

# BrainSeq ncRNA Prediction Overview

Arif Harmanci, Andrea Sboner,  
John Lu

# Overview

- Divided hg19-mm9 alignment into 150 bp windows with 75 bp overlaps
- 13,436,780 windows are generated, covering 1,815,529,081 base pairs
- The training set:
  - Get the expressed windows:
    - Threshold the maximum expression (RPM) at a window over all samples and regions
  - Overlap the expressed windows with Gencode v3c annotations and remove overlapping windows for
    - ncRNAs (50% overlap with window): 1037 windows
    - CDSs (95% overlap with window): 125786 windows
    - UTRs (95% overlap with window): 264647 windows