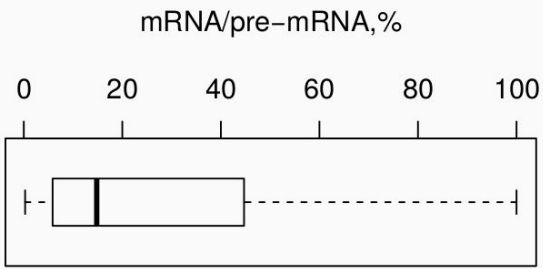
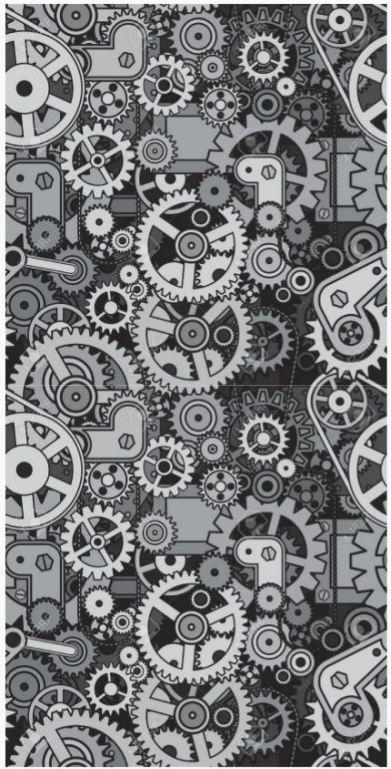
On the practical use of enhanced crosslinking and immunoprecipitation (eCLIP) data

Dmitri Pervouchine
Skolkovo Institute for Science and Technology

Eukaryotic RNA Processing is Incredibly Complex

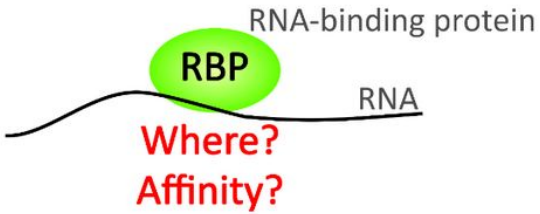


- **Most of nascent RNA is going to waste**
- Splicing, editing, cleavage and polyadenylation are co-transcriptional and intricately coupled with each other
- RNA is densely coated by proteins (RBP)
- RNA forms secondary and tertiary structure that affect all processing steps

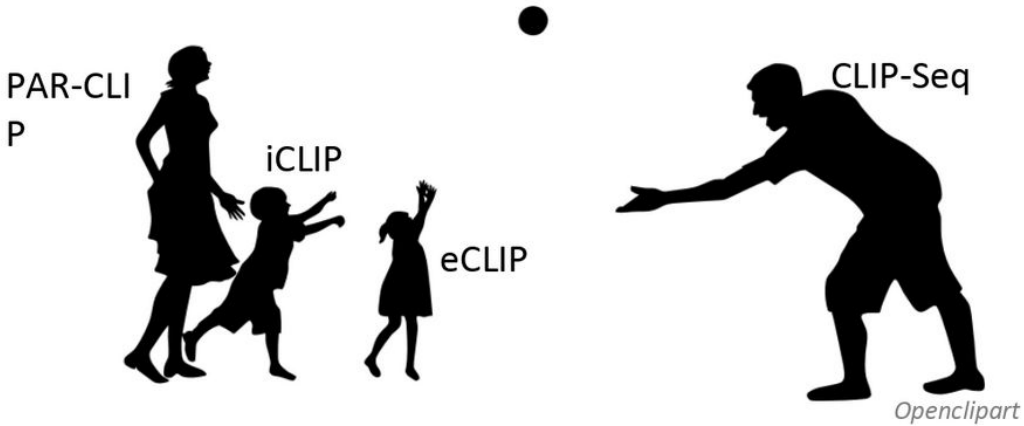


Crosslinking and immunoprecipitation family (CLIP)

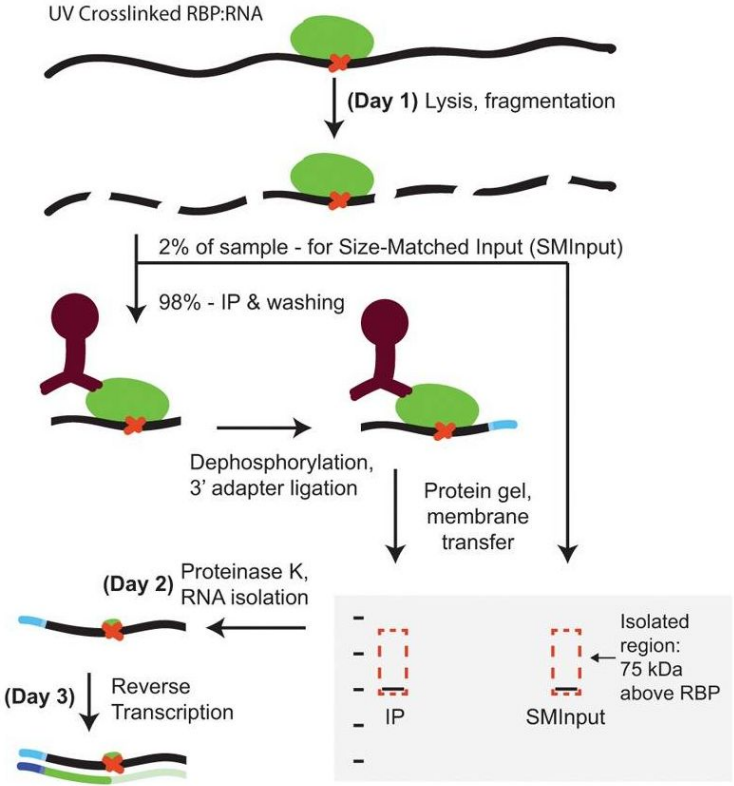
- **CLIP-Seq**

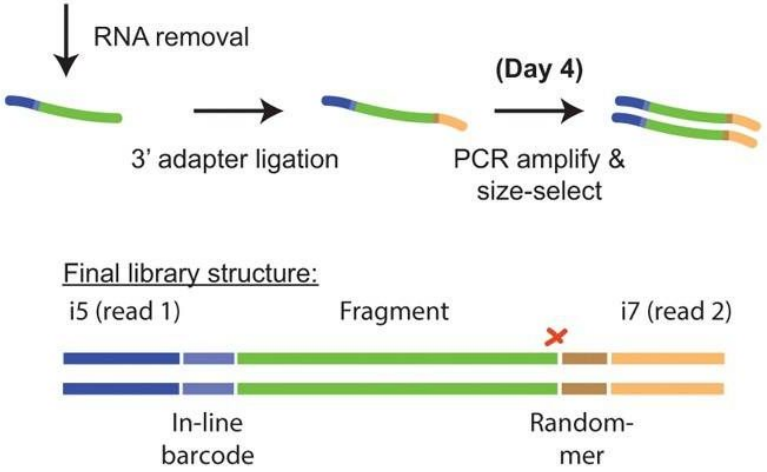


- **PAR-CLIP** (**P**hoto**a**ctivatable **R**ibonucleoside-Enhanced)
- **iCLIP** (**I**ndividual-Nucleotide Resolution)
- **eCLIP** (**E**nhanced)



Enhanced crosslinking and immunoprecipitation (eCLIP)





From: Van Nostrand et al, Nat Methods. 2016 Jun; 13(6): 508–514.

Large Panel of Expression and Binding Assays

- ENCODE: shRNA-KD+RNA-seq, eCLIP in HepG2 and K562

	CRISPR	eCLIP	shRNA		CRISPR	eCLIP	shRNA
AUH		✓	✓	NCBP2		✓	✓
BUD13		✓	✓	PRPF8		✓	✓
CSTF2T		✓	✓	PTBP1	✓	✓	✓
DDX21	✓		✓	QKI		✓	✓
DDX3X		✓	✓	RBFOX2		✓	✓
DDX55		✓	✓	RBM15		✓	✓
DDX6		✓	✓	RBM22		✓	✓
DHX30		✓	✓	SF3B4		✓	✓
EFTUD2		✓	✓	SLTM		✓	✓
FAM120A		✓	✓	SMNDC1		✓	✓
FASTKD2		✓	✓	SND1		✓	✓
GTF2F1		✓	✓	SRSF1		✓	✓
HNRNPA1		✓	✓	SRSF7	✓	✓	✓
HNRNPF	✓		✓	TAF15		✓	✓
HNRNPK		✓	✓	TBRG4		✓	✓
HNRNPM		✓	✓	TIA1		✓	✓
HNRNPU		✓	✓	TRA2A		✓	✓
HNRNPUL1		✓	✓	TROVE2		✓	✓
IGF2BP1	✓	✓	✓	U2AF1		✓	✓
ILF3		✓	✓	U2AF2		✓	✓
LARP4		✓	✓	UCHL5		✓	✓
LARP7		✓	✓	XRCC6		✓	✓
LIN28B		✓	✓	XRN2		✓	✓
LSM11		✓	✓				

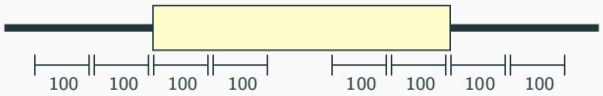
- ENCODE: RNA-seq (nuclear/cytosolic compartments) in HepG2 and K562

Prediction of Exon Inclusion from RBP binding

- Input: n exons and k RNA-binding proteins (RBP)
- $n = 125,000$, RNA-seq K562 and HepG2 (Gingeras, ENCODE)
- $k = 90$ eCLIPs, enhanced crosslinking and IP (Yeo, ENCODE)
- $\Psi = \text{PSI} = \underline{\text{P}}\text{ercent-}\underline{\text{S}}\text{pliced-}\underline{\text{I}}\text{n} \in [0, 1]$
- RBP_{ix} = eCLIP enrichment over control, $i = 1 \ldots k, x = 1 \ldots n$
- We want a classifier to predict exon inclusion Ψ_x from RBP_{ix}

$$\begin{cases} \Psi_1 = f(RBP_{11}, RBP_{21}, \ldots, RBP_{k1}) \\ \vdots \\ \Psi_x = f(RBP_{1x}, RBP_{2x}, \ldots, RBP_{kx}) \\ \vdots \\ \Psi_n = f(RBP_{1n}, RBP_{2n}, \ldots, RBP_{kn}) \end{cases}$$

Random Forest Classifier Predicts Ψ with high accuracy



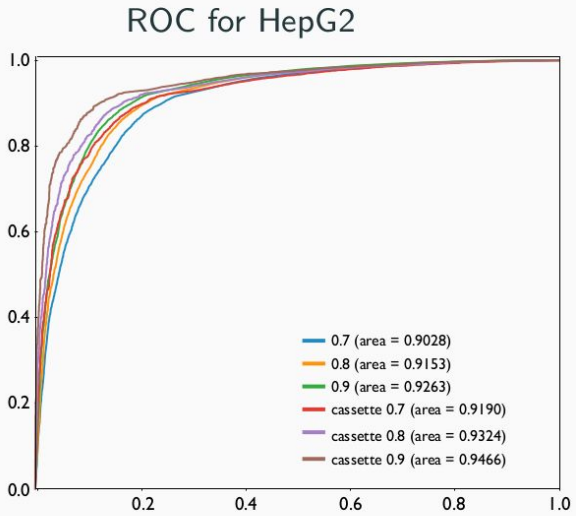
eCLIP is a compound signal

$RBP_{ix} = (z_1, \ldots, z_w)$

Ψ_x is discretized based on a threshold

$n_1 = 125,000$ annotated exons

$n_2 = 24,000$ cassette exons

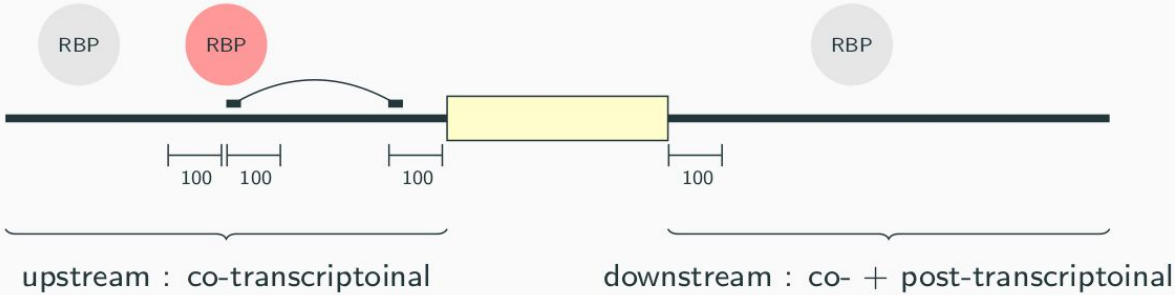


- Up to 90-95% accuracy
- Better than SVM and LR

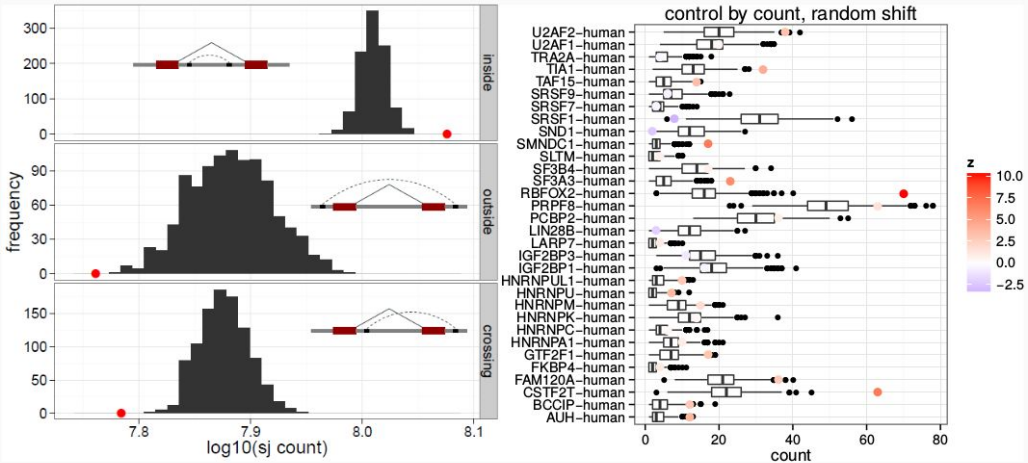
Convolution of eCLIP data with long-range RNA structure

- Without secondary structure: $RBP_i(x) = \text{eCLIP } i \text{ near site } x$

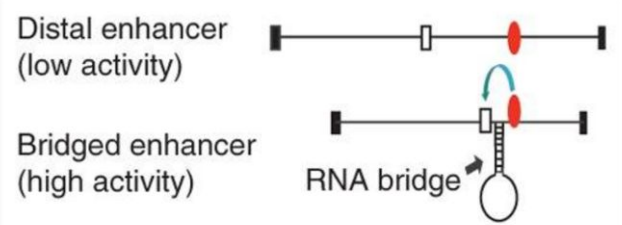
- With secondary structure : $RBP'_i(x) = \int_{TSS}^{TTS} RBP_i(y)p(x,y)dy$



Long-range RNA structure improves RF models



RBFOX2 regulates splicing through RNA bridges[2]

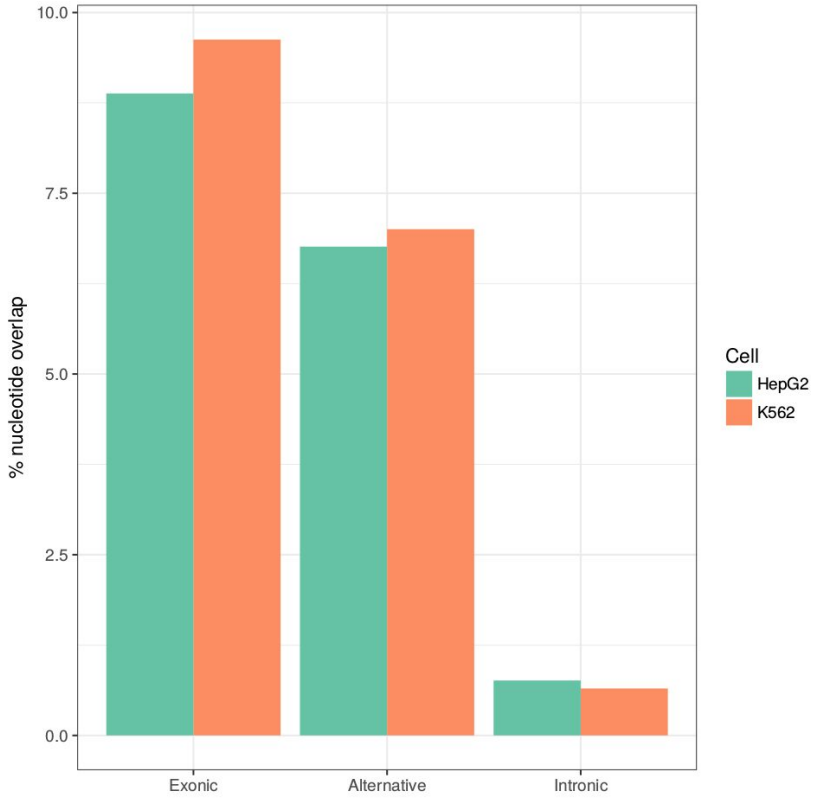


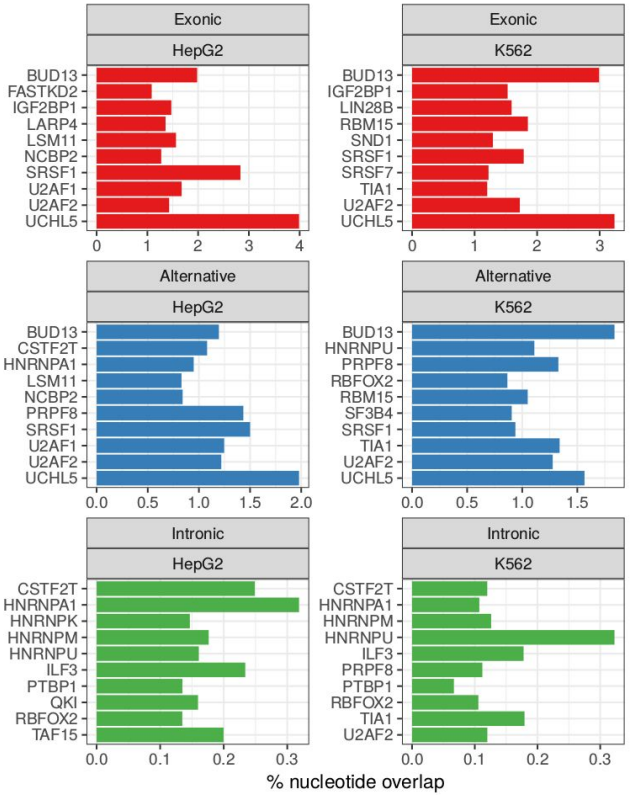
[2] Lovci *et al.* Nat Struct Mol Biol. 2013 Dec;20(12):1434-42

Now a warning

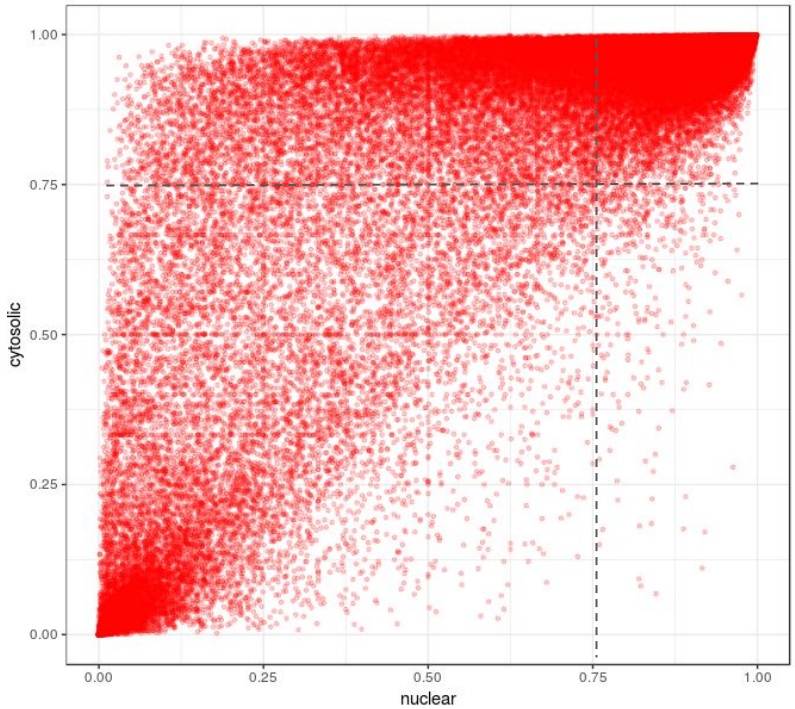
- **Exonic features are much more important than intronic features**
- **Factors with the highest importance are non-specific**
- **Reactivity to splicing factor KD is inconsistent with feature importance**
- **Feature importance strongly correlates with the number of eCLIP peaks**

eCLIP peaks are strongly depleted in introns





Co-transcriptional and post-transcriptional splicing



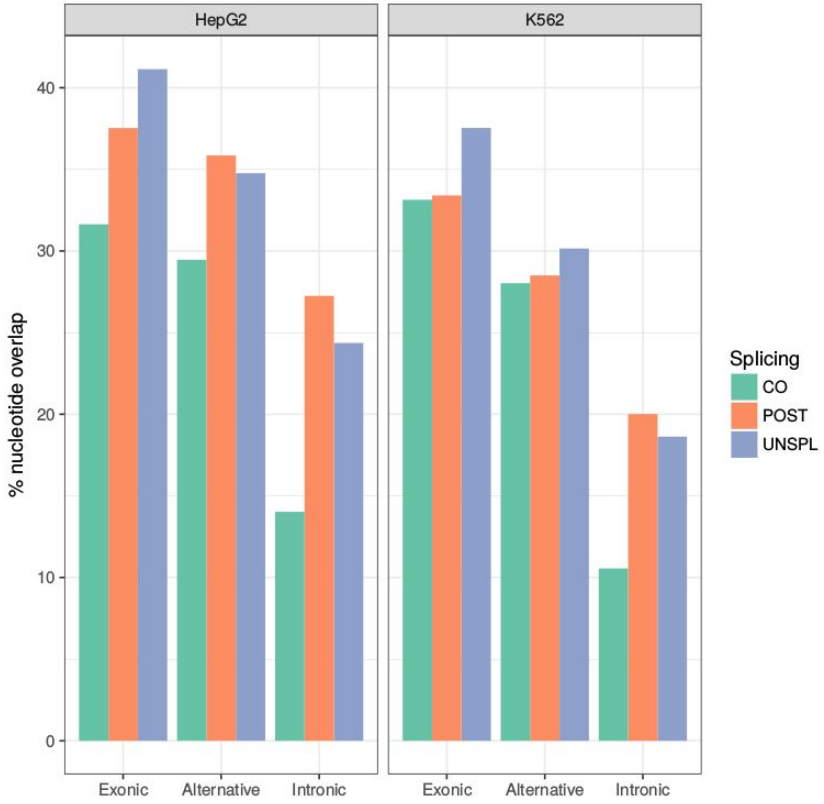
CoSI index = completeness of splicing

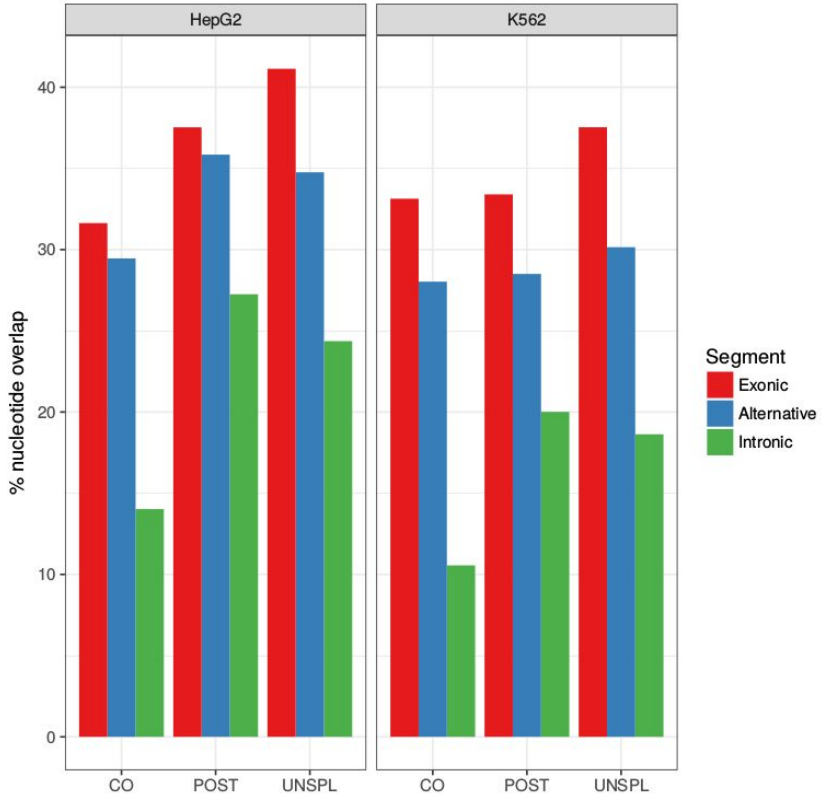
CoSI > threshold => spliced

CoSI < threshold => unspliced

		nucleus	
		unspliced	spliced
cytosol	spliced	**POST**	**CO**
	unspliced	**UNSPL**	**artifact**

eCLIP peaks are most strongly depleted in CO introns





Conclusions (points to keep in mind)

1. eCLIP signal is strongly confounded by co-transcriptional splicing
2. Co-transcriptionally-spliced introns have less chance to be sampled in eCLIP
3. The imbalance between co- and post-transcriptional introns is different for different factors, and also varies between cell lines
4. Peak calling has to be done differently in exons and in introns
5. Different significance and logFC thresholds are needed for different factors
6. Other data flavours (such as RNA duplex maps) may help interpret eCLIPs

Acknowledgements

Skolkovo Institute for Science and Technology

Dasha Romanovskaya
Artem Baranosky
Timofei Ivanov
Stepan Denisov
Alexey Samosyuk
Dmitri Svetlichnyy
Ekaterina Khrameeva
Alexander Tashkeev
Svetlana Kalmykova
Vera Rybko
Timofei Zatsepin

Higher School of Economics

Alexey Mironov
Kim Adameiko
Sudhanshu Sharma
Natalia Didkovskaya

Centre de Regulació Genòmica

Roderic Guigó
Emilio Palumbo
Beatrice Borsari
Lourdes Pena-Castillo
Diego Garrido

Moscow State University

Yaroslav Popov
Olga Vasutkina
Marina Kalinina
Dima Skvorzov
Olga Dontsova
Zoya Chervontseva
Andrei Mironov
Mikhail Gelfand

Other Institutes

Thomas Gingeras (CSHL)
Alex Dobin (CSHL)
Olga Kalinina (MPI für Informatik)
Rory Johnson (Bern University)
Andrés Lanzós (Bern University)
Marina Granovskaya (University College Dublin)
Barbara Uszczyńska (Warsaw University)
Charles Steward (Congenica LTD)
Adam Frankish (EBI)
Andy Berry (EBI)