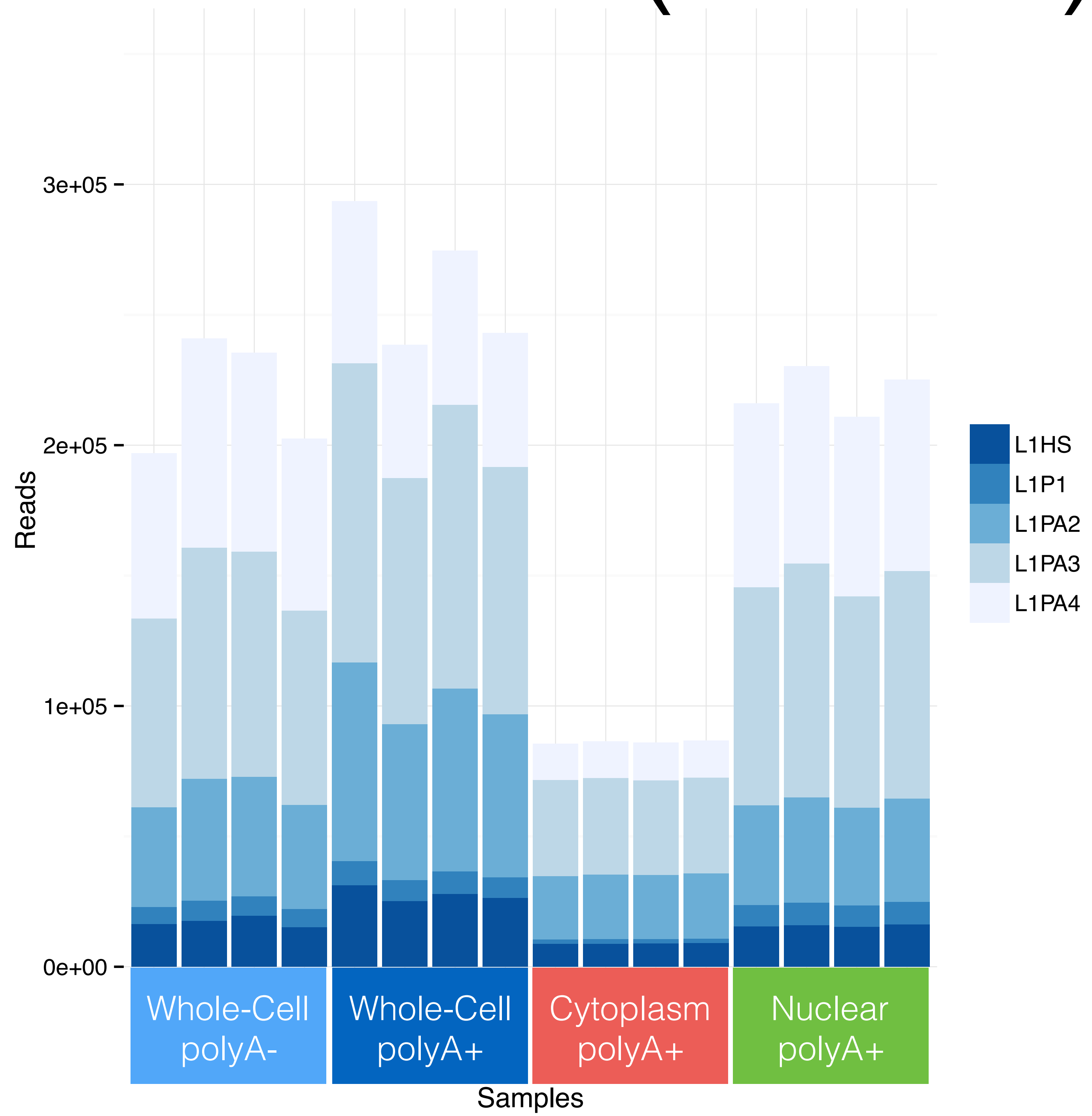


- TeXP can be used as a Makefile or as a Docker Image (cloud compatible).
- Source code is (not) available on GitHub. github.com/fabiocpn/TeXP
- And (not) available in DockerHub (fnavarro/texp)

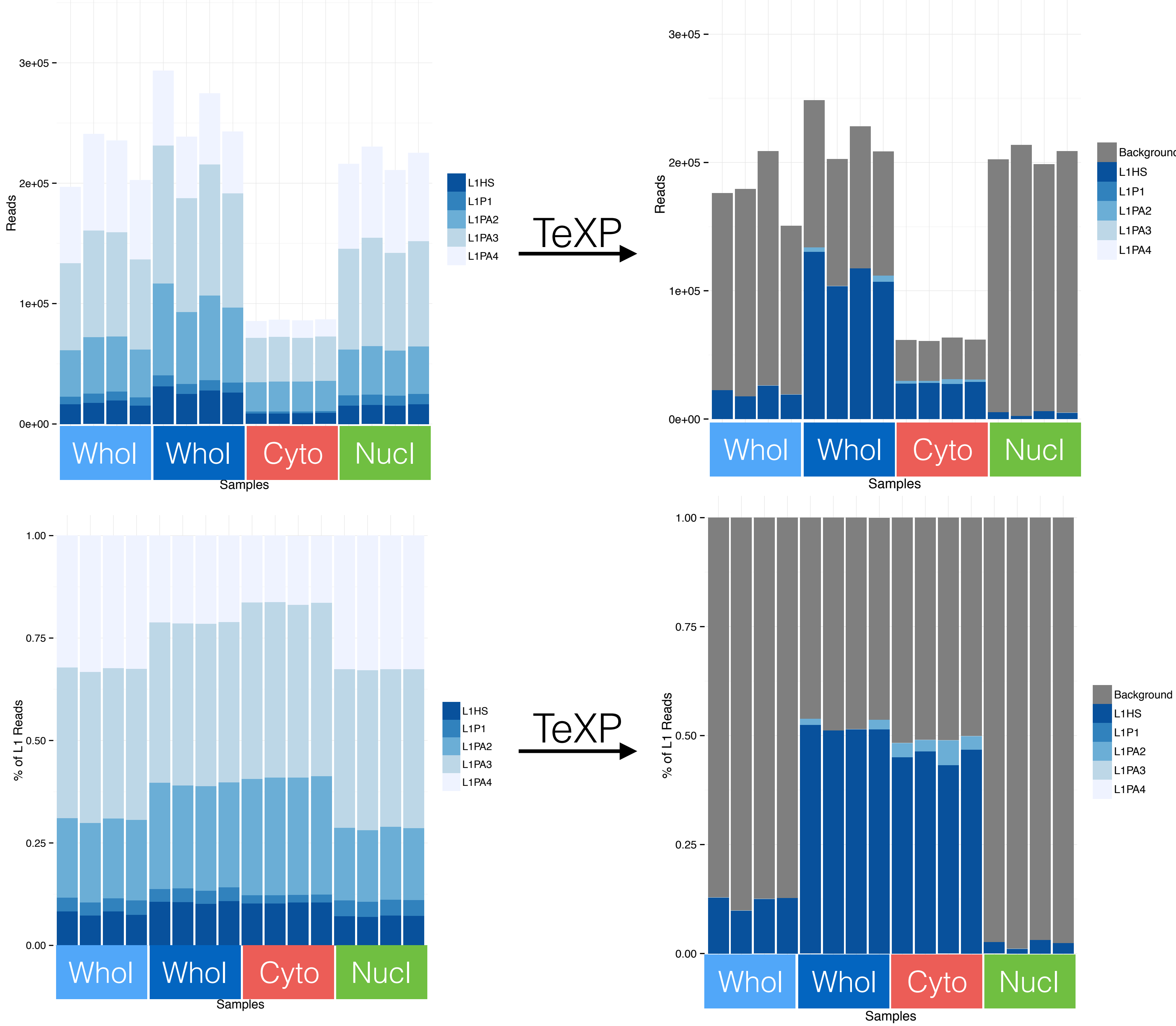
L1XP?
S'more?

TeXP in cancer cell lines

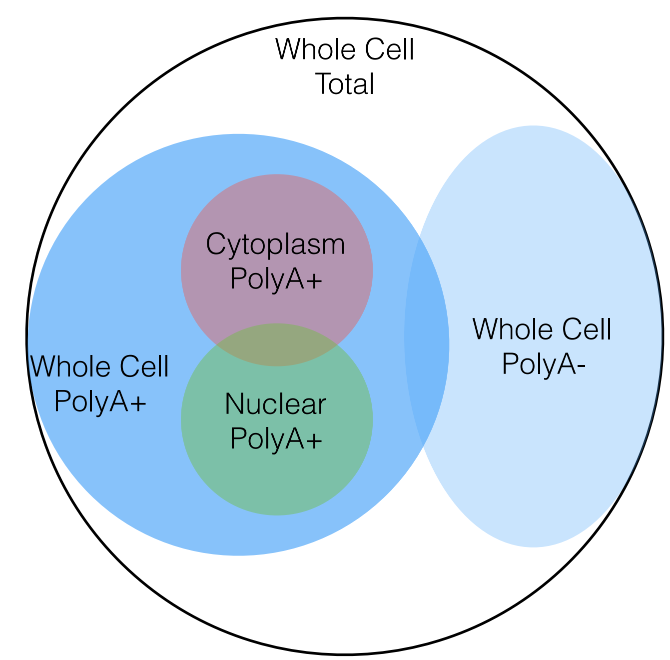
Analysis of ENCODE cell-lines (MCF7)



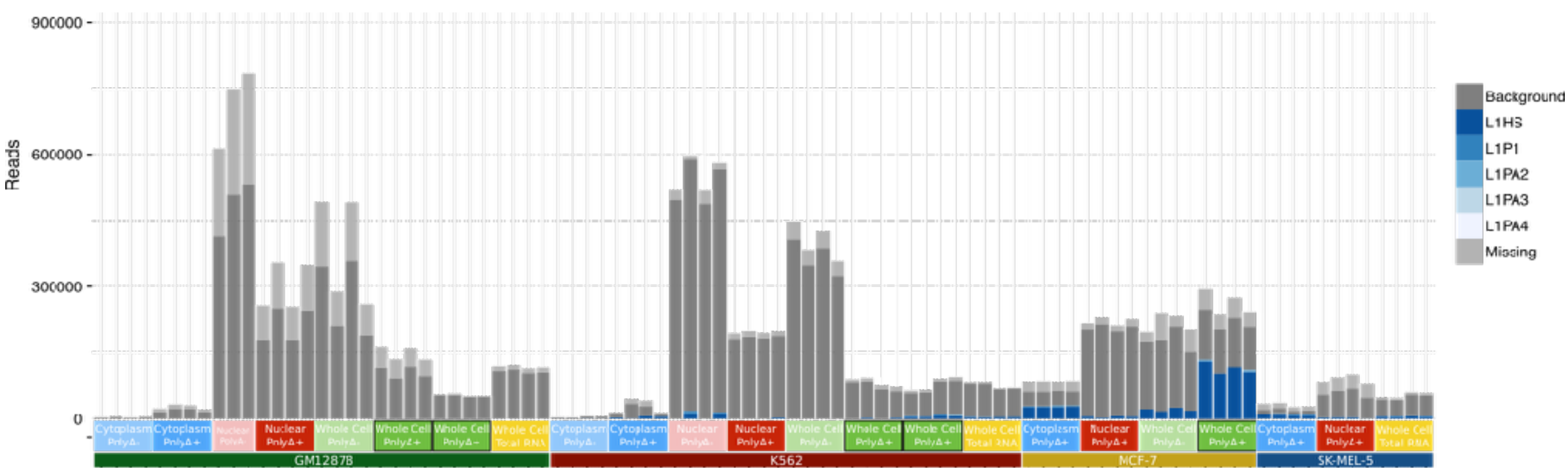
Analysis of ENCODE cell-lines (MCF7)



	L1Hs (RPKM)	Signal/Noise
WholeCell PolyA-	48.5	0.12
WholeCell PolyA+	180.7	0.51
Cytoplasm PolyA+	33.2	0.45
Nuclear PolyA+	6.2	0.02



Analysis of ENCODE cell-lines

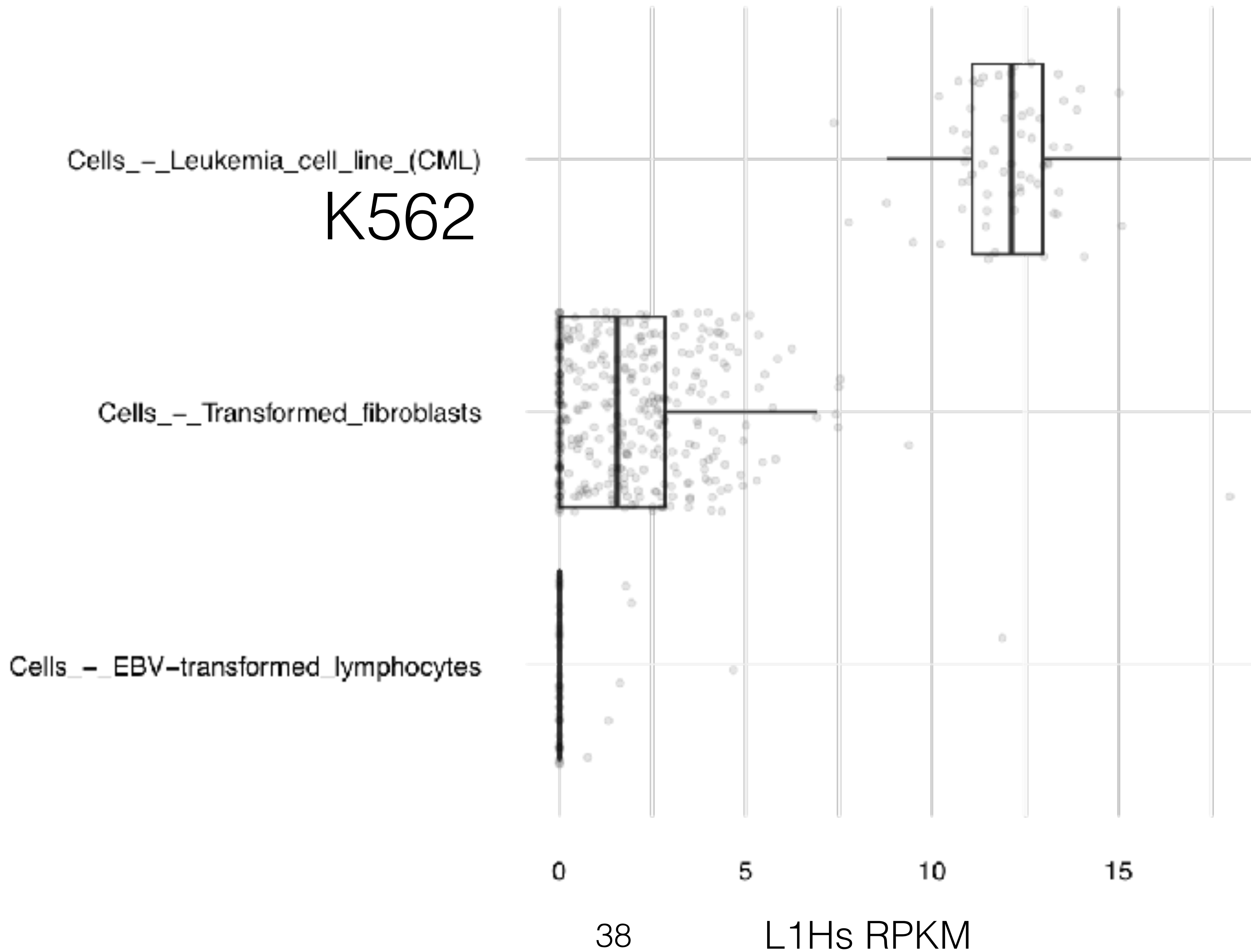


L1Hs autonomous transcription in healthy tissues

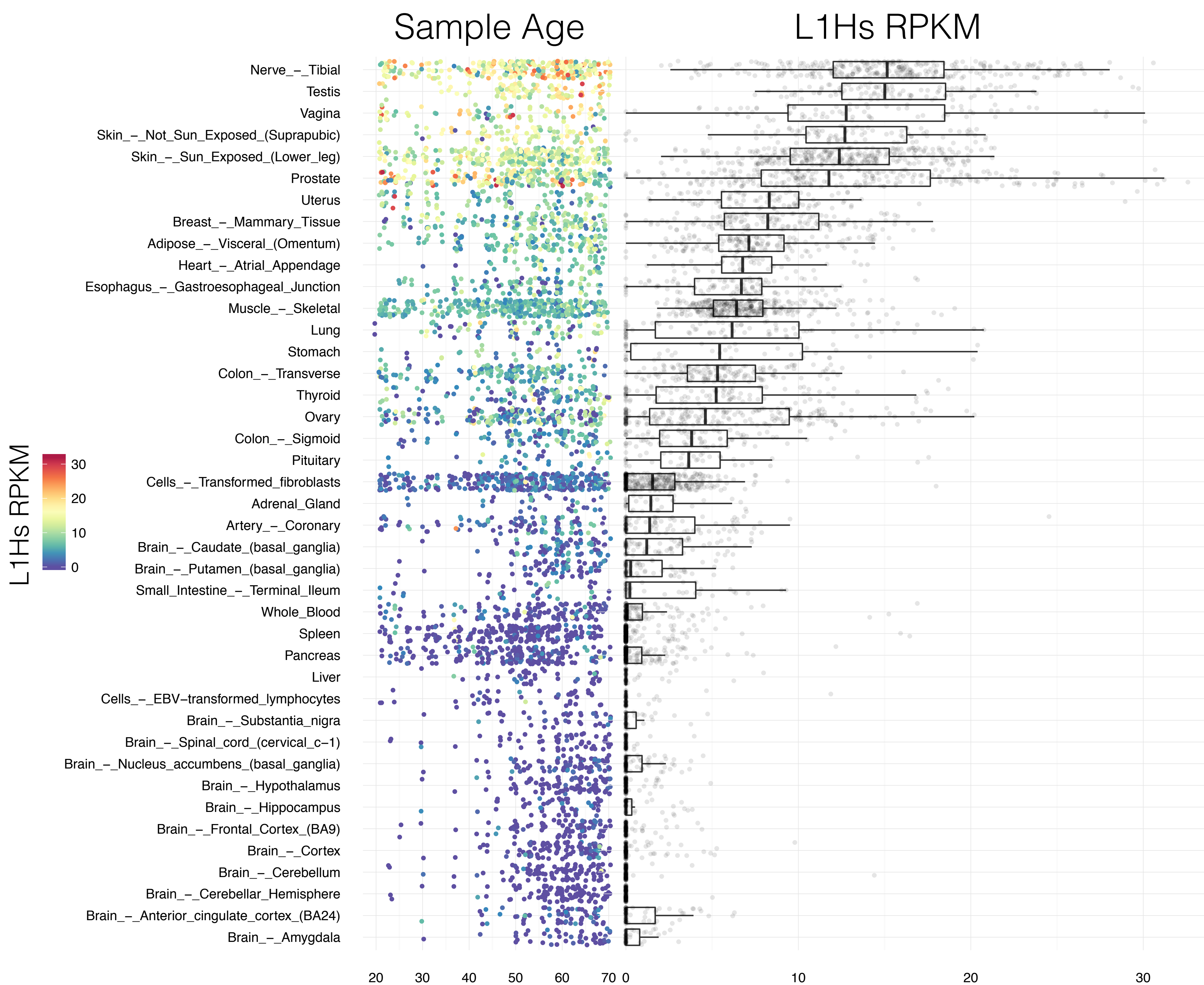
GTEx processed samples

N Samples	Tissue
234	Adipose_-_Visceral_(Omentum)
146	Adrenal_Gland
123	Artery_-_Coronary
11	Bladder
81	Brain_-_Amygdala
99	Brain_-_Anterior_cingulate_cortex_(BA24)
133	Brain_-_Caudate_(basal_ganglia)
115	Brain_-_Cerebellar_Hemisphere
145	Brain_-_Cerebellum
132	Brain_-_Cortex
117	Brain_-_Frontal_Cortex_(BA9)
102	Brain_-_Hippocampus
103	Brain_-_Hypothalamus
123	Brain_-_Nucleus_accumbens_(basal_ganglia)
103	Brain_-_Putamen_(basal_ganglia)
76	Brain_-_Spinal_cord_(cervical_c-1)
71	Brain_-_Substantia_nigra
200	Breast_-_Mammary_Tissue
132	Cells_-_EBV-transformed_lymphocytes
78	Cells_-_Leukemia_cell_line_(CML)
300	Cells_-_Transformed_fibroblasts
142	Colon_-_Sigmoid
178	Colon_-_Transverse
151	Esophagus_-_Gastroesophageal_Junction
192	Heart_-_Atrial_Appendage
36	Kidney_-_Cortex
136	Liver
280	Lung
69	Minor_Salivary_Gland
468	Muscle_-_Skeletal
335	Nerve_-_Tibial
108	Ovary
193	Pancreas
124	Pituitary
119	Prostate
271	Skin_-_Not_Sun_Exposed_(Suprapubic)
395	Skin_-_Sun_Exposed_(Lower_leg)
104	Small_Intestine_-_Terminal_Ileum
118	Spleen
205	Stomach
199	Testis
355	Thyroid
90	Uterus
88	Vagina
449	Whole_Blood

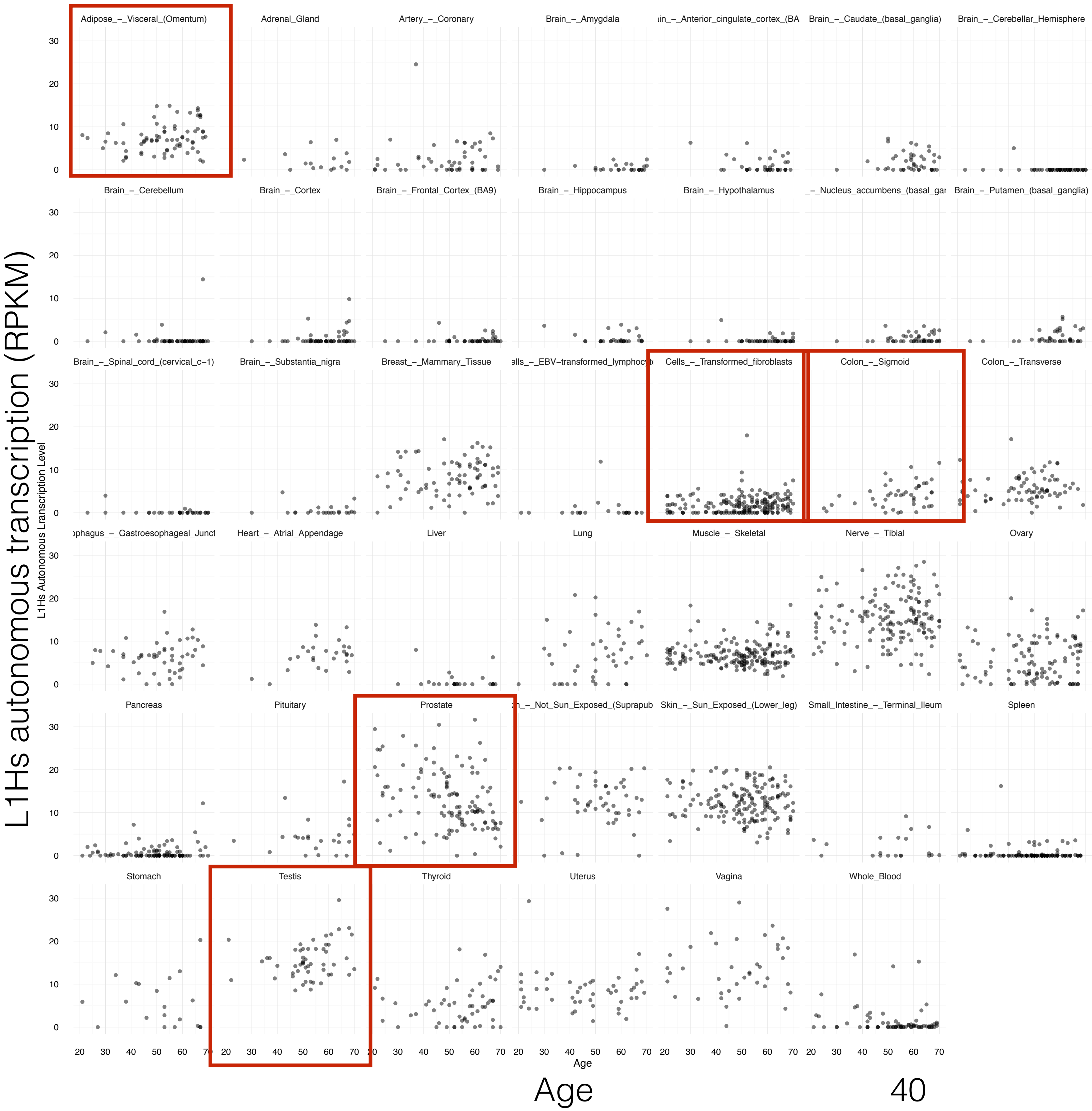
GTEx cell lines



L1Hs activity in the soma

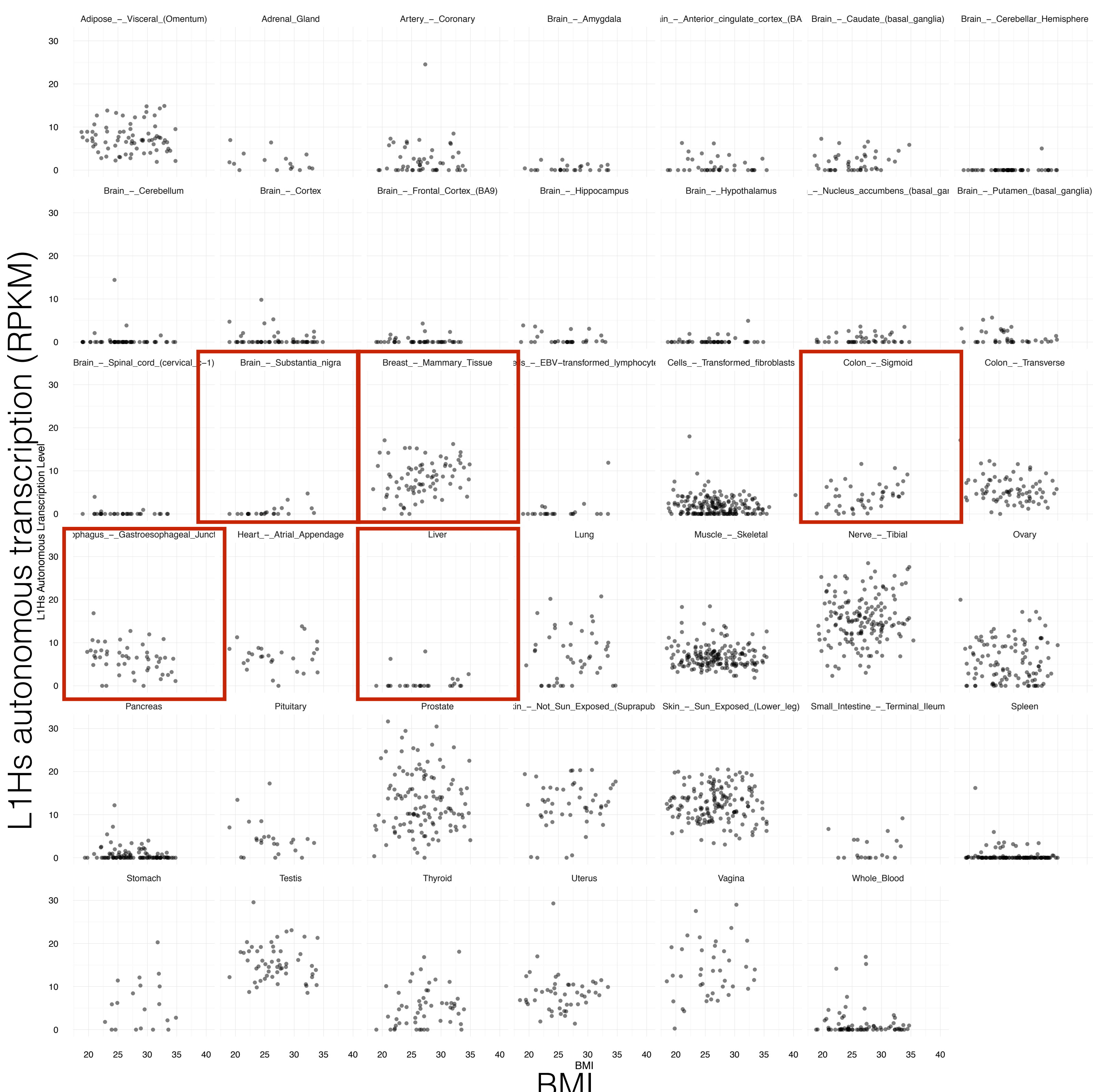


L1 activity with Age



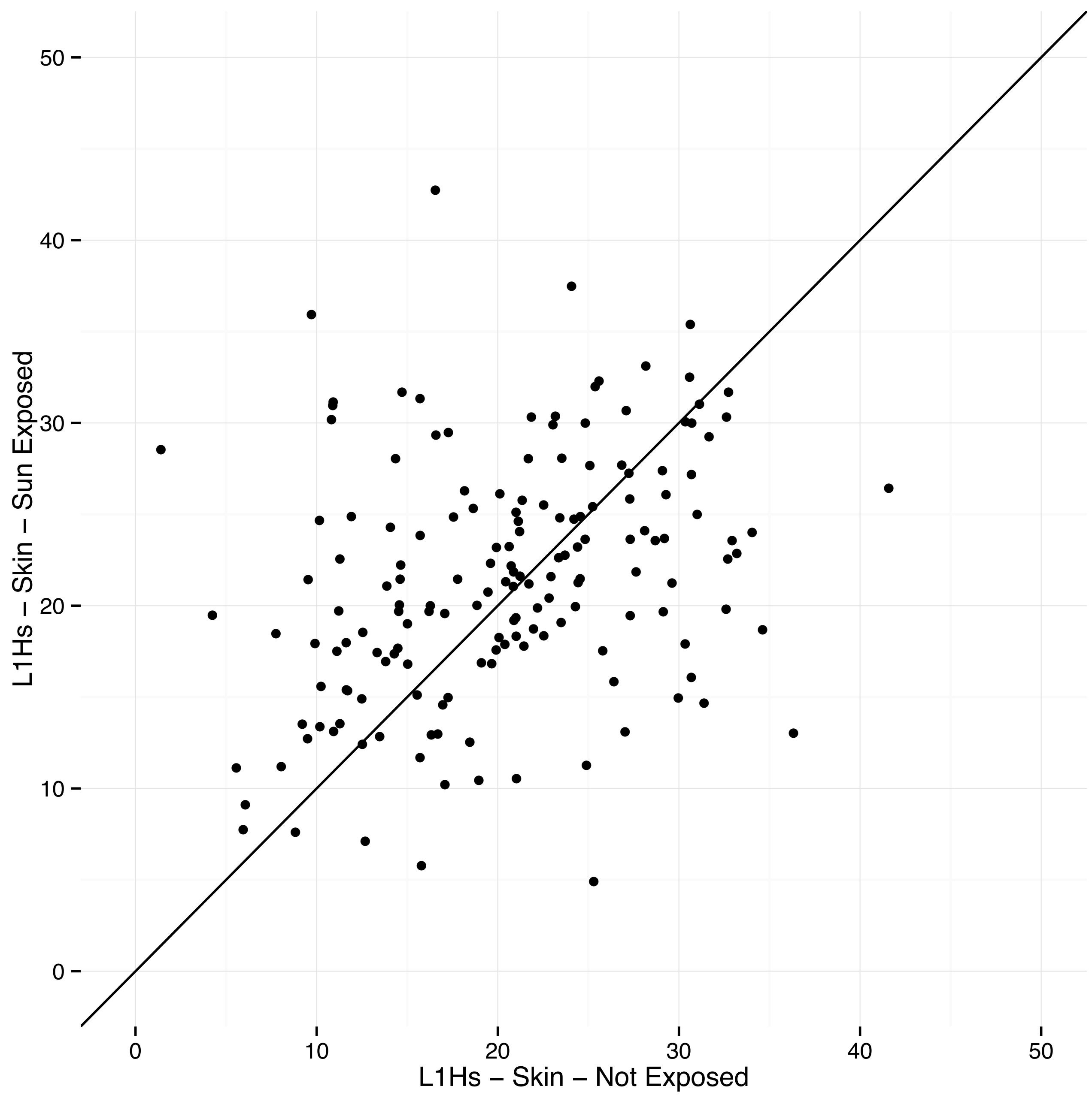
Tissue	p-value	corr
Prostate	1.65E-07	-0.333419429
Cells - Transformed	0.002835973	0.157327685
Adipose - Visceral	0.00359633	0.242811085
Colon - Sigmoid	0.005020719	0.301663825
Testis	0.008528213	0.263034162

L1 activity with BMI



Tissue	p-value	corr
Brain - Substantia	9.55E-05	0.583854047
Colon - Sigmoid	0.000275892	0.384870358
Esophagus - Gastroesophage	0.000506324	-0.343172644
Breast - Mammary Tissue	0.000633781	0.282362623
Liver	0.00701708	0.409920392

Sun effect on L1Hs activation

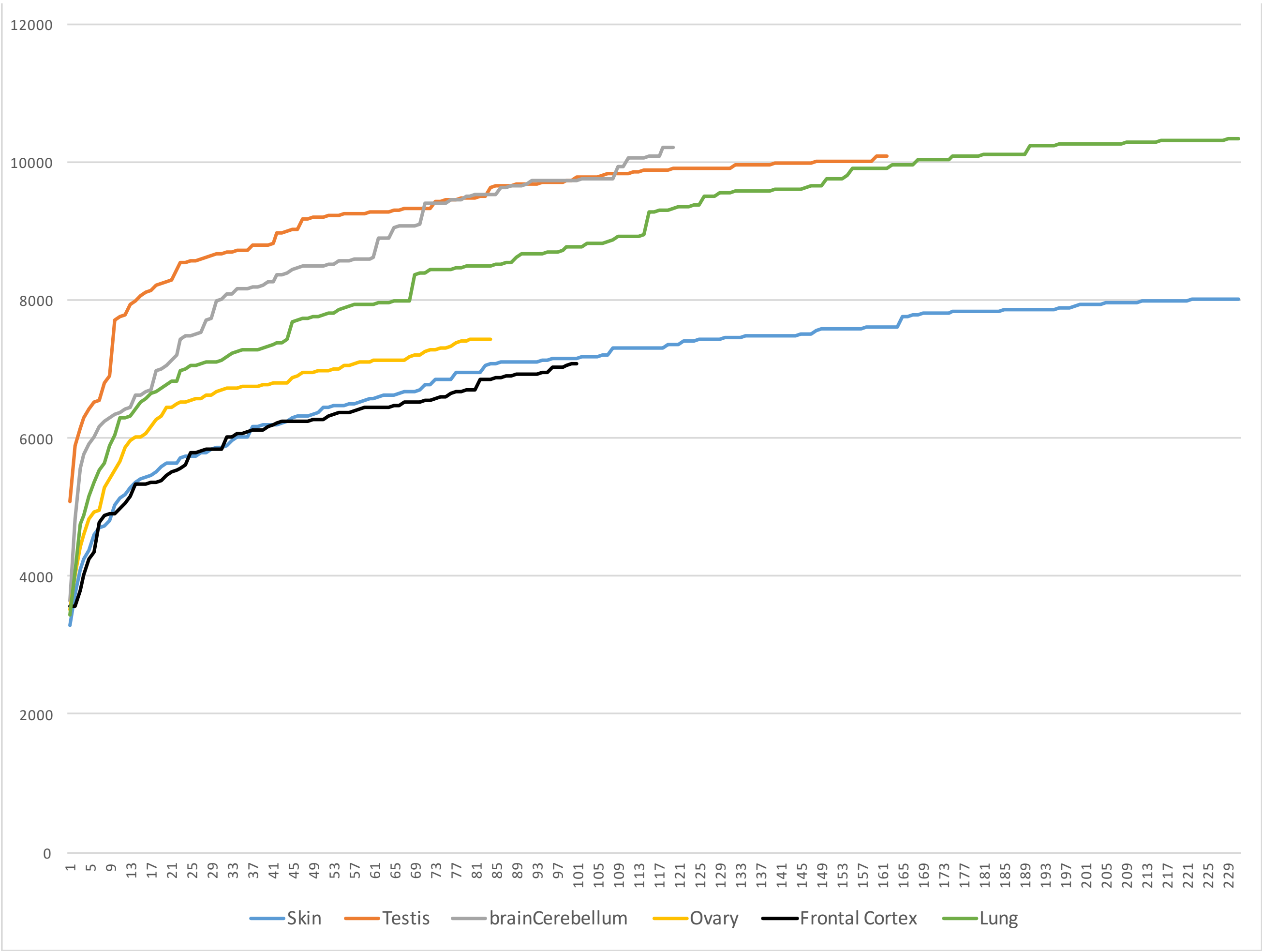


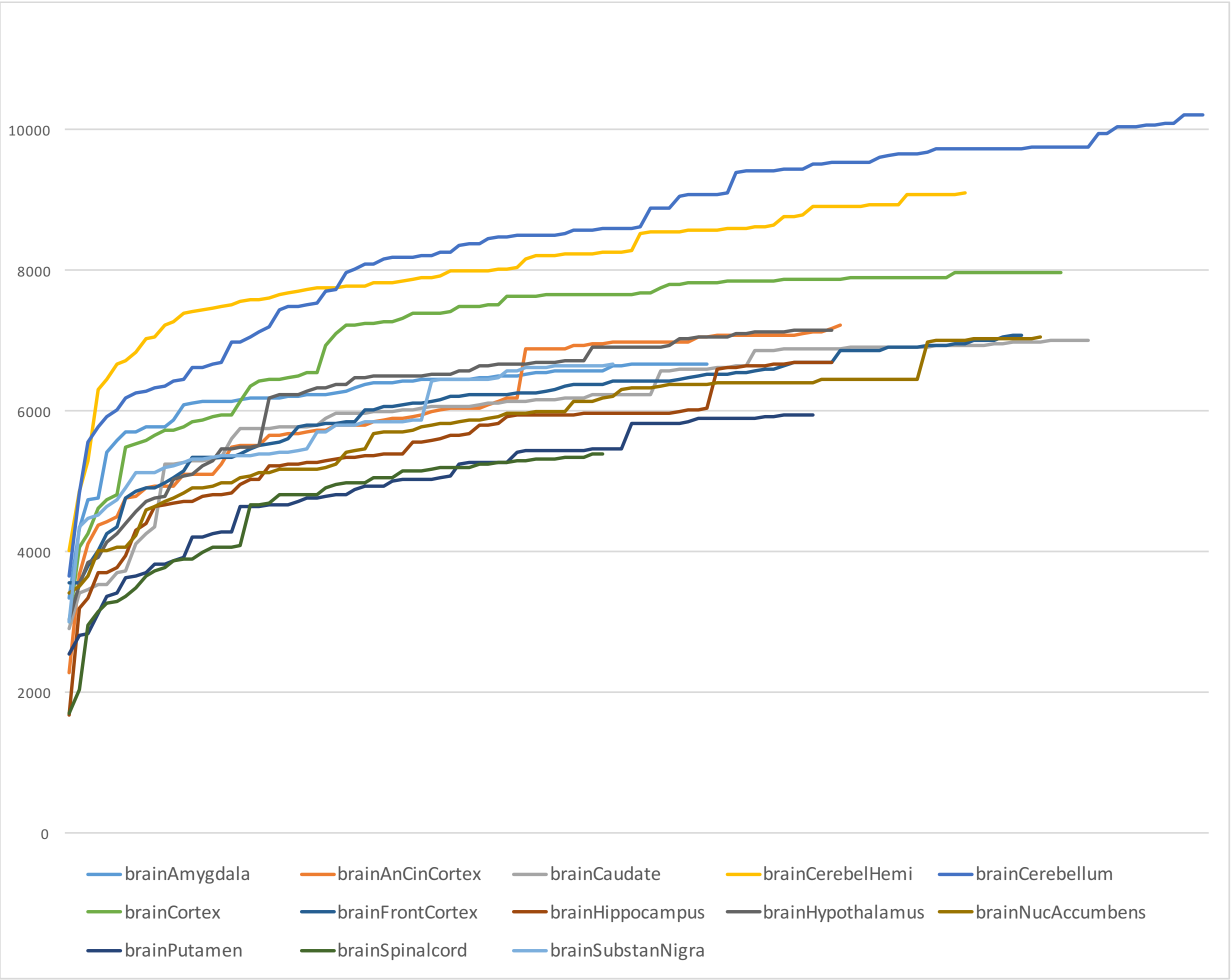
Pervasive transcription index

- Intuitively this is the number of reads originating from pervasive transcription normalized by reads from RNA sequencing experiments

- Equivalent to RPM

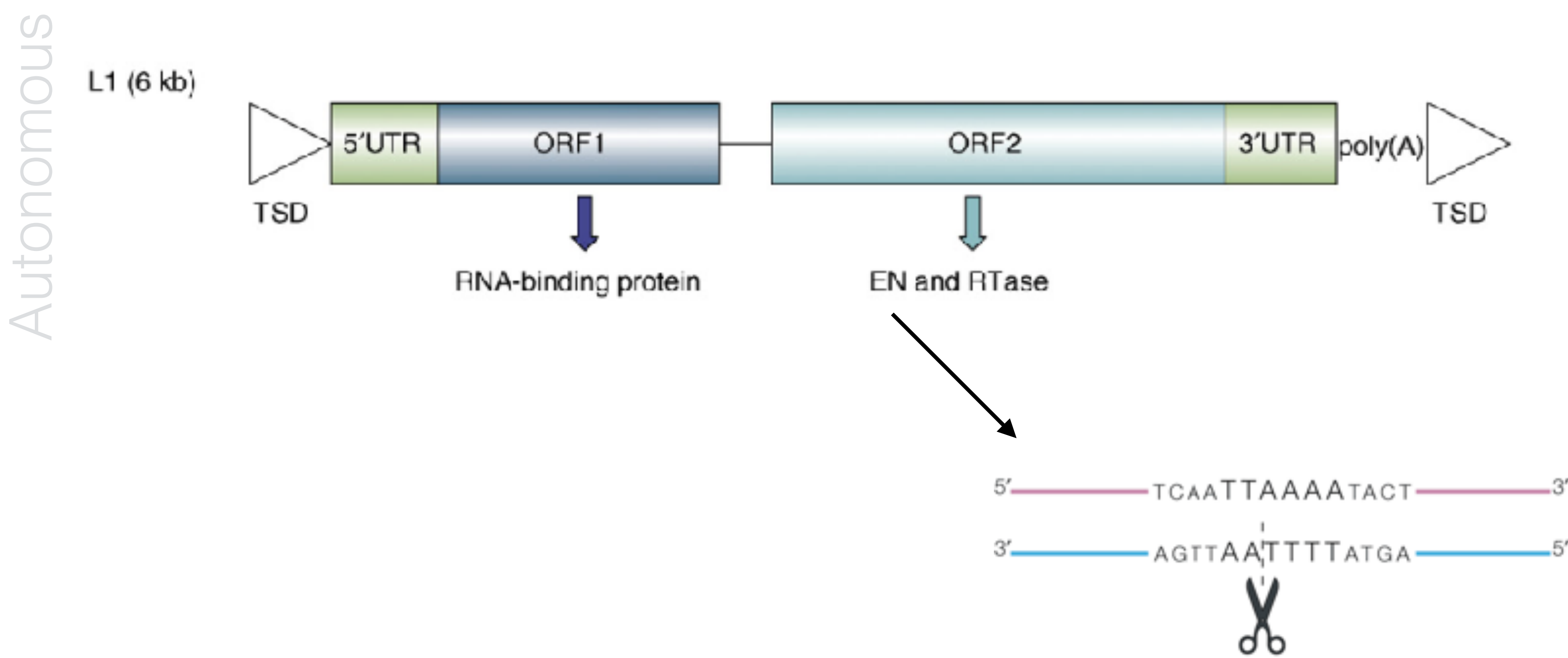
TARs diversity





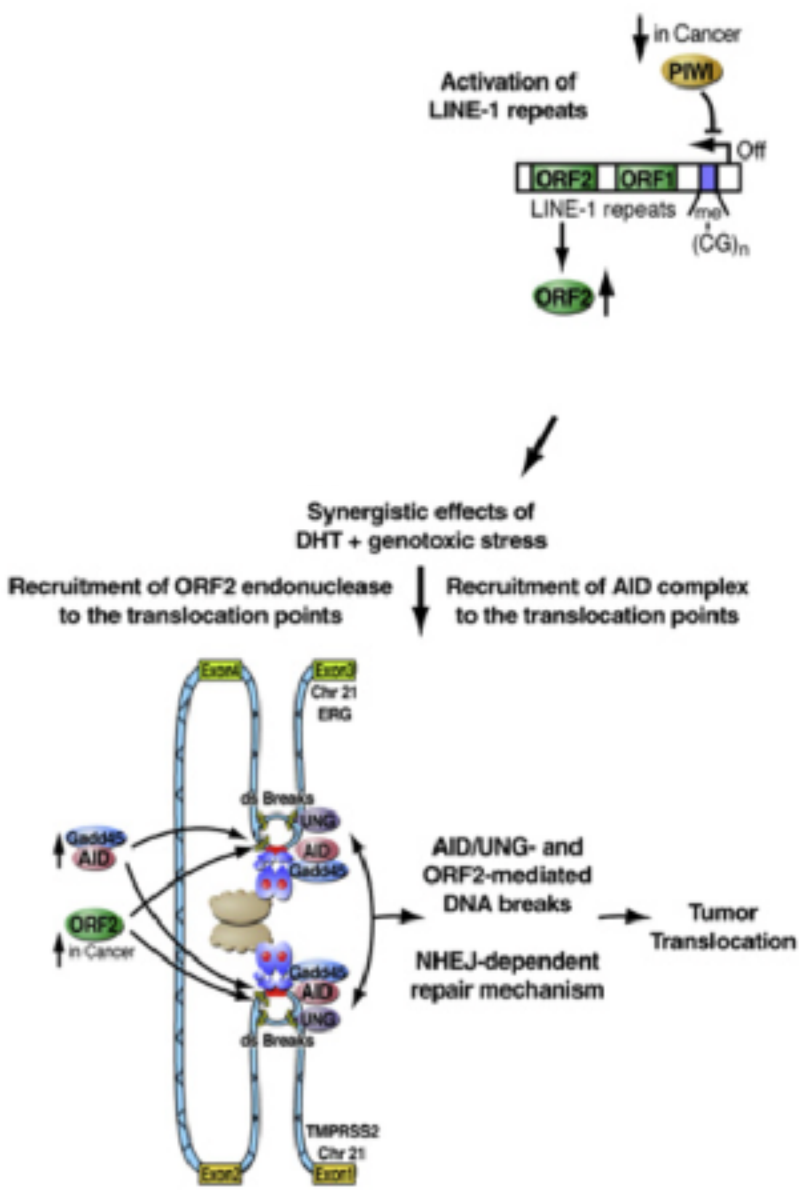
L1Hs autonomous transcription in tumors

LINE endonuclease activity



Mills, R. E., Bennett, E. A., Iskow, R. C., & Devine, S. E. (2007). Which transposable elements are active in the human genome? *Trends in Genetics : TIG*, *23*(4), 183–191. http://doi.org/10.1016/j.tig.2007.02.006

L1 Endonuclease promotes DSB



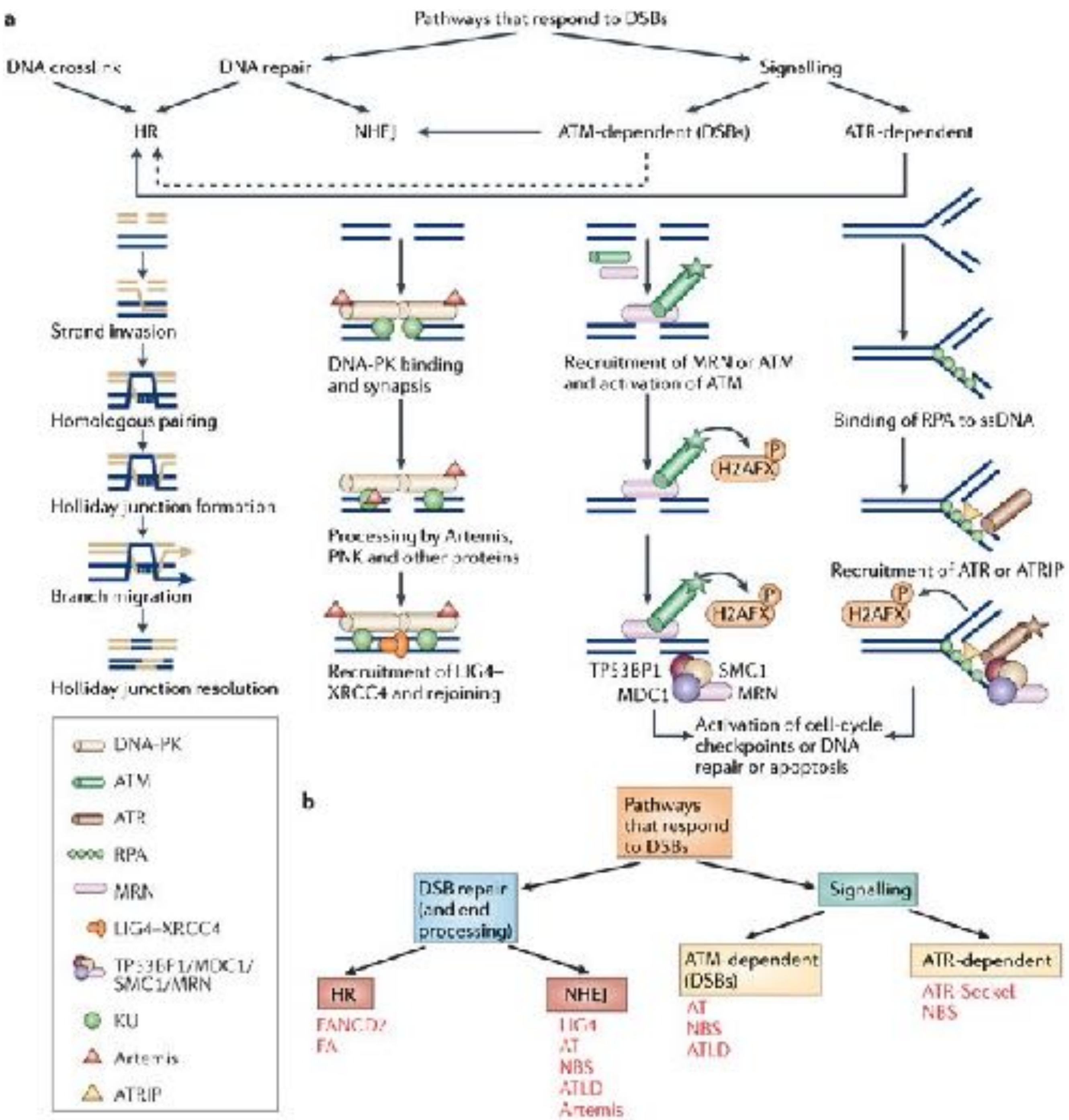
- Further- more, overexpression of the LINE-1 ORF2 induced a marked increase in both intra- and interchromosomal translocations, whereas the endonuclease-inactive mutant partially blocked DHT+IR-induced translocations.

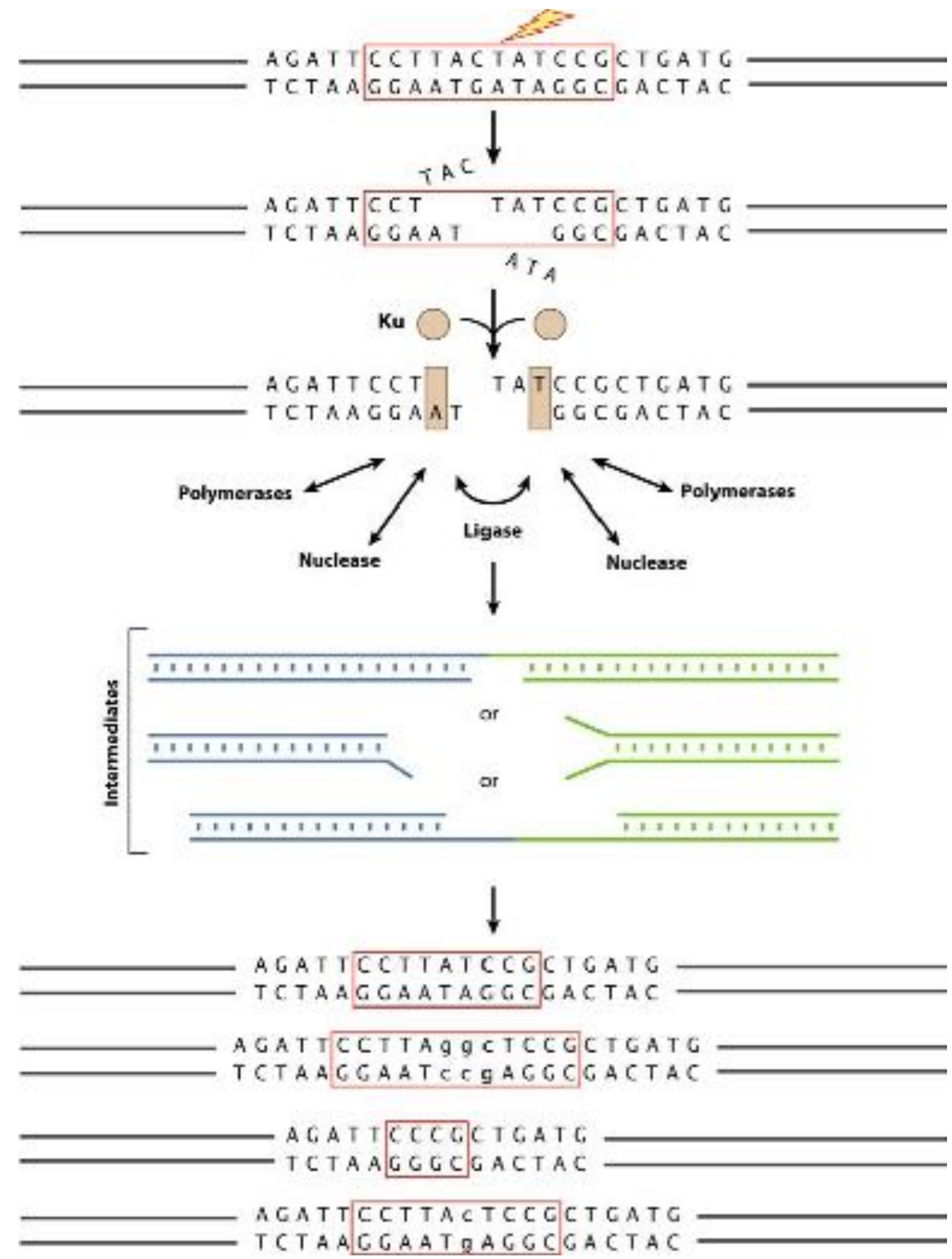
- To our surprise, even in the absence of genotoxic stress, the ORF2 endonuclease appears to be capable of targeting DNA breakage, […] generating DSBs at the translocation sites.

Lin, C., Yang, L., Tanasa, B., Hutt, K., Ju, B.-G., Ohgi, K., et al. (2009). Nuclear receptor-induced chromosomal proximity and DNA breaks underlie specific translocations in cancer. *Cell*, *139*(6), 1069–1083. http://doi.org/10.1016/j.cell.2009.11.030

Gasior, S. L., Wakeman, T. P., Xu, B., & Deininger, P. L. (2006). The human LINE-1 retrotransposon creates DNA double-strand breaks. *Journal of Molecular Biology*, *357*(5), 1383–1393. http://doi.org/10.1016/j.jmb.2006.01.089

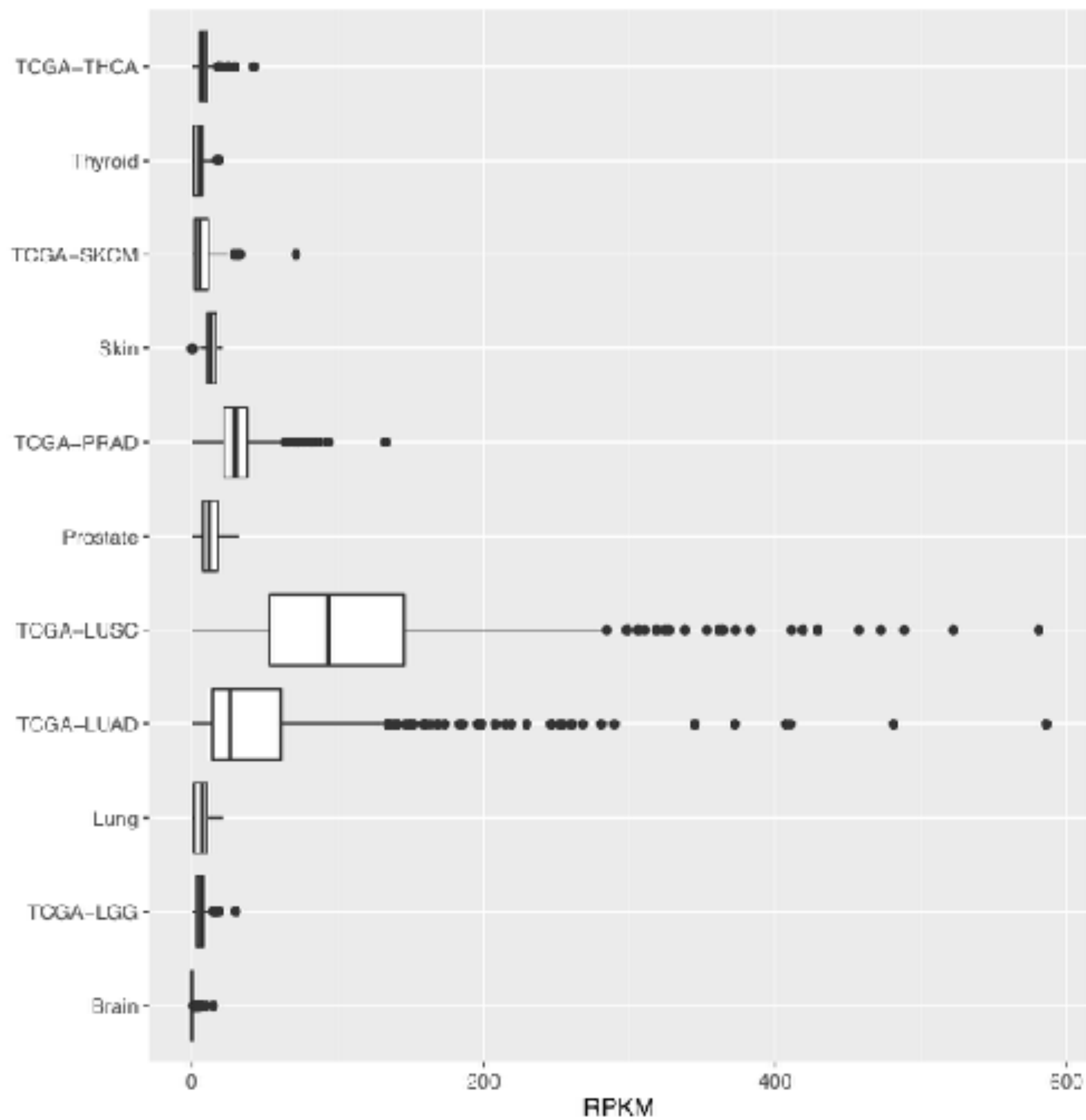
NHEJ is error prone



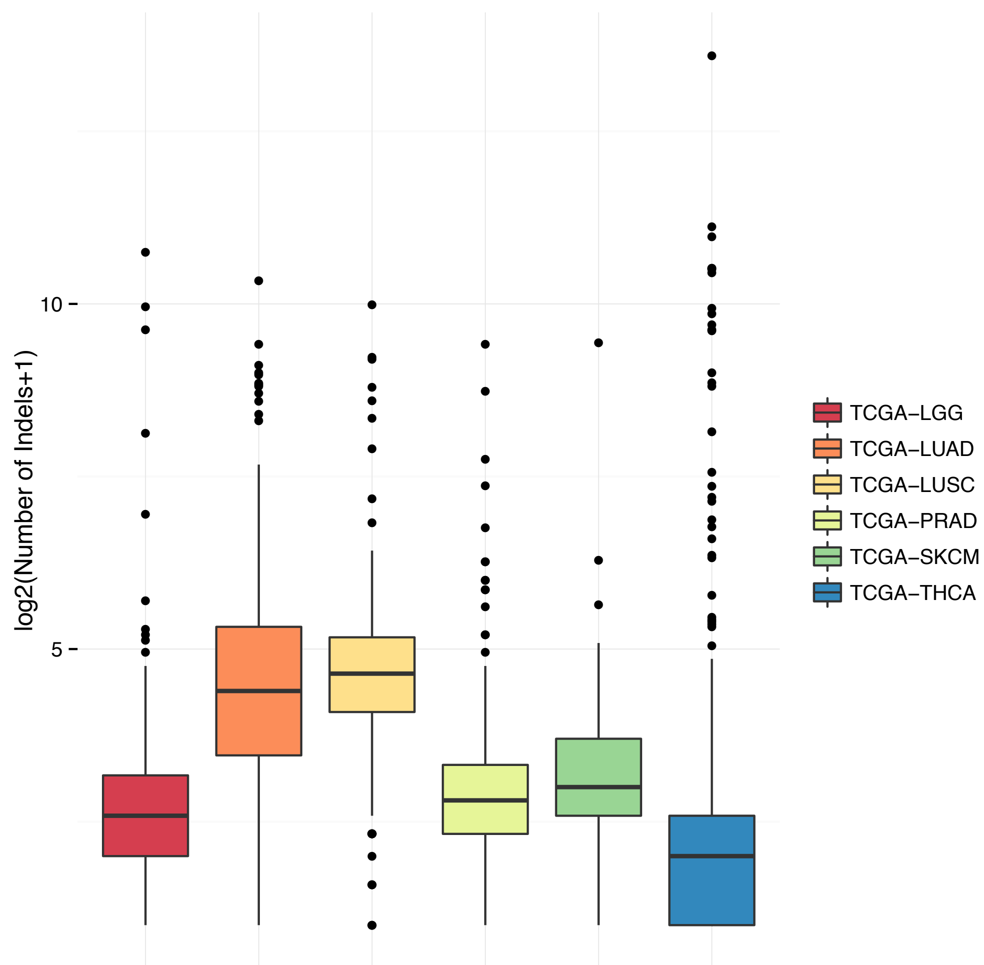


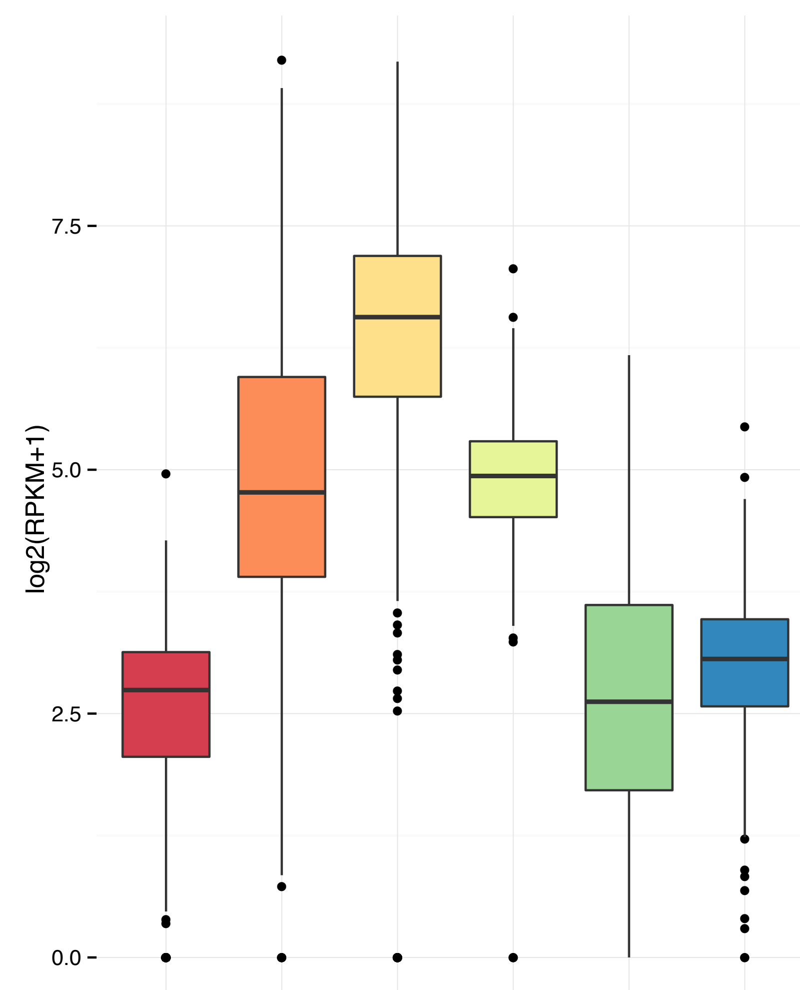
Lin, C., Yang, L., Tanasa, B., Hutt, K., Ju, B.-G., Ohgi, K., et al. (2009). Nuclear receptor-induced chromosomal proximity and DNA breaks underlie specific translocations in cancer. *Cell*, *139*(6), 1069–1083. http://doi.org/10.1016/j.cell.2009.11.030

Tumor vs Normal L1Hs autonomous transcription level

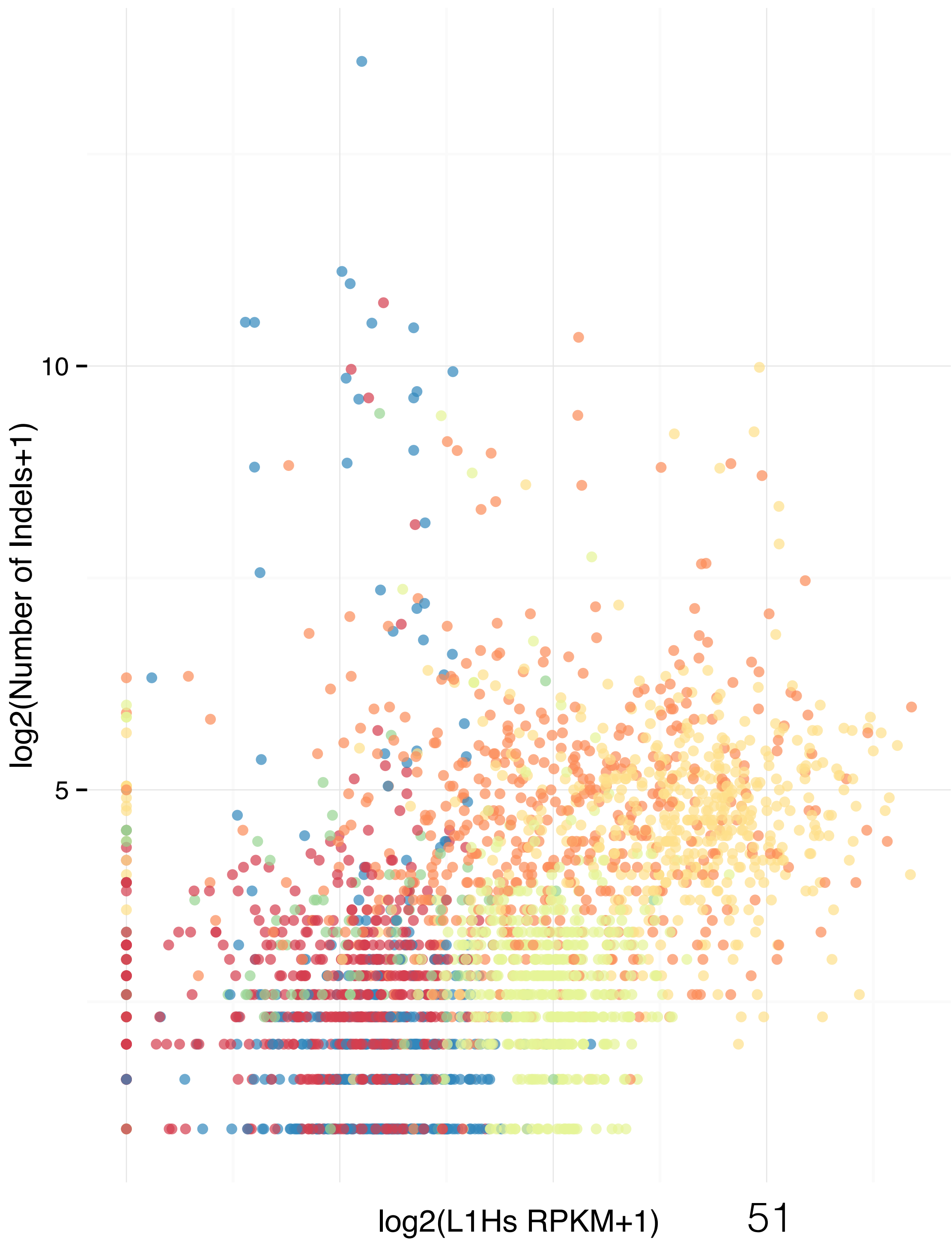


L1Hs Vs indels counts



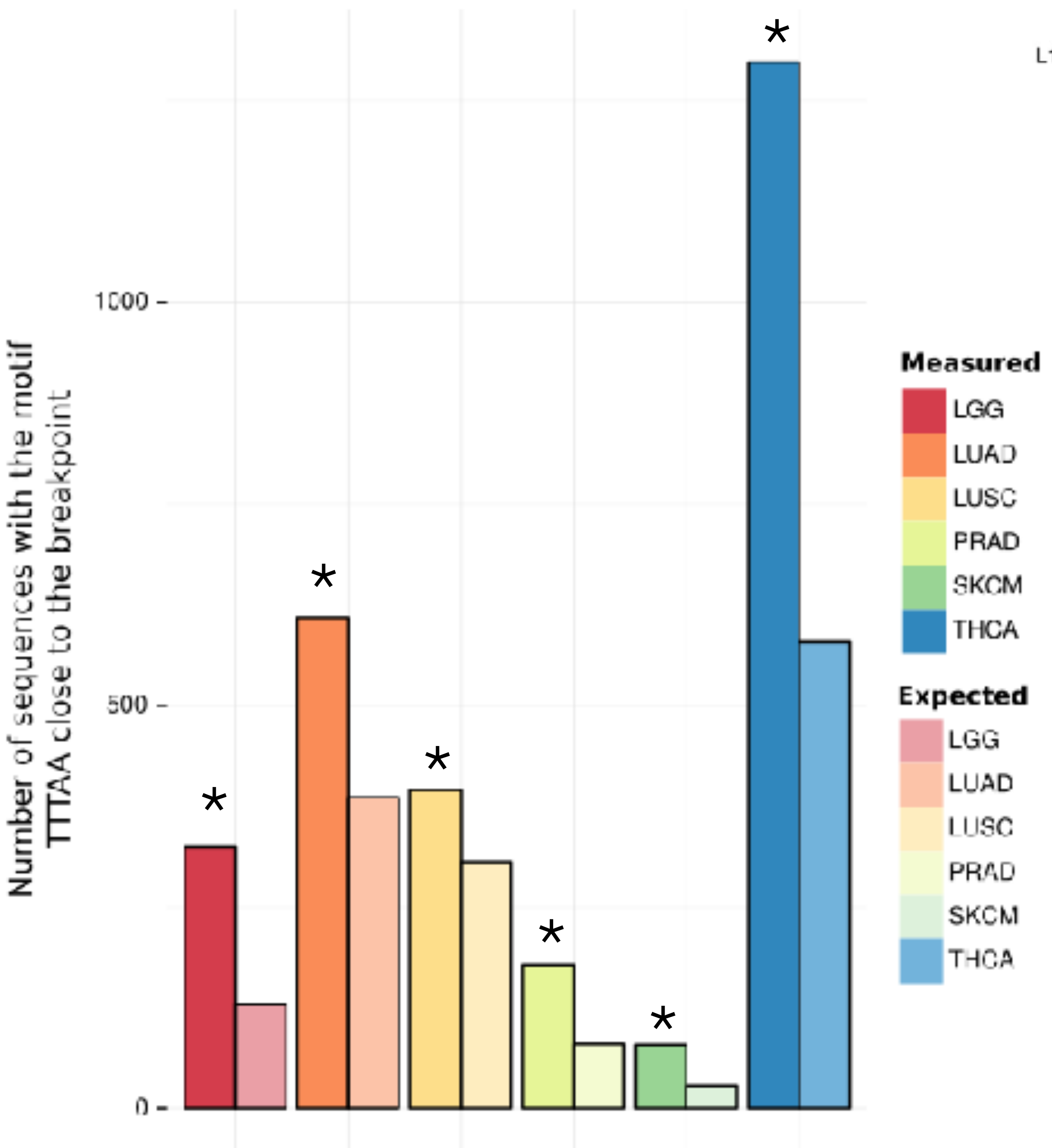


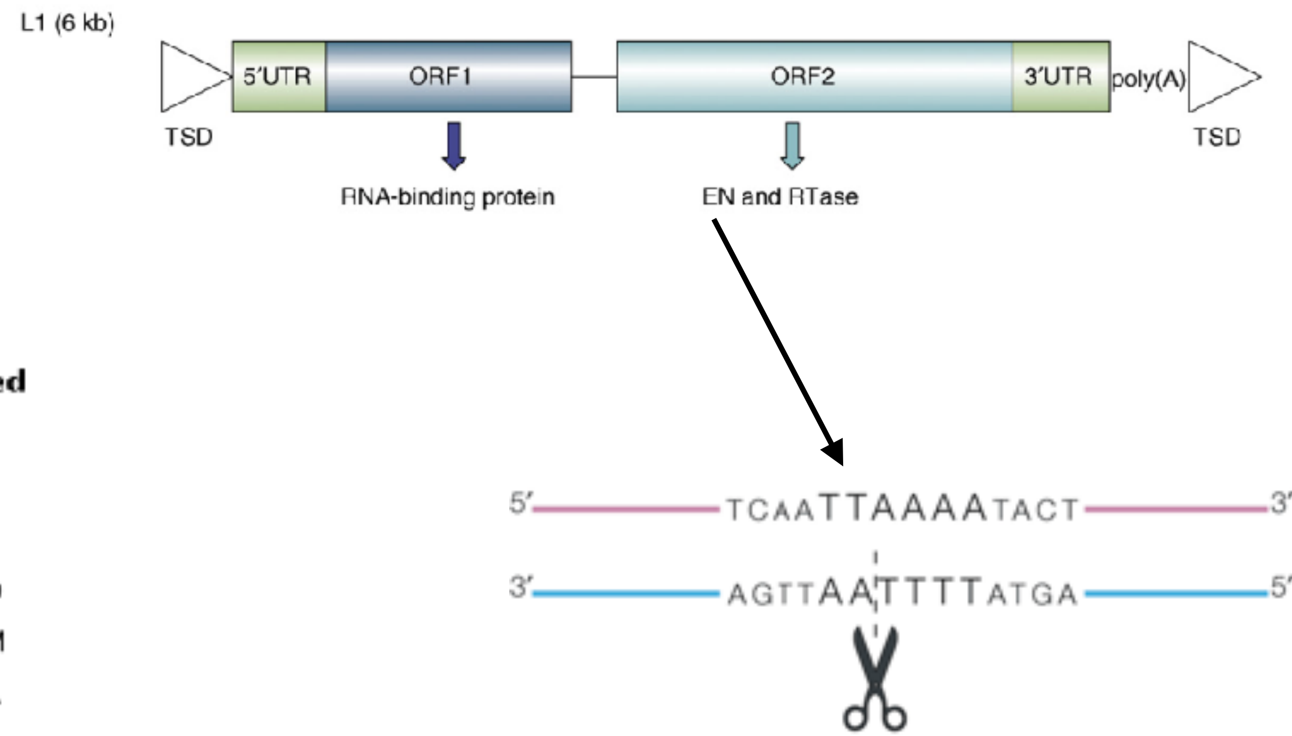
L1Hs correlates to indels counts



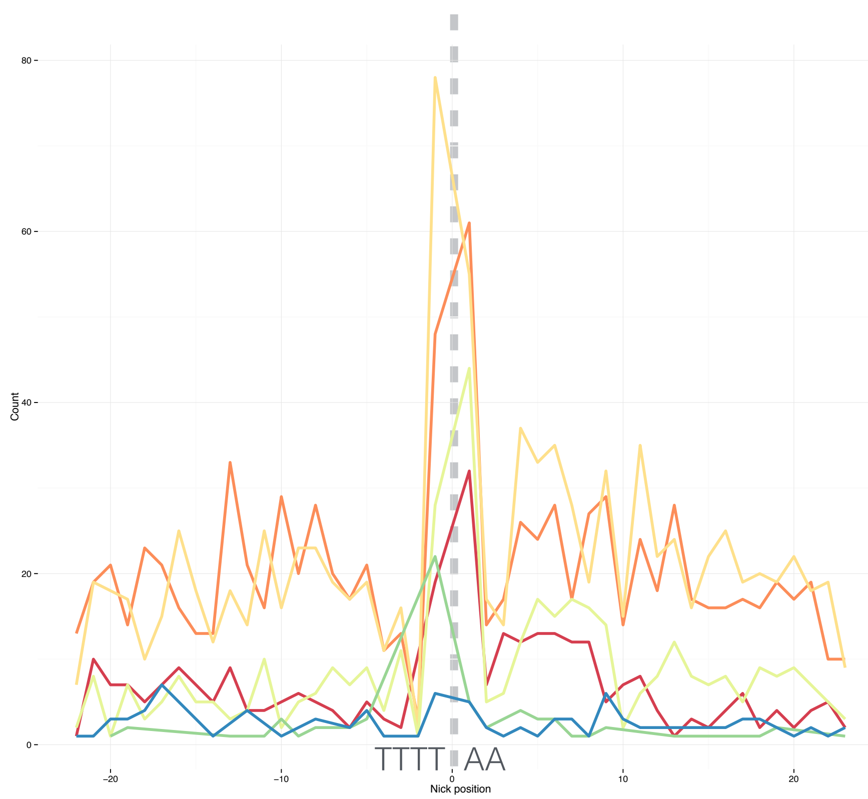
rho = 0.4662948
p-value < 2.2e-16

Enrichment of L1 endonuclease motif close to indels

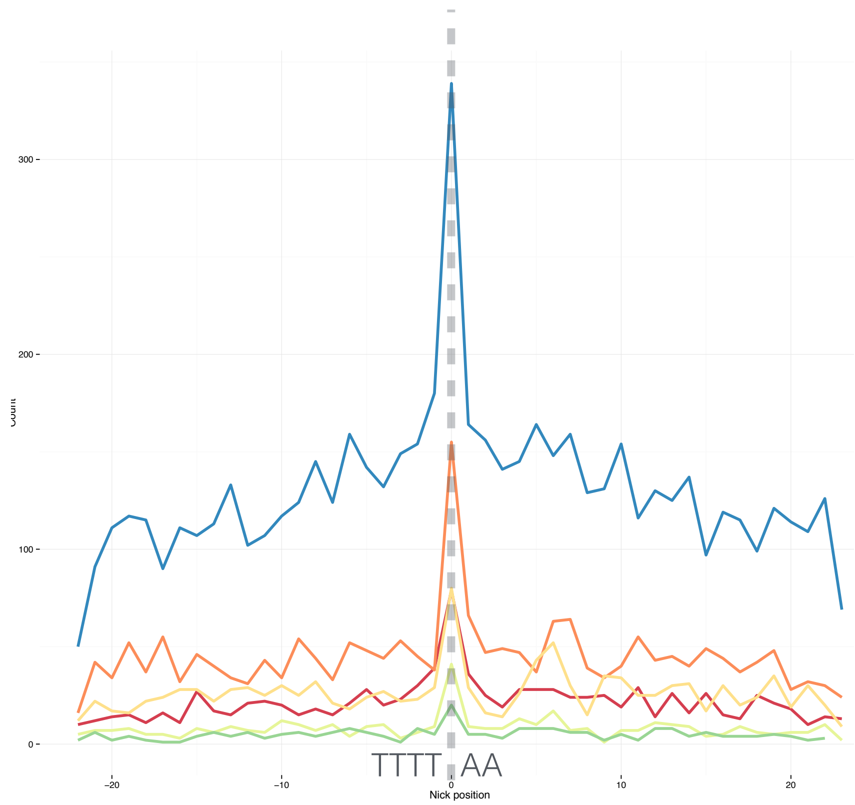




TTTAA is enriched in the actual index



(insertions)



(deletions)

Conclusions

- TeXP is a method to decouple the signal of pervasive and autonomous transcription on RNA sequencing experiments;

- There is autonomous transcription of L1Hs (and only L1Hs) in healthy somatic tissue;

- In a few tissues, this expression correlates to Age and BMI;

- Using a pervasive transcription index, we ranked tissues based on their level of pervasive transcription;

- L1 endonuclease might be a source of genome instability creating double strand breaks and, therefore, translocations (not shown) and indels in the tumor genome.

Acknowledgements

- Timur Galeev
- Jinrui Xu
- Joel Rozowsky
- Arif Harmanci
- Sushant Kumar
- Shantao Li
- Mark Gerstein

Thank you!