Protein – RNA interactions: Analysis of iCLIP-seq data

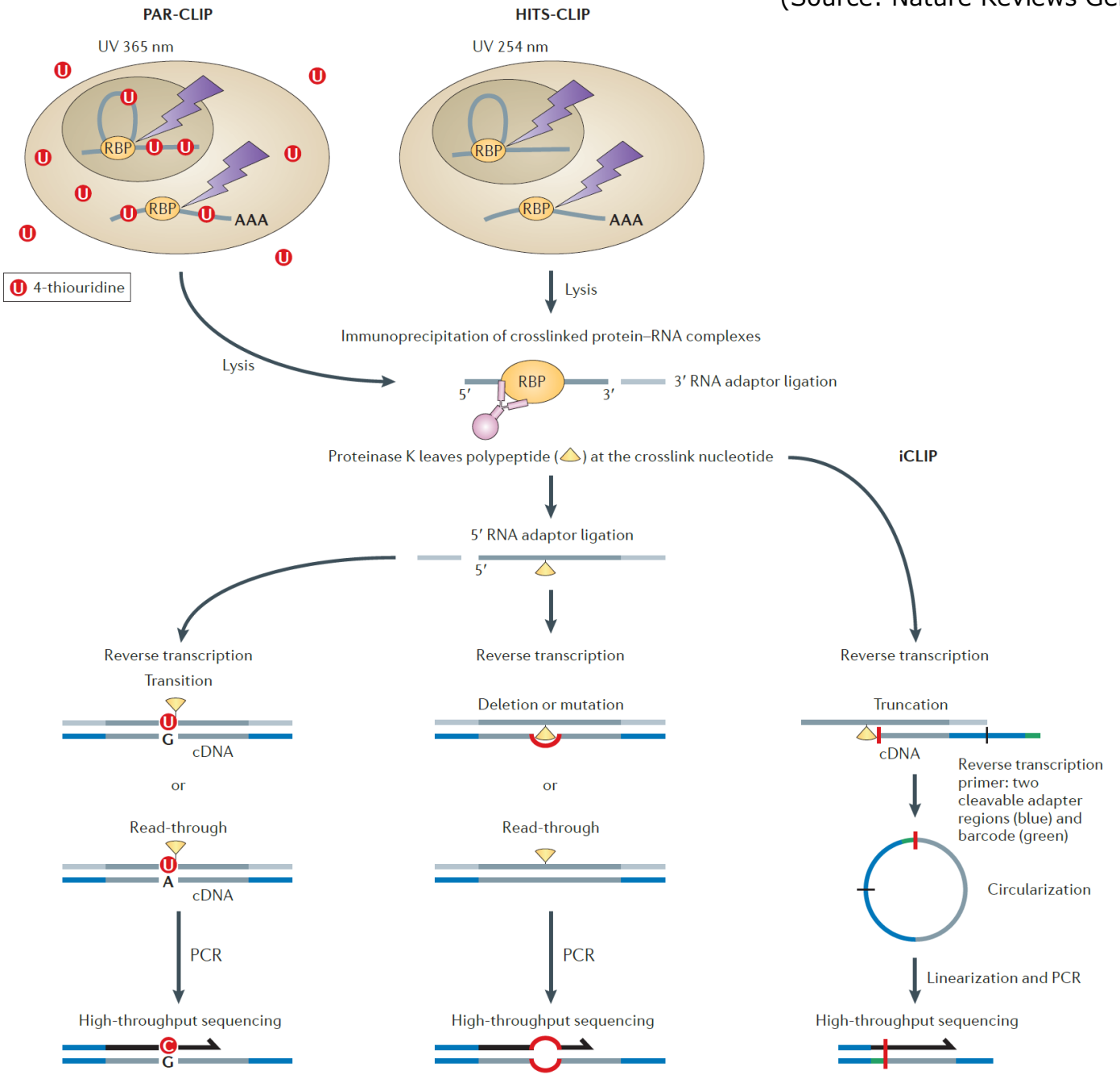
FU Berlin
Seminar RNA Bioinformatics, WS 14/15
Sabrina Krakau

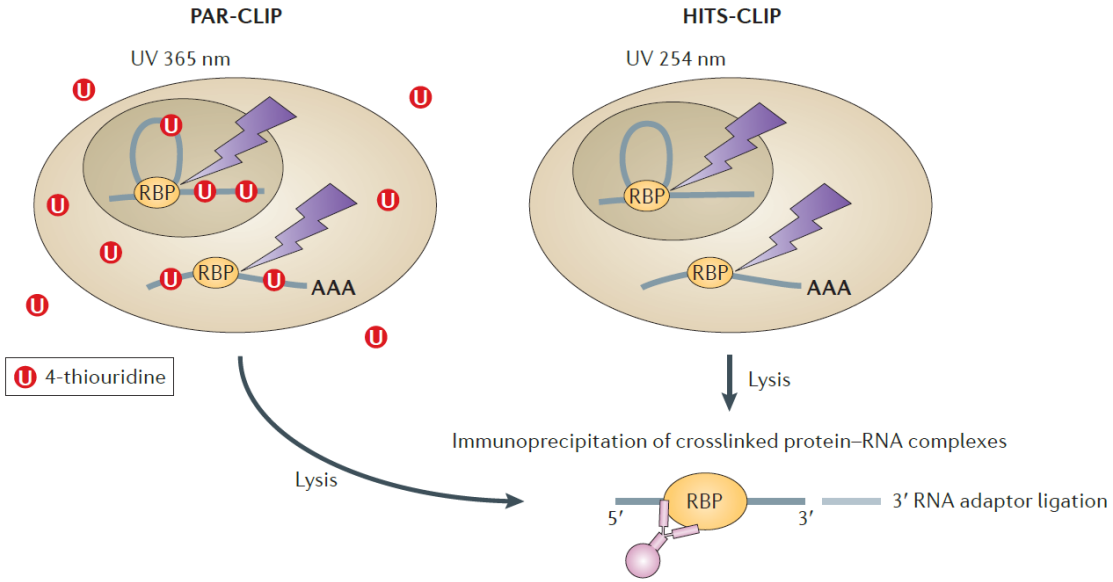
Protein – RNA interactions

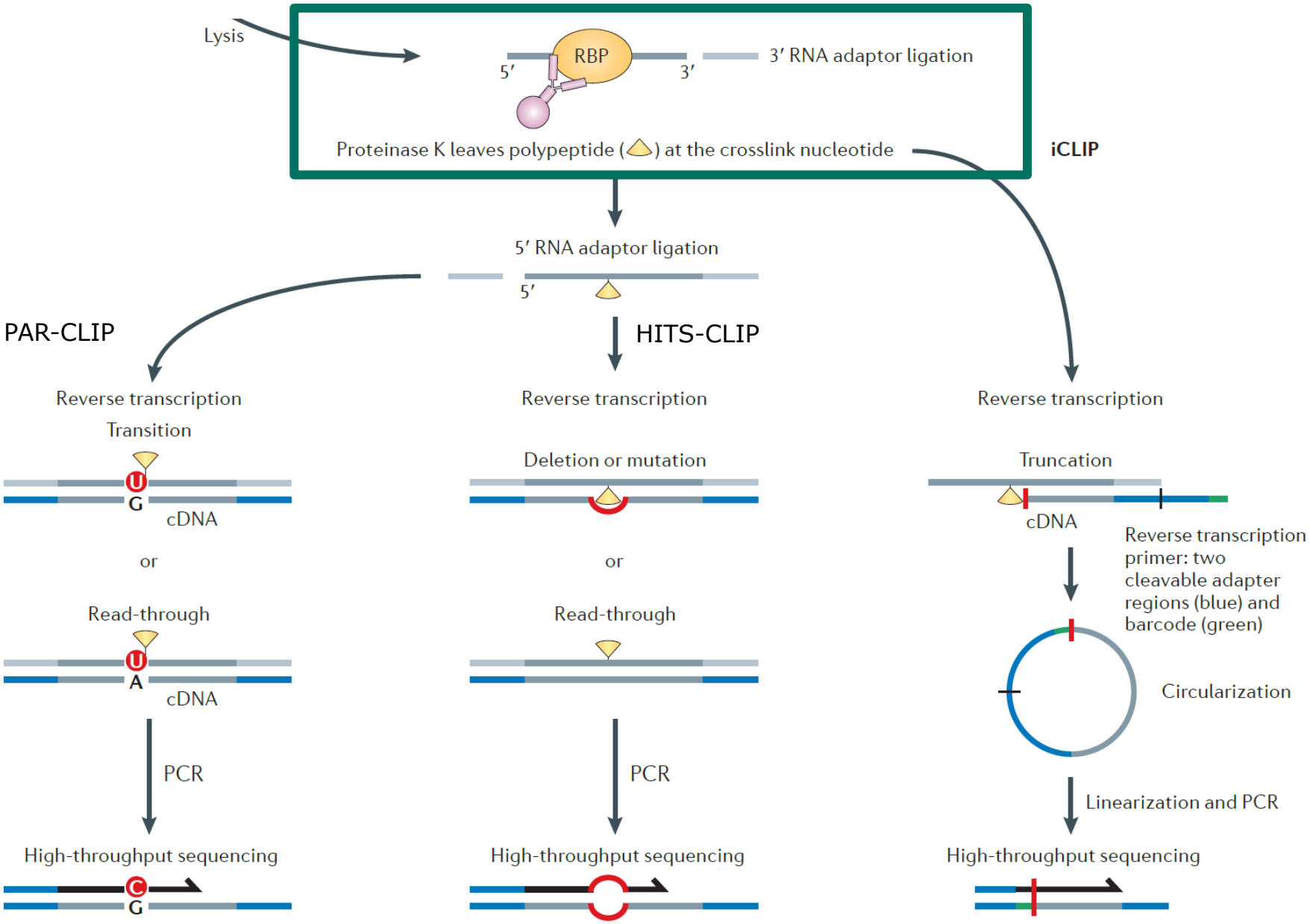
- Core of post-transcriptional regulation
- RNA binding proteins (RBPs) often bind several sites on most RNAs
  → landcape of interactions

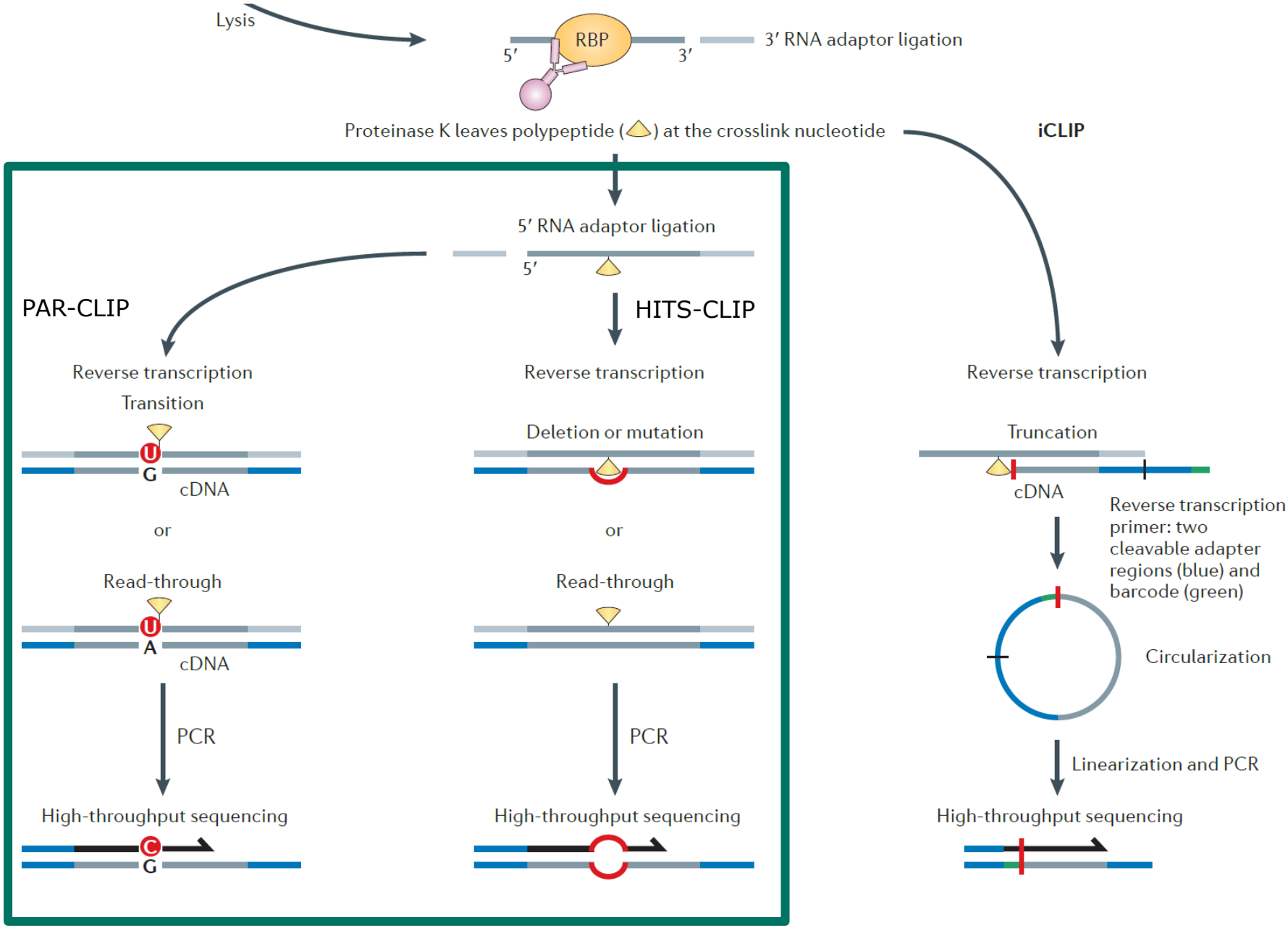
- CLIP-seq (**c**ross-**l**inking **i**mmuno**p**recipitation combined with HTS)
  - Binding site detection with high-resolution for a given RBP
  - Transcriptome-wide analysis

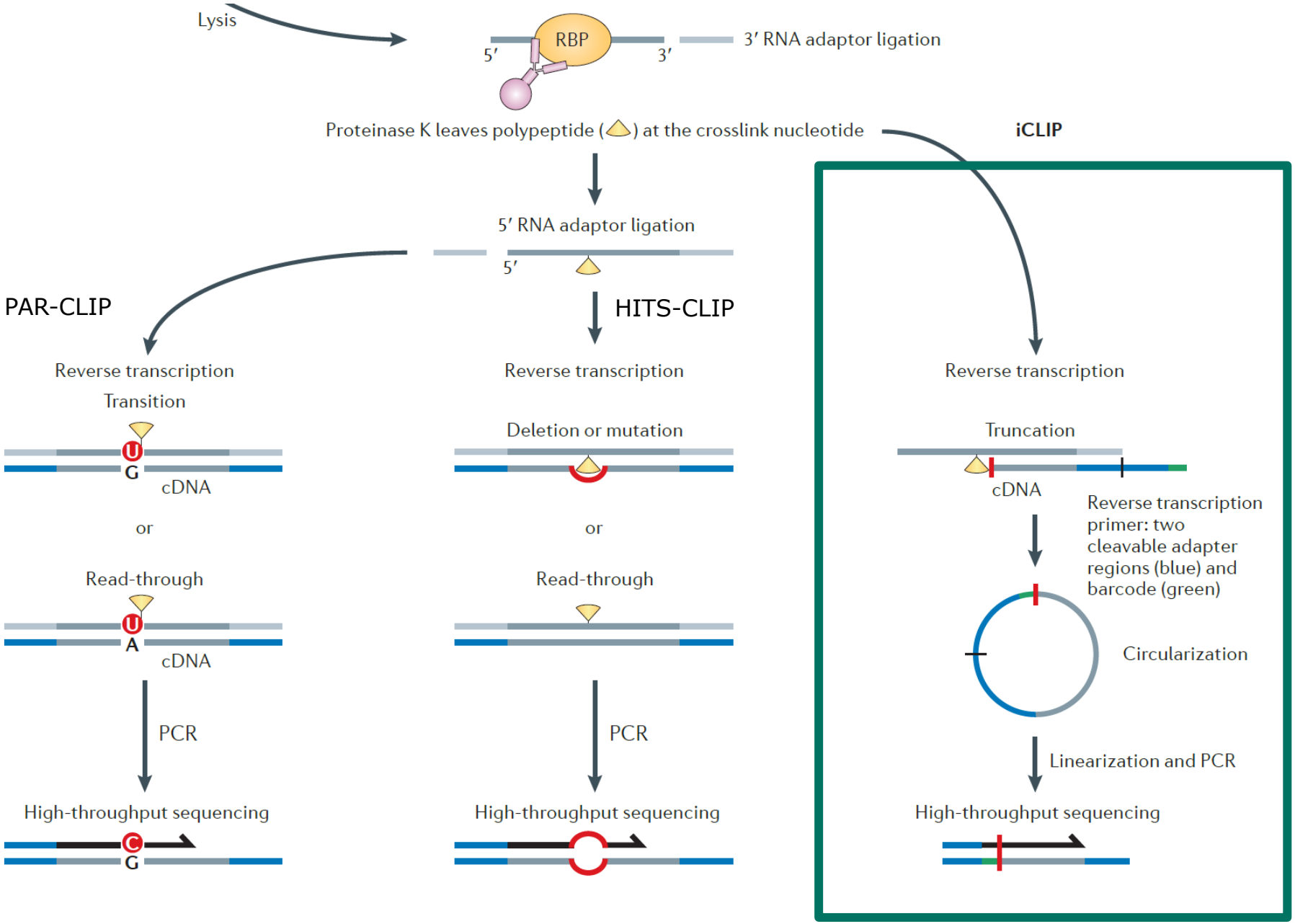
(Source: Nature Reviews Genetics 13, 77-83)



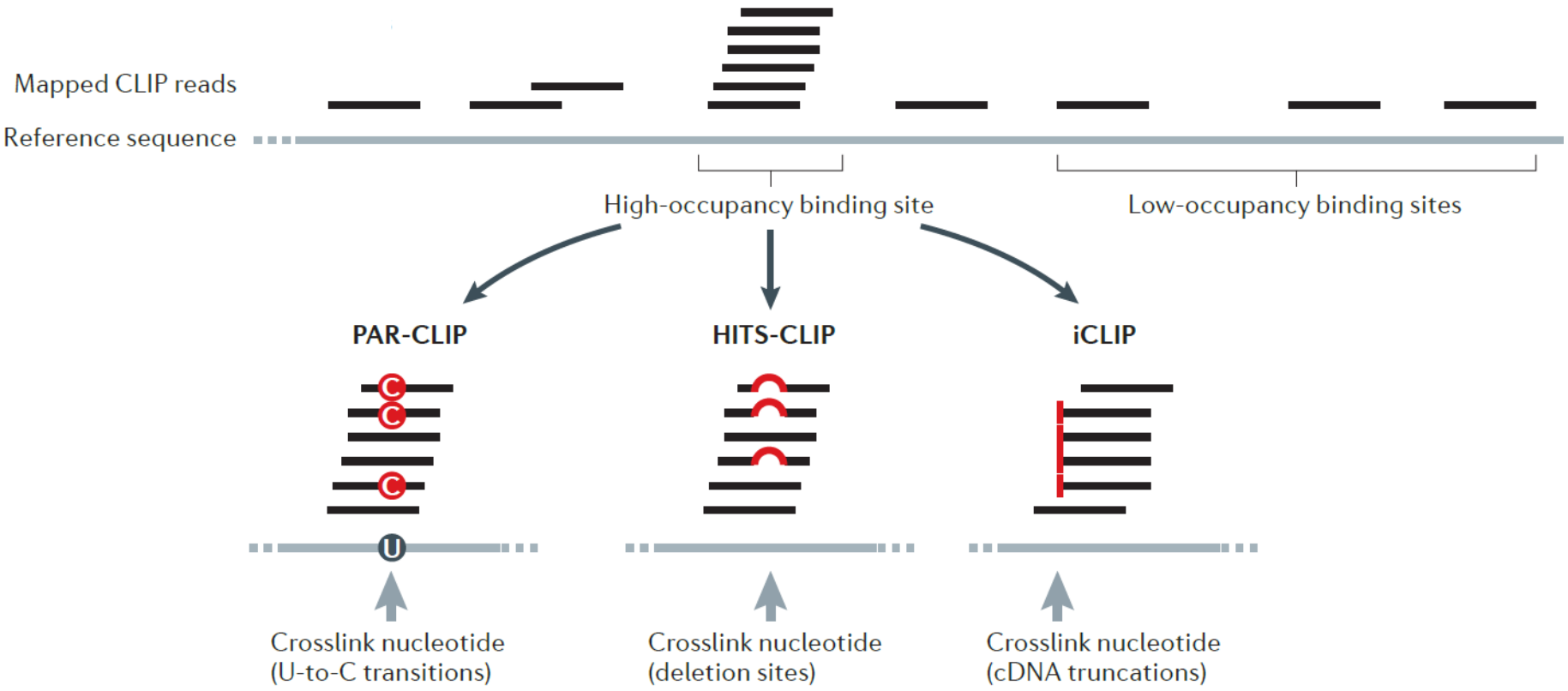

PAR-CLIP
UV 365 nm
HITS-CLIP
UV 254 nm
RBP
AAA
4-thiouridine
Lysis
Lysis
Immunoprecipitation of crosslinked protein–RNA complexes
5′
3′
3′ RNA adaptor ligation


Lysis
RBP
5′
3′
3′ RNA adaptor ligation
Proteinase K leaves polypeptide (△) at the crosslink nucleotide
iCLIP
5′ RNA adaptor ligation
5′
PAR-CLIP
HITS-CLIP
Reverse transcription
Reverse transcription
Reverse transcription
Transition
U
G
cDNA
Deletion or mutation
Truncation
cDNA
Reverse transcription primer: two cleavable adapter regions (blue) and barcode (green)
or
or
Read-through
Read-through
U
A
cDNA
Circularization
PCR
PCR
Linearization and PCR
High-throughput sequencing
High-throughput sequencing
High-throughput sequencing
C
G


Lysis
RBP
5′
3′
3′ RNA adaptor ligation
Proteinase K leaves polypeptide (△) at the crosslink nucleotide
iCLIP
5′ RNA adaptor ligation
5′
PAR-CLIP
HITS-CLIP
Reverse transcription
Transition
U
G
cDNA
or
Read-through
U
A
cDNA
PCR
High-throughput sequencing
C
G
Reverse transcription
Deletion or mutation
or
Read-through
PCR
High-throughput sequencing
Reverse transcription
Truncation
cDNA
Reverse transcription primer: two cleavable adapter regions (blue) and barcode (green)
Circularization
Linearization and PCR
High-throughput sequencing


Lysis
RBP
5′
3′
3′ RNA adaptor ligation
Proteinase K leaves polypeptide (△) at the crosslink nucleotide
iCLIP
5′ RNA adaptor ligation
5′
PAR-CLIP
HITS-CLIP
Reverse transcription
Transition
U
G
cDNA
or
Read-through
U
A
cDNA
PCR
High-throughput sequencing
C
G
Reverse transcription
Deletion or mutation
or
Read-through
PCR
High-throughput sequencing
Reverse transcription
Truncation
cDNA
Reverse transcription primer: two cleavable adapter regions (blue) and barcode (green)
Circularization
Linearization and PCR
High-throughput sequencing

Identification of binding sites



> 80% truncated

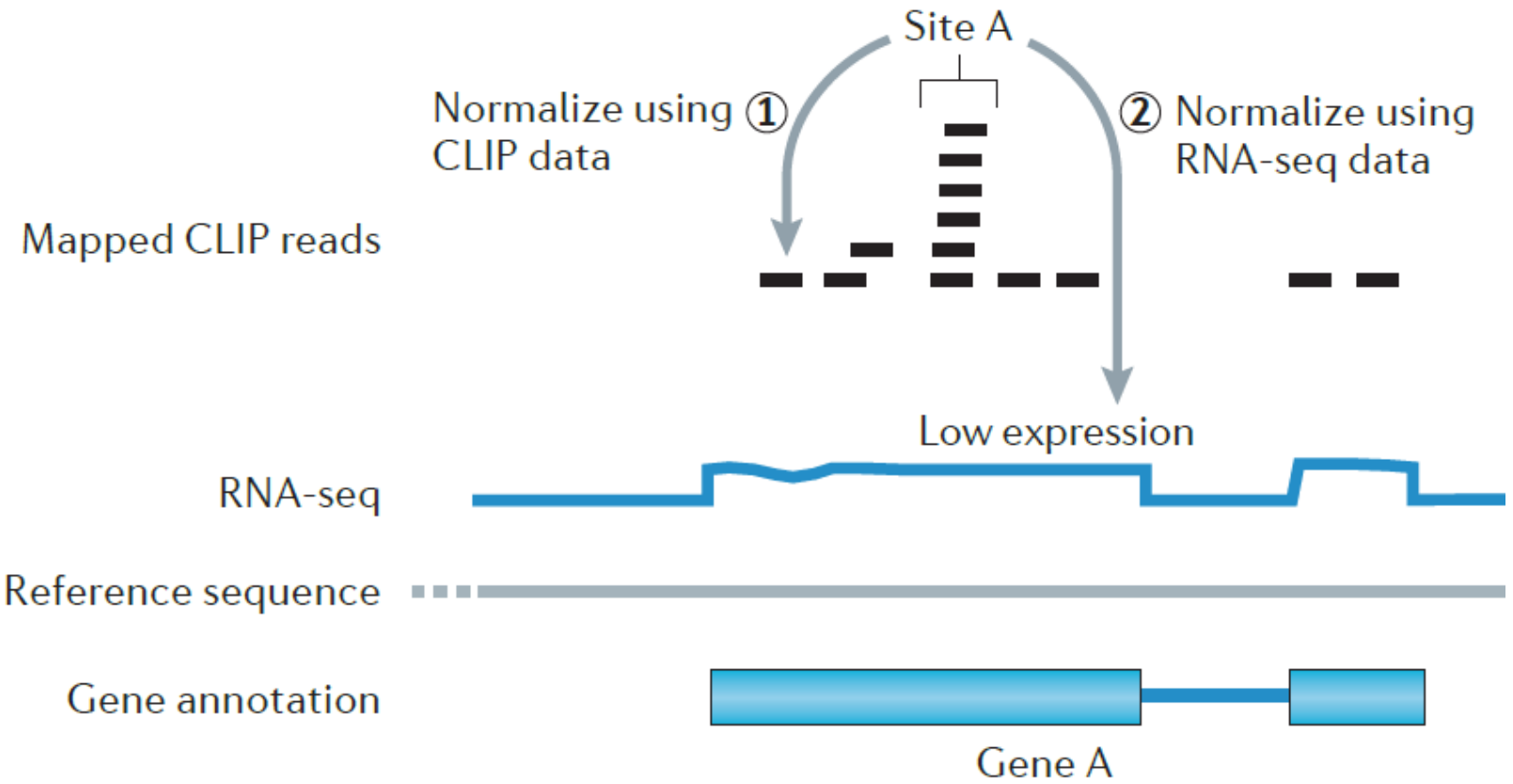
→ diagnostic events (DEs)

(Source: Nature Reviews Genetics 13, 77-83)

Read counts ⇔ RBP binding affinity?

Normalization

Read count depends on expression level:



(Source: Nature Reviews Genetics 13, 77-83)

Which peaks are significant?

Binding site calling

- Model underlying read count distribution to distinguish background from binding site

- Take DEs into account

Binding site calling

Piranha (2012)

- Models read count distribution of bins using the ZTNB (zero truncated negative binomial) distribution
  - Given (untruncated) mean read count μ

  → find dispersion parameter maximizing the ZTNB log-likelihood function

- External data as covariates X (e.g. transcript abundances, DEs)

  → ZTNB regression model: $\mu_i = \exp[\vec{\beta}^T \overrightarrow{X_i}]$

  → Find dispersion and regression parameters β that maximize the log-likelihood function

Binding site calling

PIPE-CLIP (2014)

Calling peaks/enriched clusters:

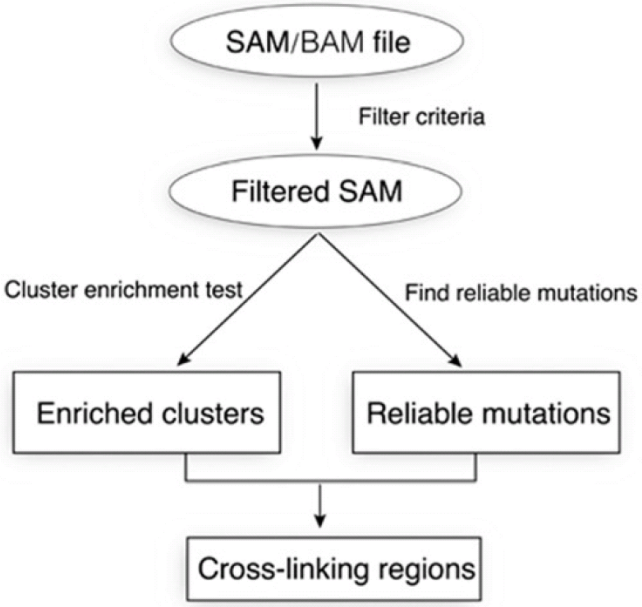
- ZTNB regression model for read counts of cluster

  → p-value → FDR

Detecting cross-linking sites:

- Number of DEs is modeled with binomial distribution (no. of mapped reads, DEs and global success rate)

  → p-value → FDR



→ Combine p-values for final calling (using Fisher‘s method)

- No normalization for transcript abundances!

Binding site calling

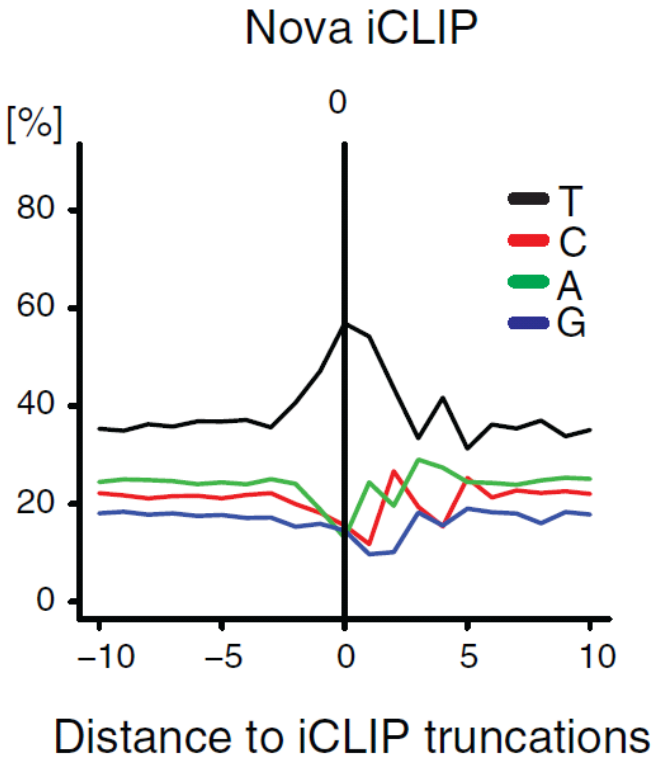
dCLIP (2014)

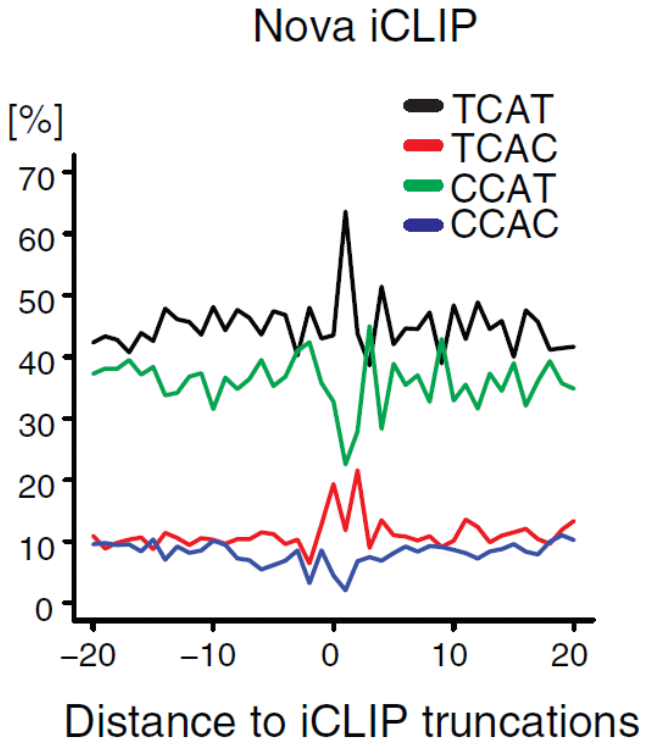
- Comparative CLIP-seq analysis
- Normalization: MA-plot (assuming a large number of common binding sites with similar binding strengths)
- Detection of RBP sites using HHM:
  - Differential binding vs. non-differential binding site

Is it that simple?

Sequence bias

UV-C induced cross-linking preferentially occurs at Us (Sugimoto et al., 2012):





→ Bias can be avoided by analysis of motifs enriched in the vicinity

(Source: Sugimoto et al. *Genome Biology* 2012)

Background binding

1) Binding to proteins != RBP of interest
2) False cross-linking events

Friedersdorf et al., 2014:

- 8 – 45% of reads from published PAR-CLIP datasets overlap with background sites from FLAG-GFP PAR-CLIP
- Background reads are mostly derived from direct protein-RNA interactions → DEs

→ Use control CLIP with unspecific protein (or publicly available results in GEO for PAR-CLIP) for correction

GC bias

Read counts depend on GC content:

- GC rich and poor sequences are underrepresented

  (due to different melting temperatures in PCR)

→ GC normalization

Motifs:

Refining binding sites and characterization

Motif discovery

- RBP binding sites:
  - Shorter than TF binding sites
  - Characteristic secondary structures (not trivially determined by sequence)!
  - Low sequence specificity in some RBPs

MEMEris:	uses RNA secondary structure to guide motif search towards single-stranded regions
RNAcontext:	learning RBP-specific sequence and structural preferences
RNAmotifs:	identifies multivalent regulatory motifs (clusters of short and degenerate sequences)
GraphProt:	learning sequence and structural preferences

Simultaneous binding site location and motif discovery

Zagros (Bahrami-Samani et al., 2014)

- Simultaneous motif characterization and binding site localization
- EM algorithm:
  - estimate parameters motif model *M* and background model *f*
  - Taking sequence, structure and DEs into account
  - Recompute motif occurence indicators at each iteration
    → binding sites
- Improved motif discovery compared to methods taking only sequence into account

Conclusion

Split-read mapping:
- TopHat
- STAR

Normalization:
- Piranha
- dCLIP

Peak calling:
- Piranha
- PIPE-CLIP
- dCLIP

Sequence bias and noise reduction?

Motif recovery:
- RNAcontext, RNAmotifs, GraphProt
- Zagros

Functional analysis

Conclusion

Open problems

- Accurate quantitative analysis remains challenging
- Need for computational methods taking sequence bias, background noise into account

Future

- Combinatorial interactions of proteins on RNAs?
- Interactions with DNA?
- How does RNA editing or epigenetic modifications influence these interactions or vice versa?

References

- Sugimoto, Yoichiro, et al. "Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions." Genome Biol 13.8 (2012): R67.
- Friedersdorf, Matthew B., and Jack D. Keene. "Advancing the functional utility of PAR-CLIP by quantifying background binding to mRNAs and lncRNAs." Genome Biol 15 (2014): R2.
- Reyes-Herrera, Paula H., and Elisa Ficarra. "Computational Methods for CLIP-seq Data Processing." Bioinformatics and Biology insights 8 (2014): 199.
- Uren, Philip J., et al. "Site identification in high-throughput RNA–protein interaction data." Bioinformatics 28.23 (2012): 3013-3020.

References

- Bahrami-Samani, Emad, et al. "Leveraging cross-link modification events in CLIP-seq for motif discovery." Nucleic acids research (2014): gku1288.
- Wang, Tao, Yang Xie, and Guanghua Xiao. "dCLIP: a computational approach for comparative CLIP-seq analyses." Genome Biol 15 (2014): R11.
- Chen, Beibei, et al. "PIPE-CLIP: a comprehensive online tool for CLIP-seq data analysis." Genome Biol 15 (2014): R18.