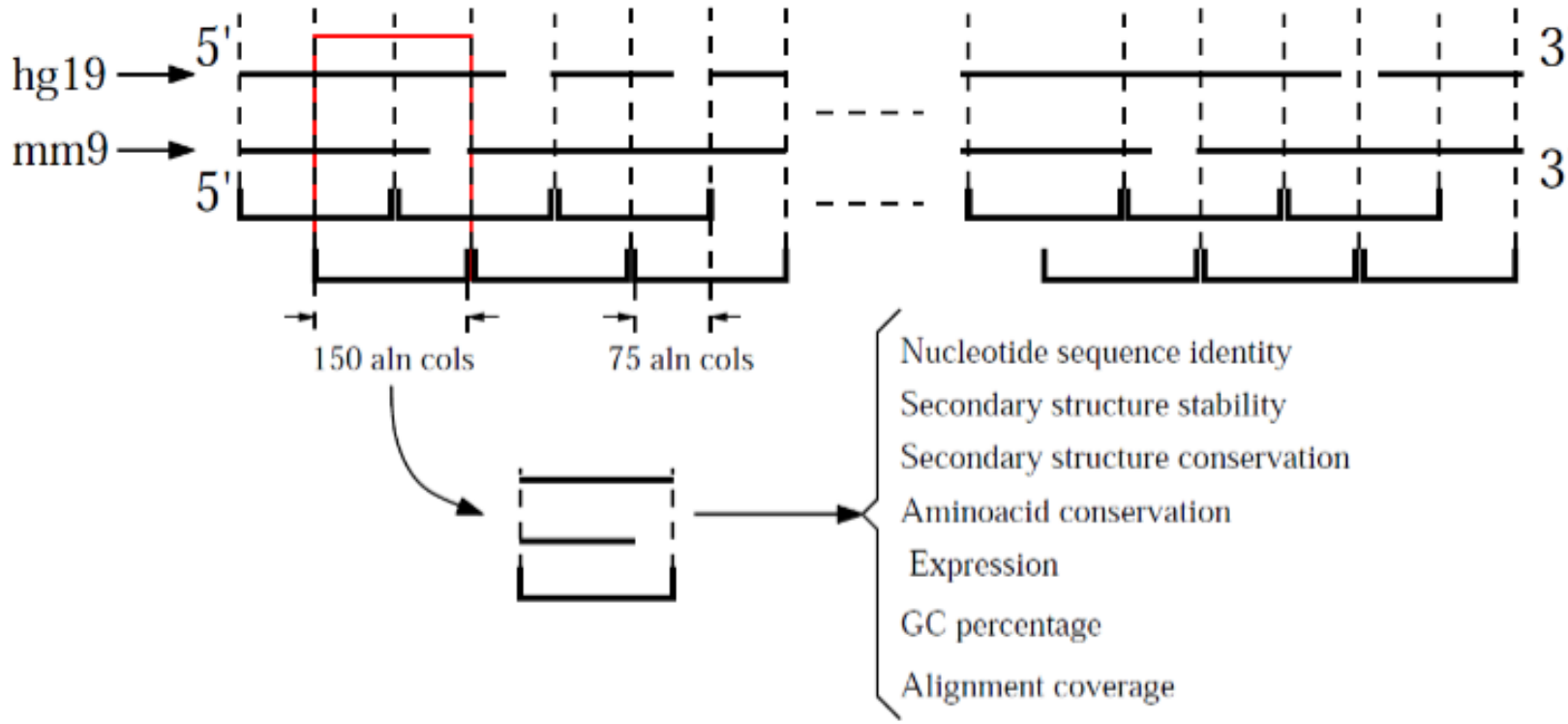
BrainSpan ncRNAs Update

Arif Harmanci, Andrea Sboner,
John Lu

Outline

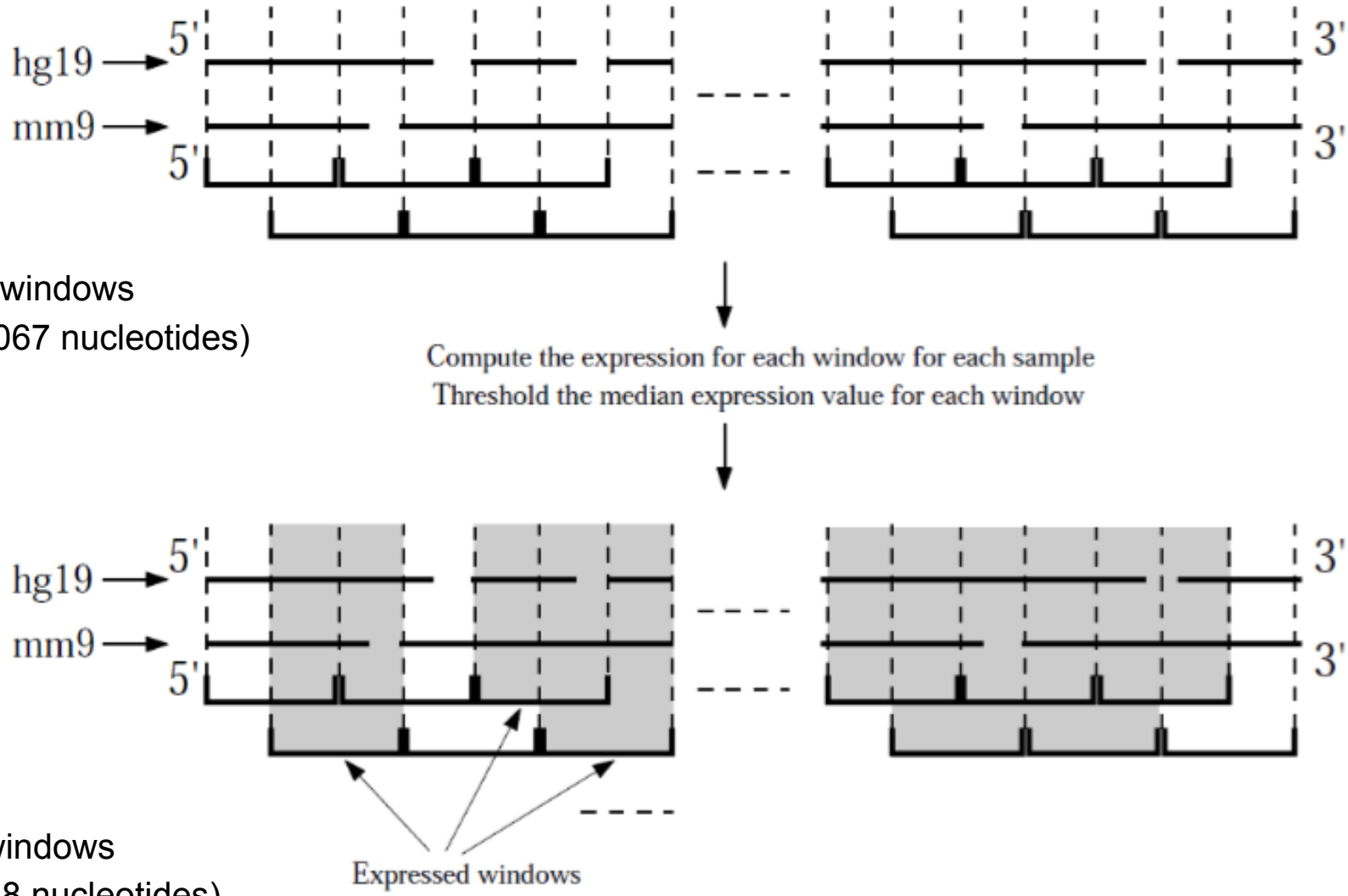
- Aim:
  - Identify novel ncRNAs
  - Characterize the novel ncRNAs and identify candidates that are differentially expressed between different regions/individuals
- Approach:
  - incRNA
    - Classification of ncRNAs, CDSs, UTRs
      - Sequence and structure homology based features

Window Generation



- Parse Human-Mouse alignment into windows of 150 alignment columns with steps of 75 columns
- Remove windows that contain less than 75 nucleotides for one species
- Each window is characterized by 7 different features

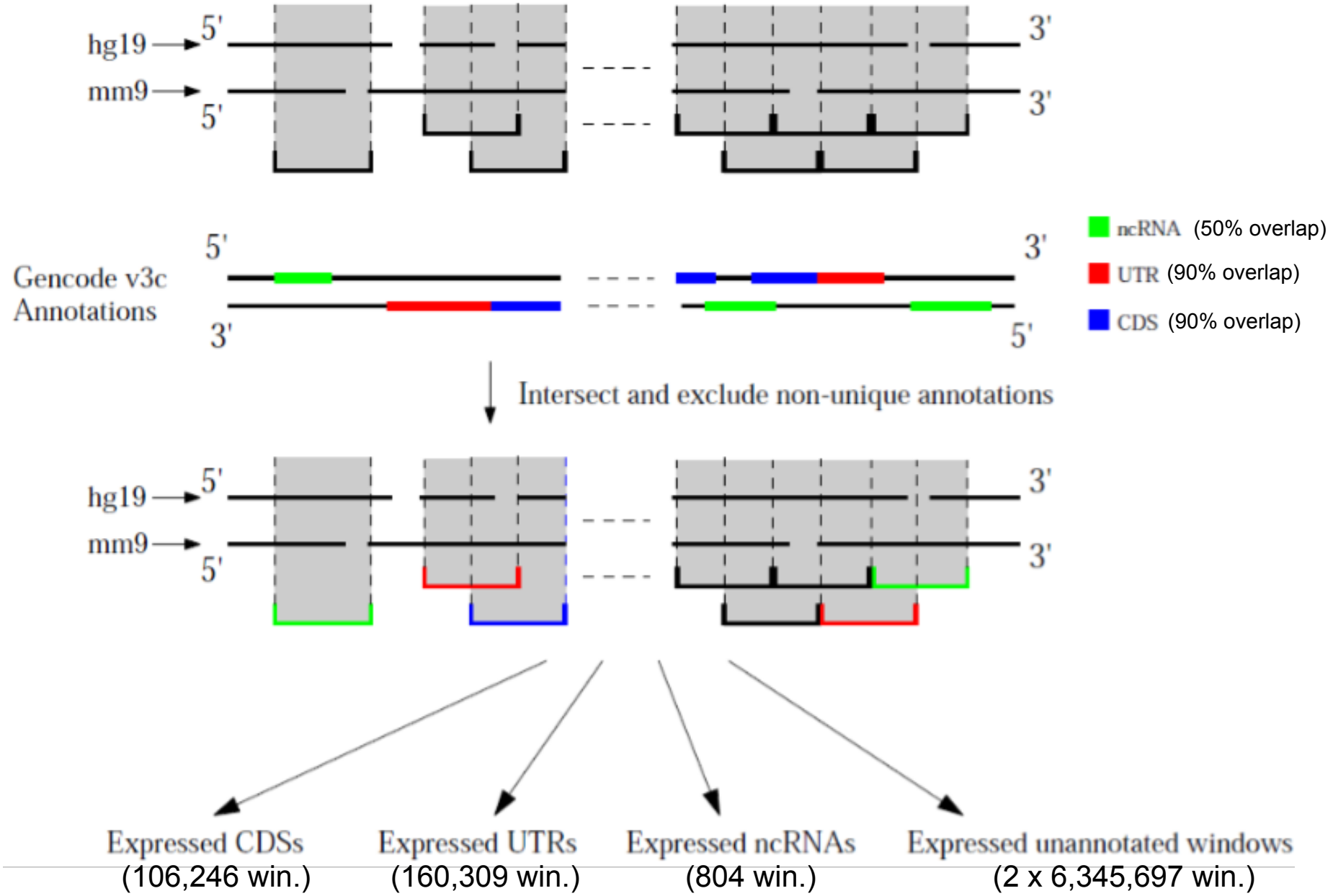
Identify the Expressed Windows



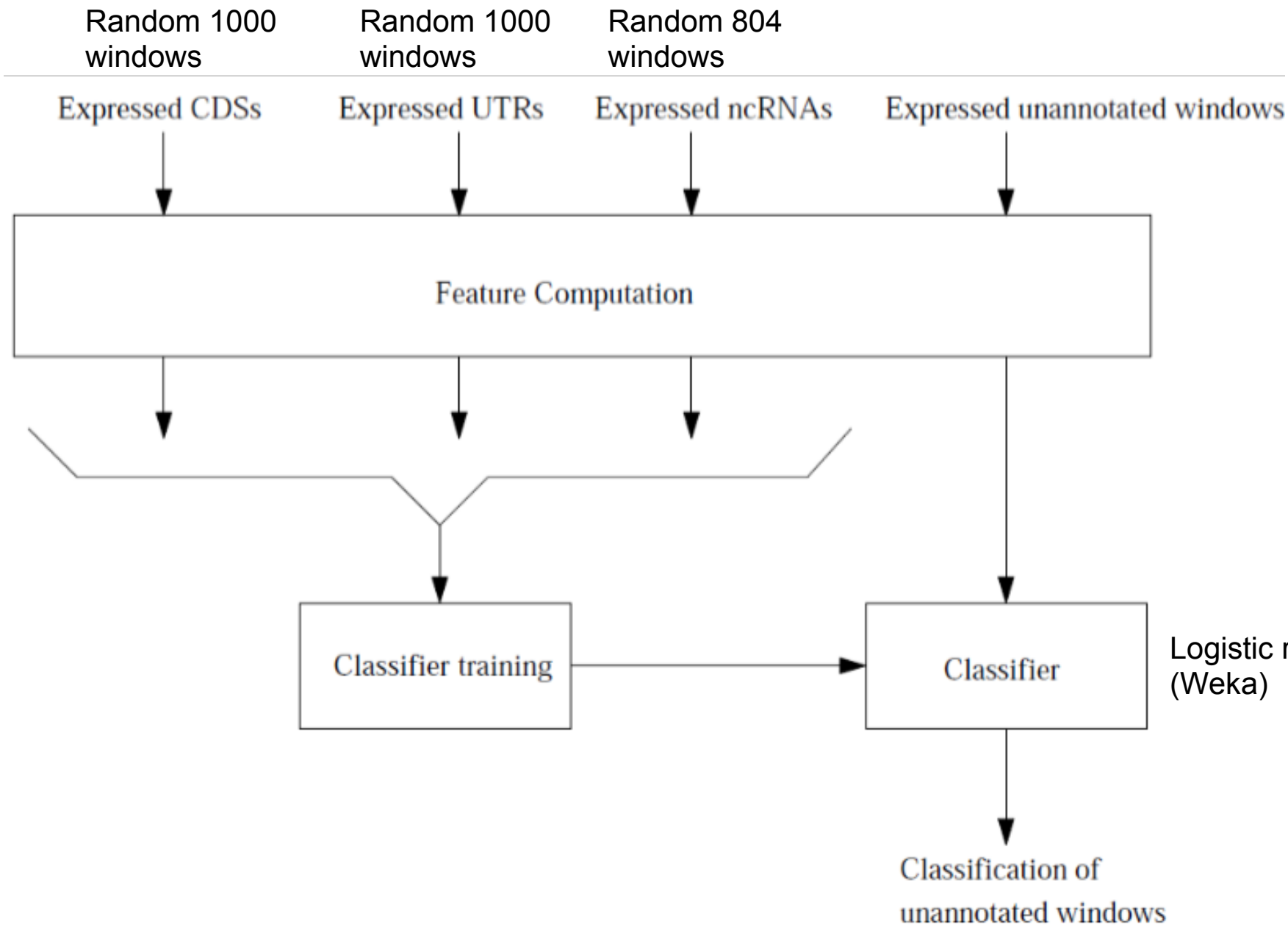
13,882,553 windows
(1,063,222,067 nucleotides)

8,328,173 windows
(684,621,038 nucleotides)

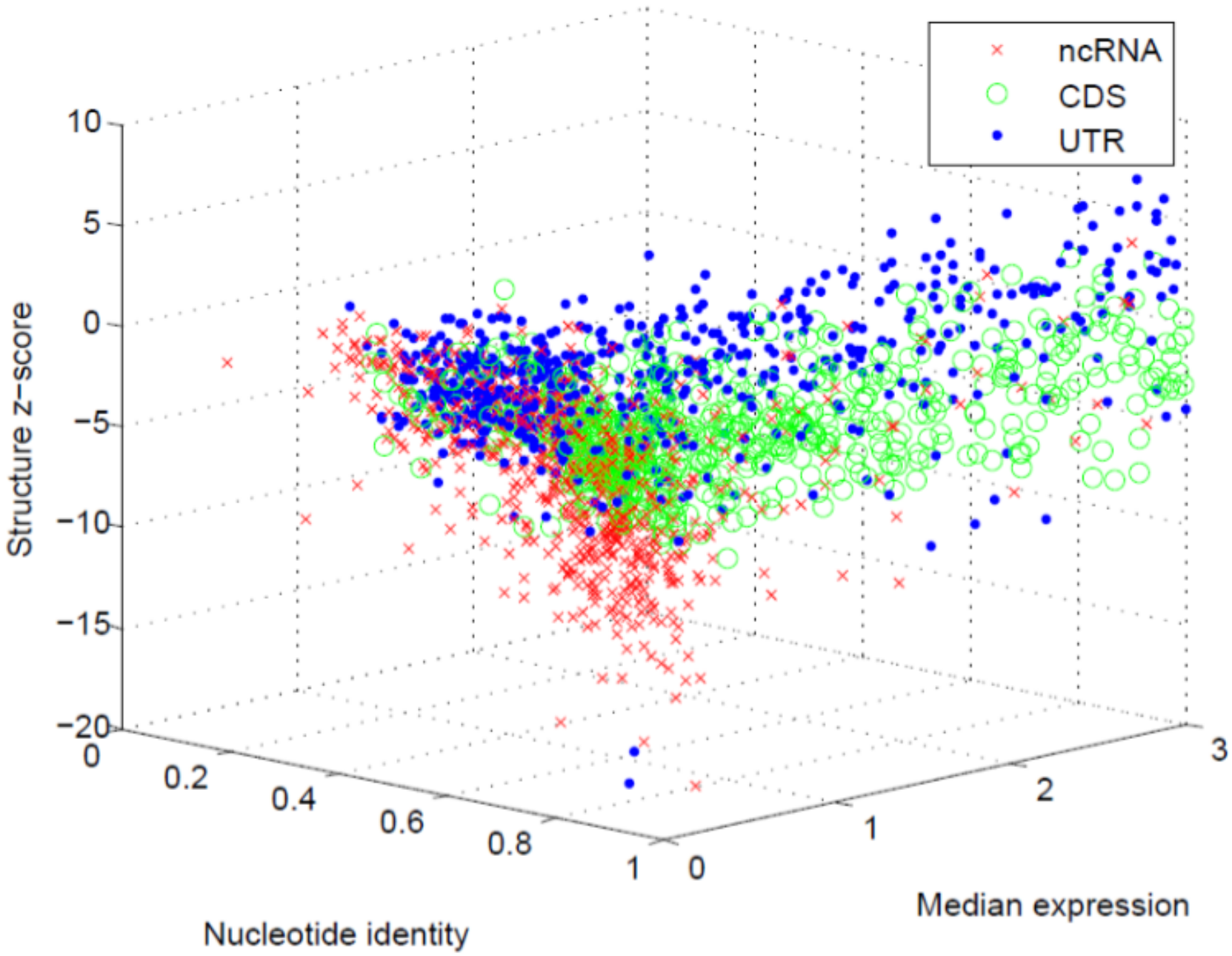
Identify Annotated and Unannotated Expressed Windows



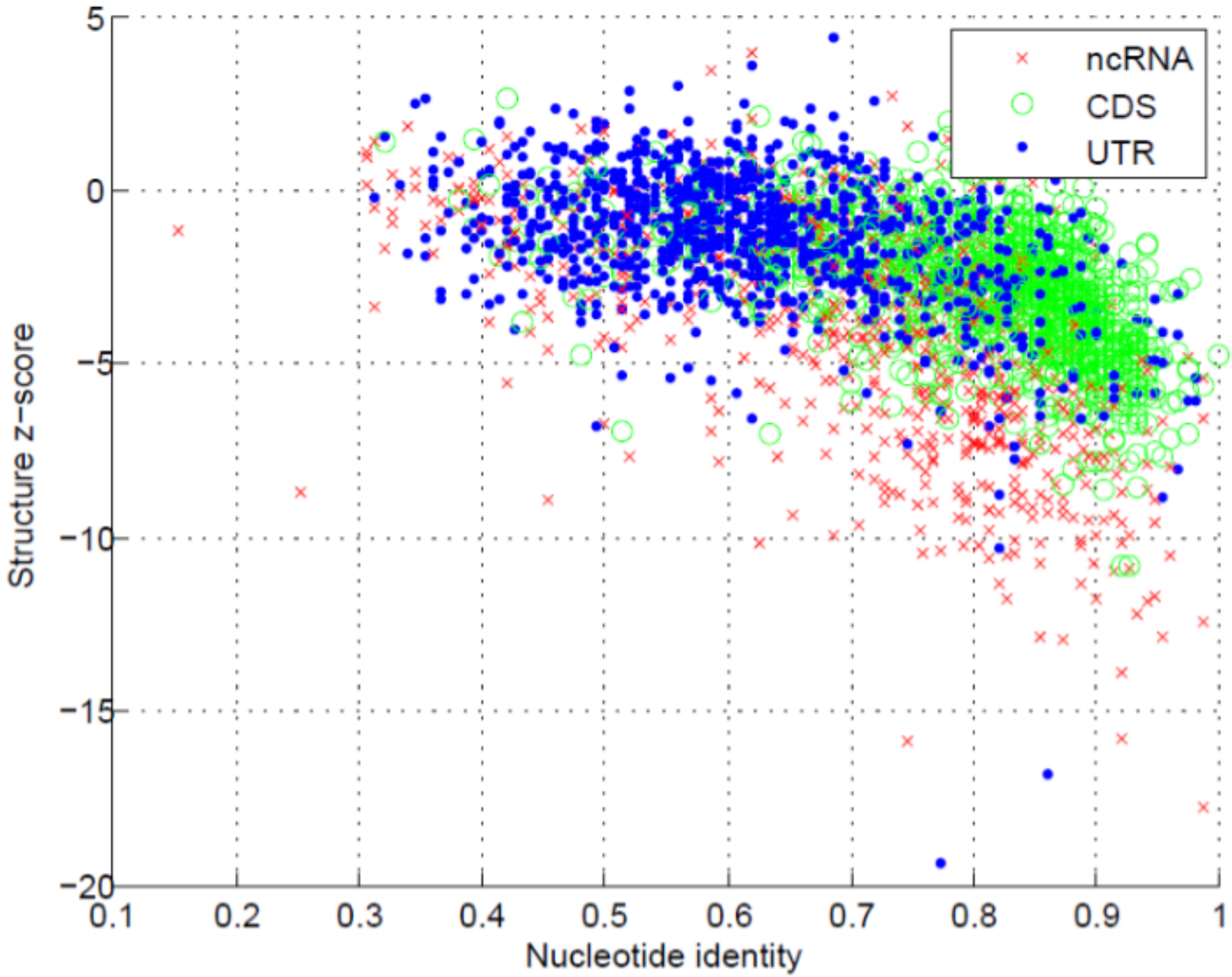
Classification of Unannotated windows



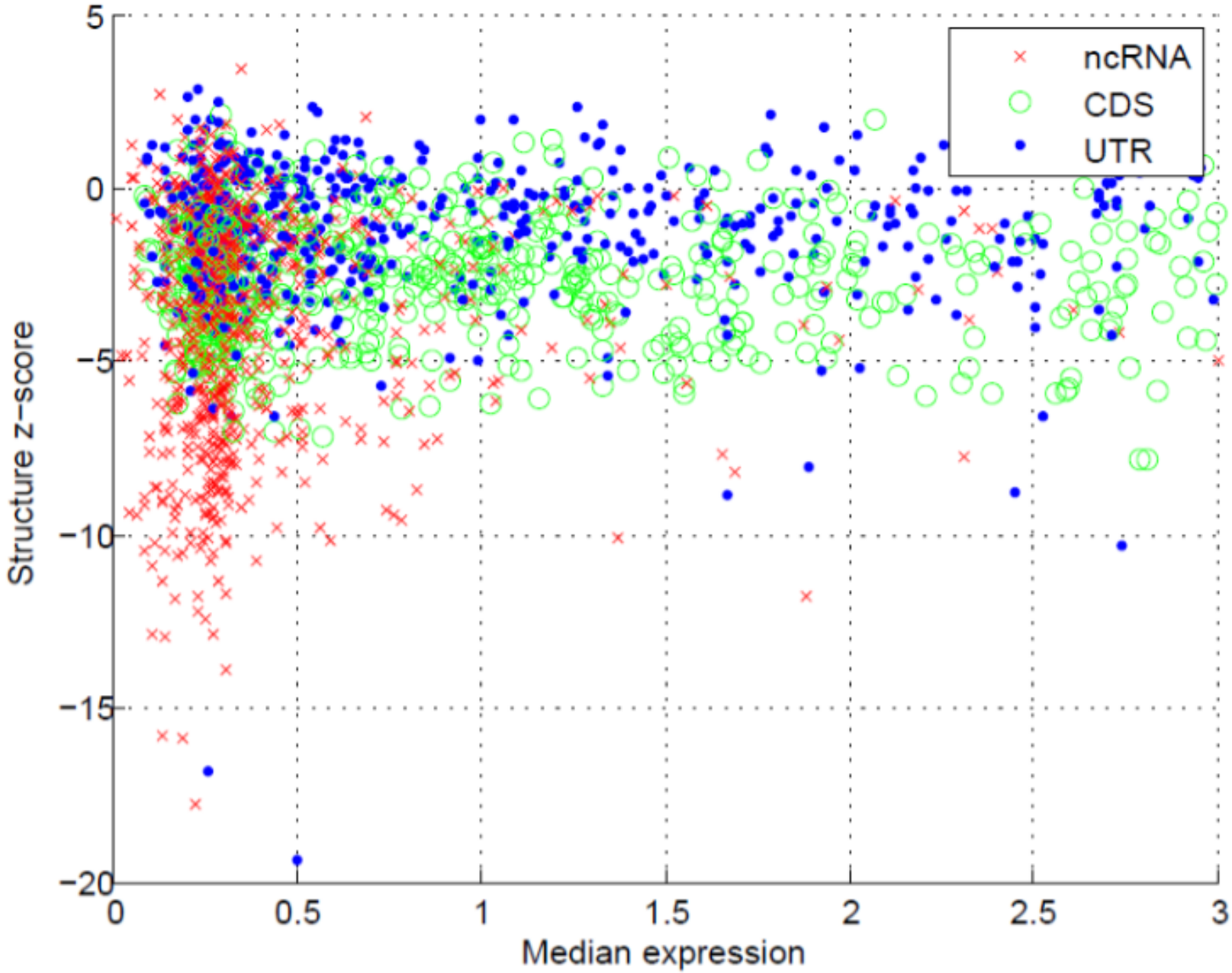
Distribution of Features



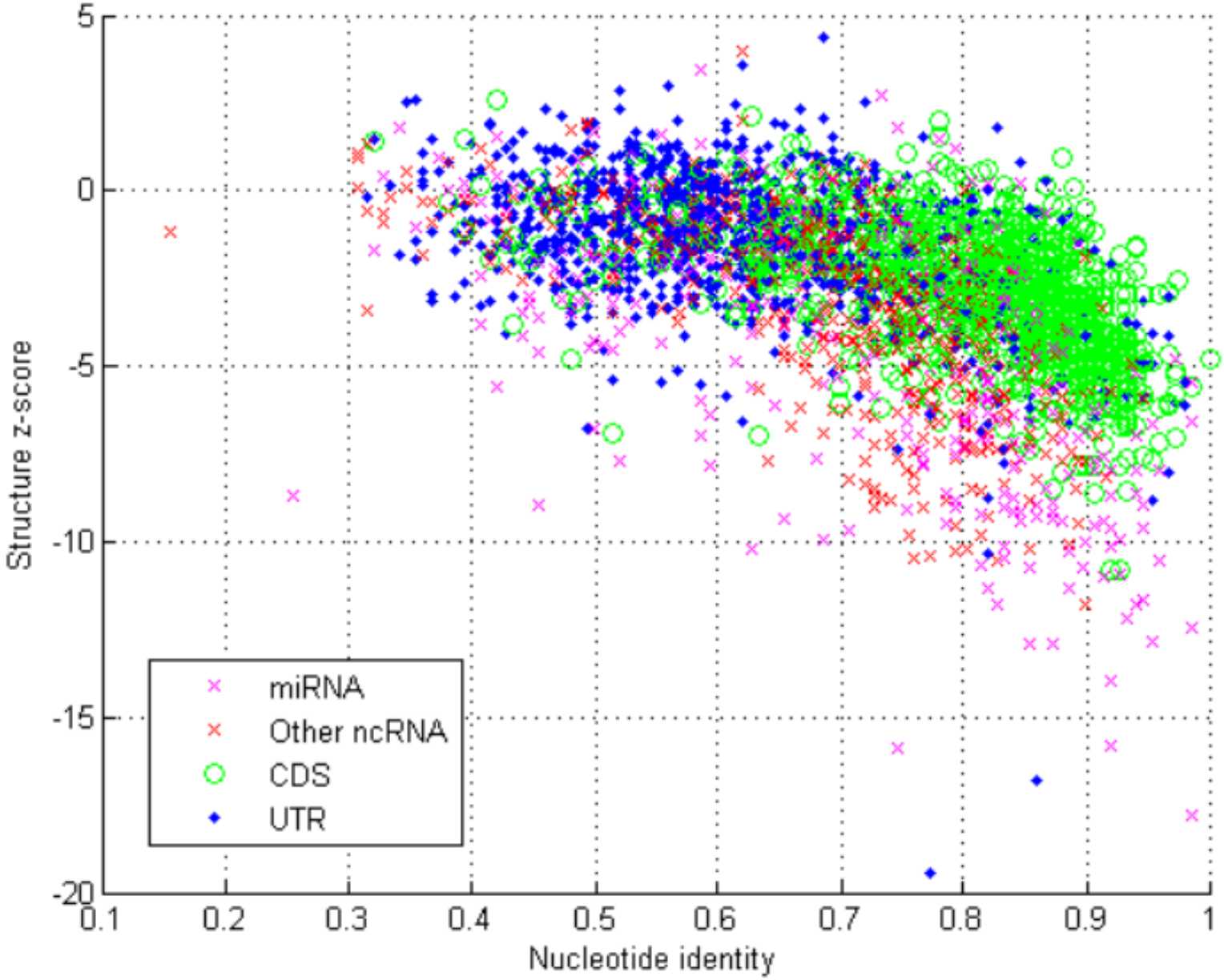
Distribution of Features



Distribution of Features



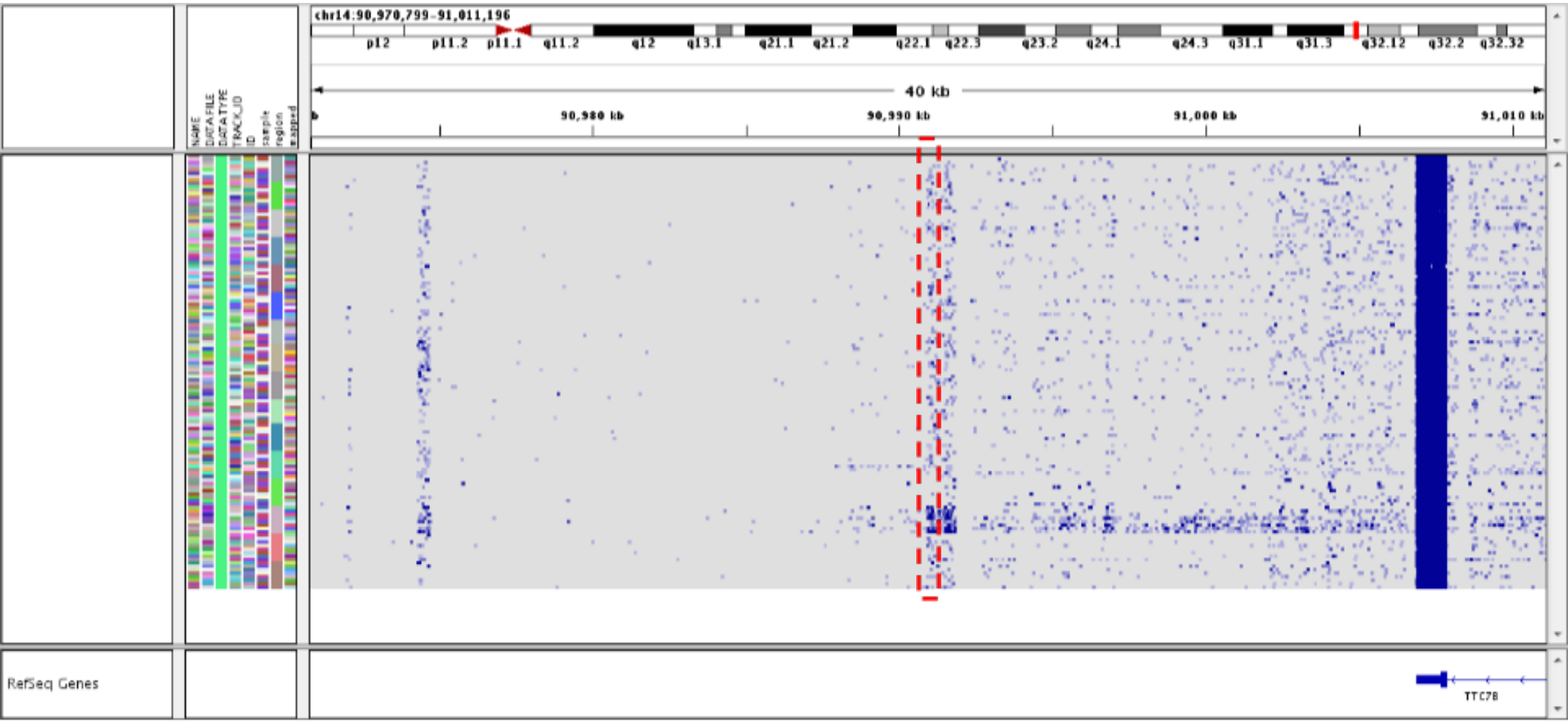
Distribution of Features



Training, Testing

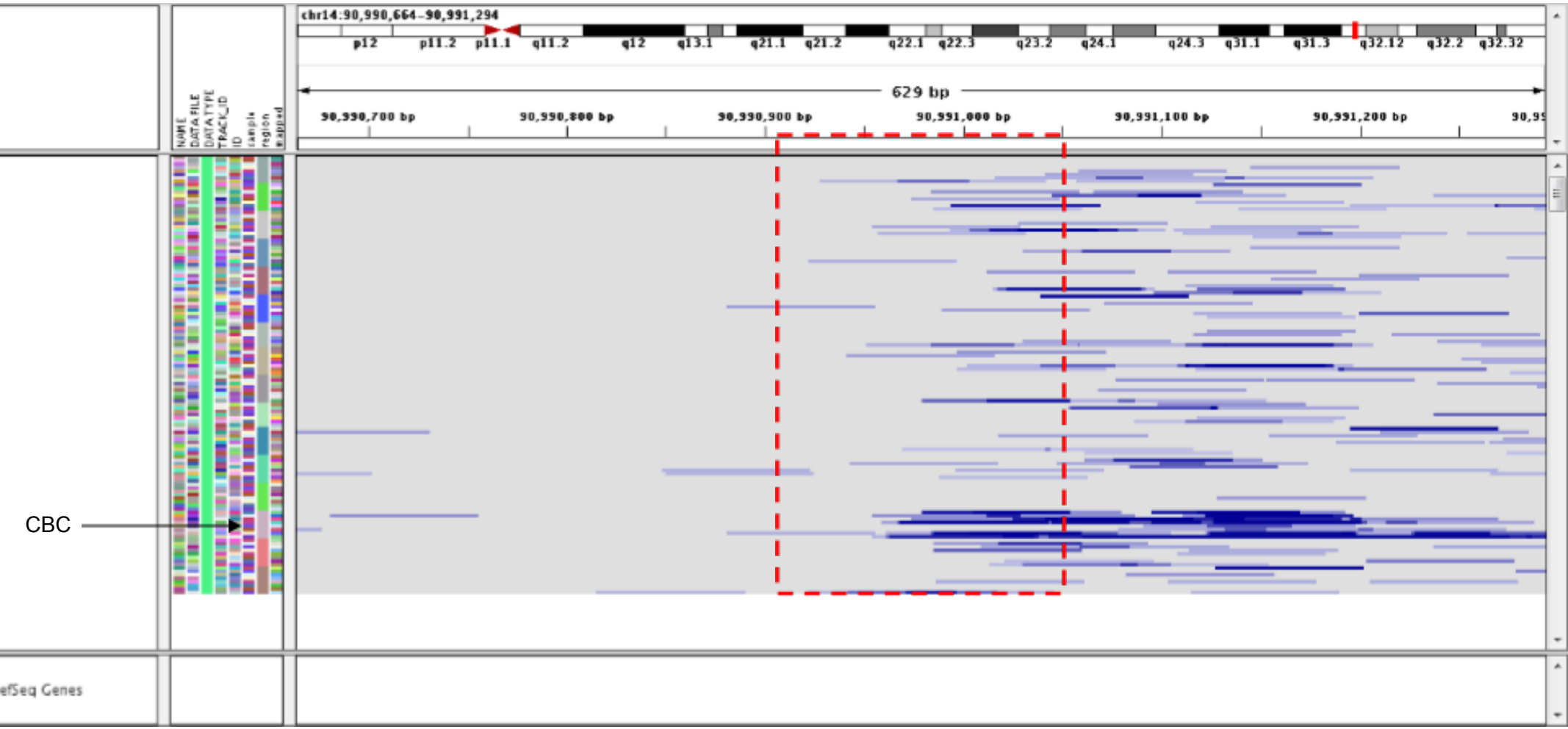
- Trained logistic regression classifier in Weka
- ~70% accuracy in 10-fold cross-validation
- Applied the classifier to all of the remaining windows in all the chromosomes
- Results can be downloaded from BrainSeq wiki

Chr14:90990909-90991050



- Expression is not correlated with the coding region

Chr14:90990909-90991050

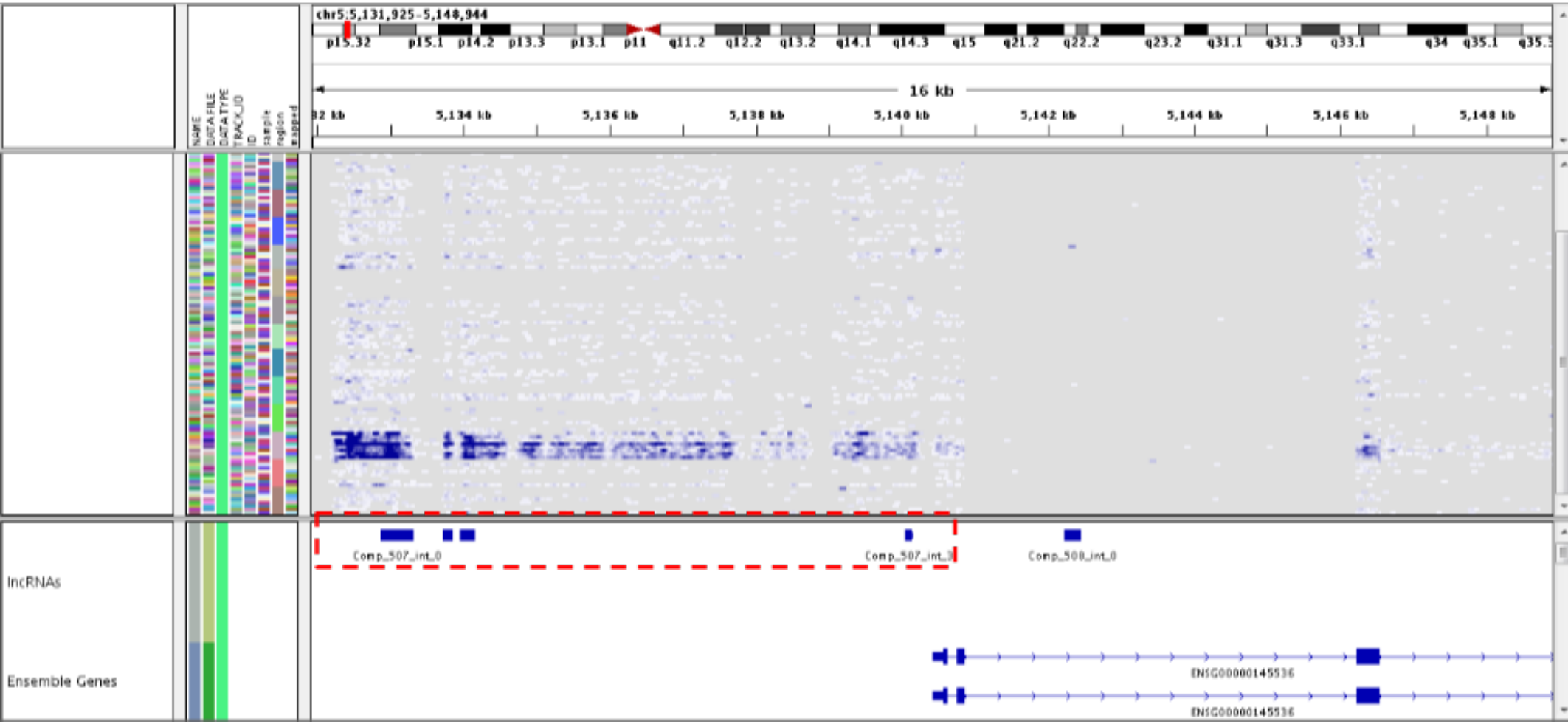


- Max signal value set to 0.0016

GENCODE v7 Long ncRNAs

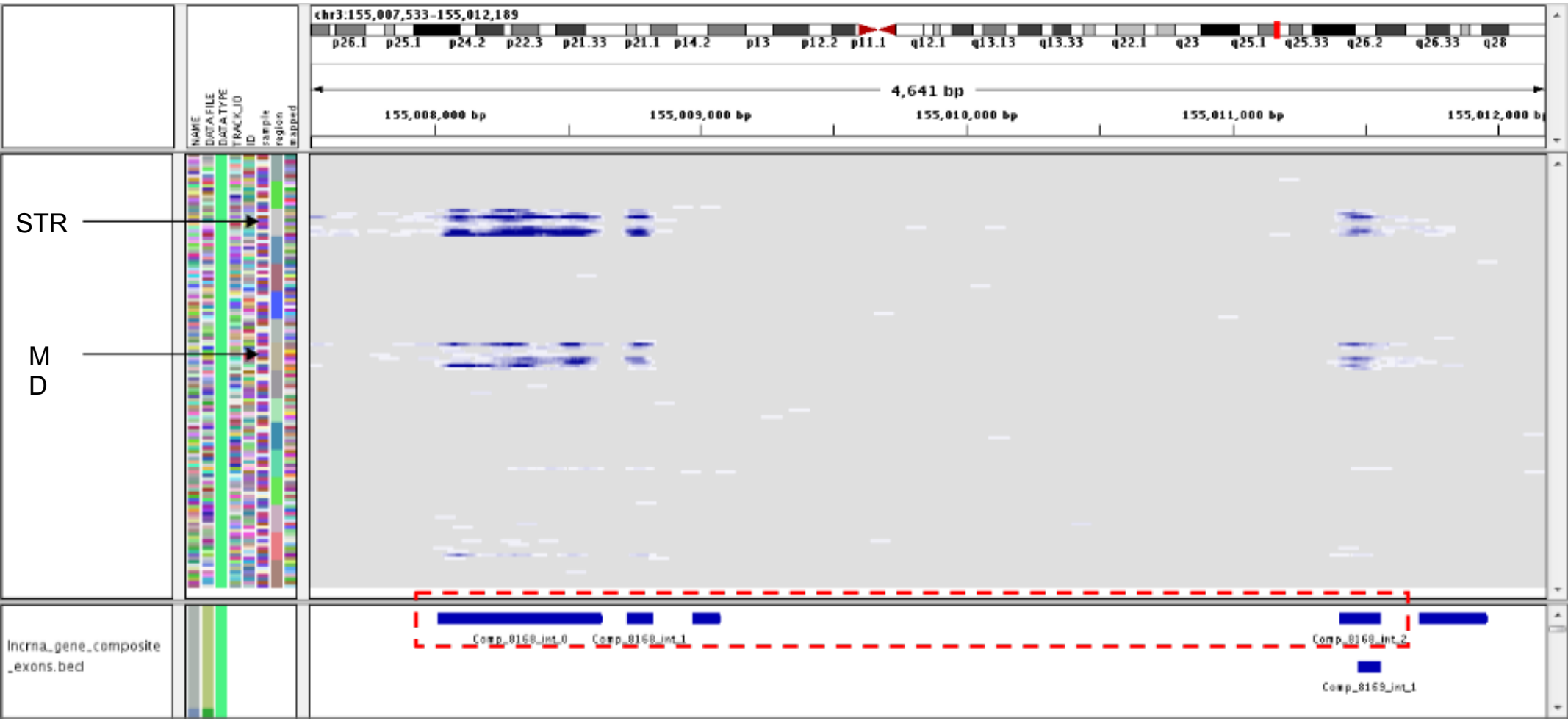
- Put together the processed transcript and lincRNA type entries
  - Includes antisense transcripts
- Total of ~9,400 entries
- Built the composite transcripts using RSeqTools
- Computed RPKMs over all samples/regions
- Did Wilcoxon test for identifying putative differentially expressed regions

GENCODE v7 Long ncRNAs



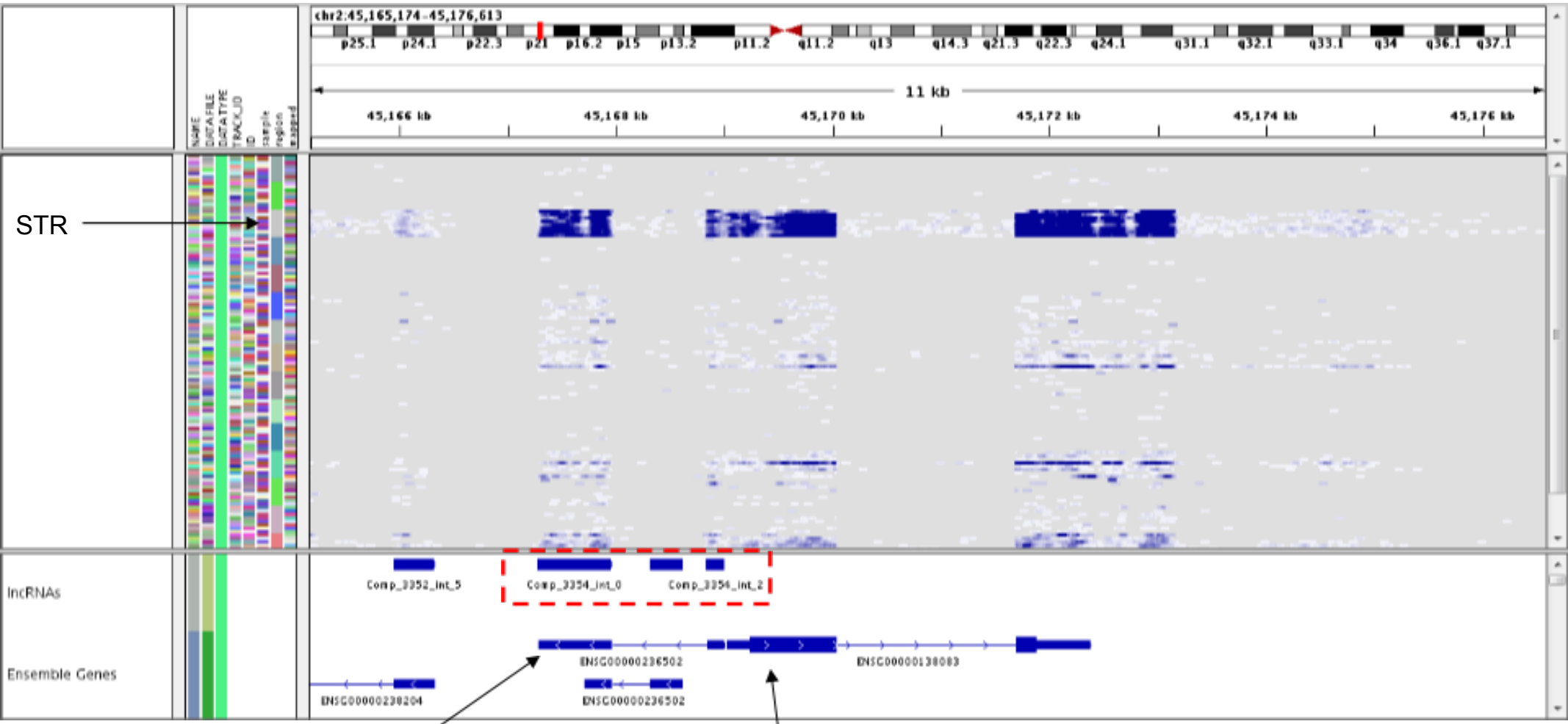
- ENSG00000250579 (CTD-2297D10.2)

GENCODE v7 Long ncRNAs



- ENSG00000240045 (RP11-451G4.2)

GENCODE v7 Long ncRNAs



- ENSG0000236502
- ENSG00000138083 (SIX3)