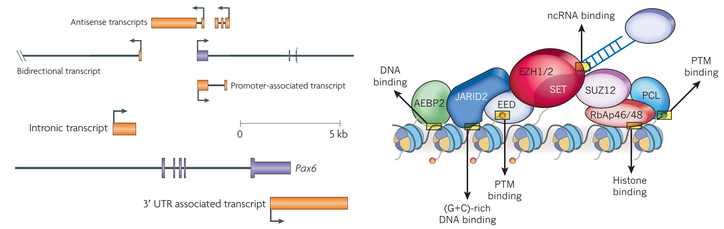
RIPSeeker: a statistical package for identifying protein-associated transcripts from RIP-Seq experiments

Yue Li[1], Dorothy Yanling Zhao [2], Jack Greenblatt[2,3] and Zhaolei Zhang[1,2,3]

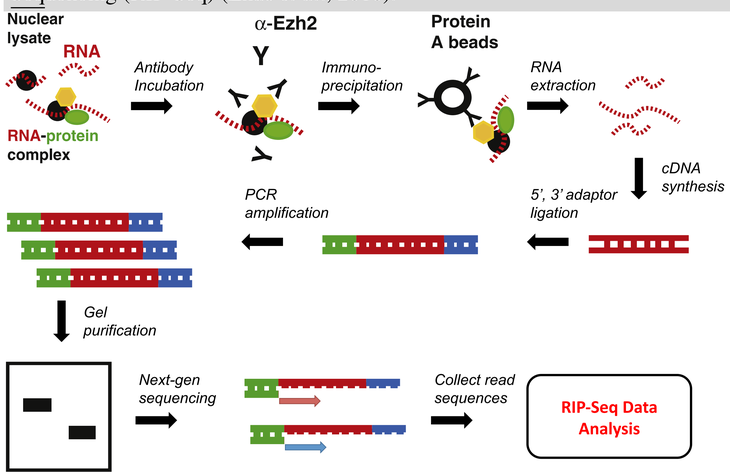
[1]*Department of Computer Science,* [2]*Department of Molecular Genetics,* [3]*Banting and Best Department of Medical Research, University of Toronto*

I. Introduction: genome-wide identification of long non-coding RNAs interacting with chromatin regulators

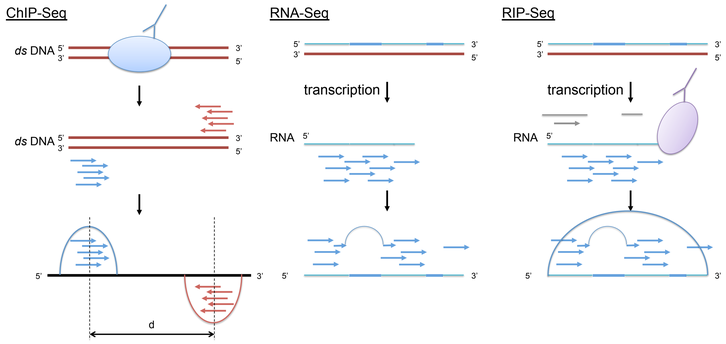
Noncoding RNA (ncRNA) biogenesis (Left; Mercer *et al.* 2009); long ncRNA partners with polycomb repressive complex 2 (PRC2) in gene regulation (Right; Margueron *et al.* 2011)



Ribonucleoprotein ImmunoPrecipitation (IP) followed by high-throughput Sequencing (RIP-Seq) (Zhao *et al.*, 2010):



Comparison of RIP-Seq with ChIP-Seq and RNA-Seq:

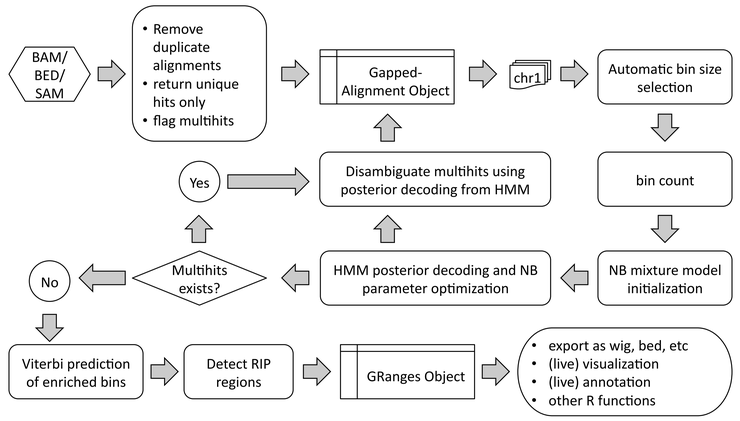


II. Motivation:

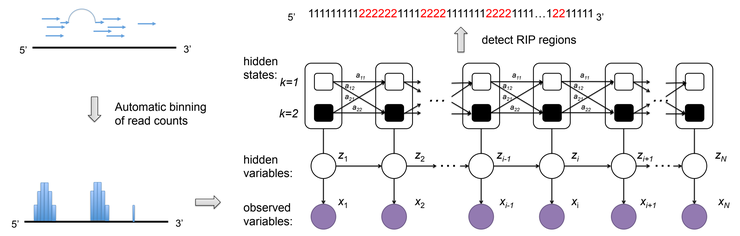
RIP-Seq measures genome-wide protein-RNA interactions. Despite similarity shared with ChIP- and RNA-Seq, RIP-Seq presents unique properties and challenges. Currently, no statistical tool is dedicated to RIP-Seq analysis.

III. Methods: probabilistic inference to disambiguate multihits and derive statistical-confidence RIP regions

RIPSeeker workflow:

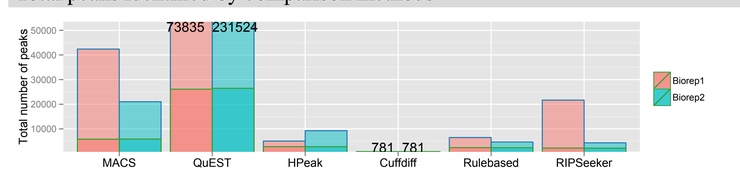


Inferring RIP regions, $p(z_i|\boldsymbol{X})$, using two-state HMM with negative binomial (NB) emission probability $p(x_i|z_i = k) = NB(a_k, b_k)$ on automatically discretized chromosome sequences:



IV. Results: analyzing PRC2 RIP-seq dataset

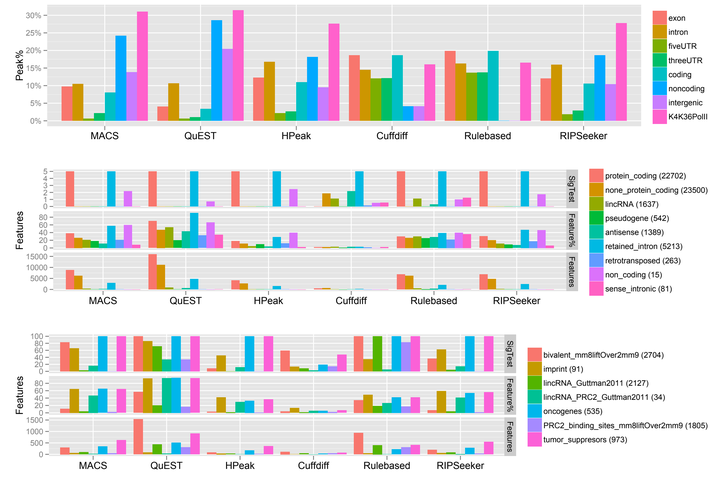
Total peaks identified by comparison methods



Pair-wise comparison of shared peaks:

PRC2	Cuffdiff	Rulebased	HPeak	RIPSeeker	MACS	QuEST
Cuffdiff	100%	2%	4%	4%	4%	2%
Rulebased	23%	100%	28%	26%	23%	12%
HPeak	26%	20%	100%	26%	19%	3%
RIPSeeker	45%	32%	55%	100%	40%	6%
MACS	54%	41%	39%	58%	100%	12%
QuEST	81%	69%	64%	71%	71%	100%

Comparison in biological contexts of various genomic and epigenetic features:



Detecting known PRC2-ncRNA: *Xist*, *Kcnq1ot1*, *Meg3*

V. Conclusion:

RIPSeeker is a self-contained software package written in R and specifically tailored to efficiently analyze RIP-Seq data with statistical rigor. RIPSeeker demonstrates its sensitivity by identifying the canonical PRC2- and CCNT1-associated (not shown) ncRNA with high statistical confidence and reasonable resolution. Additionally, RIPSeeker incorporates several existing R packages to automatically annotate RIP regions via Ensembl database, perform GO enrichments, and launch UCSC genome browser with putative RIP regions as custom tracks for visualization.

Because our current knowledge of protein-associated ncRNA is largely unknown (unlike TFBS), it is difficult to evaluate the specificity of RIPSeeker predictions. However, the ability to prioritize candidate genes with rigorous statistical assessment allows RIPSeeker to generate valuable information from RIP-Seq data for formulation of subsequent (more focused) experimental and computational strategy.