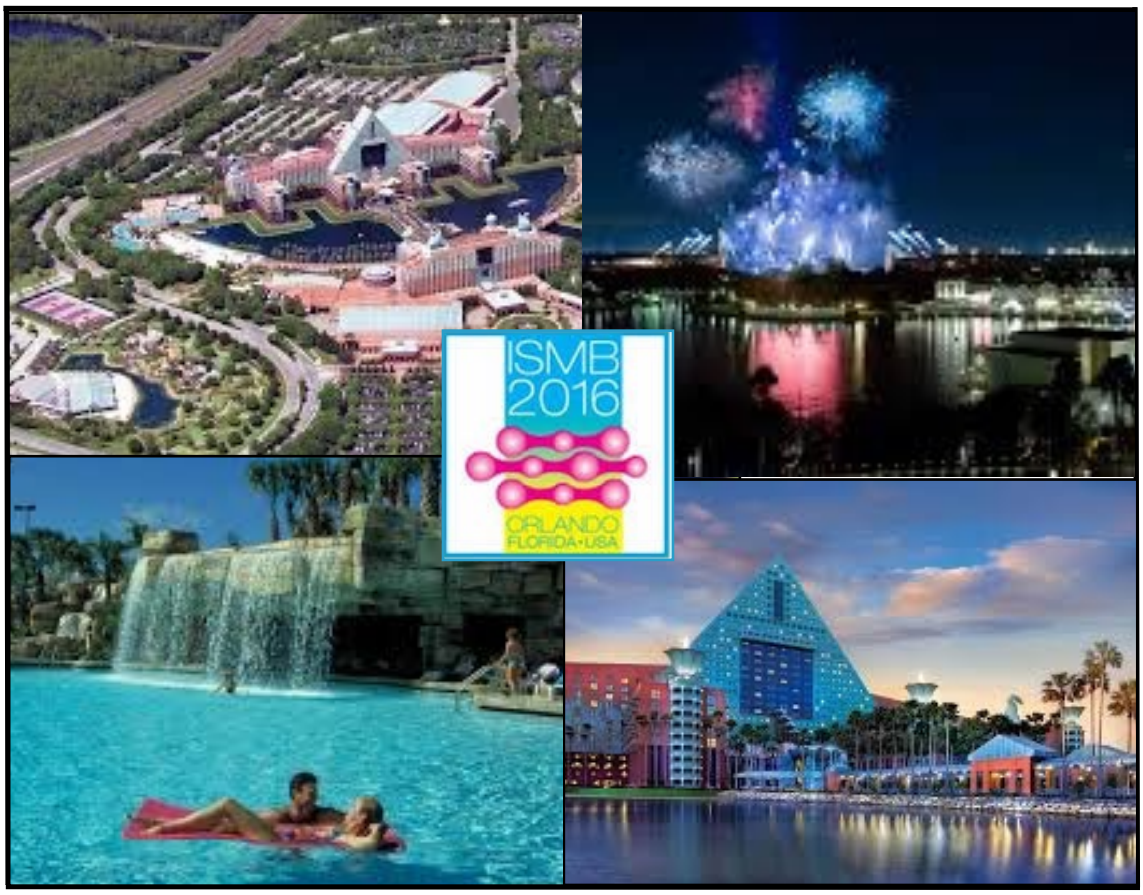


Integrative RNA Biology
Special Interest Group Meeting

July 8th, 2016
Orlando, Florida, USA

IRB
Integrative RNA Biology
group

Content

Program - Friday, July 8th

7:30 a.m. – 8:30 a.m.	Registration
SESSION 1	RNA PROCESSING
08:30 - 08:40	Opening notes
08:40 - 09:15	**Invited speaker: Christine Mayr.** 3'UTR-mediated protein complex formation determines protein functions
09:15 - 09:30	**Michael Hamilton.** *Alternative and differential polyadenylation detection from single molecule long sequencing*
09:30 - 09:45	**Alina Selega.** *Statistical modelling of protein–RNA binding time series reveals widespread control of mRNA degradation in stress response*
09:45 - 10:00	**Meredith Corley.** *An RNA structure mediated post– transcriptional model of α–1–antitrypsin deficiency*
10:00 - 10:15	**Steven Brenner.** *Polysome fractionation analysis reveals nonsense–mediated mRNA decay targets are monosomal and supports model That 3' UTR length is a weak predictor of NMD*
10:15 - 10:45	Coffee Break
SESSION 2	RNA IN DIFFERENTIATION, CELL FUNCTION, AND DISEASE
10:45 - 11:00	**Yi Xing.** *SURVIV: Survival Analysis of mRNA Isoform Variation*
11:00 - 11:15	**Jorge Vaquero.** *A new view of transcriptome complexity and regulation through the lens of local splicing variations*
11:15 - 11:30	**Elena Sotillo.** *Alternative Splicing of CD19 in relapsed B–cell acute lymphoblastic leukemia*
11:30 - 11:45	**Benjamin Cieply.** *From culture to clinic: single cell profiling of the mammalian immune system.*
11:45 - 12:20	**Invited speaker: Simon (Hualin) Xi.** *From Genome Biology to Drug Discovery: Integrative Mining of Human Transcriptomics Data for Novel Therapeutic Targets*
12:30 - 13:45	Lunch
SESSION 3	CHARACTERIZATION OF RNA SEQUENCE AND STRUCTURE
13:45 - 14:00	**Steve Mount.** *Nano–exon discovery*
14:00 - 14:15	**Jeffrey Zuber.** *Precision in RNA secondary structure prediction: A sensitivity analysis of the folding nearest neighbor parameters*
14:15 - 14:30	**Tamir Tuller.** *Multiple roles of local RNA folding along the coding region in gene expression regulation*
14:30 - 14:45	**Katrina Kutchko.** *Identifying conserved RNA structures in alphaviruses: An integrated approach*
14:45 - 15:20	**Invited speaker: Hagen Tilgner.** *Getting the entire message: Coordination of variable sites on RNA transcripts*
15:30 - 16:00	Coffee Break (and start of poster session)
16:00 - 17:30	***Poster session***
SESSION 4	RNA REGULATION IN NORMAL AND DISEASE CONDITIONS
17:30 - 17:45	**Azim Dehghani-Amirabad.** *Why asking for miRNA– gene interactions is wrong: A new paradigm for miRNA target prediction with expression data*
17:45 - 18:20	**Invited speaker: Olga Troyanskaya.** *Integrative, genome-wide characterization of the human disease landscape*
18:20 – 18:35	Concluding notes, poster prize announcement
19:30	**IRB-SIG dinner** (Il Mulino, Walt Disney World Swan Resort).

Sponsors

RNA Society

Abstracts

3'UTR-mediated protein complex formation determines protein functions

Christine Mayr[1]
[1]Memorial Sloan Kettering Cancer Center, New York, NY

At least half of human genes use alternative cleavage and polyadenylation (ApA) to generate mRNA transcripts that differ in the length of their 3' untranslated regions (3'UTRs) while producing the same protein. We recently discovered a new role for 3'UTRs. We observed that 3'UTRs can mediate protein-protein interactions, and thus, can determine protein localization and protein functions. For example, we showed that the long 3'UTR of CD47 is required for the formation of a protein complex between CD47 and SET. The interaction with SET enables efficient cell surface expression of CD47 protein. In contrast, CD47 protein that was generated by the short CD47 3'UTR does not interact with SET, but instead interacts with other partner proteins. Thus, the alternative 3'UTRs of CD47 determine the formation of alternative protein complexes of CD47. This has functional consequences, as the protein interaction partners determine the functions of CD47 membrane protein. We showed that CD47 that was generated by the long 3'UTR (CD47-LU) has a pro-survival role, whereas CD47 that was generated by the short 3'UTR (CD47-SU) promotes cell death. Importantly, alternative protein complex formation mediated by alternative 3'UTRs is not restricted to CD47. It seems that all long 3'UTRs have the capacity to mediate protein-protein interactions. Intriguingly, our quantitative mass spectrometry data on several candidates revealed that the protein interaction partners recruited by 3'UTRs are highly specific and are determined by the sequence of the 3'UTR as well as by the amino acid sequence of the nascent protein. This shows that 3'UTR-mediated protein complex formation is very widespread and suggests that alternative 3'UTRs have evolved to enable multi-functionality of proteins.

Alternative and differential polyadenylation detection from single molecule long sequencing

Michael Hamilton[1], Salah E. Abdel-Ghany[2], Anireddy S.N. Reddy[2], and Asa Ben-Hur[1*]

[1] Computer Science Department, Colorado State University[2] Department of Biology, Program in Molecular Plant Biology, Program in Cell and Molecular Biology, Colorado State University

*To whom correspondence should be addressed: asa@cs.colostate.edu

BACKGROUND

The recent extension of single molecule sequencing from Pacific Biosciences to transcriptome sequencing, known as isoform sequencing (Iso-Seq), is providing opportunities to elucidate full length splice forms and their regulation [1]. A key regulatory process of transcripts is the location of the polyadenylation (poly-A) cleavage site. It has recently been shown that in plants, alternative poly-A (APA) is prolific and has tissue specificity [2].

RESULTS

Iso-Seq data provide a means to interrogate the 3' ends of transcripts, but this has not been previously explored. We have recently proposed a computational method for identifying and quantifying alternative polyadenylation using such data and demonstrated its effectiveness [1]. We analyzed around 11,000 sorghum genes with identifiable poly-A sites from 800,000 Iso-Seq reads. Of these genes, we found that around 70% exhibit APA, i.e. have multiple poly-A sites. An example of APA detected from the data is shown in Figure 1 (A) where the number of reads supporting a cleavage site is quantified. Finally, the results were confirmed in a comparison with cleavage sites inferred from over 200,000 expressed sequence tags (ESTs), and also from experimental validation using 3' Rapid Amplification of cDNA Ends (RACE) in five randomly selected genes. Differential APA in sorghum in response to drought treatment was also found, as evidenced in Figure 1 (B).

CONCLUSIONS

We find that our method accurately identifies poly-A cleavage sites and detects substantial levels of APA with limited sequencing depth, and without the need for a specialized assay for studying transcript 3' ends. With increasing read depth, decreasing cost, and additional applications such as APA detection, the use of Iso-Seq may become more widespread.

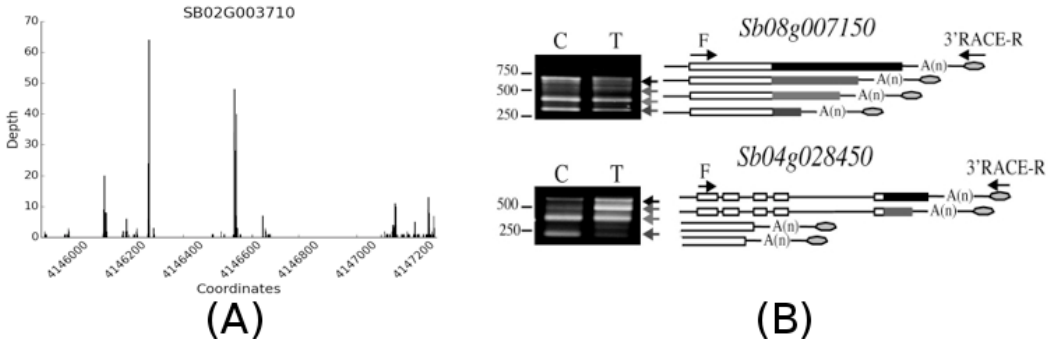


Figure 1: **(A)** Poly-(A) sites detected from IsoSeq. **(B)** 3' RACE validation and evidence of differential APA in drought treatment.

REFERENCES

1. Abdel-Ghany S., Hamilton M., Jacobi J., Ngam P., Devitt N., Schilkey F., Ben-Hur A., and Reddy A. A survey of the sorghum transcriptome using single-molecule long reads. Nature Communications. (2016). (Accepted).
2. Wu X., Liu M., Downie B., Liang C., Ji G, Li Q., and Hunt A. Genome-wide landscape of polyadenylation in Arabidopsis provides evidence for extensive alternative polyadenylation. Proc Natl Acad Sci. **108:30**. (2011).

Statistical modelling of protein-RNA binding time series reveals widespread control of mRNA degradation in stress response

Alina Selega[1*], Sander Granneman[2], and Guido Sanguinetti[1]

[1] School of Informatics, University of Edinburgh, [2] Centre for Synthetic and Systems Biology, University of Edinburgh

*To whom correspondence should be addressed: alina.selega@ed.ac.uk

---

BACKGROUND

Cells regulate gene expression through a complex coordination of processes. Understanding the relative balance and dynamical control of production and decay of macromolecules is a central challenge for unravelling the mechanisms of gene expression. Model-based efforts using labelled RNA time series of localization data on polymerase are hampered by a lack of direct measurements of degradation dynamics, and frequently resort to a simplifying assumption of constant degradation rates [1, 2].

RESULTS

Here we use an unpublished data set assaying the binding of RNA degradation factor Nab3 in a rapid time series after the imposition of a nutrient shift on exponentially growing Saccharomyces cerevisiae. The data were obtained using a modified version of the CRAC protocol [3], which greatly reduces the UV irradiation time. This permits the collection of time points as early as 1 minute, detailing the dynamic changes in protein binding transcriptome-wide during the crucial early stages of stress response. We statistically model treatment and control time series using Gaussian Processes, a non-parametric class of models for Bayesian regression [4]. This enables us to compute Bayes factors to test the hypotheses that the binding of degradation factors (and hence by implication, RNA degradation) changes dynamically during stress. Our results show that a large fraction of the transcriptome does indeed display highly dynamic binding behavior. We also find that a subset of non-coding RNAs, cryptic unstable transcripts (CUTs), are significantly over-represented among the dynamically binding transcripts.

CONCLUSIONS

Our data provide the first high-resolution time series measurements of the action of RNA degradation factors during the imposition of stress in any organism. Statistical analysis of the data reveals a highly dynamic behavior, calling into question earlier modeling assumptions and shedding new light on the kinetics of gene expression regulation.

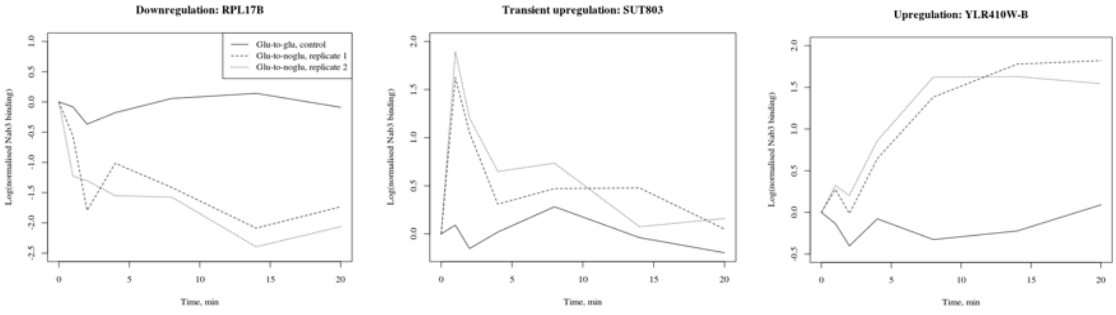


Figure 1: Binding profiles with Nab3 during stress of example transcripts RPL17B (left), SUT803 (middle), and YLR410W-B (right) demonstrate downregulation, transient upregulation, and upregulation, correspondingly.

REFERENCES

1. Rabani, Michal, et al. "Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells." *Nature biotechnology* 29.5 (2011): 436-442.
2. Marguerat, Samuel, et al. "Contributions of transcription and mRNA decay to gene expression dynamics of fission yeast in response to oxidative stress."*RNA biology* 11.6 (2014): 702-714.
3. Granneman, Sander, et al. "Identification of protein binding sites on U3 snoRNA and pre-rRNA by UV cross-linking and high-throughput analysis of cDNAs." *Proceedings of the National Academy of Sciences* 106.24 (2009): 9613-9618.
4. Äijö, Tarmo, et al. "Methods for time series analysis of RNA-seq data with application to human Th17 cell differentiation." *Bioinformatics* 30.12 (2014): i113-i120.

An RNA structure mediated post-transcriptional model of α-1-antitrypsin deficiency

Meredith Corley[1], Amanda Solem[2], Gabriela Phillips[3], Lela Lackey[1], Benjamin Ziehr[4], Nathanial Moorman[1], and Alain Laederach[1*]

[1] University of North Carolina at Chapel Hill, [2] Hastings College, [3] Oragenics, Inc, [4] University of Wisconsin-Madison

*To whom correspondence should be addressed: alain@unc.edu

BACKGROUND

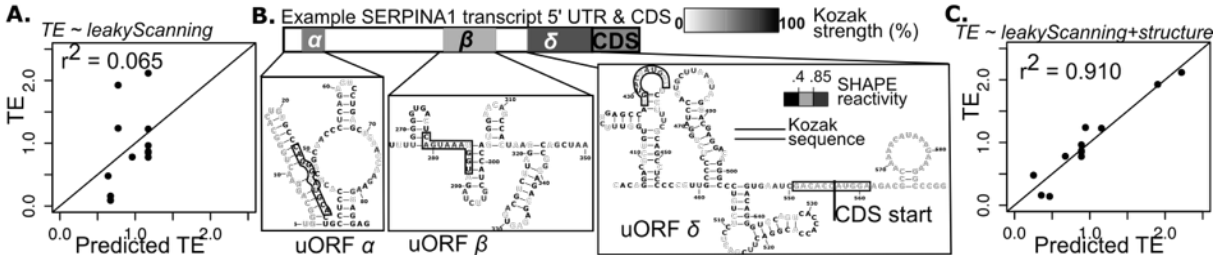
In human tissues the protein expression phenotype is only very weakly correlated to transcript expression [1] due to post-transcriptional layers of regulation that are poorly understood, indicating a need for integrative models that bridge this divide. In this work we aim to develop a quantitative post-transcriptional model that predicts the experimentally measured expression of a clinically important protein, α-1-antitrypsin. The parent gene, SERPINA1, transcribes a total of eleven transcript isoforms that only differ in their 5' untranslated region (UTR) and therefore all produce the same peptide. However, the transcripts have distinct translational efficiencies, which direct the *amount* of peptide each transcript produces. α-1-antitrypsin is expressed in a variety of human tissues as a necessary protectant. Deficient levels α-1-antitrypsin can cause COPD, adult liver disease and infantile cirrhosis [2].

RESULTS

The SERPINA1 transcripts contain distinct secondary structural features as well as different combinations of upstream open reading frames (uORFs), short protein-coding sequences in the 5' UTR known to inhibit translation. In luciferase reporter assays some of the transcripts show repressed translational efficiency, which increases five-fold when we mutate specific uORFs. We first model translational efficiency in the SERPINA1 transcripts with a leaky scanning model of translation that combines the Kozak sequence strengths of the uORFs and coding sequence (CDS) [3]. This model fails to predict our data (Figure 1A), indicating that other factors in the 5' UTRs, like secondary structure, regulate the uORFs' effect on translation. We constructed RNA structural models of the SERPINA1 transcripts with whole-transcript structure probing SHAPE-MaP experiments [4]. Integrating Kozak sequence strength with the ΔG of folding around each Kozak sequence (Figure 1B), we develop a model that predicts translational efficiency in our data with 91% accuracy (Figure 1C).

CONCLUSIONS

In this work we report that α-1-antitrypsin output in human tissues is affected by uORFs that selectively inhibit translation in several SERPINA1 transcripts, dependent upon structure. Our predictions show that a combination of secondary structure and Kozak sequence strength determines the translational control wielded by each uORF in a structure-revised leaky scanning model of translation.



REFERENCES

1. Maier, T., Guell, M. & Serrano, L. Correlation of mRNA and protein in complex biological samples. FEBS Letters **583** (2009).
2. Primhak, R. A. & Tanner, M. S. Alpha-1 antitrypsin deficiency. Archives of Disease in Childhood **85** (2001).
3. Ferreira, J. P., Overton, K. W. & Wang, C. L. Tuning gene expression with synthetic upstream open reading frames. Proceedings of the National Academy of Sciences USA **110** (2013).
4. Siegfried, N. A., Busan, S., Rice, G. M., Nelson, J. A. & Weeks, K. M. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). Nature Methods **11** (2014).

Polysome fractionation analysis reveals nonsense-mediated mRNA decay targets are monosomal and supports model that 3' UTR length is a weak predictor of NMD

James P. B. Lloyd[1,2,3], Courtney E. French[4], and **Steven E. Brenner[1,2,3,4*]**

[1] Center for RNA Systems Biology, [2] Department of Plant and Microbial Biology, [3] QB3, [4] Department of Molecular and Cell Biology, University of California, Berkeley, CA, USA

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

---

BACKGROUND

Alternative splicing of human genes generates vast transcriptomic and proteomic diversity. At least a fifth of alternatively spliced human genes produce alternate transcript isoforms that are targeted to nonsense-mediated mRNA decay (NMD). The NMD pathway targets and degrades transcripts with a premature termination codon, which is recognized during translation termination. Models propose that premature termination codons are determined by the presence of a downstream exon-exon junction or a long 3' UTR. Recent work from our group showed that globally, in humans, exon-exon junctions downstream of a stop codon, but not long 3' UTRs, determined NMD-sensitivity of transcripts. Thus we found many NMD targets through inhibition of NMD and identification of responsive transcripts that have a downstream exon-exon junction.

RESULTS

Polyribosomal fractionation can identify transcripts bound to a single ribosome (the monosome) or multiple ribosomes. Here we use recently published polysome fractionation RNA-seq data to characterize differences between NMD targeted and non-targeted transcript isoforms. We found an enrichment of NMD targeted transcripts in the monosome and low polysome fractions (<3 ribosomes), while non-targeted isoforms were enriched in higher polysome fractions (>5 ribosomes). Combining polysome fraction data with RNA-seq data can therefore help to refine NMD target identification. We also found that long 3' UTR transcripts were not more likely than short 3' UTR transcripts to be enriched in the monosome fraction, in contrast to transcripts with a downstream exon-exon junction, which are enriched in the monosome.

CONCLUSIONS

In summary, we have demonstrated that many identified NMD targeted transcripts are predominantly found in the monosome fraction, fitting with the model that NMD targets are degraded after the pioneer round of translation. Additionally, these findings are consistent with our recent work that highlighted the importance of downstream exon-exon junctions, but not long 3' UTRs, in determining the NMD-sensitivity of transcript isoforms.

SURVIV: Survival analysis of mRNA isoform variation

Shihao Shen[1], Yuanyuan Wang[2], Chengyang Wang[3], Ying Nian Wu[4], Yi Xing[1,*]

[1]Department of Microbiology, Immunology & Molecular Genetics, [2]Department of Molecular and Medical Pharmacology, [3]Bioinformatics Interdepartmental Graduate Program, [4]Department of Statistics, University of California, Los Angeles, Los Angeles, CA 90095, USA

*To whom correspondence should be addressed: yxing@ucla.edu

BACKGROUND

The rapid accumulation of clinical RNA-seq datasets has provided the opportunity to associate mRNA isoform variations to clinical outcomes. Here we report a statistical method SURVIV (Survival analysis of mRNA Isoform Variation), designed for identifying mRNA isoform variation associated with patient survival time. A unique feature and major strength of SURVIV is that it models the measurement uncertainty of mRNA isoform ratio in RNA-seq data.

RESULTS

Simulation studies suggest SURVIV outperforms the conventional Cox regression survival analysis, especially for datasets with modest sequencing depth. We applied SURVIV to TCGA RNA-seq data of invasive ductal carcinoma as well as five additional cancer types. Alternative splicing based survival predictors consistently outperform gene expression based survival predictors, and the integration of clinical, gene expression and alternative splicing profiles leads to the best survival prediction.

CONCLUSIONS

We anticipate SURVIV will have broad utilities for analyzing diverse types of mRNA isoform variation in large-scale clinical RNA-seq projects.

A new view of transcriptome complexity and regulation through the lens of local splicing variations

Jorge Vaquero-Garcia[1,2], Matthew Gazzara[1,3], Alejandro Barrera[1,2], Kristen Lynch[3], Yoseph Barash[1,2*]

[1] Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, United States; [2]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, United States, [3]Department of Biochemistry and Biophysics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, United States
*To whom correspondence should be addressed: yosephb@upenn.edu

---

BACKGROUND

As much as 95% of multi-exon genes undergo alternative splicing (AS) and this regulated process has been implicated in numerous aspects of gene regulation, development, and disease[123]. However, accurate mapping and quantification of AS using RNA-seq data has remained a challenge, particularly for complex splicing patterns that do not fit the mold of classically defined AS events (e.g. cassette exons, alternative 5' splice sites).

RESULTS

Recently, we described a new computational framework that defines and quantifies AS in units of local splicing variations (LSVs). The LSV formulation captures both classical AS events as well as more complex (non-binary) patterns of splicing that other commonly used tools fail to characterize. We have analyzed LSVs in over 250 RNA-seq experiments from diverse vertebrates, tissues, developmental stages, and conditions with altered splicing factor expression. Our analysis reveals that complex splicing variations are much more prevalent than previously appreciated, comprising approximately a third of variations observed in the human transcriptome. Such complex splicing variations are further enriched among differentially spliced LSVs, suggesting that complex LSVs are an important aspect of gene regulation and thus need to be accurately defined, quantified, and visualized.

To this end, we developed an in-house pipeline centered around our recently released software packages, MAJIQ and VOILA (freely available to the community). The pipeline allows us to detect *de novo* junctions, assess differentially spliced LSVs in and between groups of experiments, test the level of reproducibility of differentially spliced LSVs between similar experiments, and create a visual summary of how different pairs or groups agree or disagree on differentially spliced LSVs. We find that MAJIQ is more accurate than other methods for classical events. And it accurately defines and quantifies experimentally verified, complex splicing patterns (e.g. a 6-way LSV in *Eif4g3*) some of which result in novel, tissue-specific isoforms.

CONCLUSIONS

The importance of LSVs formulation is manifested in how common complex LSVs are in diverse metazoans, ranging from lizard to human. Complex LSVs are also enriched for regulated splicing when analyzing over thirty datasets across different tissues, developmental stages, splice factor knockdowns and neurodegenerative disease. In addition, LSV formulation can be used to investigate substructures of the transcriptome.

Immediate applications of the novel LSV framework and the MAJIQ software cover a wide spectrum. Examples include improved disease studies where transcriptome variations play a role, enhancing predictive models for splicing and for the effect of genetic variants, studying the regulatory underpinning of complex LSVs, and examining their evolutionary history. At the most basic level, our results illustrate the potential for novel discoveries in reanalyzing previously published data with the new LSV based methods

REFERENCES

1. Cooper, T. A., Wan, L. & Dreyfuss, G. RNA and Disease. *Cell* **136,** 777–793 (2009).
2. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* **40,** 1413–1415 (2008).
3. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456,** 470–476 (2008).

Alternative Splicing of CD19 in relapsed B-cell acute lymphoblastic leukemia

Elena Sotillo[1], Kathryn L. Black[1], Asen Bagashev[1], Ammar Naqvi[1], Deanne Taylor[2], Matthew R. Gazzara[3,4], Nicole M. Martinez[4], Yoseph Barash[3], Kristen W. Lynch[4], Andrei Thomas-Tikhonenko[1,5 *]

[1]Division of Cancer Pathobiology and [2]Department of Biomedical and Health Informatics,
The Children's Hospital of Philadelphia;
Departments of [3]Genetics, [4]Biochemistry & Biophysics, and [5]Pathology & Laboratory Medicine, Perelman School of Medicine at the University of Pennsylvania,
Philadelphia, Pennsylvania.
*To whom correspondence should be addressed: andreit@mail.med.upenn.edu

---

BACKGROUND

The CD19 antigen expressed on most B-cell acute lymphoblastic leukemias (B-ALL) can be targeted with CD19-directed immunotherapeutics, but relapses with epitope loss occur in 10% to 20% of pediatric responders. However, many of them still express an N-terminally truncated CD19 variant, which fails to trigger killing by CART-19 but partly rescues defects associated with CD19 loss [1].

RESULTS

To understand the nature of this variant, we analyzed the RNA-seq data for alternative splicing using the recently developed MAJIQ algorithm [2]. We discovered and validated alternatively spliced CD19 mRNA species in relapsed B-ALL, including the one lacking exon 2 which encodes the N-terminal epitope. Using protein/RNA pull-down assays, we detected specific binding of the SRSF3 splicing factor to CD19 exon 2. Furthermore, siRNA knockdown experiments identified SRSF3 as a splicing factor involved in exon 2 retention. Finally, using genome editing, we demonstrated that forced skipping of exon 2 results in expression of the N-terminally truncated CD19 isoform.

CONCLUSIONS

SRSF3 insufficiency could lead to alternative splicing of CD19 and confer resistance to immunotherapy. Based on these results, we are currently developing a new transcriptome-wide map of differential splicing in B-ALL and investigating additional regulatory elements.

REFERENCES

1. Sotillo et al, Convergence of Acquired Mutations and Alternative Splicing of CD19 Enables Resistance to CART-19 Immunotherapy. Cancer Discovery, 2015, 5(12):1282-95
2. Vaquero-Garcia et al, A new view of transcriptome complexity and regulation through the lens of local splicing variations. Elife, 2016, 5:e11752

Multiphasic and dynamic changes in alternative splicing during induction of pluripotency are coordinated by numerous RNA binding proteins

Benjamin Cieply[1*], Juw Won Park[2,3,4*], Yi Xing[2], and Russ P. Carstens[1]

[1]Department of Medicine and Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, 19104, USA; [2]Department of Microbiology, Immunology, & Molecular Genetics, University of California Los Angeles, Los Angeles, California, 90095, USA; [3]Department of Computer Engineering and Computer Science, [4]KBRIN Bioinformatics Core, University of Louisville, Louisville, Kentucky, 40292, USA.

Correspondence: russcars@.upenn.edu and yxing@ucla.edu

BACKGROUND

Alternative splicing (AS) plays a critical role in development, disease, cell fate transitions and pluripotency including somatic cell reprogramming. Recent studies have characterized the contrasting AS patterns between somatic and pluripotent stem cells on a genome wide scale but a detailed temporal analysis of induced pluripotency was lacking. [1] [2] We profiled transcriptome-wide AS changes that occur during the stepwise reprogramming of fibroblasts to pluripotency.

RESULTS

This analysis revealed distinct phases of AS during reprogramming, ranging from very early to a splicing program that is unique to transgene-independent iPS cells. In accordance with this, clusters of RNA binding proteins (RBPs) also demonstrated temporally regulated expression patterns. Changes in the expression of alternative splicing factors Zcchc24, Mbnl1/2 and Rbm47 were demonstrated to contribute to phase-specific AS. RNA binding motif enrichment analysis near alternatively spliced exons provided further insight into the combinatorial regulation of AS during reprogramming by different RNA binding proteins including. Celf's, Rbfox2, and Rbm24. Alternative splicing factor Esrp1 was shown to be a regulator of AS at the mesenchymal to epithelial transition phase. Ectopic expression of Esrp1 enhanced reprogramming, in part by modulating the AS of the epithelial specific transcription factor Grhl1.[3]

CONCLUSIONS

These data represent a comprehensive temporal analysis of the dynamic regulation of AS during the acquisition of pluripotency. While we implicate several RBPs, including Esrp1, as contributing to these AS patterns, it is also clear that additional factors are critical and this data can be used as a resource to inform future studies into novel roles for RBPs and alternative splicing in pluripotency and cell fate determination.

REFERENCES

1. Han, H., et al., *MBNL proteins repress ES-cell-specific alternative splicing and reprogramming.* Nature 2013. **498**(7453): p. 241-5.
2. Ohta, S., et al., *Global splicing pattern reversion during somatic cell reprogramming.* Cell Rep, 2013. **5**(2): p. 357-66.
3. Cieply, B., et al., *Multiphasic and Dynamic Changes in Alternative Splicing during Induction of Pluripotency Are Coordinated by Numerous RNA-Binding Proteins.* Cell Rep, 2016. **15**(2): p. 247-55.

From genome biology to drug discovery: Integrative mining of human transcriptomics data for novel therapeutic targets

Simon Xi[1*]

[1]Neuroscience Bioinformatics Lead, Computational Sciences, Pfizer, Cambridge, MA

*To whom correspondence should be addressed: hualin.xi@pfizer.com

---

Recent advances of genome technologies have revolutionized the ways to study gene expression and transcriptional regulations across human tissues and cell types. It also created tremendous opportunities for uncovering novel therapeutic targets for human diseases. In this talk, I will describe how the transcriptomics platform has been impacting in our drug discovery processes. I will use GTEx project as an example to illustrate the computational approaches we have taken to mine large-scale human transcriptome datasets for tissue selective gene expression and splicing signatures, and further integrate these wealth of genomics findings with human genetics and brain imaging data to generate novel therapeutic hypotheses for CNS related diseases.

Nano-exon discovery

Yifei Shi[1] and **Stephen M. Mount[1*]**

[1] Dept. of Cell Biology and Molecular Genetics and Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland 20742

*To whom correspondence should be addressed: smount@umd.edu

BACKGROUND

Exons can be arbitrarily small. In fact, an exon of 1 nt. has been observed in *Arabidopsis thaliana* [1] and sites of recursive splicing are accurately described as exons with length zero [2]. Because nano-exons (exons with length less than 4 nt.) permit gapped alignment of cDNAs and sequencing reads to the genome, or to transcripts lacking them, nano-exons often go undetected and the true number of nano-exons remains unknown.

RESULTS

We have examined a single, high-quality, Illumina HiSeq RNA-seq data set in depth using a bioinformatics pipeline to discover and quantify nano-exons. These data are derived from Drosophila hemocytes, both wild-type and expressing an A2bp1 RNAi transgene. A2bp1 is a member of the highly conserved Fox-1 family of RNA-binding proteins. Human homologs RBFOX1, RBFOX2 and RBFOX3 have been linked to brain development, cardiac function, and autism spectrum disorders, and have been implicated in the regulation of microexons less than 51 nt. [3].

Reads were mapped to the reference transcriptome allowing no mismatches and a single indel error. We determined the background indel error rate for this data (which incorporates errors from RNA polymerase, reverse transcriptase and sequencing). We find that the rate is low (0.00003) and indels are limited to sites with short homopolymer runs. Indels mapping near splice sites were examined for compatibility with splice site consensus. Our analysis has revealed new potential nano-exons in Drosophila, including eight previously undescribed nano-exons with strong support. We also found three likely cases of recursive splicing with deletion that can be considered to be exons of size -1, meaning that sequences within an intron (including the sequence AGT), result in the removal of one nucleotide due to recursive splicing.

CONCLUSIONS

Nano-exons remain to be discovered, even in well-annotated species.
Exons of size -1 are not only possible, but have been observed.
Ongoing research explores the occurrence of nano-exons in other Drosophila data sets and in other species.

REFERENCES

1. Guo L. and Liu, C.M. A single-nucleotide exon found in Arabidopsis. Scientific Reports **5:** 180887 (2015).
2. Grellscheid, S,N. and Smith, C.W.. An apparent pseudo-exon acts both as an alternative exon that leads to nonsense-mediated decay and as a zero-length exon. Mol. Cell. Biol. **26:** 2237-46 (2006).
3. Li, Y.I., Sanchez-Pulido, L., Haerty, W. and Ponting, C.P.. RBFOX1 and PTBP1 proteins regulate the alternative splicing of micro-exons in human brain transcripts. Genome Research **25:** 1-13 (2015).

Precision in RNA secondary structure prediction: A sensitivity analysis of the folding nearest neighbor parameters

Jeffrey Zuber[1], Hongying Sun[1], Xiaoju Zhang[1], Iain J. McFayden[2], and David H. Mathews[1*]

[1] University of Rochester Medical Center, Department of Biochemistry and Biophysics, Center for RNA Biology,
[2] Moderna Therapeutics, Cambridge, Massachusetts

*To whom correspondence should be addressed: David_Mathews@urmc.rochester.edu

BACKGROUND

RNA nearest neighbor parameters are a set of parameters for estimating the folding energy changes of RNA secondary structures. These parameters approximate the folding free energy of a secondary structure as the sum of the free energies of neighboring structural motifs. These parameters are derived using linear regression on a database of folding stabilities determined by optical melting data of small model RNA structures. These parameters are used widely in numerous RNA analysis algorithms. Despite their widespread application, a comprehensive review of the impact of each parameter on the precision of calculations had not been conducted.

RESULTS

To identify the parameters with greatest impact, we performed a sensitivity analysis on the 290 independent parameters that compose the 2004 version of the nearest neighbor rules. Each parameter was modulated by increments of either experimental uncertainty or a fixed value and the effect of this parameter change on predicted base-pair probabilities and secondary structures was observed for an archive of sequences that included rRNA, tRNA, mRNA, randomized sequences, and other families of RNA (Figure 1). This identified a class of parameters and energetic models of specific secondary structure motifs that should be updated in order to improve the accuracy of RNA structure prediction. In particular, the bulge loop initiations, multibranch loop parameters, and AU/GU closure terms stood out as particularly important. An analysis of parameter usage during folding free energy calculations of stochastic samples of predicted RNA secondary structures revealed a correlation between parameter usage and impact on structure prediction.

CONCLUSIONS

This work developed several new insights into the prediction of RNA secondary structure. First, there are nearest neighbor parameters for which errors in the estimates have little impact on the precision of base pairing probability estimates. Second, there are parameters that are crucial for high quality base pair estimates, and those parameters that could be addressed with experiments to improve the accuracy of the nearest neighbor parameters. The results of this analysis can be also used to inform which parameters are the most important to determine with high precision for the expansion of nearest neighbor parameters to nucleotides with modified chemistries.

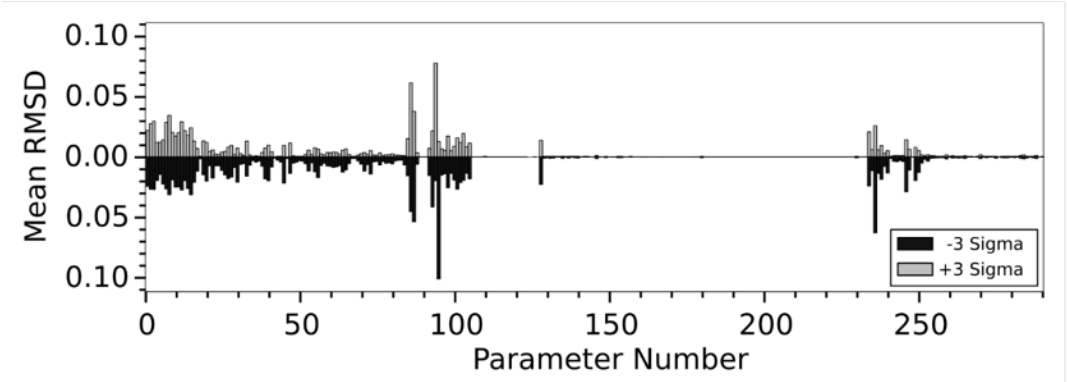


Figure 1: Sensitivity Analysis Using Constant Errors. In each panel, independent parameters are along the x-axis. Mean Pair Probability RMSD for the entire sequence archive except randomized sequences was calculated for each parameter for ±1.5 kcal/mol parameter deviations. The RMSDs for +1.5 kcal/mol are shown above the x-axis (grey), while the RMSDs for -1.5 kcal/mol are shown below the x-axis (black).

Multiple roles of local RNA folding along the coding region in gene expression regulation

The Laboratory of Computational Systems and Synthetic Biology, Biomedical Engineering, Tel Aviv University
*To whom correspondence should be addressed: tamirtul@post.tau.ac.il

---

BACKGROUND

The coding region codon composition affects not only the amino acid content of the encoded protein, but also various additional phenomena related to gene expression and protein functionality. Among others, it was shown that the folding of the mRNA/pre-mRNA in different regions along the coding region can modulate/affect translation initiation and elongation, splicing efficiency, protein folding, and more.

RESULTS

In recent years we performed a multidisciplinary research (see, for example, [1-6] and unpublished results), based on tools from computational biology, synthetic biology, and molecular evolution, to decipher the way the folding of the mRNA/pre-mRNA regulates different gene expression steps.
Specifically, among others, we quantified the effect of mRNA folding near the beginning of the coding region and downstream of it on translation initiation and elongation; we also quantified the related selection on transcripts via large scale analyses of endogenous genes and heterologous libraries [3,4,6].
In addition, via the analysis of endogenous genes and heterologous libraries we demonstrated, for the first time, that the pre-mRNA folding near intronic splice sites undergoes selection for weak folding to improve splicing efficiency in fungi [2,5].
Furthermore, we demonstrate that there is widespread signatures of local mRNA folding structure selection in various viruses (e.g. four Dengue virus serotypes, HIV, Ebola, and more); we experimentally validated some of these signals (yet unpublished results).
Finally, we demonstrate how local pre-mRNA local folding can be integrated to improve biophysical predictive models related to mRNA translation and protein levels [4,7].

CONCLUSIONS

Our results emphasize the importance of the local mRNA folding in different parts of the coding region in organisms and viruses across the tree of life. The results should contribute towards developing improved approaches for gene expression, understanding, modeling and engineering. In addition we report a detailed map of hot-spots along the coding region that undergo selection for strong/weak pre-mRNA/mRNA folding at a resolution of single nucleotides, providing an important layer of understanding of transcript evolution.

REFERENCES

1. E. Goz, T. Tuller. Evidence of a direct evolutionary selection for strong folding and mutational robustness within HIV coding regions. To appear in Journal of Computational Biology. 2016.
2. Goz E, Tuller T. Widespread signatures of local mRNA folding structure selection in four Dengue virus serotypes.. BMC Genomics. 2015;16 Suppl 10:S4.
3. Zafrir Z, Tuller T. Nucleotide sequence composition adjacent to intronic splice sites improves splicing efficiency via its effect on pre-mRNA local folding in fungi. RNA. 2015 Oct;21(10):1704-18.
4. Ben-Yehezkel T, Atar S, Zur H, Diament A, Goz E, Marx T, Cohen R, Dana A, Feldman A, Shapiro E, Tuller T. Rationally designed, heterologous S. cerevisiae transcripts expose novel expression determinants. RNA Biol. 2015;12(9):972-84.
5. Tuller T, Zur H. Multiple roles of the coding sequence 5' end in gene expression regulation. Nucleic Acids Res. 2015 Jan;43(1):13-28.
6. Yofe I, Zafrir Z, Blau R, Schuldiner M, Tuller T, Shapiro E, Ben-Yehezkel T. Accurate, model-based tuning of synthetic gene expression using introns in S. cerevisiae. PLoS Genet. 2014 Jun 26;10(6):e1004407.
7. Tuller T, Veksler-Lublinsky I, Gazit N, Kupiec M, Ruppin E, Ziv-Ukelson M. Composite effects of gene determinants on the translation speed and density of ribosomes. Genome Biol. 2011 Nov 3;12(11):R110

Identifying conserved RNA structures in alphaviruses: An integrated approach

Katrina M. Kutchko[1,2*], Kenneth Plante[3], Clayton Morrison[3], Emily Madden[4], Wes Sanders[4,5], Alain Laederach[1,2], Mark T. Heise[3], and Nathaniel J. Moorman[4,5]

[1] Curriculum in Bioinformatics and Computational Biology, [2] Department of Biology, [3] Department of Genetics, [4] Department of Microbiology and Immunology, [5] Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill

*To whom correspondence should be addressed: kutchko@email.unc.edu

---

BACKGROUND

Alphaviruses, such as chikungunya virus (CHIKV), use a single-stranded positive-sense RNA genome to infect both arthropod vectors and vertebrate hosts. In addition to encoding the viral proteins, RNA virus genomes contain important RNA structural elements that control virus genome transcription and protein synthesis, as well as allow the virus to avoid antiviral control by the host immune system. Very few of these elements, however, have been identified in alphaviruses. To find conserved and functional RNA structures in alphavirus genomes, we present a new method integrating experimental RNA structural probing via SHAPE-MaP [1], secondary structure prediction by free energy minimization, and covariation models for evolutionary support.

Using experimentally-informed structures, we apply a stochastic context free grammar to identify divergent but conserved structures across alphavirus genomes in CHIKV, sindbis virus (SINV), and Venezuelan equine encephalitis virus (VEEV). Comparative structural metrics quantify covariation in these candidate structures. From these metrics, we select the most conserved structures to test experimentally by introducing mutations within each structured genomic region that preserve the protein sequence but disrupt the RNA structure [2].

RESULTS

Our method identify 49 potential structured candidate regions across CHIKV, SINV, and VEEV. We verify our method by confirming that the 5′ end of the CHIKV genome and the packaging signal in SINV, both known functional elements in alphaviruses, are candidate structured regions and disruption of these structures reduces viral yields in our *in vitro* tests. We also find novel functional structures in CHIKV and SINV, which we are currently testing to determine their biological regulatory mechanism.

CONCLUSIONS

Our new, alternative approach to finding conserved RNA structures combines experimental structure data with multiple bioinformatic approaches to identify and validate functional structural elements in alphavirus genomes. Using RNA probing data, structure prediction, and phylogenetic analysis together presents a novel method for generating and applying experimentally-informed models to the discovery of functional viral RNA structures. This approach successfully finds important structures in these viral genomes. The successful identification of functional structures gives us further insight into the biology of alphaviruses and provides a foundation for future vaccine development.

REFERENCES

1. Siegfried N.A., Busan S., Rice G.M., Nelson J.A.E., and Weeks K.M. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). Nature Methods **11** (2014).
2. Jorge, D.M.d.M., Mills, R.E., and Lauring, A.S. CodonShuffle: a tool for generating and analyzing synonymously mutated sequences. Virus Evolution **1** (2015).

Getting the entire message: Coordination of variable sites on RNA transcripts

Hagen Tilgner*

Brain and Mind Institute, Weill Cornell Medical College, New York City

---

Alternative splice-site-, RNA-edit-, TSS- and polyA-site-usage introduce multiple variable sites into RNAs for many genes. Randomly combining variable sites allows for extreme isoform complexity – yet the question of whether variables sites are randomly combined remained underexplored due to the lack of deep long-read sequencing. Based on experiences with a variety of technologies1-4, I will describe our discovery and mapping of (i) new isoforms and (ii) non-randomly paired RNA-processing events. We generally observe that ~15-20% of the RNA molecules in multiple cell types and tissues represent previously unknown isoforms. To address (ii), we have introduced4 a facile way of RNA-sequencing, synthetic long-read-sequencing (SLR-RNA-Seq), in which long RNA-sequences are deduced from pools of cDNAs. This technology allows the full-length sequencing of millions of high quality, on average ~2kb long molecules. Using a variety of published and unpublished datasets, we can find hundreds of pairs of distant (that is separated by constitutive exons) alternative exon pairs that are paired in non-random ways to form complete RNA isoforms – although a huge number of pairs appear randomly combined. I will further discuss so far unpublished types of non-randomly paired sites. Under the assumption that these isoforms are equally translated these findings give an unprecedented view into the full-length proteome.

REFERENCES

1. Tilgner et al. G3. 2013
2. Sharon et al. Nat Biotechnol. 2013
3. Tilgner et al. PNAS 2014
4. Tilgner et al. Nat Biotechnol. 2015

Why asking for miRNA-gene interactions is wrong: A new paradigm for miRNA target prediction with expression data

Azim Dehghani Amirabad[1,2,3], Marcel H. Schulz[1,2,*]

[1] Cluster of Excellence for Multimodal Computing and Interaction, Saarland University, Saarbrücken

[2] Max Planck Institute for Informatics, Saarbrücken

[5] International Max Planck Research School for Computer Science, Saarbrücken

*To whom correspondence should be addressed: mschulz@mmci.uni-saarland.de

---

BACKGROUND

Deregulation of miRNAs is implicated in many diseases in particular cancer, where miRNAs can act as tumor suppressors or oncogenes. As sequence-based miRNA target predictions do not provide context-specific information, many algorithms combine expression data for miRNAs and genes for prioritization of miRNA targets. However, common strategies prioritize miRNA-gene associations, although a miRNA may only target a subset of the alternative transcripts produced by a gene and thus may have a differential influence on different proteins generated by the gene. Therefore, current approaches are suboptimal, e.g., for the analysis of miRNAs and their interplay with protein drug targets or the composition of protein complexes. Here we address for the first time the problem of transcript and not gene based miRNA target prioritization.

RESULTS

We introduce a new paradigm for prioritizing miRNA target associations that works on the level of individual transcripts instead of the (virtual) gene. Our approach makes full use of available transcript level expression data and annotation to estimate miRNA interaction strength by formulating a multi-task problem that allows to borrow information from all transcripts of the same gene that share the same miRNA binding site.

Performance evaluation on both synthetic and real RNA-seq cancer data shows that we are able to accurately predict miRNA targets at transcript level and with better performance compared to established methods that work on the gene level. We show that noisy transcript expression estimates using RNA-seq can be accounted for using our formulation. We show that miRNAs are able to explain around 20% of the observed variability in transcript expression levels.

CONCLUSIONS

Our new approach opens the door for prioritizing miRNA-targets in transcripts for available expression datasets, that previously only used aggregated gene expression measurements. Ultimately, predicting at the transcript level is not only likely to improve performance but also gives a direct link to the affected protein and thus enables new studies for example estimating miRNA effects on drug targets.

Integrative, genome-wide characterization of the human disease landscape

Olga Troyanskaya*
Department of Computer Science and Lewis-Sigler Institute of Integrative Genomics, Princeton University, USA
*To whom correspondence should be addressed: ogt@genomics.princeton.edu

---

Complex diseases are driven by multiple genetic changes and characterized by genome-wide perturbations of cellular pathways and functions. Gene expression profiling experiments have been potent in shedding light on the molecular pathology of diseases. Most studies typically focus on a single disease and contrast disease samples to their normal controls. However, such "one disease at a time" approaches disregard similarities and differences in pathological deregulations underlying different complex diseases and are thus unable to identify attributes unique to each particular disease, which is critical for developing targeted therapy. We have developed a unified probabilistic framework URSAHD to identify and quantify distinctive disease signals based on gene expression profiles of clinical samples. Our framework can be used to distinguish between closely-related diseases, identify discerning genes and processes, associate rare-diseases to the nearest well-studied disease, and track the effectiveness of therapy. No curated set of genes were used in our data-driven approach, and so it can easily be extended to any human disease for which high-throughput expression data can be generated. We found that the most predictive genes identified by our method are significantly under-studied in the biomedical literature, demonstrating that many key biological processes underlying human pathophysiology are in fact in critical need of further investigation.

CFIm25 promotes human MSC osteogenesis by regulating 3' processing of mRNAs encoding BMP signaling pathways

Chengguo Yao[1]*, Chunliu Huang[1], Peng Xiang[1]

[1]Center for Stem Cell Biology and Tissue Engineering, Zhongshan School of Medicine,

Sun Yat-sen University, Guangzhou 510080, China.

*To whom correspondence should be addressed: yaochguo@mail.sysu.edu.cn

---

BACKGROUND

CFIm25 is emerging as a master regulator of mRNA 3' UTR length in human cells and has recently been linked to neurite outgrowth and glioblastoma tumorigenicity [1-2]. However, most of the reports used immortal cell line and the functional significance of CFIm25 in human physiological development remains largely unknown.
Human bone marrow-derived MSC could be expanded and induced into osteoblasts in vitro, most of previous mechanism studies focused on transcriptional regulatory network during this process [3], whether post-transcriptional regulation especially 3' processing contribute to this process remains elusive.

RESULTS

We investigated the role of CFIm25 during human mesenchymal stem cells (hMSC) osteogenesis. Our results showed that overexpression of CFIm25 promotes human MSC osteogenesis and the expression of CFIm25 is upregulated during this process. Further PolyA-site and RNA-IP sequencing revealed that CFIm 25 preferentially targets mRNAs encoding BMP signaling pathway proteins, which provides mechanism insight into CFIm 25-regulating hMSC osteogenesis.

CONCLUSIONS

1.CFIm 25 promotes hMSC osteogenesis.
2.CFIm 25 is upregulated during hMSC osteogenesis.
3.CFIm 25 preferentially targets mRNAs encoding BMP signaling pathway proteins in hMSC.

REFERENCES

1. Masamha CP, Xia Z, Yang J, Albrecht TR, Li M, Shyu AB, Li W, Wagner EJ. CFIm25 links alternative polyadenylation to glioblastoma tumour suppression.Nature. 2014,510(7505):412-6
2. Fukumitsu H, Soumiya H, Furukawa S.Knockdown of pre-mRNA cleavage factor Im 25 kDa promotes neurite outgrowth.Biochem Biophys Res Commun. 2012,425(4):848-53
3. Lian JB, Stein GS, Javed A, van Wijnen AJ, Stein JL, Montecino M, Hassan MQ, Gaur T, Lengner CJ, Young DW.Networks and hubs for the transcriptional control of osteoblastogenesis.Rev Endocr Metab Disord. 2006 7(1-2)

The contribution of Alu exons to the human proteome

Lan Lin[1], Peng Jiang[2], Juw Won Park[1], Jinkai Wang[1], Zhi-xiang Lu[1], Maggie PY Lam[3], Peipei Ping[3,4], Yi Xing[1,*]

[1]Department of Microbiology, Immunology & Molecular Genetics, University of California, Los Angeles, Los Angeles, CA 90095, USA, [2]Regenerative Biology, Morgridge Institute for Research, Madison, WI 53707, USA, Departments of [3]Physiology, [4]Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA

*To whom correspondence should be addressed: yxing@ucla.edu

BACKGROUND

Alu elements are major contributors to lineage-specific new exons in primate and human genomes. Recent studies indicate that some Alu exons have high transcript inclusion levels or tissue-specific splicing profiles, and may play important regulatory roles in modulating mRNA degradation or translational efficiency. However, the contribution of Alu exons to the human proteome remains unclear and controversial. The prevailing view is that exons derived from young repetitive elements, such as Alu elements, are restricted to regulatory functions and have not had adequate evolutionary time to be incorporated into stable, functional proteins.

RESULTS

We adopt a proteotranscriptomics approach to systematically assess the contribution of Alu exons to the human proteome. Using RNA sequencing, ribosome profiling and proteomics data from human tissues and cell lines, we provide evidence for the translational activities of Alu exons and the presence of Alu exon derived peptides in human proteins. These Alu exon peptides represent species-specific protein differences between primates and other mammals, and in certain instances between humans and closely related primates. In the case of the RNA editing enzyme ADARB1, which contains an Alu exon peptide in its catalytic domain, RNA sequencing analyses of A-to-I editing demonstrate that both the Alu exon skipping and inclusion isoforms encode active enzymes. The Alu exon derived peptide may fine tune the overall editing activity and, in limited cases, the site selectivity of ADARB1 protein products.

CONCLUSIONS

Our data indicate that Alu elements have contributed to the acquisition of novel protein sequences during primate and human evolution.

REFERENCES

1. Lin L., Jiang P., Park JW., Wang J., Lu ZX., Lam MP., Ping P., Xing Y. (2016) The contribution of Alu exons to the human proteome. Genome Biology, 17:15.

Robust quantification of local splice variations for large heterogeneous datasets

Scott Norton[12*], Jordi Vaquero-Garcia[2], and Yoseph Barash[2]

[1] Genomics and Computational Biology Graduate Group, Biomedical Graduate Studies, Perelman School of Medicine, University of Pennsylvania, Philadelphia PA 19104 USA, [2] BioCiphers Laboratory Group, Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia PA 19104 USA

*To whom correspondence should be addressed: scnorton@upenn.edu

---

BACKGROUND

Over 90% of the human exome is alternatively spliced. To fully understand the complexity of splicing regulation, one needs to quantify relative splice form abundance within and between samples. Traditional methods for quantifying splicing are limited by defining only simple, binary events, which correspond to approximately 70% of observed alternative splicing events [1]. The remaining 30% are complex events consisting of more than two junctions, and are often discarded in the literature.

The recently-published model of alternative junction inclusion quantification (MAJIQ) algorithm addresses the valuable information contained within these complex splicing events by redefining the question of alternative splicing quantification in terms of local splice variations (LSVs). The current iteration of MAJIQ assumes that a group of biological or technical RNA-seq replicates from the same test condition generally agrees on the underlying value of PSI (percent splice index) for each detected LSV. When that assumption does not hold, MAJIQ's estimates of PSI might be skewed. Given the prevalence of large heterogeneous datasets (e.g. patients vs. controls in disease studies), it is therefore important to address both efficiency and heterogeneity in estimating PSI and how it changes (delta PSI) between conditions or experimental groups.

RESULTS

We are developing MAJIQ-het, a generalization of the MAJIQ model which handles within-group heterogeneity to give robust estimates of PSI and delta PSI. Briefly, MAJIQ-het assigns weights to each experiment representing a posterior belief in the relevance or group membership.

Consequently, MAJIQ-het converges to the MAJIQ model on well-behaved datasets with no detectable loss of power, yet is still capable of detecting and down-weighting outlier samples. We consider two alternative weighing schemes, termed the inside and outside models, which correspond to different probabilistic assumptions about the underlying distributions. We compare MAJIQ to both the inside and outside MAJIQ-het on synthetic and real-life data, demonstrating significant gain in both reproducibility and sensitivity for detecting differentially-spliced LSVs.

CONCLUSIONS

Based on our preliminary analysis, the MAJIQ-het prototype offers a promising approach for handling high-variance RNA-seq datasets. We expect it to be highly relevant for large cohorts to assess alternative splicing in heterogeneous populations.

REFERENCES

1. Jorge Vaquero-Garcia, Alejandro Barrera, Matthew R Gazzara, Juan González-Vallinas, Nicholas F Lahens, John B Hogenesch, Kristen W Lynch, Yoseph Barash. A new view of transcriptome complexity and regulation through the lens of local splicing variations. eLife 2016; 5e11752
2. Thomas M. Keane, Leo Goodstadt, Petr Danecek, Michael A. White, Kim Wong, Binnaz Yalcin, Andreas Heger, Avigail Agam, Guy Slater, Martin Goodson, Nicholas A. Furlotte, Eleazar Eskin, Christoffer Nellåker, Helen Whitley, James Cleak, Deborah Janowitz, Polinka Hernandez-Pliego, Andrew Edwards, T. Grant Belgard, Peter L. Oliver, Rebecca E. McIntyre, Amarjit Bhomra, Jérôme Nicod, Xiangchao Gan, Wei Yuan, Louise van der Weyden, Charles A. Steward, Sendu Bala, Jim Stalker, Richard Mott, Richard Durbin, Ian J. Jackson, Anne Czechanski, José Afonso Guerra-Assunção, Leah Rae Donahue, Laura G. Reinholdt, Bret A. Payseur, Chris P. Ponting, Ewan Birney, Jonathan Flint, David J. Adams. Mouse genomic variation and its effect on phenotypes and gene regulation. Nature 2011(477): 289-294

Thousands of targets of nonsense-mediated mRNA decay revealed by transcriptome analysis offer clues to the mechanism in multiple species

Courtney E. French[1], Anna Desai[2], Gang Wei[3], James P. B. Lloyd[3,4], Thomas L. Gallagher[5], Darwin S. Dichmann[1], Maki Inada[6], Sharon L. Amacher[5], Richard M. Harland[1], and **Steven. E. Brenner[1,3*]**

[1] Department of Molecular and Cell Biology, [2] Department of Comparative Biochemistry, [3] Department of Plant and Microbial Biology, [4] Center for RNA Systems Biology, University of California, Berkeley, CA. [5] Department of Molecular Genetics, Ohio State University, Columbus, OH. [6] Department of Biology, Ithaca College, Ithaca, NY.

*To whom correspondence should be addressed: 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 necessarily in other eukaryotes. There is also evidence that a longer 3' UTR triggers NMD in yeast, plants, flies, and mammals [5,6,7,8,9].

RESULTS

To survey the targets of NMD genome-wide in a number of diverse eukaryotic species, we have analyzed RNA-Seq data from cells where NMD has been inhibited via knockdown or knockout of UPF1, a critical protein in the degradation pathway. We found that hundreds to thousands of genes produce alternative isoforms that are potentially degraded by NMD in each of the species tested (human, mouse, zebrafish, frog, fly, S. pombe, and Arabidopsis). This includes 20-40% of the genes alternatively spliced in a given species. These genes, potentially subject to regulation through NMD, are involved in many functional categories and, in at least human and fly, are significantly enriched for RNA splice factors.

We also gained 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 the other species tested, with the exclusion of S. pombe. In contrast, we found little to no correlation between the likelihood of degradation by NMD and 3' UTR length in any of the species, independent of the 50nt Rule.

CONCLUSIONS

Overall, we found that the mechanism of regulation via alternative splicing coupled with NMD is prevalent in each species tested including animals, plants, and fungi.

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. Kerenyi, Z., et al. EMBO Journal. **27** (2008)
6. Hansen, K., et al. PLoS Genetics. **5** (2009)
7. Hogg, J and Goff, S. Cell. **143** (2010)
8. Yepiskoposyan, H., et al. RNA. **17** (2011)
9. Hurt, J.A., et al. Genome Research. **23** (2013)

PGRN network-wide project: Transcriptome analysis of pharmacogenes in human tissues

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

[1] University of California, Berkeley, CA [2] University of California, San Francisco, CA [3] Vanderbilt University, Nashville, TN [4] Baylor College of Medicine, Houston, TX [5] Children's Hospital Oakland Research Institute, Oakland, CA [6] Ohio State University, Columbus, OH [7] Brigham and Women's Hospital, Boston, MA [8] St. Jude Children's Research Hospital, Memphis, TN [9] Northwestern University Feinberg School of Medicine, Chicago, IL

*To whom correspondence should be addressed: 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 annotated 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 (≥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 demonstrate that important pharmacogenes are variably expressed across tissues of pharmacologic relevance, and across different individuals, and that the vast majority is alternatively spliced.

Network of splice factor regulation by alternative splicing coupled with nonsense mediated mRNA decay

Anna Desai[1], James P. B. Lloyd[1], Courtney E. French[1], and **Steven. E. Brenner[1*]**

[1] 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.

Dissecting the DNA and RNA bound proteome of human embryonic stem cells

Shlomi Dvir [1*] and Yael Mandel-Gutfreund [2]

[1,2] Faculty of Biology, Technion-Israel Institute of Technology, Haifa 32000, Israel

*To whom correspondence should be addressed: shlomidv@campus.technion.ac.il

---

BACKGROUND

Accumulating evidence supports the existence of dual-function DNA- and RNA- binding proteins (DRBPs) that coordinate multiple steps of gene expression programs, adding an additional level of complexity to gene regulation. Despite extensive research on DNA-binding proteins (DBPs) and RNA-binding proteins (RBPs), little is known about the identity and function of the dual binders.

RESULTS

Using RNA-sequencing and mass-spectrometry, we generated a global map of gene expression profile in undifferentiated and differentiated cells and found that the vast majority of a literature-curated set of DRBPs are expressed in human embryonic stem cells (hESCs). Notably, DRBPs have higher mRNA and protein abundance compared with non-DNA and non-RNA binding genes and, surprisingly, also when contrasted with DBPs and RBPs. Differential expression analysis further shows that a large fraction of DRBPs are significantly up-regulated in hESCs, suggesting that DRBPs may shape the stem cell state. As a step towards the experimental identification of DRBPs, we performed RNA interactome [1-3] to capture the poly(A)-RNA bound proteome of hESCs. We show that the hESC interactome is enriched for known RBPs, including a substantial number of potential DRBPs.

CONCLUSIONS

Here we present the first, to our knowledge, RNA interactome of human ESCs. We further propose a novel solution to systematically uncover the *in vivo* repertoire of DNA- and RNA- binding proteins of hESCs. While identifying the full catalog of DRBPs is technically challenging, it has the potential to lead to the detection of key factors that have been overlooked when studying DNA and RNA regulation separately, hence contributing to both fields of gene regulation and stem cell biology.

REFERENCES

1. Castello, A. *et al.* Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins. *Cell* **149,** 1393–1406 (2012).
2. Baltz, A. G. *et al.* The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol. Cell* **46,** 674–90 (2012).
3. Kwon, S. C. *et al.* The RNA-binding protein repertoire of embryonic stem cells. *Nat. Struct. Mol. Biol.* **20,** 1122–30 (2013).

Analysis of differential splicing suggests different modes of short-term splicing regulation

Hande Topa[1,2*] and Antti Honkela[2*]

[1] Department of Computer Science, Helsinki Institute for Information Technology HIIT, Aalto University, Espoo, 00076, Finland

[2] Department of Computer Science, Helsinki Institute for Information Technology HIIT, University of Helsinki, Helsinki, 00014, Finland

*To whom correspondence should be addressed: hande.topa@helsinki.fi, antti.honkela@helsinki.fi

---

BACKGROUND

Alternative splicing is an important mechanism in which the regions of pre-mRNAs are differentially joined in order to form different transcript isoforms. It is involved in the regulation of normal physiological functions but also linked to the development of diseases such as cancer. Here we use a Gaussian process (GP)-based method in order to analyse differential expression and splicing using RNA-seq time series in three different settings: overall gene expression levels, absolute transcript expression levels and relative transcript expression levels. As demonstrated in [1], incorporation of available variance information into GP models improves the performance of the GP-based ranking methods, where one would like to rank features according to their temporal activity levels. Using a similar approach and obtaining the expression level estimates by BitSeq [2], we evaluate the performance of our method with replicated and unreplicated experiment designs where the variances are obtained or used in different ways. Then we apply the best-performing method to real RNA-seq time series data from [3], which was obtained from MCF7 breast cancer cell lines at 10 different time points during estrogen receptor-alpha signaling response.

RESULTS

Evaluation of our method under different experiment designs highlights the importance of replication and suggests an L-shaped design in which the replication is done only at the first time point if full replication is not possible. L-shaped design leads to a better performance than the fully unreplicated design by enabling to model the mean-expression-dependent variances by using the replicates from the first time point and propagating them to the rest of the time series.

In the real data analysis, our GP-based method identifies genes with differential splicing and/or differentially expressed transcripts. For example, we find out genes with consistent changes in alternative splicing independent of changes in absolute expression and genes where some transcripts change while others stay constant in absolute level.

CONCLUSIONS

Application of our method to the analysis of splicing patterns during estrogen receptor-alpha signaling response in a human breast cancer cell line lead to the discovery of classes of genes with different modes of splicing and expression changes. These examples suggest a link between regulation of gene expression and splicing, but further research with careful controls is needed to assess how common this phenomenon is. The findings nevertheless suggest that alternative splicing analyses need to combine both absolute and relative transcript expression analyses.

REFERENCES

1. Topa, H., Jónás, A., Kofler, R., Kosiol, C. and Honkela, A. Gaussian process test for high-throughput sequencing time series: application to experimental evolution. Bioinformatics **31** (2015).
2. Glaus, P., Honkela, A. and Rattray, M. Identifying differentially expressed transcripts from RNA-seq data with biological variation. Bioinformatics **28**(13) (2012).
3. Honkela, A., Peltonen, J., Topa, H., Charapitsa, I., Matarese, F., Grote, K., Stunnenberg, H. G., Reid, G., Lawrence, N. D. and Rattray, M. Genome-wide modeling of transcription kinetics reveals patterns of RNA production delays. PNAS **112** (2015).

Differential expression method for highly related samples

Natalie Davidson[1*], Kjong-Van Lehmann[1], and Gunnar Rätsch[1]

[1] ETH Zürich

*To whom correspondence should be addressed: natalie.davidson@inf.ethz.ch

---

BACKGROUND

Large-scale efforts to measure genomic and transcriptomic patterns across several cancer types have helped to identify the genetic diversity of cancer [1,2]. The degree of intra-cancer variability greatly complicates the analyses of differences between cancer types. Identifying differentially-expressed genes between cancer types or normal samples is especially confounded since expression variability within a single cancer type is large. When the variance within a single condition is large, or the samples are highly correlated, the typical fixed-effects model can lead to an increase in genes falsely identified as differentially-expressed [3].

RESULTS

Our method identifies differentially-expressed genes between cancer types with high expression variability by using a mixed-effects model that incorporates relatedness between samples to account for variance within cancer types. Relatedness is calculated by the correlation of somatic and germline variants.

We validated on simulated and real data that exhibited high relatedness between samples, comparing our model against a baseline fixed-effects model. Our simulation shows, that in cases where samples are highly correlated and there are greater than 20 samples within a sample group, the mixed-effects model achieves a false positive rate of 0.012 and a false negative rate of 0.19. Comparatively, the fixed-effects model has a false positive rate of 0.02 and a false negative rate of 0.63. This clearly shows that a mixed--model approach is able to account for structured variability within cancer types and identifies less false positives and significantly less false negatives.

We applied this approach on TCGA samples of uterine carcinosarcoma and uterine corpus endometrial carcinoma. Our mixed-effects model identifies 5505 genes that are significantly different between cancers, while also accounting for relatedness. GO analysis revealed they are enriched for processes related to cell adhesion, a known difference between epithelial cancers and carcinosarcomas. Conversely, the top differentially-expressed genes from the fixed-effects model were not enriched for any cell adhesion related processes.

CONCLUSIONS

We validated our model on simulated data and successfully applied it to real data to identify differentially expressed genes between cancer types. We recommend accounting for sample relatedness in differential-expression analysis, especially in the context of cancer.

REFERENCES

1. Weinstein, John N., et al. "The cancer genome atlas pan-cancer analysis project." Nature genetics 45.10 (2013): 1113-1120.

2. Verhaak, Roel G.W. et al. “An Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR and NF1.” Cancer cell 17.1 (2010): 98. PMC.

3. Kang, Hyun Min et al. “Efficient Control of Population Structure in Model Organism Association Mapping.” Genetics 178.3 (2008): 1709–1723. PMC. Web. 29 Apr. 2016.

Exploring the relationship between intron retention and DNase I hypersensitivity in plants

Fahad Ullah[1], ASN Reddy[2], and Asa Ben-Hur[1*]

[1] Department of Computer Science, Colorado State University, Fort Collins CO

[2] Department of Biology and Program in Molecular Plant Biology, Colorado State University, Fort Collins CO

*To whom correspondence should be addressed: asa@cs.colostate.edu

BACKGROUND

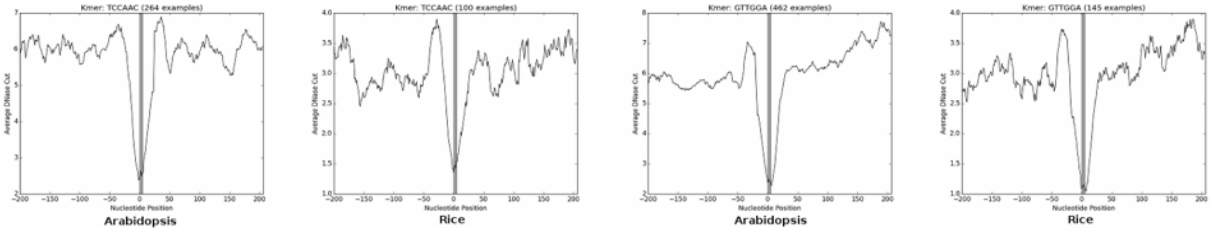
Intron retention (IR) is the most prevalent form of alternative splicing in plants [1]. IR, like other forms of alternative splicing has an important role in increasing gene product diversity and regulating transcript functionality [2]. The process of IR is not well understood. In particular, very few of the regulatory elements that control splice site choice in plants are known. Splicing is known to occur co-transcriptionally and is influenced by the speed of transcription which in turn, is inversely correlated with nucleosome occupancy [3]. It follows that chromatin structure may have an important role in the regulation of IR. Therefore, in this study we use genome-wide DNase I-seq data to investigate the relationship between IR and DNase I Hypersensitive Sites (DHSs): regions of chromatin with very high likelihood of cleavage by the DNase I enzyme [4]. DHSs represent highly open and accessible regions of chromatin where most of the active genes are found. Pertinent to that, *cis*-acting regulatory elements can be identified in those regions using protein footprinting [5] where a regulatory protein protects its binding site from the cleavage of the DNase I enzyme. We developed a continuous hidden Markov Model (HMM) to detect footprints associated with DHSs that occur around retained introns.

RESULTS

Our results show that IR events are highly enriched in DHSs compared to intron excision (IE) events in both *Arabidopsis* (p-value = $1.54e^{-47}$) and rice (p-value = $8.90e^{-67}$). This implies that the chromatin is more open in retained introns. Thus, the speed of transcription is higher in those regions of the genome, giving less time to the spliceosomal machinery to recognize and splice out those introns.

Next, to identify potential protein binding sites, we use an HMM to discover hexamers with footprints. Over 70 hexamers exhibit a "dip" in the average DNase I-seq coverage in both *Arabidopsis* and rice, indicating a possible footprint left by a binding protein. The figure below shows two such candidate protein footprints left at the two hexamers TCCAAC and GTTGGA.

Figure: footprints detected at the hexamers TCCAAC and GTTGGA in *Arabidopsis* and rice.



CONCLUSIONS

In this study we have established an association between IR and DNase I hypersensitive sites in *Arabidopsis* and rice. We have presented a mechanistic hypothesis that explains the observed association from the perspective of the co-transcriptional nature of splicing. Finally, using DNase I-seq data, we have identified several IR and IE hexamers that are likely to be footprints left by regulatory proteins that affect splicing.

REFERENCES

1. Reddy A.S.N., Rogers M.F., Richardson D.N., Hamilton M., Ben-Hur A. Deciphering the plant splicing code. Frontiers in Plant Science **3** (2012).
2. Reddy, A.S. Alternative splicing of pre-messenger RNAs in plants in the genomic era. Annual Review of Plant Biology **58** (2007).
3. Naftelberg S., Schor I.E., et al. Regulation of alterantive splicing through coupling with transcription and chromatin structure. Annual Review Biochemistry **84** (2015).
4. Boyle A.P., Davis S., Shulha H.P., et al. High-resolution mapping and characterization of open chromatin across the genome. Cell **132** (2008).
5. Boyle A.P., Song L., Lee B.K., et al . High resolution genome-wide in vivo footprinting of diverse transcription factors in human cells**.** Genome Research **24** (2010).

Predictive modeling framework for splice factor knockdown experiments

Anupama Jha[1], Matthew Gazzara[2,3], Jorge Vaquero-Garcia[1,2] and Yoseph Barash[1,2*]

[1] Department of Computer and Information Science, University of Pennsylvania, Philadelphia, United States; [2]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, United States; [3]Department of Biochemistry and Biophysics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, United States

*To whom correspondence should be addressed: yosephb@upenn.edu

---

BACKGROUND

Advancements in sequencing technologies have highlighted the role of alternative splicing in increasing transcriptome complexity. They have also demonstrated the prevalence of differential splicing patterns among tissues, cell types, and developmental stages motivating two streams of research, experimental and computational. The first involves RNA-Seq and CLIP-Seq to identify putative targets of splice factors that regulate splicing. The second involves probabilistic models that infer regulatory mechanisms and predict splicing outcome directly from genomic sequence. Till date, these streams of research, while useful for the study of gene regulation, development, and disease, have mostly been disconnected.

RESULTS

Here, we propose a computational framework that extends the work from [1] for deriving predictive splicing code models so that it can leverage the vast amounts of experimental data for splice factor knockdowns. Exploiting recent advances in alternative splicing quantification offered by MAJIQ [2], we define a new target function for splicing code quality based on a mixture of posterior beta distributions. We explore several approaches for optimizing this target function efficiently given a large set of putative regulatory features and demonstrate the usefulness of this new modeling framework on several datasets involving RNA-Seq and separate knockdown experiments of CELF1, CELF2 and MBNL1 in mouse heart and knockdown of MBNL1, RBFOX1 and RBFOX2 in mouse brain.

CONCLUSIONS

Overall, our novel approach offers a scalable solution to extend splicing code modeling for a new type of experimental data that is becoming increasingly common and yet lacks predictive modeling.

REFERENCES

1. Xiong, H.Y., Barash, Y., and Frey, B.J. Bayesian prediction of tissue-regulated splicing using RNA sequence and cellular context. Bioinformatics. 2011 Sep 15;27(18):2554-62
2. Vaquero-Garcia J., Barrera A., Gazzara M.R., González-Vallinas J., Lahens N.F., Hogenesch J.B., Lynch K.W., Barash Y. A new view of transcriptome complexity and regulation through the lens of local splicing variations. Elife. 2016 Feb 1;5;e111752

Characterization of RNA processing alterations in small cell lung cancer

Juan L. Trincado[1], Babita Singh[1], JunYokota[2], Eduardo Eyras[1,3*]
[1] Universitat Pompeu Fabra, E08003 Barcelona, Spain, [2] Institute of Predictive and Personalized Medicine of Cancer (IMPPC), Badalona, Barcelona E08916, Spain, [3] Catalan Institution for Research and Advanced Studies, E08010 Barcelona, Spain
*To whom correspondence should be addressed: eduardo.eyras@upf.edu

---

Small cell lung cancer (SCLC) accounts for 15% of all lung cancers. Previous studies have shown high frequency of mutations in TP53 and RB1 [1], and amplification of MYC [1,2]. However, no targeted therapies have been approved for use in treatment of SCLC, contrary to other lung cancer types like adenocarcinoma. Accordingly, chemotherapy remains the only treatment, which is initially effective but is inexorably followed by rapid relapse in the majority of the patients. Understanding the molecular mechanisms underneath this disease is thus necessary for improving treatment. We have analyzed RNA-seq from 73 RNA-seq SCLC patient samples from [1] and characterized the transcriptomic changes between tumor and normal tissues. We have validated these changes on other 2 cohorts of 31 and 19 RNA-seq SCLC patient samples [3,4]. In order to identify those changes specific of SCLC, and to account for the fact that SCLC tumors have different cell type of origin than other lung tumors, we performed comparisons against more than 400 non-small cell lung samples from The Cancer Genome Atlas and against neuroendocrine lung carcinoid tumors [5]. Additionally, using 71 WGS SCLC samples [1], we looked for somatic mutations disrupting intronic and exonic splicing regulatory motifs that could be responsible for these changes in the transcriptome. This large analysis of RNA processing alterations in SCLC could potentially uncover novel targets of therapy.

REFERENCES

1. Peifer et al. Nature Genetics (2012)
2. George et al. Nature (2015
3. Rudin et al. Nature Genetics (2012)
4. Iwakawa et al. Genes Chromosomes Cancer (2013)
5. Fernandez-Cuesta et al. Nature Communications (2013)

The landscape of alternative splicing alterations in human cancer

Babita Singh[1], Endre Sebestyén[1], Juan L. Trincado[1], Belén Miñana[1,2], Francesca Mateo[3], Miguel Angel Pujana[3], Juan Valcárcel[1,2,4], Eduardo Eyras[1,4,*]

[1]Universitat Pompeu Fabra, Dr. Aiguader 88, E08003 Barcelona, Spain; [2]Centre for Genomic Regulation, Dr. Aiguader 88, E08003 Barcelona, Spain; [3]Program Against Cancer Therapeutic Resistance (ProCURE), Catalan Institute of Oncology (ICO), Bellvitge Institute for Biomedical Research (IDIBELL), E08908 L'Hospitalet del Llobregat, Spain; [4]Catalan Institution for Research and Advanced Studies, Passeig Lluís Companys 23, E08010 Barcelona, Spain

*To whom correspondence should be addressed: eduardo.eyras@upf.edu

---

Alternative splicing enables genes to produce multiple RNA and protein isoforms with different and often opposite functions, and is regulated by a large number of core and auxiliary ribonucleoprotein complexes that bind to specific sites on the pre-mRNA and serve as enhancers or repressors of the splicing reaction. Notably, almost all biological processes important during the neoplastic transformation are profoundly influenced by alternative splicing (AS) of mRNAs [1]. However, the specific alternative splicing mechanisms that are disrupted in tumors are not yet exhaustively characterized.

We systematically analyzed mutation, copy number and gene expression patterns of 1348 RNA-binding protein (RBP) genes in 11 solid tumor types, together with alternative splicing changes in these tumors and the enrichment of binding motifs in the alternatively spliced sequences. Our comprehensive study reveals widespread alterations in the expression of RBP genes, as well as novel mutations and copy number variations in association with multiple alternative splicing changes in cancer drivers and oncogenic pathways. Remarkably, the altered splicing patterns in several tumor types recapitulate those of undifferentiated cells. These patterns are predicted to be mainly controlled by MBNL1 and involve multiple cancer drivers, including the mitotic gene NUMA1. We show that NUMA1 alternative splicing induces enhanced cell proliferation and centrosome amplification in non-tumorigenic mammary epithelial cells [2].

Furthermore, we have studied the somatic mutations on exons and introns genome-wide using whole genome sequencing (WGS) data to obtain a catalogue of RNA binding and splicing regulatory motifs frequently altered in tumors. We found that binding sites for specific RBPs are frequently mutated in tumors and associated significantly with splicing changes as measured from RNA sequencing information, providing new mechanisms of splicing alterations in cancer.

Our results provide a rich resource of information about novel networks of RBPs that trigger common and specific alternative splicing changes in several solid tumors that may be relevant to understand the molecular basis of, and potentially reverse, the oncogenic properties of tumor cells.

REFERENCES:

1. David, C. J., & Manley, J. L. (2010). Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. Genes & development, 24(21), 2343-2364.

2. Sebestyén, E., Singh, B., Miñana, B., Pagès, A., Mateo, F., Pujana, M.A., Valcarcel, J. and Eyras, E., (2016). Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. Genome Research

Fast and accurate computation of differential splicing in plants and animals across multiple conditions

J.C Entizne[1], M. Skalic[1], C. P. G. Calixto[2], R. Zhang[3], A. Pagès[1], J.L. Trincado[1], J. W. S. Brown[2,3], E. Eyras

[1]Universitat Pompeu Fabra, E08003 Barcelona, Spain; [2]College of Life Sciences, University of Dundee, Invergowrie, Dundee, DD2 5DA, UK; [3]The James Hutton Institute, Invergowrie, Dundee, DD2 5DA, UK; [4]Catalan Institution for Research and Advanced Studies, E08010 Barcelona, Spain

---

Alternative splicing plays an essential role in many cellular processes in eukaryotes and bears major relevance in development and disease. High-throughput RNA sequencing allows unbiased genome-wide studies of splicing across multiple conditions. However, the increasing number of available data sets represents a major challenge in terms of computation time and storage requirements. Additionally, there are no dedicated tools for the study of splicing changes across multiple conditions and time series.

We describe SUPPA [1], a computational tool to calculate relative inclusion values of alternative splicing events. SUPPA achieves accuracies comparable or higher than current methodologies but it is thousand times faster [1]. We extended SUPPA to study alternative splicing in plants, which present frequent overlapping events and different properties of exon-intron structures compared to animals. We validate its high accuracy by using RT-PCR for over 40 events in different conditions. We further show a new method to calculate differential splicing across multiple conditions with biological replicates. We applied this method to RNA-Seq data for a time-course of Arabidopsis plants transferred from 20°C to 4°C to examine the effects of low temperature and determine new patterns in circadian clock genes. Additionally, SUPPA uses a novel density-based clustering algorithm to determine groups of events with similar patterns across conditions. We apply this to data across different stages of neuronal differentiation in human and mouse to uncover a novel brain-specific splicing regulatory networks. Assessing the variability in terms of the choice of annotation shows the importance of current efforts aimed at completing the transcript annotation in plants and animals [2].

SUPPA calculates alternative splicing profiles and differential splicing patterns across multiple conditions and from a large number of samples at a much higher speed than existing methods without compromising accuracy, thereby facilitating the systematic analysis of very large data sets with limited computational resources [3]. SUPPA is available at https://bitbucket.org/regulatorygenomicsupf/suppa.

REFERENCES

1. Alamancos et al. RNA 21(9):1521-31.
2. Zhang et al. New Phytol 208(1):96-101
3. Sebestyen et al. Genome Research 2016

Cross-platform normalization of microarray and RNA-seq data for machine learning applications

Jeffrey A. Thompson[1,2], Jie Tan[1,3], and Casey S. Greene[1,4,5,6*]

[1] Department of Genetics, Geisel School of Medicine at Dartmouth College, [2] Quantitative Biomedical Sciences Program, Geisel School of Medicine at Dartmouth College, [3] Molecular and Cellular Biology Program, Geisel School of Medicine at Dartmouth College, [4] Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine at the University of Pennsylvania, [5] Institute for Biomedical Informatics, Perelman School of Medicine at the University of Pennsylvania, [6] Institute for Translational Medicine and Therapeutics, Perelman School of Medicine at the University of Pennsylvania

*To whom correspondence should be addressed: csgreene@mail.med.upenn.edu

---

BACKGROUND

A wealth of legacy microarray gene expression experiments exist in private and public databases, relevant to a large range of research questions. While RNA-seq data are now relatively inexpensive to generate, discarding the existing corpus of gene expression experiments wastes a valuable opportunity and biological tissues, particularly in the case of rare diseases for which it is hard to obtain samples. Therefore, we developed the Training Distribution Matching (TDM) method, which performs cross-platform normalization, enabling machine learning algorithms to build models on data from one platform and apply them to data from another.

RESULTS

TDM was compared to quantile normalization [1], the nonparanormal transformation [2], and simple log2 transformation for its ability to aid classification using models trained on microarray data and tested on RNA-seq data. Tests were performed on a simulated dataset with four artificial treatments at increasing levels of noise, using unsupervised clustering, and on three biological datasets, using LASSO multinomial logistic regression [3] to classify samples by tumor subtype. For the simulated data, TDM showed itself to be robust to increasing noise. For the biological datasets, TDM normalized data had the most consistent performance, with the highest accuracy for one dataset and second highest for two others. The results for the third biological dataset were notable, because test data were available with both microarray and RNA-seq for the same samples, which were from breast cancer biopsies collected by TCGA [4]. The independent microarray training data were from METABRIC [5]. This allowed us to compare classification performance on RNA-seq data to actual microarray data for the same samples. For this dataset, quantile normalization performed the best, with an accuracy of .84 and Kappa of .77, but TDM was nearly the same with an accuracy of .83 and Kappa of .76. On microarray test data the accuracy was .85 and Kappa was .78, demonstrating that a microarray-trained model can perform about the same on RNA-seq as it does on microarray.

CONCLUSIONS

We have developed a new method for cross platform normalization, TDM, which is specifically focused on machine learning applications. RNA-seq data normalized by TDM resulted in the most consistent performance. Our results suggest that models built on data from one platform can be applied to another to generate meaningful predictions. We also provide a TDM package for the R programming language, available at: https://github.com/greenelab/TDM.

REFERENCES

1. Benjamin M. Bolstad, Rafael A. Irizarry, Magnus Astrand, and Terence P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics, 19(2):185–193, 2003.
2. Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. The Journal of Machine Learning Research, 10:2295–2328, 2009.
3. Robert Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), pages 267–288, 1996.
4. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature, 490(7418):61–70, 2012.
5. Christina Curtis, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature, 486(7403):346–352, 2012.

The reference Trypanosoma cruzi transcriptome generated by de novo assembly of RNA-Seq data

Mainá Bitar, Eddie Imada, Helaine Grazielle dos Santos Vieira, Priscila Grynberg, Michele Araújo, André Martins dos Santos Reis, William Prado, Dominik Kaczorowski, Andréa M. Macedo, Carlos R. Machado, Martin A. Smith, John S. Mattick, Glória R. Franco

Universidade Federal de Minas Gerais, QIMR Berghofer Institute, Brazil

*To whom correspondence should be addressed: validée@biof.ufrj.br

---

BACKGROUND

Trypanosoma cruzi is a trypanosomatid and the etiologic agent of Chagas disease, which is currently estimated to affect at least 6 million people worldwide. In 2005 El-Sayed and collaborators sequenced the genome of the hybrid T. cruzi CL Brener clone, currently available as a partially assembled set of 82 chromossomes (41 per haplotype).

RESULTS

Given the lack of a completely assembled genomic data, RNA-Seq experiments should rely on assembled transcriptomes. We therefore decided to provide a de novo assembled, curated and analyzed version of the T. cruzi CL Brener epimastigote transcriptome, based on high-quality RNA-Seq data to serve as reference. We applied a polyA capture-based strand-specific Illumina RNA-Seq methodology to generate ~70 million paired-end reads that should reflect the total RNA content of epimastigotes. Several filters were applied to these sequences in order to eliminate contaminants, discard lower quality reads and reads derived from spike-in RNAs. After the entire curation process, nearly 15% of the reads were discarded, thus representing a relevant curation step prior to de novo transcriptome assembly. Reads were assembled by Trinity and the number and length of assembled transcripts reflected the annotated sequences on T. cruzi genome. A database will shortly be available to store and allow access to the assembled transcripts, their annotations, expression levels and basic statistics.

CONCLUSIONS

We consider this a pioneer approach for the investigation of gene expression in trypanosomatids and a very useful resource for scientists working with neglected tropical diseases caused by these parasites.

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
Yoseph Barash
for
IRB-SIG 2016