Advanced Genome Bioinformatics

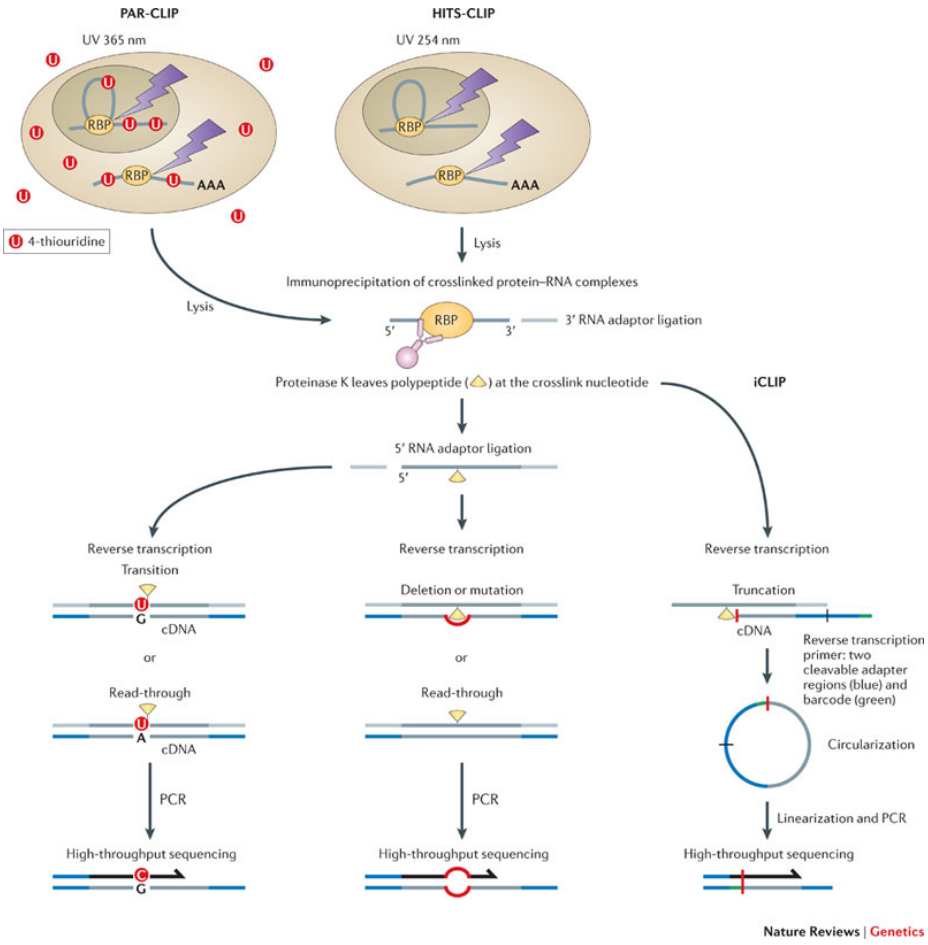
Markov model Assignment

Modeling RBFOX2 RNA binding sites with a Markov model

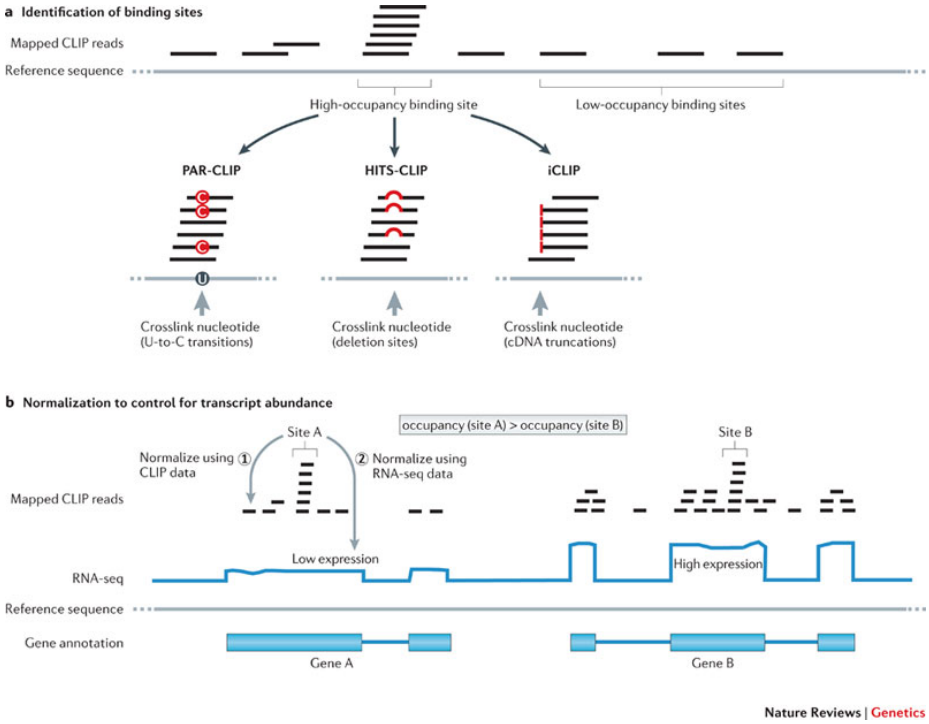
Introduction

Nearly all protein-encoding human genes have multiple exons that are combined in alternative ways to produce distinct mRNAs, often in an organ-specific, tissue-specific, or cell type specific manner. Definition of intron borders often requires the collaboration of RNA-binding proteins (RBPs), such as serine arginine (SR) and heterogeneous nuclear RNPs (hnRNPs), which interact with specific exonic or intronic sequence elements usually located in the vicinity of splice sites. Additionally, there are RBPs that are particularly relevant in development or in tissue-specific splicing regulation. This is the particular case of the RBFOX1 (A2BP1) and RBOFX2 (RBM9) proteins, whose binding sites have been found in the introns proximal to differentially spliced exons in brain and muscle tissues. In general, these binding sites were found more frequently in introns upstream of exons that were differentially excluded in those tissues, and in introns downstream of exons that were differentially included.

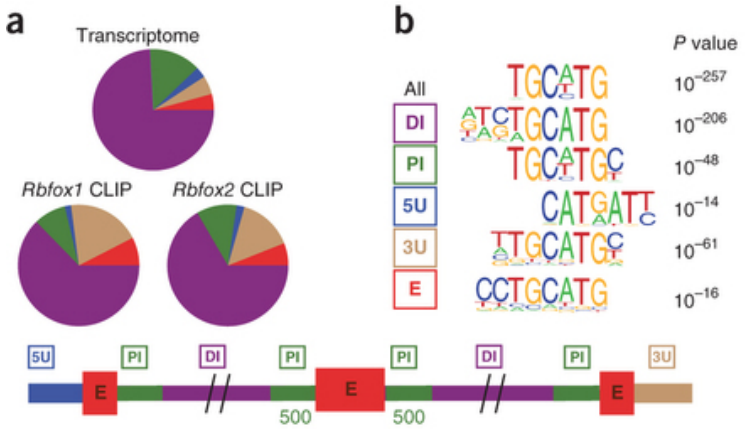
Multiple biochemical and computational strategies have been developed to study how RBPs bind RNA. A strategy that is frequently applied consists in biochemically purifying a specific RBP from a cell extract and identifying its associated RNA targets using microarrays or RNA deep sequencing. There are several variations of this approach. The most common approach, CLIP-Seq, involves cross-linking with UV light of the RBP to its cognate RNAs, followed by immunoprecipitation and deep sequencing of RNA (Konig et al. 2012):



Sequencing reads are mapped back to the genome or the transcriptome. Significant clusters of reads, also called peaks, are calculated by estimating the enrichment with respect to a background of RNA expression, or with respect to an expected distribution of CLIP-Seq reads in the same region (the pre-mRNA where the RBP is expected to bind). This allows the determination of candidate targets for an RBP (See Konig et al. 2012):



These techniques have been applied before to RBFOX proteins to describe their binding specificities. Multiple studies have determined that RBFOX proteins bind to RNA through a highly conserved sequence motif `TGCATG` (Lovci et al. 2013):



Objective

The objective of this exercise is to implement a search algorithm for RBFOX2 binding sites using a Markov model of order k, making use of the Log-likelihood ratio between the signal model and the background model, to identify RBFOX binding sites. In this assignment, we are not so much intersted in describing the actual binding motif, but to study whether we can predict the binding specificity of RBFOX2. This is expected to include the known RBFOX2 motif `TGCATG`, but possibly other motifs and dependencies among them, providing information to understand the specificity of binding.

Guide

1. As training and testing sets we will use regions that have been identified as biding RBFOX2 using eCLIP-Seq, a modified protocol of CLIP-Seq that have been used in the ENCODE project to study more than hundred RBPs.
   - Select the significant regions for eCLIP of RBFOX2 from the ENCODE web page. You can use one replicate only: ENCODE RBFOX eCLIP. Select the peaks file in bigBed format (bigBed format is described here). You can download a binary to convert the bigBed into Bed format from this page.
   - Modify the BED file to add 50nt on either side of each CLIP binding region.
   - Use Galaxy to upload the Bed file, and then extract the nucleotide sequences for the significant CLIP regions. Use the region (TAB delimited) format. Now you have a TAB delimited file with the sequences for the CLIP binding regions (variable length) plus 50nt flanking at either side. Something like this:

     ```
     chr1 17451 17528 ENSG00000227232.4_0_57 494 - -1 -1 3.73715129938e-50 17489 GCACTTCGATC....CAGGTACAGCACATAG
     ...
     ```

     the bed plus the sequence

     The columns of this file contain the following information:

     1. chromosome

2. start of CLIP region in hg19 assembly
3. end of CLIP region in hg19 assembly
4. gene where this CLIP region is.
5. number of CLIP reads found in this region.
6. strand
7. not used
8. not used
9. p-value of significance for each CLIP region.
10. position of the middle point of the peak.
11. sequence for 50nt + CLIP region + 50nt.

2. Separate the datasets into training and testing.

3. Develop a program that will read these sequences and will estimate a background Markov model of order k from the flanking sequences, and a signal Markov model of order k for the CLIP regions. Note that Markov models of order `k`, with `k>1`, can be seen as a Markov model of order 1 over the alphabet of k-tuples. For instance, a Markov model of order 2 for RNA can be treated as a Markov model of order 1 over the alphabet:

   ```
   AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT
   ```

   In this case, it is important to process one character at a time. Accordingly, the sequence `ACGGT` would be processed:

   ```
   AC CG GG GT
   ```

   Accordingly, we will calculate the signal and the background model as probability distributions over `k`-mers, using the parts of the sequence corresponding to the CLIP regions and the parts corresponding to the 50nt flanking regions. The output of this first program will be the list of probabilities for k-mers for signal and background. That is, inside and outside the CLIP regions. This first program should accept `k` as input.

4. Develop a second program to scan the test sequences to predict RBFOX2 binding regions with your Markov model. This program will read the estimated k-mer probabilities from the previous step and the set of test sequences (different from the training sequences) and will run a sliding window of length `l` along the sequence to score for the potential to detect an RBFOX2 binding site. This program should also accept a length `l` as input. Decide on a score cutoff to determine the windows that can be considered potential RBFOX2 binding sites.

   sigma??

5. Calculate the precision and recall of the predictions using the following definitions:
   - TP (true positives): number of bases that are correctly predicted as part of an RBFOX2 binding region.
   - FP (false positives): those bases that are incorrectly predicted as RBFOX2 binding positions.
   - FN (false negatives): those bases that are missed (False negatives),
   - TN (true negatives): those that are correctly not predicted as RBFOX2 binding site.

6. Calculate the accuracy of the method as a function of the length `l` of the search window and as a function of the order `k` of the Markov model. Can you find an optimal window length or an optimal k?

References

Konig J, Zarnack K, Luscombe NM, Ule J. Protein-RNA interactions: new genomic technologies and perspectives. Nat Rev Genet. 2012 Jan 18;13(2):77-83. doi: 10.1038/nrg3141. Review. Erratum in: Nat Rev Genet. 2012 Mar;13(3):220. PubMed PMID: 22251872.

Yeo GW, Coufal NG, Liang TY, Peng GE, Fu XD, Gage FH. An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. Nat Struct Mol Biol. 2009 Feb;16(2):130-7.

Lovci MT, Ghanem D, Marr H, Arnold J, Gee S, Parra M, Liang TY, Stark TJ, Gehman LT, Hoon S, Massirer KB, Pratt GA, Black DL, Gray JW, Conboy JG, Yeo GW. Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. Nat Struct Mol Biol. 2013 Dec;20(12):1434-42.