Some PeakSeq Updates

Arif Harmanci

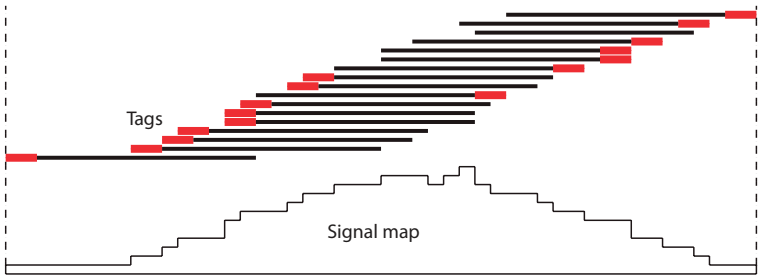
February 3rd, 2011

PeakSeq revisited

- PeakSeq is re-coded
- A little faster than previous C version
- Accepts BAM, ELAND, tagAlign formats
- Uses a configuration file to specify input
  - Easier to run than previous version
- Running on EBI for peak calling on Jan 2011 data freeze

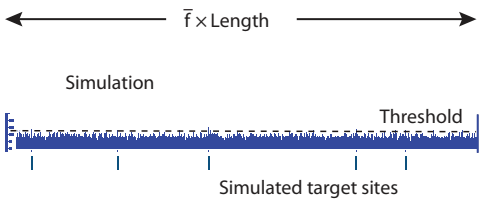
Satisfying the FDR in PeakSeq

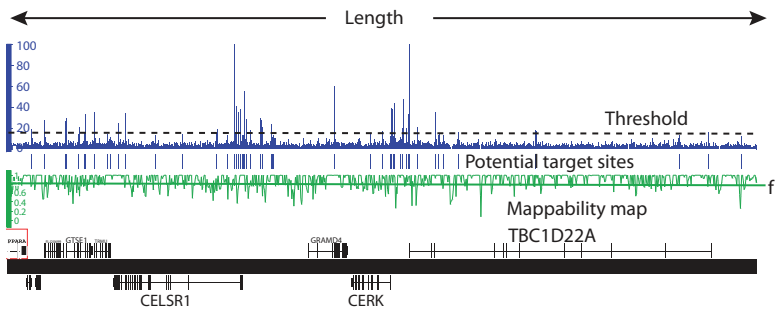
1. Constructing signal maps



- Extend mapped tags to DNA fragment
- Map of number of DNA fragments at each nucleotide position

2. First pass: determining potential binding regions by comparison to simulation





- Simulate each segment
- Determine a threshold satisfying the desired initial false discovery rate
- Use the threshold to identify potential target sites

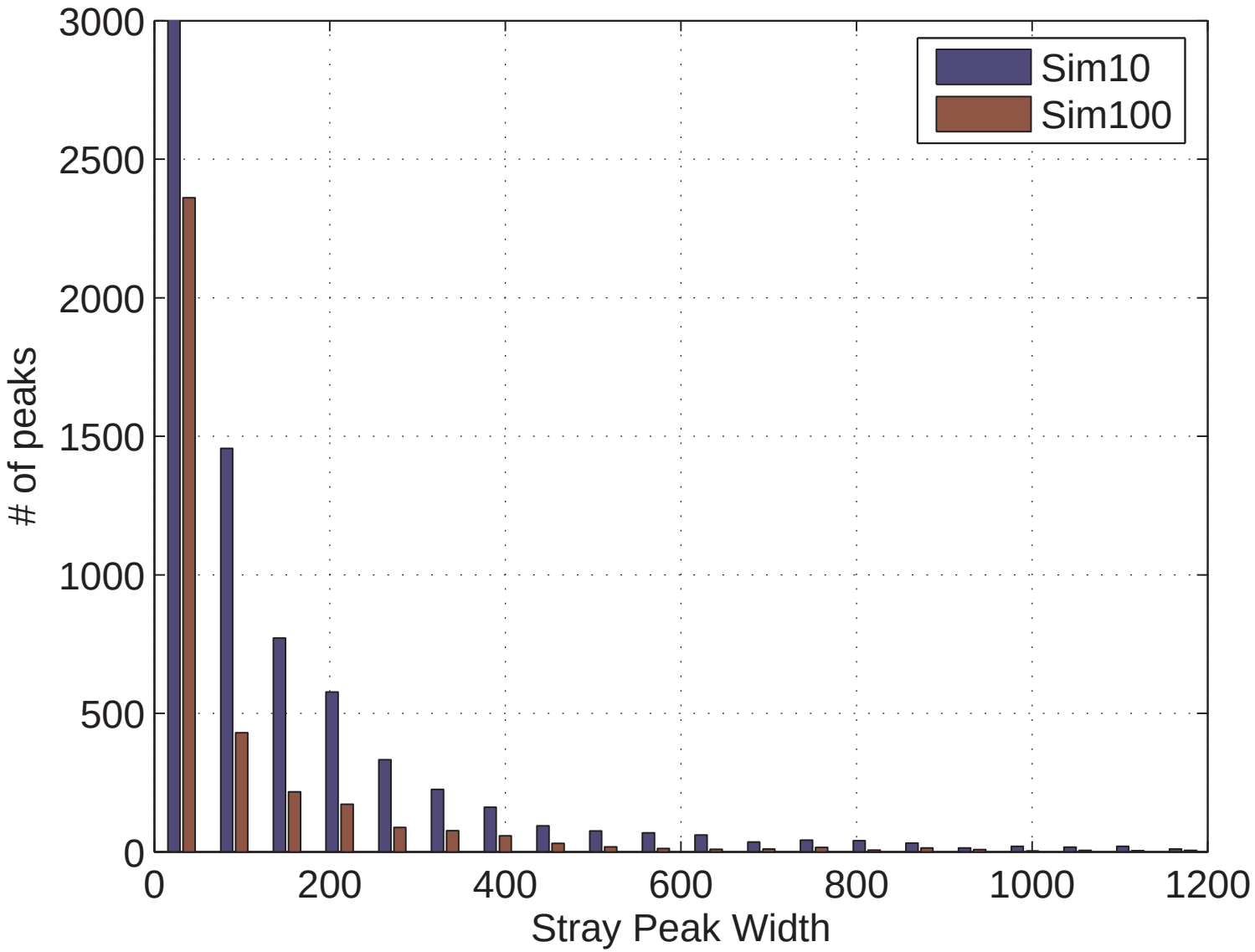
Simulations for Threshold Selection

- Simulate the random reads in each 1 Megabase window in the chromosome:
- For 10 times do:
  - Randomly generate r_{total} reads in the *uniquely mappable* window uniquely mappable window of n_{unique} nucleotides
    - r_{total}: Number of ChIP-Seq reads
    - n_{unique}: Number of uniquely mappable nucleotides in the window
  - For thresholds $[1 - 100]$ count the number of peaks, $n_{fp}(sim, threshold)$ for each threshold
- Take average of $n_{fp}(sim, threshold)$ over 10 sim's to get estimate of $n_{fp}(sim)$.
- What is the effect of simulations on the resulting peaks?

Effect of Simulations on Called Peaks

- Called the peaks for ChIP-Seq data for NA12878 maternal chromosome (part of AlleleSeq project):
  - Twice with 10 simulations: $peaks^1_{sim10}$, $peaks^2_{sim10}$
  - Twice with 100 simulations: $peaks^1_{sim100}$, $peaks^2_{sim100}$.
- Count the *stray* peaks: Count the peaks in $peaks^1_{sim10}$ that do not overlap with any other peak in the $peaks^2_{sim10}$ with at least 90% overlap:
  - 11925 peaks out of 107977 peaks, % (widths between 1 to 9676 base pairs)
    - Very long stray peaks, $\approx$ 30 kbases, pop up.
- Count the stray peaks for $peaks^1_{sim100}$ and $peaks^2_{sim100}$
  - 3594 peaks out of 107854 peaks, 3% (between 1 to 6000 base pairs long)

Histograms of Widths of Stray Peaks



Alternatives

- Get rid of simulations:
  - Cons:
    - No FDR control.
    - Too many peaks to process if there is no pre-selection of the peaks.
  - Pros:
    - Becomes much faster (at least twice)
    - How important is FDR control? The peaks are eventually scored by p-values

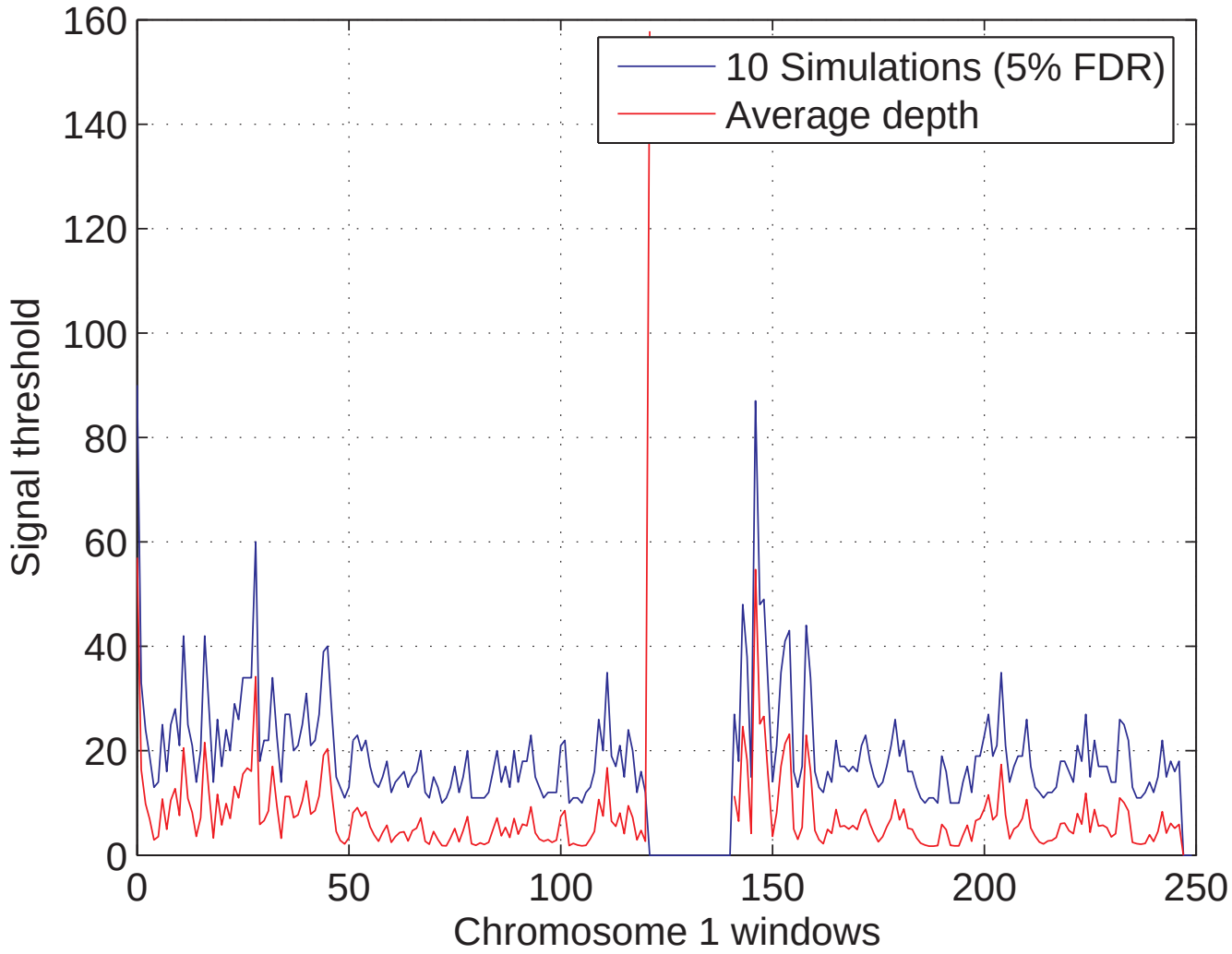
Alternative Threshold Selection

Use the average depth as a threshold:

$$thr(win) = \frac{r_{\text{total}} \cdot l_{fragment}}{n_{\text{unique}}} \tag{1}$$

- Average signal depth with random fragment generation

Compare with thresholds from simulations



Acknowledgements

- Mark Gerstein
- Joel Rozowsky
- Alexej Abyzov
- ...