Transcription of transposable elements in human brain

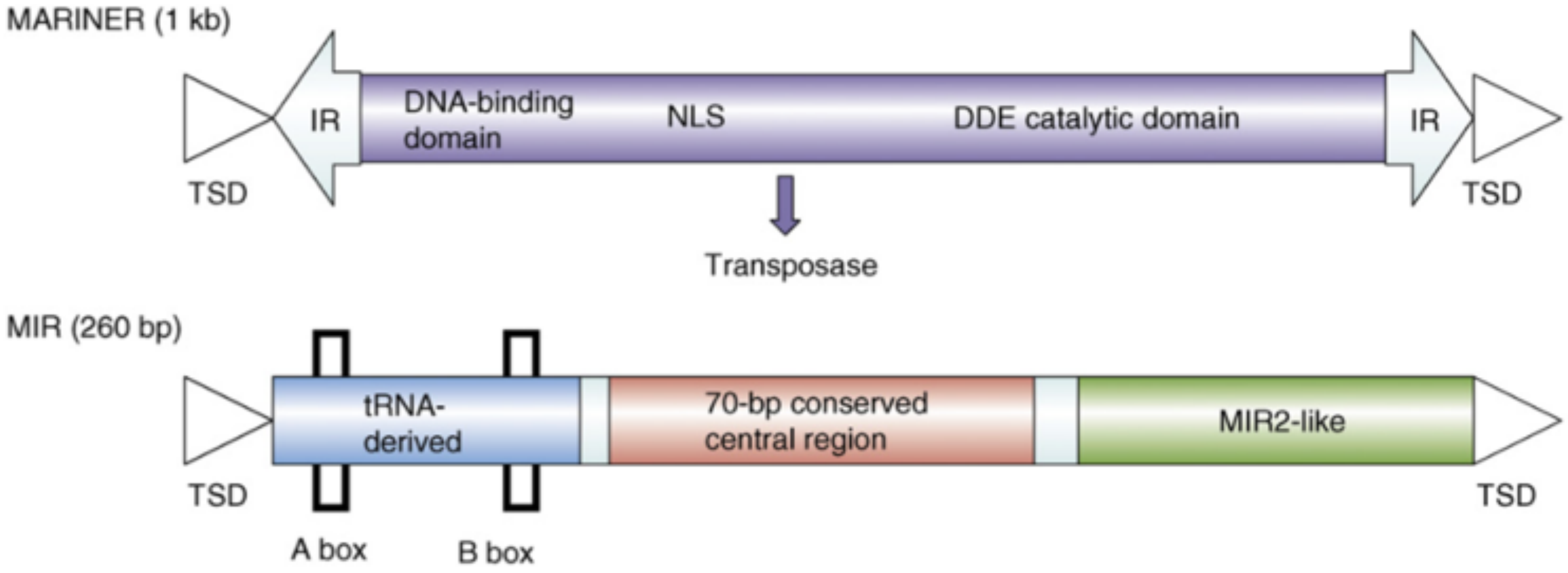
Fábio Navarro - Group Meeting
2015

Overview

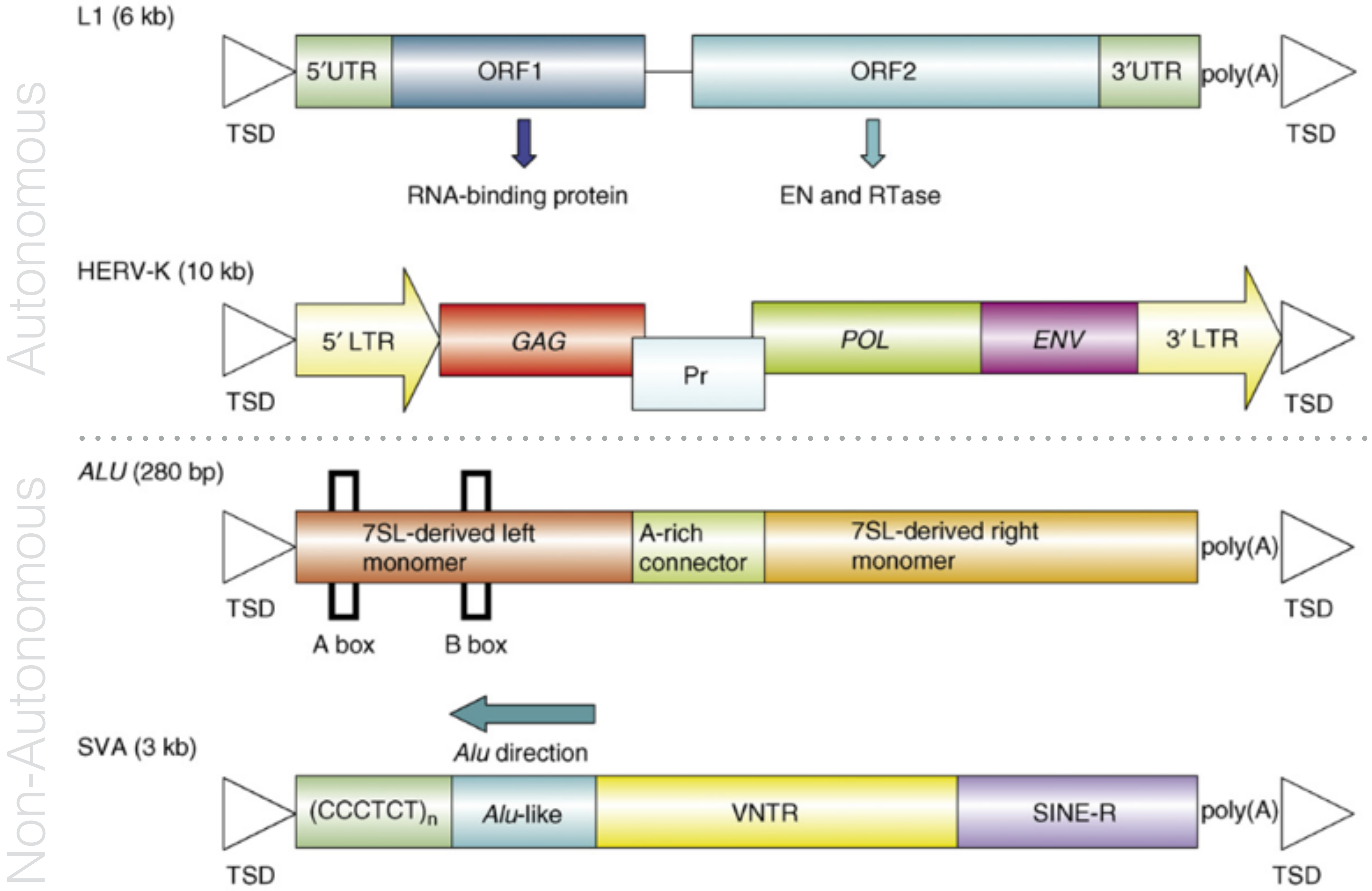
- Classification of transposable elements
- Somatic activity of L1 elements in the human brain
- Methods to assign short reads to transposable elements
- Preliminary results
    - P1- Mappable TEs
    - P2 - Unmappable TEs

Dead TE in human genome

Autonomous

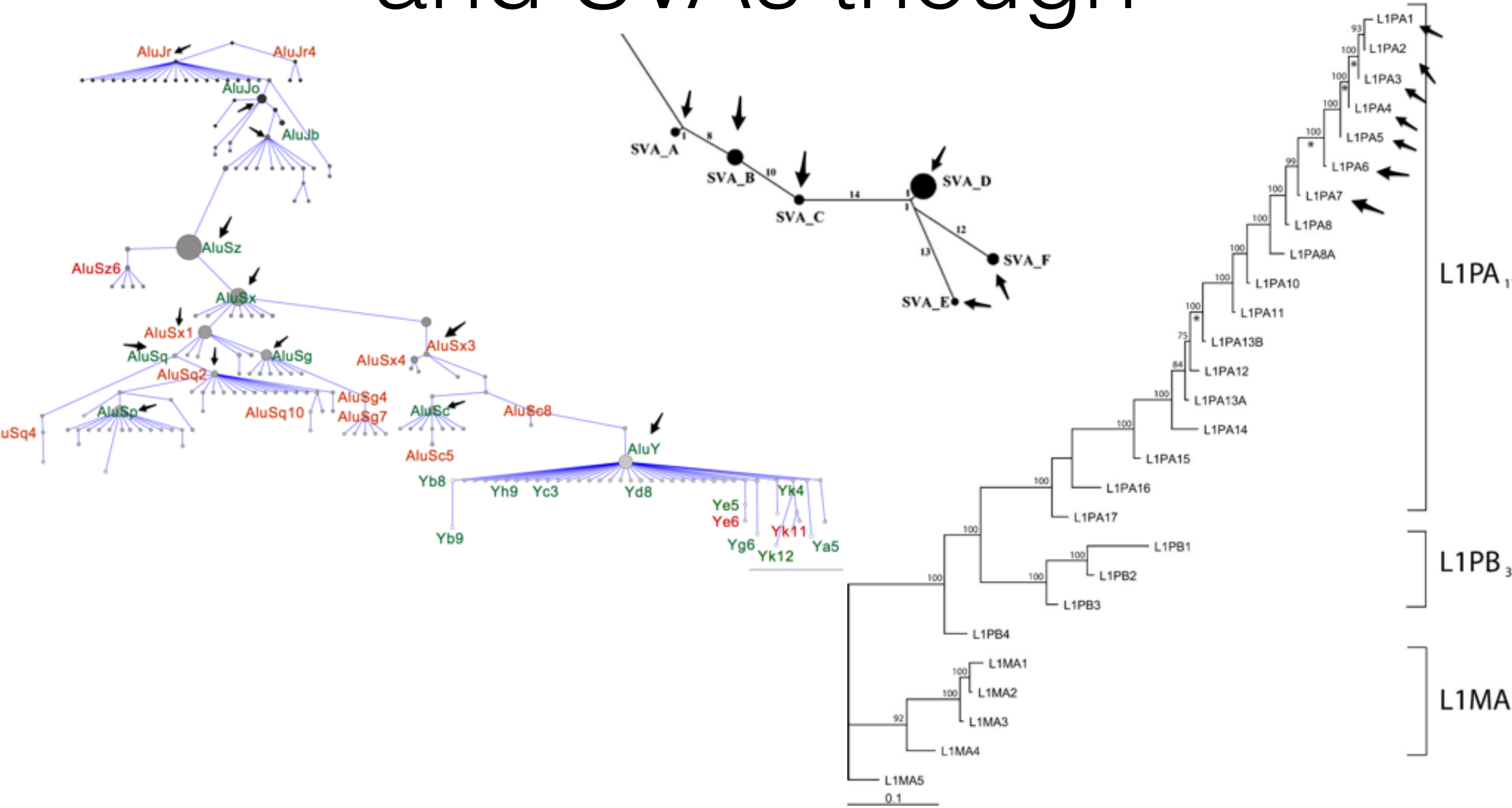


Active TE in human genome



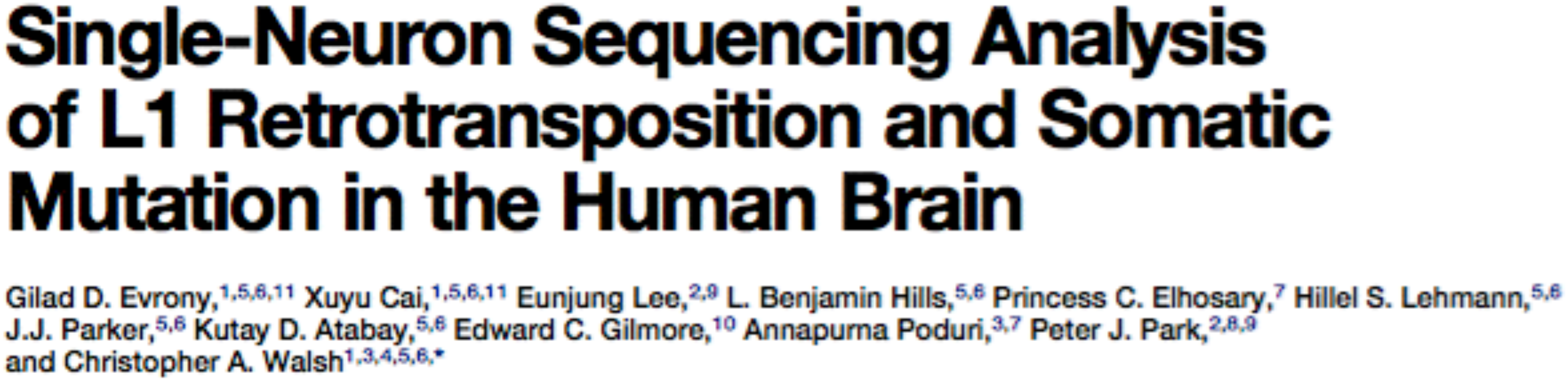
Mills, R. E., Bennett, E. A., Iskow, R. C., & Devine, S. E. (2007). Which transposable elements are active in the human genome? *Trends in Genetics : TIG*, *23*(4), 183–191. http://doi.org/10.1016/j.tig.2007.02.006

Not all LINEs,HERVs, ALUs and SVAs though



Brouha, B., Schustak, J., Badge, R. M., Lutz-Prigge, S., Farley, A. H., Moran, J. V., & Kazazian, H. H. (2003). 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*(9), 5280–5285. http://doi.org/10.1073/pnas.0831042100

Somatic activity of L1 in the brain



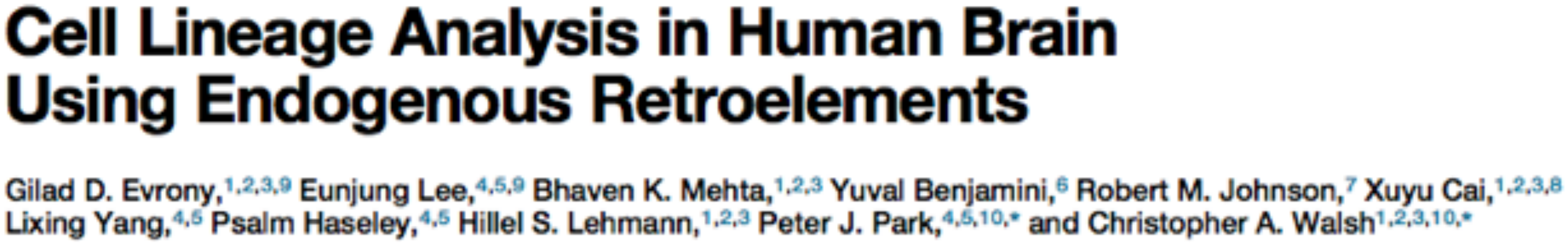
Single-Neuron Sequencing Analysis of L1 Retrotransposition and Somatic Mutation in the Human Brain

Gilad D. Evrony,[1,5,6,11] Xuyu Cai,[1,5,6,11] Eunjung Lee,[2,9] L. Benjamin Hills,[5,6] Princess C. Elhosary,[7] Hillel S. Lehmann,[5,6] J.J. Parker,[5,6] Kutay D. Atabay,[5,6] Edward C. Gilmore,[10] Annapurna Poduri,[3,7] Peter J. Park,[2,8,9] and Christopher A. Walsh[1,3,4,5,6,*]

- Sequenced 300 neurons (Single cell)
- L1-IP library
- 0.6 ± 1.5 (SD) candidate unique insertions per neuron (after validation: 0.07 ± 0.15 (SD) and 0.04 ± 0.10 (SD) insertions per neuron)
- 82% of 1-neuron samples had no detectable unique somatic insertions.

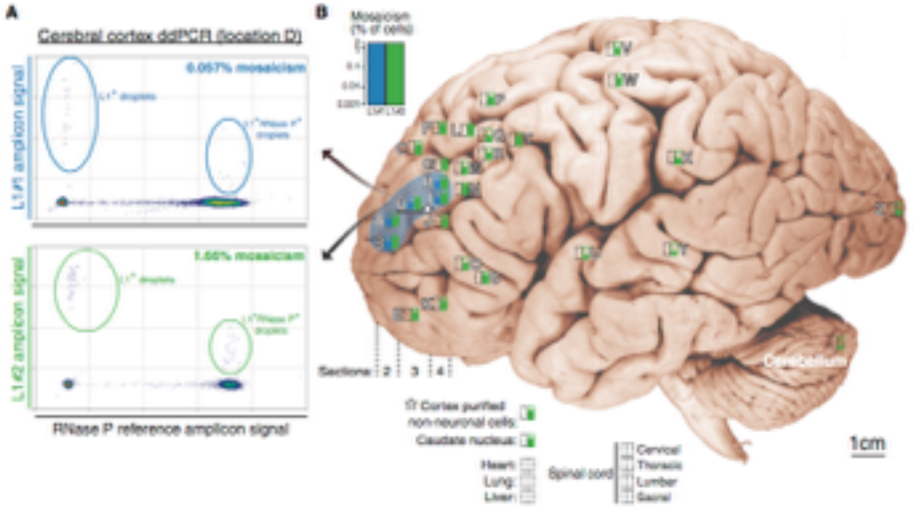
Evrony, G. D., Cai, X., Lee, E., Hills, L. B., Elhosary, P. C., Lehmann, H. S., et al. (2012). Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell*, *151*(3), 483–496. http://doi.org/10.1016/j.cell.2012.09.035

Somatic activity of L1 in the brain



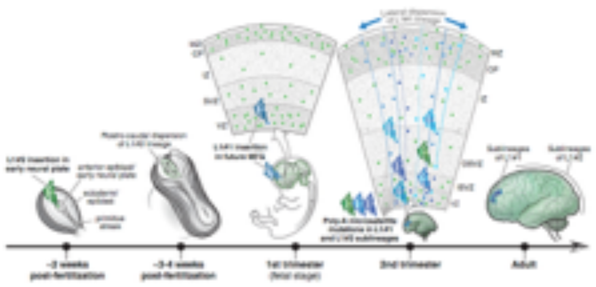
Cell Lineage Analysis in Human Brain Using Endogenous Retroelements

Gilad D. Evrony,[1,2,3,9] Eunjung Lee,[4,5,9] Bhaven K. Mehta,[1,2,3] Yuval Benjamini,[6] Robert M. Johnson,[7] Xuyu Cai,[1,2,3,8] Lixing Yang,[4,5] Psalm Haseley,[4,5] Hillel S. Lehmann,[1,2,3] Peter J. Park,[4,5,10,*] and Christopher A. Walsh[1,2,3,10,*]



- High-coverage whole-genome sequencing of single neurons from human brain (N=16 frontal gyrus of the dorsolateral prefrontal cortex).

- 2 somatic insertions.


~2 weeks post-fertilization
~3-4 weeks post-fertilization
1st trimester (fetal stage)
2nd trimester
Adult

- Spatial tracing of cell lineages in human brain using somatic retrotransposon insertions

- Somatic mutations reveal patterns of clonal dispersion and focal mutation in normal brain

Evrony, G. D., Lee, E., Mehta, B. K., Benjamini, Y., Johnson, R. M., Cai, X., et al. (2015). Cell lineage analysis in human brain using endogenous retroelements. *Neuron*, *85*(1), 49–59. http://doi.org/10.1016/j.neuron.2014.12.028

Somatic activity of L1 in the brain

Ubiquitous L1 Mosaicism in Hippocampal Neurons

Kyle R. Upton,[1,6] Daniel J. Gerhardt,[1,6] J. Samuel Jesuadian,[1,6] Sandra R. Richardson,[1] Francisco J. Sánchez-Luque,[1] Gabriela O. Bodea,[1] Adam D. Ewing,[1] Carmen Salvador-Palomeque,[1] Marjo S. van der Knaap,[2] Paul M. Brennan,[3] Adeline Vanderver,[4] and Geoffrey J. Faulkner[1,5,*]

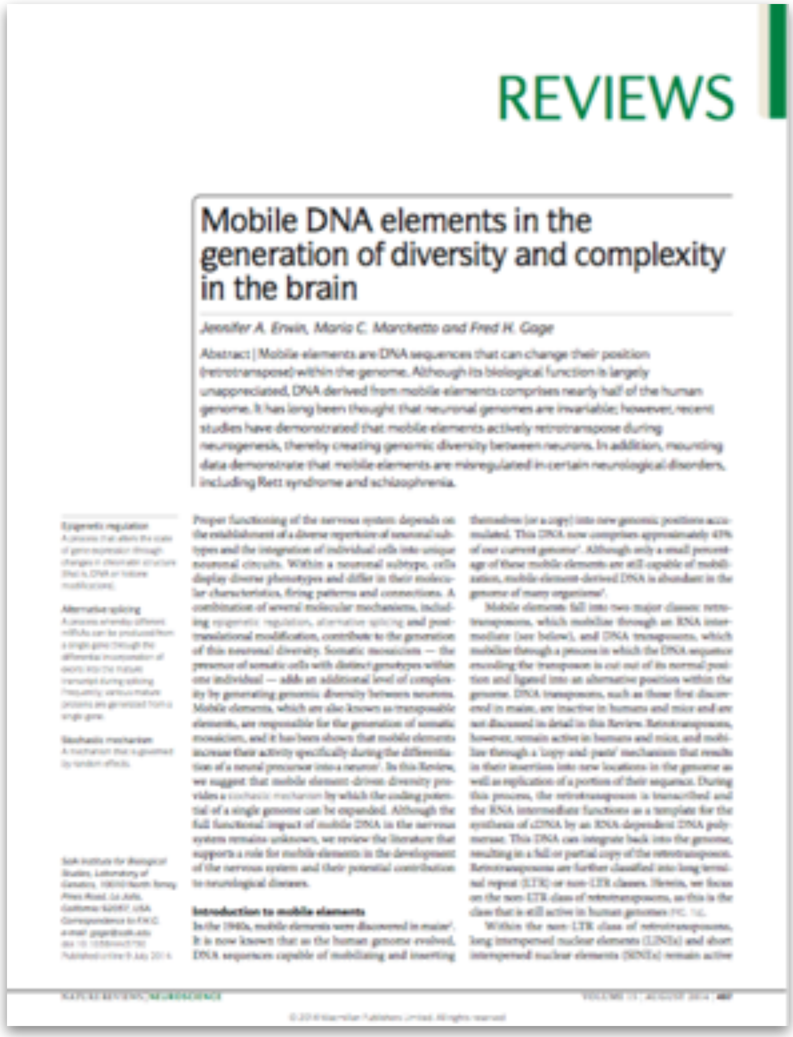
- Single-cell RT-Seq: 92 individual neuronal nuclei from hippocampal neuron.

- Estimated 13.7 somatic L1 insertions occur per hippocampal neuron.

- "Developmental timing of L1 mobilization in the brain remains unclear", but few events are across many neurons.

Upton, K. R., Gerhardt, D. J., Jesuadian, J. S., Richardson, S. R., Sánchez-Luque, F. J., Bodea, G. O., et al. (2015). Ubiquitous L1 Mosaicism in Hippocampal Neurons. *Cell*, *161*(2), 228–239. http://doi.org/10.1016/j.cell.2015.03.026

Mobile DNA elements in the generation of diversity and complexity in the brain



REVIEWS

Mobile DNA elements in the generation of diversity and complexity in the brain

Jennifer A. Erwin, Maria C. Marchetto and Fred H. Gage

Abstract | Mobile elements are DNA sequences that can change their position (retrotranspose) within the genome. Although its biological function is largely unappreciated, DNA derived from mobile elements comprises nearly half of the human genome. It has long been thought that neuronal genomes are invariable; however, recent studies have demonstrated that mobile elements actively retrotranspose during neurogenesis, thereby creating genomic diversity between neurons. In addition, mounting data demonstrate that mobile elements are misregulated in certain neurological disorders, including Rett syndrome and schizophrenia.

1. Rate of retrotransposition in different regions of the brain?

2. Which cell types are more prone to retrotransposition?

3. Different individuals have different rates of retrotransposition?

4. What are the mechanism regulating their activity?

5. When they are active?

—

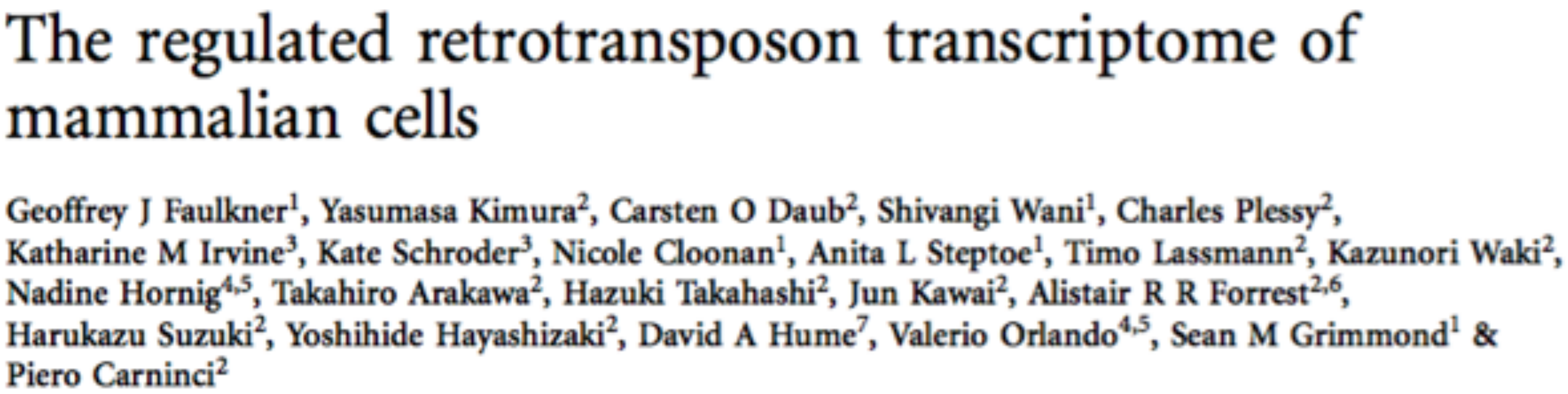
6. Which elements are active?

7. Can we reliably use RNA-seq to access their activity?

8. Is transcription a good proxy to measure TE activity?

Erwin, J. A., Marchetto, M. C., & Gage, F. H. (2014). Mobile DNA elements in the generation of diversity and complexity in the brain. Nature Publishing Group, 15(8), 497–506. http://doi.org/10.1038/nrn3730

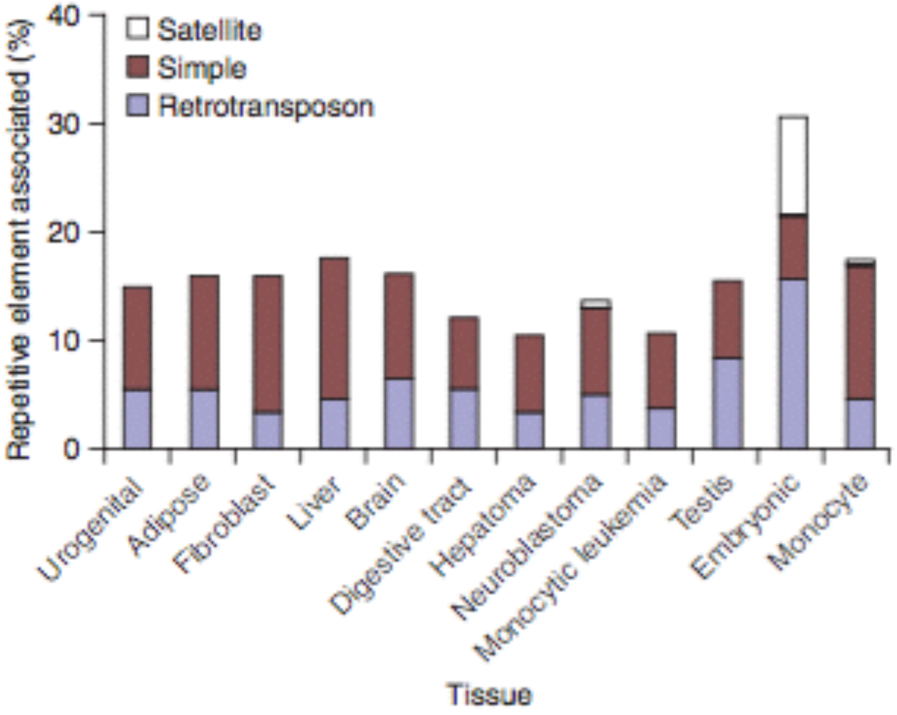
Transcription of TEs

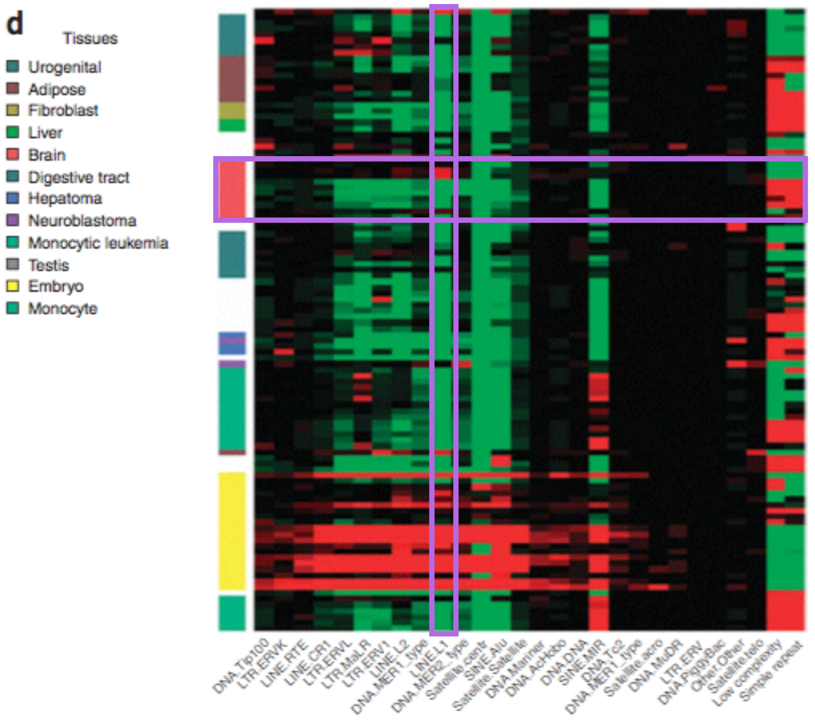


The regulated retrotransposon transcriptome of mammalian cells

Geoffrey J Faulkner[1], Yasumasa Kimura[2], Carsten O Daub[2], Shivangi Wani[1], Charles Plessy[2], Katharine M Irvine[3], Kate Schroder[3], Nicole Cloonan[1], Anita L Steptoe[1], Timo Lassmann[2], Kazunori Waki[2], Nadine Hornig[4,5], Takahiro Arakawa[2], Hazuki Takahashi[2], Jun Kawai[2], Alistair R R Forrest[2,6], Harukazu Suzuki[2], Yoshihide Hayashizaki[2], David A Hume[7], Valerio Orlando[4,5], Sean M Grimmond[1] & Piero Carninci[2]

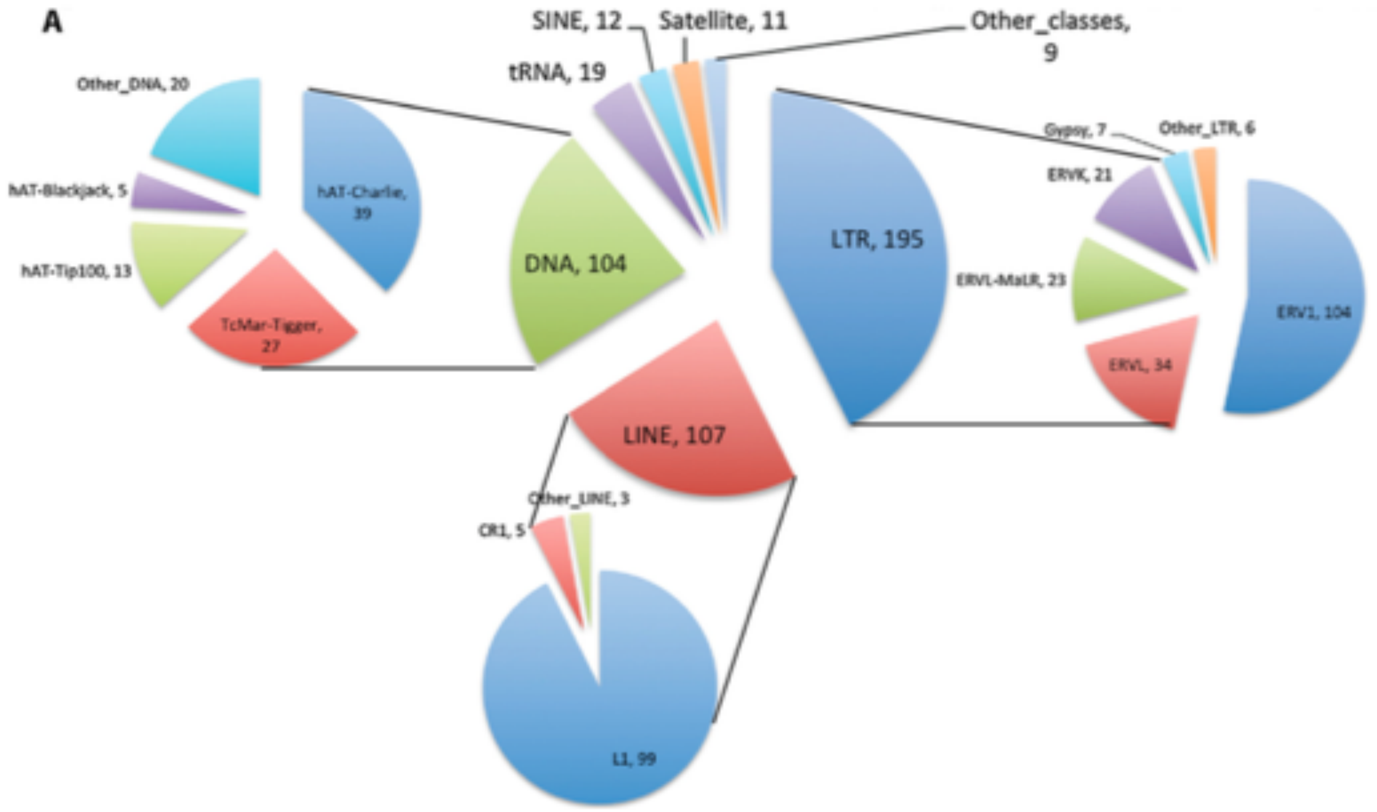
CAGE (Cap Analysis Gene Expression) - 20nt tags!





Faulkner, G. J., Kimura, Y., Daub, C. O., Wani, S., Plessy, C., Irvine, K. M., et al. (2009). The regulated retrotransposon transcriptome of mammalian cells. *Nature Genetics*, *41*(5), 563–571. http://doi.org/10.1038/ng.368

Transcription of TEs



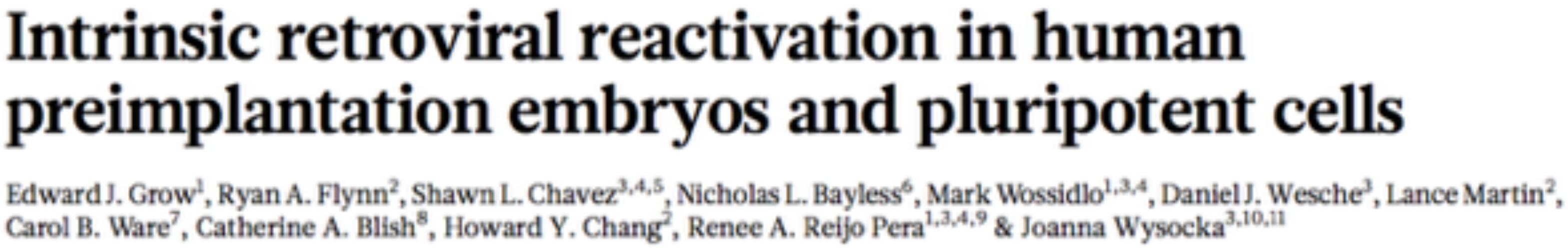
Classes and families of repetitive elements differentially expressed in prostate cancer tumor tissue versus normal tissue. The number next to each class and family name corresponds to the number of differentially expressed subfamilies (FDR < 0.05).

"Prevalently from the LTR, LINE and DNA classes."

https://github.com/nerettilab/RepEnrich

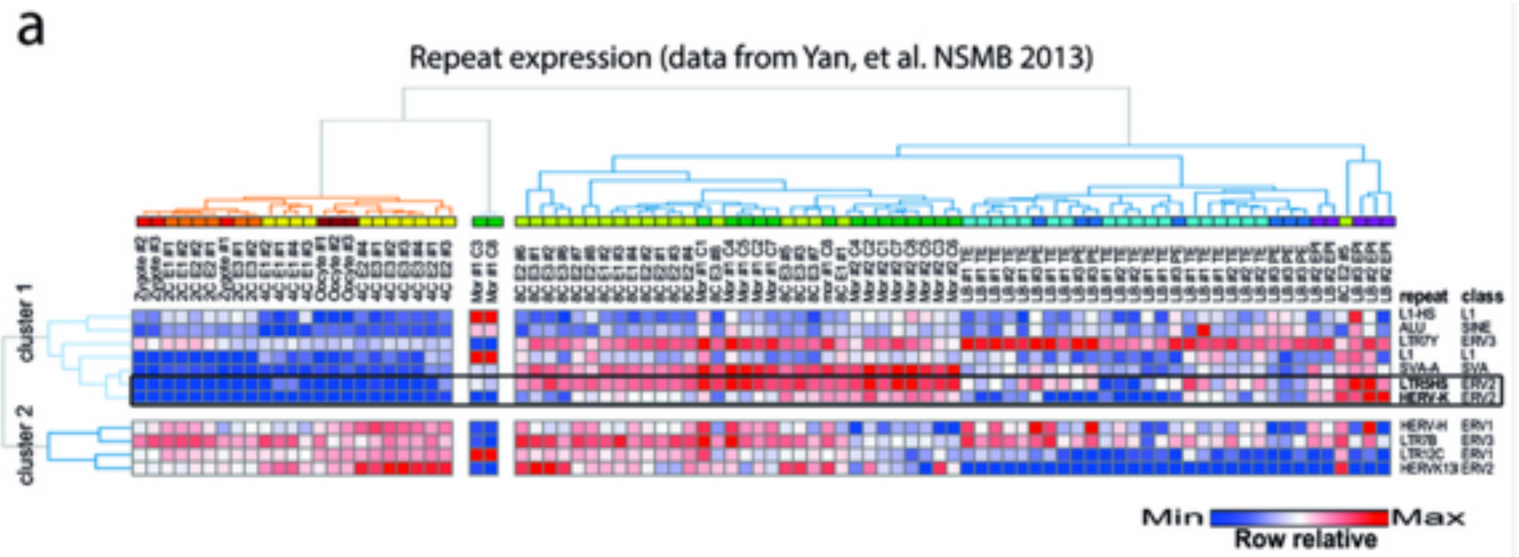
Criscione, S. W., Zhang, Y., Thompson, W., Sedivy, J. M., & Neretti, N. (2014). Transcriptional landscape of repetitive elements in normal and cancer human cells, *15*(1), 1–17. http://doi.org/10.1186/1471-2164-15-583

Transcription of TEs



Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells

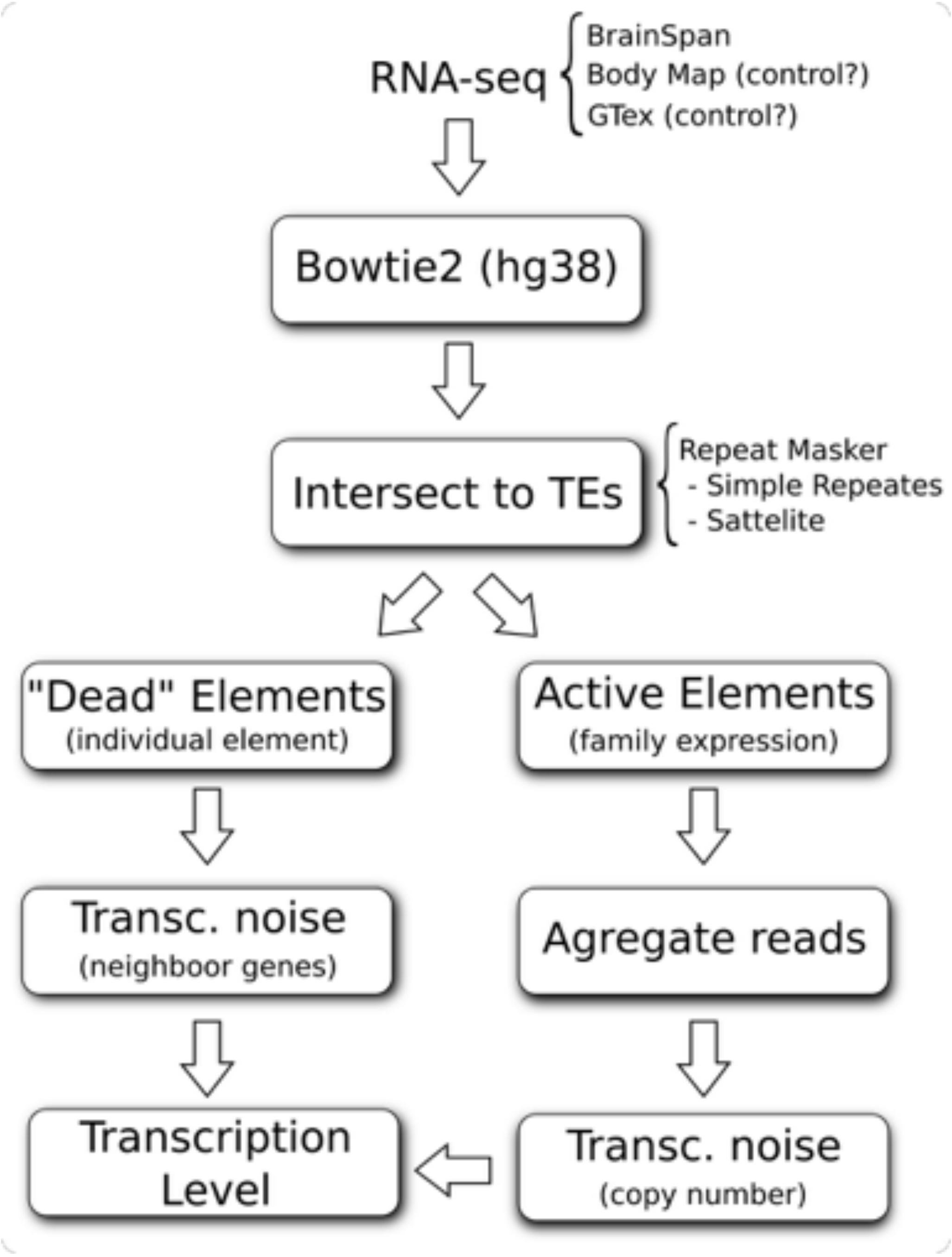
Edward J. Grow[1], Ryan A. Flynn[2], Shawn L. Chavez[3,4,5], Nicholas L. Bayless[6], Mark Wossidlo[1,3,4], Daniel J. Wesche[3], Lance Martin[2], Carol B. Ware[7], Catherine A. Blish[8], Howard Y. Chang[2], Renee A. Reijo Pera[1,3,4,9] & Joanna Wysocka[3,10,11]



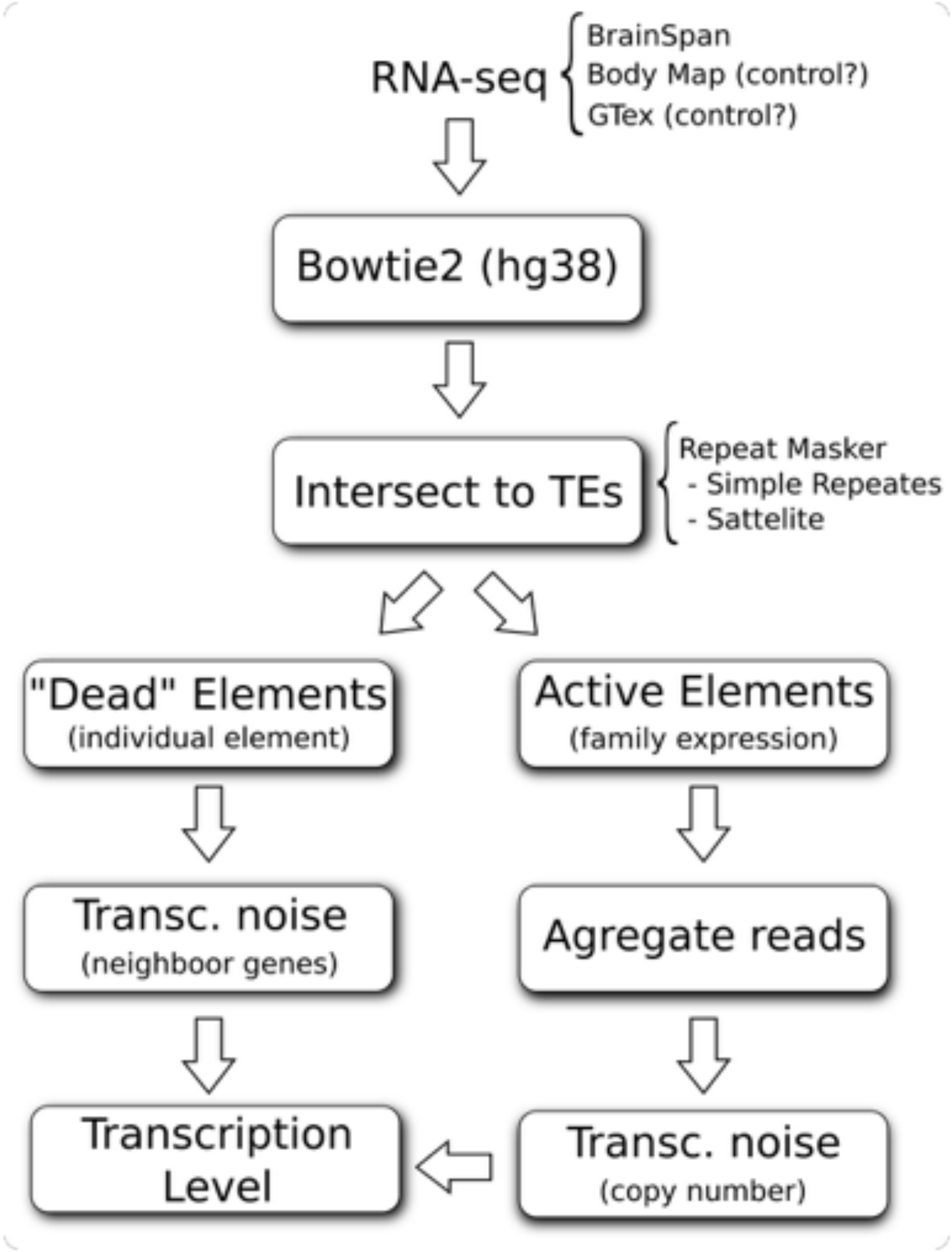
FASTQ files were aligned to repbase consensus sequences (downloaded from RepBase) with bowtie using the command “bowtie -q -p 8 -S -n 2 -e 70 -l 28-maxbts 800 -k 1 -best”. These bowtie parameters ensure that only the best alignment (highest scores) is reported, furthermore only one alignment per read is reported, that is, these settings do not allow multiple-matching.

Grow, E. J., Flynn, R. A., Chavez, S. L., Bayless, N. L., Wossidlo, M., Wesche, D. J., et al. (2015). Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature*. http://doi.org/10.1038/nature14308

Pipeline


RNA-seq
BrainSpan
Body Map (control?)
GTex (control?)
Bowtie2 (hg38)
Intersect to TEs
Repeat Masker
- Simple Repeates
- Sattelite
"Dead" Elements
(individual element)
Active Elements
(family expression)
Transc. noise
(neighboor genes)
Agregate reads
Transcription Level
Transc. noise
(copy number)

Pipeline



Body Map dataset:

- Illumina, poly(A), paired-end (75bp)
- 16 Human tissues (~30 samples)

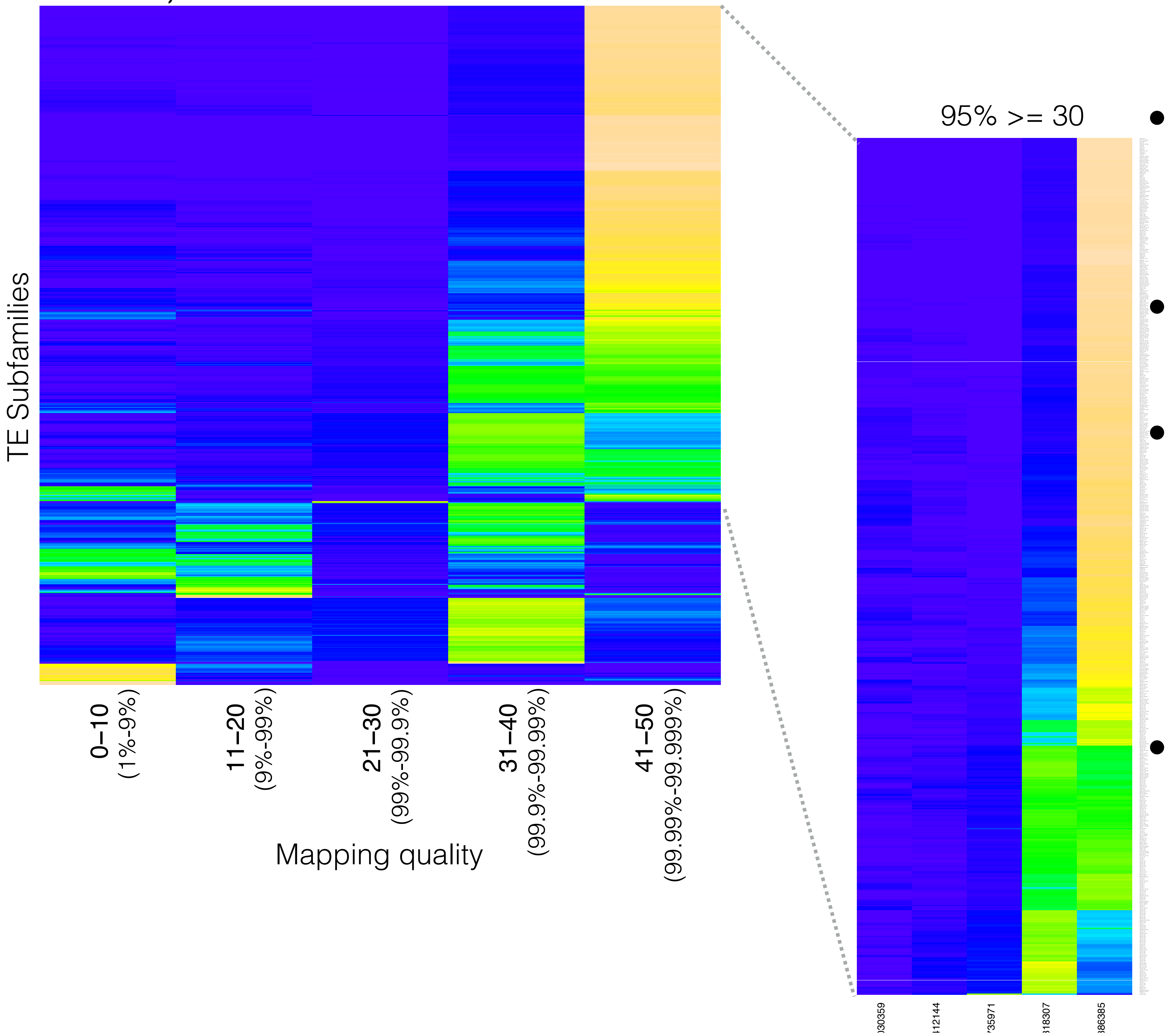
Brain Span dataset:

- ~600 samples
- Illumina, poly(A), shotgun (75bp)
- Human brain regions across many development periods.

Mappability of TEs

(Brain Span)

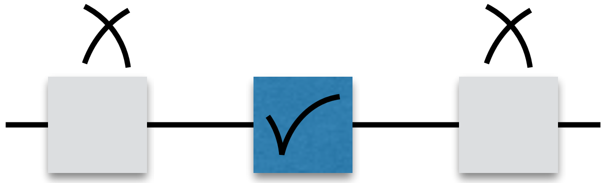
1,281 TE Subfamilies



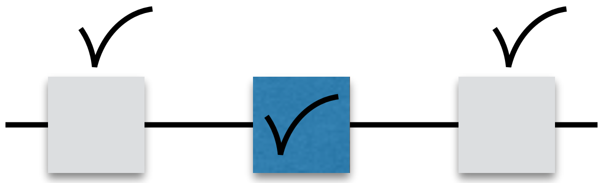
- 637 Subfamilies can be reliably mappable
- 1,286,924 loci
- 2,239 expressed loci
  5% of the samples
  with
  RPKM >= 1
- Quantile normalization

P1: Mappable TEs

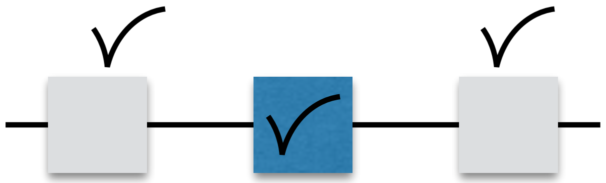
2,239 mappable expressed loci



Neighbor genes not expressed

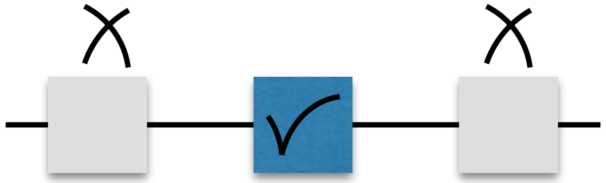


Neighbor genes not correlated

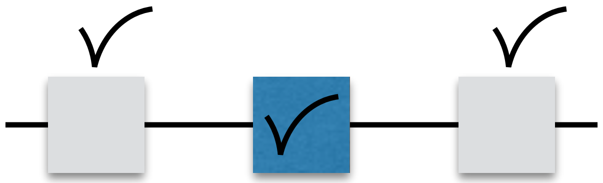


Correlated neighbor genes

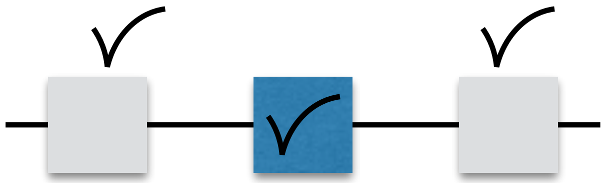
2,239 mappable expressed loci



Non-expressed neighbor genes



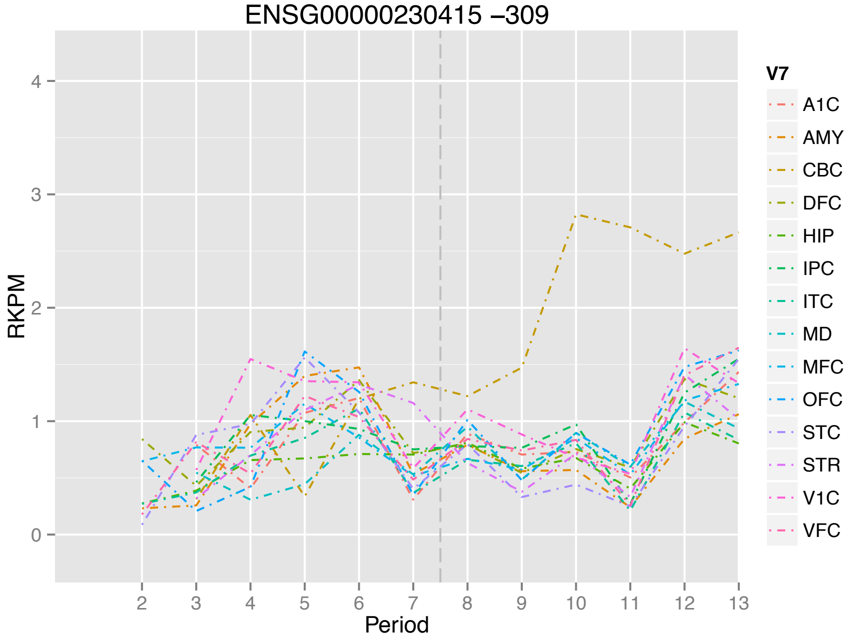
Uncorrelated neighbor genes

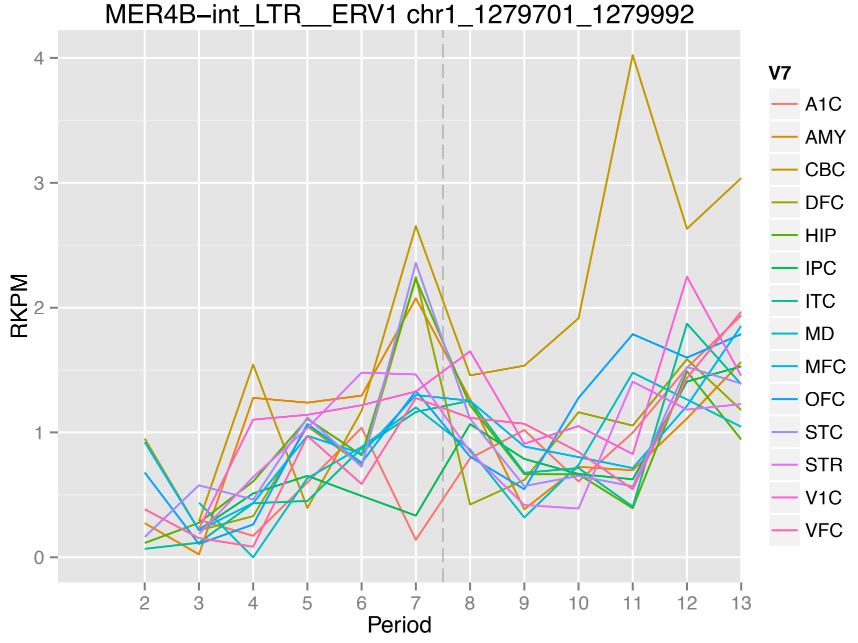


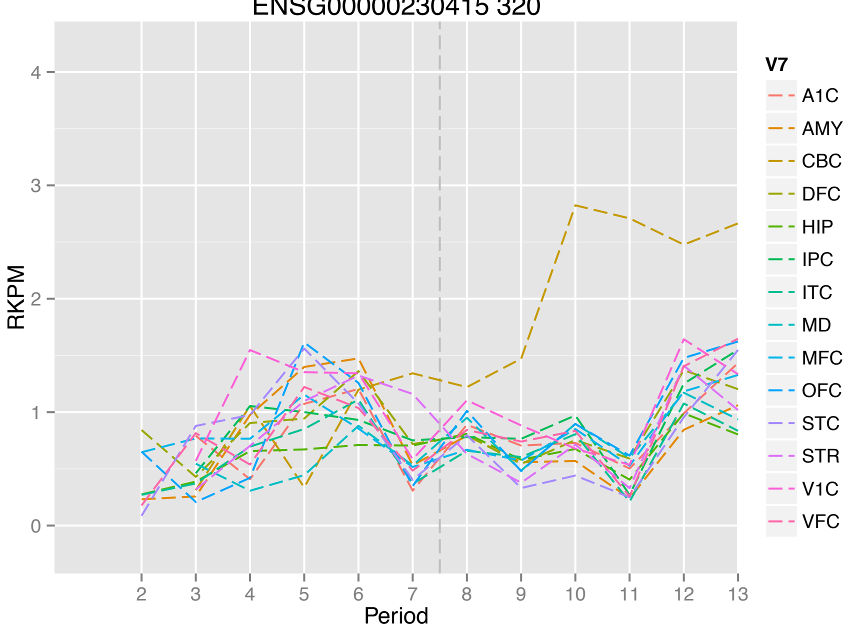
Correlated neighbor genes

~2,046 ✗

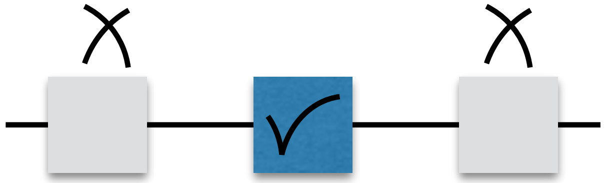
~91%



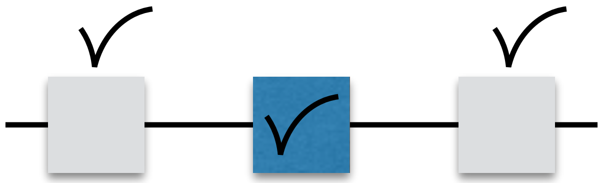




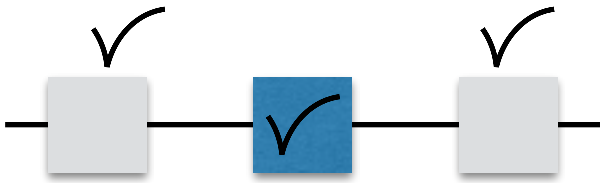
2,239 mappable expressed loci



Non-expressed neighbor genes

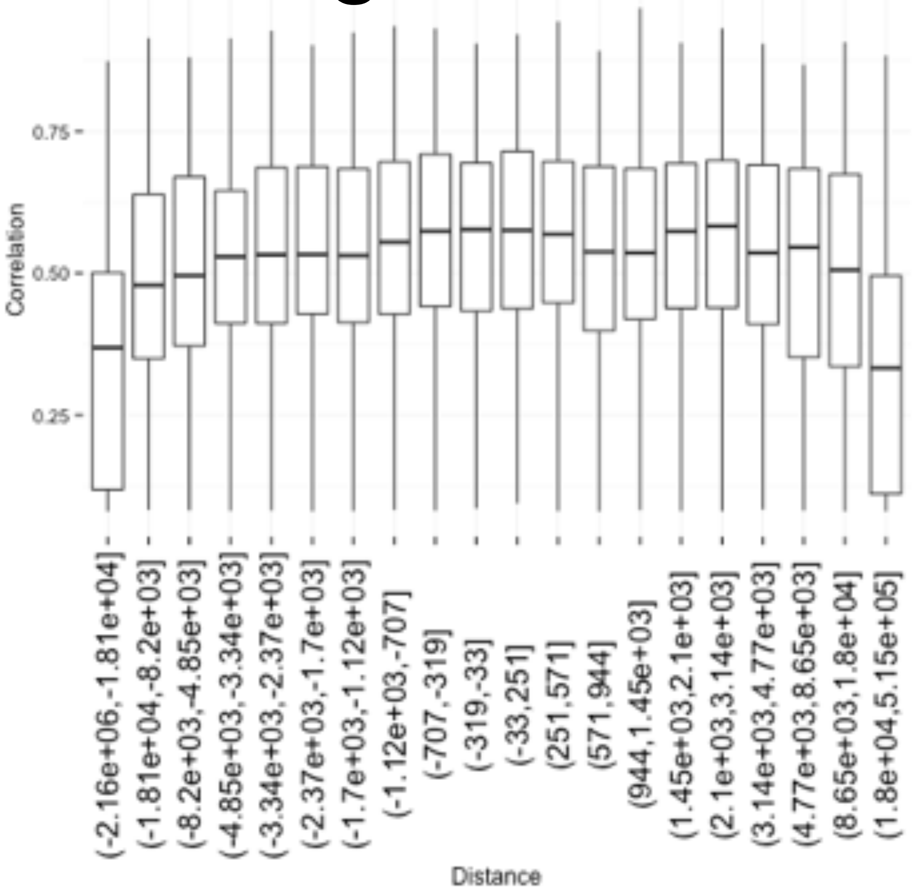


Uncorrelated neighbor genes

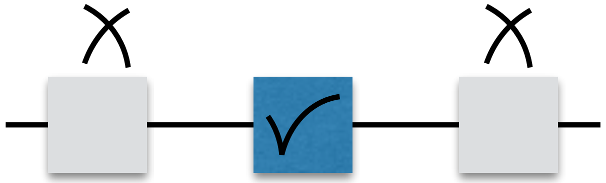


Correlated neighbor genes

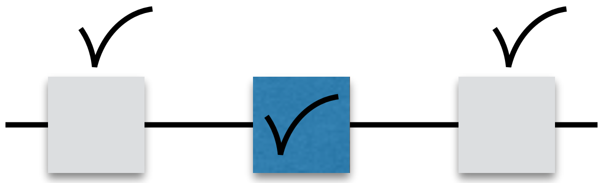
~2,046 ✗



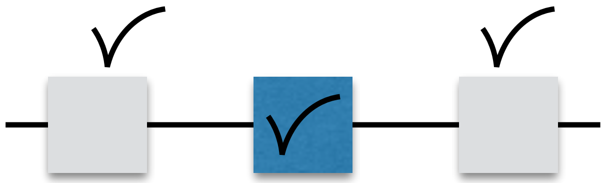
2,239 mappable expressed loci



Non-expressed neighbor genes



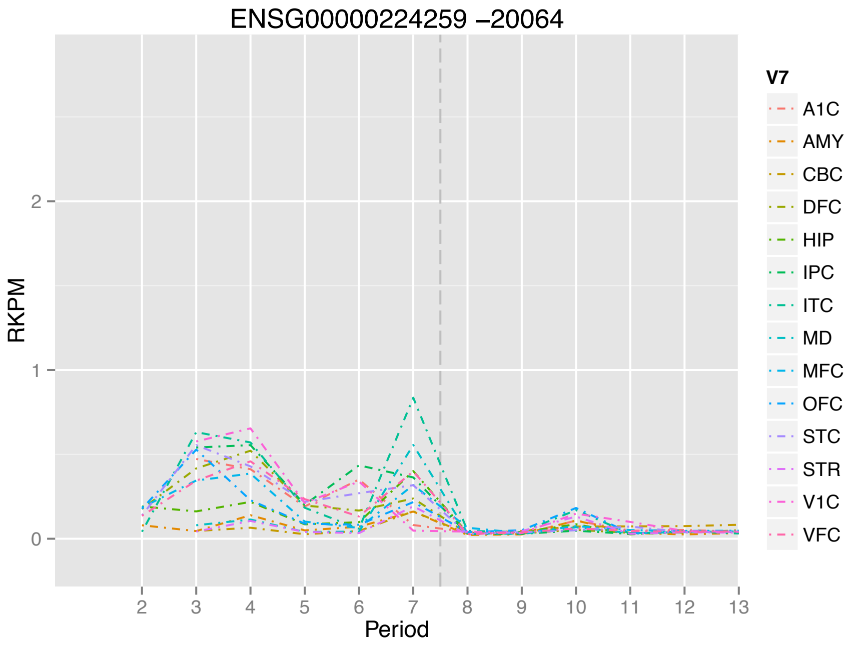
Uncorrelated neighbor genes

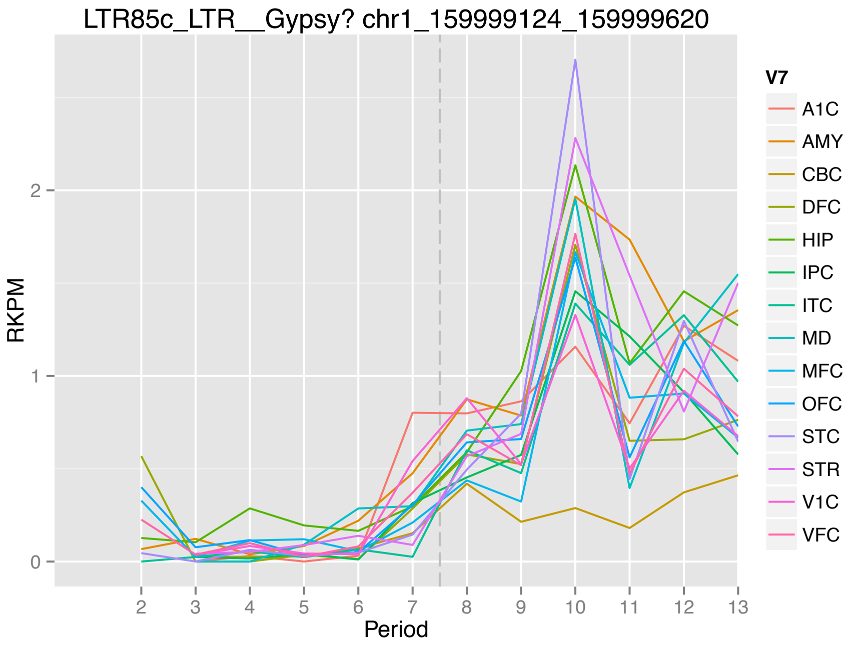


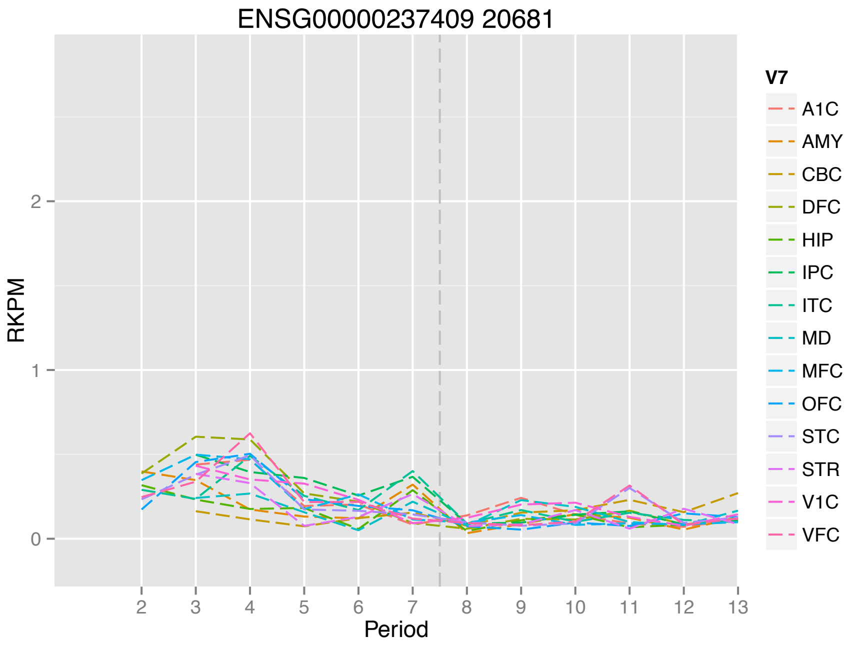
Correlated neighbor genes

~104 ✓

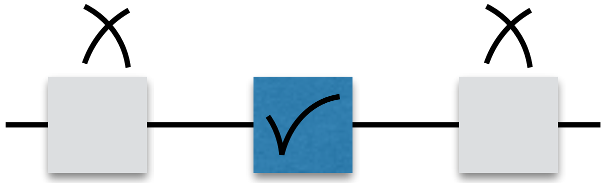
~2,046 ✗





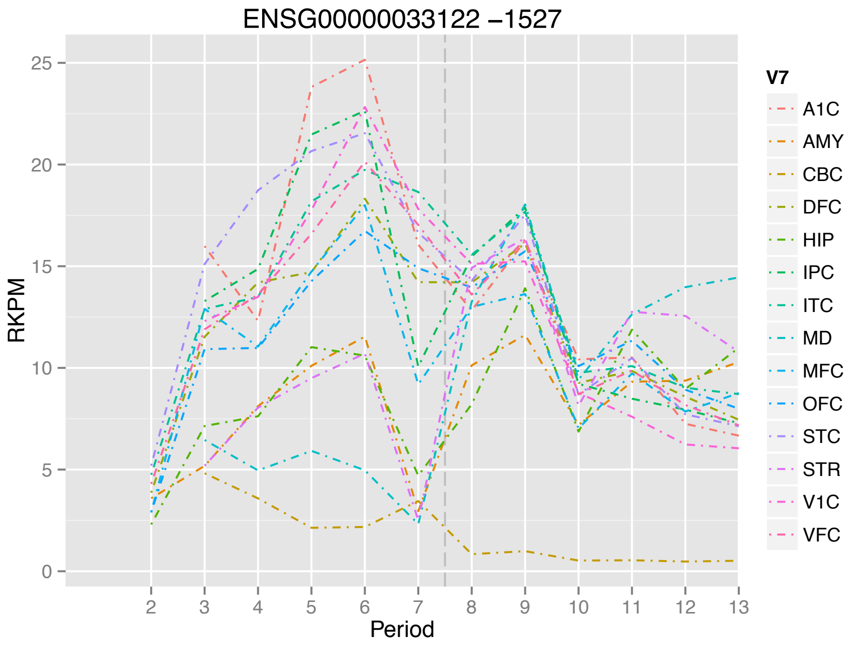


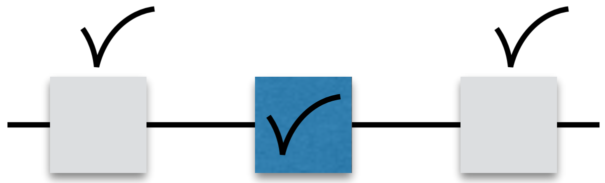
2,239 mappable expressed loci



Non-expressed neighbor genes

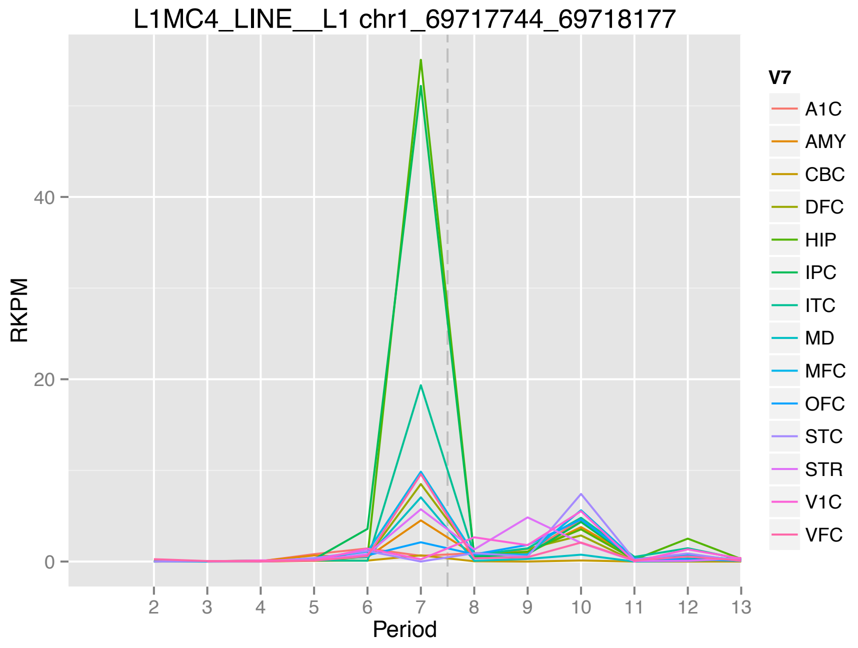
~104✓

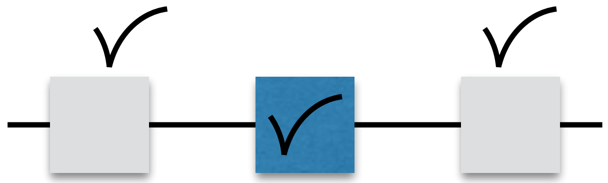

ENSG00000033122 −1527



Uncorrelated neighbor genes

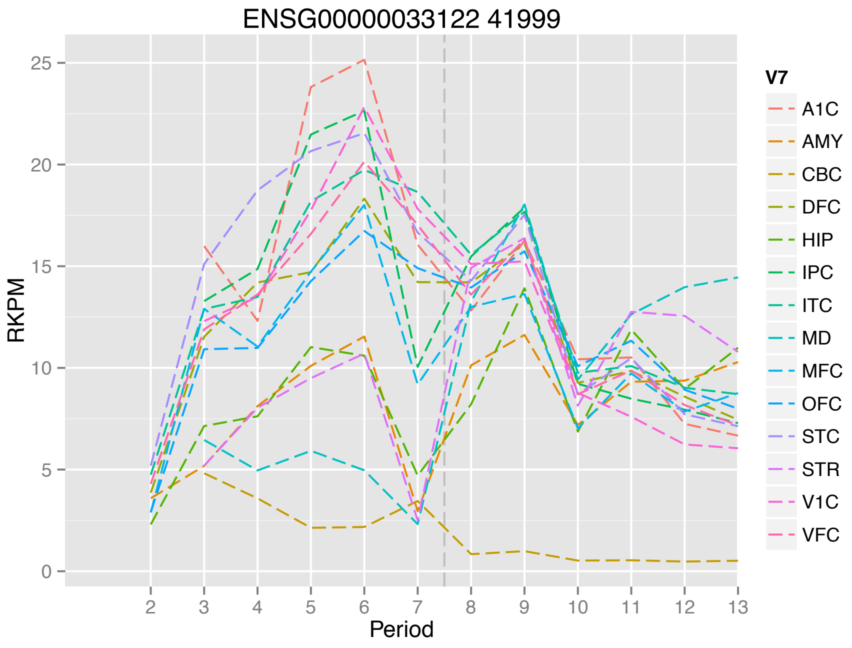
~189✓


L1MC4_LINE__L1 chr1_69717744_69718177



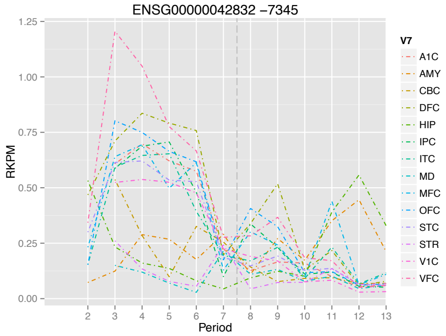
Correlated neighbor genes

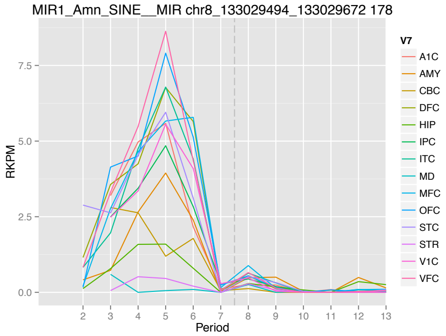
~2,046✗

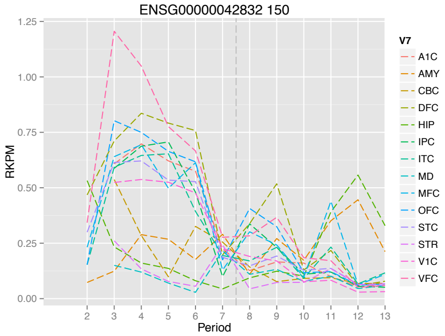

ENSG00000033122 41999

Contingent on the thresholds

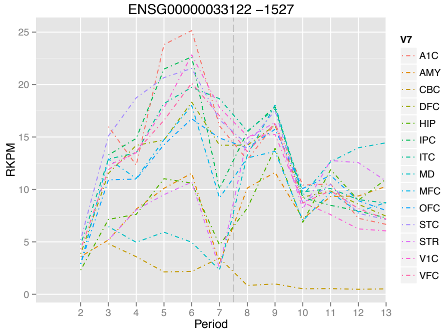
- Select TE in 20% of the samples with RPKM >= 1
  - Non-expressed neighbor genes: 33 (4% was 104) ✓

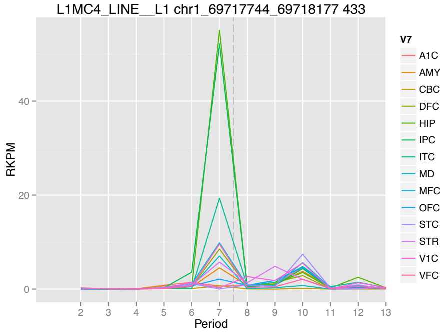


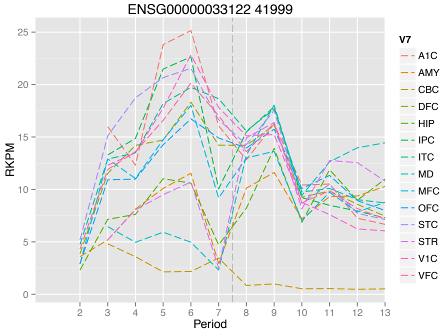




  - Uncorrelated neighbor genes: 44 (5.3%was 189) ✓

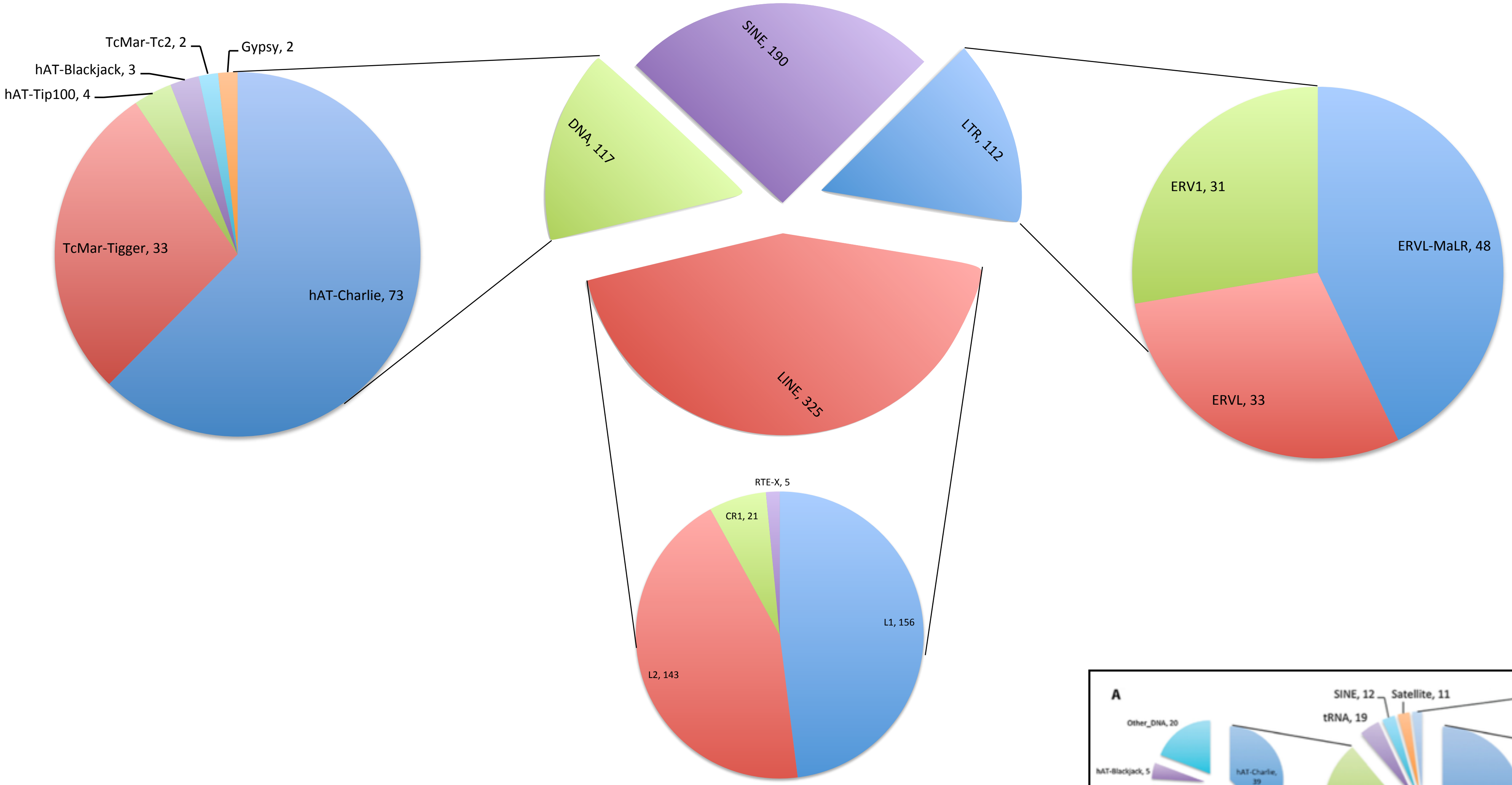


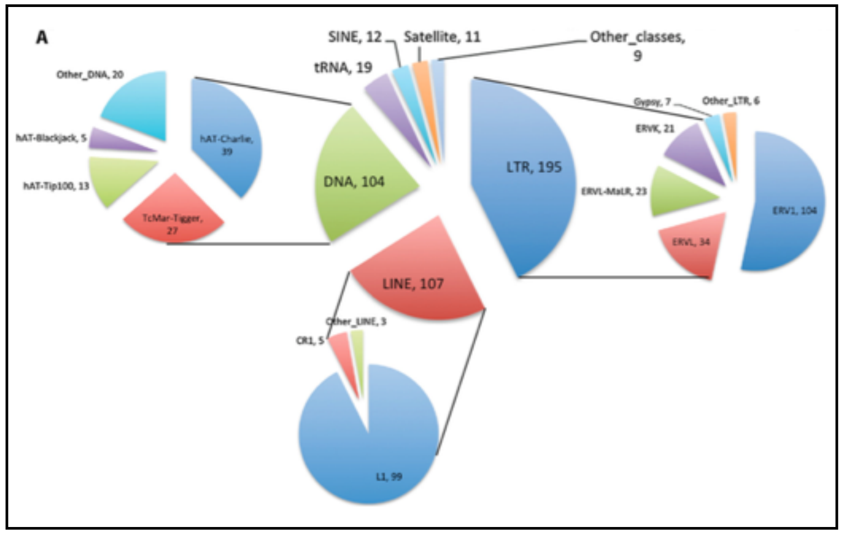




  - Correlated neighbor genes: 747 (90.7% was 2,046)

Transcription of TEs





Criscione et. al. may be accessing differentially expressed genes by indirectly evaluating the expression of TEs close to expressed genes.

Transcription of TEs

- Contingent on the number of samples and expression threshold… If I chose more stringent parameters, just a few TE are reported as independently expressed.

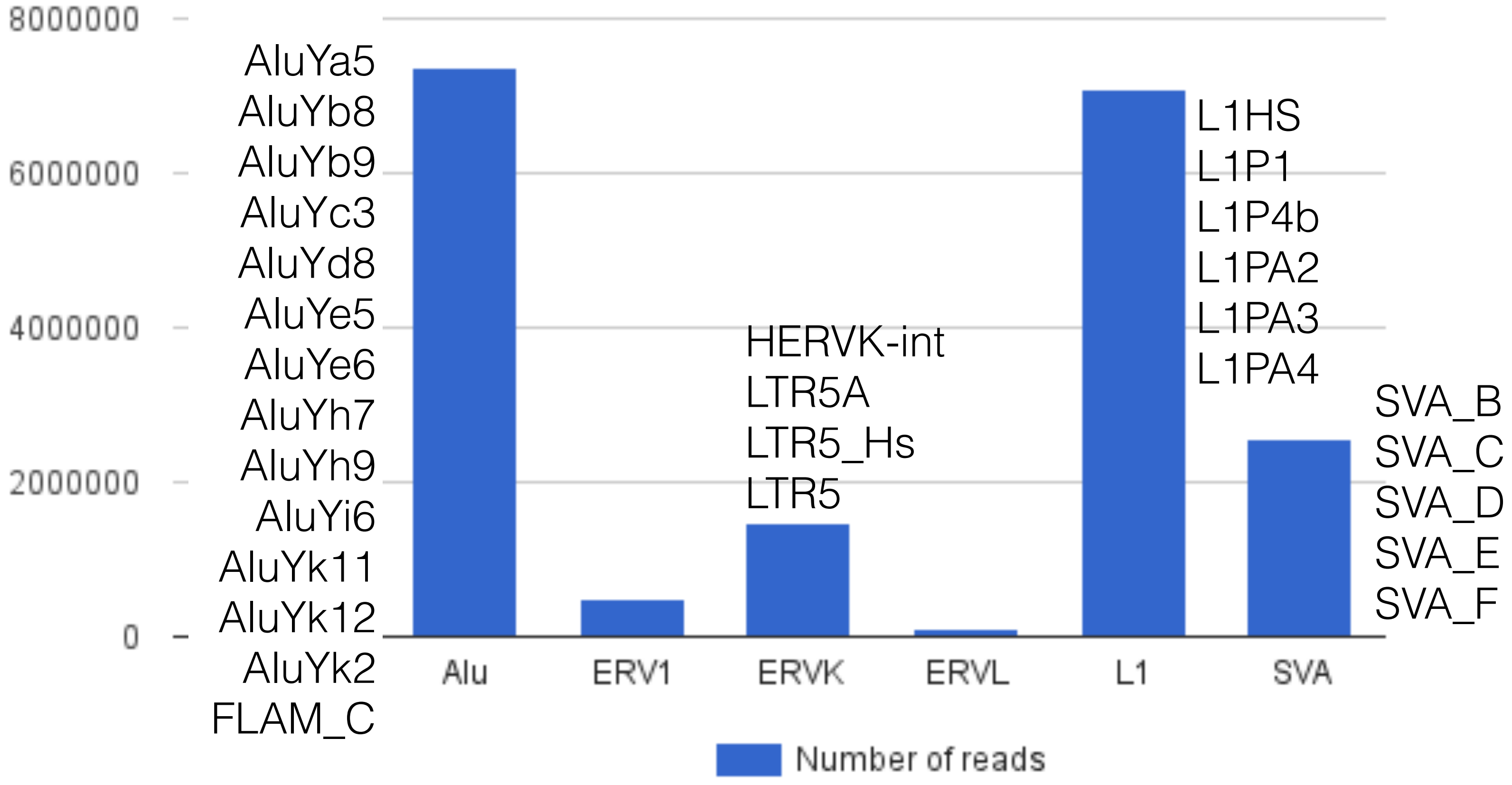
- Are there any TE independently expressed? What are the mechanisms?

    - RNA-seq+Chip-seq from ENCODE cell-lines?

- Are these transcripts functional? Probably not… But that makes sense. Remember: these are the dead elements!

P2: “Unmappable" TEs

- At least 75% of the reads aligned to the reference genome with mapping quality < 20



AluYa5
AluYb8
AluYb9
AluYc3
AluYd8
AluYe5
AluYe6
AluYh7
AluYh9
AluYi6
AluYk11
AluYk12
AluYk2
FLAM_C

HERVK-int
LTR5A
LTR5_Hs
LTR5

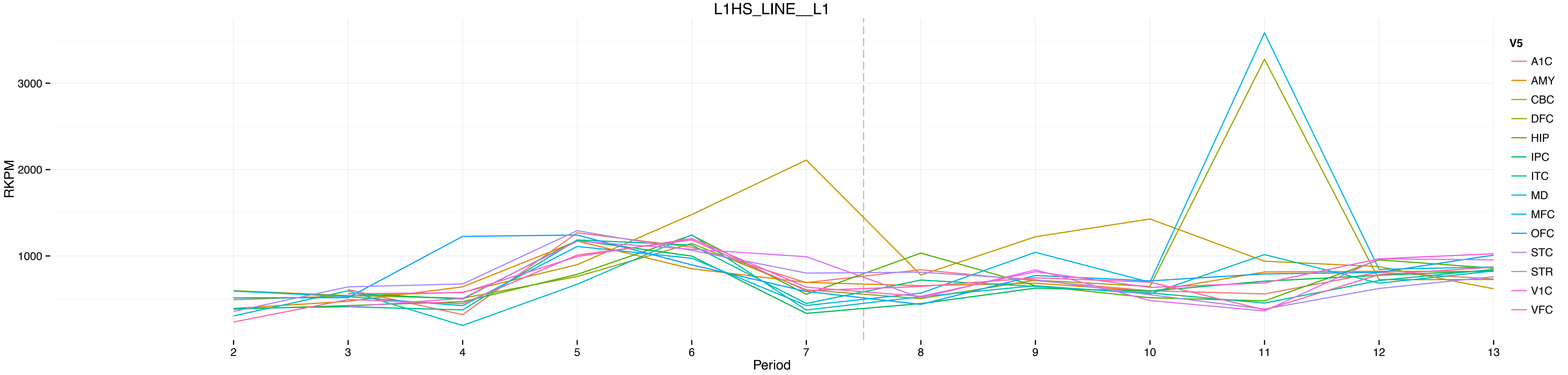
L1HS
L1P1
L1P4b
L1PA2
L1PA3
L1PA4

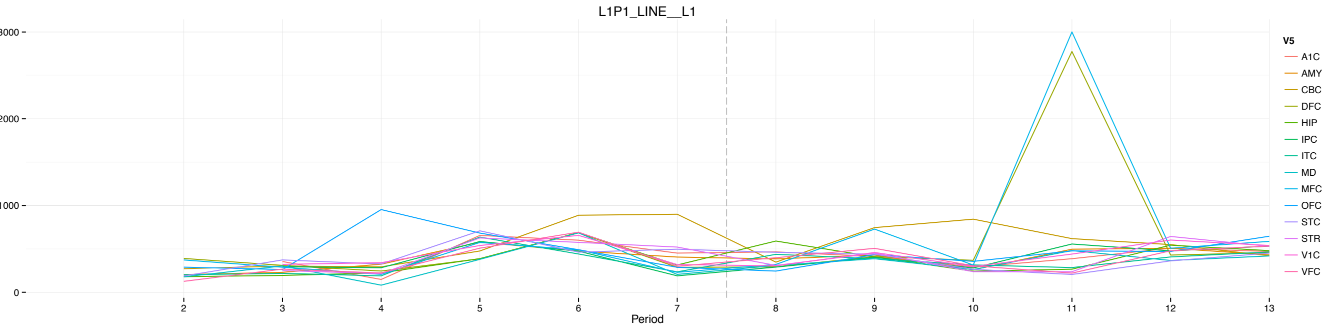
SVA_B
SVA_C
SVA_D
SVA_E
SVA_F

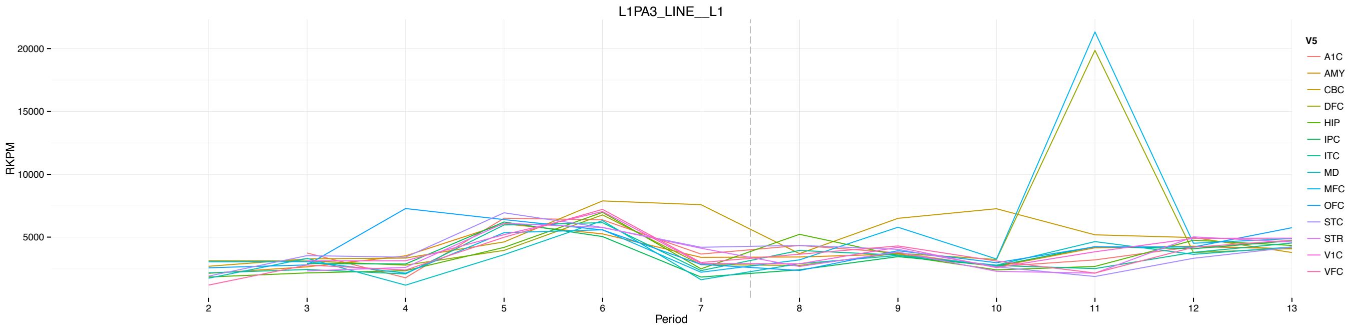
- L1, HERV-K/LTR, AluY, FLAM_C, SVA

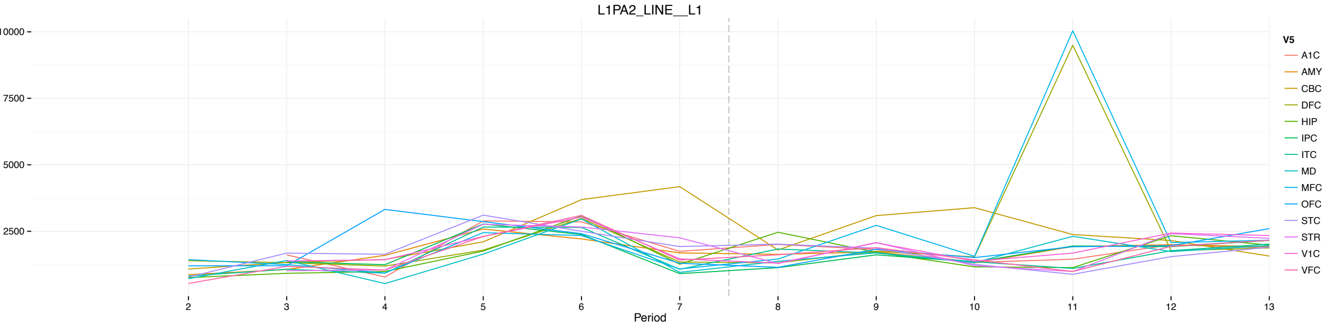
~80% of the alignments over L1Hs have mapping quality = 0

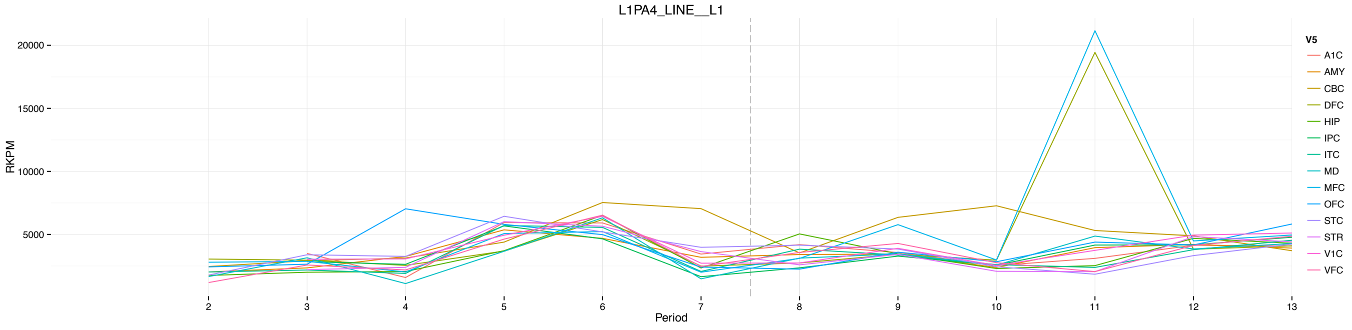
L1


L1HS_LINE__L1

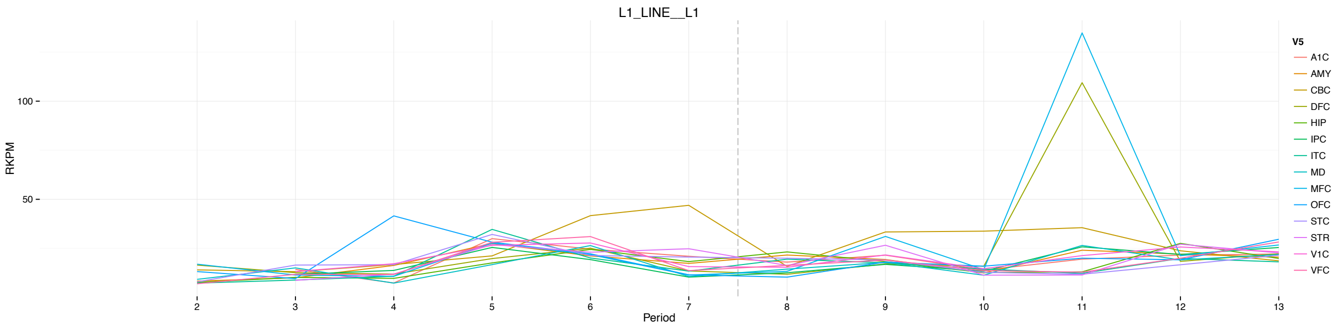

L1P1_LINE__L1

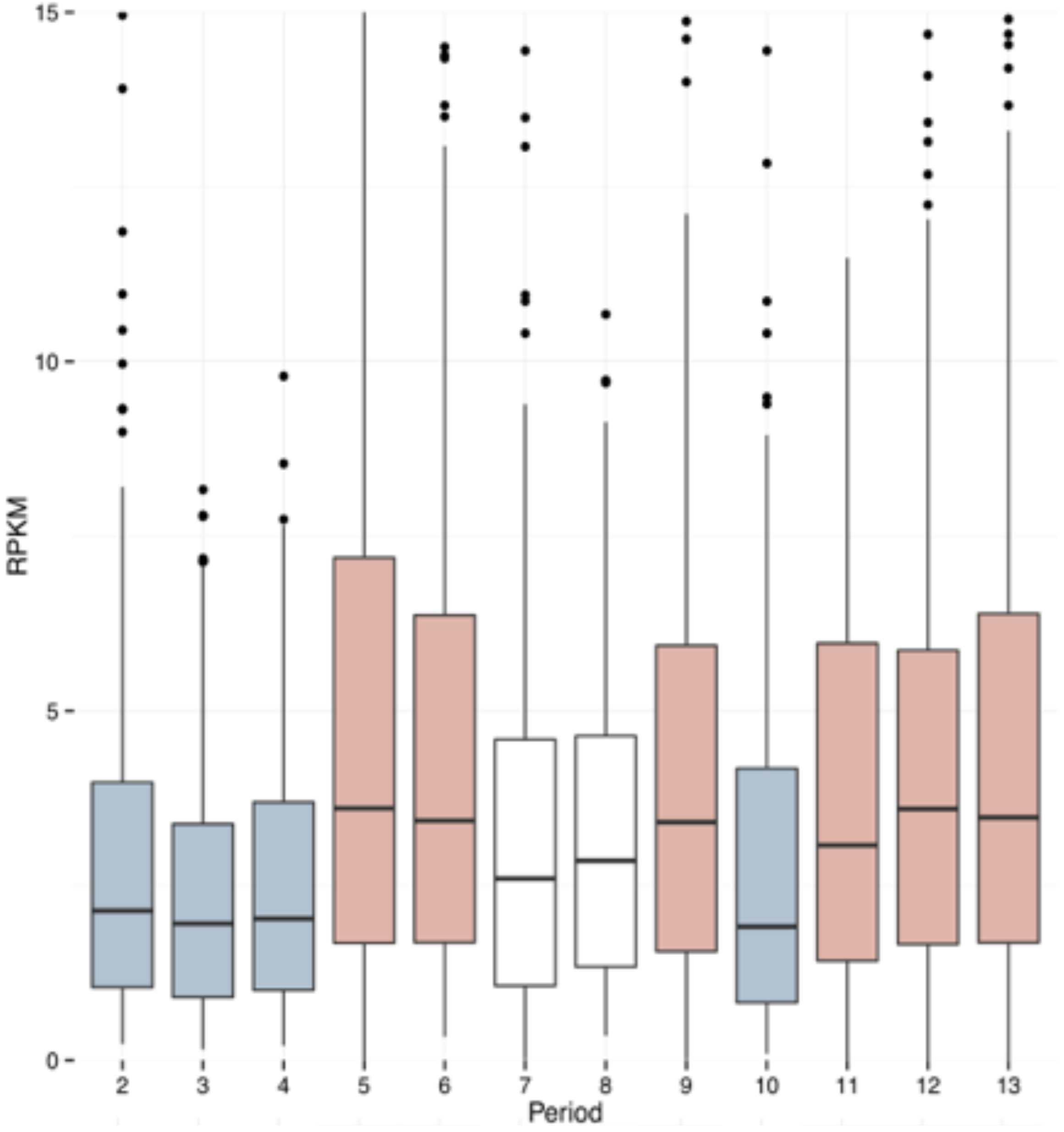

L1PA3_LINE__L1


L1PA2_LINE__L1


L1PA4_LINE__L1

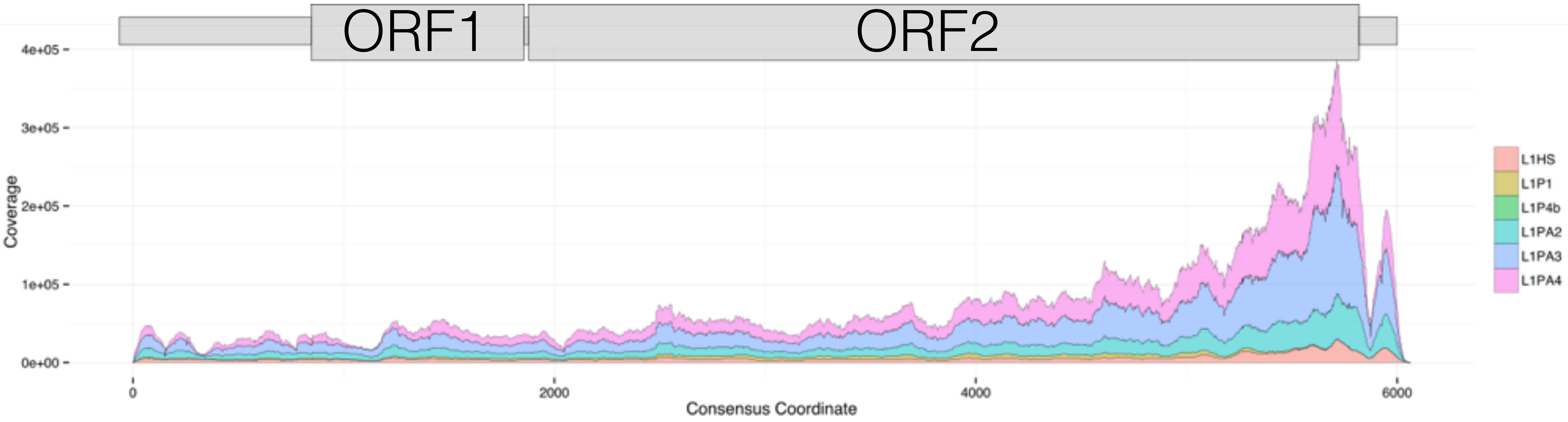
L1


L1_LINE__L1
RKPM
100
50
Period
2
3
4
5
6
7
8
9
10
11
12
13
V5
A1C
AMY
CBC
DFC
HIP
IPC
ITC
MD
MFC
OFC
STC
STR
V1C
VFC


RPKM
15
10
5
0
Period
2
3
4
5
6
7
8
9
10
11
12
13

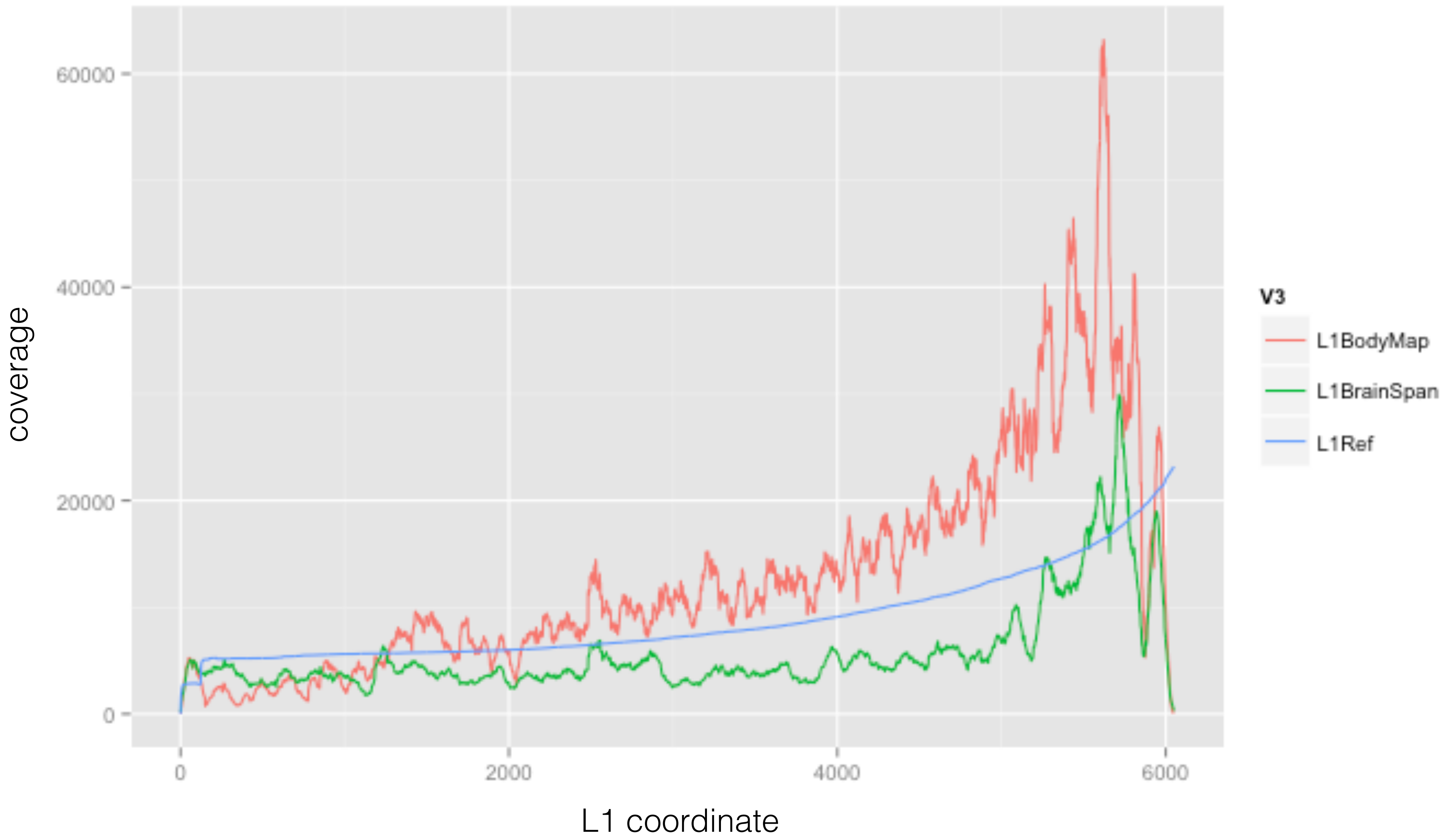
L1

Fetch all reads on L1HS, L1PA2, L1PA3 and L1PA4 and align to a reference L1HS.

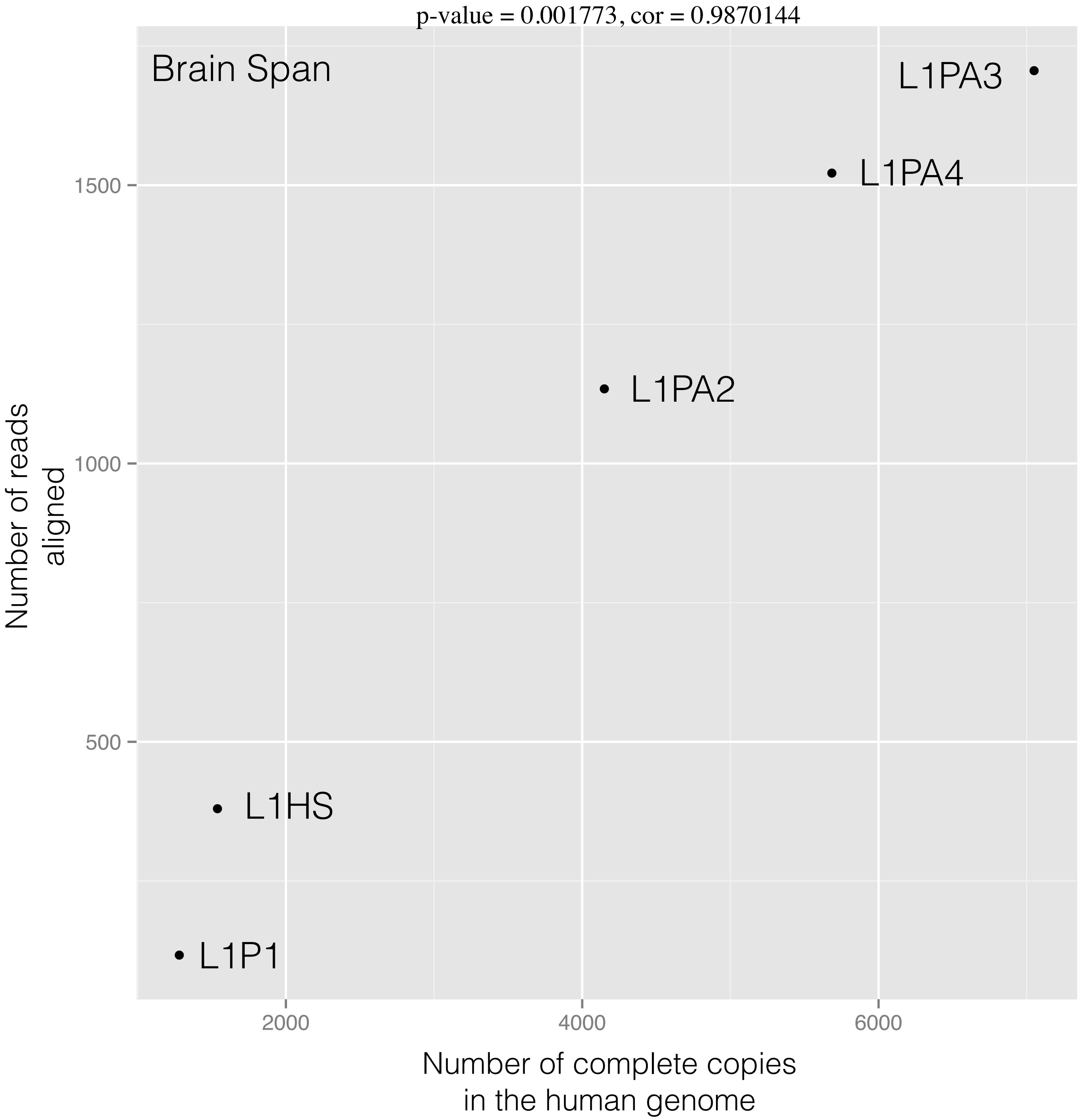


L1 background transcription

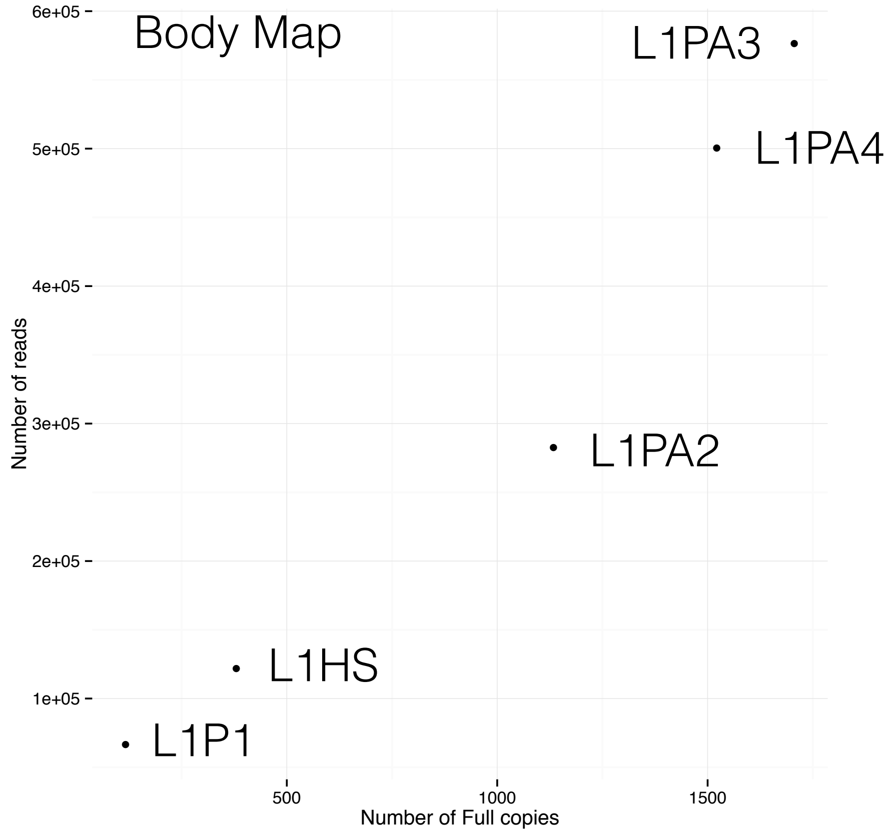
(copy number)



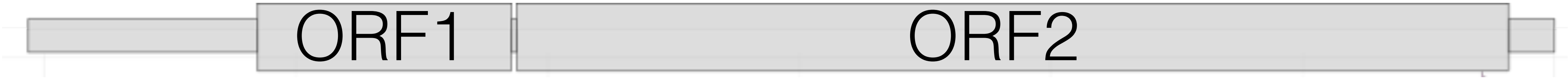
L1 background transcription

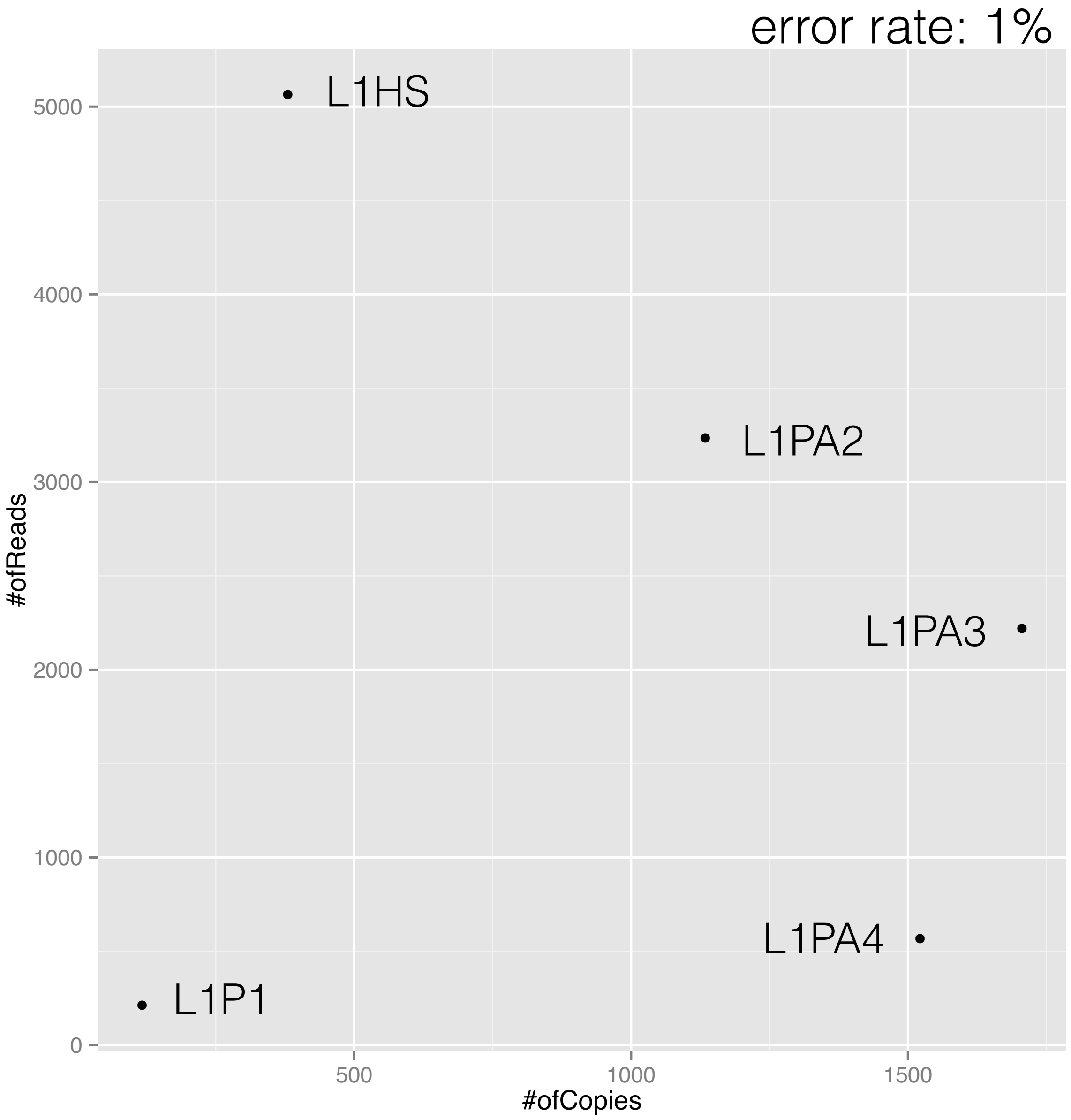


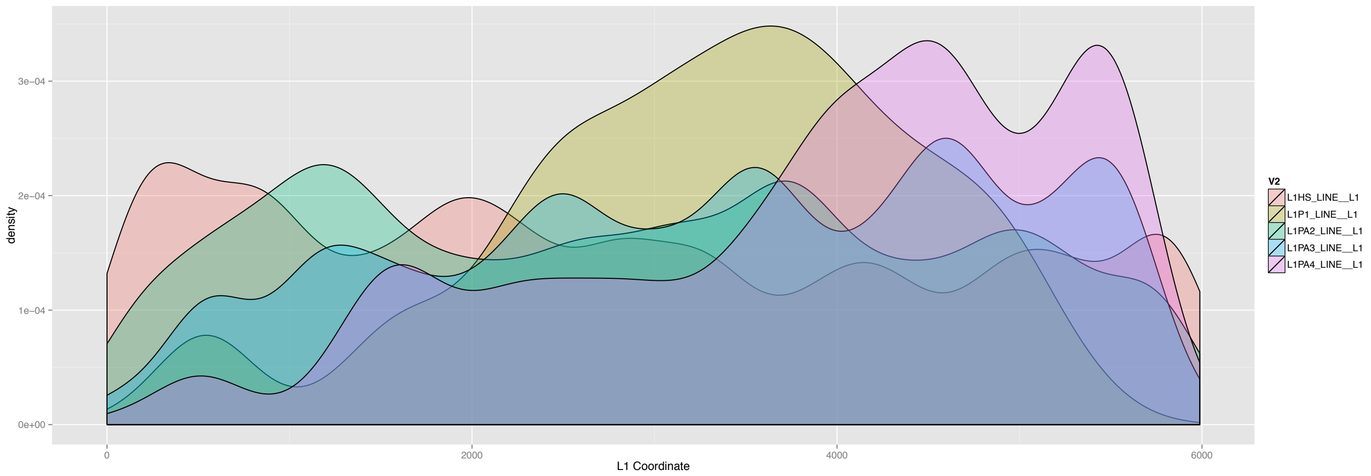
```
Mismatches
x$V10: L1HS_LINE__L1
  . 1st Qu. Median Mean 3rd Qu. Max.
  000 0.0000 0.0000 0.9118 1.0000 11.0000
 ----------------------------------------------------------------
 '10: L1P1_LINE__L1
  . 1st Qu. Median Mean 3rd Qu. Max.
  00 1.000 3.000 2.828 4.000 13.000
 ----------------------------------------------------------------
 '10: L1PA2_LINE__L1
  . 1st Qu. Median Mean 3rd Qu. Max.
  00 0.000 1.000 1.604 2.000 12.000
 ----------------------------------------------------------------
 '10: L1PA3_LINE__L1
  . 1st Qu. Median Mean 3rd Qu. Max.
  00 1.000 2.000 2.324 3.000 14.000
 ----------------------------------------------------------------
 '10: L1PA4_LINE__L1
  . 1st Qu. Median Mean 3rd Qu. Max.
  00 2.000 3.000 3.144 4.000 15.000
```

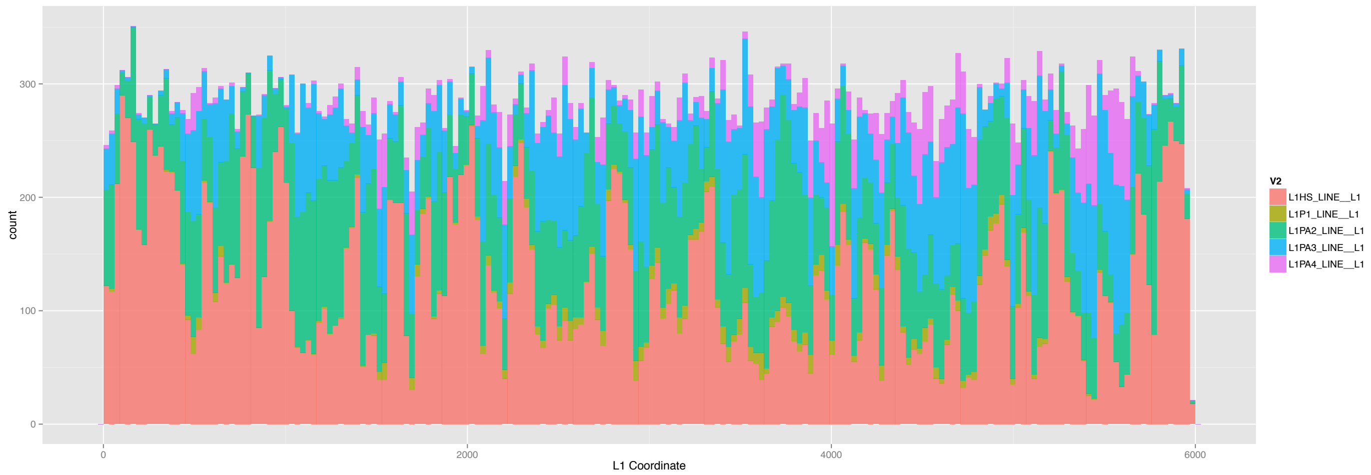


Simulated L1 transcripts alignment (wgsim)



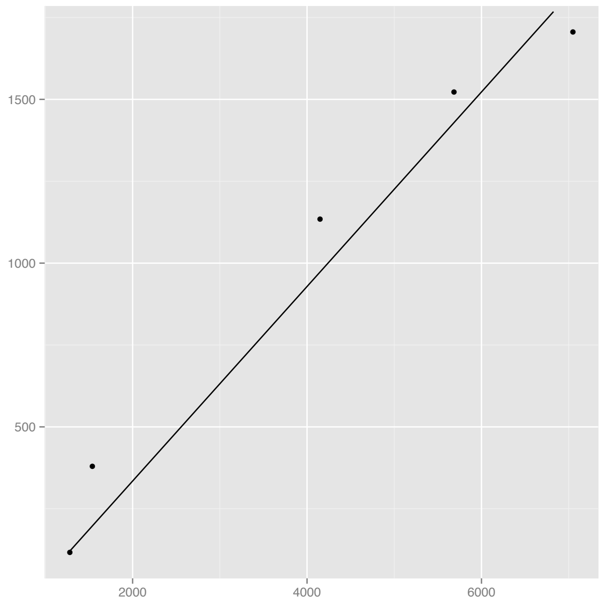






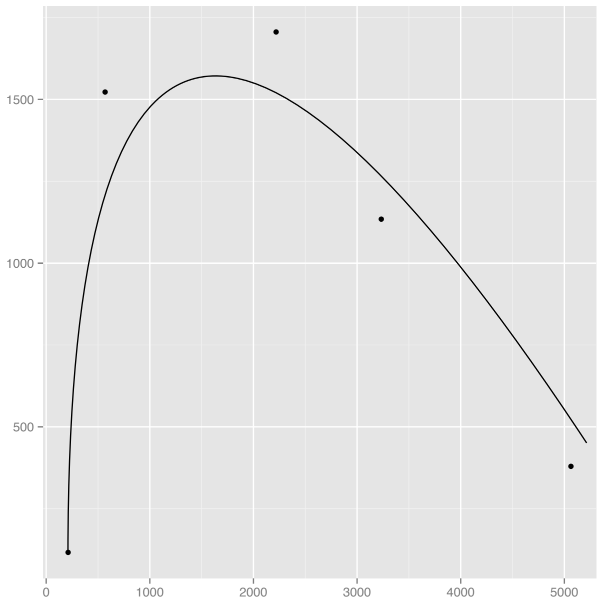
Overcoming background transcription

[N]%

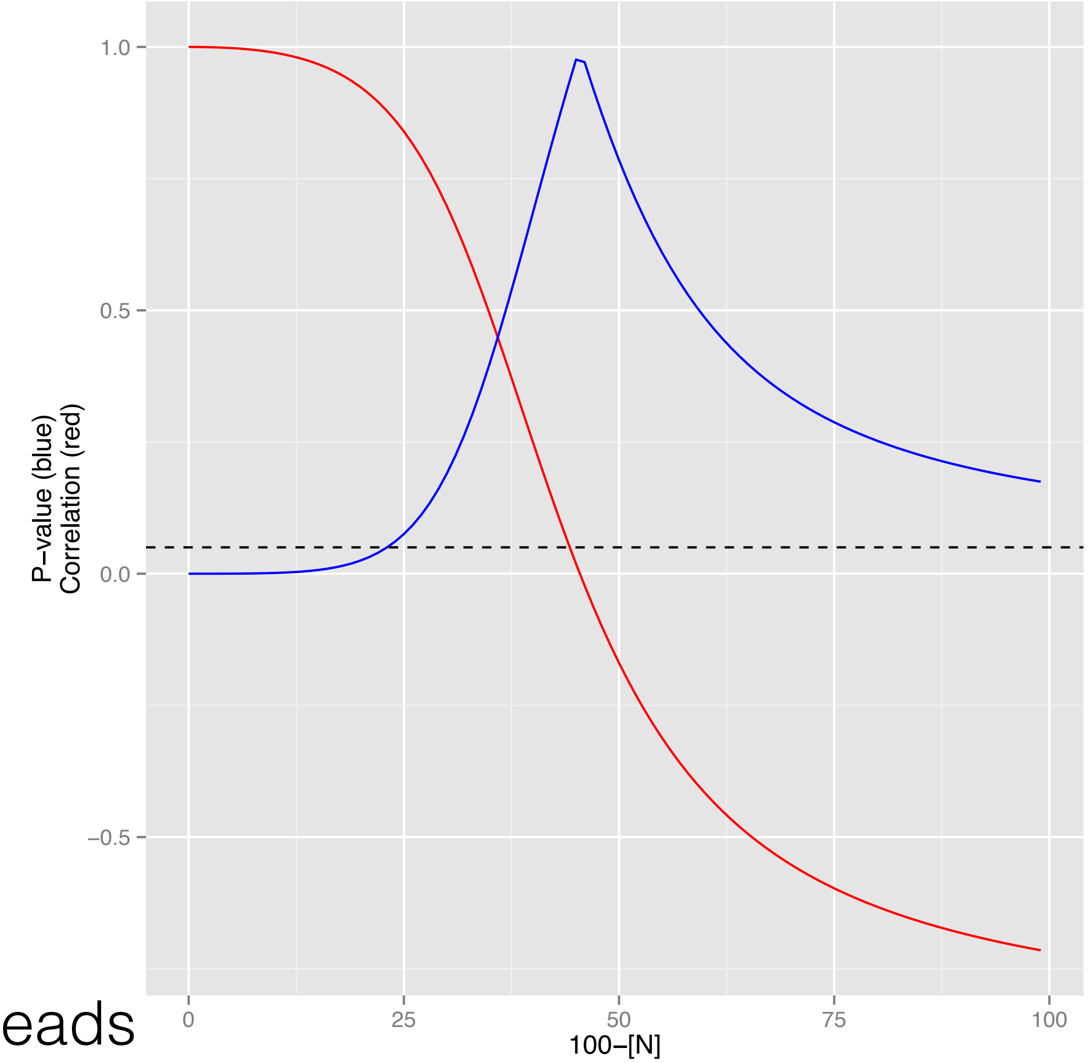


Background

100-[N]%



L1
Transcription



~24%;40% of the reads
would have to originate from L1HS transcripts
to overcome the background transcription noise

~15;~25 RPKM

(Body Map average expression)

Is it possible to evaluate L1 expression?

- All runs from Brain Span, Body Map and Lung cancer have a high significant correlation between the number of copies and the number of reads mapped on L1 subfamilies.

- Look into embryonic stem single cell transcriptome data to evaluate the expression of L1 (evidence of L1 activity).

Conclusions

- Most of the repetitive elements are transcribed by background activity of RNA pol II.

- Use other samples with known L1 activity as positive control. Suggestions?

- Start working on a small preprint with negative results.

- Carefully interpret results from highly duplicated regions. These observations may be also pertinent to pseudogene expression, chip-seq data and sRNA.