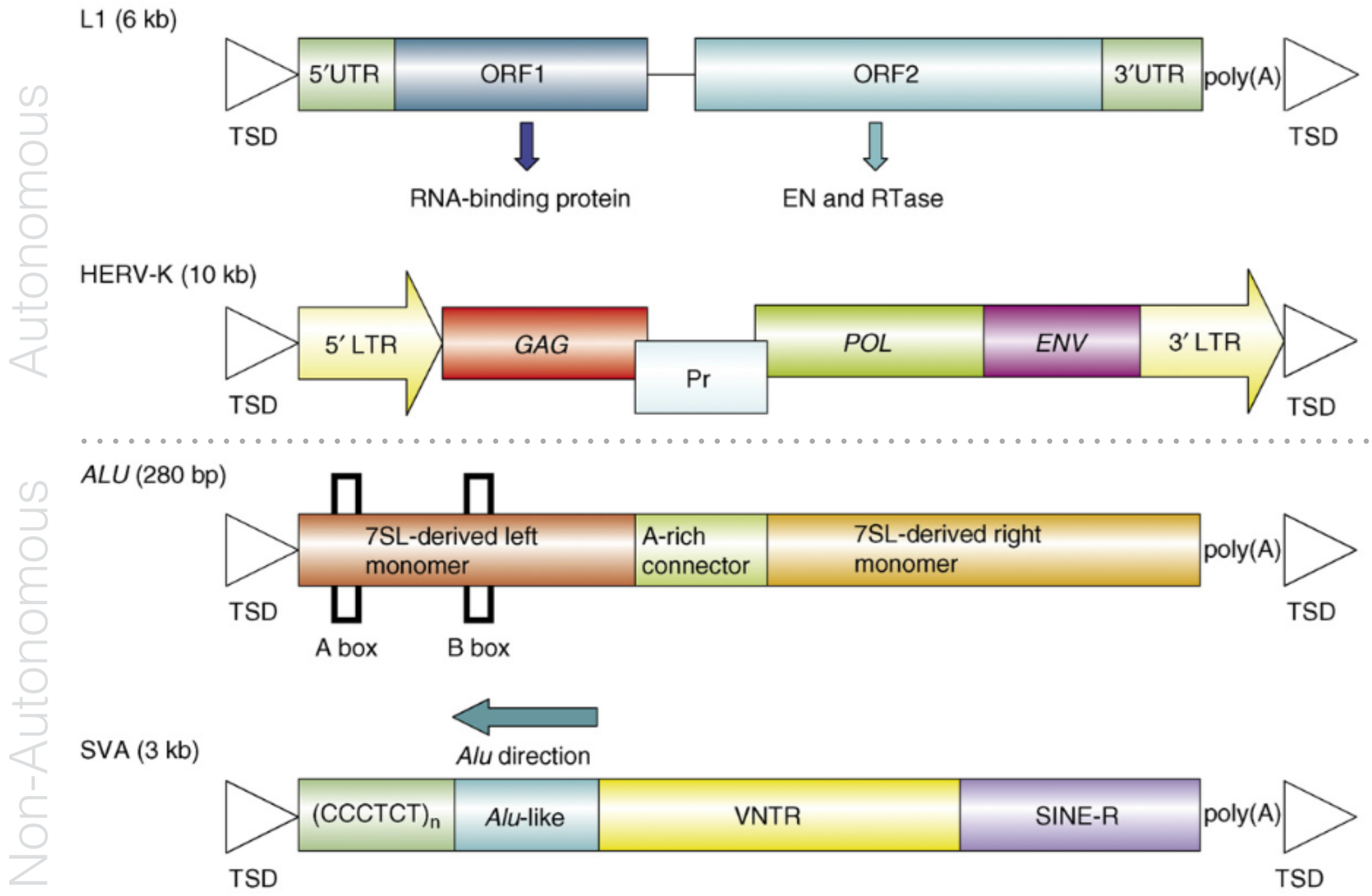
Genomic mobilization of ALUs in tumoral samples

Group Meeting - January 2016

FN

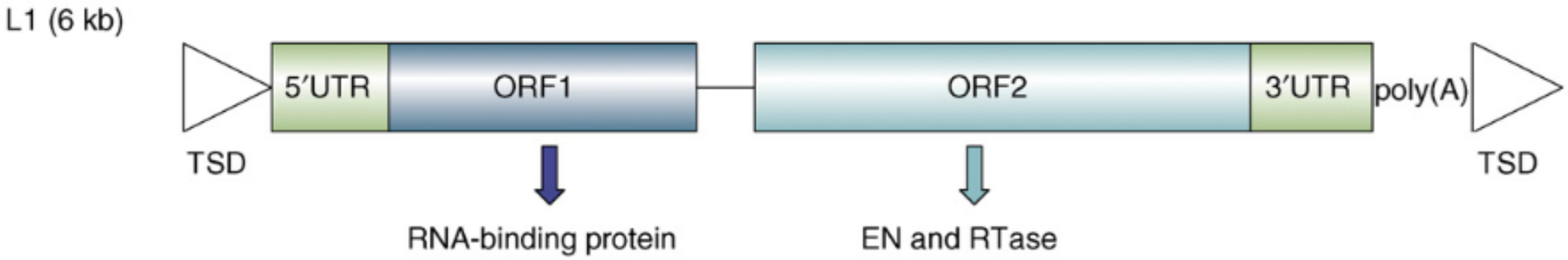
Repetitive Elements

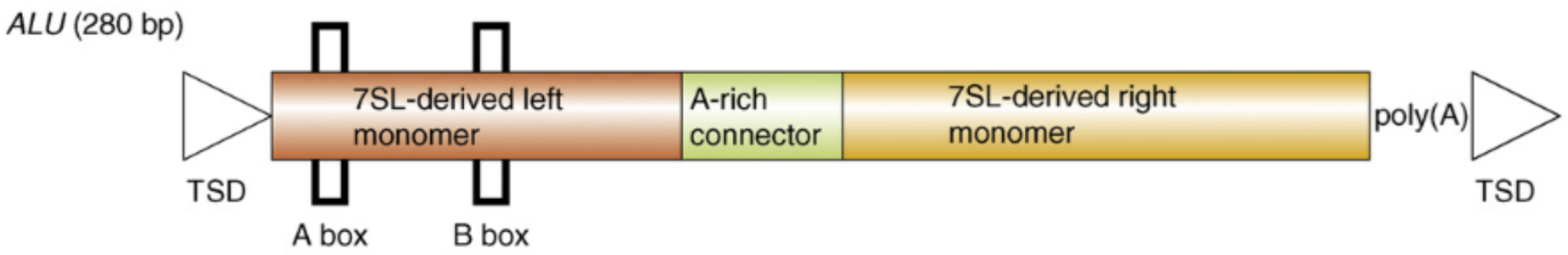


Mills, R. E., Bennett, E. A., Iskow, R. C., & Devine, S. E. (2007). Which transposable elements are active in the human genome? *Trends in Genetics : TIG*, *23*(4), 183–191. http://doi.org/10.1016/j.tig.2007.02.006

Repetitive Elements

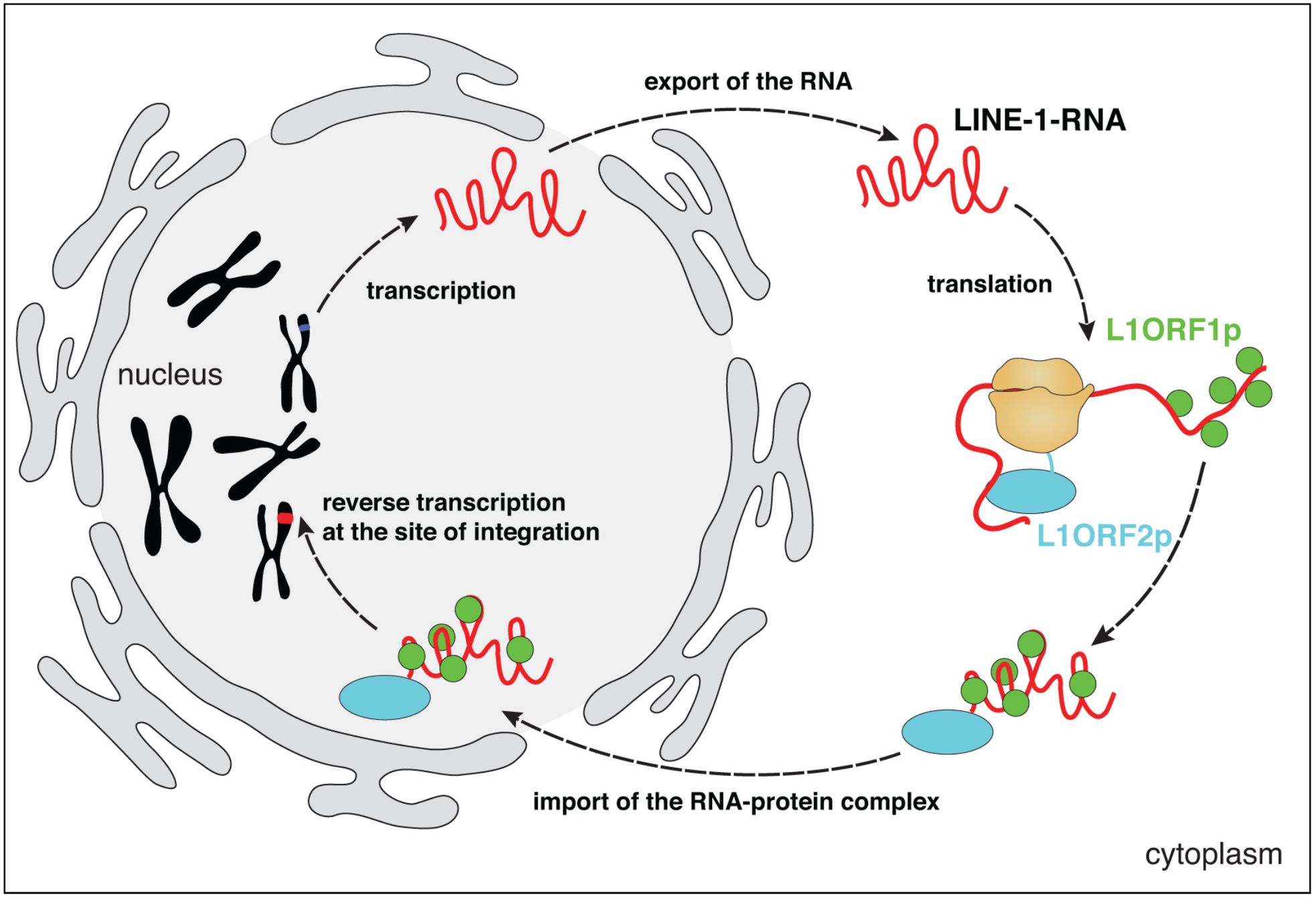
Non-Autonomous





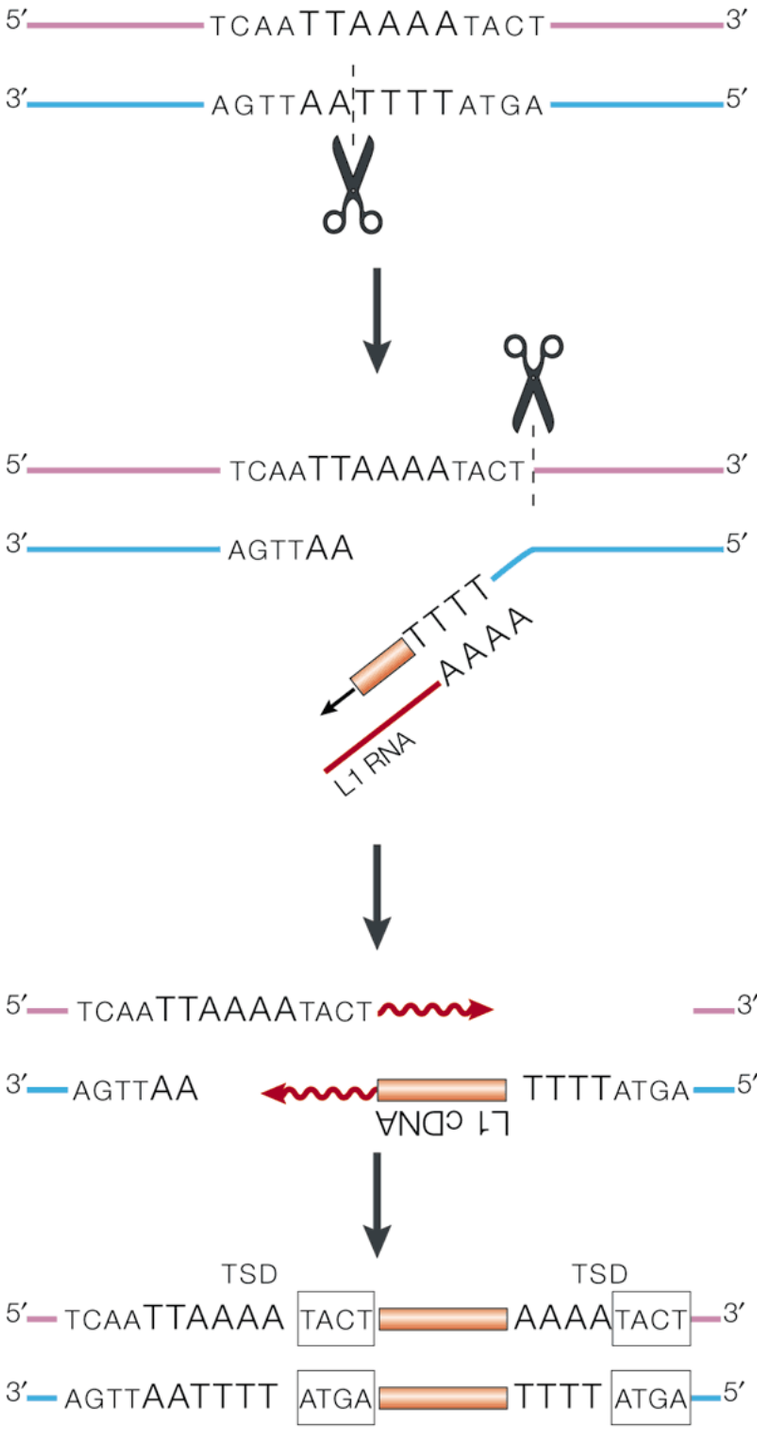
Mills, R. E., Bennett, E. A., Iskow, R. C., & Devine, S. E. (2007). Which transposable elements are active in the human genome? *Trends in Genetics : TIG*, *23*(4), 183–191. http://doi.org/10.1016/j.tig.2007.02.006

L1 and ALU cycle



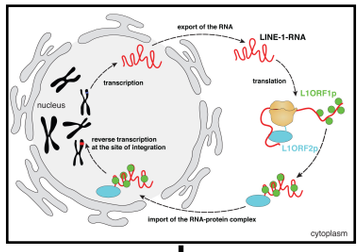
Non-LTR retrotransposons encode noncanonical RRM domains in their first open reading frame. E. Khazina, et al. PNAS, 2009.

Target Site Duplication
Target primed reverse transcription



Nature Reviews | Genetics

Consequences of L1 activity in tumors



L1

SVA?

ALU

Coding transcripts

Iskow, R. C., McCabe, M. T., Mills, R. E., Torene, S., Pittard, W. S., Neuwald, A. F., et al. (2010). **Natural mutagenesis of human genomes by endogenous retrotransposons.** ***Cell***, *141*(7), 1253–1261. http://doi.org/10.1016/j.cell.2010.05.020

"60 samples that were examined […], 650 novel L1 insertions were identified […] An additional 403 novel *Alu* insertions were identified from the *Alu* assay"

~11 L1 ins/tumor

~6.7 ALU ins/tumor

Lee, E., Iskow, R., Yang, L., Gokcumen, O., Haseley, P., Luquette, L. J., et al. (2012). **Landscape of somatic retrotransposition in human cancers.** ***Science*** *(New York, N.Y.)*, *337*(6097), 967–971. http://doi.org/10.1126/science.1222077

"43 colorectal, prostate, ovarian, multiple myeloma, and glioblastoma cancer patients. 194 high-confidence somatic TE insertions (183 L1s, 10 Alus, and 1 ERV)"

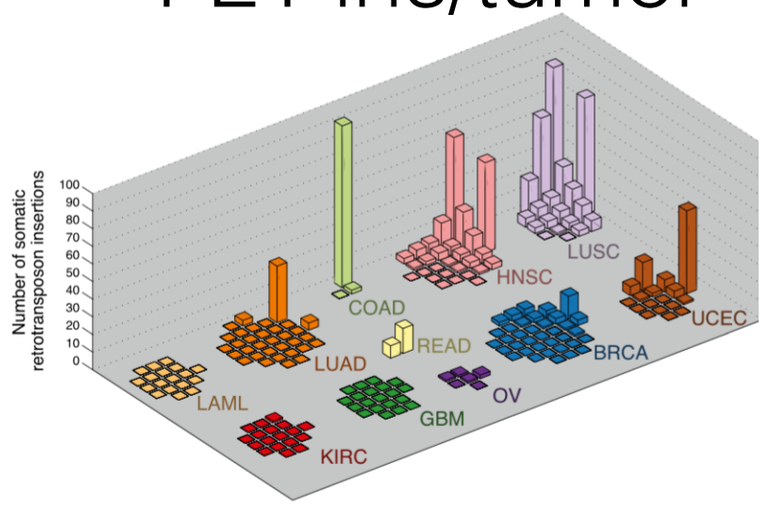
~4.3 L1 ins/tumor

~.23 ALU ins/tumor

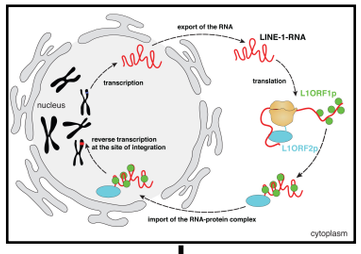
Helman, E., Lawrence, M. L., Stewart, C., Sougnez, C., Getz, G., & Meyerson, M. (2014). **Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing.** ***Genome Research***, *24*(7), 1053–1063. http://doi.org/10.1101/gr.163659.113

"200 tumor and matched normal samples […], 810 putative retrotransposon insertions […]"

~4 L1 ins/tumor



Consequences of L1 activity in tumors

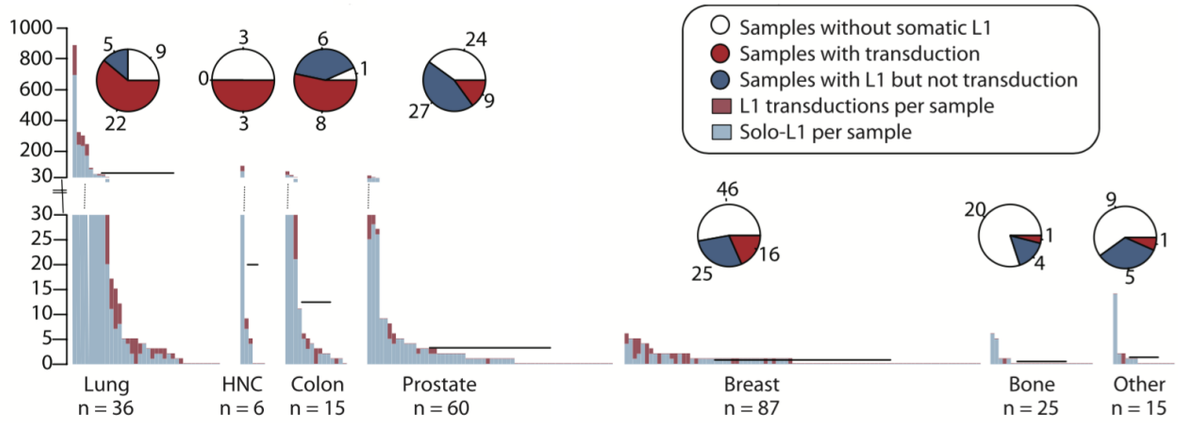


L1 SVA? ALU Coding transcripts

Tubio, J. M. C., Li, Y., Ju, Y. S., Martincorena, I., Cooke, S. L., Tojo, M., et al. (2014). **Mobile DNA in cancer.Extens ive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes.** ***Science*** *(New York, N.Y.)*, *345*(6196), 1251343. http://doi.org/10.1126/science.1251343

"Across the 290 samples, we identified 2756 somatic L1 retrotranspositions."

~9.5 L1 ins/tumor

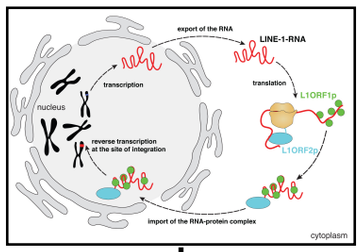


Ewing, A. D., Gacita, A., Wood, L. D., Ma, F., Xing, D., Kim, M.-S., et al. (2015). **Widespread somatic L1 retrotransposition occurs early during gastrointestinal cancer evolution.** ***Genome Research***, gr.196238.115. http://doi.org/10.1101/gr.196238.115

"104 somatic heterozygous L1Hs insertions were validated by PCR and Sanger sequencing in the 18 GI cancer patients"

~5.7 L1 ins/tumor

Consequences of L1 activity in tumors



L1

SVA?

ALU

Coding transcripts

Cooke, S. L., Shlien, A., Marshall, J., Pipinikas, C. P., Martincorena, I., Tubio, J. M. C., et al. (2014). **Processed pseudogenes acquired somatically during cancer development.** ***Nature Communications***, *5*, 3644. http://doi.org/10.1038/ncomms4644

"We identified 42 somatically acquired pseudogenes in 14 out of 629 primary cases".

.06 ins/tumor

What could contribute to retrotransposition activation?

- Overall hypomethylation of intergenic regions.
  - And, therefore, hypomethylation of L1 promoters.

(Tubio, J. M. C., et. al., 2014)

- miRNA deregulation - miR128

(Hamdorf, M., et. al., 2015)

- Somatic mutation disrupting retrotransposition regulation
  - ZNF 91/93

(Jacobs, F. M. J., et al., 2014)

- Integration and reverse transcriptase disruption
  - APOBEC3 mutation

(Marchetto, M., et al., 2013)

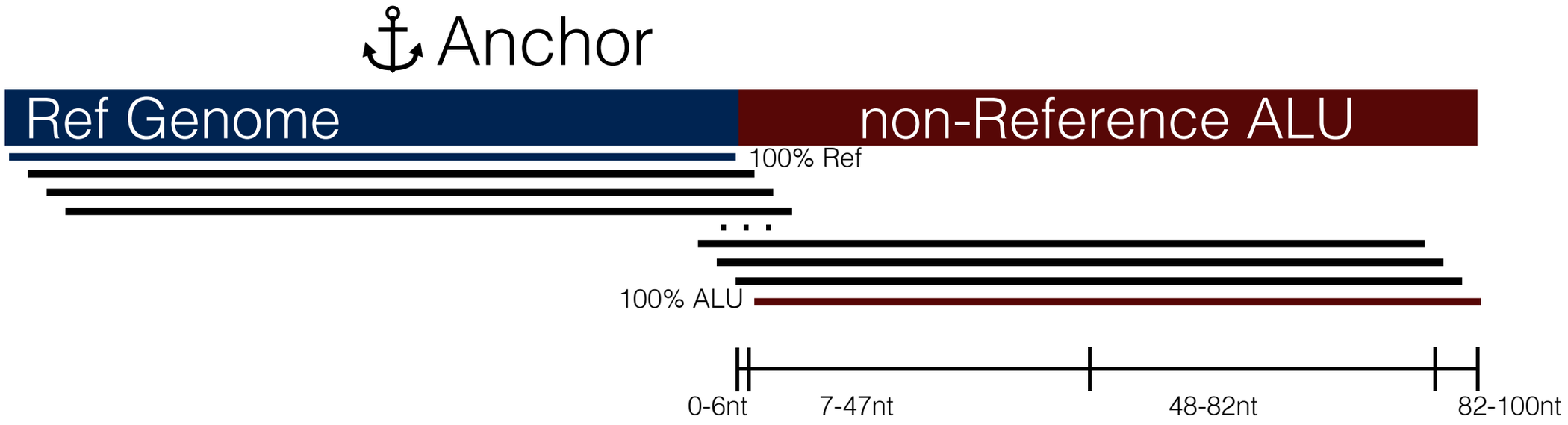
PCAWG Pilot 63 - WGS Dataset

- 55 Tumoral samples from CGHub;
- 53 Normal (paired) samples from CGHub;
    - Some “normal” samples are peripheral tissue, some are blood samples.
- 80x-100x real coverage paired-end reads.

Analysis constrains

- To date PCAWG is comprised of >2000 WGS experiments.
  - Inviable to download raw data (fastq).
  - Might be inviable to download unmapped reads.
  - Rely on mapped reads:
    - Paired-end discrepancy.
    - Partially aligned reads.

Partially Aligned Reads



0-6nt: Anchor only alignment

7-47nt: Soft clip (able to recover ALU sequence)

48-82nt: Hard clip (unable to recover ALU sequence)

82-100nt: No anchor alignment

Partially Aligned Reads

0-6nt: Anchor only alignment

```
0_forward   0     17  7579341  60  100M     *  0  0  AATGCAAGAAGCCCAGACGGAAACCGTAGCTGCCCTGGTAGGTTTTCTGGGAAGGGACAGAAGATGACAGGGGCCAGGAGGGGGCTGGTGCAGGGGCCGC  *  NM:i:0
                                                                                          MD:Z:100     AS
1_forward   0     17  7579341  60  100M     *  0  0  AATGCAAGAAGCCCAGACGGAAACCGTAGCTGCCCTGGTAGGTTTTCTGGGAAGGGACAGAAGATGACAGGGGCCAGGAGGGGGCTGGTGCAGGGGCCGG  *  NM:i:1
                                                                                          MD:Z:99C0    AS
2_forward   0     17  7579341  60  100M     *  0  0  AATGCAAGAAGCCCAGACGGAAACCGTAGCTGCCCTGGTAGGTTTTCTGGGAAGGGACAGAAGATGACAGGGGCCAGGAGGGGGCTGGTGCAGGGGCCGG  *  NM:i:1
                                                                                          MD:Z:99C0    AS
```

7-47nt: Soft clip (able to recover ALU sequence)

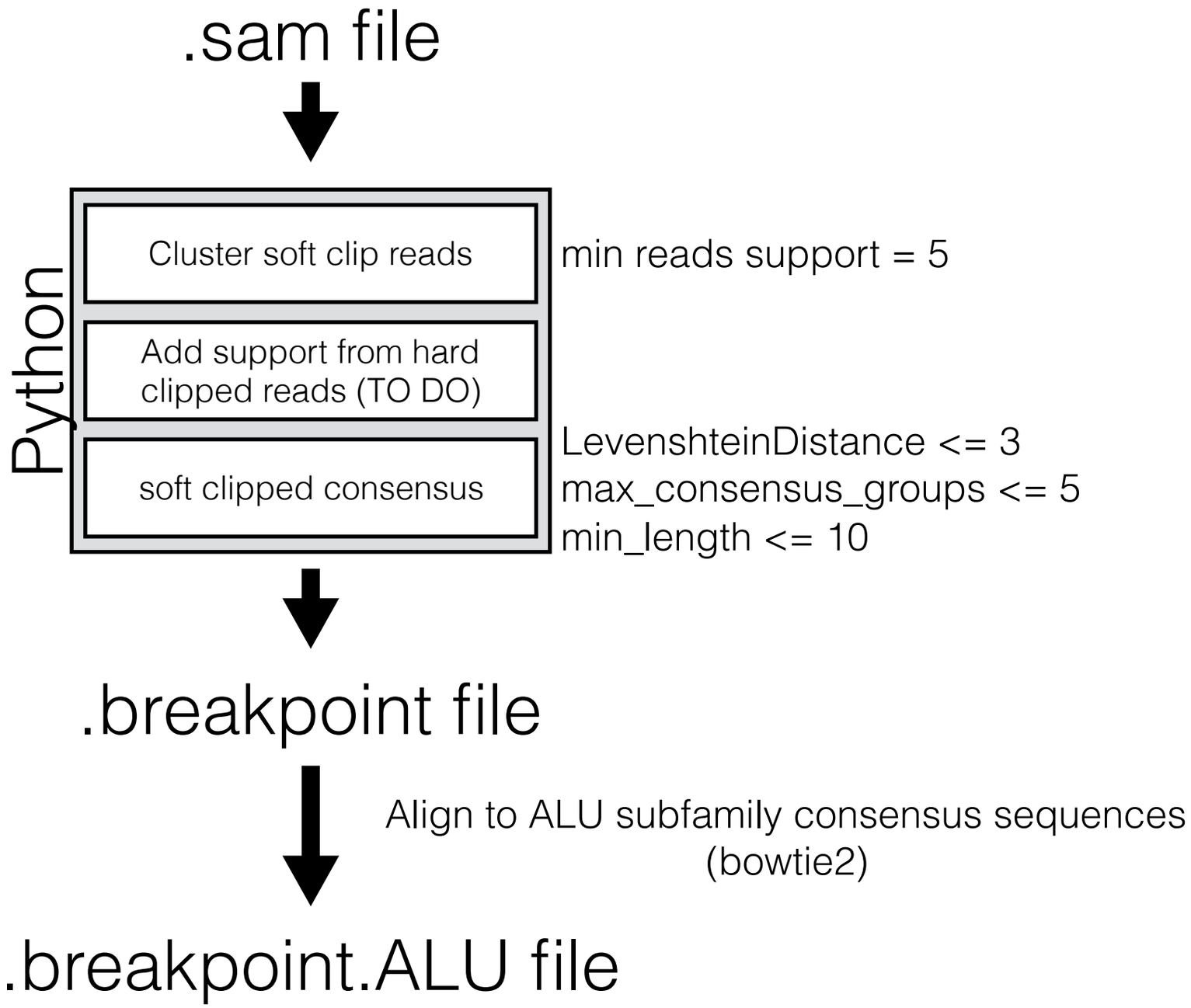
```
17_forward  0     17  7579341  60  84M16S   *  0  0  AATGCAAGAAGCCCAGACGGAAACCGTAGCTGCCCTGGTAGGTTTTCTGGGAAGGGACAGAAGATGACAGGGGCCAGGAGGGGGGCCGGGCGCGGTGGCT  *  NM:i:0
                                                                                          MD:Z:84      AS
18_forward  0     17  7579341  60  85M15S   *  0  0  AATGCAAGAAGCCCAGACGGAAACCGTAGCTGCCCTGGTAGGTTTTCTGGGAAGGGACAGAAGATGACAGGGGCCAGGAGGGGGCCGGGCGCGGTGGCTC  *  NM:i:0
                                                                                          MD:Z:85      AS
19_forward  0     17  7579341  60  83M17S   *  0  0  AATGCAAGAAGCCCAGACGGAAACCGTAGCTGCCCTGGTAGGTTTTCTGGGAAGGGACAGAAGATGACAGGGGCCAGGAGGGGCCGGGCGCGGTGGCTCA  *  NM:i:0
                                                                                          MD:Z:83      AS
```

48-82nt: Hard clip (unable to recover ALU sequence)

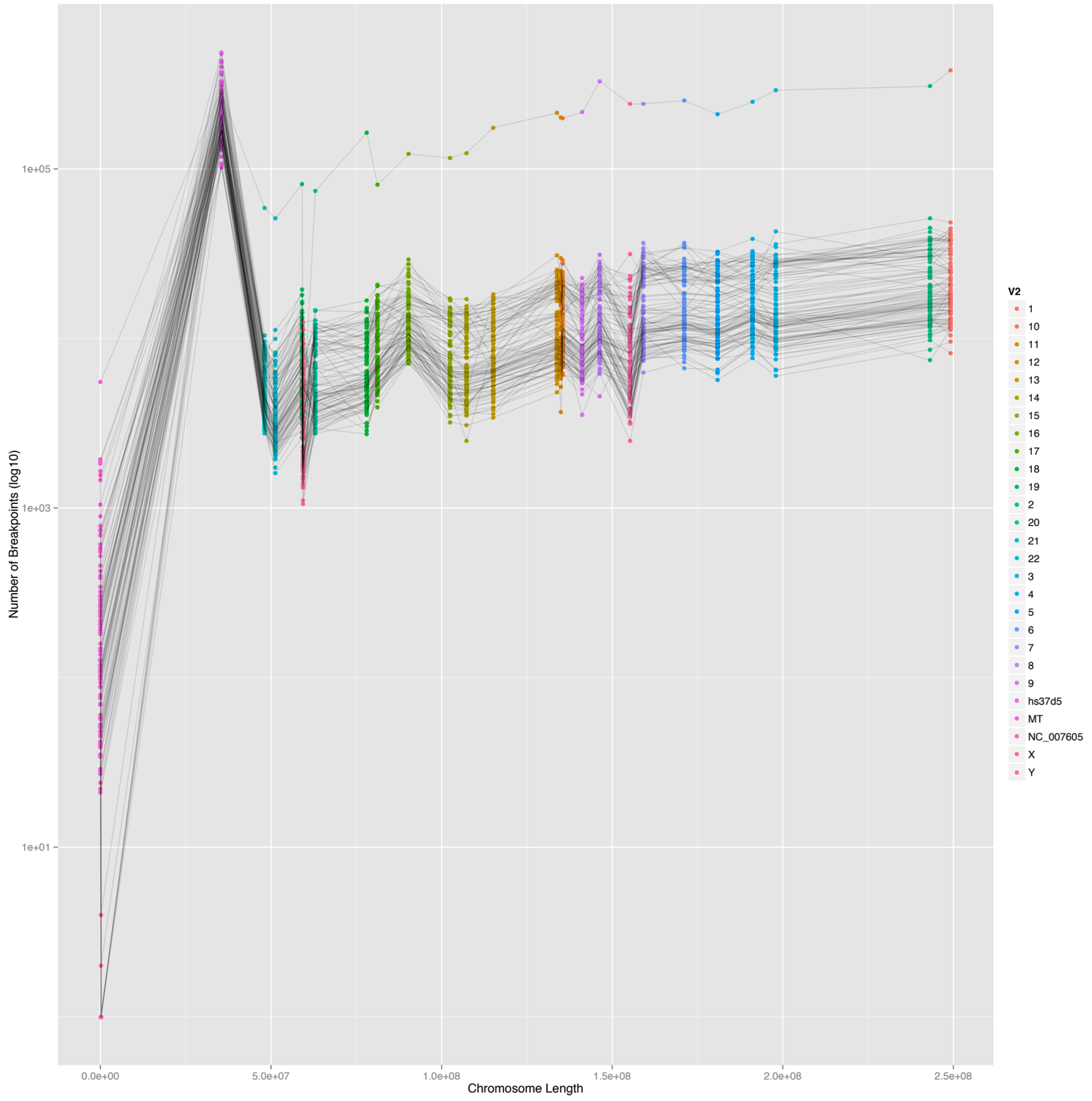
```
52_forward  2048  17  7579341  0   50M50H   *  0  0  AATGCAAGAAGCCCAGACGGAAACCGTAGCTGCCCTGGTAGGTTTTCTGG                                                          *  NM:i:0
                                                                                          MD:Z:50      AS
53_forward  2048  17  7579341  0   47M53H   *  0  0  AATGCAAGAAGCCCAGACGGAAACCGTAGCTGCCCTGGTAGGTTTTC                                                           *  NM:i:0
                                                                                          MD:Z:47      AS
54_forward  2048  17  7579341  0   46M54H   *  0  0  AATGCAAGAAGCCCAGACGGAAACCGTAGCTGCCCTGGTAGGTTTT                                                            *  NM:i:0
                                                                                          MD:Z:46      AS
```

82-100nt: No anchor alignment

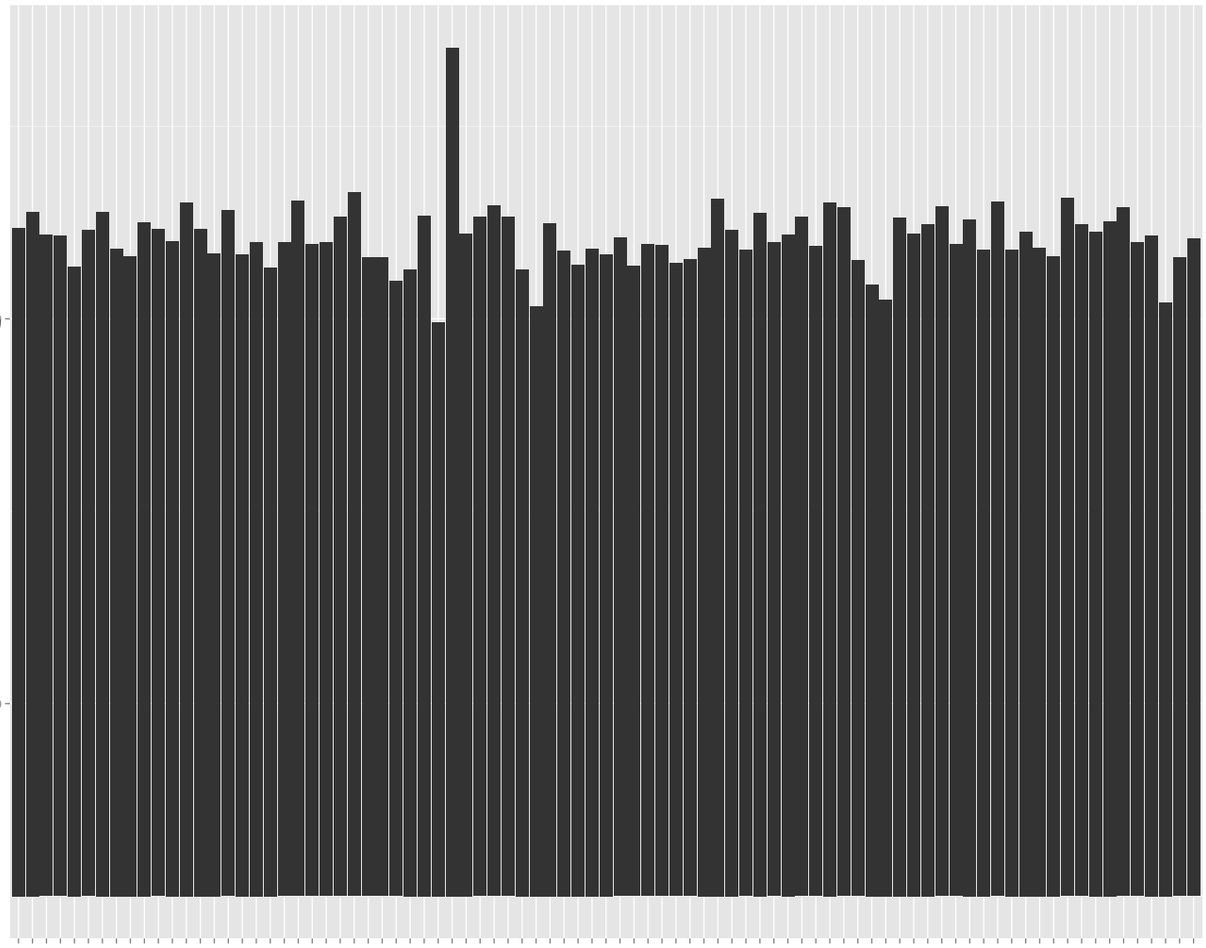
Overall workflow

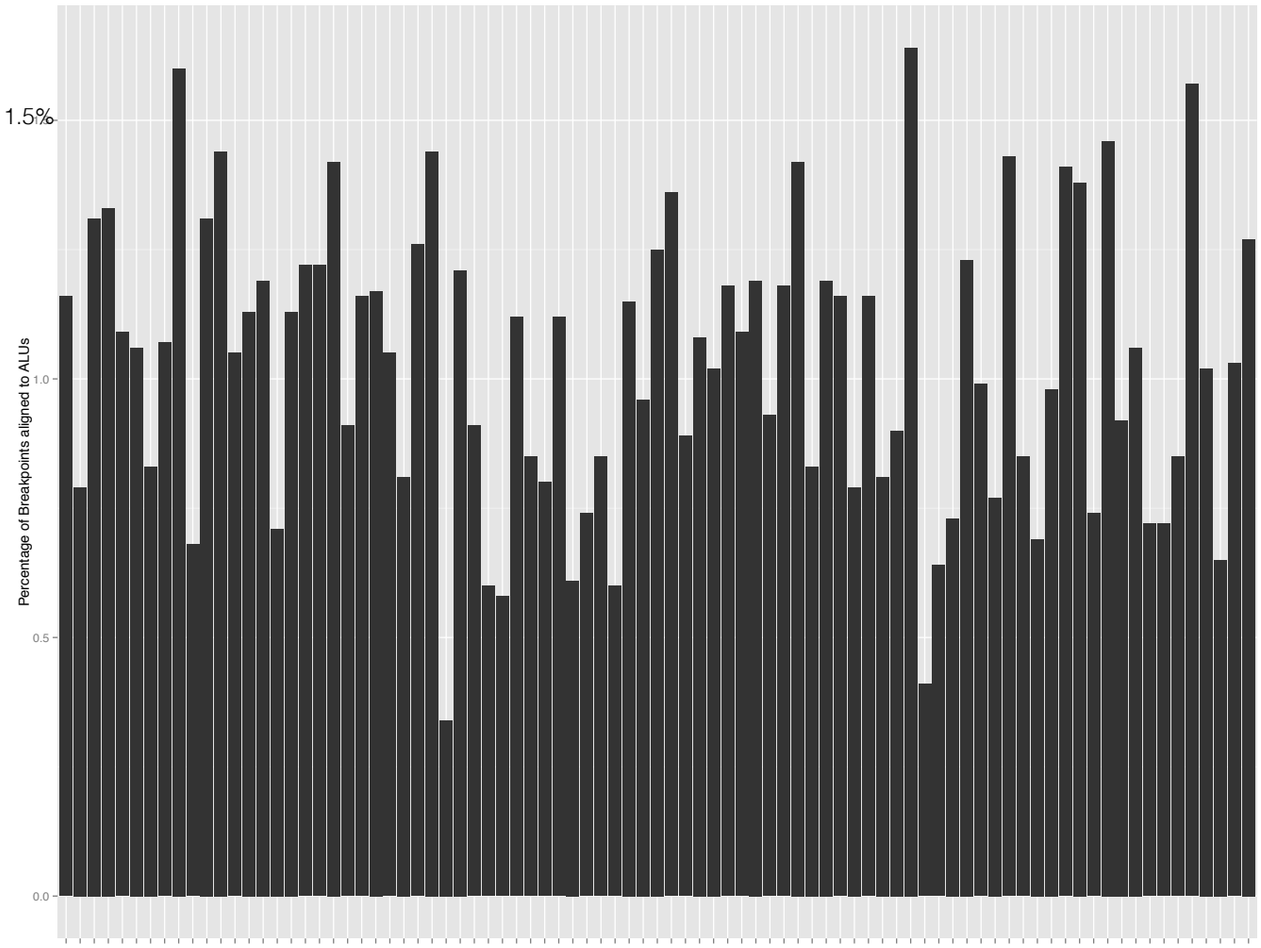


Sanity check: Breakpoints per chromosome

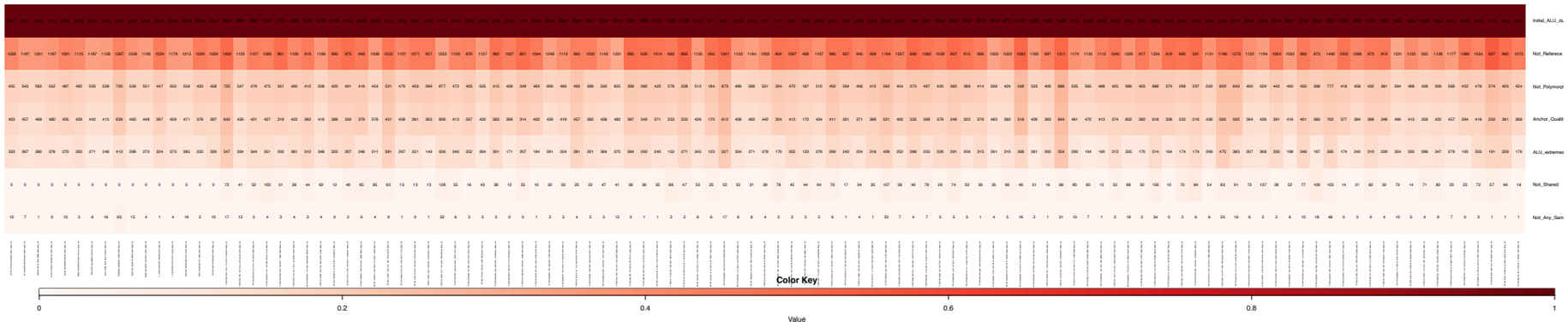


Sanity check: Breakpoints with ALU sequences





Filters: Selecting new ALU insertions



Total breakpoints aligned to ALUs

(rm) overlapping ALUs in ref

(rm) overlapping polymorphic ins.

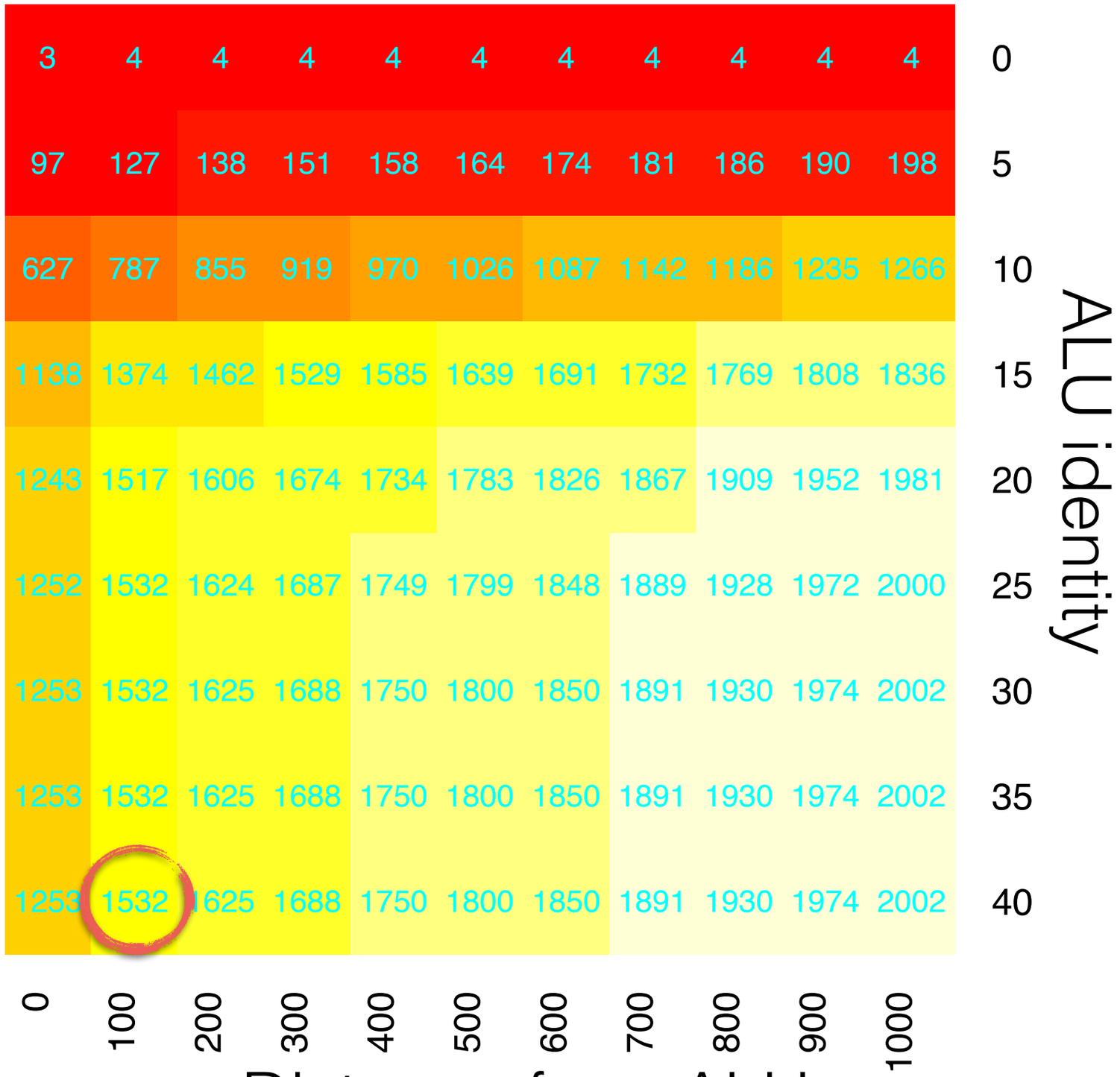
(rm) with dubious anchor

insertions at ALU extremities

(rm) shared with other samples

2532	1170	1984	2451	ALU call
1054	627	960	1073	Not in reference genome
479	274	425	454	Not in polymorphic DB
416	252	381	388	Good anchor quality
355	191	329	176	ALU extremities
72	57	46	18	Not in paired sample
3	1	1	1	Not in any analyzed sample

(rm) Breakpoints overlapping with ALUs in the reference genome

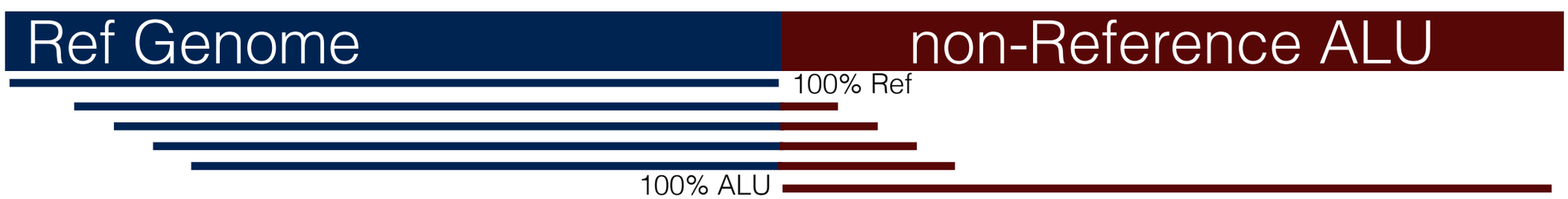


(rm) Breakpoints overlapping polymorphic insertions

- Two polymorphic repetitive element databases
  - 1000 Genomes phase3 (Eugene Gardner)
    12,786 polymorphic ALU insertions
    3,048 polymorphic L1 insertions

- dbRIP
  2,086 polymorphic ALU insertions

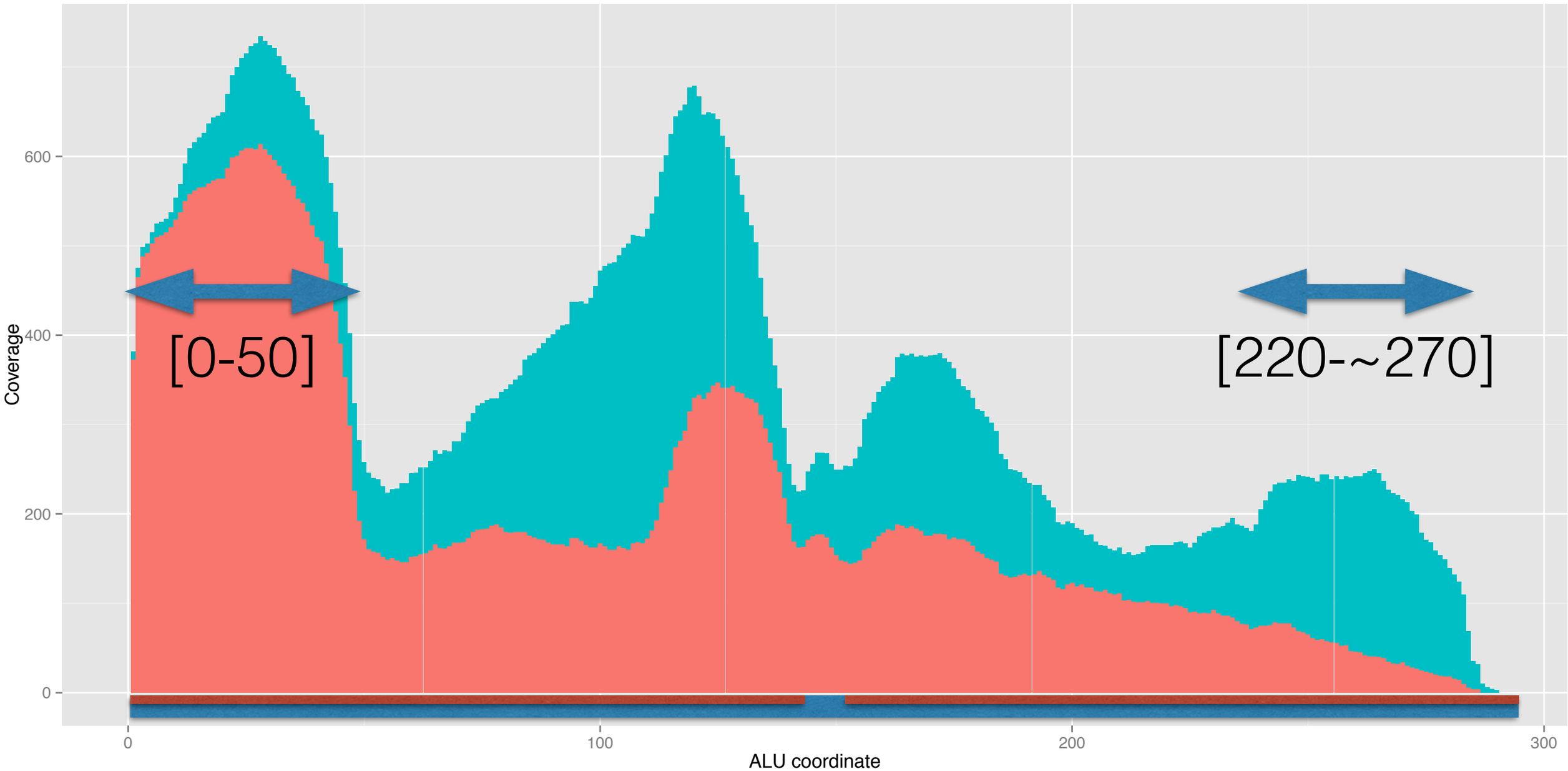
(rm) Dubious anchor alignments



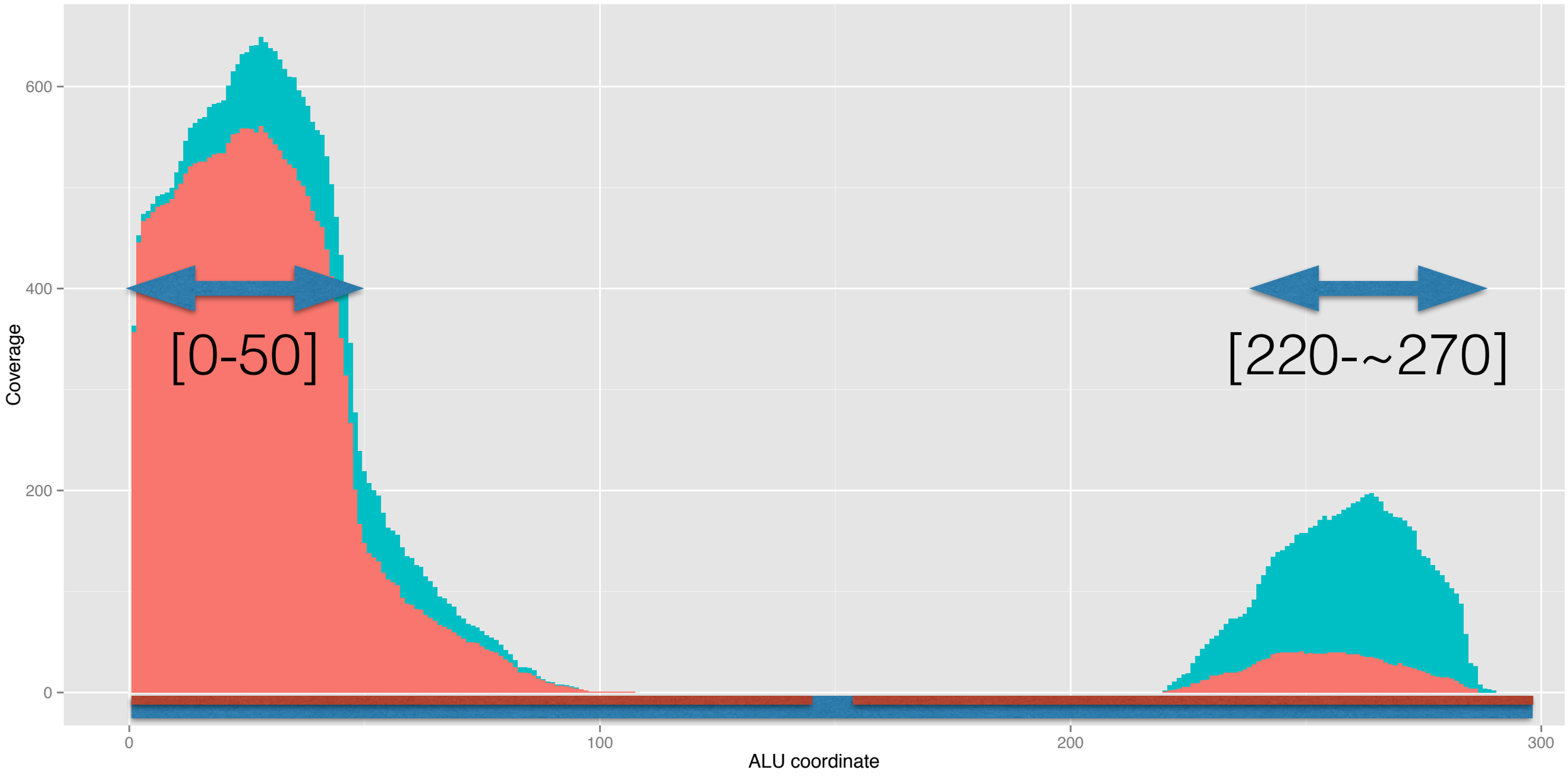
- At least two reference genome segments aligned to the reference genome with alignment quality higher than 20.

- Remove reference genome segments with low sequencing quality (fastQC)

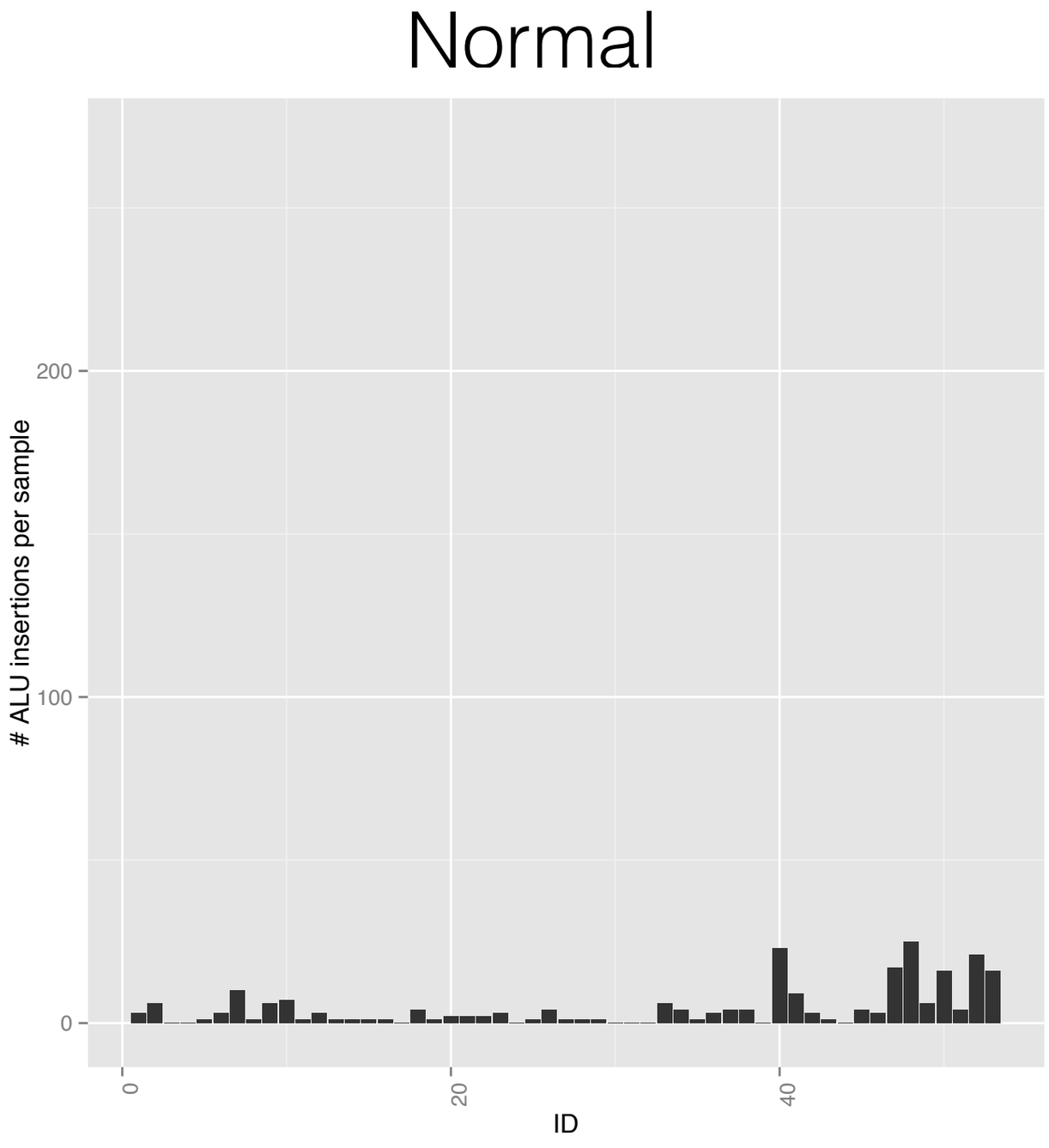
(rm) ALU consensus coverage (all samples)

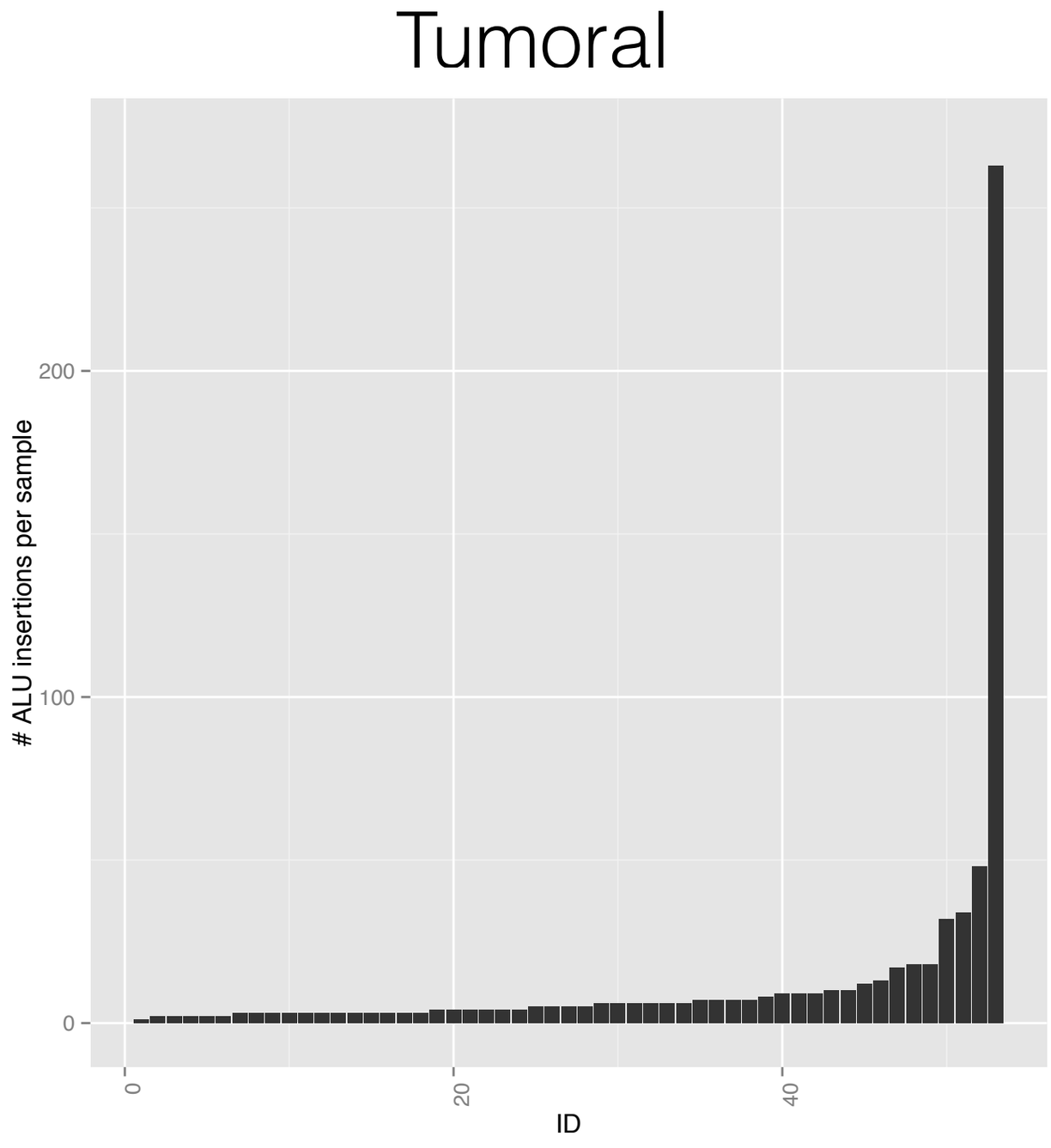


(rm) ALU consensus coverage (all samples)

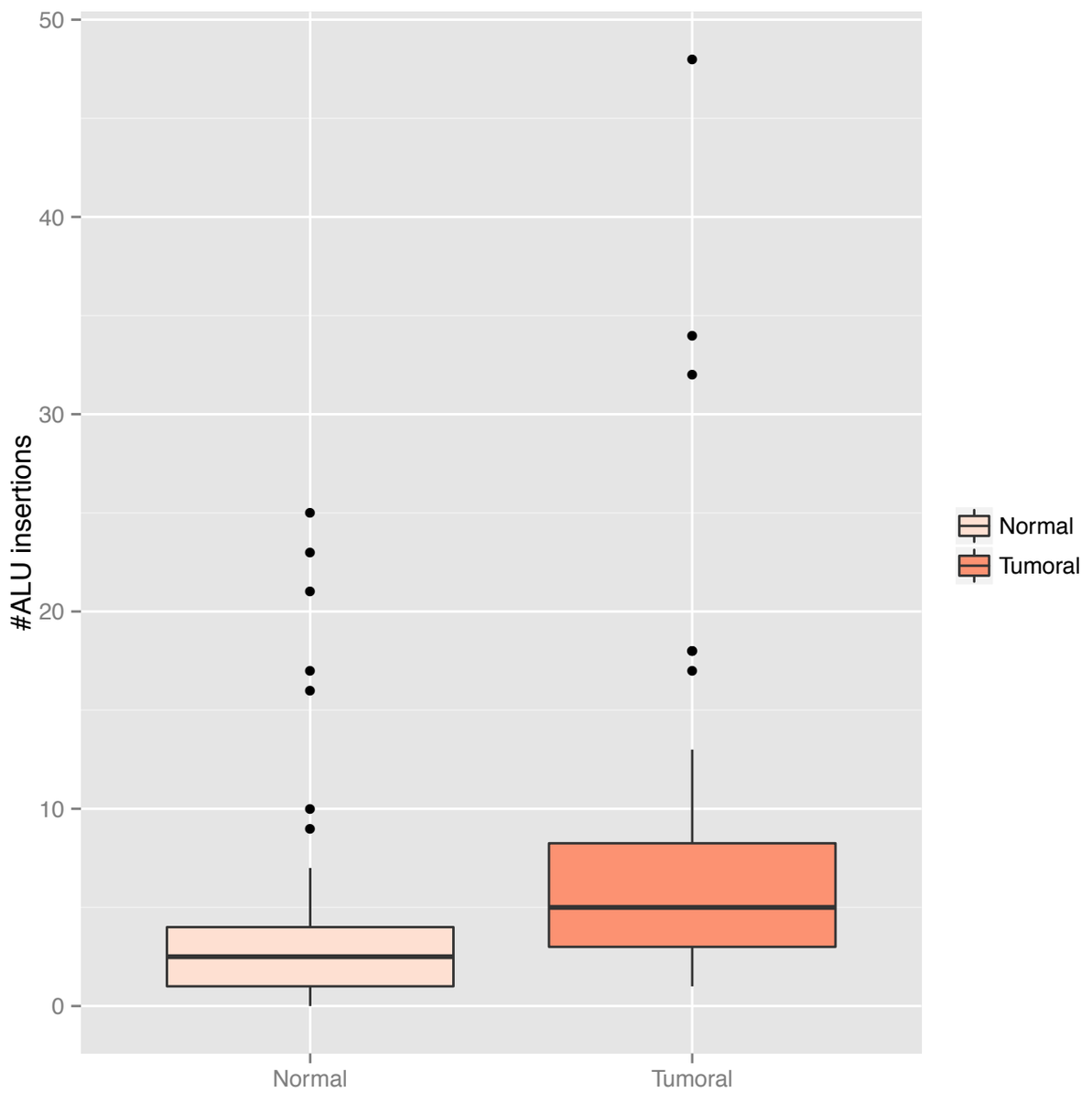


Number of ALU somatic insertions per sample

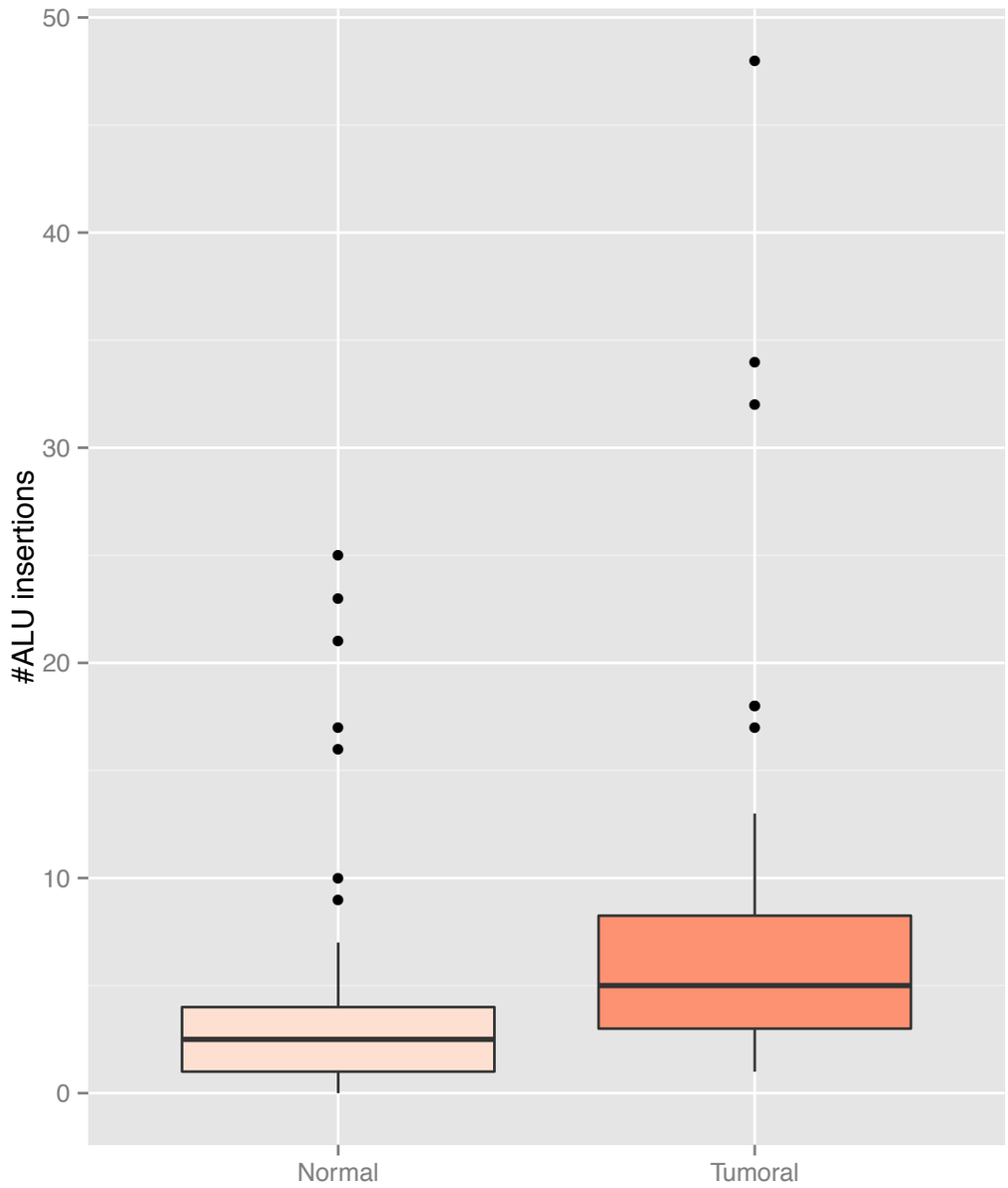




Number of ALU somatic insertions



Number of ALU somatic insertions



~12.7 ins/tumor

or

~8 ins/tumor

ALU insertion point annotation

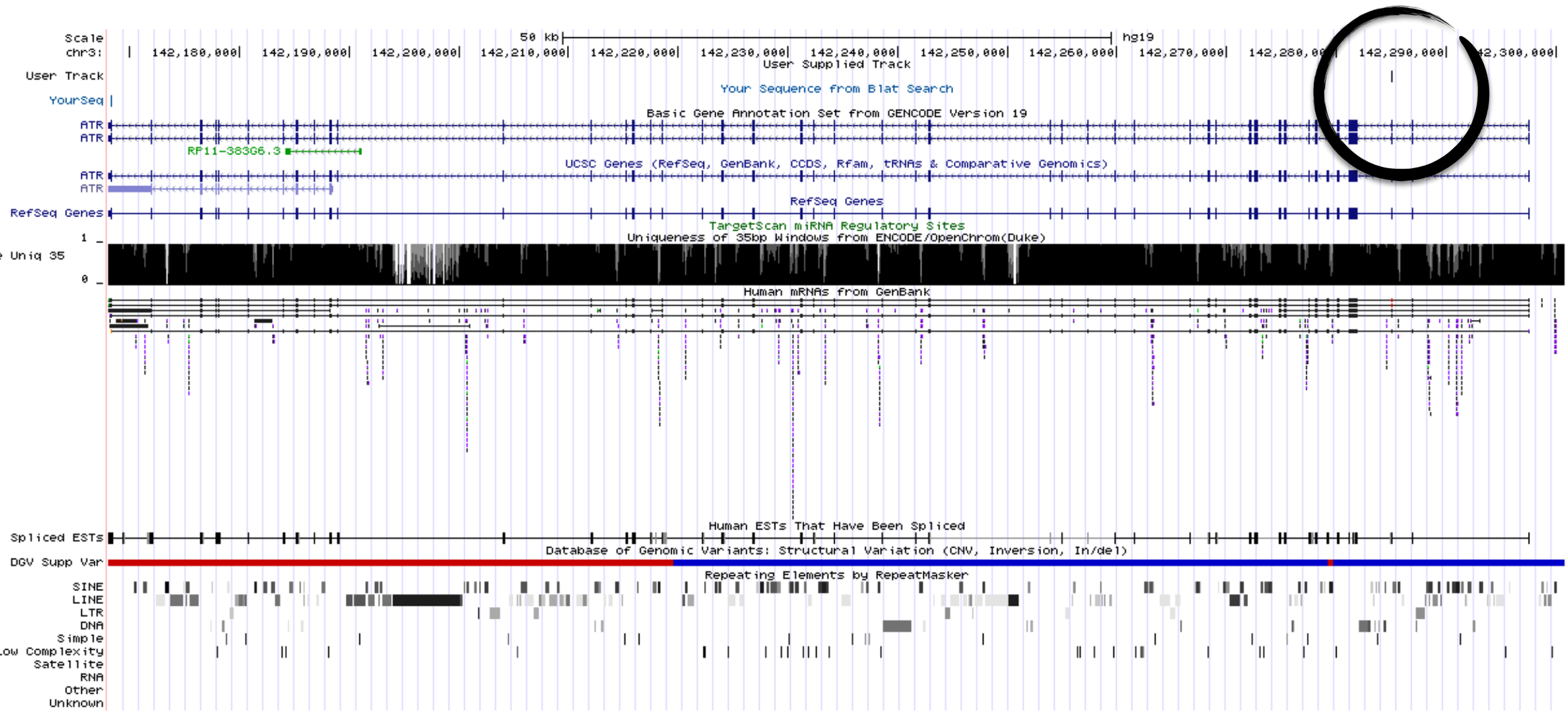
	Total	Intragenic	Exonic	Coding
Observed	699	386	23	9
Expected (gencode)	-	364 (50.49%)	27 (3.923%)	7 (1.132%)
p-value		0.0958	0.4324	0.4474

ALU insertion point annotation

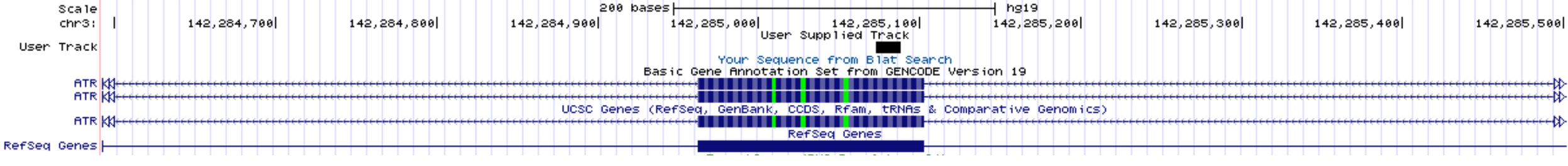
	Total	Intragenic	Exonic	Coding
Observed	699	386		
Expected (gencode)	-	364 (50.49%)		
p-value		0.0958		

ACCS protein_coding
APEX2 protein_coding
ATR protein_coding
F11 protein_coding
HEATR1 protein_coding
KIAA0430 protein_coding
SPI1 protein_coding
USP19 protein_coding
ZBTB40 protein_coding

ALU insertion: ATR Gene



ALU insertion: ATR Gene



Inhibition of ATR expression in cultured cancer cells has been demonstrated to increase sensitivity to chemotherapeutic drugs, including the DNA-crosslinking agent cisplatin.

The insertion point doesn't overlap any of the three ATR domains: (FAT, PI3_4_KINASE_3, FATC) but the insertion could create a frameshift of premature stop codon (insertion in the third exon).

Unfortunately, there is no information on patient chemotherapy. Cisplatin is a common treatment for breast cancer and might have been used.

Sangster-Guity, N., Conrad, B. H., Papadopoulos, N., & Bunz, F. (2011). ATR mediates cisplatin resistance in a p53 genotype-specific manner. *Oncogene*, *30*(22), 2526–2533. http://doi.org/10.1038/onc.2010.624

Hurley, P. J., Wilsker, D., & Bunz, F. (2007). Human cancer cells require ATR for cell cycle progression following exposure to ionizing radiation. *Oncogene*, *26*(18), 2535–2542. http://doi.org/10.1038/sj.onc.1210049

ALU insertion point annotation

	Total	Enhancer	TFP	DHS
Observed	699	80	96	109
sim. median (sd)	-	67 (7.72)	86 (8.82)	100 (8.82)
p-value		0.12	0.26	0.43

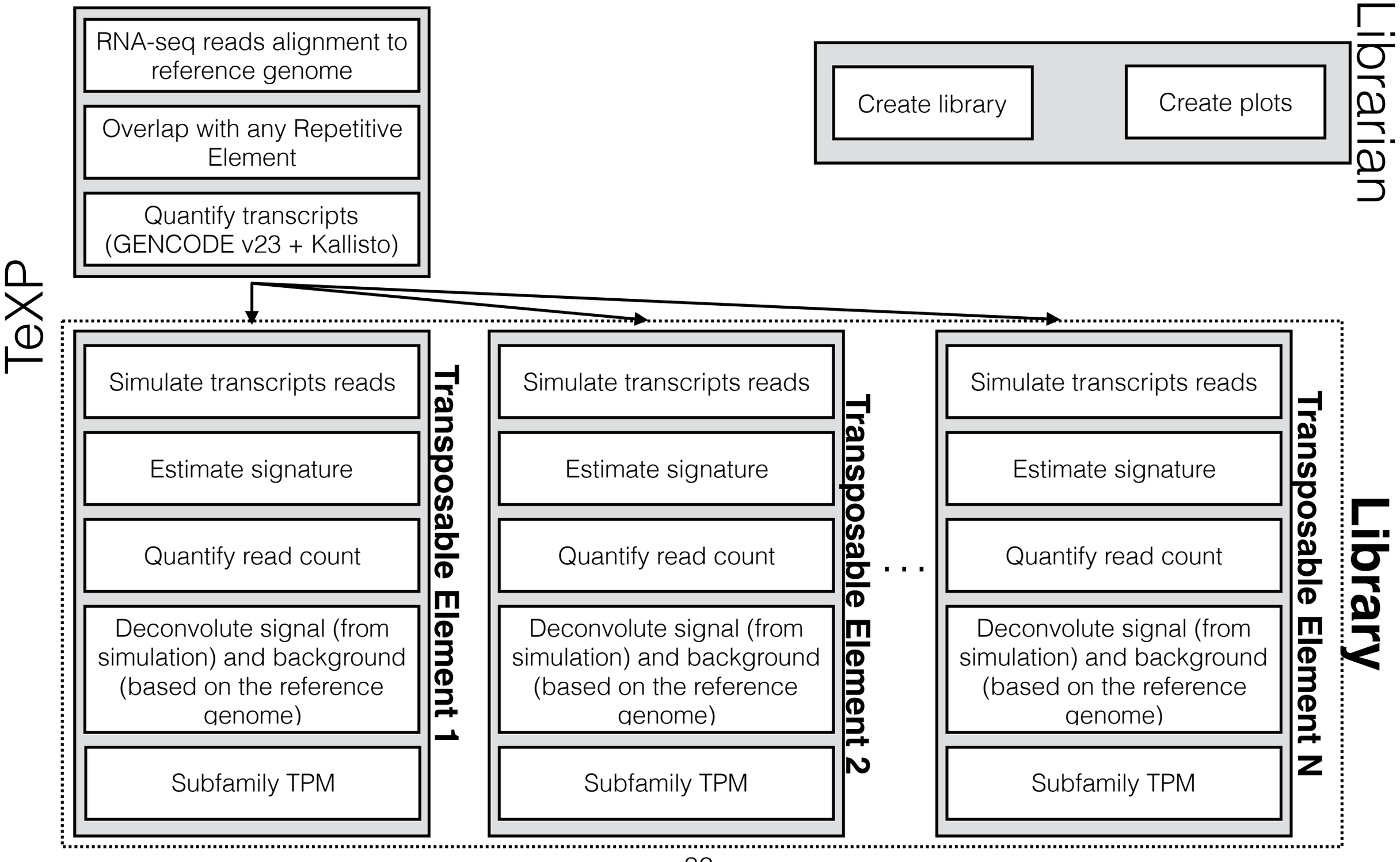
Transcriptome data

41 donors (matched with WGS data)

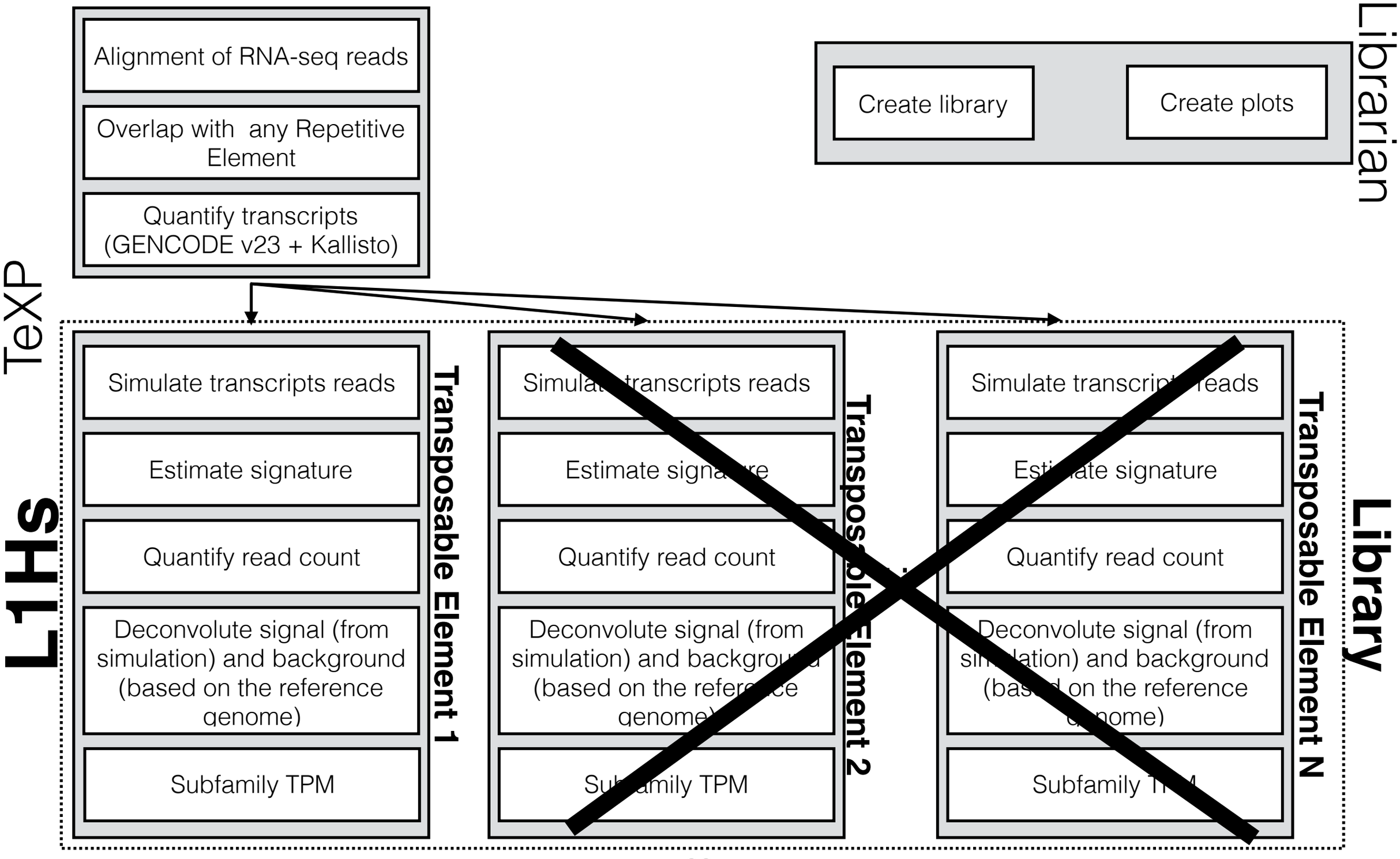
40 paired-end and 1 single end RNA-seq
- Most samples are poly(A) enriched (oligo dT primers)
- KIRP sample is the only total RNA sample

Reads outside annotated regions
- Average 91.8% reads align to annotated regions.
- Average 5M reads mapped to non annotated regions.
- Average 850k reads mapped to Repetitive Elements.

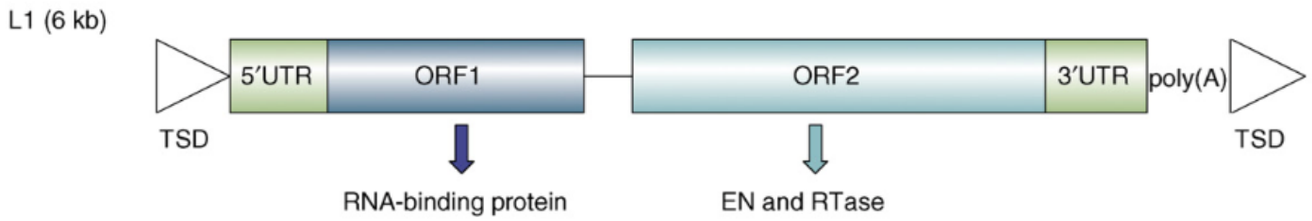
TExp


RNA-seq reads alignment to reference genome
Overlap with any Repetitive Element
Quantify transcripts (GENCODE v23 + Kallisto)
Librarian
Create library
Create plots
TExP
Library
Transposable Element 1
Simulate transcripts reads
Estimate signature
Quantify read count
Deconvolute signal (from simulation) and background (based on the reference genome)
Subfamily TPM
Transposable Element 2
Simulate transcripts reads
Estimate signature
Quantify read count
Deconvolute signal (from simulation) and background (based on the reference genome)
Subfamily TPM
. . .
Transposable Element N
Simulate transcripts reads
Estimate signature
Quantify read count
Deconvolute signal (from simulation) and background (based on the reference genome)
Subfamily TPM

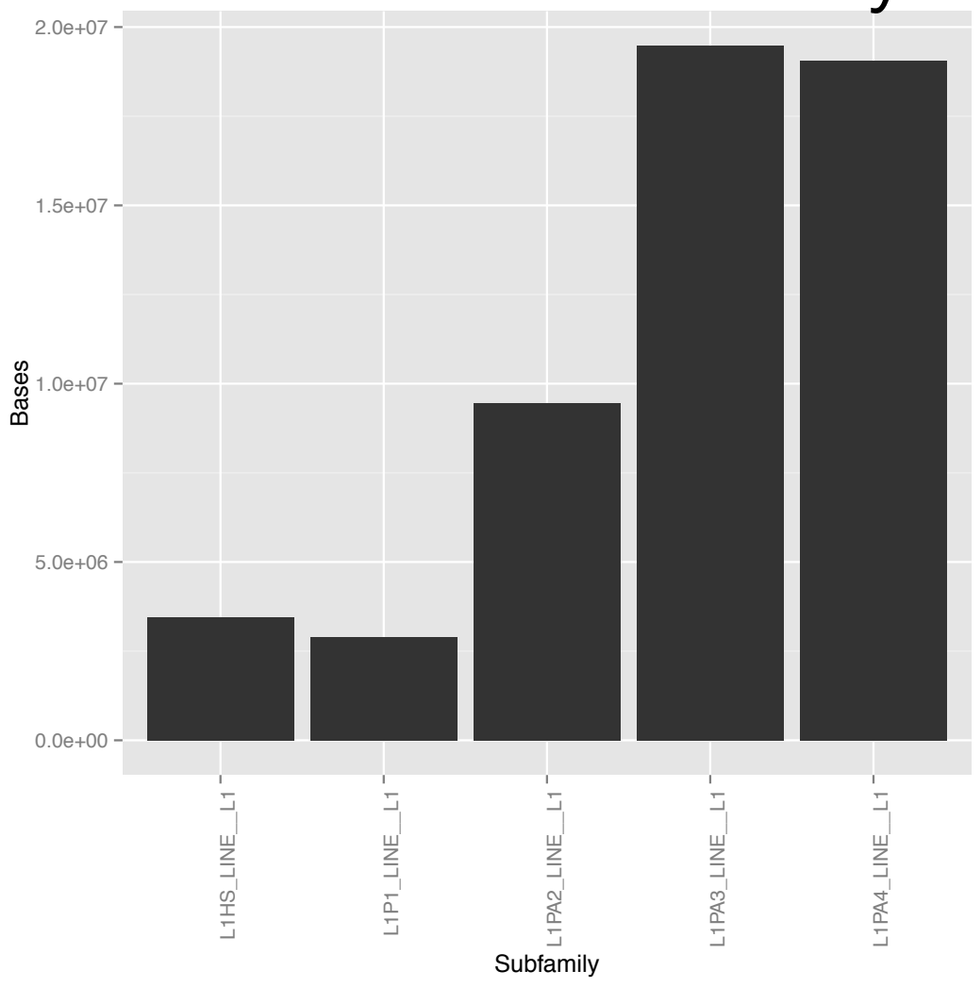
TExp


Alignment of RNA-seq reads
Overlap with any Repetitive Element
Quantify transcripts (GENCODE v23 + Kallisto)
Librarian
Create library
Create plots
TeXP
L1HS
Transposable Element 1
Simulate transcripts reads
Estimate signature
Quantify read count
Deconvolute signal (from simulation) and background (based on the reference genome)
Subfamily TPM
Transposable Element 2
Transposable Element N
Library

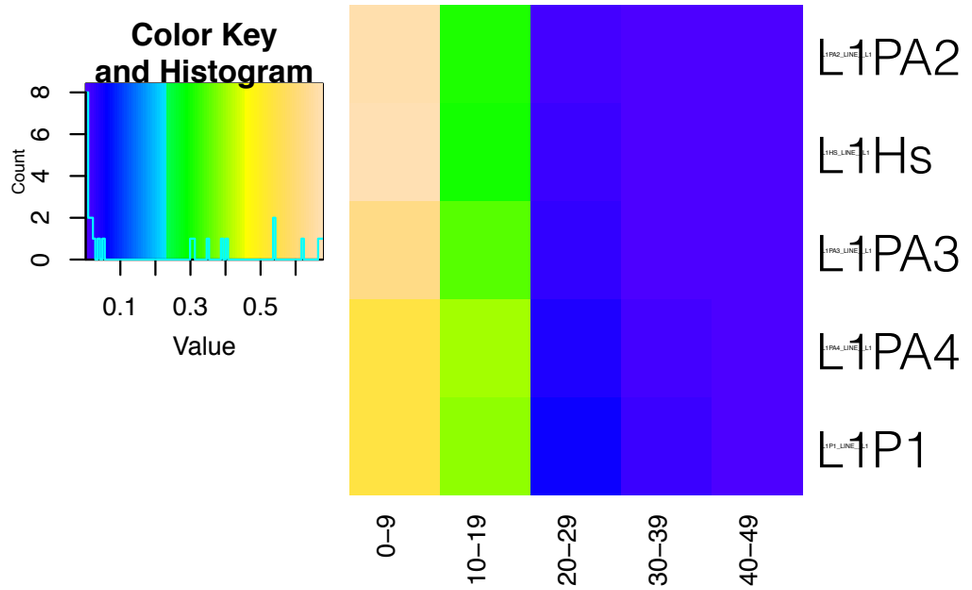
TExp: L1Hs



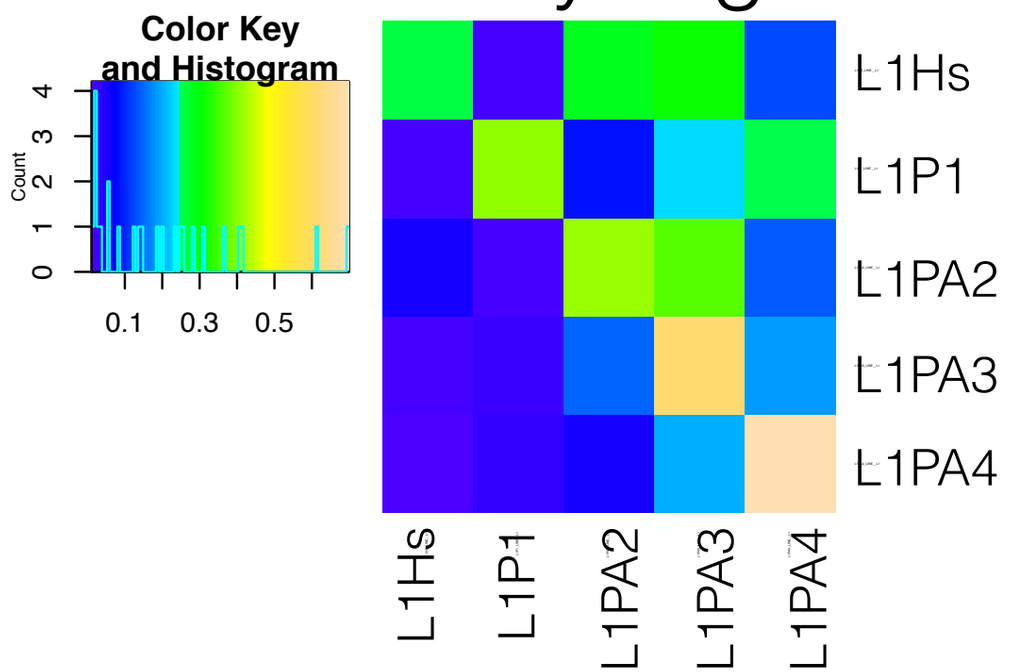
Number of bases of each subfamily



Alignment Quality



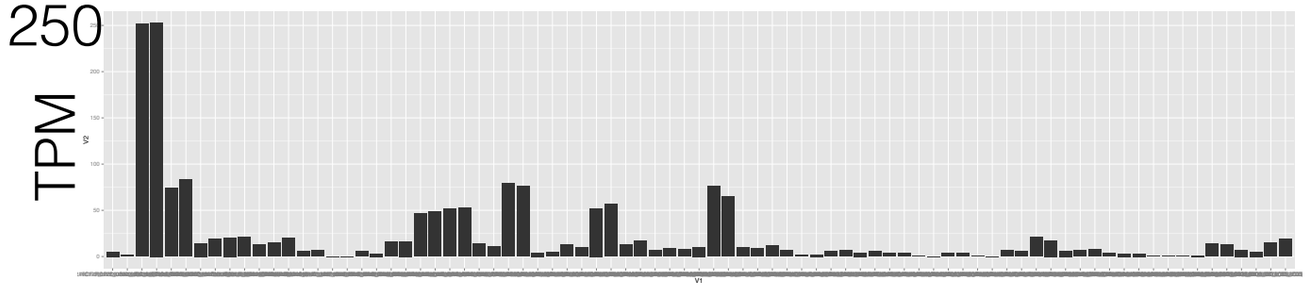
Cross subfamily alignment



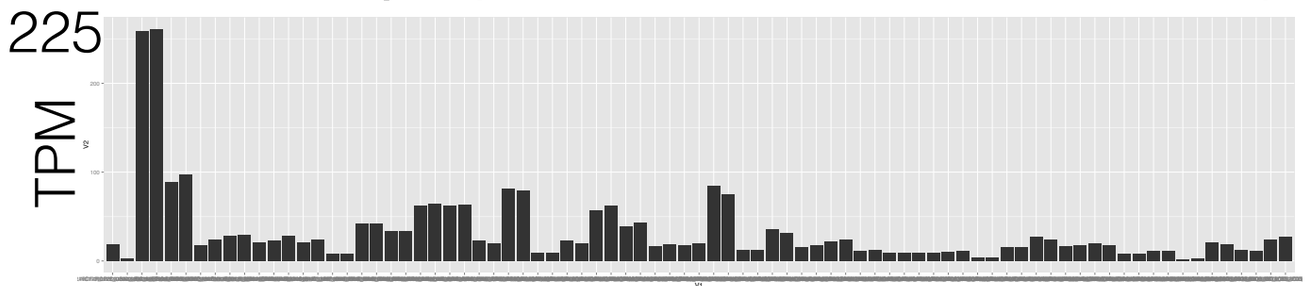
TExp: L1Hs transcription in cancer samples

Deconvolute transcription signal (from simulation) from background transcription (based on the reference genome)

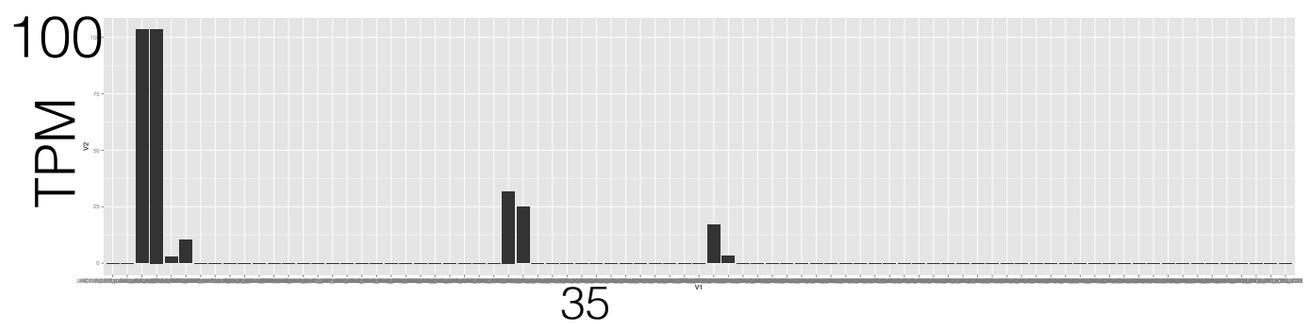
→ Least Squares with Equalities and Inequalities (lsei)



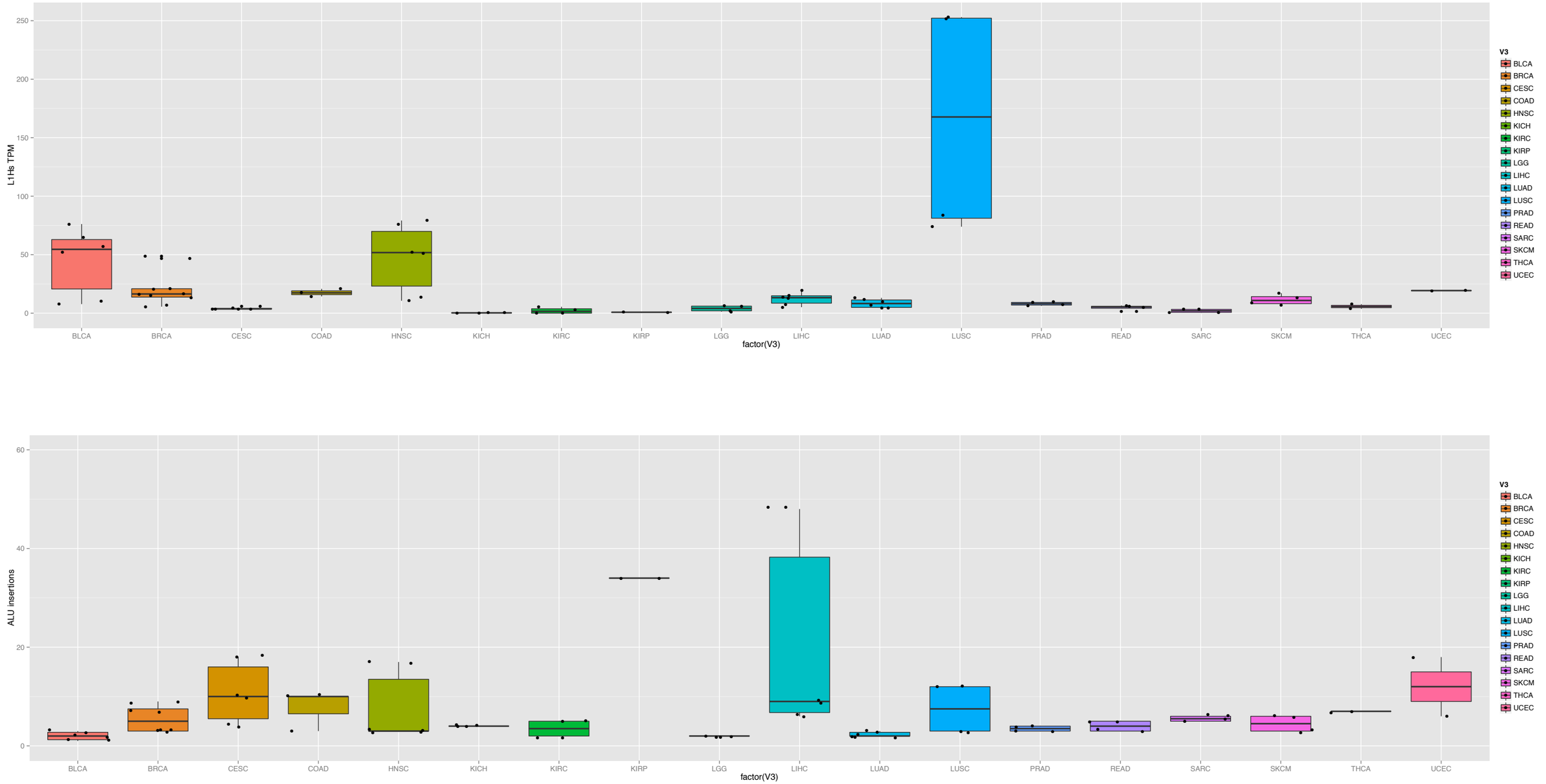
→ Mixed Membership (mixedMem - Erosheva et al (2004))



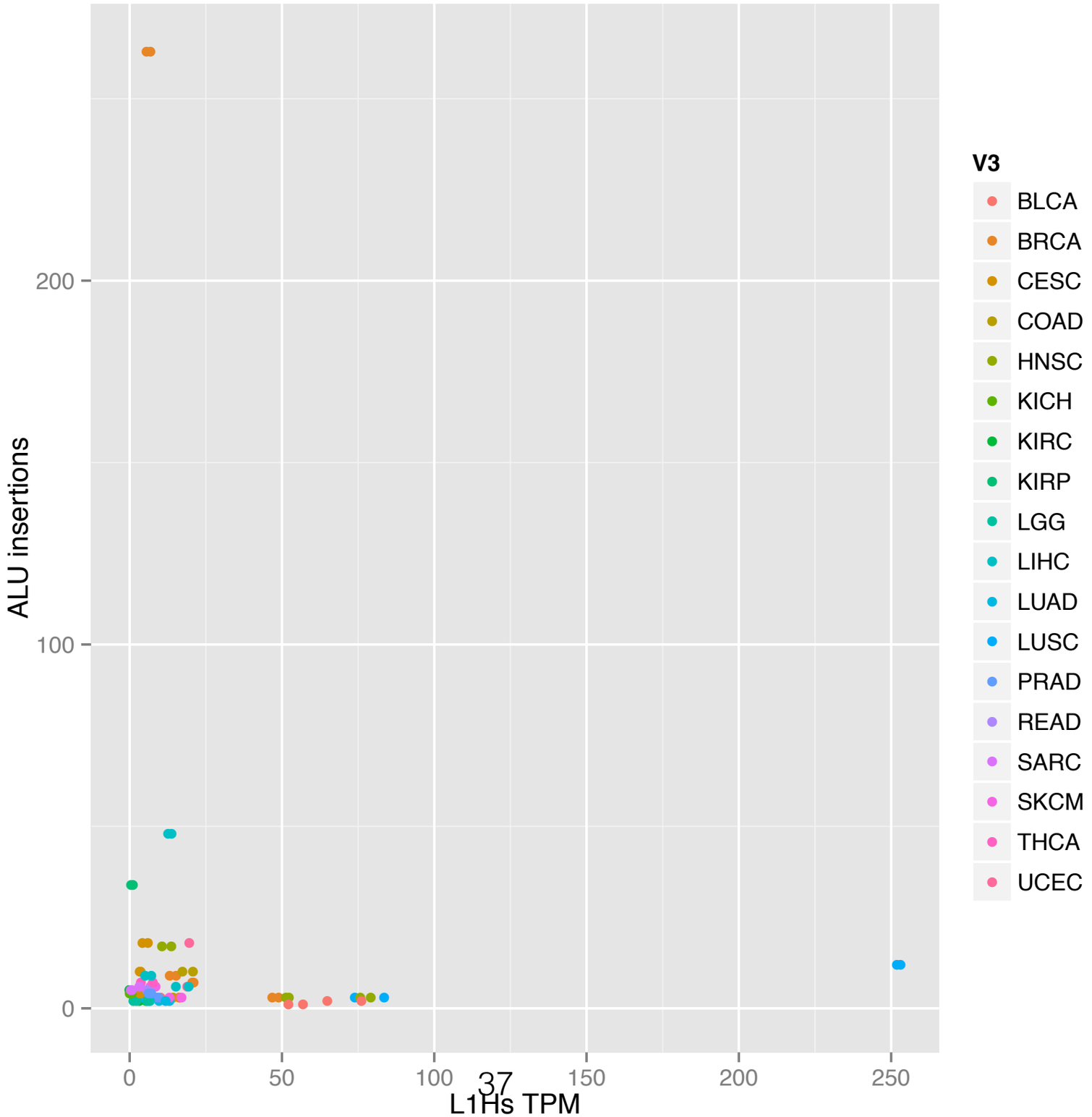
→ LASSO (glmnet)



LSEI regression for L1Hs transcripts



Transcription vs Retrotranspostion



Future steps

- Include hard clipped reads as breakpoint support.
- Include paired-end information, focusing on multi-mapping reads.
- Add samples at annai (complete the 63 pilot samples).
- Correlate insertion to other genomic features.
- Implement a step forward regression to estimate the number of reads from L1Hs transcripts.

Acknowledgments

Genome Tech

Mark Gerstein

Questions?