



# Integrative RNA Biology Special Interest Group Meeting

July 10th 2015



# Content

<b>Content</b> .....	<b>3</b>
<b>Program - Friday, July 10</b> .....	<b>5</b>
<b>Sponsors</b> .....	<b>7</b>
<b>Abstracts</b> .....	<b>9</b>
RNA splicing regulation in development and disease	9
Predicting the functional consequences of alternative splicing variations with Exon Ontology	11
Integrative analysis of nineteen protein-RNA interaction CLIP data sets using orthogonal matrix factorization	12
Combining structure probing experiments on RNA mutants with evolutionary information reveals RNA-binding interfaces	14
Genome-wide analysis of co-transcriptional splicing	16
Long-non coding RNAs as a source of new peptides	17
Ribosome motion revealed by RNA sequencing	18
Apa-db.org: an integrative polyadenylation research platform	19
From the translome to the small proteome	21
Integrative analysis of allele-specific RNA biology	22
PGRN Network-wide Project: Transcriptome Analysis of Pharmacogenes in Human Tissues	23
Widespread disruption of transcription termination in HSV-1 infection	24
Relative abundances of transcript isoforms are predictive of tumor staging and survival in 12 cancer types	26
Genomic alterations dysregulate cancer genes by modulating microRNA activity	27
A highway to the sites of RNA silencing: Cell entry routes of exosomes as a novel paradigm for therapeutic RNA delivery	29
Probabilistic modelling of RNA structure probing data	30
Bayesian methods for transcript expression estimation and differential expression calling from RNA-Seq data	32
ASpli: an integrative R package for the analysis of alternative splicing using RNA-Seq	33
The combinatorial effects of RNA Polymerase II elongation rate, nucleosome occupancy and chromatin organization on alternative splicing	34
RNA aptamers and methods for their optimal utilization	36
RNA binding by <i>C. elegans</i> splicing factors in vitro and in vivo	38
The Human Inosinome Atlas	40
Transcriptome analysis reveals thousands of targets of nonsense-mediated mRNA decay that offer clues to the mechanism in different species	42
Network of Splice Factor Regulation by Alternative Splicing Coupled with Nonsense Mediated mRNA Decay	44
Skip and bin: Widespread exon-skipping triggers degradation by nuclear RNA surveillance	45
Mutant U2AF1 alters splicing in hematopoietic cells in vitro and in vivo	46
Lysinylation and Asparaginylation Identity Tangling Through a Rapidly-Evolving Ion-Binding Pocket in <i>Drosophila</i> tRNA	48
The expanding landscape of snoRNAs: a tale of the trials and tribulations of deep-sequencing structured small RNAs	49
The importance of conservation and homology in alternative splicing at the protein level	51
STATegra: Statistical and Bioinformatics tools for multi-omics data integration	53
Novel Approach to Identifying Cleavage Sites of Bacterial Toxin-Antitoxin Systems	54
Complementary mapping approaches improve circRNA detection	55
Dumpster Diving: Finding the source of every last RNA-Seq read	57
<b>Organizing committee</b> .....	<b>59</b>



# Program - Friday, July 10

## Session 1 Mechanisms of regulation and function

- 08:30 - 08:40 Opening notes
- 08:40 - 09:15 **Invited speaker: David Elliot.** *RNA splicing regulation in development and disease.*
- 09:15 - 09:30 **Léon-Charles Tranchevent.** *Predicting the functional consequences of alternative splicing variations with Exon Ontology.*
- 09:30 - 09:45 **Martin Strazar.** *Integrative analysis of nineteen protein-RNA-interaction CLIP data sets using orthogonal matrix factorization.*
- 09:45 - 10:00 **Vladimir Reinhartz.** *Combining structure probing experiments on RNA mutants with evolutionary information reveals RNA-binding interfaces.*
- 10:00 - 10:15 **Maria Carmo-Fonseca.** *Genome-wide analysis of co-transcriptional splicing.*
- 10:15 - 10:45 Coffee Break

## Session 2 RNA variation and translation

- 10:45 - 11:20 **Invited speaker: Mar Alba.** *Long non-coding RNAs as a source of new peptides.*
- 11:20 - 11:35 **Liana F. Lareau.** *Ribosome motion revealed by RNA sequencing.*
- 11:35 - 11:50 **Juan Mata.** *From the translome to the small proteome.*
- 11:50 - 12:20 **Invited speaker: Cathal Soighe** *Integrative analysis of allele specific RNA biology.*
- 12:30 - 13:30 Lunch

## Session 3 RNA and disease

- 13:45 - 14:00 **Steven Brenner.** *PGNR Network-wide Project: Transcriptome Analysis of Pharmacogenes in Human Tissues.*
- 14:00 - 14:15 **Caroline C. Friedel.** *Widespread disruption of transcription termination in HSV-1 infection.*
- 14:15 - 14:30 **Juan L. Trincado.** *Relative abundances of transcript isoforms are predictive of tumor staging and survival in 12 cancer types.*
- 14:30 - 14:45 **Pavel Sumazin.** *Genomic alterations dysregulate cancer genes by modulating microRNA activity.*
- 14:45 - 15:20 **Invited speaker: Nicole Meisner.** *A highway to the sites of RNA silencing: Cell entry of exosomes as a novel paradigm for therapeutic RNA delivery.*
- 15:20 - 16:00 Coffee Break (and start of poster session)
- 16:00 - 17:30 **Poster session**

## Session 4 Probabilistic modeling of RNA

- 17:30 - 17:45 **Alina Selega.** *Probabilistic modeling of RNA structure probing data.*
- 17:45 - 18:20 **Invited speaker: Magnus Rattray.** *Bayesian methods for transcript expression estimation and differential expression calling from RNA-Seq data.*
- 18:20 - 19:00 Concluding notes, poster prize announcement
- 19:30 IRB-SIG dinner**



# Sponsors

RNA Society







# Abstracts

## RNA splicing regulation in development and disease

David J. Elliott\*

Institute of Genetic Medicine, Newcastle University

\*To whom correspondence should be addressed: David.Elliott@ncl.ac.uk

---

### BACKGROUND

My group is investigating alternative RNA splicing in normal development and disease.

### RESULTS

Recent results have included the identification of transcriptome-wide targets of two important RNA splicing regulators. Firstly, we recently globally identified Tra2 $\beta$ -target exons in breast cancer. The RNA splicing regulator Tra2 $\beta$  is conserved in all animals, and in humans becomes up-regulated in some breast cancers. Tra2 $\beta$  target exons include the key cell cycle regulator CHK1, and also exons in other genes implicated in DNA and chromosome biology. We showed Transformer2 $\beta$  is essential for expression of CHK1 protein, and for this reason helps breast cancer cells prevent accumulation of replication stress [1]. Secondly, we recently used a mouse genetics approach to identify splicing targets of the T-STAR RNA binding protein in the mouse brain. This revealed that T-STAR controls regional splicing patterns of the Neurexin1-3 genes that encode proteins with important roles in the synapse [2]. We have also monitored widespread changes in the splicing landscape that occur during male meiosis in the mouse [3].

### CONCLUSIONS

I will present some of our recent work on these projects.

### REFERENCES

- [1] Human Tra2 proteins jointly control a CHEK1 splicing switch among alternative and constitutive target exons. Best A, James K, Dalgliesh C, Hong E, Kheirolah-Kouhestani M, Curk T, Xu Y, Danilenko M, Hussain R, Keavney B, Wipat A, Klinck R, Cowell IG, Cheong Lee K, Austin CA, Venables JP, Chabot B, Santibanez Koref M, Tyson-Capper A, Elliott DJ. Nat Commun. 2014 Sep 11;5:4760. doi: 10.1038/ncomms5760.
- [2] The tissue-specific RNA binding protein T-STAR controls regional splicing patterns of neurexin pre-mRNAs in the brain. Ehrmann I, Dalgliesh C, Liu Y, Danilenko M, Crosier M, Overman L, Arthur HM, Lindsay S, Clowry GJ, Venables JP, Fort P, Elliott DJ. PLoS Genet. 2013 Apr;9(4):e1003474. doi: 10.1371/journal.pgen.1003474. Epub 2013 Apr 25
- [3] The splicing landscape is globally reprogrammed during male meiosis. Schmid R, Grellscheid

SN, Ehrmann I, Dalgliesh C, Danilenko M, Paronetto MP, Pedrotti S, Grellscheid D, Dixon RJ, Sette C, Eperon IC, Elliott DJ.  
Nucleic Acids Res. 2013 Dec;41(22):10170-84. doi: 10.1093/nar/gkt811. Epub 2013 Sep 12.

# Predicting the functional consequences of alternative splicing variations with Exon Ontology

Léon-Charles Tranchevent<sup>1,2,3,\*</sup>, Emilie Chautard<sup>1,2,3,4</sup>, François-Olivier Desmet<sup>1,2,3</sup>, Hussein Mortada<sup>1,2,3</sup>, Marion Dubarry<sup>1,2,3</sup>, Clara Benoit-Pilven<sup>1,2,3</sup>, and Didier Auboeuf<sup>1,2,3</sup>

<sup>1</sup> Inserm UMR-S1052, CNRS UMR5286, Cancer Research Centre of Lyon, Lyon, France, <sup>2</sup> Université de Lyon 1, Villeurbanne, France, <sup>3</sup> Centre

Léon Bérard, Lyon, France, <sup>4</sup> CNRS UMR5558, Laboratoire de Biométrie et Biologie Evolutive, INRIA Bamboo, Villeurbanne, France

\*To whom correspondence should be addressed: leon-charles.tranchevent@inserm.fr

---

## BACKGROUND

Alternative splicing allows the production of several protein isoforms from a single gene. These isoforms have different protein sequence, and often diverse or even antagonistic functions [1]. In the recent years, the use of high-throughput technologies has revealed that alternative splicing is massively deregulated in many experimental conditions [2]. However, it is often difficult to decipher the functional consequences of the numerous observed splicing variations because of the lack of functional information at the exon level.

To circumvent that problem, we introduce a computational strategy termed Exon Ontology that relies on the functional annotation of the exons in order to predict the consequences of their inclusion or skipping.

## RESULTS

We have first associated human exons to functional features using reference databases and tools. The Exon Ontology that structures these features thus describes catalytic domains, binding domains, structural elements, localization motifs, post-translational modifications, and physicochemical features.

We have then developed three web based interfaces that use this Exon Ontology. The first interface allows users to visualize, for a single gene, the features associated to its exons. With the second interface, users can investigate a list of joint alternative splicing variations in order to detect the global consequences. The last interface is dedicated to the analysis of joint alternative splicing variations through network and pathway based methods.

This strategy has been used to study the alternative splicing variations observed between epithelial and mesenchymal cells [3]. Our computational tools predict that numerous variations change the localization of the associated proteins, which was then experimentally validated for some cases.

## CONCLUSIONS

We proposed a strategy to decipher the functional consequences of alternative splicing variations. In addition, we demonstrated its usefulness through a comparison of distinct cell types showing that protein localization can indeed be modulated by alternative splicing.

## REFERENCES

1. Kelemen O, Convertini P, Zhang Z, Wen Y, Shen M, Falaleeva M, Stamm S., Function of alternative splicing., *Gene*. (2013)
2. Oltean S, Bates DO., Hallmarks of alternative splicing in cancer., *Oncogene*. (2014)
3. Mallinjoud P, et al., Endothelial, epithelial, and fibroblast cells exhibit specific splicing programs independently of their tissue of origin., *Genome Res*. (2014)

# Integrative analysis of nineteen protein-RNA interaction CLIP data sets using orthogonal matrix factorization

Martin Stražar<sup>1\*</sup>, Jernej Ule<sup>2</sup>, and Tomaž Curk<sup>1</sup>

<sup>1</sup> University of Ljubljana, Faculty of Computer and Information Science, Večna pot 113, SI 1000, Slovenia,

<sup>2</sup> Department of Molecular Neuroscience, UCL Institute of Neurology, Queen Square, London, WC1N 3BG, UK

\*To whom correspondence should be addressed: author@institute.edu

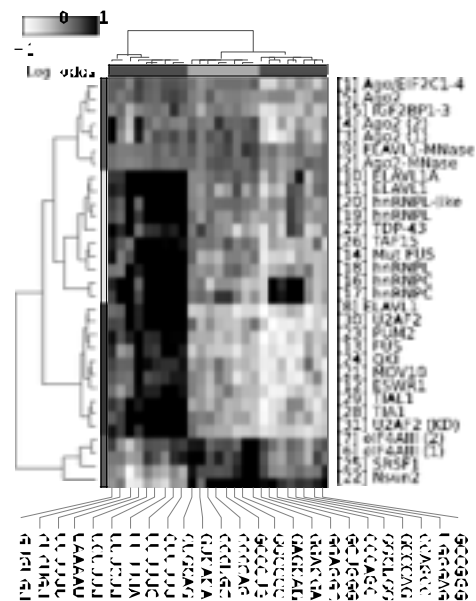
## BACKGROUND

RNA binding proteins (RBPs) regulate different of post-transcriptional control mechanisms of gene expression, including splicing, RNA editing, and other. Contemporary methods modeling RBP interaction on RNA assume precise structural information, which available for few proteins [1, 2]. On the other side, increasingly available experimental data enables simultaneous integrative study of different data sources on a number of RBPs.

## RESULTS

We have developed an algorithm for orthogonal non-negative matrix factorization (ONMF) and used it to integrate various omics data on protein-RNA interactions describing nucleotide regions. The algorithm enables fusion of multiple data sources by joint dimensionality reduction, enabling the discovery of multiple, non-overlapping modules within the data. Compared to alternative non-negative matrix factorization models, ONMF infers more accurate predictive models and enables interpretation of models, due to the orthogonality constrains. We analyzed publicly available protein-RNA interaction data on 19 RBPs, the largest such study to date. We confirmed a number of biologically relevant patterns and discovered a number of novel patterns describing protein-RNA interaction sites. These include RNA structure, RNA sequence motifs and positioning, cooperation and competition in binding among proteins to same target RNAs, affinity to interact with specific types of gene regions, and functional annotation of targeted genes. We used these patterns to cluster RBPs and discover functionally related groups. For example, we find that hnRNP proteins bind to U-rich motifs in introns to regulate splicing (hnRNPs, U2AF2, ELAVL1, TDP-43, TAF15, FUS, QKI). Those that regulate splicing (SR), spliced mRNA (eIF3E3) or 3'UTR (Ago, IGF2BP) mRNA bind to GC-rich motifs (see Figure).

Type to enter text



For example, we find that hnRNP proteins bind to U-rich motifs in introns to regulate splicing (hnRNPs, U2AF2, ELAVL1, TDP-43, TAF15, FUS, QKI). Those that regulate splicing (SR), spliced mRNA (eIF3E3) or 3'UTR (Ago, IGF2BP) mRNA bind to GC-rich motifs (see Figure).

## CONCLUSIONS

The inferred motifs and other features included in the predictive models agree with a published *in vitro* study (Ray *et al.*, 2013) and identified a number of promising candidates for further investigation. Our experimental findings confirm the utility of ONMF for data integration where sparse, modular and interpretable models are desired. Data on RNA binding proteins is growing rapidly, emphasizes the need of integrative analyses that jointly consider all available information sources [3]. ONMF-like methods provide means for such integrative analysis.

## REFERENCES

1. T. Puton, *et al.*, “Computational methods for prediction of protein-RNA interactions,” *J. Struct. Biol.*, vol. 179, no. 3, pp. 261–8, Sep. 2012.
2. D. Cirillo, *et al.*, “Predictions of protein-RNA interactions,” *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, vol. 3, β
3. J. König, *et al.* “Protein–RNA interactions: new genomic technologies and perspectives,” *Nat. Rev. Genet.*, vol. 13, no. February, pp. 77–83, Feb. 2012.
4. D. Ray, *et al.* “A compendium of RNA-binding motifs for decoding gene regulation,” *Nature*, vol. 499, pp. 172–7, Jul. 2013.

# Combining structure probing experiments on RNA mutants with evolutionary information reveals RNA-binding interfaces

Vladimir Reinharz<sup>1</sup>, Yann Ponty<sup>2</sup>, and Jérôme Waldispühl<sup>1\*</sup>

<sup>1</sup> School of Computer Science, McGill University, Montreal, Canada,

<sup>2</sup> Laboratoire d'informatique, école Polytechnique, Palaiseau, France

\*To whom correspondence should be addressed: jerome.waldispuhl@mcgill.ca

---

## BACKGROUND

In vitro methods such as selective 2-hydroxyl acylation analyzed by primer extension (SHAPE) [1] and parallel analysis of RNA structure (PARS) [2], provide transcriptome-wide measurements of RNA structure at single-nucleotide resolution in vitro. These data enabled us to significantly improve the accuracy of RNA secondary structure prediction methods [3, 4]. Recently, R. Das and colleagues introduced the mutate-and-map protocol, which consists in obtaining the SHAPE data of one RNA and all (or a large number of) 1-point mutants [5], which can be used to enhance the structure prediction. A distinct, yet complementary, approach to analyze RNAs makes use of the evolutionary information encapsulated within multiple sequence alignments (MSAs). The latter provides an alternate signal which is often key to understand and characterize the origin and structure of functional motifs [3, 4]. To date, both approaches have not been combined and even less reconciled. Nonetheless, systematic mutations such as those conducted in the mutate-and-map protocol enable us to probe the evolutionary landscape of a molecule, which in turn can be used to reveal nucleotide patterns in the fitness landscape. To capture this signal, it is essential to design a formal framework that calculates correlations between the genetic robustness of the structural profile (obtained from mutate-and-map experiments) and the evolutionary information available for this molecule (usually contained in MSAs).

## RESULTS

The main contribution of this work is to show that neutral theory principles can be combined with structure probing experiments to calculate complex evolutionary signals embedded in ncRNA sequences. We implement our model in a software named aRNhAck which aims to be a model for a new family of RNA sequence/structure analysis techniques. We analyze mutate-and-map data sets available on the RNA Mapping Database [6]. Our experiments reveal non-trivial long-range dependencies within ncRNA primary structures of 5S ribosomal RNA and the c-di-GMP riboswitch. We investigate the biological significance of these patterns by looking at the distribution of these nucleotides on the RNA 3D structures. Interestingly, we find significant correlations between the sets of nucleotides produced by our method and those identified as participating in RNA-RNA and RNA-protein interfaces.

## CONCLUSIONS

Programs predicting RNA-Protein or RNA-RNA interactions aim to identify potential molecular targets from a library, and predict the best fits. By contrast, aRNhAck focuses on the sole biochemical and evolutionary properties of the RNA being analyzed. It enables, for the first time without prior knowledge of potential partners, the identification of RNA-RNA and RNA-Protein interfaces, i.e. sets of critical nucleotides possibly implicated in the molecular functions. This information can be in turn used to identify molecular targets. This result illustrates of the usefulness of the signal extracted by aRNhAck, but only constitutes a first step towards a more comprehensive analysis of the interplay between structure and evolution. For instance, we envision to use the nucleotide networks detected with aRNhAck to predict non-canonical interactions and 3D motifs of an RNA molecules.

## REFERENCES

1. Wilkinson K. A. et al. Selective 2-hydroxyl acylation analyzed by primer extension (shape): quantitative RNA structure analysis at single nucleotide resolution. *Nat Protoc*, 2006.
2. Kertesz M. et al. Genome-wide measurement of RNA secondary structure in yeast. *Nature*, 2010.
3. Washietl S., et al. RNA folding with soft constraints: reconciliation of probing data and thermodynamic secondary structure prediction. *Nucleic Acids Res*, 2012.
4. Deigan K. E. et al. Accurate SHAPE-directed RNA structure determination. *Proc Natl Acad Sci USA*, 2009.
5. Kladwang W. et al. A two-dimensional mutate-and-map strategy for non-coding RNA structure. *Nat Chem*, 2011.
6. Cordero P. et al. An RNA mapping database for curating RNA structure mapping experiments. *Bioinformatics*, 2012.

# Genome-wide analysis of co-transcriptional splicing

Tomás Gomes<sup>1</sup>, Takayuki Nojima<sup>2</sup>, Ana Rita Grosso<sup>1</sup>, Nicholas J. Proudfoot<sup>2</sup> and Maria Carmo-Fonseca<sup>1\*</sup>

<sup>1</sup>Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, 1649-028 Lisboa, Portugal

<sup>2</sup>Sir William Dunn School of Pathology, University of Oxford, Oxford OX1 3RE, UK

\*To whom correspondence should be addressed: carmo.fonseca@medicina.ulisboa.pt

---

## BACKGROUND

Expression of genetic information in eukaryotes involves a series of interconnected processes that ultimately determine the quality and amount of proteins in the cell. In particular, there is extensive cross-talk between transcription and pre-mRNA splicing. To better understand the mechanisms of co-transcriptional splicing we carried out a genome-wide analysis of nascent RNA complexes immunoprecipitated by antibodies against endogenous Pol II with distinct CTD modifications (1).

## RESULTS

We found splicing intermediates associated with Pol II CTD Ser5P, suggesting that during co-transcriptional splicing exons cleaved at the 5' splice site remain tethered to Pol II harboring the Ser5P modification. We further observed spliced products associated with both Pol II CTD Ser5P and Ser2P, suggesting that co-transcriptional splicing is coupled to dynamic shifts in CTD modification. Mapping the position of Pol II complexes containing fused exons indicates that introns are excised as the 3' splice site emerges from the polymerase exit channel. We also detected selective accumulation of Pol II over spliced exons, suggesting that exons are transcribed slower than introns in a splicing-dependent manner.

## CONCLUSIONS

Taken together, these results underscore the importance of Pol II CTD modifications for the splicing reaction.

## REFERENCES

Nojima T., Gomes T., Grosso A.R., Kimura H., Dye M.J., Dhir S., Carmo-Fonseca M., and Proudfoot N. Mammalian NET-seq reveals genome-wide nascent transcription coupled to RNA processing. *Cell* 161: 526-540 (2015).



# Long-non coding RNAs as a source of new peptides

Mar Albà

Universitat Pompeu Fabra, Dr. Aiguader 88, E08003 Barcelona, Spain <sup>2</sup>Centre for Genomic Regulation, Dr. Aiguader 88, E08003 Barcelona, Spain

## BACKGROUND

High throughput sequencing of the eukaryotic transcriptome has resulted in the identification of many transcripts that lack long conserved open reading frames and which have been classified as long non-coding RNAs (lncRNAs). The majority of them are expressed at low levels and do not have a known function. Despite having been annotated as non-coding RNAs, the sequencing of ribosome-protected RNA fragments (ribosome profiling) has revealed that many of them are scanned by ribosomes and are likely to translate short peptides. We have examined single nucleotide polymorphism (SNP) data and have found evidence of purifying selection on the corresponding coding regions. We propose that lncRNAs provide the necessary raw material for the evolution of new functional proteins during evolution

# Ribosome motion revealed by RNA sequencing

Dustin H Hite<sup>1</sup>, Gregory J Hogan<sup>1</sup>, Robert J Tunney<sup>2</sup>, Patrick O Brown<sup>1</sup>, and **Liana F Lareau**<sup>3\*</sup>

<sup>1</sup> Department of Biochemistry and Howard Hughes Medical Institute, Stanford University, <sup>2</sup> Center for Computational Biology, University of California, Berkeley,

<sup>3</sup> California Institute for Quantitative Biosciences, University of California, Berkeley

\*To whom correspondence should be addressed: lareau@berkeley.edu

---

## BACKGROUND

During translation elongation, the ribosome ratchets along its mRNA template, incorporating each new amino acid and translocating from one codon to the next. The elongation cycle requires dramatic structural rearrangements of the ribosome.

## RESULTS

We show here that deep sequencing of ribosome-protected mRNA fragments reveals not only the position of each ribosome but also, unexpectedly, its particular stage of the elongation cycle [1]. Sequencing reveals two distinct populations of ribosome footprints, 28–30 nucleotides and 20–22 nucleotides long, representing translating ribosomes in distinct states, differentially stabilized by specific elongation inhibitors. We find that the balance of small and large footprints varies by codon and is correlated with translation speed.

## CONCLUSIONS

The ability to visualize conformational changes in the ribosome during elongation, at single-codon resolution, provides a new way to study the detailed kinetics of translation and a new probe with which to identify the factors that affect each step in the elongation cycle.

## REFERENCES

1. Lareau, L.F., Hite, D.H., Hogan, G.J, and Brown, P.O. Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments. *eLife* **3** (2014).

# Apa-db.org: an integrative polyadenylation research platform

Gregor Rot<sup>1\*</sup>, Zhen Wang<sup>4</sup>, Ina Huppertz<sup>2</sup>, Miha Modic<sup>5</sup>, Tomaž Curk<sup>3</sup>, Christian von Mering<sup>1</sup>,  
Jernej Ule<sup>2</sup>

<sup>1</sup>Institute of Molecular Life Sciences and Swiss Institute of Bioinformatics, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland,

<sup>2</sup>Department of Molecular Neuroscience, UCL Institute of Neurology, Queen Square, London WC1N 3BG, UK,

<sup>3</sup>Faculty of Computer and Information Science, University of Ljubljana, Traska Cesta 25, 51-1000 Ljubljana, Slovenia,

<sup>4</sup>Institut de Biologie de l'ENS, (IBENS), 46 rue d'Ulm, Paris F-75005, France,

<sup>5</sup>Institute of Stem Cell Research, German Research Center for Environmental Health, Helmholtz Center Munich, 85764 Neuherberg, Germany

\*To whom correspondence should be addressed: gregor.rot@uzh.ch

## BACKGROUND

The 3'-end of most protein-coding and lncRNA genes is cleaved and polyadenylated. More than half of the genes have two or more poly-A sites resulting in alternative polyadenylation [1]. Different isoforms are produced with altered coding or 3'UTR regions, leading to potential changes in function, localization and translation efficiency. In recent years increasing amounts of next-generation sequencing data on samples prepared with protocols specifically targeting 3'-ends of mRNA is becoming available (poly-A data) [2].

## RESULTS

We propose an integrative web application (apa-db) for bioinformatics analysis of poly-A data. Apa-db is a collection of interconnected software tools used to perform various analysis steps: 1. mapping of reads to the reference genome with filtering of internal priming events, 2. identifying differentially polyadenylated sites, 3. gene ontology analysis; 4. integration of RNA-protein binding data using RNA-maps, 5. motif analysis with RNA-motifs [3], 6. prediction of site-usage using SVM and other models. The front-end is written in Javascript and the server-side is coded in Python.

## CONCLUSIONS

Apa-db is a research platform for the analysis of data from targeted next-generation sequencing of mRNA 3'-ends. The integration with RNA-protein binding data and the construction of an unified polyA-database provides opportunity for the platform to become a reference point for scientists interested in the study of alternative polyadenylation.



Figure 1: apa-db.org with differential polyadenylation results

KEYWORDS: alternative polyadenylation, database, web-app, [poster]

## REFERENCES

1. de Klerk, E. et al., Alternative mRNA transcription, processing, and translation: insights from RNA sequencing, *Trends Genet.* 2015;31(3):128–139
2. Ni, T. et al., Distinct polyadenylation landscapes of diverse human tissues revealed by a modified PA-seq strategy *BMC Genomics.* 2013;14(1):615
3. Cereda, M. et al., RNAmotifs: prediction of multivalent RNA motifs that control alternative splicing, *Genome Biol.* 2014;15(1):R20

# From the translome to the small proteome

Juan Mata<sup>1\*</sup> and Caia Duncan<sup>1</sup>

<sup>1</sup>Department of Biochemistry, University of Cambridge, UK

\*To whom correspondence should be addressed: [jm593@cam.ac.uk](mailto:jm593@cam.ac.uk)

---

## BACKGROUND

Sexual development in the fission yeast *Schizosaccharomyces pombe* culminates in meiosis and sporulation. We used ribosome profiling to investigate the translational landscape of this process..

## RESULTS

We showed that the translation efficiency of hundreds of genes is regulated in complex patterns, often correlating with changes in RNA levels. Ribosome-protected fragments show a three-nucleotide periodicity that identifies translated sequences and their reading frame. Using this property, we found novel translated genes, frequent translation of RNAs annotated as non-coding, and hundreds of translated upstream open reading frames (uORFs) in leader sequences.

## CONCLUSIONS

Overall we identified over 900 unannotated translated regions of 20 codons or more, suggesting that the complexity of the fission yeast proteome is much higher than expected [1].

## REFERENCES

1. Duncan C and Mata J. Nature Structural & Molecular Biology (2014) 21, 641–647. The translational landscape of fission-yeast meiosis and sporulation

# Integrative analysis of allele-specific RNA biology

Cathal Seoighe<sup>1\*</sup>, Thong Nguyen<sup>2</sup>, and Ngoc Nguyen<sup>1</sup>

<sup>1</sup> National University of Ireland Galway, <sup>2</sup> Genentech, San Francisco

\*To whom correspondence should be addressed: Cathal.Seoighe@nuigalway.ie

---

## BACKGROUND

The two copies of human autosomal genes may be expressed at very different levels within tissues and even more so within individual cells. This imbalance can have several sources. For imprinted genes, the copy inherited from one parent is exclusively or preferentially expressed, random monoallelic expression is observed for some autosomal genes and, for many genes, one allele is consistently expressed at a higher level than the other. Allelic imbalance resulting from the presence of *cis*-acting genetic variants can give rise to expression quantitative trait loci (eQTLs); although, due to the buffering effects of regulatory feedback loops genetic variants can cause imbalance between alleles without affecting the overall mRNA expression level. We hypothesized that genetic variants affecting pre- and post-transcriptional processes can have distinct implications for how allelic imbalance is manifested in individual cells.

## RESULTS

We have analyzed allele-specific gene expression in single-cell RNA-Seq data from human cell lines, using mixture models to discriminate between allelic variation in the proportion of expressing cells and variation in the abundance of mRNA from each allele in the cells in which the gene is expressed. By comparing to ChIP-Seq data we test the hypothesis that pre-transcriptional sources of allelic imbalance are more likely to affect proportions of expressing cells rather than gene expression levels within cells. We found that mRNA stability has a strong effect on the pattern of allelic imbalance between cells, with long-lived transcripts significantly more likely to support a model of imbalance in the expression level within cells rather than in the proportion of expressing cells. This reflects the contribution of transcriptional bursting to variation in mRNA expression of short-lived genes as well as the greater impact of genetic variants with post-transcriptional effects on the expression of long-lived transcripts. We also report an integrative analysis of RNA-Seq and ribosome footprinting data as a means to identify allelic variation in the rate of mRNA translation.

## CONCLUSIONS

We have carried out an exploration of the origins and implications of different types of allelic imbalance in gene expression and show that integrative analysis of multiple genomics data types can be used to distinguish genetic variants affecting distinct stages of the gene expression process.

# PGRN Network-wide Project: Transcriptome Analysis of Pharmacogenes in Human Tissues

Courtney E. French<sup>1</sup>, Aparna Chhibber<sup>2</sup>, Sook Wah Yee<sup>2</sup>, Eric R. Gamazon<sup>3</sup>, Xiang Qin<sup>4</sup>, Elizabeth Theusch<sup>5</sup>, Amy Webb<sup>6</sup>, Scott T. Weiss<sup>7,8</sup>, Marisa W. Medina<sup>5</sup>, Erin G. Schuetz<sup>9</sup>, Alfred L. George, Jr.<sup>10</sup>, Ronald M. Krauss<sup>5</sup>, Christine Q. Simmons<sup>11</sup>, Steven E. Scherer<sup>4</sup>, Nancy J. Cox<sup>3</sup>, Kathleen M. Giacomini<sup>2</sup>, and **Steven. E. Brenner**<sup>1\*</sup>

<sup>1</sup> University of California, Berkeley, CA <sup>2</sup> University of California, San Francisco, CA <sup>3</sup> University of Chicago, Chicago, IL <sup>4</sup> Baylor College of Medicine, Houston, TX <sup>5</sup> Children's Hospital Oakland Research Institute, Oakland, CA <sup>6</sup> Ohio State University, Columbus, OH <sup>7</sup> Brigham and Women's Hospital, Boston, MA <sup>8</sup> Harvard Medical School, Boston, MA <sup>9</sup> St. Jude Children's Research Hospital, Memphis, TN <sup>10</sup> Northwestern University Feinberg School of Medicine, Chicago, IL <sup>11</sup> Vanderbilt University, Nashville, TN  
\*To whom correspondence should be addressed: [brenner@compbio.berkeley.edu](mailto:brenner@compbio.berkeley.edu)

---

## BACKGROUND

Gene expression variation is crucial to the etiologies of common disorders and the molecular underpinnings of pharmacologic traits; however, the nature and extent of this variation remains poorly understood. The NIH Pharmacogenomics Research Network (PGRN) Network-wide RNA-seq project aims to create a community resource containing quantitative information on known and novel isoforms of genes involved in therapeutic and adverse drug response (pharmacogenes).

## RESULTS

Using 18 samples from each of 5 tissues of pharmacologic importance (liver, kidney, adipose, heart, and lymphoblastoid cell lines [LCLs]), we performed transcriptome profiling by RNA-Seq with the goal of determining differences in expression of pharmacogenes across tissues and between individuals. The data were analyzed for expression quantification, and we used the JuncBASE tool developed by members of our consortium to identify and quantify splicing events.

In each of the tissues and LCLs, 11,223-15,416 genes were expressed at a substantial level. In pairwise comparisons of tissues, 105-211 pharmacogenes were differentially expressed ( $\geq 2$ -fold difference,  $FDR < 0.1$ ). For example, as expected, the CYP enzymes CYP2C19 and CYP2D6 were 10-fold and 100-fold more highly expressed in the liver than in other tissues. Other important drug metabolizing enzymes such as DPYD and TPMT showed more balanced gene expression patterns. In general, pharmacogenes were among the most variably expressed between individuals.

We also observed that 72-93% of pharmacogenes are alternatively spliced within each tissue. There was substantial variation in both annotated and novel splicing events both between tissues and between individuals. For example in SLC22A7, a gene encoding a transporter for various drugs, we found evidence of a novel alternative last exon that is variably spliced between individuals. LCLs are important pre-clinical models for human genetic studies, but they highly express less than half of pharmacogenes as compared with the 66-83% expressed at a substantial level in each of the physiological tissues. However, a number of genes like BRCA2 and SLC6A4 are much higher in LCLs than the tissues, as are alternative splice events of many genes.

## CONCLUSIONS

These studies provide mechanistic insights into pharmacogenomic findings and facilitate an understanding of the factors that lead to inter-individual differences in drug response.

# Widespread disruption of transcription termination in HSV-1 infection

Andrzej J. Rutkowski<sup>1</sup>, Florian Erhard<sup>2</sup>, Anne L'Hernault<sup>1</sup>, Thomas Bonfert<sup>2</sup>, Markus Schilhabel<sup>3</sup>, Colin Crump<sup>4</sup>, Philip Rosenstiel<sup>3</sup>, Stacey Efstathiou<sup>4</sup>, Ralf Zimmer<sup>2</sup>, **Caroline C. Friedel<sup>2\*</sup>**, and Lars Dölken<sup>1,5</sup>

<sup>1</sup>Department of Medicine, University of Cambridge, UK,

<sup>2</sup>Institut für Informatik, Ludwig-Maximilians-Universität München, Germany,

<sup>3</sup>Institut für Klinische Molekularbiologie, Christian-Albrechts-Universität Kiel, Germany,

<sup>4</sup>Division of Virology, University of Cambridge, Tennis Court Road, UK,

<sup>5</sup>Institut für Virologie, Julius-Maximilians-Universität Würzburg, Germany.

\*To whom correspondence should be addressed: caroline.friedel@bio.ifi.lmu.de

---

## BACKGROUND

Herpes simplex virus 1 (HSV-1) is an important human pathogen causing both common cold sores as well as life-threatening infections. During lytic infection, HSV-1 rapidly shuts down host gene expression, making it a paradigm for virus-induced ‘host shut-off’. In this study, we combined sequencing of 4-thiouridine (4sU)-labelled newly transcribed RNA (4sU-RNA) and ribosome profiling to study changes in RNA synthesis and processing and their impact on translation during the full course of HSV-1 lytic infection [1]. Here, 4sU-labelling was performed in one-hour intervals during the first 8 hours of infection and Ribosome profiling was performed at 0, 1, 2, 4, 6 and 8h post infection (p.i.).

## RESULTS

Lytic HSV-1 infection induced dramatic changes in transcriptional activity, with 75% of cellular protein-coding genes down-regulated in 4sU-RNA and 5.8% genes up-regulated. Surprisingly, increased transcriptional activity of cellular genes was not matched by a respective increase in translational activity. Only 0.34% of translated genes showed increased translational activity at 8h p.i. and 77% of transcriptionally induced protein-coding genes were not translated at all at 8h p.i.

When analyzing genes that were transcriptionally induced but not translated, we observed massive transcriptional activity upstream of their 5'-ends at late times of infection originating from neighboring upstream genes. Thus, the transcription termination and cleavage machinery was no longer properly functioning at the termination signals of upstream genes, resulting in transcription into downstream regions by >100,000nt (denoted as ‘read-out’). By 8h p.i., read-out was >15% of the gene’s transcription for 64% of cellular genes and >75% for 26% of genes. Furthermore, the extent of read-out correlated with the prevalence of different polyadenylation [poly(A)] signal sequences within 50nt of gene 3'-ends, indicating that non-canonical and likely weaker poly(A) signals were more strongly affected by disrupted transcription termination.

Late in infection, read-out commonly extended over thousands of nucleotides into downstream genes (denoted as ‘read-in’). At least 32.6% of genes showed read-in >15% of the gene’s transcription at 7-8h p.i. and read-in was inversely correlated with the distance to the next upstream gene. For genes with low or no transcription in uninfected cells, read-in often exceeded endogenous transcript level, resulting in seeming ‘induction’. Indeed, 36% of genes with >75% read-in appeared to be up-regulated in 4sU-RNA, in contrast to ~2.6% of genes with ≤5% read-in. Furthermore, HSV-1 infection induced aberrant splicing events, which were enriched among genes with high read-out. Thus, splicing was already affected upstream of poly(A) sites suffering from read-out. Interestingly, 44% of the induced splice junctions were novel and 11% of these represented intergenic splicing between two neighboring genes



connected by read-out and subsequent read-in. These intergenic splicing events conclusively demonstrate that disruption of transcription termination resulted in large RNA molecules spanning two or more cellular genes.

## **CONCLUSIONS**

We report the surprising observation that HSV-1 specifically disrupts transcription termination of cellular genes, which has never been reported before for any pathogen or in a wild-type setting, i.e. without knock-down of individual genes. This study thus substantially advances our understanding of a common human pathogen and establishes HSV-1 as a model system for studying transcription termination.

## **REFERENCES**

1. Rutkowski, A.J., Erhard, F., L'Hernault, A., Bonfert, T., Schilhabel, M., Crump, C., Rosenstiel, P., Efstathiou, S., Zimmer, R., Friedel, C.C., Dölken, L. Wide-spread disruption of host transcription termination in HSV-1 infection. *Nature Comm.*, accepted (2015).

# Relative abundances of transcript isoforms are predictive of tumor staging and survival in 12 cancer types

Juan L. Trincado<sup>1</sup>, Amadis Pagés<sup>1,2</sup>, Endre Sebestyén<sup>1</sup>, Eduardo Eyras<sup>1,3</sup>

<sup>1</sup>Universitat Pompeu Fabra, Dr. Aiguader 88, E08003 Barcelona, Spain

<sup>2</sup>Centre for Genomic Regulation, Dr. Aiguader 88, E08003 Barcelona, Spain

<sup>3</sup>Catalan Institution for Research and Advanced Studies, Passeig Lluís Companys 23, E08010 Barcelona, Spain

Correspondence to: [eduardo.eyras@upf.edu](mailto:eduardo.eyras@upf.edu)

---

## BACKGROUND

Establishing the stage of a tumor is crucial to select the appropriate therapeutic strategy and to determine patient prognosis. Molecular signatures that accurately predict the clinical outcome of individuals with cancer are essential for appropriate therapy selection. Alterations in RNA processing are emerging as important novel signatures to understand tumor formation and to develop new therapeutic strategies. However, it is not yet known whether specific patterns of transcript isoform expression in tumors can be associated to clinical stage.

## RESULTS

Using a machine learning approach, we integrate data from RNA sequencing from 12 cancer types from The Cancer Genome Atlas (TCGA) project and build predictive models of tumor staging and clinical outcome. We show that this models can separate with high accuracy early from late stage cancer patients and show significant difference in survival. Applied to patients of unknown stage, we show a significant difference in survival between late-stage and early-stage predicted patients. We further provide evidence that cancer type specific models have better accuracies than generic models across multiple types. In addition, we compare with gene expression classifiers and show that the accuracy obtained through transcript isoforms is comparable to models based on gene expression. We also find significant isoforms changes that separate different prognosis according to the expression of the estrogen receptor gene in breast cancer samples

## CONCLUSIONS

We conclude that transcripts isoform relative abundances could be used as another useful tool in clinical decision-making.

# Genomic alterations dysregulate cancer genes by modulating microRNA activity

Hua-Sheng Chiu<sup>1</sup>, Andrea Califano<sup>2</sup>, and Pavel Sumazin

<sup>1</sup> Texas Children's Cancer Center, Baylor College of Medicine, and <sup>2</sup> Columbia Department of Systems Biology, Columbia University, USA;

Correspondence: sumazin@bcm.edu

---

## BACKGROUND

Analyses of established cancer genes in eight cancer types, profiled by The Cancer Genome Atlas (TCGA), suggests that the vast majority of cancer genes are aberrantly expressed in tumors where their DNA loci have no mutations and no copy-number or methylation abnormalities. For example, GATA3 and PIK3CA are aberrantly expressed while having normal DNA loci in 40-45% of profiled breast cancer; IDH1 and RUNX1 in 95% of glioblastoma, and STAT3 in 90% of prostate cancer tumors [1-3]. The identification of trans-acting genomic variants that are predictive of this dysregulation is a step towards clinically interpret genomic data and identifying causal variants with pathophysiological relevance. It will assist interpreting recurrent focal copy-number alterations with unknown function, and lead to new approaches for targeting cancer genes. Recently, modulators of miRNA activity, including competing endogenous RNA (ceRNA) [4] species that can regulate the abundance of other RNAs in trans by competing for common regulating miRNAs, have been shown to alter the expression of key cancer genes. Up- or down-regulation of ceRNAs alters the expression of their cognate targets, and alterations in copy number and methylation at ceRNA loci are integrated and propagated in trans by ceRNA, resulting in pathophysiological dysregulation of tumor suppressor and oncogene expression [4].

## RESULTS

Our analyses suggest that the expression of hundreds of tumor suppressors and oncogenes are altered by genomic variants at the loci of their ceRNA regulators in each of eight cancer types. We provide evidence to suggest that ceRNA interactions are near independent of individual miRNA abundance, resulting in a near context-independent pan-cancer ceRNA interaction network (henceforth PCI). We validated the PCI using high- and low-throughput biochemical assays, and showed that key cancer genes are mechanistically dysregulated by concerted genomic alterations at their cognate ceRNA-interacting genes in samples where their genomic loci are intact. Conclusions from our analysis were confirmed using molecular profiles of 14,240 tumors from 129 additional cohorts. Focusing on specific tumor suppressors and oncogenes, we've shown that tumor suppressors, including PTEN, RB1, and P53 are able to regulate each other on the RNA level; that the same is true for oncogenes, including HIF1A, CCND1 and HMGA2; that genomic alterations at APC and ESR1 ceRNA regulators were predictive of their dysregulation in colon and breast cancer tumors, respectively, and that perturbations that target these ceRNA regulators alter target expression and mimic corresponding cell and tumor phenotypes.

## CONCLUSIONS

Altogether, our results suggest that the PCI represents a key resource for cancer genomic studies and that its further study may elucidate critical pathophysiological mechanisms.

## REFERENCES

1. TCGA-Consortium. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*. Oct 23 2008;455(7216):1061-1068.
  2. Curtis C, Shah SP, Chin SF, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*. Jun 21 2012;486(7403):346-352.
-

3. Taylor BS, Schultz N, Hieronymus H, et al. Integrative genomic profiling of human prostate cancer. *Cancer Cell*. Jul 13 2010;18(1):11-22.
4. Tay Y, Rinn J, Pandolfi PP. The multilayered complexity of ceRNA crosstalk and competition. *Nature*. Jan 16 2014;505(7483):344-352.

# **A highway to the sites of RNA silencing: Cell entry routes of exosomes as a novel paradigm for therapeutic RNA delivery**

Nicole Meisner

Novartis Institutes for Biomedical Research, Novartis Campus, 4002 Basel, Switzerland. [nicole-claudia.meisner@novartis.com](mailto:nicole-claudia.meisner@novartis.com)

---

## **BACKGROUND**

Delivery remains a major challenge for the clinical viability of therapeutic RNA. In addition to pharmacokinetics, tissue distribution and cell uptake, subcellular trafficking emerges as an additional hurdle to overcome. In particular we have recently shown that RNA silencing is nucleated at the rough endoplasmic reticulum, and that current state of the art liposomal delivery vehicles are highly inefficient in delivering siRNA to the ER, thereby limiting loading into the RNA silencing machinery. An improvement of subcellular targeting should therefore directly improve the therapeutic index and thus the clinical viability of siRNA therapies. Here we will discuss our recent data that reveal how exosomes have evolved a specific route for cell entry with efficiency reminiscent of highly infective pathogens, which licenses them for directed subcellular transport to the sites of translation and RNA silencing - thereby providing a role model for next generation RNA delivery vehicles.

# Probabilistic modelling of RNA structure probing data

Alina Selega<sup>1\*</sup>, Christel Sirocchi<sup>2</sup>, Sander Granneman<sup>2</sup>, and Guido Sanguinetti<sup>1</sup>

<sup>1</sup>School of Informatics, University of Edinburgh,

<sup>2</sup>Centre for Synthetic and Systems Biology, University of Edinburgh

\*To whom correspondence should be addressed: [alina.selega@ed.ac.uk](mailto:alina.selega@ed.ac.uk)

---

## BACKGROUND

RNA molecules control a wide variety of cellular processes. The folding of RNA into highly complex shapes fundamentally determines its biological function, making the inference of its secondary structure an important problem. Recently, RNA structure probing methods have been increasingly combined with high-throughput sequencing [1, 2, 3], generating a wealth of data, currently unmatched by thorough statistical efforts to analyse it. Most of the existing algorithms [4, 5] consider a single experimental replicate, are agnostic to biases of the technology, and require setting semi-arbitrary thresholds for calling modified regions. We address these issues by presenting a statistical machine learning pipeline for modelling ChemModSeq [1] data, which specifically focuses on capturing biological variability and intrinsic biases.

## RESULTS

ChemModSeq [1] data describe the drop off of the reverse transcriptase (RT) during the complementary DNA synthesis, reflecting chemical modifications of the flexible parts of the RNA. To measure the degree of chemical modification at a nucleotide, we define its drop off rate as the ratio between the total number of drop offs at that nucleotide over its coverage. However, as RT can drop off randomly, our pipeline takes into account the variability between experimental replicates in order to determine whether a certain drop off rate is significantly above the expected background levels in control conditions. The algorithm computes probabilities of chemical modification at each nucleotide based on empirical comparisons of drop off rates in treated samples to the (adjusted) null distribution of drop off rates in control replicates. We propose empirical strategies to transform the data in order to correct for the coverage and sequence-dependent biases that we identified. Our probabilistic model introduces spatial correlations between neighbouring nucleotides by using a Hidden Markov Model (HMM), in which the emission probabilities come from a mixture Uniform-Beta model, reflecting the unmodified and modified states of the hidden nodes. We optimise the shape parameters of the Beta distribution component using the expectation-maximisation algorithm.

Our pipeline demonstrates competitive runtime when applied to the full transcriptome data from *S. cerevisiae* and produces high-quality reconstructions of RNA secondary structures that stay reliable at considerably lower coverage levels than previously reported. We evaluated the algorithm's performance on the ribosomal RNA 18S with a known structure, demonstrating prediction accuracy comparable or exceeding the results obtained from using selective 2'-hydroxyl acylation (SHAPE) reactivities [6].

## CONCLUSIONS

We present what to our knowledge is the first attempt at a statistical model of ChemModSeq data, accounting for biases and intrinsic variability in the data. Our algorithm is efficient, which makes it suitable for the transcriptome-level studies, and requires lower levels of coverage for reliable predictions. When applied to a solved structure, our results are comparable to or exceed the accuracy of threshold-based approaches.

## REFERENCES

1. Hector, R.D., et al. Snapshots of pre-rRNA structural flexibility reveal eukaryotic 40S assembly dynamics at nucleotide resolution. *Nucleic acids research* **42** (19) (2014).
2. Lucks, J.B., et al. Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proceedings of the National Academy of Sciences* **108** (27) (2011).
3. Underwood, J.G., et al. FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nature methods* **7** (12) (2010).
4. Aviran, S., et al. Modeling and automation of sequencing-based characterization of RNA structure. *Proceedings of the National Academy of Sciences* **108** (27) (2011).
5. Talkish, J., et al. Mod-seq: high-throughput sequencing for chemical probing of RNA structure. *RNA* (2014).
6. Wilkinson, K.A., et al. Selective 2 [prime]-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nature protocols* **1** (3) (2006).

# Bayesian methods for transcript expression estimation and differential expression calling from RNA-Seq data

Magnus Rattray

Faculty of Life Sciences, University of Manchester, Manchester, United Kingdom

---

## BACKGROUND

I will describe recent work on Bayesian methods for transcript expression estimation and differential expression calling. In the case of expression estimation we find that posterior mean (PM) estimates that can be obtained from methods such as BitSeq and RSEM perform very well in benchmarks on real and simulated data. However, these PM estimates are typically obtained using MCMC which can be prohibitively slow. We have therefore developed a natural gradient variational Bayes (VB) method which is shown to speed up inference ten fold compared to MCMC on typical datasets (Hensman et al. 2014). The natural gradient VB method is shown to be much faster than the more standard VBEM algorithm. In the case of differential expression estimation we show how to extend the model to multiple conditions and we develop an MCMC algorithm for joint estimation of expression level and differential expression calling (Papastamoulis & Rattray 2015).

## REFERENCES

- J. Hensman, P. Papastamoulis, P. Glaus, A. Honkela, M. Rattray "Fast and accurate approximate inference of transcript expression from RNA-seq data" arXiv:1412.5995, 1014.
- P. Papastamoulis, M. Rattray "A Bayesian model selection approach for identifying differentially expressed transcripts from RNA-Seq data" arXiv:1412.3050, 1014.



# ASpli: an integrative R package for the analysis of alternative splicing using RNA-Seq

Estefania Mancini<sup>1\*</sup>, Ariel Chernomoretz<sup>1</sup> and Marcelo J Yanovsky<sup>1</sup>

<sup>1</sup> Fundación Instituto Leloir \*emancini@leloir.org.ar

---

## BACKGROUND

Alternative splicing (AS) is a prevalent mechanism of post transcriptional gene regulation in multicellular eukaryotes. It allows a single gene to increase functional and regulatory diversity, through the synthesis of multiple mRNA isoforms encoding structurally and functionally distinct proteins. AS occurs via 4 main events: intron retention (IR), exon skipping (ES) and alternative use of donor and acceptor sites (alt 5' and alt 3'). The development of novel high-throughput sequencing methods for RNA (RNA-Seq) provided a very powerful mean to study alternative splicing under multiple conditions at unprecedented depth. As long as new studies on post-transcriptional regulation arises, there are an increasing evidence than AS frequency is higher than expected. Despite It has become the new standard for studying gene and transcription expression, the use of RNA-seq for the study of transcripts repertoire in a given condition is not trivial.

## RESULTS

Here we introduce a very flexible and easy-to-use R package named ASpli. We propose a count based integrative method taking into account gene expression, exon and intron differential usage and their relationship with junctions spanning those features.

## CONCLUSIONS

Using an annotated transcriptome we are able to classify subgenic features into alternative or not alternative regions. ASpli is intended to facilitate the analysis of RNAseq data for the quantification and discovery of AS events and it has been used in many recent publications from our lab. Results of the analysis are presented in a user friendly manner, including plots of the most relevant AS events discovered.

# The combinatorial effects of RNA Polymerase II elongation rate, nucleosome occupancy and chromatin organization on alternative splicing

Eisenberg A.<sup>1\*</sup>, Tamer L.<sup>1</sup>, Iannone C.<sup>2</sup>, Valcarcel J.<sup>2</sup> and Ast G.<sup>1</sup>

<sup>1</sup> Department of Human Molecular Genetics & Biochemistry, Sackler Faculty of Medicine, Tel-Aviv University.

<sup>2</sup> Centre for Genomic Regulation, Barcelona.

\* annalitik.savchenko@gmail.com

---

## BACKGROUND

Introns were identified almost 40 years ago, and yet we still do not fully understand how the splicing machinery locates short exons embedded between long flanking intron sequences and splices them out to generate a mature mRNA molecule. There are two proposed models for how the splicing machinery recognizes introns and exons – intron definition and exon definition. When introns are short (<300 nt), the splicing machinery recognizes the introns as the spliced unit<sup>1</sup>. When introns are longer than ~800 nt, the basal machinery is placed across the exon<sup>2,3</sup>. However, exon-intron architecture by itself fails to fully explain the recognition of exons. We have recently identified two exon-intron GC content architectures that reflect the two mechanisms by which splicing signals are recognized. In our research we examine whether the splicing pattern depends on the genomic location and the differences between GC rich regions and GC poor regions. The second aspect of our research involves spatial proximity between the 5' splice site and the 3' splice site located across an exon and the role of 'nucleosome marking' in the exon selection process.

## RESULTS

In order to examine structural, transcriptional, and epigenetic differences between varying GC content environments we integrated GC-rich and GC-poor gene segments within the same genomic location. The low GC content genes showed full exon inclusion. Surprisingly, two of the three high GC content minigenes, which under transient transfection conditions were unspliced, exhibited exon inclusion in this genomic context. In order to distinguish the molecular mechanism that differentiates between intron retention and exon skipping we generated several minigenes in which we affect the recognition of their second intron's 5' splice-sites. Our results show that a stably transfected mutated low GC content minigene undergoes exon skipping, whereas mutated high GC-content minigenes exhibit mainly intron retention. Our results imply that exons are under selection to maintain a length similar to that of DNA wrapped around a nucleosome (147 nucleotides). To determine whether the link between exon inclusion and the length of the exon depends on chromatin organization, we used an in vitro splicing reaction, which is DNA and histone free, and occurs independently of transcription. Our results indicate that the absence of splice site proximity around a nucleosome leads to a splicing shift from exon inclusion to skipping.

## CONCLUSIONS

In our research we demonstrated that the genomic environment plays a prominent role in efficient splicing for genes from a high GC-content environment. The results of 5`splice site recognition experiments strongly support that intron retention is associated with high GC-content genes and exon definition is related to genes from the low GC-content regions. We conclude that the correlations we have found between the spatial proximity of splice sites and increased inclusion of the exon is of major importance to the regulation of splicing. Finally, we demonstrated by the in-vitro splicing experiment that the link between exon inclusion and the length of the exon depends on chromatin organization.

## REFERENCES

- 1 Fox-Walsh, K. L. *et al. Proc Natl Acad Sci U S A* **102**, 16176-16181 (2005).
- 2 De Conti, L., Baralle, M. & Buratti, E. *Wiley Interdiscip Rev RNA* **4**, 49-60 (2012). 3 Berget, S. M. *J Biol Chem* **270**, 2411-2414 (1995).

# RNA aptamers and methods for their optimal utilization

Jan Hoinka<sup>1</sup>, Puong Dao<sup>1</sup>, Alexey Berezhnoy<sup>2</sup>, Zuben E Sauna<sup>3</sup>, Eli Gilboa<sup>2</sup>, Teresa M. Przytycka<sup>1\*</sup>

<sup>1</sup> National Center of Biotechnology Information, NIH, Bethesda MD, US <sup>2</sup> Department of Microbiology & Immunology, University of Miami Miller School of Medicine, Miami, FL, US; Laboratory of Hemostasis, Division of Hematology, Center for Biologics Evaluation and Research, FDA, Silver Spring, MD, US

\*To whom correspondence should be addressed: przytyck@ncbi.nlm.nih.gov

---

## BACKGROUND

Systematic Evolution of Ligands by EXponential Enrichment (SELEX) is a well established experimental procedure to identify aptamers - synthetic single-stranded (ribo)nucleic molecules that bind to a given molecular target. Aptamers have a broad spectrum of applications and are increasingly being used to develop new therapeutics and diagnostics. High Throughput (HT) SELEX combines with massively parallel sequencing technologies. It produces unprecedented amount of data that require suitable computational methods to analyze it. HT-SELEX opened the field to new computational opportunities and challenges that are yet to be addressed. Over the years, we have developed several computational methods [1-3] to aid the analysis of the results of HT-SELEX and to advance the understanding of the selection process itself.

## RESULTS

To address emerging need for developing efficient algorithms for HT-SELEX data analysis we new clustering algorithm, AptaCluster, which is, to the best of our knowledge, the only currently available tool capable of efficiently clustering entire aptamer pools of more than 20 Million unique sequences [1]. With this tool at hand we were interested in understanding the role of mutations, by which we understand nucleotide sequence errors arising at any stage of the selection experiment, including amplification, and for RNA aptamers, transcription. While the principles of mutagenesis during traditional SELEX and as a means of post-selection optimization of binding affinity have previously been described, the lack of high-throughput sequencing of entire aptamer pools posed a natural limit to the resolution of the available data and consequent analysis.

We specifically asked: (i) is the distribution of mutants consistent with the random mutation model, and (ii) is it possible to computationally identify mutations that improve binding affinity [3]. Our study was informed by high-throughput sequencing data from five rounds of HT-SELEX developing aptamers against the Interleukin 10 receptor alpha chain (IL-10RA). IL-10 is considered to be a master regulator of immunity to infection and is an important therapeutic molecular target. To address the first question, we utilized our new clustering algorithm, AptaCluster to obtain families of aptamer sequences related to each other by mutations. Interestingly, we found that similar to a number of phenomena in life and social sciences, the distribution of aptamers in these families follows a scale-free distribution. We obtained the same distribution using an in-silico aptamer evolution program, AptaSim, which we have developed. Next we developed AptaMut, a computational approach to identify binding enhancing mutations.. We discuss the practical implications of these findings for predicting binding affinity.

## **CONCLUSIONS**

Our results demonstrate that new computational methods cannot only aid the elucidation of under-appreciated properties of the SELEX procedure but can ultimately lead to uncovering new practical predictive methods and aptamers of desired binding affinity.

## **REFERENCES**

1. Hoinka J, Zotenko E, Friedman A, Sauna ZE, Przytycka TM. Identification of sequence-structure RNA binding motifs for SELEX-derived aptamers. *Bioinformatics*. 2012 Jun 15;28(12):i215-23.
2. Hoinka J., Berezhnoy A., Sauna Z.E., Gilboa E., Przytycka T.M. AptaCluster – A Method to Cluster HT-SELEX Aptamer Pools and Lessons from Its Application; 18th Annual International Conference on Research in Computational Molecular Biology. Pittsburgh, Pennsylvania: Springer; 2014 p. 115-128.
3. Hoinka J, Berezhnoy A, Dao P, Sauna ZE, Gilboa E, Przytycka TM. Large scale analysis of the mutational landscape in HT-SELEX improves aptamer discovery. *Nucleic Acids Res*. 2015

# RNA binding by *C. elegans* splicing factors in vitro and in vivo

Cameron Mackereth<sup>1</sup>, Samir Amrane<sup>1</sup>, Heddy Soufari<sup>1</sup>, Denis Dupuy<sup>1</sup>

<sup>1</sup> Univ. Bordeaux, Institut Europeen de Chimie et Biologie, Inserm U869, 33607 Pessac, France

\*To whom correspondence should be addressed: [c.mackereth@iecb.u-bordeaux.fr](mailto:c.mackereth@iecb.u-bordeaux.fr)

---

## BACKGROUND

Due to the process of alternative splicing, a single gene can give rise to multiple proteins, often with divergent function or changes in translation efficiency. The selective inclusion or exclusion of exons from the pre-mRNA generates this variety, and this process is regulated in response to external triggers, as well as in a time- or location-dependent manner – for example at specific points in development or in certain tissues. Regulation often depends on the restricted expression of RNA-binding splice factor proteins that recognize conserved RNA elements within the subset of target pre-mRNA molecules. We have therefore initiated a multi-disciplinary investigation on nematode splicing proteins that combines atomic level structural information, binding affinity measurements, and observation of isoform control in live worms.

## RESULTS

Our initial model system is based on the muscle-specific regulation of *egl-15* splicing by the SUP-12 protein from *C. elegans* [1][2] and we now study additional splicing factors that also give rise to tissue-specific regulation. NMR spectroscopy and X-ray diffraction methods have provided atomic details of the RNA-bound complexes. Other biophysical methods, notably isothermal titration calorimetry (ITC), help to define the critical side chains and RNA sequence motifs that are required for high affinity binding. From these in vitro data we have a precise list of the contacts between the splicing protein residues and atoms in the RNA bases that are required for association. The data also provide an important quantitative correlation between subtle changes in the RNA motif sequence and the effect on binding affinity. We have then used the molecular information to create a series of transgenic *C. elegans* strains in which wildtype or mutant RNA motifs are used within a fluorescent mini-gene reporter to detect the resulting splice isoform control in live worms.

## CONCLUSIONS

Molecular characterization of splicing factor protein-RNA complexes provides atomic details that can guide the design of mutants to precisely regulate affinity in vitro. These same quantitative effects can also be studied in vivo with fluorescent mini-gene reporters, which we have used to demonstrate that RNA sequence variation has similar effects on both biophysical measurements and in the worm. The data therefore provides not just the binding details towards an optimal RNA motif, but also the exact affinity consequence of base changes within the motif. Consequently, a more comprehensive representation of the target RNA sequences can be obtained by using the continuum of RNA motifs ranked in order of affinity. Our current study now addresses the effect of concurrent multiple domain (or protein) binding, and the RNA sequence elements that are recognized when homodimeric and heterodimeric splicing factor molecules bind to the same pre-mRNA.

## REFERENCES

1. Amrane, S., Rebora, K., Zniber, I., Dupuy, D. and Mackereth, C.D. Backbone-independent nucleic acid binding by splicing factor SUP-12 reveals key aspects of molecular recognition. *Nature Communications* **5**:4595 (2014).
2. Mackereth, C.D. Splicing factor SUP-12 and the molecular complexity of apparent cooperativity. *Worm* **3**:e991240 (2014).

# The Human Inosinome Atlas

Ernesto Picardi<sup>1,2</sup>, Caterina Manzari<sup>2</sup>, Francesca Mastropasqua<sup>1</sup>, Italia Aiello<sup>1</sup>, Anna Maria D'Erchia<sup>1,2</sup> and Graziano Pesole<sup>1,2</sup>

<sup>1</sup> Dipartimento di Bioscienze, Biotecnologie e Biofarmaceutica, Università di Bari, Bari, Italy, <sup>2</sup> Istituto Biomembrane e Bioenergetica del Consiglio Nazionale delle Ricerche, Bari, Italy

\*To whom correspondence should be addressed: graziano.pesole@uniba.it

---

## BACKGROUND

RNA editing is a post-transcriptional molecular phenomenon whereby a genetic message is modified from the corresponding DNA template by means of substitutions, insertions and/or deletions [1]. In human, it mainly involves the deamination of adenosines to inosines by the family of ADAR enzymes acting on double RNA strands [2]. A-to-I RNA editing has a plethora of biological effects depending on the RNA region involved in the modification [3]. Changes in UTRs can lead to altered expression, whereas modifications in coding protein regions can induce amino acid replacements with more or less severe functional consequences [4]. The detection of RNA editing events at genomic scale has been largely facilitated by the advent of NGS technologies. Although the computational identification of A-to-I changes in human is yet a challenging task, several attempts have been done and more than 2,4 million events have been collected [5]. RNA editing is a dynamically regulated process and a comprehensive understanding of its biological roles requires large-scale studies and the creation of a specialized atlas.

## RESULTS

To this aim, we have massively sequenced total RNA from six human tissues (brain, lung, liver, kidney, heart and muscle) in three different individuals using a HiSeq2500 Illumina sequencer and according to a stranded protocol to preserve RNA orientation. In addition, we have generated whole genome sequencing data for each donor and whole exome sequencing data for each tissue. NGS data have been analysed according to an improved RNA editing algorithm implemented in our REDIttools suite taking into account several error sources. Hyper edited reads have also discovered using a recent approach developed to rescue heavily edited RNA-Seq reads that are generally missed by current methods [6]. Overall we detected 3,041,422 events representing the largest collection of A-to-I RNA editing in human with more than 2 millions of novel positions. Of these, 97% was in repetitive regions and ~90% in Alu elements. Only a limited amount of sites fell in non-repetitive regions (3%), as expected. The number of predicted A-to-I events varied greatly among samples because of sequencing depth variation, stringent filters used to recover editing candidates and tissue specific roles of RNA editing. Nonetheless, brain appeared the most edited tissue with on average 511,733 sites per sample. In contrast, heart and muscle showed a smaller number of editing sites than other tissues with on average 79,976 and 28,620 changes, respectively. Regarding the impact on known human protein-coding genes, we discovered that 13062 loci over 20173 (65%) underwent RNA editing in their exons and/or introns. Very interestingly, we found that edited genes were consistently enriched in genes involved in neurological disorders and cancer. In addition, 74% (1842/2501) of essential genes [7] were in the edited set, confirming the relevant biological role of RNA editing in human.



## CONCLUSIONS

Here we present the largest collection of RNA editing events in human tissues. We confirm that RNA editing is pervasive in human and indispensable to pervert the cellular homeostasis. Indeed, edited genes are enriched in genes linked to cancer and neurological diseases. Our collection will facilitate the understanding of RNA editing role in normal as well as pathological conditions.

## REFERENCES

1. Gott JM, Emeson RB: **Functions and mechanisms of RNA editing.** *Annu Rev Genet* 2000, **34**:499-531.
2. Hogg M, Paro S, Keegan LP, O'Connell MA: **RNA editing by mammalian ADARs.** *Adv Genet* 2011, **73**:87-120. 3. Maas S: **Gene regulation through RNA editing.** *Discov Med* 2011, **10**(54):379-386.
4. Hood JL, Emeson RB: **Editing of Neurotransmitter Receptor and Ion Channel RNAs in the Nervous System.** *Current topics in microbiology and immunology* 2011.
5. Ramaswami G, Li JB: **RADAR: a rigorously annotated database of A-to-I RNA editing.** *Nucleic acids research* 2014, **42**(Database issue):D109-113.
6. Porath HT, Carmi S, Levanon EY: **A genome-wide map of hyper-edited RNA reveals numerous new sites.** *Nature communications* 2014, **5**:4726.
7. Dickerson JE, Zhu A, Robertson DL, Hentges KE: **Defining the role of essential genes in human disease.** *PloS one* 2011, **6**(11):e27368.

# Transcriptome analysis reveals thousands of targets of nonsense-mediated mRNA decay that offer clues to the mechanism in different species

Courtney E. French<sup>1</sup>, Gang Wei<sup>2</sup>, Angela N. Brooks<sup>3</sup>, Thomas L. Gallagher<sup>4</sup>, Li Yang<sup>5</sup>, Brenton R. Graveley<sup>6</sup>, Sharon L. Amacher<sup>4</sup>, and Steven E. Brenner<sup>1\*</sup>

<sup>1</sup> University of California, Berkeley, CA

<sup>2</sup> Fudan University, Shanghai, China

<sup>3</sup> Broad Institute, Cambridge, MA

<sup>4</sup> Ohio State University, Columbus, OH

<sup>5</sup> CAS-MPG Partner Institute for Computational Biology, Shanghai, China

<sup>6</sup> University of Connecticut Health Center, Farmington, CT

\*To whom correspondence should be addressed: [brenner@compbio.berkeley.edu](mailto:brenner@compbio.berkeley.edu)

---

## BACKGROUND

Many alternatively spliced isoforms contain a premature termination codon that targets them for degradation by the nonsense-mediated mRNA decay RNA surveillance system (NMD). Some such unproductive splicing events have a regulatory function, whereby alternative splicing and NMD act together to impact protein expression. Numerous RNA-binding proteins, including all the human SR splicing factors, are regulated by alternative splicing coupled to NMD, in conjunction with highly- or ultra-conserved elements [1,2,3]. The “50nt rule” is the prevailing model for how premature termination codons are defined in mammals, and requires a splice junction downstream of the stop codon [4]. There is evidence that this rule holds in *Arabidopsis* [5] but not in other eukaryotes including *Drosophila* [6]. There is also evidence that a longer 3' UTR triggers NMD in yeast, plants, flies, and mammals [5,7,8,9,10].

## RESULTS

To survey the targets of NMD genome-wide in human, zebrafish, and fly, we have performed RNA-Seq analysis on cells where NMD has been inhibited via knockdown of UPF1, a critical protein in the degradation pathway. We found that hundreds to thousands of genes produce alternative isoforms that are degraded by NMD in each of the three species, including over 20% of the genes alternatively spliced in human HeLa cells. These genes, potentially subject to regulation through NMD, are involved in many functional categories and, in human and fly, are significantly enriched for RNA splice factors. We also found a significant enrichment for ultraconserved elements in the human NMD targets and usually these elements overlapped a poison cassette exon.

We were able to gain insight into what defines NMD targets from our RNA-Seq data. We found that the 50nt rule is a strong predictor of NMD degradation in human cells, and also seems to play a role in zebrafish and in fly. In contrast, we found little correlation between the likelihood of degradation by NMD and 3' UTR length in any of the three species. In fly, we see no enrichment for longer 3' UTRs in isoforms degraded by NMD, unless they have an intron. Other features have also been associated with propensity for NMD. We also found that thousands of human transcripts have uORFs that seem to affect their likelihood of degradation.

## CONCLUSIONS

Ultimately, our findings demonstrate that gene expression regulation through NMD is widespread in human, zebrafish, and fly, and that NMD is strongly predicted by the 50nt rule but not by 3' UTR length.

## REFERENCES

1. Lareau, L.F., et al. *Nature* **446** (2007).
2. Ni, J., et al. *Genes and Development*. **21** (2007)
3. Lareau, L.F. and Brenner, S.E. *Molecular Biology and Evolution*. **32** (2015)
4. Nagy, E. and Maquat, L. *Trends in Biochemical Science*. **23** (1998)
5. Kerényi, Z., et al. *EMBO Journal*. **27** (2008)
6. Gatfield, D., et al. *EMBO Journal*. **22** (2003)
7. Hansen, K., et al. *PLoS Genetics*. **5** (2009)
8. Hogg, J and Goff, S. *Cell*. **143** (2010)
9. Yepiskoposyan, H., et al. *RNA*. **17** (2011)
10. Hurt, J.A., et al. *Genome Research*. **23** (2013)

# Network of Splice Factor Regulation by Alternative Splicing Coupled with Nonsense Mediated mRNA Decay

Anna Desai<sup>1</sup>, James P. B. Lloyd<sup>1</sup>, Courtney E. French<sup>1</sup>, and Steven. E. Brenner<sup>1\*</sup>

<sup>1</sup> University of California, Berkeley, CA

\*To whom correspondence should be addressed: brenner@compbio.berkeley.edu

---

## BACKGROUND

Nonsense-mediated mRNA decay (NMD) is an RNA surveillance pathway that degrades aberrant transcripts harboring premature termination codons. However, this pathway also has physiological targets: many genes produce alternative isoforms containing premature termination codons. In this mode of regulation, a splicing factor can induce splicing of an alternative isoform with an early stop codon. These isoforms will be degraded by NMD, resulting in lower protein expression. Regulation of alternative splicing involves complex interactions between many splice factors, and so splice factor levels must be carefully regulated. Splicing coupled to NMD allows for an additional level of post-transcriptional regulation for these genes. For example, splicing factors such as SRSF1, SRSR2, SRSF3, and SRSF7 are known to regulate their own expression and expression of other splice factors by coupling alternative splicing and NMD. hnRNP L, hnRNP LL, PTBP1, and PTBP2 are regulated in the same manner.

## RESULTS

After an extensive literature search, we generated a splicing factor regulatory network that encompasses current knowledge of splice factor regulatory interactions. The currently available data shows that the majority of the SR proteins and a few hnRNP splicing factors are known to be regulated by another splicing factor via alternative splicing coupled with NMD. Since all the SR proteins and many hnRNP splicing factors produce isoforms degraded by NMD, we predict that this mode of regulation is pervasive in this dense splicing factor regulatory network. In addition, CLIP-seq data reveals extensive splicing factor-mRNA interactions, providing an additional hint that many more splicing factors might be regulated by other splicing factors via alternative splicing coupled with NMD. Further work will establish the true extent of regulation by alternative splicing coupled to NMD of splicing factors by building a comprehensive regulatory network model.

## CONCLUSIONS

A network built from the currently available data on splicing factor regulatory interactions indicates extensive auto-and cross-regulation through alternative splicing coupled with NMD. In this dense and robust regulatory network, there does not seem to be a hierarchy in which certain splicing factor are “master regulators” of splicing.

# Skip and bin: Widespread exon-skipping triggers degradation by nuclear RNA surveillance

Danny A. Bitton<sup>1</sup>, Sophie R. Atkinson<sup>1</sup>, Charalampos Rallis<sup>1</sup>, Graeme C. Smith<sup>1</sup>, David A. Ellis<sup>1</sup>, Yuan Y.C. Chen<sup>1,4</sup>, Michal Malecki<sup>1</sup>, Sandra Codlin<sup>1</sup>, Jean-François Lemay<sup>2</sup>, Cristina Cotobal<sup>3</sup>, François Bachand<sup>2</sup>, Samuel Marguerat<sup>1,5</sup>, Juan Mata<sup>3</sup>, and Jürg Bähler<sup>1\*</sup>

<sup>1</sup>University College London, Research Department of Genetics, Evolution & Environment and UCL Genetics Institute, London, WC1E 6BT, UK.

<sup>2</sup>Université de Sherbrooke, Department of Biochemistry, Sherbrooke, Quebec J1H 5N4, Canada <sup>3</sup>Department of Biochemistry, University of Cambridge, Cambridge, UK.

<sup>4</sup>Current address: Duke-NUS Graduate Medical School, Singapore 169857.

<sup>5</sup>Current address: Imperial College London, MRC Clinical Sciences Centre, London W12 0NN, UK.

\* j.bahler@ucl.ac.uk

---

## BACKGROUND

Exon-skipping is considered a principal mechanism by which eukaryotic cells expand their transcriptome and proteome repertoires. To study the extent and the functional relevance of exon skipping in fission yeast, we analyzed RNA-seq data from 116 transcriptomes, covering multiple physiological conditions as well as transcriptional and RNA processing mutants.

## RESULTS

We applied brute-force algorithms to detect all possible exon-skipping events, which were widespread but rare compared to normal splicing events. Exon-skipping increased in cells deficient for the nuclear exosome or the 5'-3' exonuclease Dhp1, and also during late meiosis. The pervasive exon-skipping transcripts did not increase in specific physiological conditions and were mostly present at <1 copy per cell. These exon-skipping transcripts are therefore unlikely to be functional and may reflect splicing errors that are actively removed by nuclear RNA surveillance. The average splicing rate by exon-skipping was ~0.24% in wild-type and ~1.75% in nuclear exonuclease mutants. We detected ~250 circular RNAs, derived from single or multiple exons, which were rare and stochastic. Using an exhaustive search algorithm, we also uncovered thousands of unknown splice sites, indicating pervasive splicing, yet most of these splicing variants were cryptic and increased in nuclear degradation mutants.

## CONCLUSIONS

This study highlights widespread, but low frequency alternative or aberrant splicing events, which are targeted by nuclear RNA surveillance [1].

## REFERENCES

1. Bitton, D.A. et al. Widespread exon-skipping triggers degradation by nuclear RNA surveillance in fission yeast. *Genome Research*, gr. 185371.185114 (2015).

# Mutant U2AF1 alters splicing in hematopoietic cells in vitro and in vivo

Brian S. White<sup>1,2\*</sup>, Cara Lunn Shirai<sup>1</sup>, James N. Ley<sup>1</sup>, Sanghyun Kim<sup>1</sup>, Matthew Ndonwi<sup>1</sup>, Brian Wadugu<sup>1</sup>, Theresa Okeyo-Owuor<sup>1</sup>, Timothy A. Graubert<sup>3</sup>, and Matthew J. Walter<sup>1</sup>

<sup>1</sup>Department of Medicine, Washington University,

<sup>2</sup>The Genome Institute, Washington University<sup>3</sup>

Massachusetts General Hospital/Harvard Medical School

\*To whom correspondence should be addressed: bwhite@dom.wustl.edu

---

## BACKGROUND

Myelodysplastic syndromes (MDS) are the most common myeloid malignancy of the elderly. 11% of MDS patients have mutations in the U2AF1 splicing factor, making it one of the most commonly mutated genes in this disease. The majority of these mutations affect codon S34. Several RNA-seq studies comparing wildtype (WT) and S34F mutant U2AF1 have revealed alternative splicing (AS) between the genotypes. Nevertheless, little overlap has been reported across studies between genes specifically dysregulated by U2AF1 (S34F).

## RESULTS

To explore the hypothesis that mutant U2AF1-induced AS results in abnormal hematopoiesis, we generated single-copy, site-specific, tetracycline-inducible transgenic mice expressing U2AF1 (WT) or U2AF1 (S34F) [1]. Lethally-irradiated wildtype recipient mice transplanted with U2AF1 (S34F) transgenic mouse bone marrow showed phenotypes resembling MDS, including leukopenia and an expansion of bone marrow common myeloid progenitors (CMP). We performed RNA-seq on CMP cells expressing U2AF1 (S34F) or U2AF1 (WT) and identified 742 dysregulated junctions in 633 genes. As observed previously in human cells, exons skipped more frequently by U2AF1 (S34F) relative to U2AF1 (WT) were enriched for uracil upstream of the AG dinucleotide.

To prioritize altered splicing events for further analysis, we intersected junctions that were significant, homologous across human and mouse, and concordantly dysregulated (same direction of log fold change) across 3 datasets: mouse CMP samples, primary human CD34+ cells over-expressing U2AF1 (S34F) or U2AF1 (WT) that we described previously [2], and acute myeloid leukemia (AML) patient samples with and without *U2AF1* mutations [3]. The intersection included 17 dysregulated junctions in 13 genes. Several of these occurred in genes mutated in MDS and AML (*GNAS*, *PICALM*) or known to be involved in stem cell biology (*H2AFY*, *MED24*). To broaden the list beyond this very conservative estimate, we performed a meta-analysis of the 3 independent datasets using Fisher's combined probability test, and identified 555 significantly dysregulated homologous junctions in 415 genes. Pathway analysis revealed that of 28 significantly dysregulated pathways/gene sets, 14 involved RNA processing, RNA splicing, RNA localization/transport, and RNA binding, while 11 involved protein translation processes and ribosomal pathways. Based on their known biological function, we selected several genes for validation in MDS patient bone marrow samples. Mutant U2AF1-induced splice isoform changes identified by RNA-seq analysis were concordant with RT-PCR of MDS patient bone marrow samples for 7 of 8 splicing events examined, including *H2AFY*, *MED24*, *GNAS*, *PICALM*, *KDM6A*, *KMT2D (MLL2)* and *BCOR*; only *EIF4A2* showed no difference.

## **CONCLUSIONS**

In this study, we provide evidence that mutant U2AF1 expression alters hematopoiesis and pre-mRNA splicing in the primary hematopoietic progenitor cells of mice. U2AF1 (S34F) expression in mice results in leukopenia in the peripheral blood and increases the frequency of progenitor cells in the bone marrow and spleen. We identify U2AF1 (S34F)-specific changes in splice isoforms in 633 genes in mouse common myeloid progenitors. Finally, through an integrative meta-analysis of 3 independent RNA-seq datasets, we identify junctions that are consistently altered across species by mutant U2AF1 expression and that are enriched in RNA processing genes and translational processes/ribosomal genes. Collectively, these results suggest that U2AF1 (S34F)-induced alternative splicing may contribute to the altered hematopoiesis characteristic of MDS.

## **REFERENCES**

1. Shirai CL, Ley JN, White BS, et al. Mutant U2AF1 Expression Alters Hematopoiesis and Pre-mRNA Splicing In Vivo. *Cancer Cell* 2015.
2. Okeyo-Owuor T, White BS, Chatrikhi R, et al. U2AF1 mutations alter sequence specificity of pre-mRNA binding and splicing. *Leukemia* 2014.
3. TCGA consortium. Genome and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med* 2013.

# Lysinylation and Asparaginylation Identity Tangling Through a Rapidly-Evolving Ion-Binding Pocket in *Drosophila* tRNA

Julie Baker Phillips<sup>1,2 \*</sup> and David H. Ardell<sup>2</sup>

<sup>1</sup> Vanderbilt University, Nashville, TN, USA,

<sup>2</sup> University of California, Merced, CA, USA

\*To whom correspondence should be addressed: dardell@ucmerced.edu

---

## BACKGROUND

Non-coding RNA gene families with high copy numbers make both genome assembly and orthology inference very challenging. Yet, provided these challenges can be overcome, these families can yield reliable, universal models for system-wide analysis of RNA-RNA and protein-RNA interaction networks. Here we combined a comprehensive molecular evolutionary analysis of orthologous tRNA gene sets from the *Drosophila* 12 genomes from Rogers *et al.* [1], with our custom estimates of function- and interaction-determining features in each species considered separately. These estimates are based on bioinformatically derived tRNA Class-Informative Features (CIFs), which predict functional sites that target interactions within the tRNA-protein interaction network [2].

## RESULTS

Nearly all of the most rapidly-evolving sites in *Drosophila* tRNAs co-occur in a structurally plastic ion-binding pocket at the core of the tRNA body. tRNA CIFs with the highest variance in functional information across species also occur at this structurally plastic, ion-binding "hot pocket." One highly informative CIF in this pocket, perfectly conserved in Lysine tRNAs in *Drosophila* and *Musca*, has evolved independently three times as a perfectly conserved CIF in *Drosophila* Arginine tRNAs. When this occurred, every Asparagine tRNA in the genome gains this CIF. Interestingly, most paralogs of Asn and Lys tRNA genes are genomically co-clustered, but parallel evolution also includes a satellite tRNA Asn paralog on a different chromosome.

## CONCLUSIONS

The tangling of Lys and Asn identities in *Drosophila* precedes and helps explain the previously reported switching of tRNA identities between these two classes. Also, the potential for Asn-Lys identity tangling may be increased compared to other identity pairs, as AsnRS and LysRS are relatively closely related as Subclass IIb aminoacyl-tRNA synthetases. More generally, Transfer RNA CIFs provide useful insights on how functions and structures turnover in macromolecular interaction networks.

## REFERENCES

1. Rogers, H.H., Bergman, C.M., and Griffiths-Jones, S. (2010). The evolution of tRNA genes in *Drosophila*. *Genome Biol Evol* 2, 467–477.
2. Amrine, K.C.H., Swingley, W.D., and Ardell, D.H. (2014). tRNA signatures reveal a polyphyletic origin of SAR11 strains among alphaproteobacteria. *PLoS Comput. Biol.* 10, e1003454.



# The expanding landscape of snoRNAs: a tale of the trials and tribulations of deep-sequencing structured small RNAs

Fabien Dupuis-Sandoval<sup>1</sup>, Gabrielle Deschamps-Francoeur<sup>1</sup>, Sherif Abou Elela<sup>2</sup>, and **Michelle S. Scott**<sup>1\*</sup>

<sup>1</sup> Department of Biochemistry and RNA group, University of Sherbrooke, Canada, <sup>2</sup> Department of Microbiology and RNA group, University of Sherbrooke, Canada

\*To whom correspondence should be addressed: michelle.scott@usherbrooke.ca

---

## BACKGROUND

Small nucleolar RNAs (snoRNAs) are a large class of small non-coding RNAs present in all eukaryotes sequenced thus far. As a family, they have been well-characterized as playing a central house-keeping role in ribosome biogenesis, guiding either the sequence-specific chemical modification of pre-rRNA or its processing [1]. However, multiple independent studies show non-canonical functions including roles in splicing regulation, chromatin architecture regulation and in stress response pathways for small subsets of snoRNAs (reviewed in [2]).

## RESULTS

To characterize the snoRNA landscape, we recently elaborated protocols to sequence the small RNA transcriptome. The analysis of the sequencing of different human cell lines in the presence or absence of the depletion of relevant snoRNA interactors allows the study of the abundance, the different sequence forms and the protein dependencies of the full complement of snoRNAs. Most human snoRNAs display strong processing conservation, and two groups of box C/D snoRNAs differing in their ends with respect to boxes C and D and in their terminal stem length are apparent, with important repercussions on their stability and protein dependencies [3]. The long forms, which display more stability in their terminal ends, and are typically located in close proximity to the next downstream exon of the host gene, are generally dependent on the expression of the core snoRNP NOP58, thought to be essential for box C/D snoRNA production. In contrast, a subset of the short forms are dependent on the splicing factor RBFOX2 which has previously been shown to interact with lncRNAs with snoRNA ends [4], suggesting the existence of at least two distinct box C/D snoRNA groups differing in their processing and binding preferences. In addition, despite their ubiquitous role in ribosome biogenesis, many human snoRNAs display tissue-specificity, supporting the concept of non-canonical functionality for subsets of snoRNAs.

## CONCLUSIONS

While many snoRNAs display canonical features and protein dependencies, a subset consistently show distinct sequence and structure features, and non-canonical protein dependencies suggesting distinct maturation pathways and supporting recent reports of non-canonical functionality. The analysis of deep-sequencing datasets gives insight into snoRNA processing and function, supporting the emerging picture of a central and versatile family of regulatory RNAs.

## REFERENCES

1. Dieci G, Preti M, Montanini B. Eukaryotic snoRNAs: a paradigm for gene expression flexibility. *Genomics* 2009, 94:83-88.
2. Dupuis-Sandoval F, Poirier M, Scott MS. The emerging landscape of small nucleolar RNAs in cell biology. *WIREs RNA* 2015.
3. Deschamps-Francoeur G, Garneau D et al Identification of discrete classes of small nucleolar RNA featuring different ends and RNA binding protein dependency *NAR* 2014, **42**(15):10073-85.
4. Yin QF et al, Long noncoding RNAs with snoRNA ends. *Mol Cell* 2012, **48**(2):219-30.

# The importance of conservation and homology in alternative splicing at the protein level

Federico Abascal<sup>1</sup>, Iakes Ezkurdia<sup>2</sup>, Juan Rodriguez-Rivas<sup>1</sup>, Jose Manuel Rodriguez<sup>3</sup>,  
Enrique Carillo-de Santa Pau<sup>1</sup>, Angela del Pozo<sup>4</sup>, Jesús Vázquez<sup>5</sup>, Alfonso Valencia<sup>1,3</sup>,  
Michael L. Tress<sup>1\*</sup>

<sup>1</sup> Structural Biology and Bioinformatics Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain,

<sup>2</sup> Unidad de Proteómica, Centro Nacional de Investigaciones Cardiovasculares, CNIC, Madrid, Spain,

<sup>3</sup> National Bioinformatics Institute (INB), Spanish National Cancer Research Centre (CNIO), Madrid, Spain,

<sup>4</sup> Instituto de Genética Médica y Molecular, Hospital Universitario La Paz, Madrid, Spain,

<sup>5</sup> Laboratorio de Proteómica Cardiovascular, Centro Nacional de Investigaciones Cardiovasculares, CNIC, Madrid, Spain

\*To whom correspondence should be addressed: mtress@cnio.es

---

## BACKGROUND

Although eukaryotic cells can express a wide range of alternatively spliced transcripts (the Ensembl version of the human genome currently houses over 70,000 alternative coding variants) it is not clear whether these variants can be defined as dominant or alternative, or whether genes can express a range of transcripts simultaneously across cells. To date large-scale investigations into the pattern of dominance at the transcript level across distinct tissues have produced contradictory results [1-2].

## RESULTS

We interrogated 8 large-scale human proteomics experiments using a rigorous peptide identification strategy and carried out an analysis of alternative splicing at the protein level. We found that the vast majority of protein-coding genes may have a single dominant protein isoform: while we identified peptides for 12,716 protein-coding genes, there was evidence for just 282 alternative splicing events. Remarkably over 20% of the splice events we identified were generated from highly conserved homologous exon substitutions and very few of these splicing events would disrupt protein functional domains [3]. We were able to identify a unique dominant proteomics isoform for 5,011 genes in the analysis [4]. These main proteomics isoforms were overwhelmingly supported by reference isoforms from two completely orthogonal sources, CCDS consensus variants, chosen by manual genome curation teams, and APPRIS principal isoforms [5], predicted automatically from protein conservation, structure and function.

## CONCLUSIONS

The agreement between three orthogonal sources significantly reinforces the probability that the main proteomics isoform is the dominant cellular isoform. In particular, the agreement with APPRIS principal isoforms demonstrates that the cellular machinery tends to express the most conserved splice isoform. The alternative isoforms generated from homologous exons were highly conserved, first appearing 460 million years ago. The combination of proteomics evidence and ancient origin indicates the importance of homologous exons in alternative splicing.

## REFERENCES

1. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **14** (2012).
2. González-Porta, M. *et al.* Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biology* **14** (2013).
3. Abascal F. *et al.* Alternatively spliced homologous exons have ancient origins and are highly expressed at the protein level. *PLOS Computational Biology* (2015).
4. Ezkurdia, I. *et al.* Most highly expressed protein-coding genes have a single dominant isoform. *Journal of Proteome Research* **14** (2015).
5. Rodriguez, J.M. *et al.* APPRIS: annotation of principal and alternative splice isoforms. *Nucleic Acids Research* **41** (2013).

# STATegra: Statistical and Bioinformatics tools for multi-omics data integration

Ana Conesa<sup>1,2\*</sup>, The STATegra Consortium<sup>3</sup>

<sup>1</sup> Genomics of Gene Expression Lab, Principe Felipe Research Center, Valencia, Spain,

<sup>2</sup> Microbiology and Cell Science, Institute for Agricultural Sciences, University of Florida, Gainesville <sup>3</sup><http://stategra.edu>

\*To whom correspondence should be addressed: [aconesa@cipf.es](mailto:aconesa@cipf.es)

---

## BACKGROUND

Next generation sequencing has speed up genome analysis and brought omics research closer to many organisms and biological scenarios. Today an increasing number of research projects propose the combined use of different omics platforms to investigate diverse aspects of genome functioning. However, standard methodologies for the integration of diverse omics data types are not yet ready and researchers frequently face post-experiment question on how to combine data of different nature, variability, and significance into an analysis routine that sheds more light than the analysis of individual datasets separately. Novel statistical and bioinformatics tools are needed for these emerging analysis scenarios. STATegra is a FP7 project aimed to address these questions and provide resources for gene expression-centered, multi-omics data analysis.

## RESULTS

We have created a multi-omics data-set for method development consisting of a controlled mouse B-cell differentiation experiment with 6 time points and treatments. We have measured up to 8 different omics: RNA-seq, microRNA-seq, single-cell-RNA-seq, RRBS-seq (DNA methylation), ChIP-seq, DNase-seq, Metabolomics and Proteomics. We have developed a set of integrative analysis tools centered on gene expression data. I will present approaches for:

- Comparison of Performance Metrics across omics types and multi-omics power analysis
- Data exploration to reveal shared and data-type specific signal properties
- Mapping chromatin signals to gene annotations.
- Pair-wise (e.g. RNA-seq vs DNase-seq) integrative analysis of time course data
- Define regulatory programs for co-expressed genes based on chromatin and post-transcriptional regulation.
- Integrate ChIP-seq, RNA-seq and Metabolomics data to study the impact of transcriptional regulation on metabolic changes.
- Pathway-level integrative analysis of multi-omics data.

Methods are being implemented either in R under the STATegRa Bioconductor package or as web-based tools such as Paintomics.

## CONCLUSIONS

STATegra provides a wide variety of resources for the analysis of multi-omics data, available to the scientific community through open access distributions

# Novel Approach to Identifying Cleavage Sites of Bacterial Toxin-Antitoxin Systems

Toomas Mets<sup>1\*</sup>, Markus Lippus<sup>2</sup>, David Schryer<sup>3</sup>, Tanel Tenson<sup>4</sup>, Niilo Kaldalu<sup>5</sup>

<sup>1,2,3,4,5</sup> Institute of Technology, University of Tartu, Nooruse 1, Tartu 50411, Estonia \*Correspondence: toomas.mets@ut.ee

---

## BACKGROUND

Most bacteria contain toxin-antitoxin (TA) systems, which are composed of growth-inhibiting toxins and corresponding antitoxins that bind to and neutralize them. The antitoxins are continuously produced in excess of the toxins yet are labile and rapidly degrade. Under conditions of stress, antitoxin production is inhibited, their concentration drops, and toxins are freed. This process is believed to help the bacteria survive stress. TA systems have also been linked to antibiotic tolerance by helping a fraction of bacteria to survive antimicrobial treatment in a dormant state, which is a serious clinical problem.

Many toxins are endoribonucleases and widely considered to cleave exclusively mRNA. However, the MazF toxin in *E. coli* has been shown to cleave a fragment from the 3'-end of the 16S rRNA. We have also observed large fragments of ribosomal RNA in our previous toxin overexpression experiments. This study aims to characterize rRNA cleavage by toxins in *E. coli*.

## RESULTS

We overexpressed MazF and MqsR toxins in *E. coli* and identified the cutting sites of these toxins using RNA-seq. It has been observed that MazF cleaves RNA at the 5' end at ACA recognition sequences, thereby producing 5'-hydroxyl and 2',3'-cyclic phosphate ends while other endoribonucleases produce 5'-phosphate and 3'-hydroxyl termini. To leverage this difference, we either left RNA samples untreated or treated them with T4 polynucleotide kinase, which acts as a 5'-kinase and 3'-phosphatase thereby allowing for ligation of RNA adapters to the 5'-ends and polyadenylation of the 3'-ends of the MazF cleavage products. This treatment enables one to differentiate between the cleavage sites of MazF and those of the other ribonucleases. After end sequencing of the 3'- and 5'-ends, we trimmed the low quality bases and aligned the reads to the *E. coli* genome using bowtie1 (with one mismatch). The ends of aligned transcripts were counted by position on both the 16S and 23S precursor rRNAs. Comparing overexpression samples to control samples we discovered that both MazF and MqsR cut rRNA to a substantial degree at specific sites and induce cleavage by other ribonucleases.

## CONCLUSIONS

In our work we demonstrated a novel method which allows us to accurately map the cleavage sites of bacterial TA systems and using it we were able to find cleavage sites specific to the toxins used in our experiments. In the future, we believe that this method will be very useful in understanding which genes and how are regulated by TA systems and also for investigating how mRNA is being degraded by the toxins.

# Complementary mapping approaches improve circRNA detection

Franziska Metge<sup>1\*</sup>, Marek Rajman<sup>2</sup>, Gerhard Schratt<sup>2</sup> and Christoph Dieterich<sup>1</sup>

<sup>1</sup> Computational RNA Biology, MPI for Biology of Ageing, Cologne, Germany,

<sup>2</sup> Institute of Physiological Chemistry University of Marburg, Marbug, Germany

\*To whom correspondence should be addressed: franziska.metge@age.mpg.de

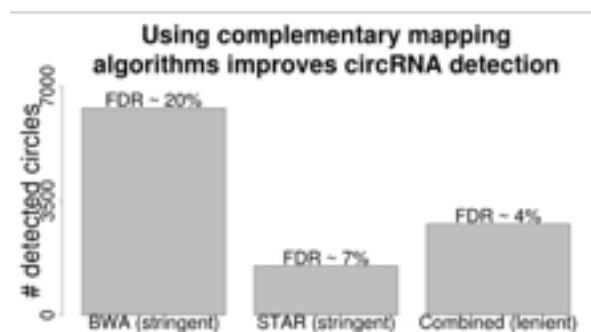
## BACKGROUND

Circular RNAs (circRNAs) emerge during the post-transcriptional processing of RNAs by a process called back-splicing. A downstream 5' splice site is linked with an upstream 3' splice site to form a circular instead of the corresponding linear host transcript (host gene) [1]. Although there is functional evidence for several circRNAs acting as miRNA sponges [2], most circRNAs' function remains unknown. Furthermore, a recent study indicates that circRNA abundances vary across different tissues and developmental stages [3]. This specific expression indicates an unknown yet meaningful purpose of circRNAs. Despite this, the lack of a standard method for circRNA detection from RNA sequencing still leads to unreproducible circRNA loci [1].

Although most methods describe similar circRNA detection strategies, the used read mapper differs from one study to the next, with each mapper having its advantages as well as known drawbacks. In all these approaches a compromise between a high false discovery rate (FDR) and a low number of detected circRNAs has to be made.

## RESULTS

We studied picrotoxin (PTX) induced synaptic downscaling in rat hippocampal neurons to identify potential regulatory mechanisms of neuronal homeostatic plasticity. Circles were identified with our new approach of mapping RNA sequencing reads using two different mappers, BWA-MEM [4] and STAR [5], identifying back-spliced reads (chimeras) indicating circRNA candidates using CIRI [6] and the in-house DCC (Detecting CircRNAs from Chimeras), keeping only candidates detected by both methods. Using this approach we identified 2784 candidates circles with a 4 % FDR. 1510 of these are shared among control and PTX treated samples.



**Figure 1. Comparison of different mapping approaches:** Even with a stringent filtering, BWA/CIRI identifies 7292 circles at the cost of a high FDR, while STAR/DCC has a considerably lower FDR at the cost of only identifying 1505 circles. Using our approach has the advantage of accepting lowly expressed circles if both mapping algorithms map a read to a circular junction, identifying 2784 circles with a 4% FDR.

## **CONCLUSIONS**

We demonstrate that combining two complementary mapping algorithms decreases the FDR and increases the amount of circRNA candidates. With this approach we achieve a five-fold lower FDR than using only the first method, while detecting two-fold more circRNAs than using only the second method.

## **REFERENCES**

1. Lasda and Parker; RNA, Volume 20, 2014
2. Hansen et al.; EMBO J, Volume 30, 2011
3. You et al.; Nature Neuroscience, Volume 18, 2015
4. Li H. and Durbin R.; Bioinformatics, Volume 25, 2009
5. Dobin et al; Bioinformatics, Volume 29, 2013
6. Gao et al.: Genome Biology, Volume 16, 2015



# Dumpster Diving: Finding the source of every last RNA-Seq read

Sergei Mangul<sup>1\*</sup>, Timothy Daley<sup>2</sup>, Nicolas Strauli<sup>3</sup>, Ryan Hernandez<sup>3</sup>, Roel Ophoff<sup>4</sup>, Andrew Smith<sup>2</sup>, Max Seibold<sup>5</sup>, Eleazar Eskin<sup>1,4</sup>, Noah Zaitlen<sup>6</sup>

<sup>1</sup>UCLA, Computer Science, Los Angeles, CA,

<sup>2</sup>Division of Biological Sciences at University of Southern California

<sup>3</sup>UCSF, Department of Bioengineering and Therapeutic Sciences, San Francisco, CA,

<sup>4</sup>UCLA, Human Genetics, Los Angeles, CA,

<sup>5</sup>National Jewish Health, Department of Pediatrics, Denver, CO,

<sup>6</sup>UCSF, Department of Medicine, San Francisco, CA

To whom correspondence should be addressed: [smangul@ucla.edu](mailto:smangul@ucla.edu)

---

## BACKGROUND

Advances in RNA sequencing technology and the ability to generate deep coverage data in the form of millions of reads provide an unprecedented opportunity to probe the universe of gene expression. Standard RNA-seq analysis protocols map reads against a host reference genome to determine the placement of the reads on the genome. Mapping-based protocols are complemented by assembly procedures to accurately profile the origin of reads condensed into isoform transcripts. Many reads are discarded by these protocols and the possibility that reads originate outside of the extant genome is usually ignored. In order to identify shortcomings of existing technologies and tools as well as identify novel uses of RNA-Seq data, we aim to profile the origin of every last read delivered by RNA sequencing. Our study reveals that the vast majority of unmapped reads are human reads discarded by the mapping protocol. Many of these discarded human reads contains sequences originating from the recombined Ig locus of B and T lymphocytes. Other unmapped human reads correspond to novel exon junctions from previously unknown isoforms. In addition to human DNA, the human body harbors a diverse microbial ecosystem. We identified a substantial number of reads mapping to non-human sequences.

## RESULTS

We used the Illumina Hiseq 2000 platform to produce peripheral blood and nasal epithelium Illumina RNA-Seq data sets from 100 individuals. The read origin protocol (ROP) was used to obtain a detailed profile of the RNA-Seq data. First we extracted the human sequences by mapping the reads against the human transcriptome and genome reference (90.6% of read pairs are mapped to the human genome via tophat2). After the reads compatible with the human reference were extracted we filtered out 2.19% of the reads, which were low-quality and/or low complexity. We attempted realignment of remaining reads to the human reference sequences using the MEGABLAST aligner. MEGABLAST was able to identify 4.64% of reads with 100% identity to the human reference missed by the short reads aligners. Among the identified human sequences 57% of reads corresponded to interspersed repeats. Relaxing the 100% identity constrain allows for identification of an additional 1.72% of human reads. The remaining non-host reads are used to perform the phylogenetic profiling of the microbial communities. We use the bacterial hyper-variable regions from the gene families as the reference to map the reads. We were able to identify 0.27% reads uniquely mapped to microbial genes. Lastly we sought to find reads derived from the immune cells. We identify the 0.44% of reads corresponding to the V(D)J recombination from lymphocytes. The remaining 0.14%(141,612 singleton reads) received no assignment. After identifying each read with its origin, we investigate the complexity of each class within each library using

capture-recapture models and the software package preseq. We observe exons to be saturated, but at most 20% of the full diversity of reads originating from the human genome have been sequenced. This indicates the presence of a large number of rare transcriptional events that comprise the full transcriptional diversity.

## **CONCLUSIONS**

Profiling all the sequencing reads helps to accurately measure all transcriptional activity of individuals and provides new insights into sample quality and environment. Detailed classification of the sequencing reads derived by the RNA-Seq protocol helps to determine the shortcomings left to tackle under different analysis strategies as well as novel uses of RNA-seq data sets.

## Organizing committee

Eduardo Eyras

Pompeu Fabra University,

Barcelona, Spain

eduardo.eyras@upf.edu



Klemens Hertel

University of California, Irvine

Irvine, CA, United States

khertel@uci.edu



Yoseph Barash

University of Pennsylvania,

Philadelphia, PA, USA

yosephb@mail.med.upenn.edu





---

Prepared by  
BioCiphers Lab  
for  
IRB-SIG 2015

---