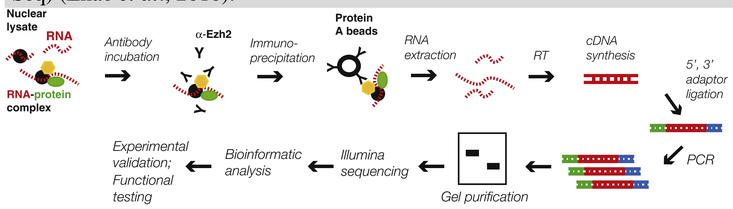
Constructing Protein-RNA Interactome using RIP-Seq

Yue Li[1,2], Dorothy Yanling Zhao [2,3], Jack Greenblatt[2,4] and Zhaolei Zhang[2,3,4]

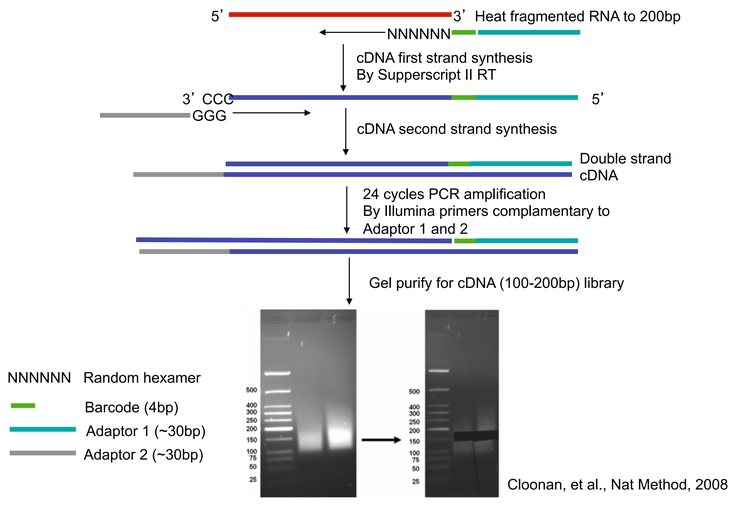
[1]*Department of Computer Science,* [2]*Donnelly Centre for Cellular and Biomolecular Research,*
[3]*Department of Molecular Genetics,* [4]*Banting and Best Department of Medical Research, University of Toronto*

I. Introduction

RNA immunoprecipitation (IP) followed by high-throughput sequencing (RIP-Seq) (Zhao *et al.*, 2010):



Strand-specific sequencing and cost-effective cDNA library prep:



II. Motivation:

RIP-Seq technology provides an unprecedented opportunity for identifying genome and transcriptome-wide protein-RNA interactions.

III. Challenges:

- Accurately binning the sample mixtures with short barcode sequences.
- Effectively filtering reads mappped to noisy (unknown) background RNAs.
- High sequence duplication and over-representation of k-mers.
- Low mappability to the corresponding genomes and/or transcriptomes.
- Lack of sequence annotations for ncRNA molecules.
- Effectively examining the mapped results.
- Inferring abundance of mapped isoforms with statistical significance.

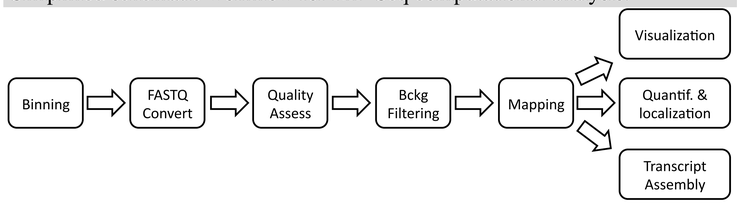
IV. Material

16 human IP from HEK293 cell lines and 4 mouse IP from mouse testes with 77,859,160 total binned reads and 3,309,899 unbinned reads.

Origin	IP Sample	Adaptor 1 (black) + 4 nucleotide barcode (red)
human	CBX2	CTTTCCCTACACGACGCTCTTCCGATCTATCC
human	CBX3	CTTTCCCTACACGACGCTCTTCCGATCTTAGC
...		
mouse	STK31	CTTTCCCTACACGACGCTCTTCCGATCTGACA
mouse	TDRKH	CTTTCCCTACACGACGCTCTTCCGATCTTGAC

V. Computational Methods

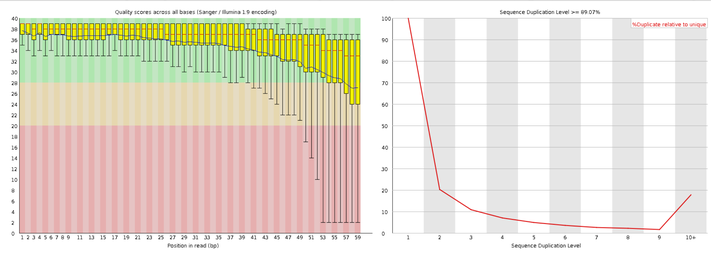
Simplified schematic workflow for RIP-Seq computational analysis:



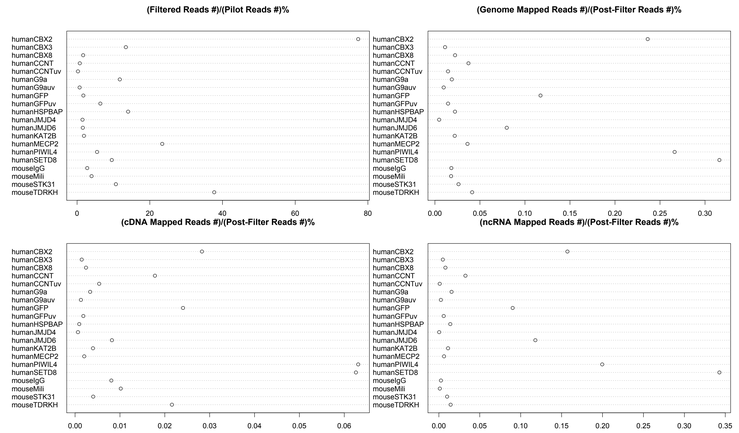
1. **Binning** based on 4-nucleotide barcode sequence
2. **Converting to FASTQ** to take into account both read sequence and Phred quality score
3. **Quality assessment** with *FASTQC*
4. **Filtering RNA background** sequences with *Bowite* aligner:
   (a) primer/adaptor
   (b) Spike-in virus/bacterial DNA genome (e.g., phiX genome)
   (c) Prevalent ribosomal, mitochondria RNA, actin mRNA, and more to discover
5. **Mapping** against reference genomes with *Bowite* aligner:
   (a) human (hg19) or mouse (mm9) genome
   (b) human or mouse transcriptome (Ensembl cDNA databases)
   (c) human or mouse ncRNA (Ensembl ncRNA databases)
   (d) 7sk and U6 RNA (known positive controls for human CCNT and human CCNT UV)
   (e) kcnq1ot1 RNA (known positive control for human G9a and human G9a UV)
6. **Visualization** of mapping results with *Integrative Genome Viewer* (IGV)
7. **Quantification and localization** of reads (an effective approach is under development)
8. **Transcript** (*de novo*) **assembly** with *Cufflinks*

VI. Preliminary Results

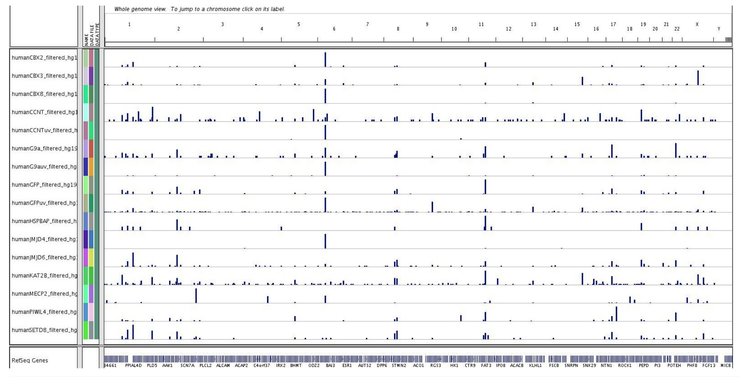
Per base Phred quality score (left) and sequence duplication levels (right) for GFP UV sample:



Filtering and mapping results:



IGV visualization of reads mapped to human genome (hg19):



- As expected, 105 reads from CCNT and < 5 reads from other samples map to 7sk RNA.
- A top list of read-enriched cDNA and ncRNA molecules were automatically generated.
- Some of them show distinctive enrichments in subsets of the 20 samples. For instance,
  - PKD2L1 (cDNA on '-' strand of chr10:102,047,903-090,243) only in CCNT and PIWIL4.
  - AC008541.1 (ncRNA on '+' strand of chr5:122,990,602-1,011) only in the 3 CBX's.
- Many unknown background RNA molecules were discovered.

VII. Works in Progress

- Filter out highly repeated sequences and newly discovered background RNA for remapping.
- Integrate previously published data with our data to distinguish noise and meaningful peaks.
- Correlate reads in HEK293 cell lines and mouse testes to gene expression and chromatin state.
- Explore frameworks (e.g., *cufflinks*) for (*de novo*) transcript assembly + infer abundances.
- Investigate statistical frameworks for significance of read counts *specifically* for RIP-Seq.

VIII. Acknowledgment

Donnelly Centre for Cellular + Biomolecular Research
UNIVERSITY OF TORONTO
UNIVERSITY OF TORONTO
NSERC CRSNG