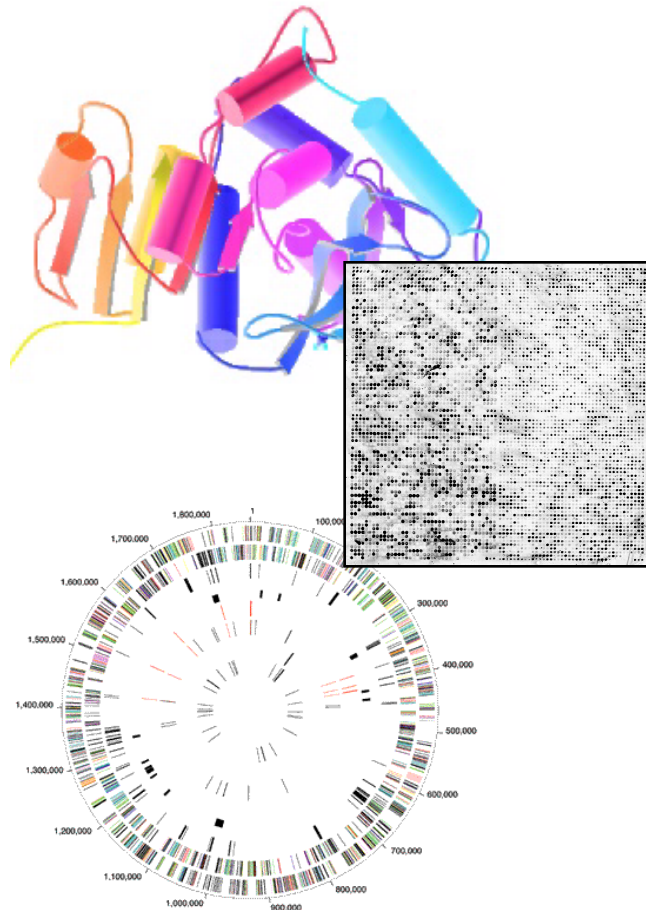


Biomed. Data Science:

# Quick RNA-seq & Chip-seq

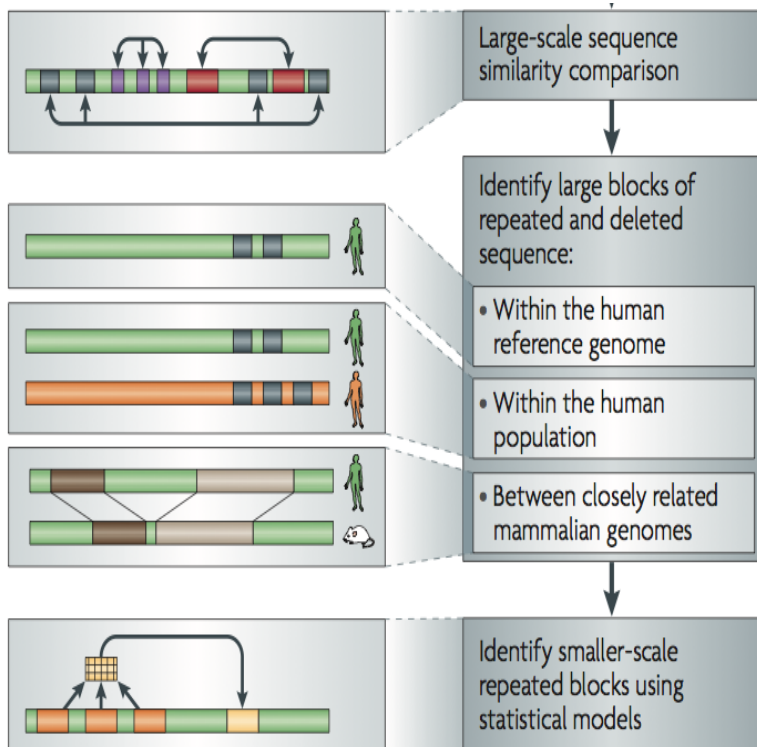


Mark Gerstein, Yale University  
[gersteinlab.org/courses/452](http://gersteinlab.org/courses/452)  
(last edit in spring '18)

# Non-coding Annotations: Overview

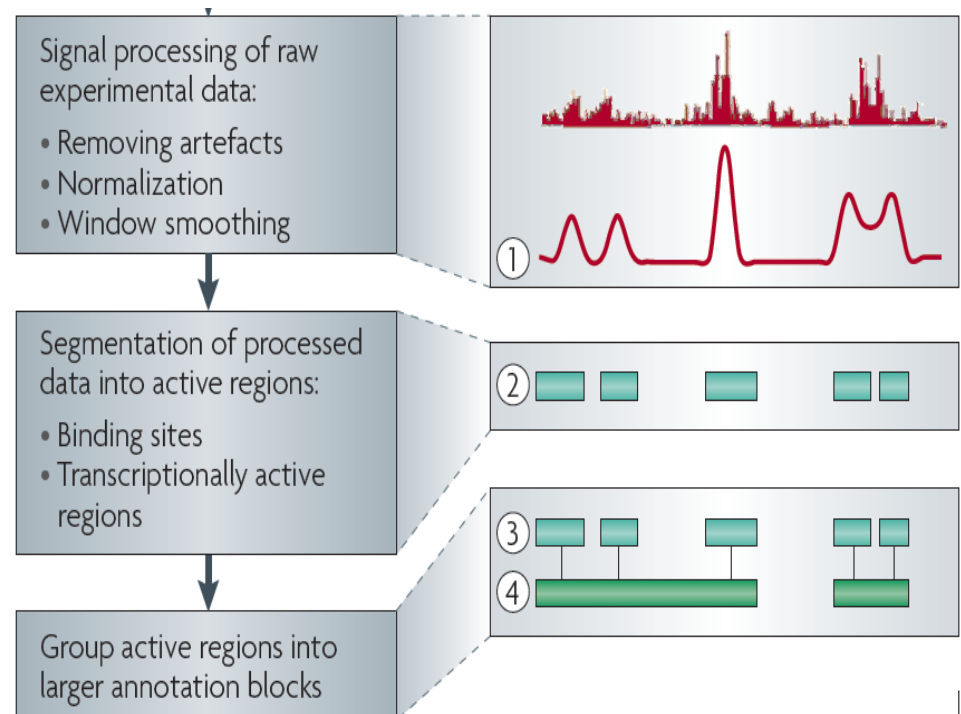
Features are often present on multiple "scale" (eg elements and connected networks)

## Sequence features, incl. Conservation



## Functional Genomics

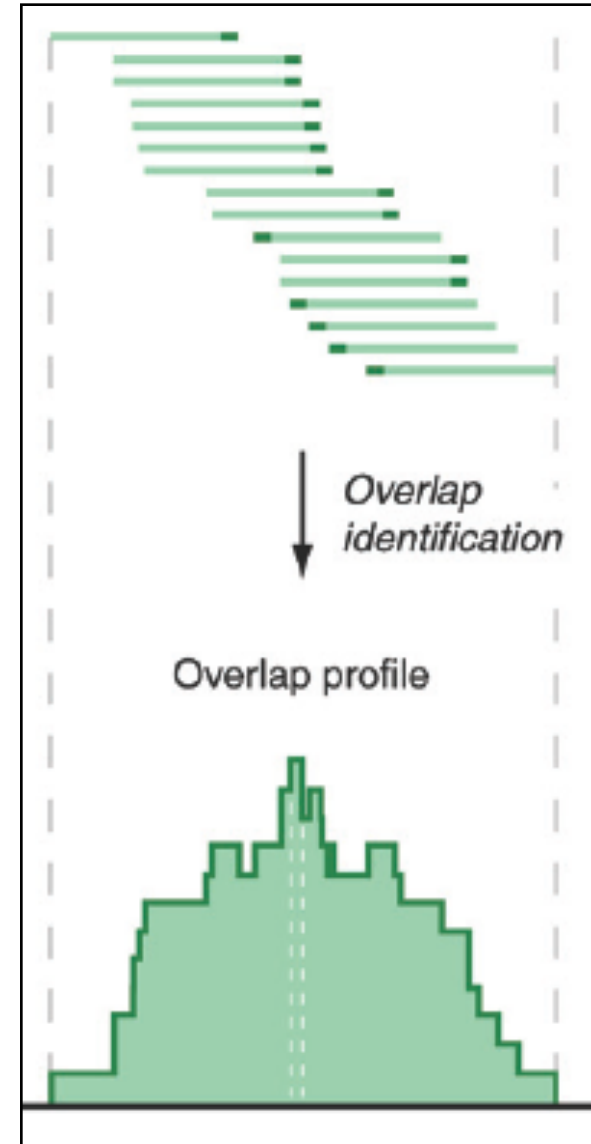
### Chip-seq (Epigenome & seq. specific TF) and ncRNA & un-annotated transcription



[Nat. Rev. Genet. (2010) 11: 559]

# Low-Level Data for RNA-seq & Chip-seq

```
@ILMN-GA001 3 208HWAAXX 1 1 110 812
ATACAAGCAAGTATAAGTTCGTATGCCGTCTT
+ILMN-GA001 3 208HWAAXX 1 1 110 812
hhhYhh]NYhhhhhhYIhhaZT[hYHNSPKXR
@ILMN-GA001 3 208HWAAXX 1 1 111 879
GGAGGCTGGAGTTGGGGACGTATGCGGCATAG
+ILMN-GA001 3 208HWAAXX 1 1 111 879
hSWhRNJ\hFhLdhVOhAIB@NFKD@PAB?N?
```



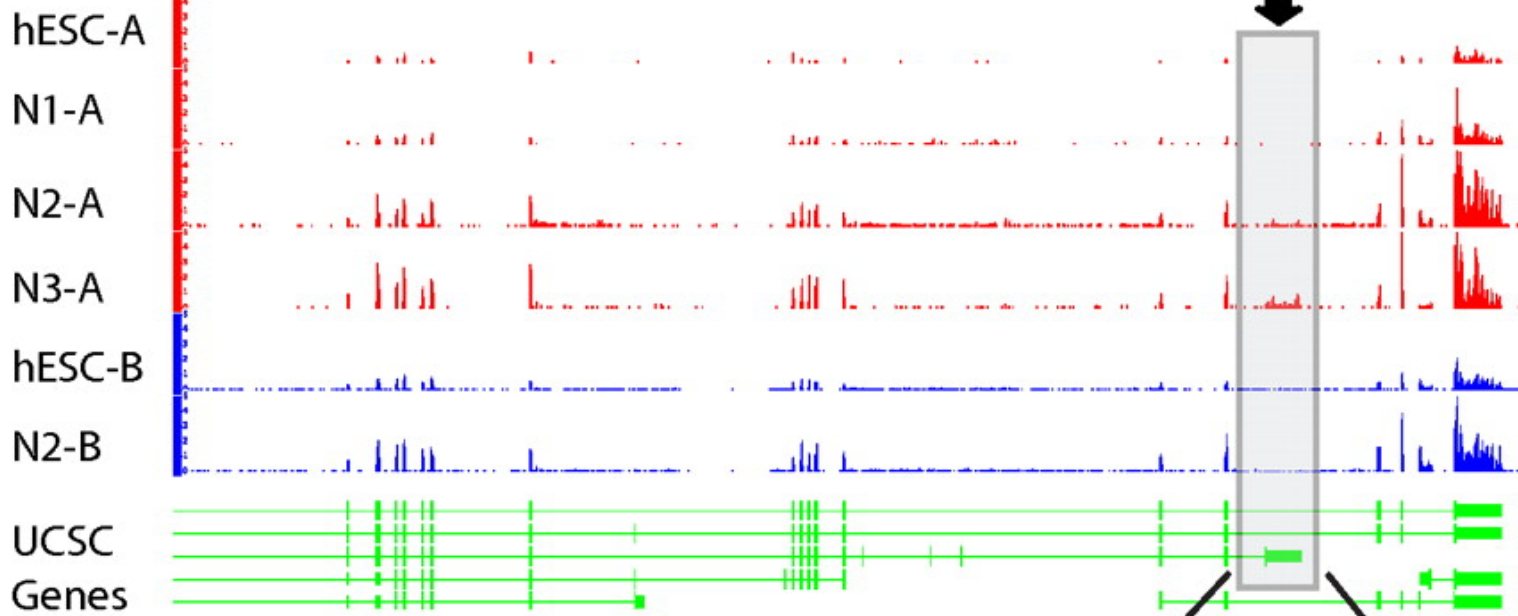
Reads (fasta)

+ quality scores (fastq)

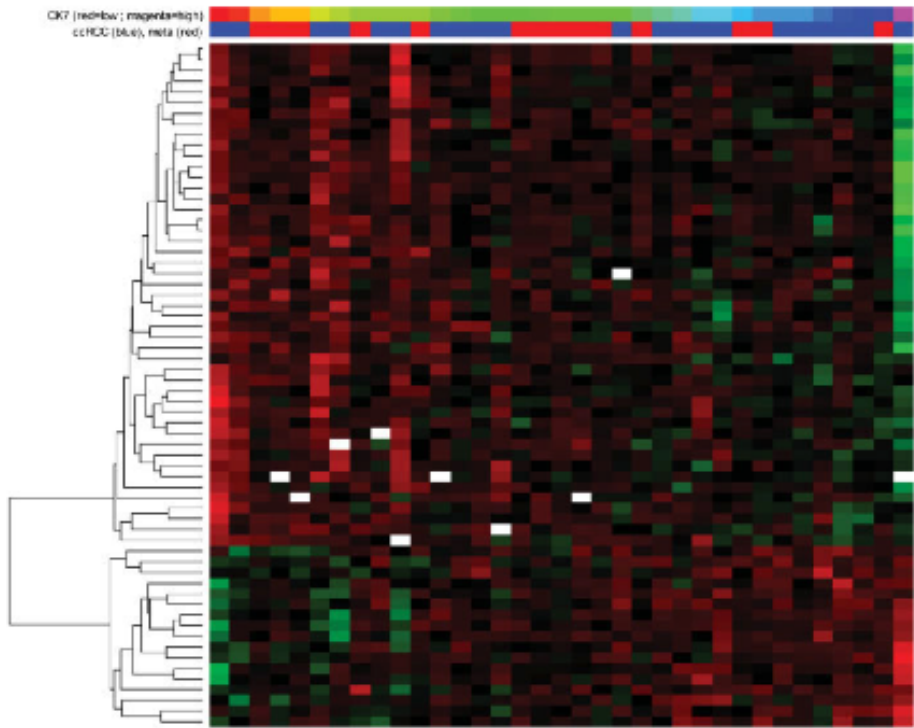
+ mapping (BAM)

Reads => Signal (Intermediate file)

Accumulating @ >1 Pbp/yr (currently),  
~20% of tot. HiSeq output



CAM1



**Information from  
RNA-seq:  
Avg. signal at exons &  
TARs (RPKMs)**

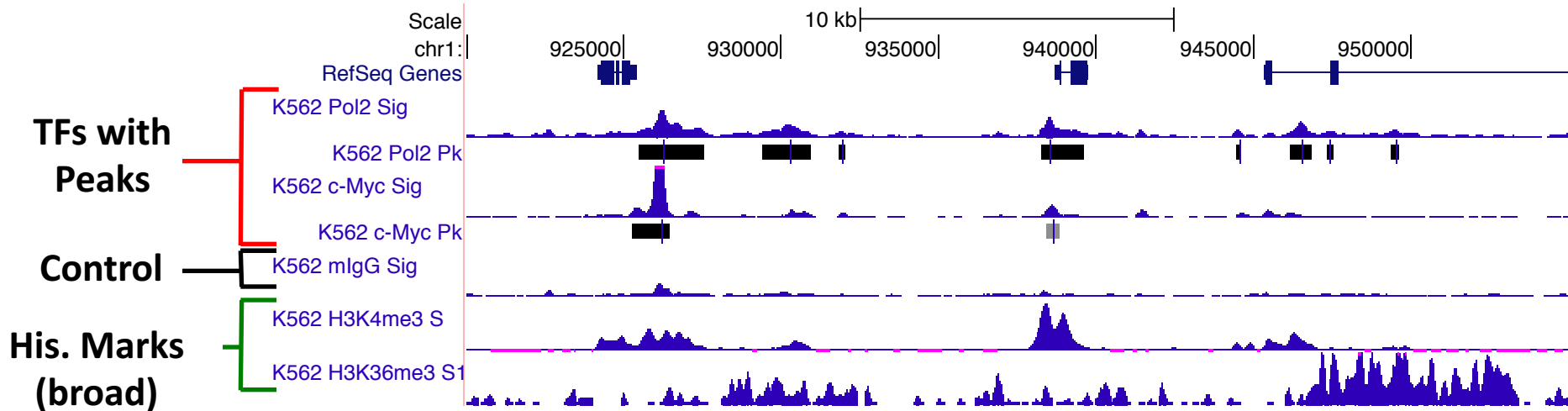
[PNAS 4:107: 5254 ; IJC 123:569]



## Activity Patterns

- RNA Seq. gives rise to activity patterns of genes & regions in the genome

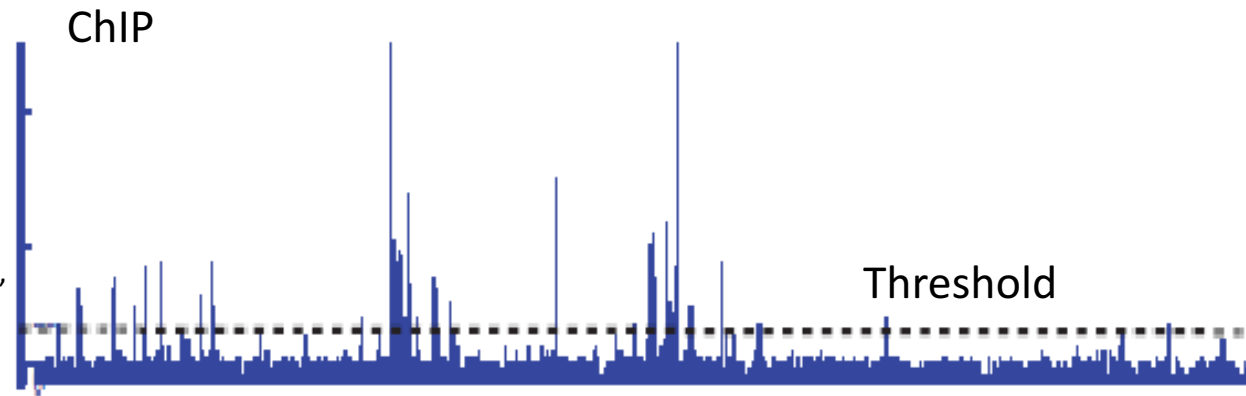
# Information from Chip-seq



[*Science* 330: 1775  
 + ENCODE Data  
 Sources  
 TFs & Control: Yale  
 HMs: UW & Broad ]

# Summarizing the Signal: "Traditional" ChipSeq Peak Calling

- Generate & threshold the signal profile to identify candidate target regions
  - Simulation (PeakSeq),
  - Local window based Poisson (MACS),
  - Fold change statistics (SPP)



Potential Targets



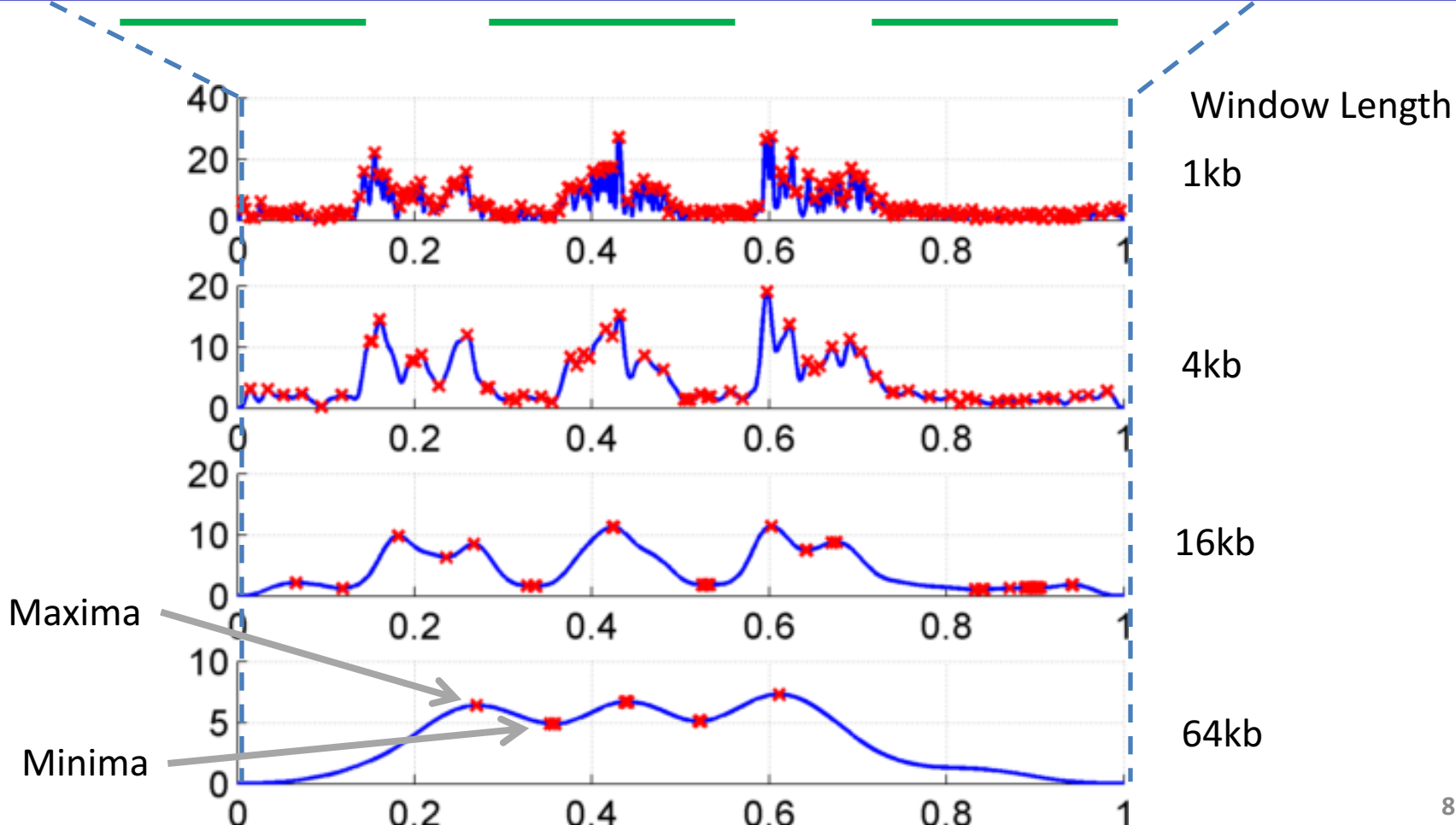
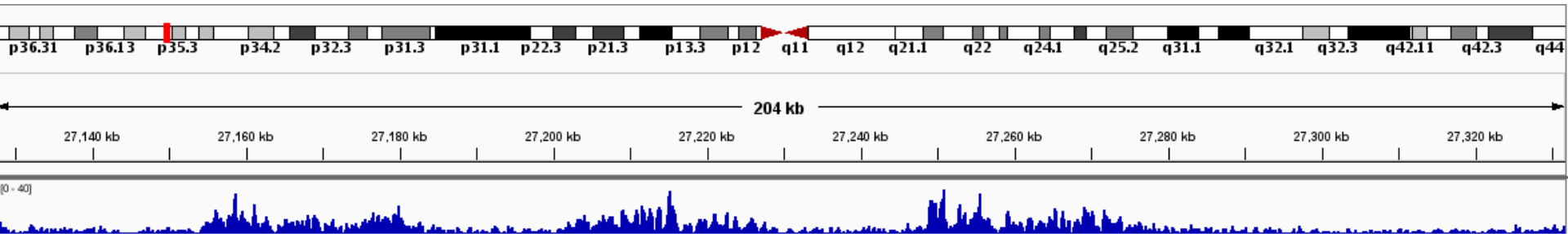
- Score against the control



Significantly Enriched targets

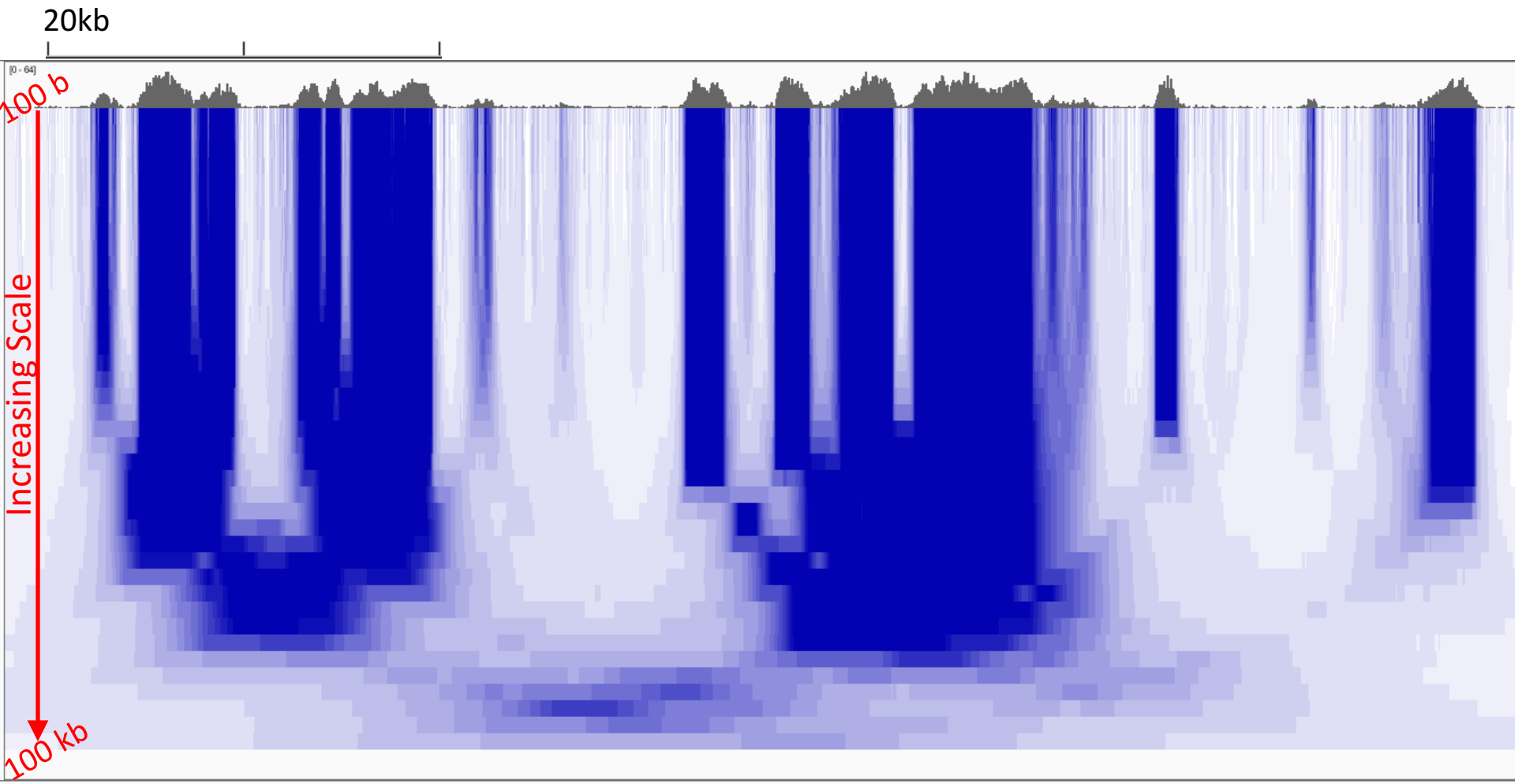


# Multiscale Analysis, Minima/Maxima based Coarse Segmentation





# Multiscale Decomposition



# Multiscale Decomposition

