Annotation Free Analysis of Recurrent Somatic Mutations

Jason Liu

Mentor: JZ

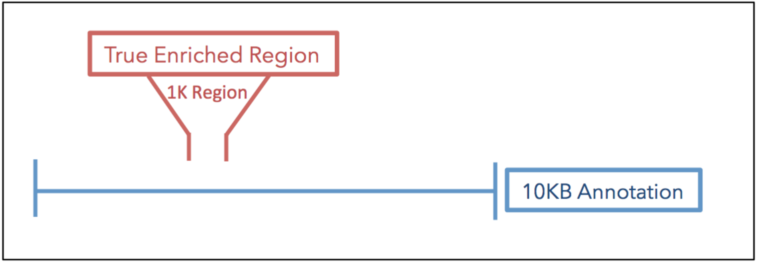
July 29, 2015

Identifying Non-coding Driver Mutations

- Non-coding variants may serve as drivers in many cancer types:
  - TERT, PLEKHS1, WDR74 and SDHD promoters
  - miRNA binding sites
- Our goal is to identity regions in the noncoding regions that are highly mutated

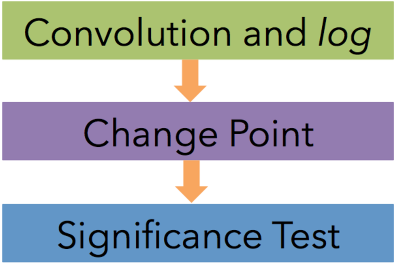
Previous Efforts

- Two papers
  - Weinhold, N. *et al.* Genome-wide analysis of noncoding regulatory mutations in cancer. *Nature Genetics*
  - Melton, C. *et al.* Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nature Genetics*
- Drawback
  - Annotations - low genome coverage
  - Small Fixed Regions - low mutation rate resolution
  - Not dynamic, not true region



Annotation Free Analysis

- Goal:
  - Auto-cluster genome into regions of enriched mutations
- 3 Steps:



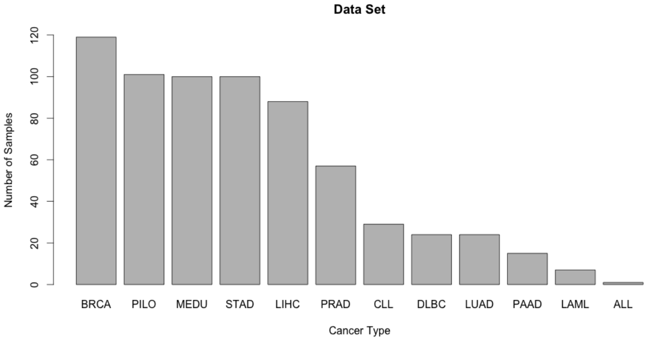
Convolution and *log*

Change Point

Significance Test

Dataset

- Somatic Mutations from:
  - 12 Cancer Types
  - 665 WGS total
- Includes *Alexandrov et al* data (WGS)

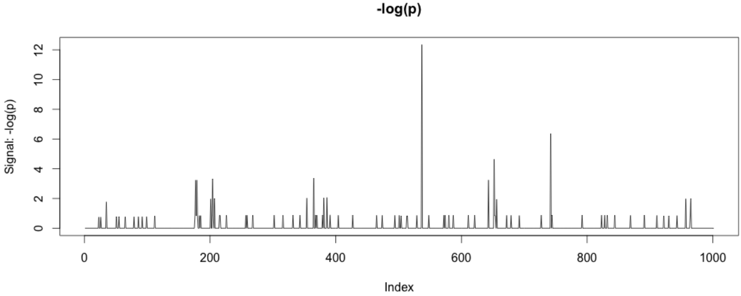


Convolution and $\log$

- Divide genome into 50bp bins
- Number of mutations in bin, k
- For a single cancer type $\sim$ Binomial
- Convolution Method: Combine discrete probabilities over all cancer types
  - $\Pr(K \geq k) = 1 - \Pr(K < k)$
  - linear combination of discrete probabilities
  - Result: single p-value for each 50bp bin

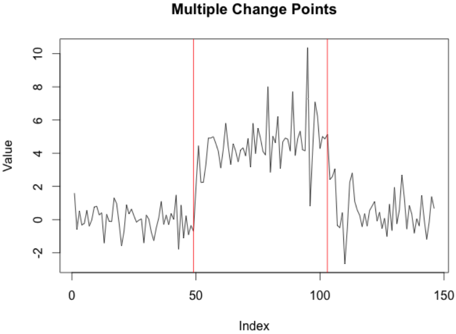
Convolution and $\log$

- For each p-value, take negative log ($-\log$)
- Creates signal for each 50bp bin, correlating to significance
- Pros
  - Amplify significant mutation count signal
  - Reduce signals that are less significant
  - Removes some noise found in mutation counts



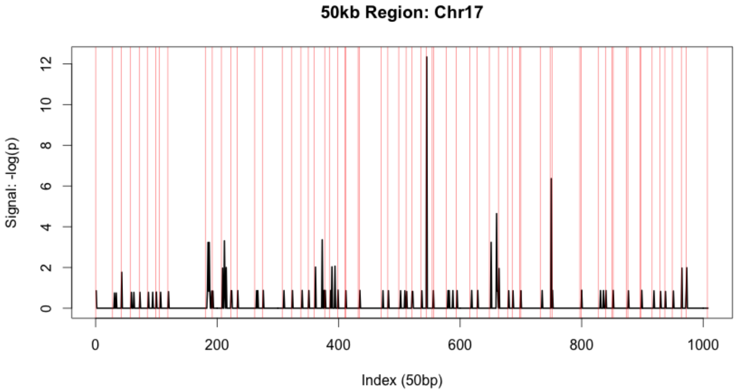
Change Point Detection

- Motivation:
  - Change points: determine start and end of region of interest
- Change in distribution before and after point
- Series of change points can be detected

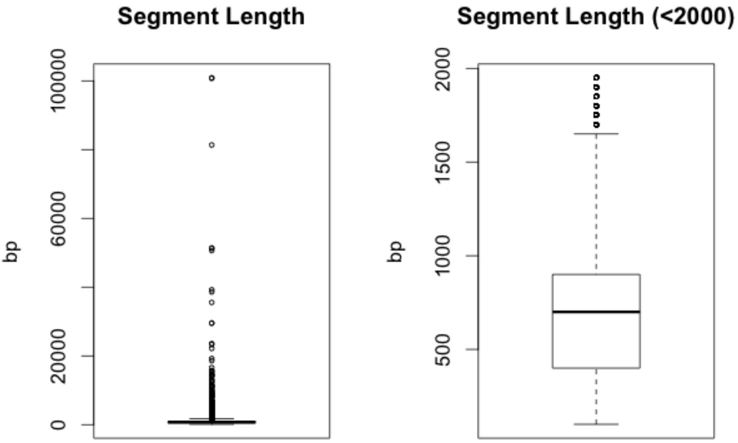


Change Point Detection: Usage

- p-values $\sim$ Uniform
- $-\log(p) \sim$ Exponential
- Apply change point algorithm to dataset of $-\log(p)$ for whole genome
- Example Result:

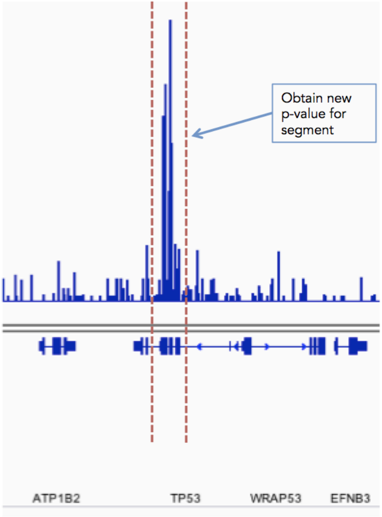


Change Point Segment Lengths

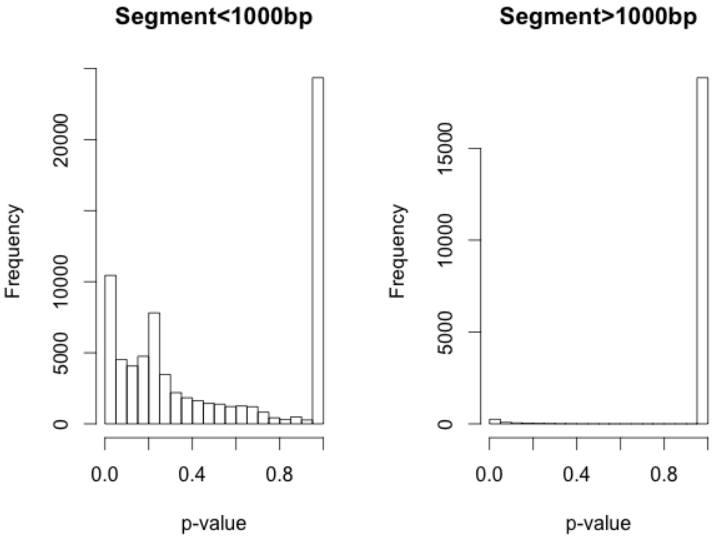


Statistical Testing

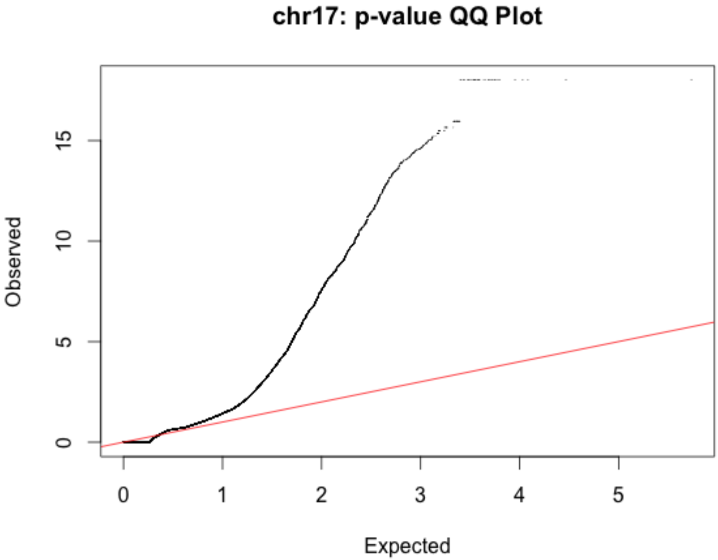
- Statistical testing on each segment
- Assess significance of segments determined by change point
- New p-value for segment (Convolution Method)



Segment P-values



Preliminary Results



Further Analysis

- Perform FDR or other p-value correction
    - Filter for significant segments
- Intersect significant segments with annotations
    - Expectation:
        - Intersections with known regulatory elements
        - Regions not contained in annotations, but also significant