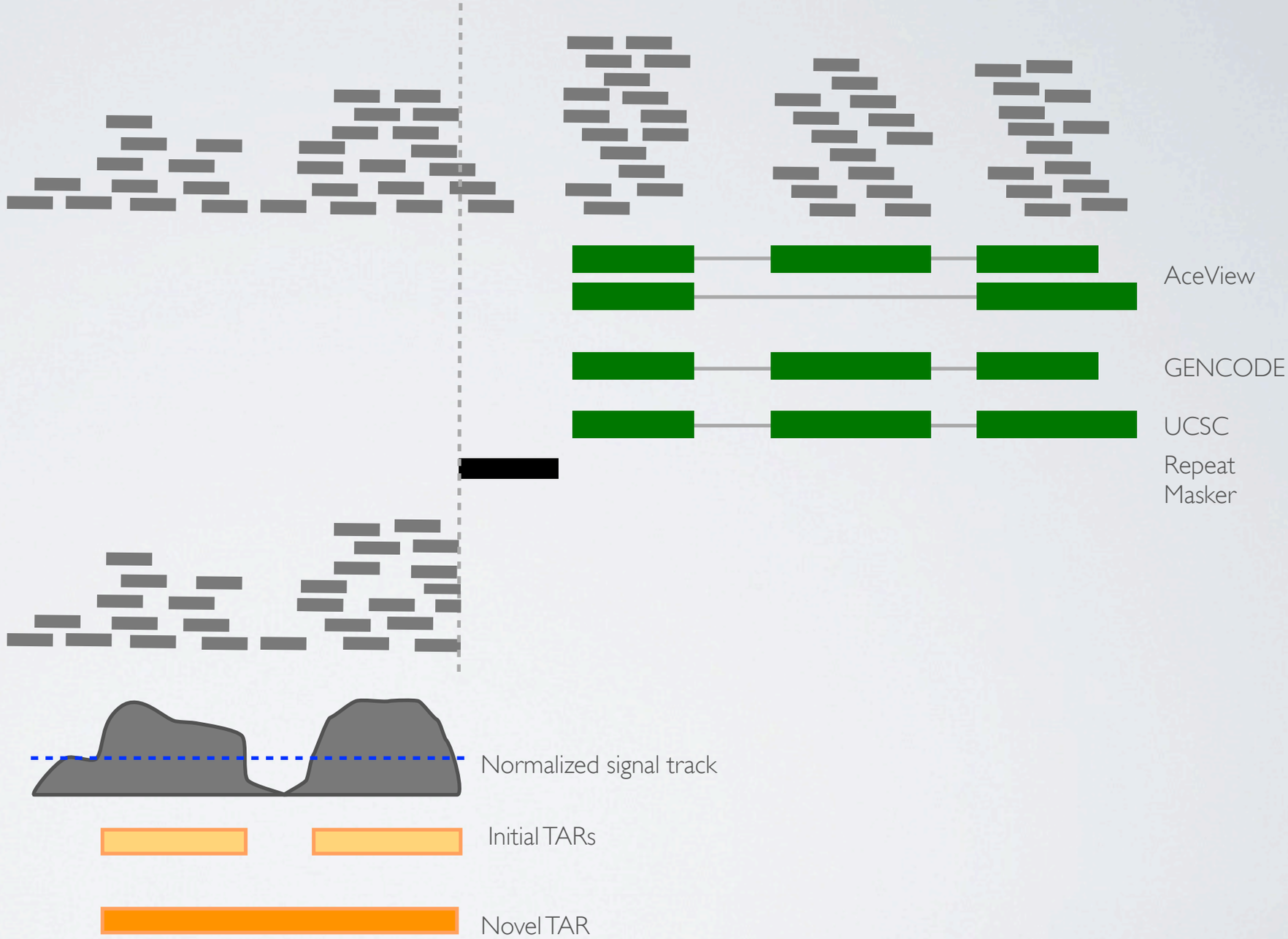
# Novel Transcriptionally Active Regions (TARs)

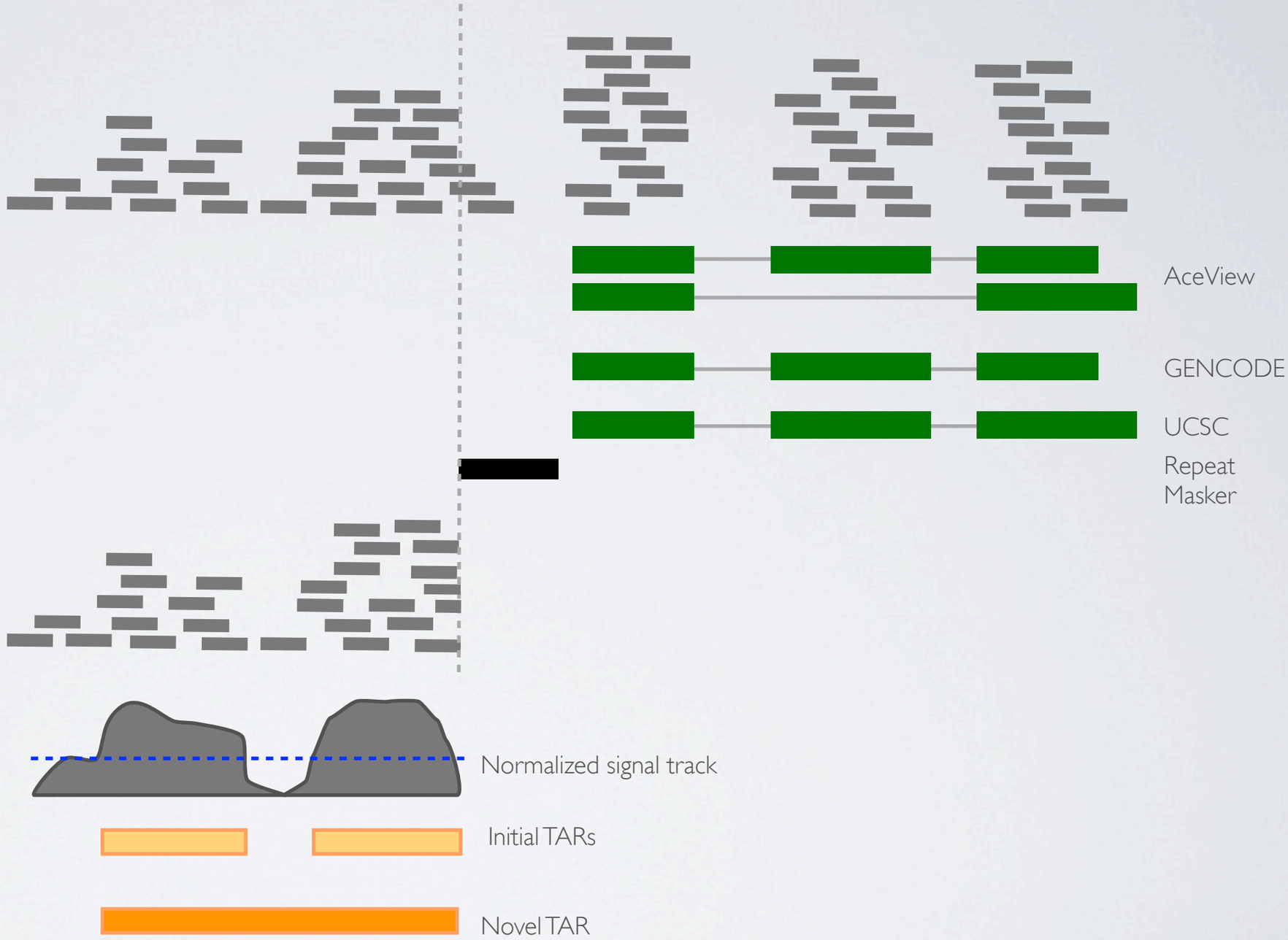
Andrea Sboner - 2011.04.08

# Discovery of Novel TARs



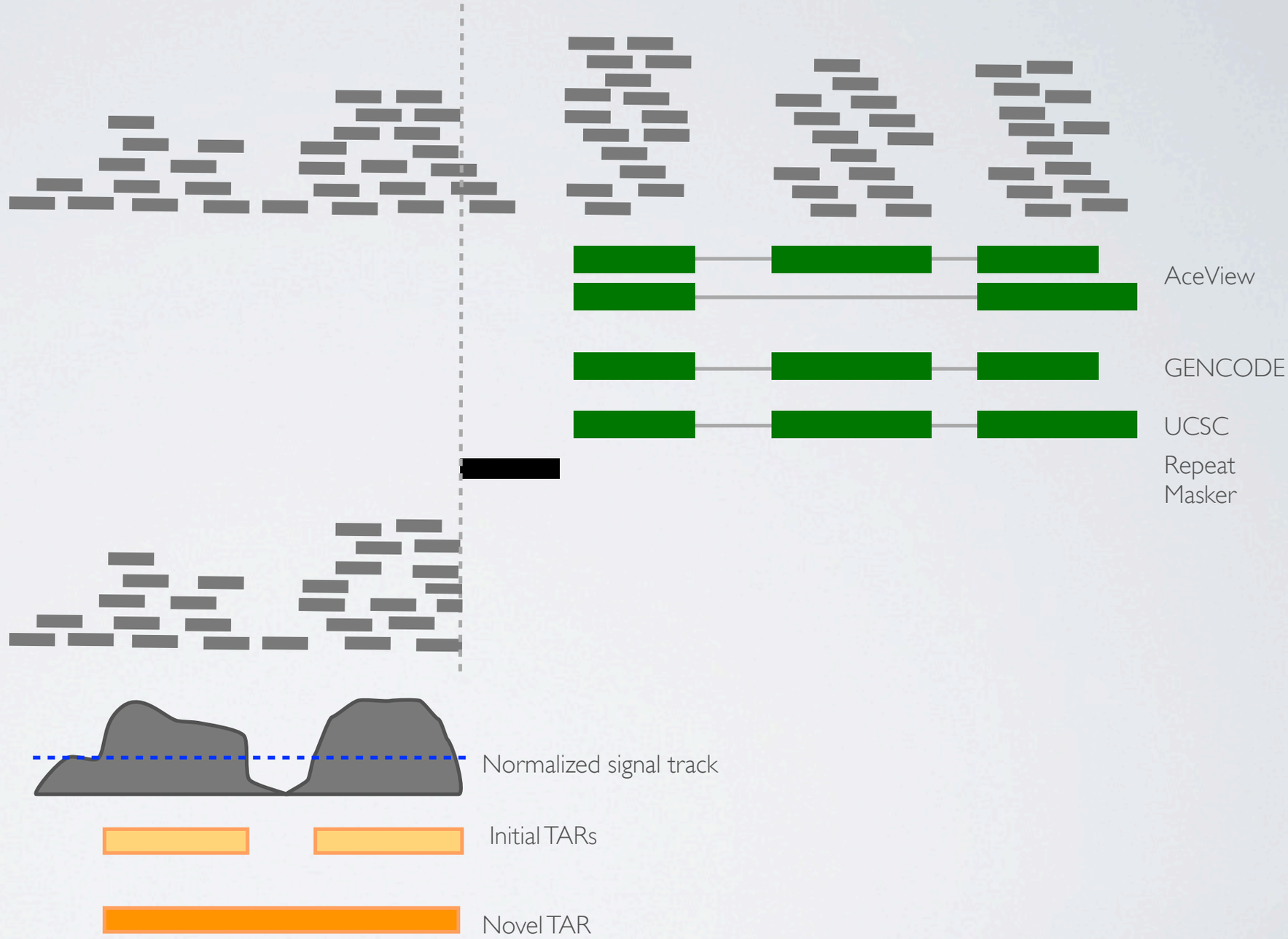
# Discovery of Novel TARs

- Reads overlapping gene annotation set -- AceView, GENCODE, UCSC, and repetitive regions are excluded



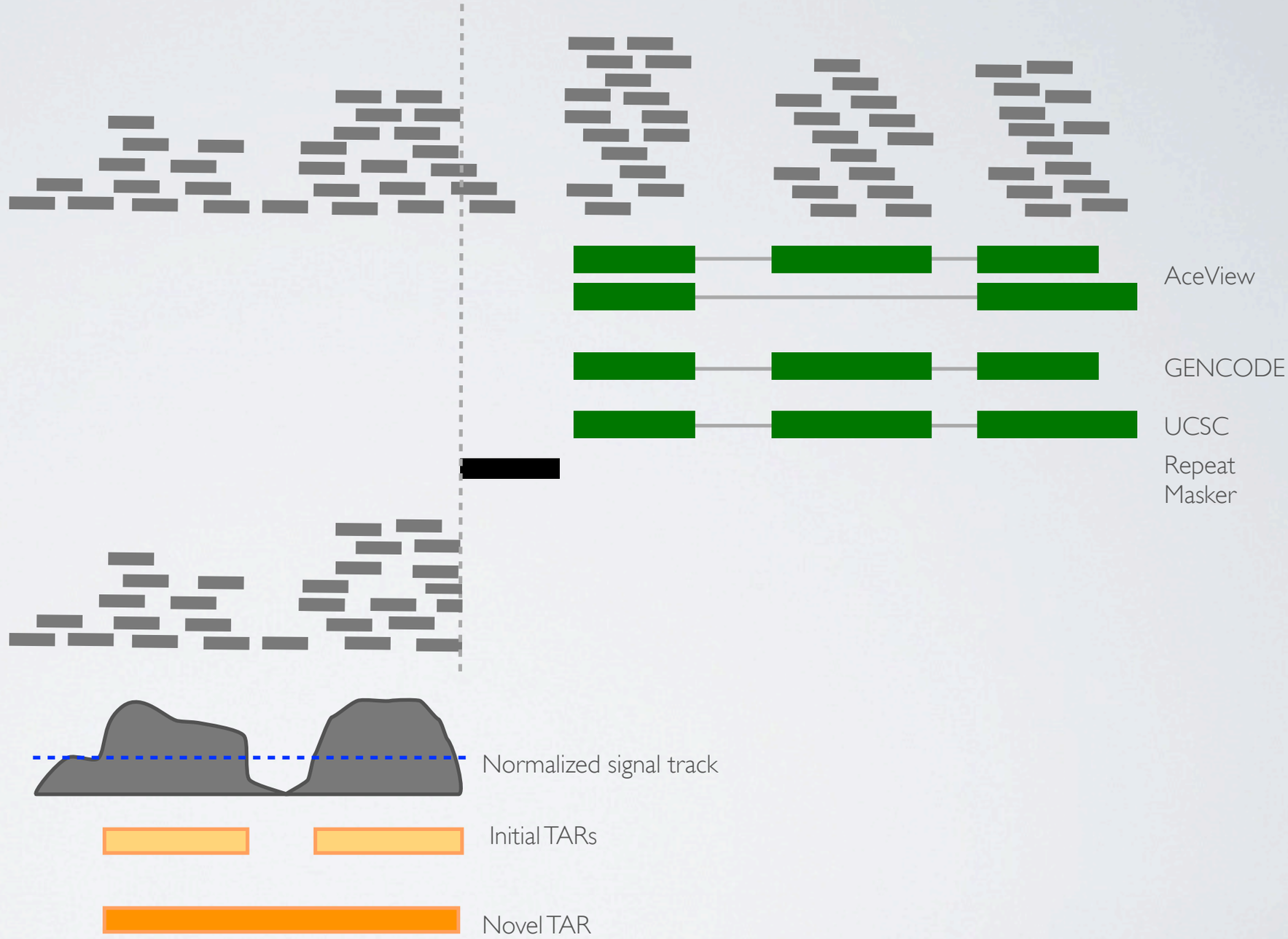
# Discovery of Novel TARs

- Reads overlapping gene annotation set -- AceView, GENCODE, UCSC, and repetitive regions are excluded
- Reads from all samples pooled together



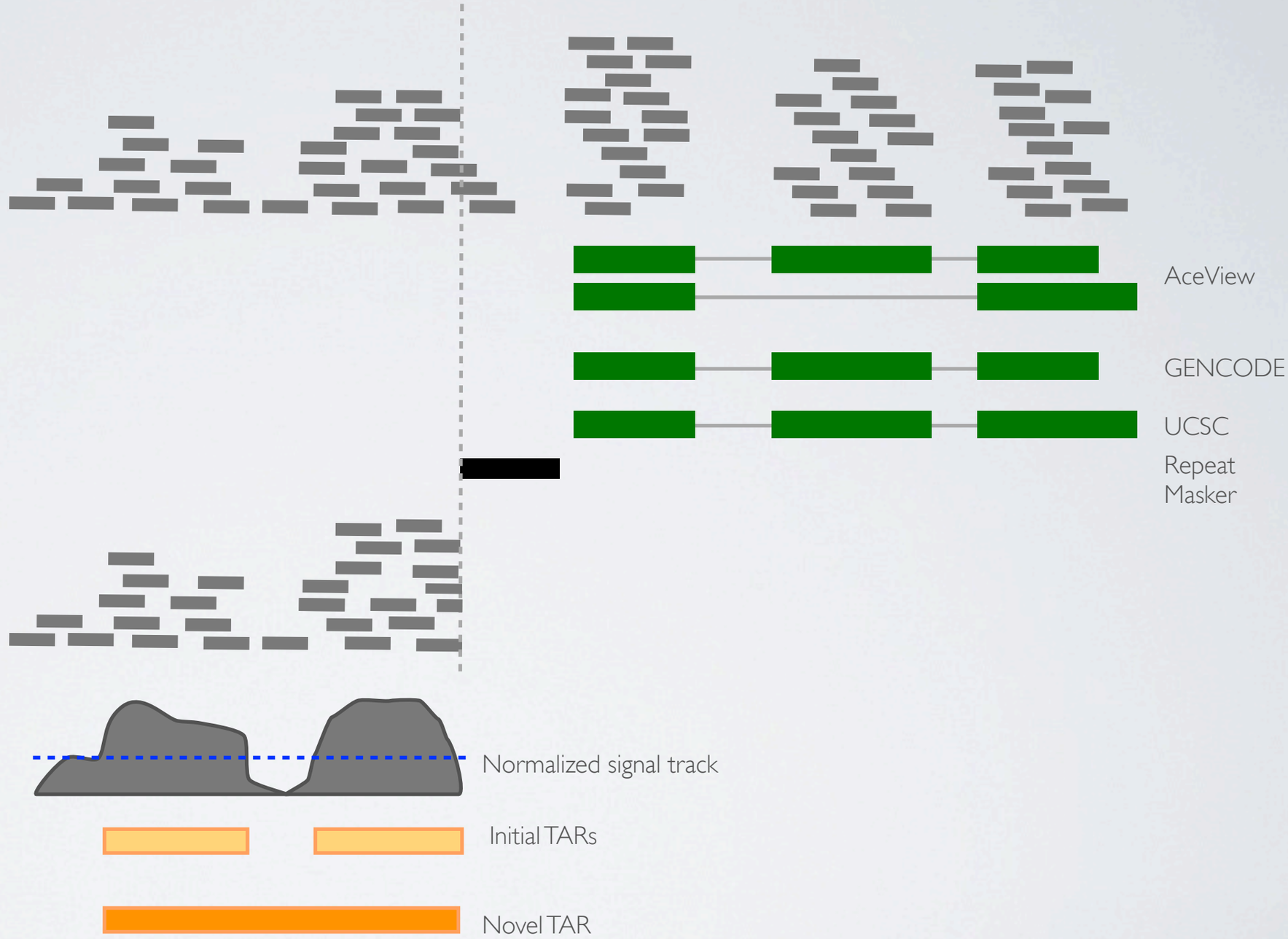
# Discovery of Novel TARs

- Reads overlapping gene annotation set -- AceView, GENCODE, UCSC, and repetitive regions are excluded
- Reads from all samples pooled together



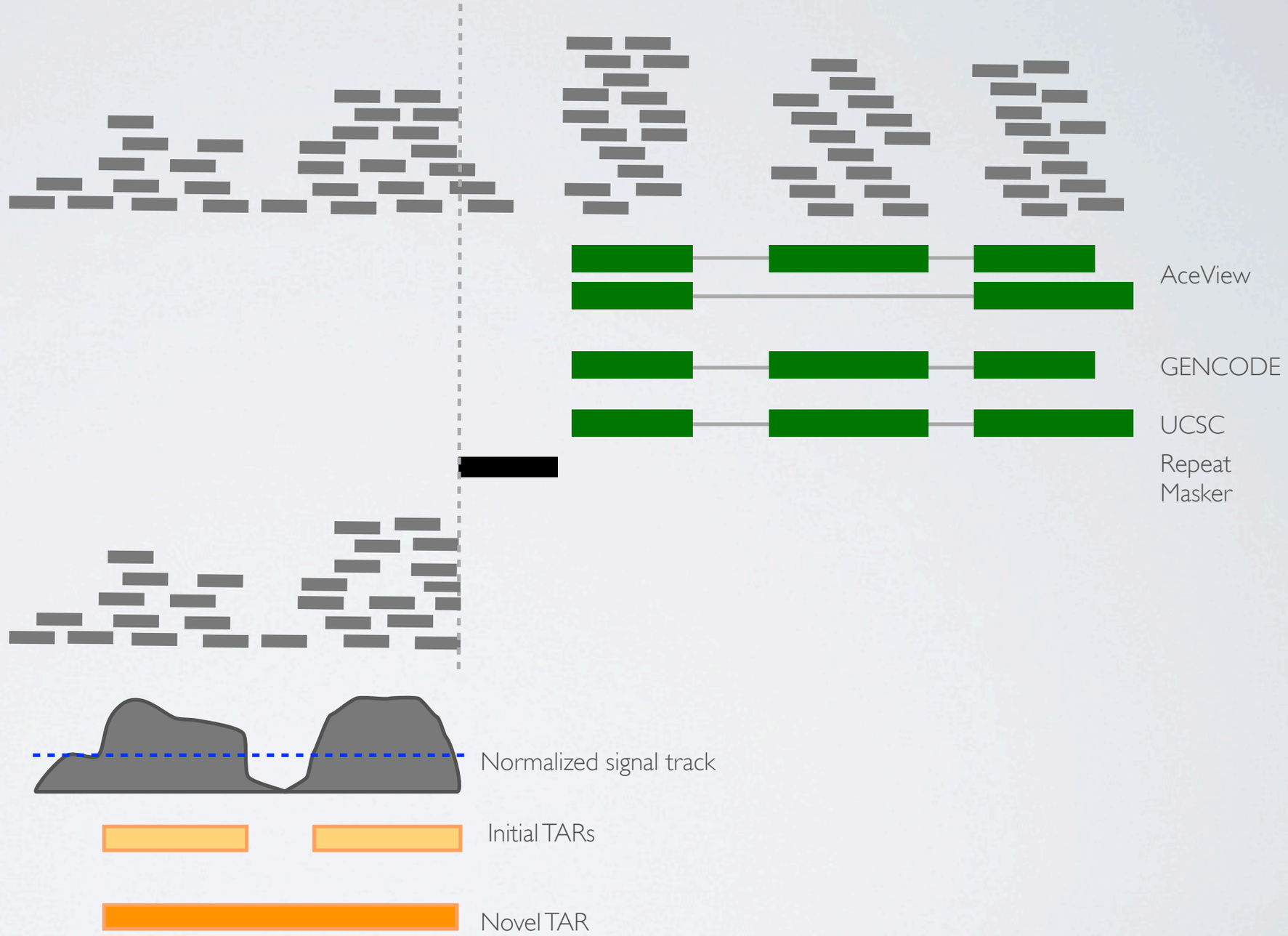
# Discovery of Novel TARs

- Reads overlapping gene annotation set -- AceView, GENCODE, UCSC, and repetitive regions are excluded
- Reads from all samples pooled together



# 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

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

# Min-GAP, MAX-Run algorithm: Threshold definition

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

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

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

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

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

# 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:

  Determine the median value of each gene: $\tilde{x}(g)$

$$x_c = quantile(\vec{\tilde{x}}, 0.05)$$

$$\forall \tilde{x}(g) > 0$$

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

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

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

# 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:
  Determine the median value of each gene: $\tilde{x}(g)$
- Computed the 5th percentile of expressed genes:

$$x_c = quantile(\vec{\tilde{x}}, 0.05)$$
$$\forall \tilde{x}(g) > 0$$
$$r_c = 2^{x_c} - 1$$
$$|\vec{L}| = 0.129Kb$$
$$t_{bgr} = |\vec{L}| \cdot r_c$$

# 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:
  Determine the median value of each gene: $\tilde{x}(g)$
- Computed the 5th percentile of expressed genes: $x_c = quantile(\vec{\tilde{x}}, 0.05)$ $\forall \tilde{x}(g) > 0$
- Compute the corresponding RPKM value: $r_c = 2^{x_c} - 1$

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

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

# 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:
  Determine the median value of each gene: $\tilde{x}(g)$
  - Computed the 5th percentile of expressed genes: $x_c = quantile(\vec{\tilde{x}}, 0.05)$ $\forall \tilde{x}(g) > 0$
  - Compute the corresponding RPKM value: $r_c = 2^{x_c} - 1$
  - Consider the median length of GENCODE exons: $|\vec{L}| = 0.129Kb$

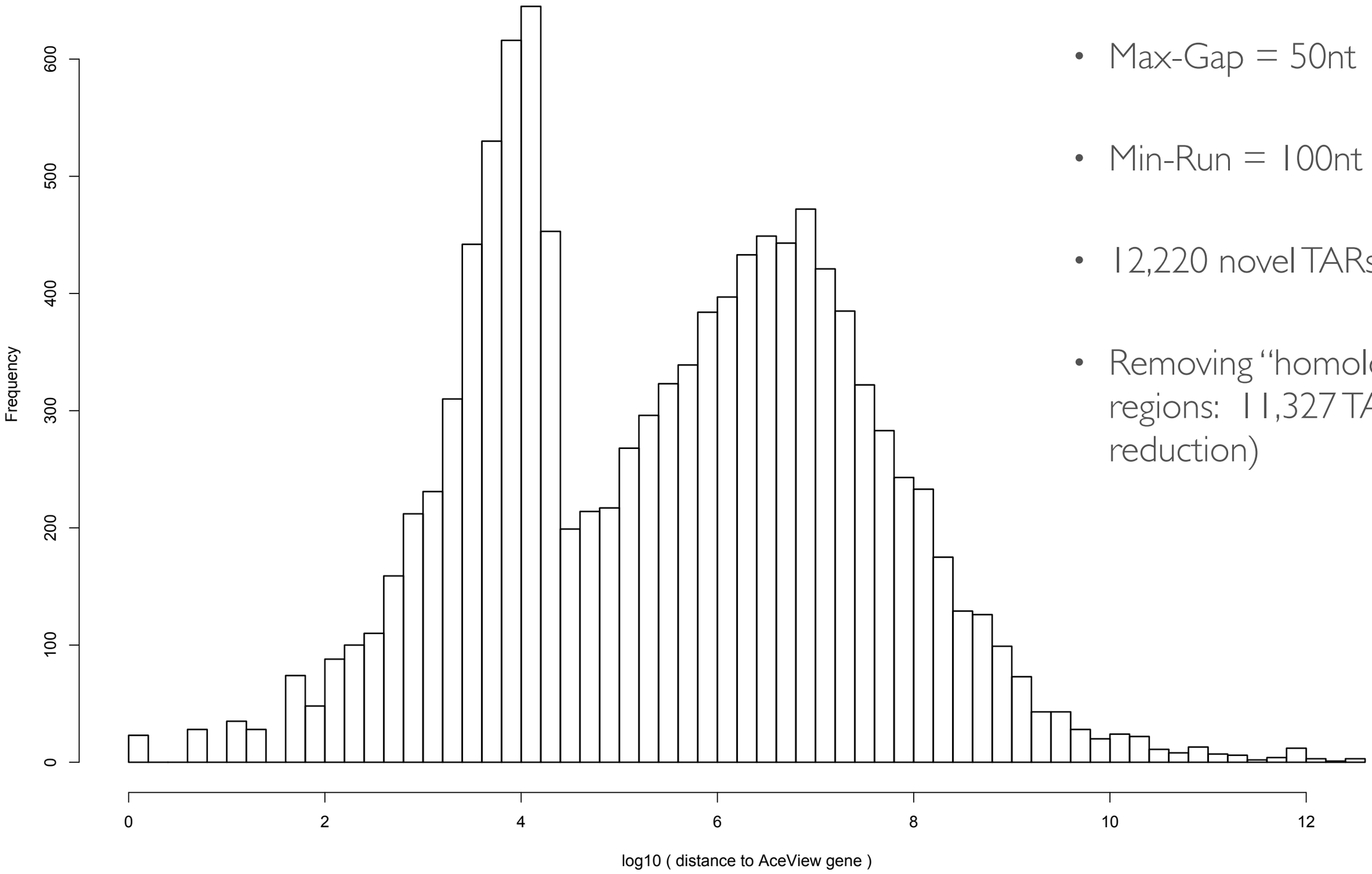
$$t_{bgr} = |\vec{L}| \cdot r_c$$

# 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:
  Determine the median value of each gene: $\tilde{x}(g)$
- Computed the 5th percentile of expressed genes: $x_c = quantile(\vec{\tilde{x}}, 0.05)$ $\forall \tilde{x}(g) > 0$
- Compute the corresponding RPKM value: $r_c = 2^{x_c} - 1$
- Consider the median length of GENCODE exons: $|\vec{L}| = 0.129Kb$
- Define *threshold* (normalized per million mapped reads): $t_{bgr} = |\vec{L}| \cdot r_c$

# RESULTS




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

# EXAMPLE


p15.2 p14 p13 p12.33 p12.2 p12.1 p11.22 p11.1 q11.21 q11.23 q21.1 q21.2 q21.3 q22.1 q22.2 q22.3 q23.1 q23.2 q23.32 q24.1 q24.31 q25.1 q25.2 q25.3 q26.12
155 kb
3,260 kb
3,280 kb
3,300 kb
3,320 kb
3,340 kb
3,360 kb
3,380 kb
VFC
V1C
STR
No expression
STC
S1C
OFC
MFC
MD
Low expression
M1C
ITC
IPC
HIP
Low expression
DFC
CBC
No expression
AMY
A1C
Low expression
RefSeq Genes
all.NOVEL.novel.0.019617
UCSC Genes
repeatMasker_hg19