BackCLIP: a tool to identify common background presence in PAR-CLIP datasets

C.A. Sierra, P.H. Reyes-Herrera,
C. Speck, S. Herrera

OUTLINE

---

1. Background
2. Research problem
3. Proposed approach
4. Results

Conclusions

OUTLINE

1.Background

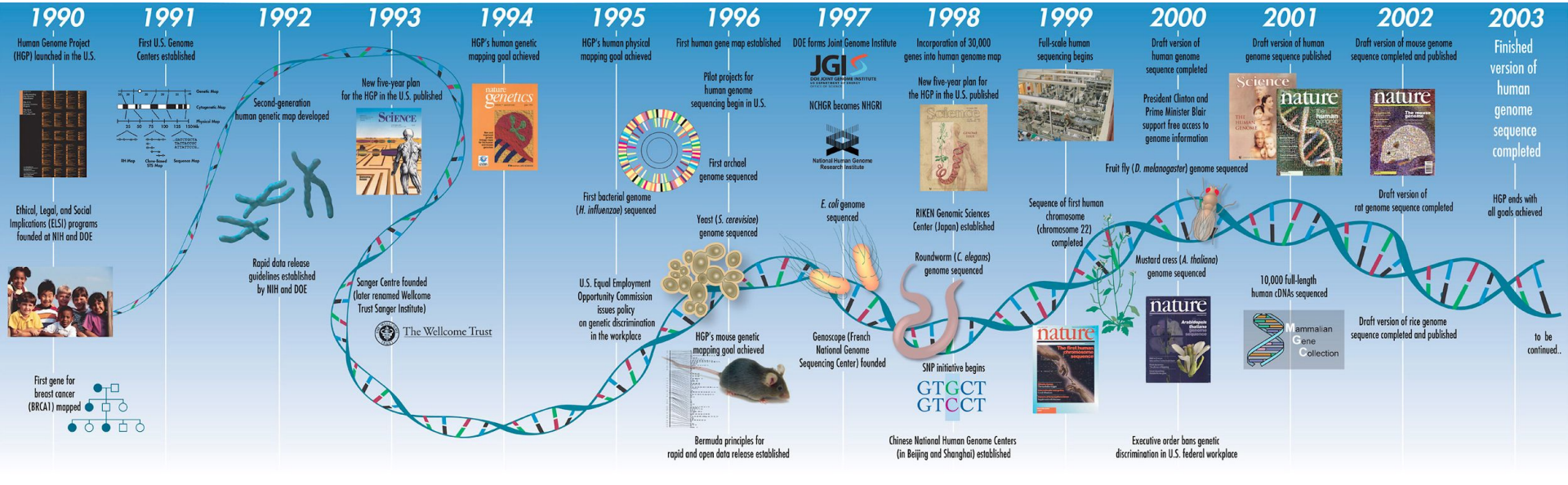
2.Research problem

3.Proposed approach

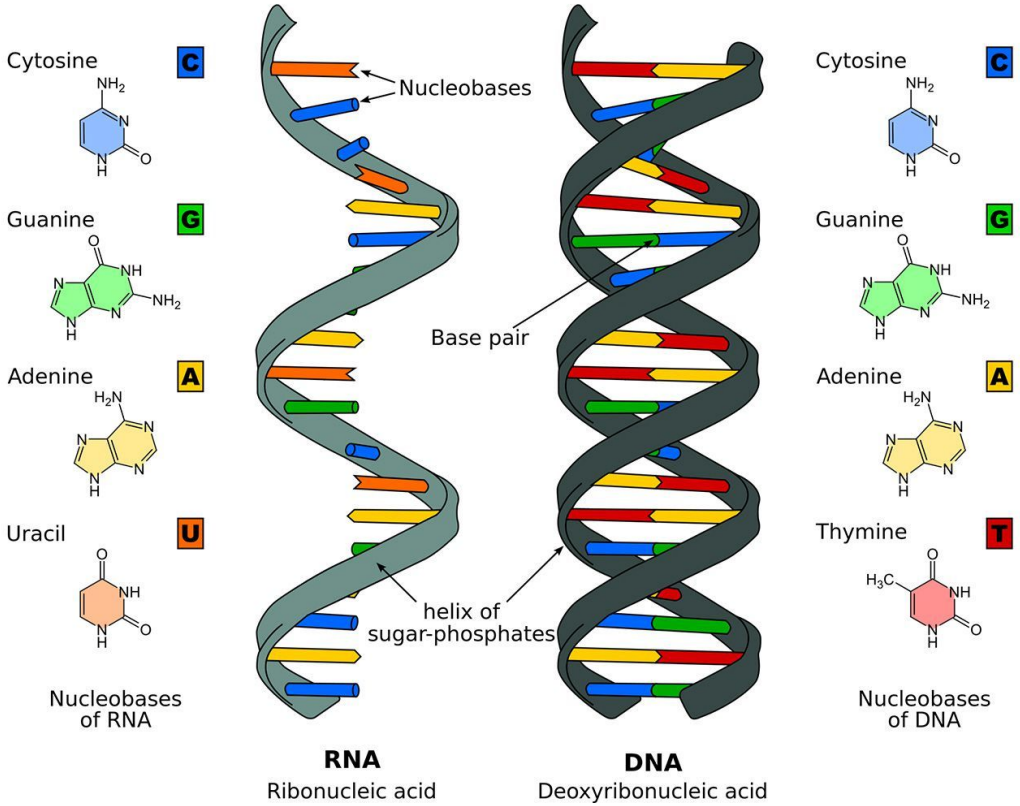
4.Results

Conclusions

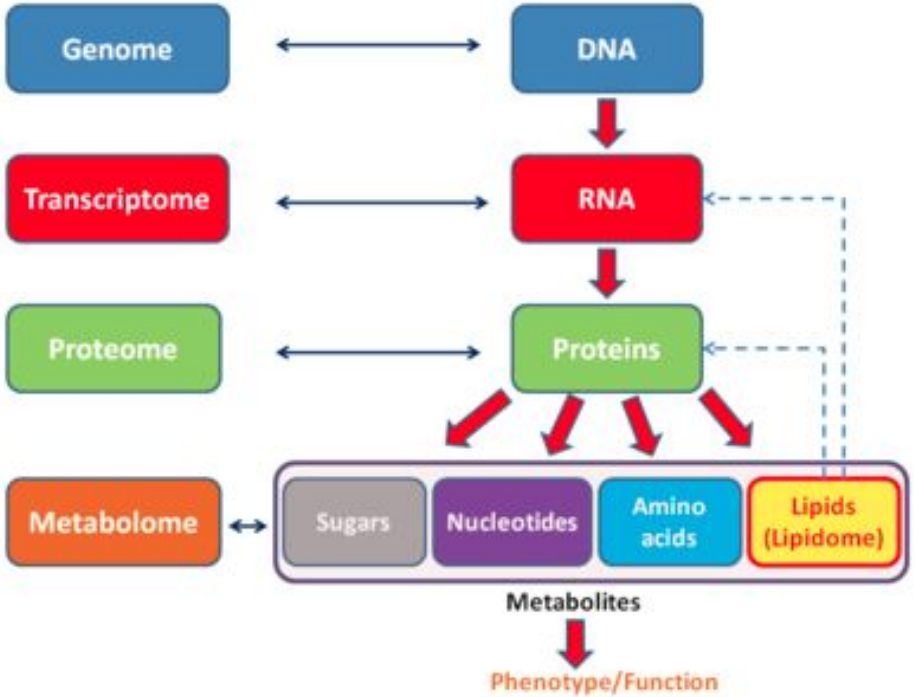
Human Genome



DNA & RNA


Cytosine
C
Guanine
G
Adenine
A
Uracil
U
Nucleobases of RNA
Nucleobases
Base pair
helix of sugar-phosphates
RNA
Ribonucleic acid
DNA
Deoxyribonucleic acid
Cytosine
C
Guanine
G
Adenine
A
Thymine
T
Nucleobases of DNA

Transcriptome

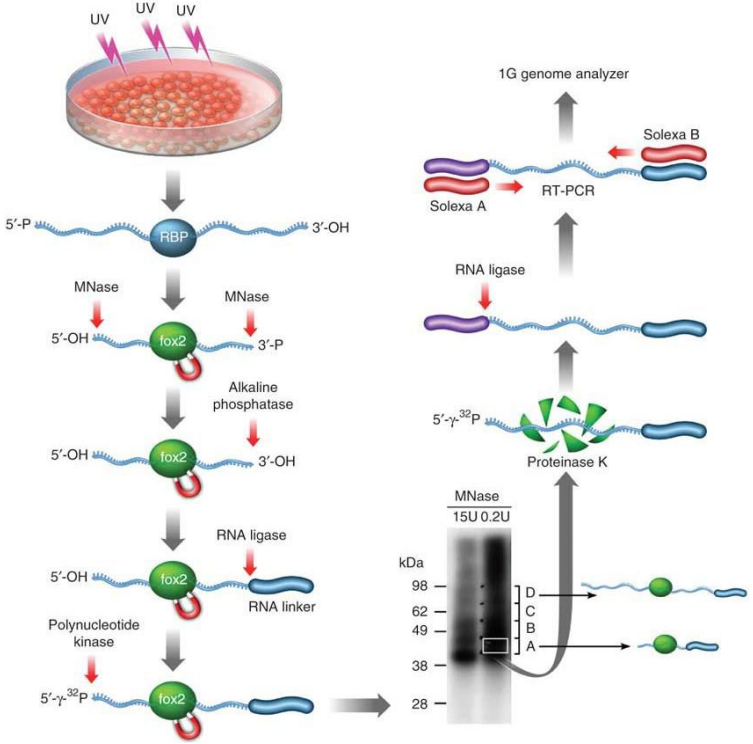


CLIP-seq protocol

Check interactions between RNA and proteins.

It uses antobiotics to affect protein binding sites.

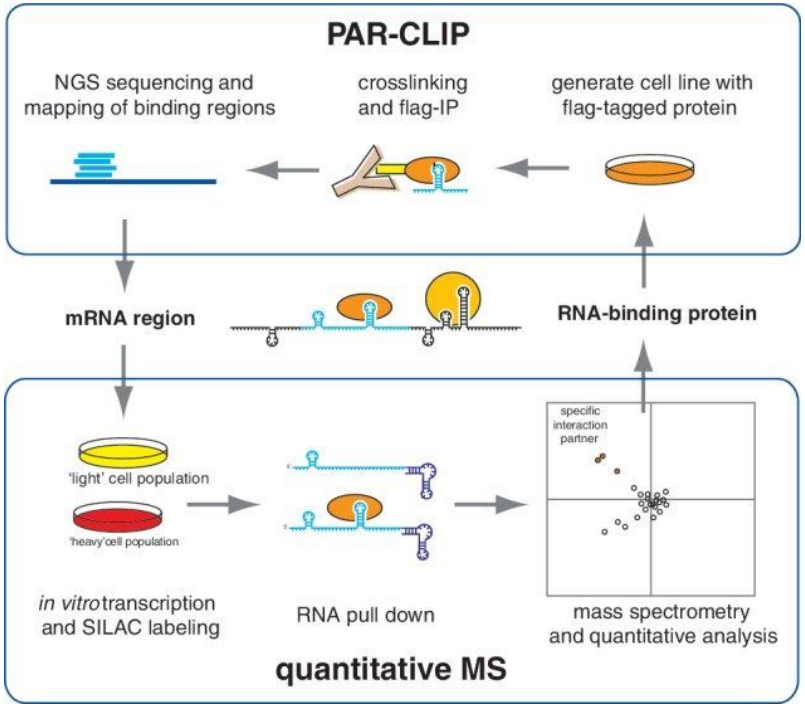
Immunoprecipitation helps to obtain specific genetic sequences.



PAR-CLIP data

PAR-CLIP, a frequently used CLIP-seq protocol, uses **photo activatable** nucleosides to label the transcripts in addition to an enhanced crosslinking (*Hafner et al., 2010*).

These modifications induce specific nucleotides transitions that facilitate the recognition of the cross-linked sites.



RBP (RNA Binding Proteins)

RNA-binding proteins (RBPs) have important roles in RNA regulation. The first step to understand RBPs' specific functions is to identify the RNA targets for each RBP.

The introduction of CLIP-seq protocols have made it possible to obtain sets of binding sites for RBPs at a transcriptome-wide scale (*Licatalosi et al., 2008*).

However, each CLIP-seq protocol introduces distinct modifications to reduce the presence of background (non-crosslinked RNA).

OUTLINE

1.Background

2.Research problem

3.Proposed approach

4.Results

Conclusions

Motivation

PAR-CLIP derives a transcriptome wide set of binding sites for RNA-binding proteins. Even though the protocol uses stringent washing to remove experimental noise, some of it remains.

A recent study measured three sets of non-specific RNA backgrounds which are present in several PAR-CLIP datasets (*Sievers et al., 2012*).

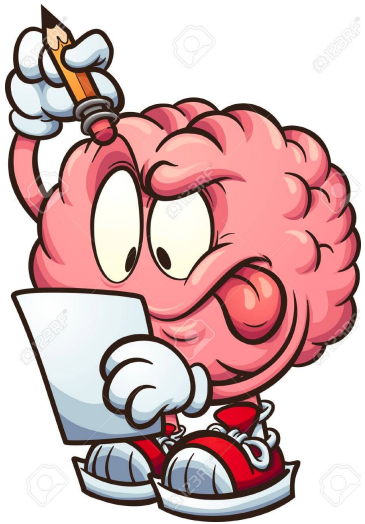
Motivation

However, a tool to identify the presence of common background in PAR-CLIP datasets is not yet available.

Non-specific RNA background must be taken into account when processing PAR-CLIP data because it can interfere with the distinction of the specific characteristics recognized by the RBPs, and therefore the identification and understanding of binding targets and protein function (*Friedersdorf and Keene, 2014*).

Research Question

Is it possible to create a tool to identify common background presence in PAR-CLIP datasets and remove it?



OUTLINE

1.Background

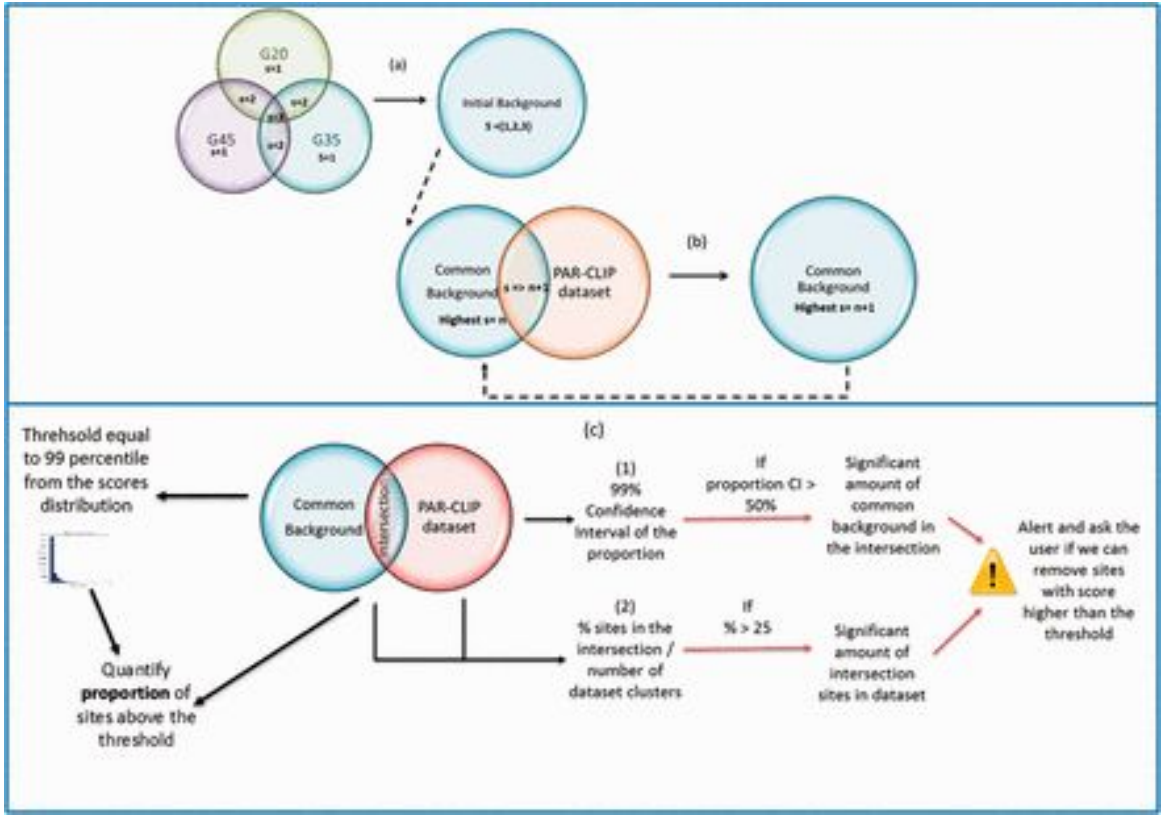
2.Research problem

3.Proposed approach

4.Results

Conclusions

Proposed Approach



Score Measure

Definition of possible motif candidates.

Clusters detection based on position, chromosome, and motif candidate.

Count occurrences of motifs into clusters based on clustering profiles.

OUTLINE

1.Background

2.Research problem

3.Proposed approach

4.Results

Conclusions

Results

We used the proposed strategy in 30 PAR-CLIP datasets from nine proteins.

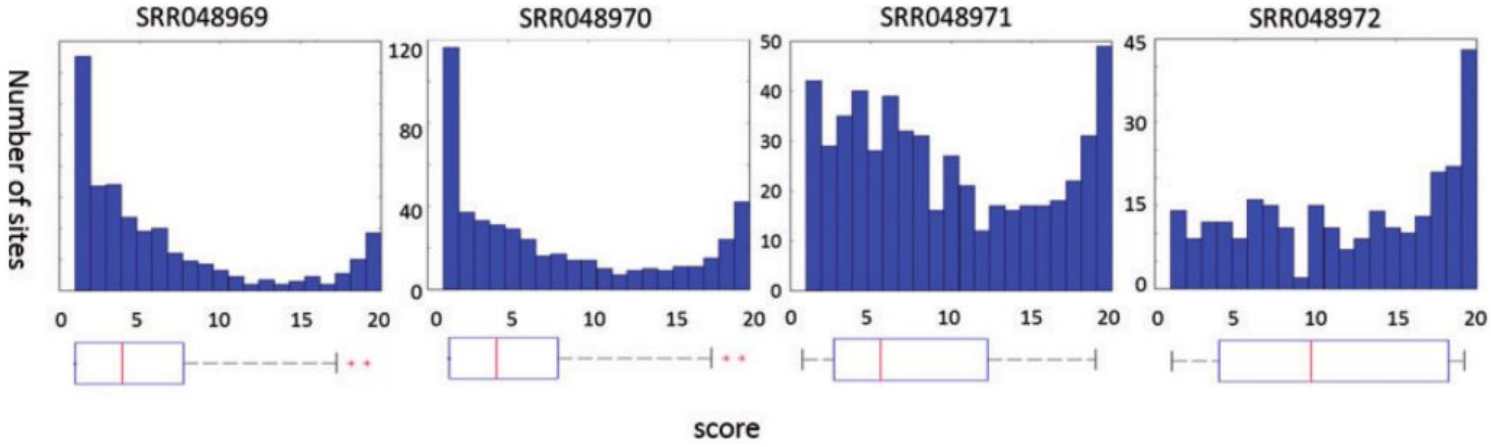
As an example, we selected Quaking (QKI) protein from the nine RBPs (30 PAR-CLIP datasetsand four of its PAR-CLIP datasets.

Results

Table 1. QKI datasets results for the PAR-CLIP dataset and for the background intersection

Dataset	Clusters	Clusters with motif (%)	Background intersection Sites	Background intersection sites with motif (%)	Background intersection sites / Clusters (%)	motifs in intersection / motifs in dataset (%)	Proportion of sites above threshold (score = 8)	BackCLIP sites identified (score ≤ 8)	BackCLIP sites with motif (%)
SRR048969	5286	44	654	11	12	3.1	[32%, 41%]	260	2
SRR048970	5091	48	479	14	9	2.7	[29%, 39%]	176	2
SRR048971	1688	23	539	5	32	7.0	[41%, 52%]	294	2
SRR048972	590	13	276	3	47	10.8	[55%, 59%]	178	1

Results



Histogram and boxplot of the scores in the intersection between the common background and four QKI PAR-CLIP datasets (Tophat alignment)

OUTLINE

---

1. Background
2. Research problem
3. Proposed approach
4. Results

Conclusions

Conclusions

It is possible to identify the presence of common backgrounds in a dataset and identify differences in datasets for the same protein.

BackCLIP is a useful tool to identify the amount of common background in any dataset. This method is the first step in the process of completely removing such backgrounds.

GitHub link: ***https://github.com/phrh/BackCLIP***

Thanks!

Questions?

casierrav@unal.edu.co