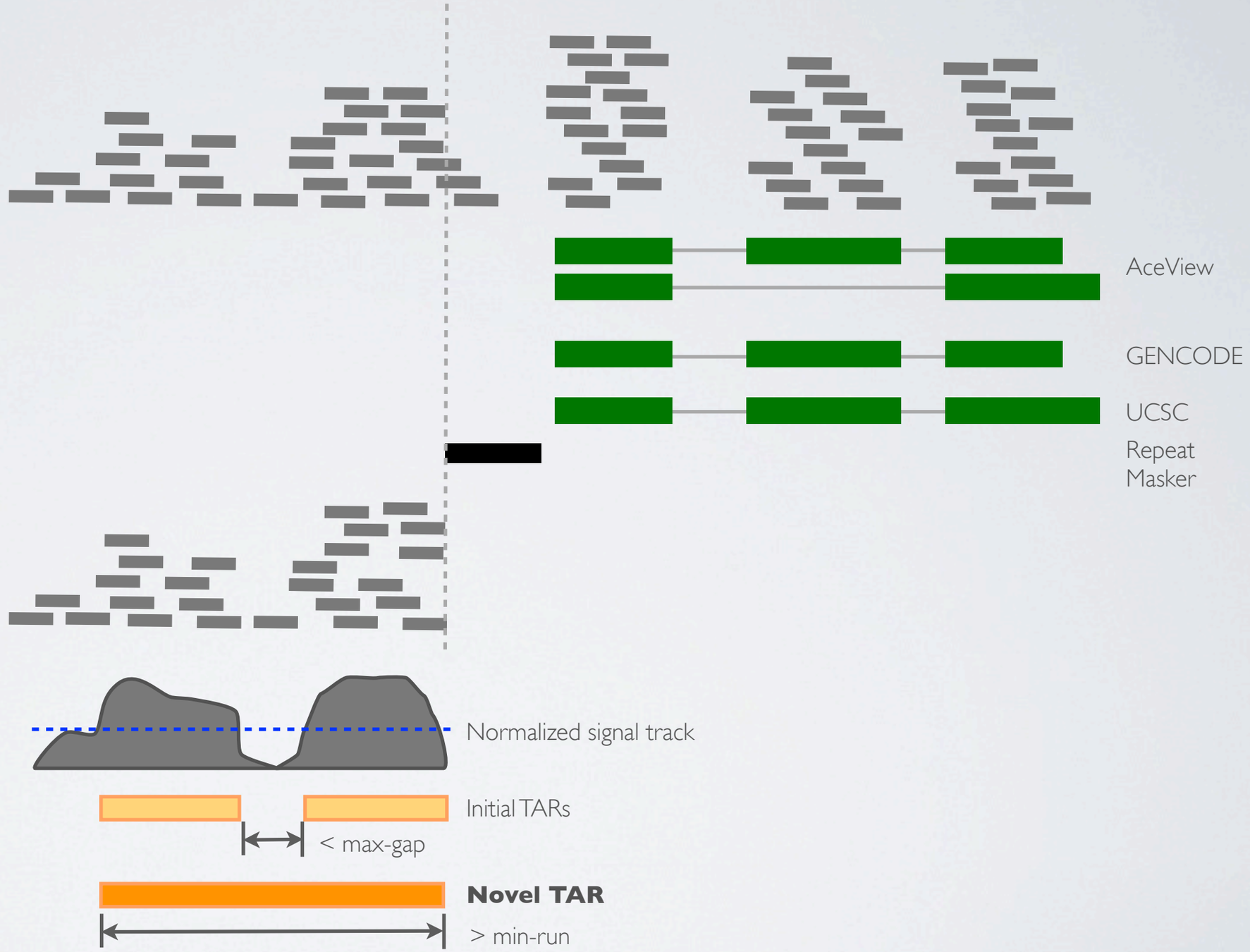
# Novel Transcriptionally Active Regions (TARs)

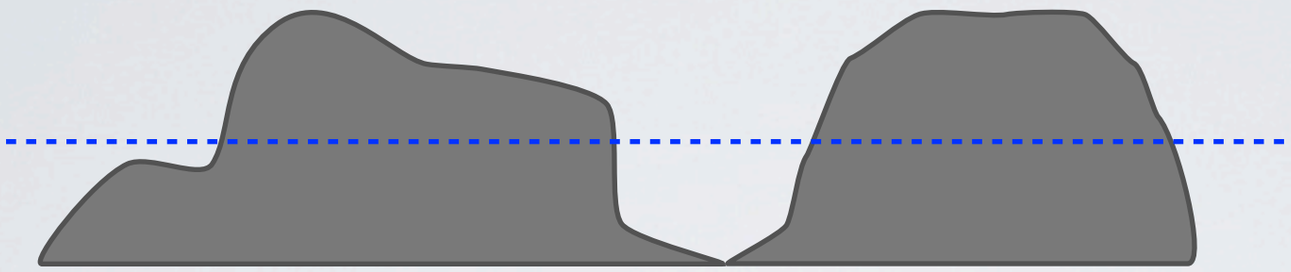
Andrea Sboner - 2011.04.15

# Discovery of Novel TARs

- Reads overlapping gene annotation set -- AceView, GENCODE, UCSC, and repetitive regions are excluded
- Reads from all samples pooled together
- Max-gap, min-run algorithm to identify TARs



# MIN-GAP, MAX-RUN ALGORITHM: THRESHOLD DEFINITION



Normalized signal track (per million mapped nucleotides)

$$r(g) = \frac{n(g)}{l(g) * M}$$

$$x(g) = log_2(r(g) + 1)$$

- From exon expression values:

$\tilde{x}(g)$

- Determine the median value of each exons:

$$x_c = quantile(\vec{\tilde{x}}, 0.05)$$
$$\forall \tilde{x}(g) > 0$$

- Compute the 5th percentile of expressed exons:

$$r_c = 2^{x_c} - 1$$

- Compute the corresponding RPKM value:

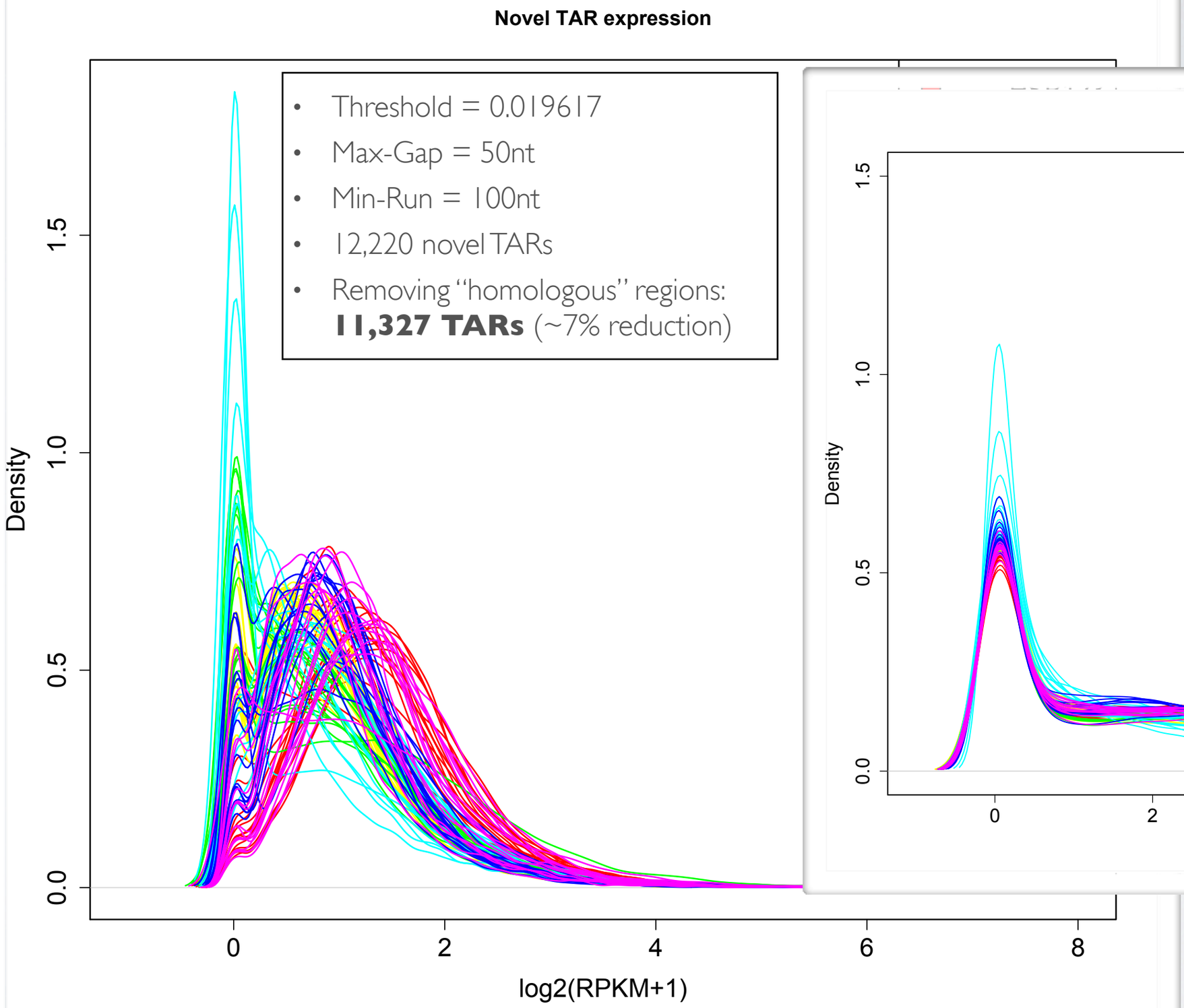
$$|\vec{L}| = 0.129Kb$$

- Consider the median length of GENCODE exons:

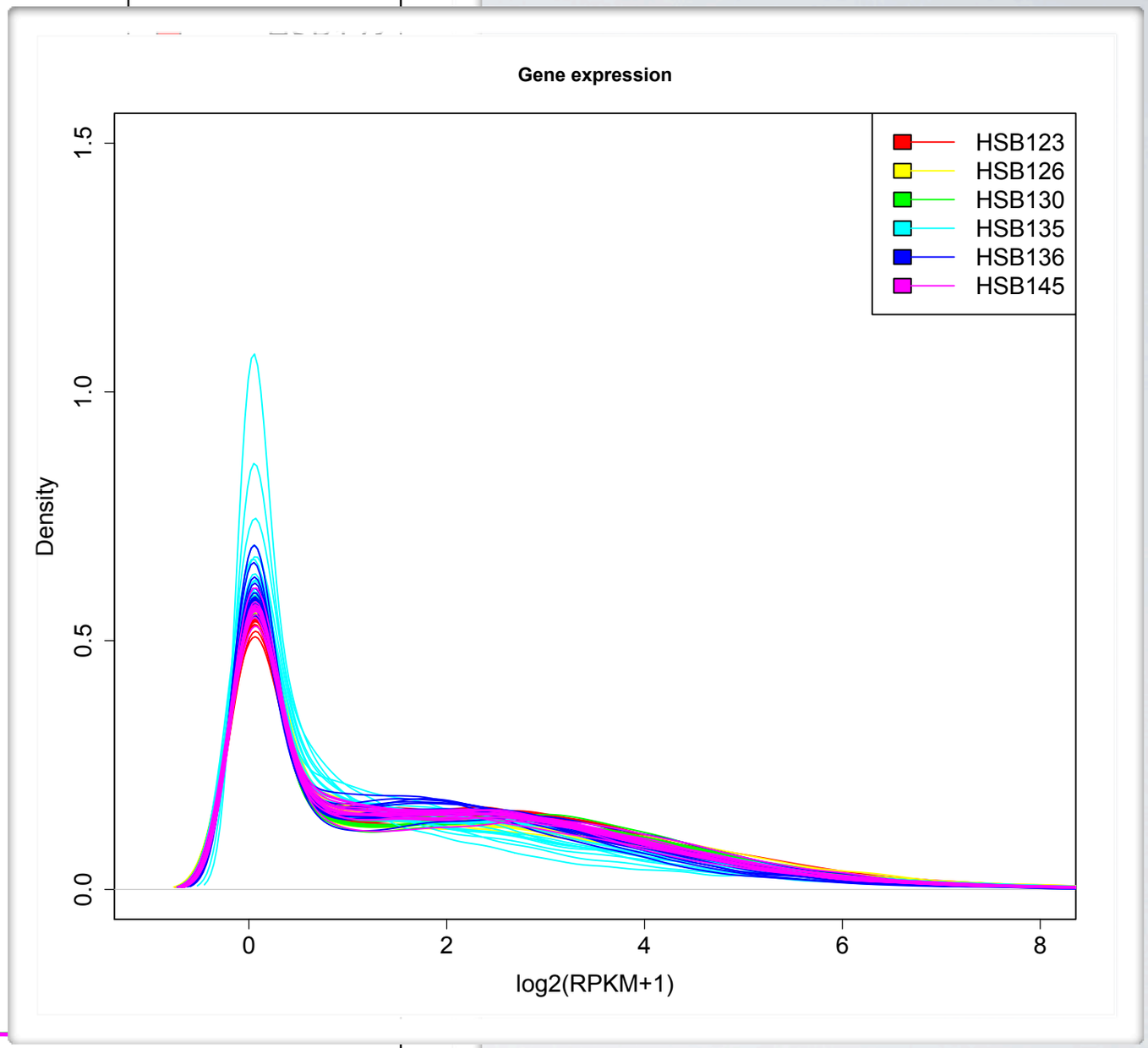
$$t_{bgr} = |\vec{L}| \cdot r_c$$

- Define *threshold* (normalized per million mapped nt):

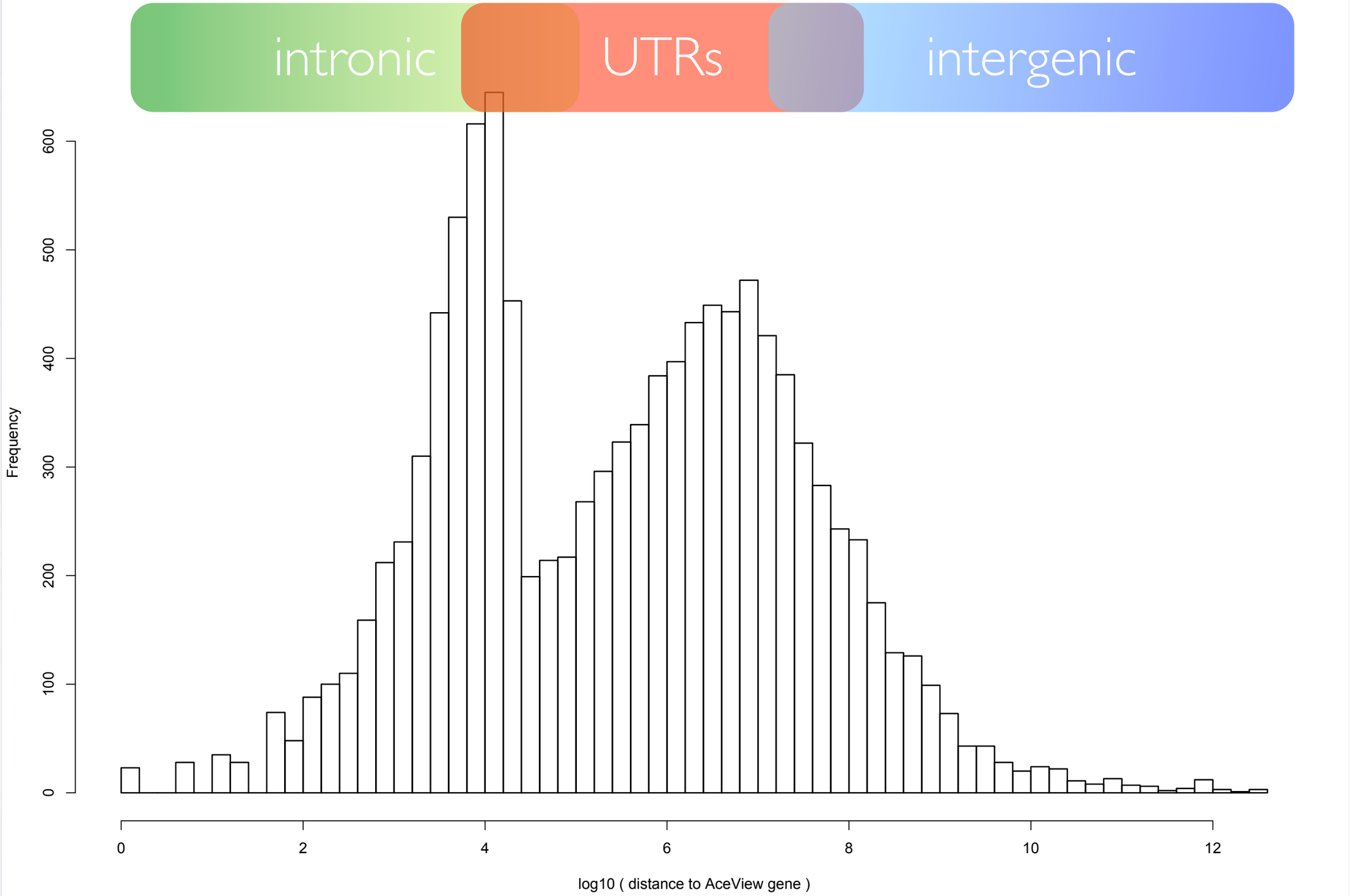
# RESULTS




- Threshold = 0.019617
- Max-Gap = 50nt
- Min-Run = 100nt
- 12,220 novel TARs
- Removing "homologous" regions: **11,327 TARs** (~7% reduction)



# Distance from Genes

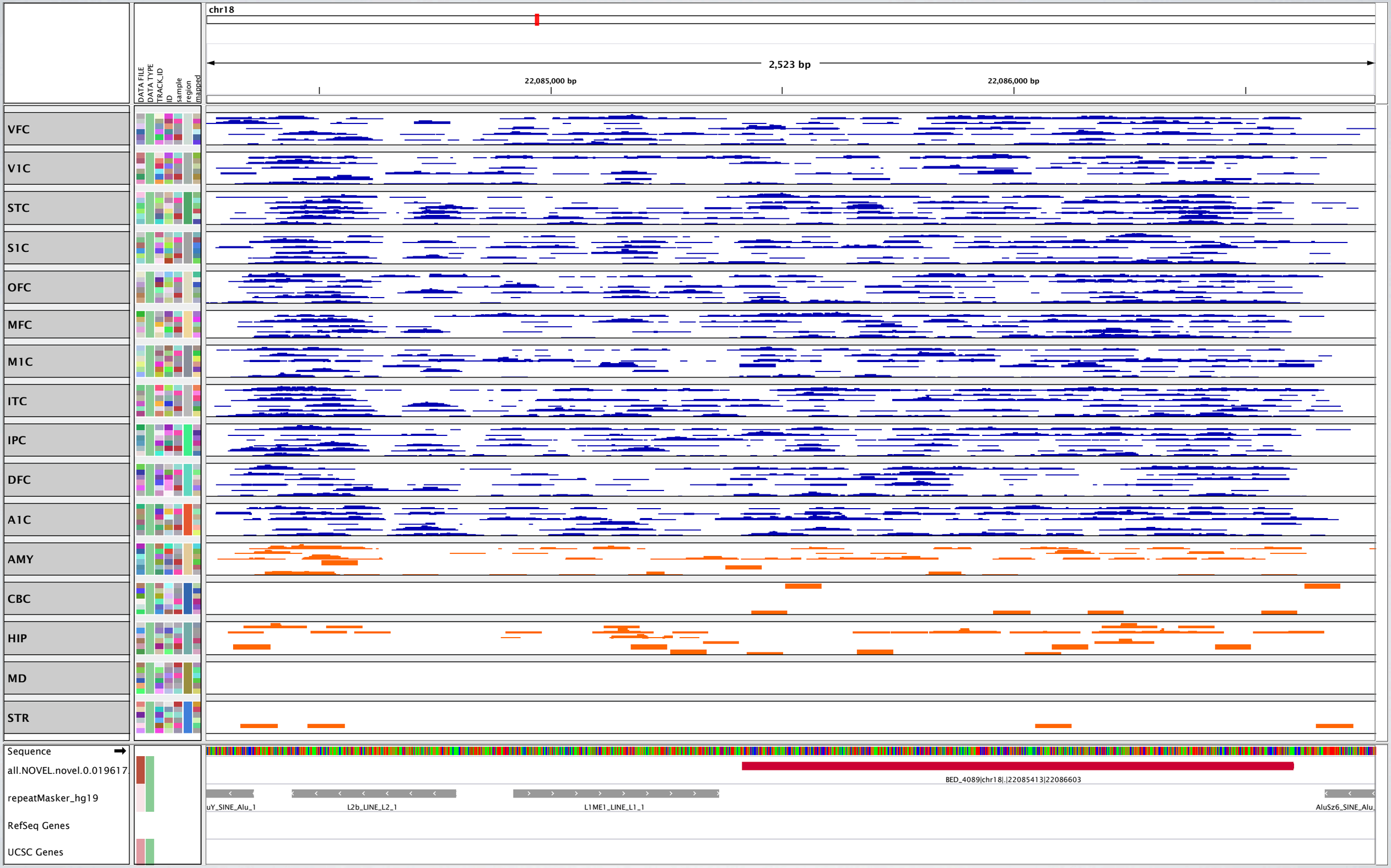


# DIFFERENTIAL EXPRESSION (PRELIMINARY)

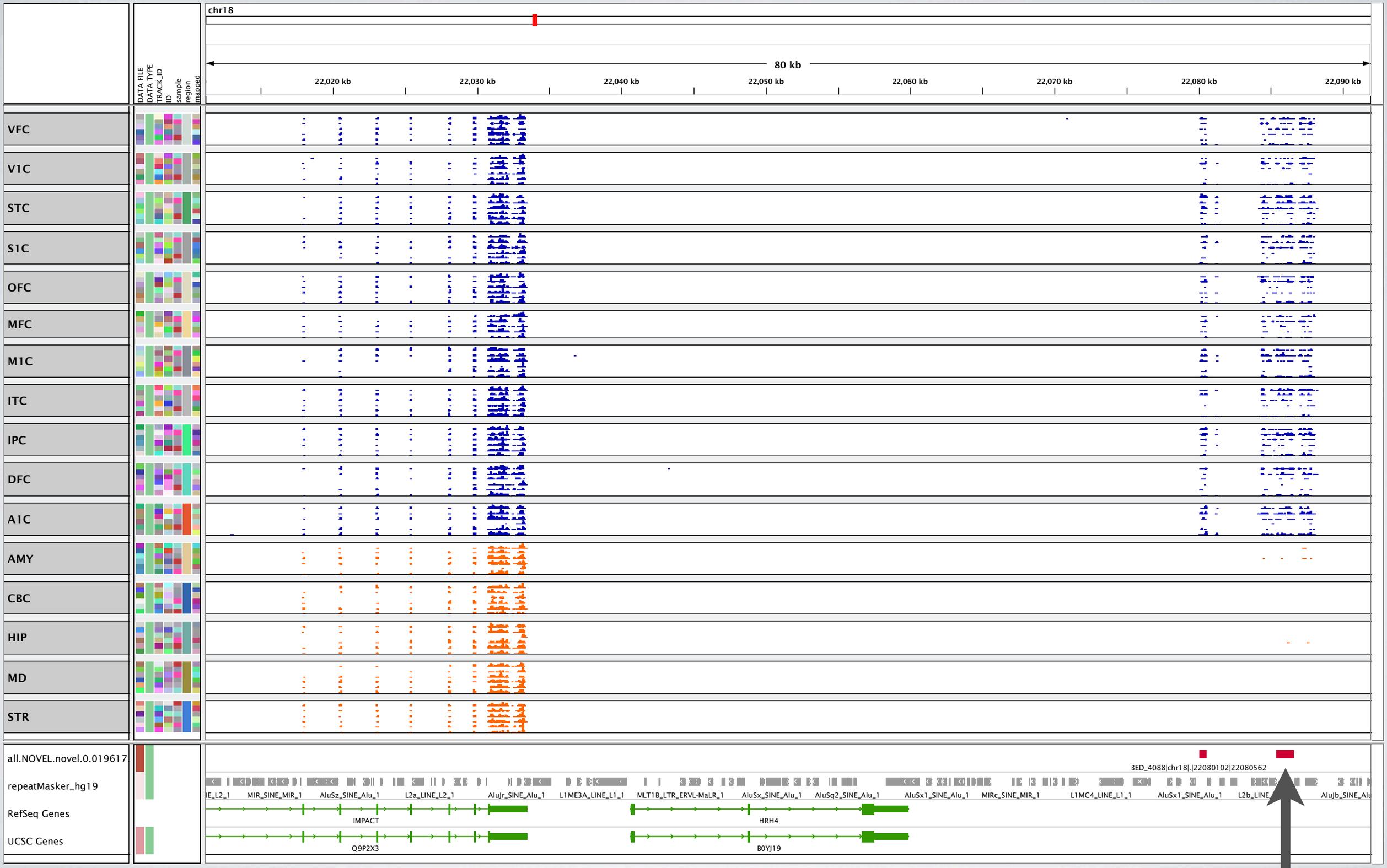
- Normalized by the number of mapped nucleotides
- t-test test between neocortex (NCX) and the "rest" (AMY, CBC, HIP, MD, STR)

| cutoff | Raw p-values | BH | Bonferroni |
|---|---|---|---|
| 0.05 | 4180 | 3242 | 1067 |
| 0.01 | 3045 | 2359 | 900 |

# EXAMPLE


chr18
2,523 bp
22,085,000 bp
22,086,000 bp
DATA FILE
DATA TYPE
TRACK_ID
ID
sample
region
mapped
VFC
V1C
STC
S1C
OFC
MFC
M1C
ITC
IPC
DFC
A1C
AMY
CBC
HIP
MD
STR
Sequence
all.NOVEL.novel.0.019617
repeatMasker_hg19
RefSeq Genes
UCSC Genes
BED_4089|chr18|.|22085413|22086603
uY_SINE_Alu_1
L2b_LINE_L2_1
L1ME1_LINE_L1_1
AluSz6_SINE_Alu_

# EXAMPLE (ZOOM-OUT)



# ANOTHER EXAMPLE

# Future Steps

- Analyze "intronic", "UTR", and intergenic novel TARs
- Classify them with incRNA to determine their "non-coding" potential

IGV – Session: /Users/andrea/Desktop/IGV/igv_session_brainSeq2.xml

Human hg19 | All | Go

DATA FILE
DATA TYPE
TRACK_ID
ID
sample
region
mapped

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y

all.NOVEL.novel.0.019617
repeatMasker.hg19
Zoom in to see features.
RefSeq Genes
UCSC Genes

chr21:9,828,152

278M of 534M