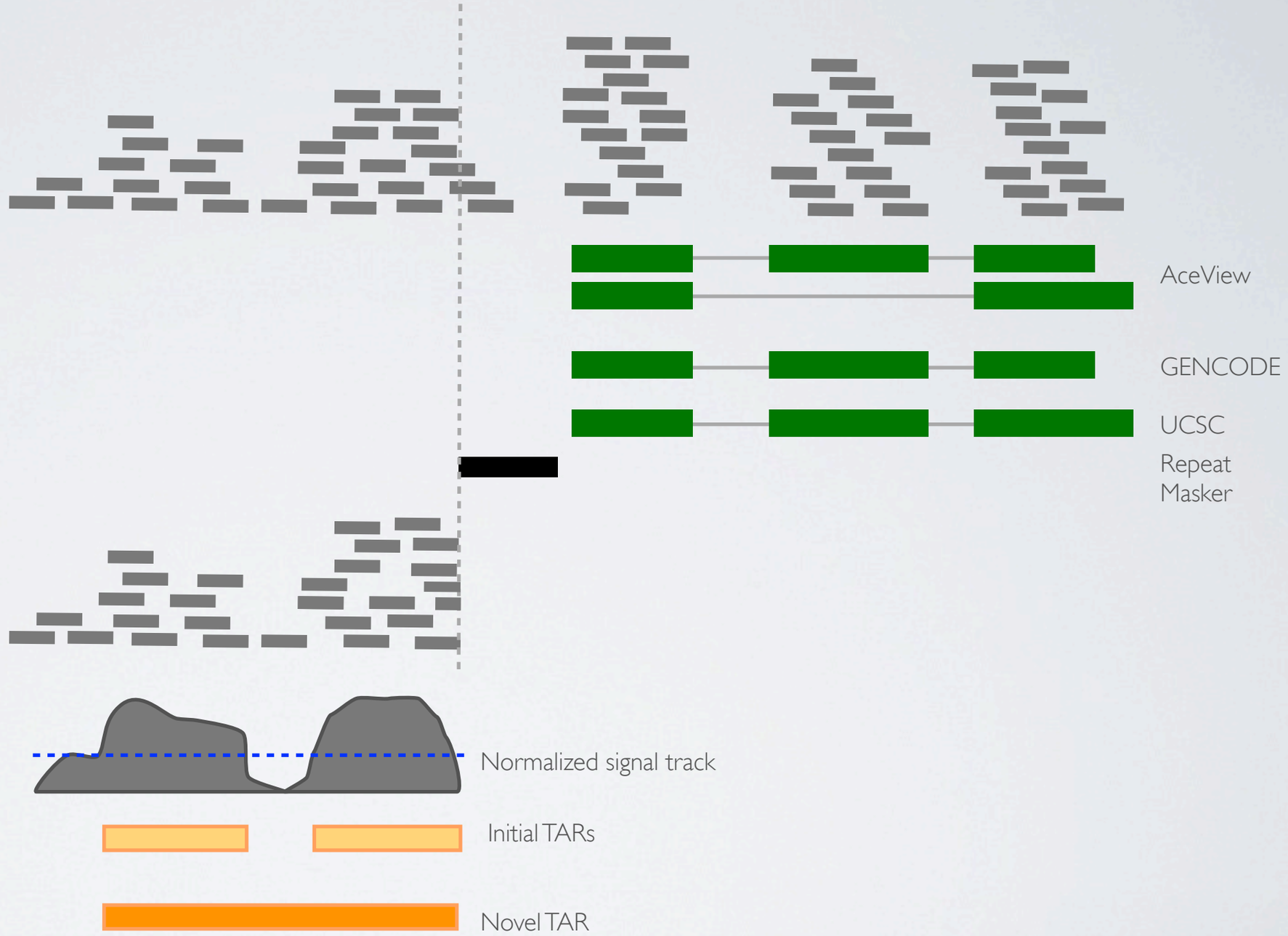
Novel Transcriptionally Active Regions (TARs)

Andrea Sboner - 2011.04.08

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

$$r(g) = \frac{n(g)}{l(g) * M} \qquad x(g) = log_2(r(g) + 1)$$

- 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:

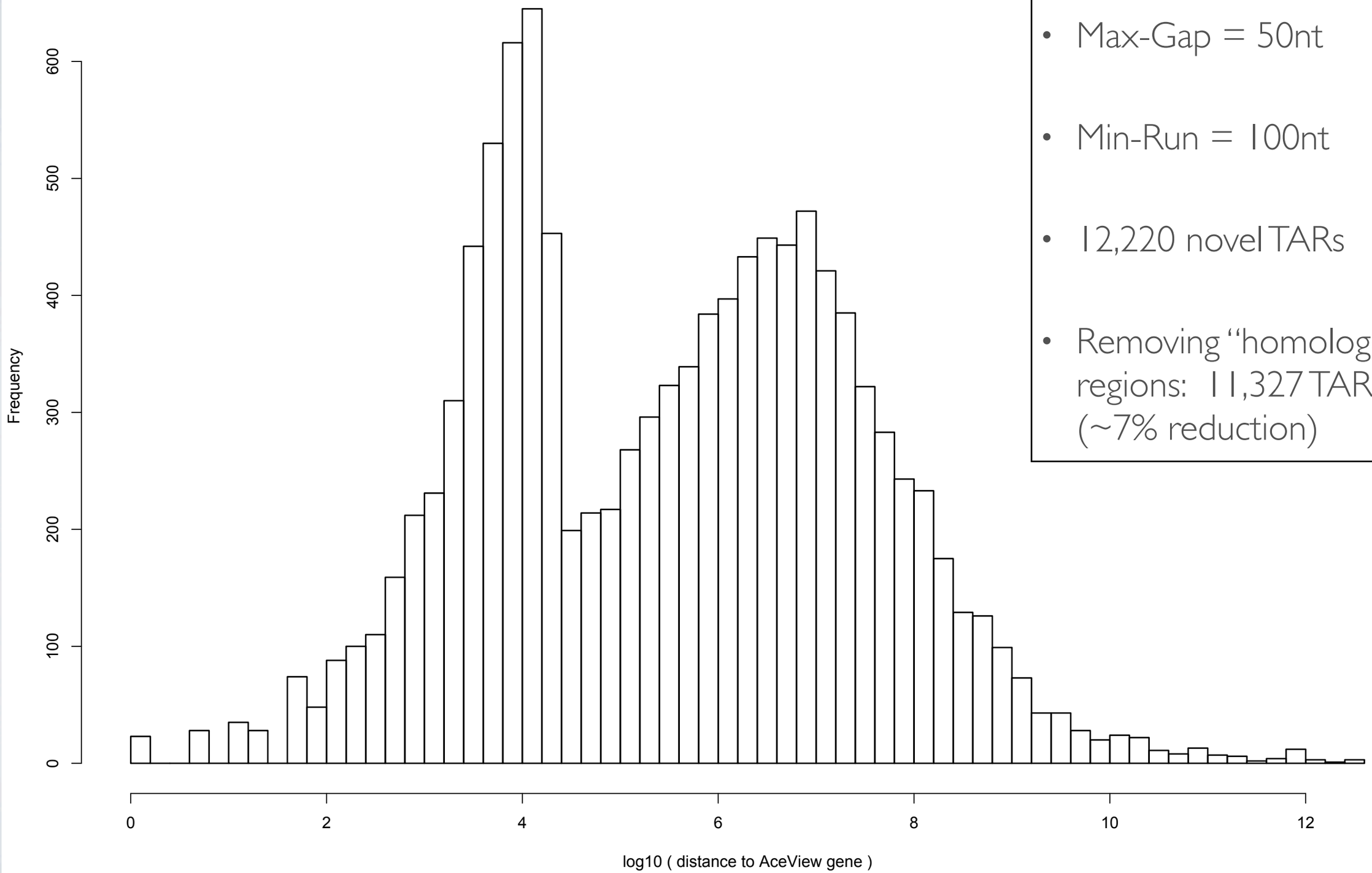
$$|\vec{L}| = 0.129Kb$$

- Consider the median length of GENCODE exons:

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

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

RESULTS



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

DIFFERENTIAL EXPRESSION (PRELIMINARY)

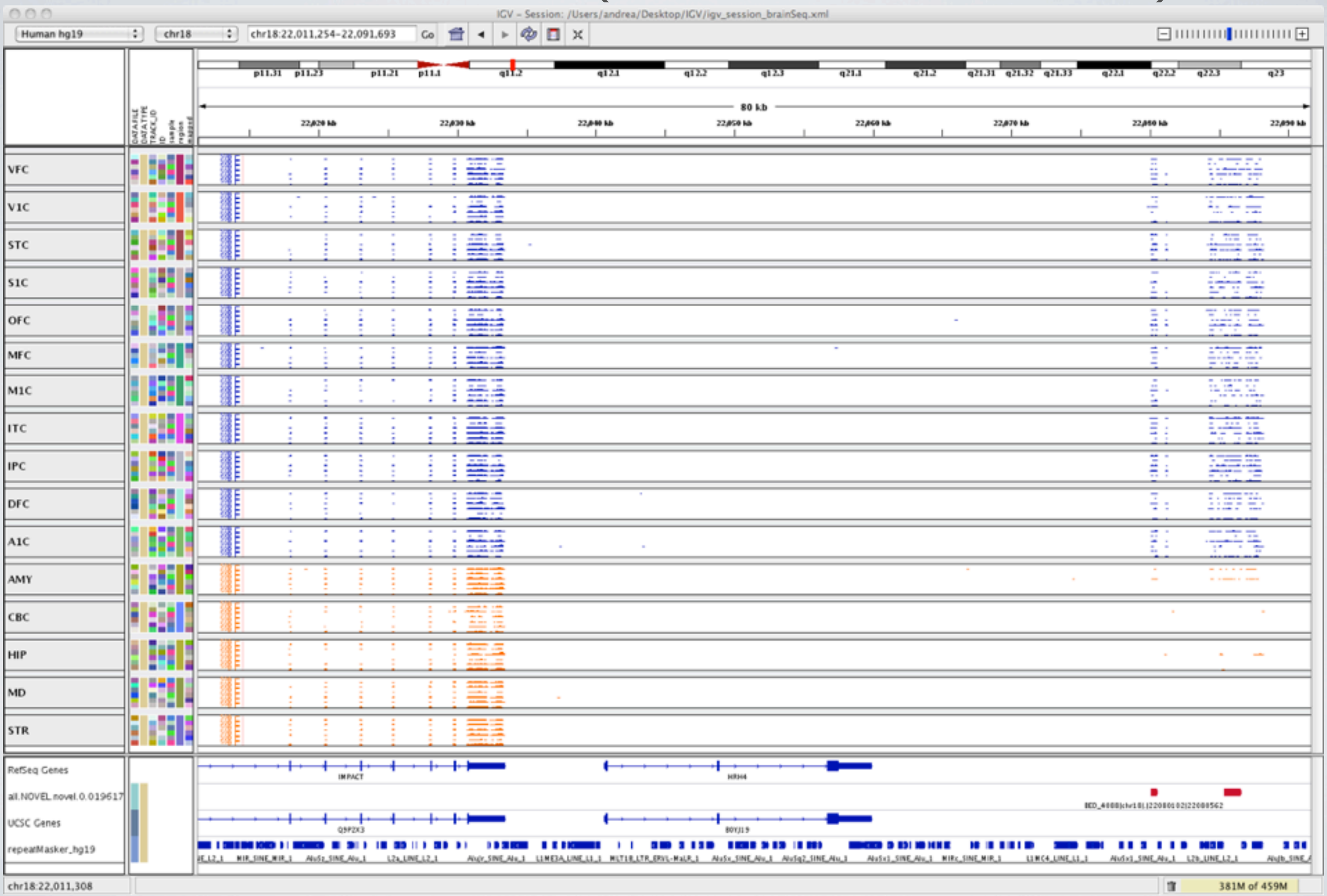
- Normalized by the number of mapped nucleotides
- t-test test between neocortex (NCX) and the "rest" (AMY, CBC, HIP, MD, STR)

cutoff	Raw p-values	BH	Bonferroni
0.05	3425	2136	403
0.01	2170	1288	295

EXAMPLE


IGV - Session: /Users/andrea/Desktop/IGV/igv_session_brainSeq.xml
Human hg19
chr18
chr18:22,078,650-22,088,704
Go
p11.31
p11.23
p11.21
p11.1
q11.2
q12.1
q12.2
q12.3
q21.1
q21.2
q21.31
q21.32
q21.33
q22.1
q22.2
q22.3
q23
10 kb
22,050 kb
22,052 kb
22,054 kb
22,056 kb
22,058 kb
VFC
V1C
STC
S1C
OFC
MFC
M1C
ITC
IPC
DFC
A1C
AMY
CBC
HIP
MD
STR
RefSeq Genes
all.NOVEL.novel.0.019617
UCSC Genes
repeatMasker_hg19
chr18:22,082,367
391M of 459M

EXAMPLE (ZOOM-OUT)



ANOTHER EXAMPLE

Future Steps

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