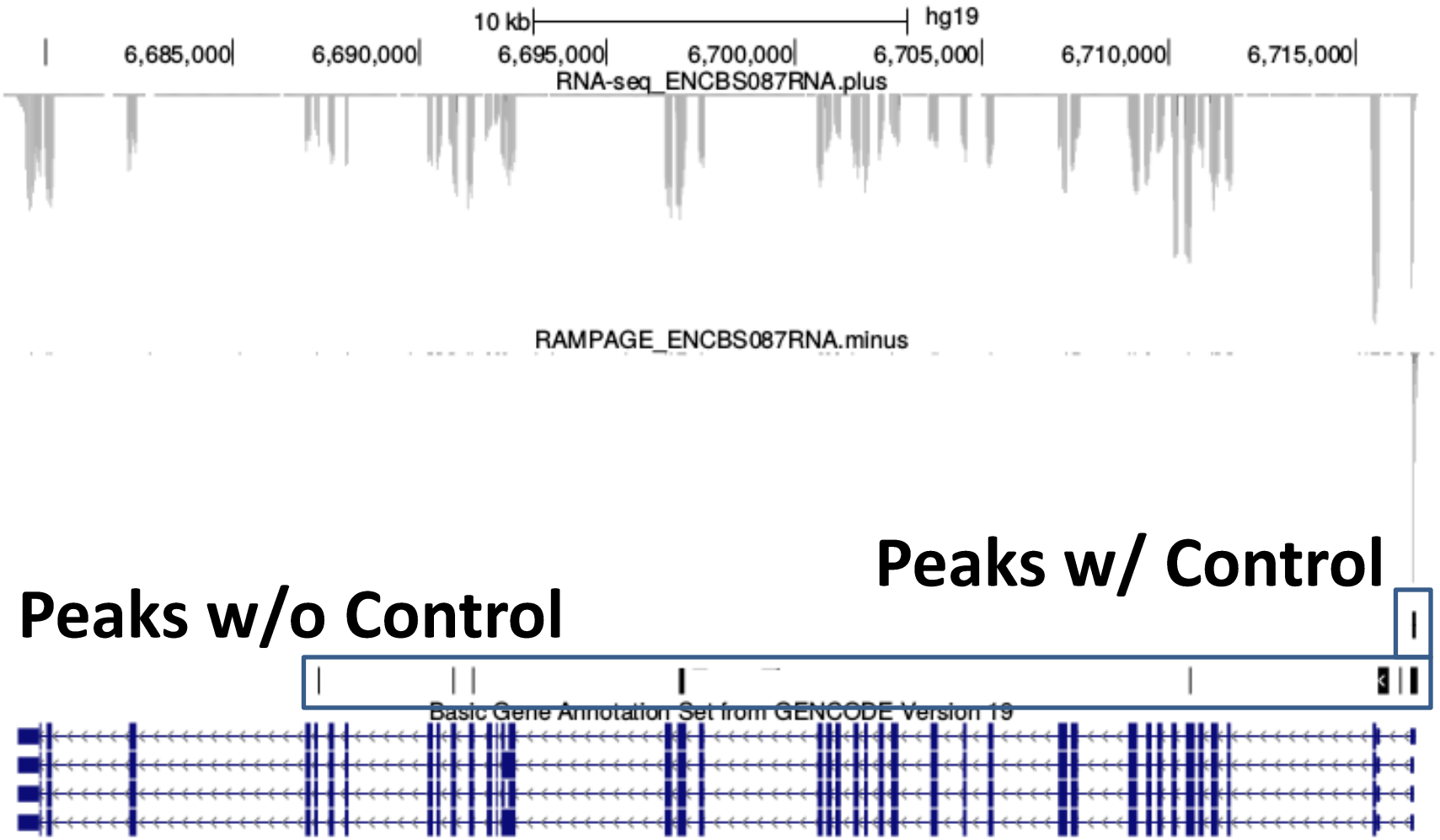
RAMPAGE Peak Calling Pipeline

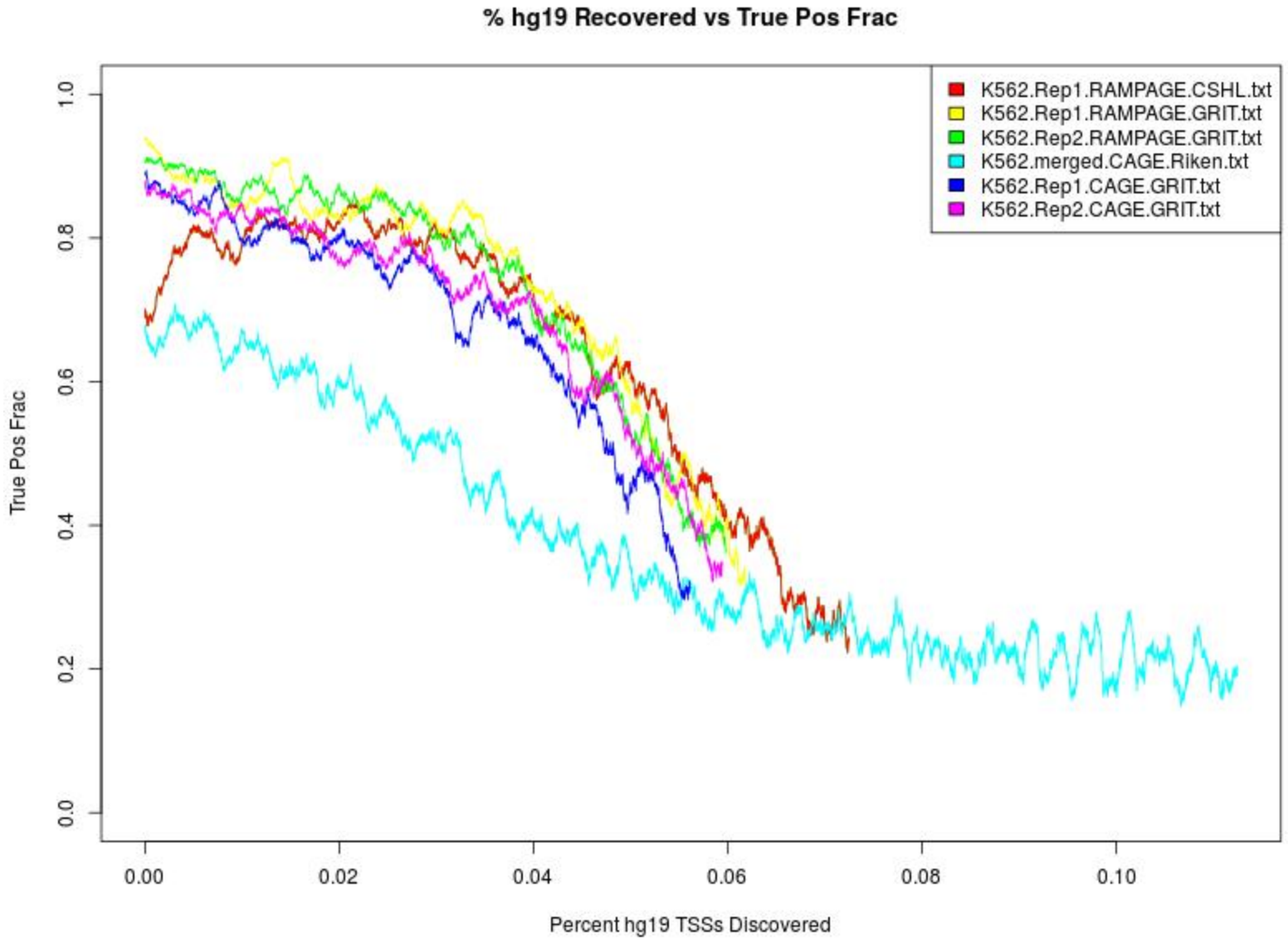
Nathan Boley**, Peter Bickel*, Anshul Kundaje**
*UC Berkeley, **Stanford

An RNA-Seq Control Allows the Removal of Experimental Artifacts



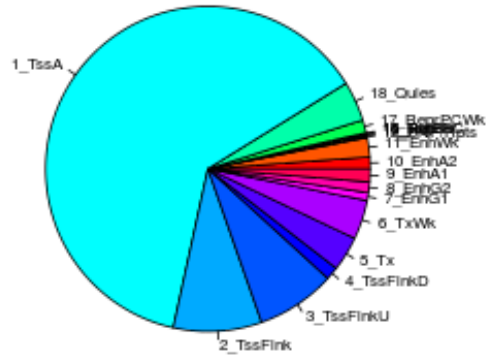
Analysis

K562 Peak Call Comparison

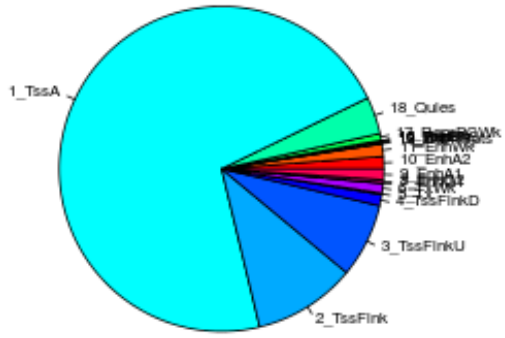


Epigenomic Roadmap Predicted States

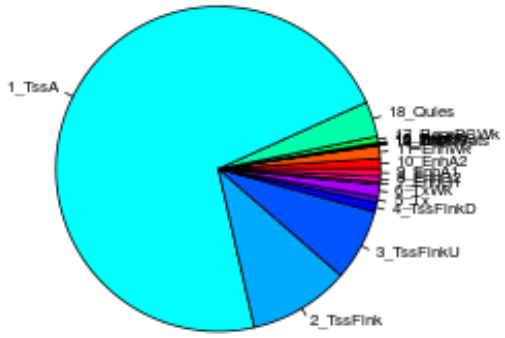
K562.rep1.CSHL.peaks_and_chromatin_state.txt



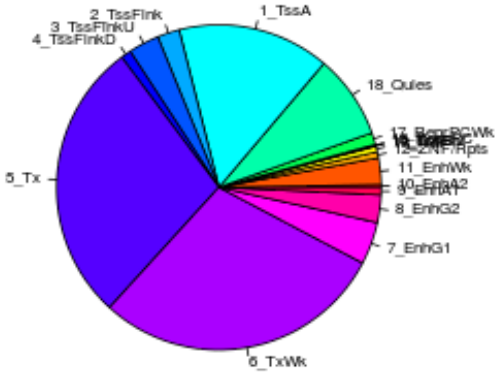
K562.GRIT.rep1.peaks_and_chromatin_state.txt



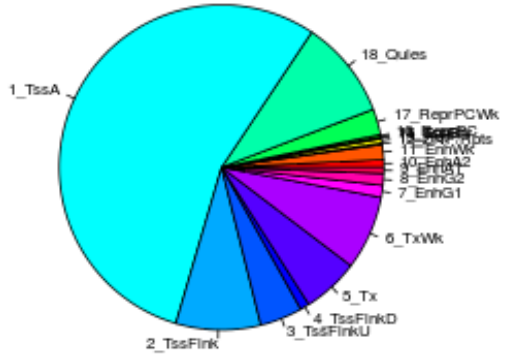
K562.GRIT.rep2.peaks_and_chromatin_state.txt



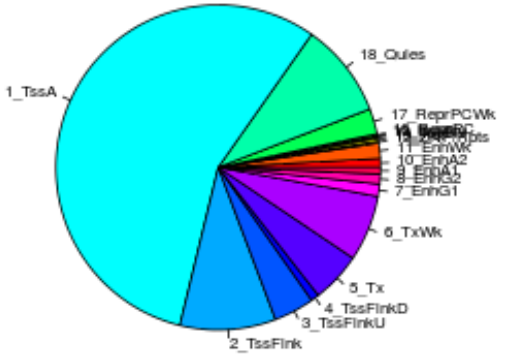
K562.CAGE.Riken.merged.peaks_and_chromatin_state.txt



CAGE.GRIT.rep1.peaks_and_chromatin_state.txt



CAGE.GRIT.rep2.peaks_and_chromatin_state.txt



Integrated Annotation

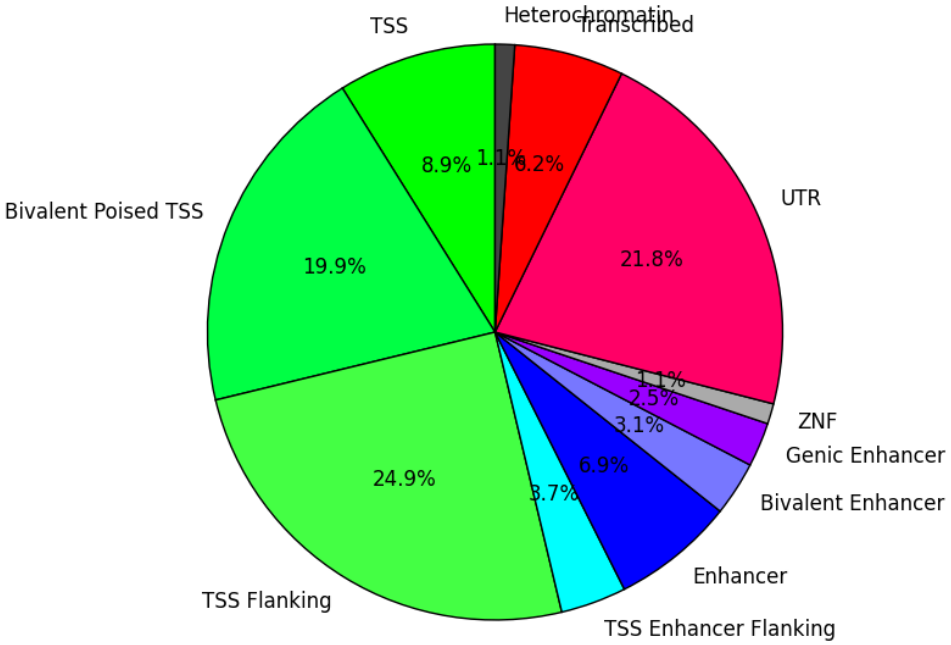
Pipeline

- Merge reproducible peaks
- Filter merged peaks
- Associate with GENCODE gene
- Find TSS peak density for each replicate
- Trim the merged peaks at +/- 5% of library depth normalized signal
- Add GENCODE TSS's corroborated by Histone Marks

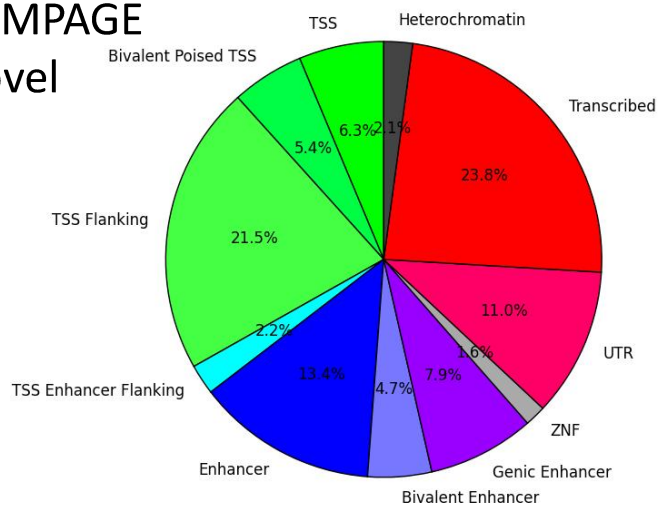
- Filtering:
  - Remove peaks from data w/o a proper control (all CAGE, a couple RAMPAGE)
  - Remove peaks that overlap a repeat and are not within 500 bp of a GENCODE TSS

Epilogos Chromatin State

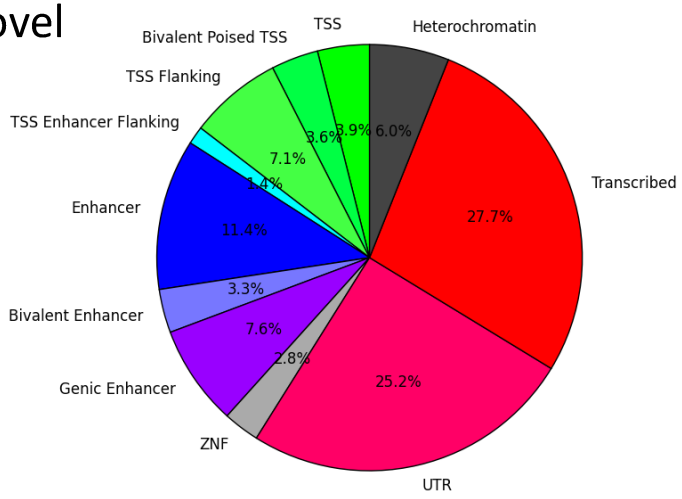
RAMPAGE/GENCODE Intersection



RAMPAGE Novel

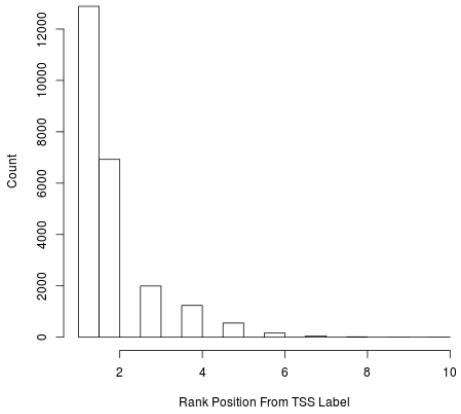


GENCODE Novel

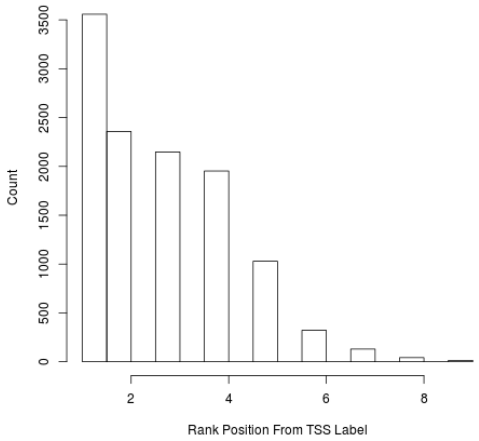


Epilogos Chromatin State

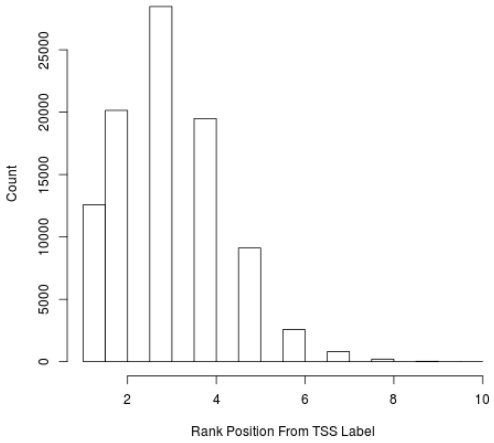
RAMPAGE + GENCODE



RAMPAGE Novel



GENCODE Novel



- Re-added all novel GENCODE TSS's with a TSS state label in >= 22% of epilogos cells
  - Corresponds to a 10% FDR

- Statistics
  - 73,781 Total TSS's
  - 36119 Distinct GENCODE genes
  - 11626 Novel Genes
  - 1746 Anti-sense TSS's
  - 1594 Anti-sense genes

Acknowledgments

- Alex Dobin
- Phillip Batut
- Carrie Davies
- Tom Gingeras
- Sue Celniker
- Ben Brown
- Roderic Guigo
- Sarah Djebali
- Tim Dreszer

Statistical Model

Statistical Model

Likelihood for the Mixture Model

- Y_i the observed count of TSS reads that start at base i
- N_i the observed count of RNAseq reads that start at base i
- s_i the probability of sampling a signal read that begins at base i
- n_i the probability of sampling a noise read that begins at base i
- λ the fraction of total reads that are noise

If we do not constrain the distribution of $\vec{s}$ and$\vec{n}$ then the log likelihood is

$$\begin{aligned} l(n, s; Y, N) &= \sum_i \{Y_i \log(\lambda n_i + (1-\lambda)s_i) - \log(Y_i!)\} + \sum_i \{N_i \log(n_i) - \log(N_i!)\} \\ s.t. \qquad & 1 = \sum_i n_i,\ 1 = \sum_i s_i,\ s_i \geq 0,\ n_i \geq 0 \end{aligned}$$

Statistical Model

Likelihood for the Mixture Model

- Given a region covering bases $[i, i+n)$
- Wish to test whether $n_i = 0$ for every position within the region
- Formally, for$\theta = \sum_{j=i}^{i+n} n_j$, test the null $\theta_0 = 0$ vs alternative $\theta_1 > 0$
- estimate n_j by maximizing $lhd(\theta_1; \vec{Y})$ subject to $\theta_1 = \sum_{j=i}^{i+n} n_j$ and $n_j \geq 0$
- Given λ, $\vec{s}$ the log likelihood is:

$$\begin{aligned}
\log\left[lhd(\theta_1;\vec{Y})/lhd(\theta_0;\vec{Y})\right] &= \sum_{j=i}^{i+n}\{Y_j \log\left(\lambda n_j + (1-\lambda)s_j\right)\} - \sum_{j=i}^{i+n}\{Y_j \log\left((1-\lambda)s_j\right)\} \\
&= \sum_{j=i}^{i+n} Y_j\left[\log\left(\lambda n_j + (1-\lambda)s_j\right)\right] - C
\end{aligned}$$

- The likelihood ratio statistic is non-dcreasing in θ, so we choose C s.t. $P\left[lhd(\theta_0;\vec{Y}) > C\right] = \alpha$

Statistical Model

Estimating the Critical Value

Estimating the critical value with the parametric bootstrap:

- Sample $N_{bootstrap}$ times from the multinomial with bin probabilities $\vec{n}$ and counts $\lambda\left(\sum Y_i\right)$
- Estimate the critical value by the α'th empirical quantile among $\sum_{j=i}^{i+n}\left\{N_j^{(k)}\log\left(n_j\right)-\log\left(N_j^{(k)}!\right)\right\}$

Aproximating the tail distribution:

- Under H_0 the distribution of counts at base i can be approximated by the binomial $Bin\left(\lambda\left(\sum Y_i\right),n_i\right)$
- n_i are typically very small
- We can efficiently calcualte the moments with a truncated series

$$Ef(X^m)=\eta_m \quad = \quad \sum_{k=0}^{\lambda\sum Y}\binom{\lambda\sum Y}{k}n_k^k(1-n_k)^{\lambda\sum Y-k}\left(k\log\left(n_i\right)-\log\left(k!\right)\right)^m$$

$$Ef(N_i)^m=\eta_m \quad = \quad \sum_{k=0}^{\lambda\sum_i Y_i}\frac{n_i^k}{k!}e^{-n_i}\left(k\log\left(n_i\right)-\log\left(k!\right)\right)^m$$

Statistical Model

Algorithm 1 Greedy identification of singal regions

Given a gene region of length L, an estimate of $\vec{s}$, a minimum region size, and a desired significance level α:

1. Initialize $\lambda_0 = 1$
2. Initialize the set of noise regions, $\mathbb{N}$, to the empty set
3. Initialize the set of signal regions, $\mathbb{S}$, to the empty set
4. Initialize the set of regions to test, $\mathbb{T}$, to contain the entire gene region
5. Until $\lambda_i = \lambda_{i-1}$
   (a) While $\mathbb{T}$ is non-empty, choose a region from $\mathbb{T}$, and test the region for significance at level $\alpha/2L$ (see 3.2.2).
      - If the region is not significant, add it to $\mathbb{N}$
      - If the region is significant and it is smaller than the minimum region size, add it to $\mathbb{S}$
      - Otherwise, split the region into 2 subregions by choosing the base with the lowest number of reads and add them to $\mathbb{T}$.
   (b) Update the estimate of lambda, setting it to

$$\lambda_i = \sum_{i=1}^{L} Y_i \mathbb{I}[i \in \mathbb{N}] / \left(\sum_{i=1}^{L} Y \sum_{ii=1}^{L} s_i \mathbb{I}[i \in \mathbb{N}] \right)$$

Model Overview

- Assumptions
  - Observed CAGE/RAMPAGE are a mixture of capped and uncapped reads
  - The fraction of noise reads is dependent on the region
  - The distribution of uncapped read starts is the same as from a matching RNA-seq experiment
- Approach
  - Estimate the uncapped fraction in each gene region
  - Given this estimate, hierarchically identify regions enriched for capped reads
  - Trim and filter peaks based upon signal coverage and realtive peak height