Intergrative RNA Biology
Special Interest Group Meeting

July 11, 2014

Content

Program - Friday, July 11

7:30 a.m. – 8:30 a.m.	Registration

SESSION 1

08:30 - 08:40	Opening notes
08:40 - 09:15	**Invited speaker: John Calarco.** *Shining a light on the diversity of messages in the nervous system*
09:15 - 09:30	**Maria Carmo-Fonseca.** *5 seconds to splice*
09:30 - 09:45	**Courtney French.** *Transcriptome analysis reveals thousands of targets of nonsense-mediated mRNA decay that offer clues to the mechanism in different species*
09:45 - 10:00	**Hyeshik Chang.** *TAIL-seq: Genome-wide determination of poly(A) tail length and 3' end modifications*
10:00 - 10:15	**Lawrence Chasin.** *Saturation mutagenesis of a human exon*
10:15 - 10:45	Coffee Break

SESSION 2

10:45 - 11:20	**Invited speaker: Kristen Lynch.** *Interplay of signaling and splicing during T cell activation*
11:20 - 11:35	**Auinash Kalsotra.** *Identification of a conserved and cell-type specific program of regulated mRNA splicing in postnatal liver development.*
11:35 - 11:50	**Fátima Sánchez-Cabo.** *Understanding the regulation of alternative splicing changes in heart disease*
11:50 - 12:20	**Invited speaker: Aviv Regev/Rahul Satija** *From culture to clinic : single cell profiling of the mammalian immune system.*
12:20 - 13:30	Lunch

SESSION 3

13:30 - 14:05	**Invited speaker: Christina Leslie.** *Context-specific 3'UTR isoform expression and miRNA regulation*
14:05 - 14:20	**Angela Brooks.** *The landscape of RNA splicing alterations in human cancers*
14:20 - 14:35	**Sara Gosline.** *Beyond Argonaute: understanding microRNA dysregulation in cancer and its effect on protein interaction and transcriptional regulatory networks*
14:35 - 14:50	**Will Fairbrother.** *Massively parallel reporter assays reveal splicing defects in 20% missense disease alleles*
14:50 - 15:25	**Invited speaker: Chris Burge.** *Origins and impacts of new exons*
15:25 - 16:00	Coffee Break (and start of poster session)

SESSION 4

16:00 - 17:30	***Poster session***
17:30 - 17:45	**Arvind Subramaniam.** *Mechanistic constraints for modelling translation from genome-wide measurements of ribosome occupancy*
17:45 - 18:00	**Alexander Junge.** *RNAalignClust: Discovering ncRNA families by sequence-structure-based clustering of multiple sequence alignments*
18:00 - 18:15	**Stephen Mount.** *Specialized reference transcriptomes for Sailfish*
18:15 - 18:50	**Invited speaker: Quaid Morris.** *The eukaryotic RNA-protein interaction code*
18:50 – 19:00	Concluding notes, poster prize announcement
19:30	**IRB-SIG dinner**

Program - Friday, July 11

7:30 a.m. – 8:30 a.m.	Registration

SESSION 1

08:30 - 08:40	Opening notes
08:40 - 09:15	**Invited speaker: John Calarco.** *Shining a light on the diversity of messages in the nervous system*
09:15 - 09:30	**Maria Carmo-Fonseca.** *5 seconds to splice*
09:30 - 09:45	**Courtney French.** *Transcriptome analysis reveals thousands of targets of nonsense-mediated mRNA decay that offer clues to the mechanism in different species*
09:45 - 10:00	**Hyeshik Chang.** *TAIL-seq: Genome-wide determination of poly(A) tail length and 3' end modifications*
10:00 - 10:15	**Lawrence Chasin.** *Saturation mutagenesis of a human exon*
10:15 - 10:45	Coffee Break

SESSION 2

10:45 - 11:20	**Invited speaker: Kristen Lynch.** *Interplay of signaling and splicing during T cell activation*
11:20 - 11:35	**Auinash Kalsotra.** *Identification of a conserved and cell-type specific program of regulated mRNA splicing in postnatal liver development.*
11:35 - 11:50	**Fátima Sánchez-Cabo.** *Understanding the regulation of alternative splicing changes in heart disease*
11:50 - 12:20	**Invited speaker: Aviv Regev/Rahul Satija** *From culture to clinic : single cell profiling of the mammalian immune system.*
12:20 - 13:30	Lunch

SESSION 3

13:30 - 14:05	**Invited speaker: Christina Leslie.** *Context-specific 3'UTR isoform expression and miRNA regulation*
14:05 - 14:20	**Angela Brooks.** *The landscape of RNA splicing alterations in human cancers*
14:20 - 14:35	**Sara Gosline.** *Beyond Argonaute: understanding microRNA dysregulation in cancer and its effect on protein interaction and transcriptional regulatory networks*
14:35 - 14:50	**Will Fairbrother.** *Massively parallel reporter assays reveal splicing defects in 20% missense disease alleles*
14:50 - 15:25	**Invited speaker: Chris Burge.** *Origins and impacts of new exons*
15:25 - 16:00	Coffee Break (and start of poster session)

SESSION 4

16:00 - 17:30	***Poster session***
17:30 - 17:45	**Arvind Subramaniam.** *Mechanistic constraints for modelling translation from genome-wide measurements of ribosome occupancy*
17:45 - 18:00	**Alexander Junge.** *RNAalignClust: Discovering ncRNA families by sequence-structure-based clustering of multiple sequence alignments*
18:00 - 18:15	**Stephen Mount.** *Specialized reference transcriptomes for Sailfish*
18:15 - 18:50	**Invited speaker: Quaid Morris.** *The eukaryotic RNA-protein interaction code*
18:50 – 19:00	Concluding notes, poster prize announcement
19:30	**IRB-SIG dinner**

Sponsors

RNA Society

Abstracts

Shining a light on the diversity of messages in the nervous system

John Calarco[1]
[1]FAS Center for Systems Biology, Harvard University

BACKGROUND

Recent transcriptome-wide analyses of multicellular organisms have identified that a significant fraction of messenger RNAs (mRNAs) are subject to tissue-specific regulation of their abundance and/or diversity. One important layer contributing to tissue-regulated mRNA diversity is the phenomenon of alternative splicing. Our group is currently exploring the mechanisms governing differential alternative splicing in the nervous system at single cell resolution.

RESULTS

Using a fluorescence microscopy-based genetic screening approach in the nematode C. elegans, we have recently identified a pair of RNA binding proteins that coordinate differential splicing patterns between GABAergic and cholinergic neurons, the two major classes of motor neurons in the animal. I will discuss our ongoing efforts towards characterizing how these factors establish this neuron-specific regulation. I will also present results suggesting that these proteins play a critical role in fine-tuning the physiological properties of these neurons. Finally, I will describe the adaptation of a method for isolating mRNAs from specific cell types in C. elegans, and the future use of this genome-wide approach to uncover tissue and neuronal-subtype specific splicing.

CONCLUSIONS

Our work demonstrates that there is an incredible amount of diversity at the level of alternative splicing regulation in distinct neuron classes in C. elegans. The approaches described here should begin to allow us to gain a deeper mechanistic understanding of how differential splicing is regulated within the nervous system, and insights into the functional consequences of this regulation.

5 seconds to splice

José Braga‡[1], José Rino‡[1], Robert M. Martin[1], Ana Paula Leite[1], Célia Carvalho[1],
Tomas Kirchhausen[2], and **Maria Carmo-Fonseca**[1*]
[1] Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Portugal, [2] Department of Cell Biology, Harvard Medical School, Immune Disease Institute and Program in Molecular and Cellular Medicine at Children's Hospital, Boston, Massachusetts, USA

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

BACKGROUND

Expression of genetic information in eukaryotes involves a series of interconnected processes that ultimately determine the quality and amount of proteins in the cell. Many individual steps in gene expression are kinetically coupled, but tools are lacking to determine how temporal relationships between chemical reactions contribute to the output of the final gene product.

RESULTS

To decipher how splicing occurs in real time, we have directly examined with single-molecule sensitivity the kinetics of intron excision from pre-mRNA in the nucleus of living human cells [1]. We also developed a probabilistic model of transcription-coupled splicing to extract kinetic parameters from time-lapse recordings of fluorescence intensity at the transcription site. We find that pausing of RNA polymerase II at the 3'-splice site is very brief, in the range of a few seconds, and it takes less than 5 s for splicing to be completed after the polymerase resumes elongation past the 3' splice site. This fast model of splicing was validated by analysis of PRO-seq data that revealed spliced transcripts associated with polymerases that have just transcribed intron-exon junctions.

CONCLUSIONS

Our results provide a high-resolution view of the temporal and spatial relationships between transcription and splicing. In contrast to the fast rate of splicing that we observe, previous studies indicated that slicing requires several minutes for completion. Reasons for such discrepancy will be discussed.

REFERENCES

1. Martin, RM, Rino, J, Carvalho, C, Kirchhausen, T, Carmo-Fonseca, M. (2013) Live-cell visualization of pre-mRNA splicing with single-molecule sensitivity. Cell Reports 4: 1144-1155 (2013).

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

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

[1] University of California, Berkeley, CA [2] Fudan University, Shanghai, China [3] Broad Institute, Cambridge, MA [4] Ohio State University, Columbus, OH [5] Partner Institute of Computational Biology, Shanghai, China [6] University of Connecticut Health Center, Farmington, CT

*To whom correspondence should be addressed: cfrench@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]. 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 [3]. There is evidence that this rule holds in *Arabidopsis* [4] but not in other eukaryotes including *Drosophila* [5]. There is also evidence that a longer 3' UTR triggers NMD in yeast, plants, flies, and mammals [4,6,7].

RESULTS

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

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

CONCLUSIONS

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

REFERENCES

1. Lareau, L., et al. **446** (2007).
2. Ni, J., et al. Genes and Development. **21** (2007)
3. Nagy E. and Maquat, L. Trends in Biochemical Science. **23** (1998)
4. Kerenyi, Z., et al. EMBO Journal. **27** (2008)
5. Gatfield, D., et al. EMBO Journal. **22** (2003)
6. Hansen, K., et al. PLoS Genetics. **5** (2009)
7. Hogg, J and Goff, S. Cell. **143** (2010)

TAIL-seq: Genome-wide determination of poly(A) tail length and 3' end modifications

Hyeshik Chang[1], Jaechul Lim[1], Minju Ha[1], and V. Narry Kim[1,*]

[1] Center for RNA Research, Institute for Basic Science, and School of Biological Sciences, Seoul National University, Seoul 151-742, Korea

*To whom correspondence should be addressed: narrykim@snu.ac.kr

BACKGROUND

The 3′ termini of eukaryotic RNAs reflect the history of transcript and play important roles in determining the fate of RNA. Despite the importance, the actual sequences of 3′ ends remain unknown for the vast majority of transcripts, and our current knowledge is based on studies of a limited number of individual genes by northern- and RT-PCR/Sanger sequencing-based techniques [1, 2]. Genome scale investigation has been hampered mainly due to the limitation of the current deep sequencing technologies which cannot determine homopolymeric sequences of longer than ~30 nt. Although microarray combined with differential elution from oligo(dT) column have been used to roughly estimate poly(A) length [3, 4], the resolution is too low for accurate measurement.

RESULTS

We here develop a method, TAIL-seq, to sequence the very end of mRNA molecules [5]. To work around the serious inaccuracy in sequencing homopolymers, we adopt the original fluorescence signals from poly(A) tails instead of the standard base calls. The signals from chemically synthesized spike-ins are learned using a machine learning algorithm, then the cDNAs produced from cells are analyzed with the model to call the length of poly(A) tails and identify additional tailing following poly(A) tails. TAIL-seq allows us to measure poly(A) tail length at the genomic scale. Median poly(A) length is 50–100 nt in HeLa and NIH 3T3 cells. Poly(A) length correlates with mRNA half-life, but not with translational efficiency. Surprisingly, we discover widespread uridylation and guanylation at the downstream of poly(A) tail. The U tails are generally attached to short poly(A) tails (<25 nt), while the G tails are found mainly on longer poly(A) tails (>40 nt), implicating their generic roles in mRNA stability control. By depleting mRNA decay enzymes, we further find that uridylation of poly(A) tails plays a central role in mRNA decay.

CONCLUSIONS

TAIL-seq is the first method that allows global survey of poly(A) length and 3′ end modification of mRNA. In designing the current version of TAIL-seq, we aimed to be as comprehensive as possible, which allowed us to discover many new exciting features such as differential poly(A) length control, uridylation, guanylation, and RNA cleavage. TAIL-seq will also be useful to solve various general issues regarding the relative dynamics of mRNA deadenylation, translation, and decay. In addition, one can examine RNA terminal modifications in diverse physiological and pathological contexts, such as in neural synapse, late oogenesis, early embryogenesis, cellular senescence and inflammation where dynamic control of cytoplasmic polyadenylation is known to play a critical role. The TAIL-seq protocol can be applied to any species and cell types with minor modifications, which will greatly expand the initial observations made in this study.

REFERENCES

1. Norbury, C. J. Cytoplasmic RNA: a case of the tail wagging the dog. Nat Rev Mol Cell Biol **14** (2013).
2. Sallés, F. J., Richards, W. G., and Strickland, S. Assaying the polyadenylation state of mRNAs. Methods **17** (1999).
3. Beilharz, T. H. and Preiss, T. Widespread use of poly(A) tail length control to accentuate expression of the yeast transcriptome. RNA **13** (2007).
4. Meijer, H. A., Bushell, M., Hill, K., Gant, T. W., Willis, A. E., Jones, P., and de Moor, C. H. A novel method for poly(A) fractionation reveals a large population of mRNAs with a short poly(A) tail in mammalian cells. Nucleic Acids Res **35** (2007).
5. Chang, H., Lim, J., Ha, M., and Kim, V. N. TAIL-seq: genome-wide determination of poly(A) tail length and 3′ end modifications. Mol Cell **53** (2014).

Saturation mutagenesis of a human exon

Shengdong Ke, Jorge Rojas-Zamalloa, and **Lawrence A. Chasin***

Department of Biological Sciences, Columbia University, New York, NY USA
*To whom correspondence should be addressed: lac2@columbia.edu

BACKGROUND

Exon splicing regulators (ESRs) are sequence motifs that contribute information for the distinction of real splice sites from the much larger number of pseudo splice sites that abound in long introns. Validated computationally predicted and experimentally selected ESR motifs are now so numerous they comprise the majority of the nucleotides found in typical constitutively or alternatively spliced exons. To what extent are these apparent motifs functional and if so how do they work together to bring about, in the case of constitutively spliced exons, a binary decision to include a real exon and skip a pseudo exon? We have used a high throughout genetic approach to probe these questions. A 51-nt central exon (a poorly spliced Wilms tumor 1 exon 5) in a 3-exon minigene was subjected to saturation mutagenesis by creating a library of synthetic oligonucleotides. Adjacent dinucleotides at each exon position were replaced with all other possible dinucleotides, resulting in 555 variations. This population of molecules was transfected into human HEK293 cells and after 24 hours the pool of successfully spliced mRNA molecules was isolated and deep sequenced. The relative number of sequence reads of each mutant molecule mirrors its splicing efficiency [1]. Due to the deprivation of its natural intronic flank sequences, the wild type exon was spliced in at only 6.5%; this level of splicing was designed to allow us to sensitively detect mutations that either increase or decrease splicing.

RESULTS

We found that 70% of the mutations resulted in changes of at least 2-fold (up or down) in mRNA levels and these included mutants with as much as a 16-fold increased and a decrease to near zero representation. To determine how many of these changes were due to splicing as opposed to a possible effect on mRNA stability, we repeated the experiment using the same mutant library constructed as intronless minigenes. Very few changes and none of comparable magnitude were observed, suggesting that the original mutant phenotypes were due to changes in splicing. We conclude that in this exon, most of the 51 nucleotides indeed are contributing to splicing efficiency. To ask whether secondary structural differences play a role in splicing efficiency, for each mutant we calculated the predicted probability that a base was unpaired. Unpairedness of two 5-nt regions representing a stem of a stem-loop structure exhibited a significant positive correlation with splicing (Pearson's r ~0.5), suggesting that disruption of this stem leads to increased splicing. We used the CISBP-RNA database of RNA-binding protein (RBP) specificities [2] to map potential RBP binding motifs in the wild type and mutant exons. Of 91 RBPs queried, at least 16 are predicted to bind to 7-nt motifs in the 51-nt exon, most of which are clustered in 3 locations. Despite the complication of overlapping binding sites, at least 5 RBPs show a good correlation (r^2 >0.5) between relative binding affinity and splicing, either positive (enhancing) or negative (silencing). To probe for evidence of combinatorial effects between motifs, we repeated this experiment with 9 variants of the wild type exon, each of which carries a different 6-nt substitution near the 5' end of the exon (positions 5 to 10). Interestingly, none of the variants showed a strong secondary structure correlation, hinting that the stem loop structure seen in the wild type may have been selected for in evolution. Mutations in the downstream half of the exons produced nearly identical phenotypes among the 10 exons. However, in the region just downstream and adjacent to the 6 nt substitutions, the double base substitutions produced very difference phenotypes in almost all of the variants compared to the wild type. This result is consistent with the idea that a combination that was functional in the wild type was lost by disrupting the 5 to 10 nt region. Alternatively, a local secondary structure difference could underlie this observation.

CONCLUSIONS

In an incompletely spliced exon, most of the nucleotides play a role in splicing efficiency. There appears to be ample binding of RBPs to mediate these effects, which are likely to be governed by both competition for binding to overlapping sites and to interactions between proteins bound to neighboring motifs but not to more distant motifs. Moreover, RNA structural effects must also be accommodated. Additional precisely designed high throughput genetic experiments based on these results should help test these ideas. Additional authors: Vincent Anquetil, Alisha Maity, Sergey Kalachikov, Irina Morozova, Jingyue Ju, all at Columbia University. Present address of S. Ke is Lab. of Molecular Neuro-oncology, Rockefeller University, NY.

REFERENCES:

1. Ke, S.,t al.. Quantitative evaluation of all hexamers as exonic splicing elements. Genome Res. 21:1360-1374 (2011)
2. Ray, D., et al., A compendium of RNA-binding motifs for decoding gene regulation. Nature. 499:172–177 (2013) and http://cisbp-rna.ccbr.utoronto.ca

Interplay of Signaling and Splicing during T cell Activation

Kristen W. Lynch*

Departments of Biochemistry and Biophysics and Genetics, Perelman School of Medicine, University of Pennsylvania

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

Alternative splicing is a ubiquitous mechanism of gene expression that is tightly regulated in response to developmental and environmental stimuli. For example, our laboratory and others have demonstrated widespread changes in alternative splicing upon T cell activation. However, major unanswered questions include how specific signaling pathways connect antigen engagement of the T cell receptor to changes in nuclear splicing regulation, and how individual or sets of co-regulated alternative splicing events influence the downstream function of T cells in an immune response.

We have recently utilized RNA-Seq, and variants thereof, to gain a comprehensive view of splicing regulation in T cells. Specifically, we have assayed differential splicing before and after antigen-induced stimulation, following treatment with various kinase inhibitors, and upon depletion of splicing factors previously shown to be important components of signal-induced alternative splicing in T cells. We have identified over 400 genes that exhibit robust changes in isoform expression upon T cell activation, and attribute the JNK and GSK3 signaling pathways to much of this regulation. We have further used computer modeling to identify significant functional and regulatory features of groups of co-regulated genes. Interestingly, we have found that at least one of the genes regulated by the JNK-signaling pathways is itself an activator of JNK kinase activity and promotes JNK effector functions. Together, our data demonstrate a reciprocal relationship between signaling and splicing in T cells, in which the initial activation of specific signaling cascades leads to changes in splicing that then further amplify the functional outcome of antigen-stimulation.

Identification of a conserved and cell-type specific program of regulated mRNA splicing in postnatal liver development

Amruta Bhate[1*], Darren John Parker[1*], Jaegyoon Ahn [2], Anthony Chau[1], Sandip Chorghade[1], Jae-Hyung Lee[2], Yaseswini Neelamraju[3], Sarath Chandra Janga[3], Xinshu Xiao[2] and **Auinash Kalsotra**[1, 4]

[1]Department of Biochemistry, [4]Institute of Genomic Biology, University of Illinois, Urbana-Champaign, [2]Department of Integrative Biology and Physiology, University of California, Los Angeles. [3]School of Informatics and Computing, Indiana University-Purdue University, Indianapolis. *Equal contribution.

*Correspondence should be addressed to kalsotra@illinois.edu

BACKGROUND

While the major regulatory programs controlling early specification and morphogenesis of liver are well studied, how the organ matures during the postnatal period is poorly understood.

RESULTS

We performed large-scale high-resolution RNA-seq analysis of mouse livers across a postnatal developmental time course. We demonstrate a pervasive and a highly coordinated shift in liver transcriptome within the first four weeks after birth. Remarkably, the genes undergoing changes in expression, alternative splicing (AS) and alternative 3'-UTR usage not only show minimal overlap but also exhibit enrichment for unique functional categories. While many DNA-binding proteins follow the same overall pattern of expression as *all* genes, we report that a large number of RNA-binding proteins are down regulated postnatally.

Direct comparison of over one hundred developmentally regulated, and variable spliced regions between mouse and human livers show that nearly 50% are evolutionarily conserved in both timing and direction. When analyzed in freshly isolated hepatocytes and non-parenchymal cells, these variable regions show cell-type specific transitions such that subsets of events follow either similar or opposite patterns of splicing through development. A computational framework to correlate the changes in variable regions at specific developmental time points and various *cis* elements identified 117 RNA-binding proteins to be significantly associated with splicing alterations. Further, a *de novo* motif analysis showed a positional enrichment for many tissue-specific splicing factors around the variably spliced regions. Detailed temporal analysis of splicing and the associated regulatory factors revealed that a majority of AS changes are co-regulated, and follow either prenatal, postnatal or biphasic patterns of change.

CONCLUSIONS

Together, our results identify a complex and a highly conserved program of mRNA processing associated with the development of the mammalian liver that supports its physiological growth and maturation.

Understanding the regulation of alternative splicing changes in heart disease

Fátima Sánchez-Cabo[1*], Alberto Gatto[2], Girolamo Giudice[2], Carlos Torroja[1] and Enrique Lara-Pezzi[2]

[1] Bioinformatics Unit, Centro Nacional de Investigaciones Cardiovasculares, Madrid, Spain [2] Cardiovascular Development and Repair Department, Centro Nacional de Investigaciones Cardiovasculares, Madrid, Spain

*To whom correspondence should be addressed: fscabo@cnic.es

BACKGROUND

Heart failure is a widespread, deadly diseases for which the molecular basis is still poorly understood. While alternative splicing and transcriptional regulation are known to play a major role, we lack a global understanding of how different regulation layers shape the transcriptional landscape following cardiac insult. Here we aimed at the identification and comparison of the key functional elements behind splicing changes in different models of cardiovascular disease: (i) *in vitro* model with cardiomyocytes (CM) and cardiac fibroblasts (CF) to reproduce the heart changes following myocardial infarction; (ii) *in vivo* mouse model one week after myocardial infarction and (iii) human patients with ischemic and non-ischemic heart failure.

RESULTS

We developed an analysis pipeline for RNA-Seq data to detect alternatively spliced genes based on the distribution of the expression of their isoforms. In all our experiments, alternatively spliced and differentially expressed genes were largely different, underlying the importance of characterizing alternative splicing changes in heart disease. Approximately a third of the genes alternatively spliced in CM and CF mimicking myocardial infarction *in vitro* were also alternatively spliced in the *in vivo* model. Genes alternatively spliced after myocardial infarction were mainly involved in proteolysis, cytoskeleton and vesicle transport, supporting the findings described in [1]. Genes alternatively spliced in human heart failure vs control samples were also enriched in proteolysis and vesicle transport.

To understand the regulation of these changes, we considered all non-constitutive cassette exons from the switching isoforms and searched for several regulatory features in their sequences and in their upstream and downstream intronic regions: (i) known motifs of RNA Binding Proteins (RBP) compiled in our in-house database (ii) CLIP-Seq binding sites from [2] and (iii) histone marks and transcription factors (TF) binding sites from ENCODE. For the enriched (RBP) and TFs, we reconstructed the regulatory network controlling alternative splicing. For all networks, the regulatory program of included and skipped exons was mainly different. The regulatory network of the mice model of MI shared different elements with the CM and the CF networks. Motifs for some known regulators of alternative splicing in the developing heart (MBNL1 and CELF proteins [3]) were found to be enriched in the exons alternatively spliced during MI. There was also an enrichment of the proteins of SRSF family, some of which were shared with the CM network.

CONCLUSIONS

Using an integrative computational approach we have been able to understand not only general and specific changes in alternative splicing in different human and animal models of heart failure but also to unravel the main regulators behind these changes.

REFERENCES

1. Giudice, J., Xia, Z., Wang, E.T., Scavuzzo, M.A., Ward, A.J., Kalsotra, A., Wang, W., Wehrens, X.H.T., Burge, C.B., Li, W., et al. (2014) Alternative splicing regulates vesicular trafficking genes in cardiomyocytes during postnatal heart development. *Nat Commun*
2. Li, J.-H., Liu, S., Zhou, H., Qu, L.-H. and Yang, J.-H. (2014) starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Research*, 42, D92–D97.
3. Kalsotra, A., Xiao, X., Ward, A.J., Castle, J.C., Johnson, J.M., Burge, C.B. and Cooper, T.A. (2008) A postnatal switch of CELF and MBNL proteins reprograms alternative splicing in the developing heart. *PNAS*, **105**, 20333–20338.

From culture to clinic: single cell profiling of the mammalian immune system

Rahul Satija[1,2], **Aviv Regev**[1,2,3]

[1]Department of Biology, Massachusetts Institute of Technology, [2]The Broad Institute, [3]Howard Hughes Medical Institute

Individual cells in a ‘homogeneous’ population can differ markedly in their molecular components, and in particular, cellular heterogeneity in RNA levels has been shown to be widespread. Of particular interest, however, are groups of genes whose expression levels exhibit co-variation across single cells, and may thus be targets of a shared upstream regulator. While, identifying these relationships in single cell RNA-seq is a promising tool for reconstructing regulatory circuitry, extensive biological and technical noise presents a core challenge. Standard correlation-based approaches are thus poorly suited for analyzing single cell data, particularly for small datasets.

Here, we utilized the Fluidigm C1 Single Cell Auto Prep system to analyze the immune response of mouse primary dendritic cells (DCs) stimulated with three pathogenic components. Our data revealed extensive heterogeneity across the responding transcriptome, with cell-to-cell differences approaching 1,000 fold in extreme cases. In parallel, we designed a statistical framework that explicitly calculated models of technical noise for each library in order to probabilistically interpret and analyze our single cell measurements. I will show how these tools enable us to characterize cellular heterogeneity, detect rare subpopulations, and identify putative transcriptional regulators that drive cellular variation. I will conclude by discussing our ongoing efforts to apply these techniques to better characterize the human immune system.

Context-specific 3'UTR isoform expression and miRNA regulation

Christina Leslie[1*]

[1]Computational Biology Program, Memorial Sloan Kettering Cancer Center

*To whom correspondence should be addressed: cleslie@cbio.mskcc.org

BACKGROUND

Alternative cleavage and polyadenylation (ApA) generates mRNA transcripts that differ in the length of their 3' untranslated region (3'UTR), altering the post-transcriptional fate of the message through inclusion or exclusion of miRNA binding sites, other regulatory elements, and localization signals in distal UTR regions.

RESULTS

We will first describe recent insights into the nature and extent of 3'UTR variation due to alternative cleavage and polyadenylation (ApA) across human tissues from analysis of 3' end sequencing (3'-seq) data. In particular, although more that half of human genes use multiple pA signals, true tissue-specific ApA events are rare; rather, genes alter the relative expression of their 3'UTR isoforms in a tissue-specific manner. We will present evidence that ApA plays a functional role in tissue-specific expression programs and contributes to context-specific miRNA regulation by altering the landscape targetable by ubiquitously expressed miRNAs. Next, to more closely examine the relationship between ApA and miRNA regulation – and to ask whether factors beyond ApA contribute to context-specific miRNA regulation – we study the miR-155 regulatory program in four different activated immune cell types through analysis of 3'-seq data in wild type and miR-155 knockout cells. Finally, time permitting, we will describe improved modeling of miRNA targeting and evidence for context-specificity from analysis of AGO CLIP-seq and CLASH-like data sets.

CONCLUSIONS

Recent next-generation sequencing technologies such as 3'-seq and CLIP-seq, together with careful statistical analysis, provide new insights into context-specific ApA and miRNA regulation.

The landscape of RNA splicing alterations in human cancers

Angela N. Brooks[1,2*], Yawei Ge[3], Kevin Chau[3], Gordon Saksena[1], Chandra Sekhar Pedamallu[1], and Matthew Meyerson[1,2,3*]

[1] Broad Institute, Cambridge MA, [2] Dana-Farber Cancer Institute, Boston, MA, [3] Harvard Medical School, Boston, MA

*To whom correspondence should be addressed: brooks@broadinstitute.org, Matthew_Meyerson@dfci.harvard.edu

BACKGROUND

Recent whole-exome sequencing studies have found that somatic mutations frequently occur in splicing factors across multiple cancer types, supporting the need to systematically and globally characterize splicing alterations across human cancers. Through the integration of mRNA and whole-exome sequencing data from The Cancer Genome Atlas, we are identifying RNA splicing alterations across ~7,000 cancer transcriptomes and investigating the underlying somatic mutations that cause these splicing alterations. To perform this analysis, we have further developed a computational pipeline called JuncBASE [1] to identify and quantify alternative splicing in cancer RNA-Seq data. In an initial study, we identified 470 altered splicing events significantly associated with mutations in the splicing factor *U2AF1* [2], including an altered splicing event in *CTNNB1*.

RESULTS AND CONCLUSIONS

Continuing our analysis of splicing alterations, we have now identified somatic mutations in splice sites using an extended annotation of splice site positions, beyond what is typically used in cancer genomic studies, and have found associated transcriptome changes at the gene expression or splicing level. As a result of this analysis, we have identified splice site mutations that are associated with expression of oncogenic isoforms, including isoforms of *MET* and *ERBB2*. Restricting our analysis to known or predicted mutations in splicing regulatory elements may miss important splicing alterations; therefore, we are also using outlier detection methods to identify additional altered splicing events. To distinguish between cancer-specific splicing alterations and normal transcriptome variation, we are utilizing RNA-Seq data from healthy individuals from the Genotype-Tissue Expression (GTEx) project. To identify splicing events that may be novel somatic driver alterations, these events have been profiled using RNA-Seq data from the Cancer Cell Line Encyclopedia and are being used as biomarkers to identify genetic vulnerabilities in high-throughput shRNA screens. This work will have a significant impact on our understanding of the role of splicing in cancer pathogenesis.

REFERENCES

1. Brooks AN, Yang L, Duff MO, et al. (2011). Conservation of an RNA regulatory map between Drosophila and mammals. *Genome Research* 21:193-202
2. Brooks AN, Choi PS, de Waal L, et al. (2014). A Pan-Cancer Analysis of Transcriptome Changes Associated with Somatic Mutations in U2AF1 Reveals Commonly Altered Splicing Events. *PLoS One*, *9*(1), e87361. doi:10.1371/journal.pone.0087361

Beyond Argonaute: understanding microRNA dysregulation in cancer and its effect on protein interaction and transcriptional regulatory networks

Sara JC Gosline[1], Gabriela Pregernig[1], Coyin Oh[1] and Ernest Fraenkel[1*]

[1] Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA

*To whom correspondence should be addressed: fraenkel-admin@mit.edu

BACKGROUND

microRNAs (miRNAs) are small (~22 nucleotide) molecules that post-transcriptionally regulate gene expression by guiding the Argonaute complex to genes possessing a complementary region in the 3' un-translated region (UTR) of the mRNA transcript, causing the premature degradation and decreased translation of miRNA-targeted genes. miRNA targets can be predicted computationally by searching for complementary 'seed' regions in putative mRNA targets, yet mRNA expression measurements following miRNA transfection or deletion identify many mRNAs that change without evidence of a predicted miRNA-mRNA interaction. These broader changes suggest that miRNAs can affect cellular signaling by targeting protein interaction networks that ultimately alter transcriptional regulation.

RESULTS

In cancer, miRNAs and mRNAs are highly dysregulated, with specific expression 'signatures' being highly correlated with poor patient prognosis. While these signatures represent a read-out of underlying dysregulation, they offer little insight into the mechanistic changes that occur. Here, we integrate miRNA and mRNA expression data from the cancer genome atlas (TCGA) with published interactome data to identify altered pathways in cancer (TCGA Network, 2013). By studying the relationship of miRNAs with predicted mRNA targets and with mRNA anti-correlated with miRNA across patient samples, we can uncover the role of miRNAs in cancer prognosis. We aim to expand previous approaches that have been limited to the study of the direct effects of miRNAs on mRNA expression or transcriptional regulation by studying all putative protein interactions.

We've collected matched miRNA and mRNA expression levels from 866 TCGA breast cancer samples with predicted protein-DNA interactions derived from MCF7-derived Dnase I hypersensitivity data from ENCODE (Birney et al., 2007) and published protein-protein interactions. We then used SAMNet, a multi-commodity flow-based algorithm (Gosline et al., 2012), to identify miRNA-specific interaction networks that best explain the relationship between miRNA and anti-correlated mRNA.

Our initial results (http://fraenkel.mit.edu/samnetweb/brca_mir) identify individual networks of influence for each of the top expressed miRNAs in breast cancer patients. We have been able to find interactions such as that between CBX5 and TRIM28 that are shared across miRNA signaling pathways. TRIM28 is known to be correlated with breast cancer prognosis and regulates cell proliferation. In addition to shared pathways, SAMNet enables the identification of unique pathways targeted by miRNAs in cancer. Our preliminary analysis identifies the miR-148 family in adhesion and angiogenesis and the miR-9 family in apoptosis and cell death

CONCLUSIONS

We aim to expand our analysis to additional cancer types for which there is available DNase I hypersensitivity data in ENCODE. In this way we can probe network activity of miRNAs for each type/ subtype of cancer to determine how miRNA activity plays a role in cancer progression and overall prognosis.

REFERENCES

1. Birney,E. et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature, 447, 799–816.
2. Gosline,S.J. et al. (2012) SAMNet: a network-based approach to integrate multi-dimensional high throughput datasets. Integr. Biol.
3. TCGA Network,T.C.G.A. (2013) Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature, 499, 43–9.

Massively parallel reporter assays reveal splicing defects in 20% missense disease alleles

Rachel Soemedi[1], Kamil Cygan[1] and **Will Fairbrother[1]***

[1] MCB Department, Biomed Division, Brown University Providence RI 02806.

*To whom correspondence should be addressed: fairbrother@brown.edu

BACKGROUND

The Human Gene Mutation Database (HGMD) archives genomic variants that have been reported in the literature to cause hereditary disease(Krawczak et al. 2000). About 85% of these entries are missense mutations. While most these mutations are thought to disrupt gene function by altering the protein sequence, a significant fraction have been hypothesized to disrupt the pre-mRNA processing by either creating cryptic sites or disrupting exonic splicing control elements(Lim et al. ; Sterne-Weiler et al.) While biochemical analysis on isolated mutations have revealed numerous examples of missense mutations disrupting splicing, there has yet to be a global survey that describes the full extent of splicing mutations in human disease(Ars et al. 2000). Here, we report an unpublished survey from a dual in *vivo/in vitro* massively parallel reporter assay that records the extent of miss-splicing phenotypes in missense mutations. We describe over 5,000 wildtype mutant comparisons and correlate their splicing efficiency with a variety of other genomic and transcript features (e.g. conservation, predicted secondary structure). In addition we utilize these assays as a drug screening platform and demonstrate surprising effects of RNA binding pharmaceuticals and translational inhibitors on splicing.

RESULTS

Approximately 20% of missense mutations that cause disease significantly disrupt splicing in vivo and in vitro by a magnitude greater than 1.5 fold (Chi square statistic at FDR 5%). Splicing disruption is mostly detected indirectly as an allelic skew in the spliced product versus pre-mRNA however approximately 3% of missense mutations cause cryptic splice site usage. Unsurprisingly, variants that are common in the human population (i.e. SNPs) disrupt splicing to a much lower degree (i.e. 2%, no cryptic usage). The availability of thousands of splicing substrates allows for the correlation of transcript features with splicing efficiency. Secondary structure that sequesters the 3'ss and branchpoint region alters splicing in vitro but not in vivo. High throughput in vitro analysis establishes multiple distinct pathways for human intron splicing. The assay can also be used as a drug screening platform. For example, G418 treatment confer broad effects on splicing which result in the restoration of wild type function to 1% of loss-of-function splicing mutants.

CONCLUSIONS

Massively parallel reporter assays offer a platform for computational analysis of the determinants of splicing. Over 5,000 disease loci were compared to wild type in multiple, highly controlled contexts. We conclude that many mutations outside of canonical splice sites disrupt spicing and small compound screening targeting splicing defects offers a future avenue for therapy.

REFERENCES

1. Ars E, Serra E, Garcia J, Kruyer H, Gaona A et al. (2000) Mutations affecting mRNA splicing are the most common molecular defects in patients with neurofibromatosis type 1. Hum Mol Genet 9(2): 237-247.
2. Krawczak M, Ball EV, Fenton I, Stenson PD, Abeysinghe S et al. (2000) Human gene mutation database-a biomedical information and research resource. Hum Mutat 15(1): 45-51.
3. Lim KH, Ferraris L, Filloux ME, Raphael BJ, Fairbrother WG Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. Proc Natl Acad Sci U S A 108(27): 11093-11098.
4. Sterne-Weiler T, Howard J, Mort M, Cooper DN, Sanford JR Loss of exon identity is a common mechanism of human inherited disease. Genome Res 21(10): 1563-157

Origins and impacts of new exons

Chris Burge[1*], Jason Merkin[1,2], Ping Chen[3] and Sampsa Hautaniemi[3]

[1]Dept. of Biology, MIT, Cambridge MA 02142, [2]Current address: Novartis, Inc., Cambridge MA 02139, [3]Univ. of Helsinki, Helsinki Finland

*To whom correspondence should be addressed: cburge@mit.edu

BACKGROUND

Mammalian genes are typically broken into several protein-coding and non-coding exons, but the evolutionary origins and functions of new exons are not well understood. Here, we analyzed patterns of exon gain using deep cDNA sequencing data from several mammals and one bird, identifying thousands of species- and lineage-specific exons.

RESULTS

While exons conserved across mammals are mostly protein-coding and constitutively spliced, species-specific exons were mostly located in 5' untranslated regions and alternatively spliced. New exons most often derived from unique intronic sequence rather than repetitive elements, and were associated with upstream intronic deletions, increased nucleosome occupancy and RNA polymerase II pausing.

CONCLUSIONS

Surprisingly, exon gain was associated with increased gene expression, but only in tissues where the exon was included, suggesting that splicing enhances steady-state mRNA levels and that changes in splicing represent a major contributor to the evolution of gene expression.

Mechanistic Constraints for Modeling Translation from Genome-wide Measurements of Ribosome Occupancy

Arvind R. Subramaniam[1, 2] and Erin K. O'Shea[1, 2, 3, 4, *]

[1]FAS Center for Systems Biology, [2]Department of Molecular and Cellular Biology, [3]Department of Chemistry and Chemical Biology, and [4]Howard Hughes Medical Institute, Harvard University, Cambridge, MA 02138.

*To whom correspondence should be addressed: erin_oshea@harvard.edu

BACKGROUND

Protein synthesis involves initiation of translation by ribosomes on mRNA templates, and is followed by a series of elongation steps during which amino acids are added to the growing polypeptide chain. Initiation is the rate-limiting step of protein synthesis during nutrient-rich growth of cells. However, during perturbations such as nutrient starvation or overproduction of proteins, elongation rate of ribosomes can decrease significantly. Computational models of translation aim to predict the effect of changes in elongation rate on the overall rate of protein synthesis. Most computational models of translation are variants of the totally asymmetric simple exclusion process (TASEP), a paradigmatic model in non-equilibrium physics [1]. These models postulate that a decrease in ribosome elongation rate on an mRNA causes a traffic jam of trailing ribosomes, which then attenuates the rate of protein synthesis. However, the relevance of ribosome traffic jams to measured *in vivo* rates of protein synthesis remains to be characterized. We recently found that starvation for single amino acids in the bacterium *E. coli* caused over a 100-fold decrease in protein synthesis rate, that was caused by the presence of certain starvation-sensitive codons in the mRNA [2]. This starvation-specific effect of codons determined the growth response of *E. coli* subjected to amino acid downshift, and the biofilm-forming ability of the bacterium *Bacillus subtilis* [3]. To determine the mechanistic basis for this large effect of synonymous codons, we measured the genome-wide ribosome density on messenger RNAs in *E. coli* during starvation for single amino acids using the ribosome profiling method [4]. We then used this measurement to rigorously constrain a whole-cell model of translation in *E. coli*.

RESULTS

Starvation for single amino acids led to ribosome pausing, but only at a subset of codons cognate to the limiting amino acid. Prediction of this codon specificity required accounting for the differential kinetics of aminoacylation among tRNA isoacceptors in addition to a supply-demand balance [5]. Pausing of a single ribosome caused a traffic jam of trailing ribosomes in accordance with the TASEP model of translation. However the TASEP model was quantitatively inconsistent with the effect of pausing on ribosome occupancy along mRNAs. Incorporating translation abortion at ribosome pause sites into our model was sufficient to predict the effect of ribosome pausing on measured ribosome density. Further, evidence indicated that the transfer-messenger RNA (tmRNA/ssrA) is the primary effector of translation abortion at ribosome pause sites during amino acid starvation. Finally, quantitative analysis of abortion suggests an optimal balance between ribosome rescue and the synthesis of full proteins during stress.

CONCLUSIONS

Our work provides an experimentally-validated set of ingredients for systematic modeling of translation. More generally, it highlights the ability of deep-sequencing approaches to rigorously constrain computational models.

REFERENCES

1. C. T. MacDonald, J. H. Gibbs, Concerning the kinetics of polypeptide synthesis on polyribosomes, Biopolymers 7, 707–725 (1969).
2. A. R. Subramaniam, T. Pan, P. Cluzel, Environmental perturbations lift the degeneracy of the genetic code to regulate protein levels in bacteria, Proc. Natl. Acad. Sci. 110, 2419–2424 (2013).
3. A. R. Subramaniam et al., A serine sensor for multicellularity in a bacterium, eLife 2 (2013).
4. N. T. Ingolia, S. Ghaemmaghami, J. R. S. Newman, J. S. Weissman, Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling, Science 324, 218–223 (2009).
5. J. Elf, D. Nilsson, T. Tenson, M. Ehrenberg, Selective charging of tRNA isoacceptors explains patterns of codon usage, Science 300, 1718–1722 (2003).

RNAalignClust: Discovering ncRNA families by sequence-structure-based clustering of multiple sequence alignments

Milad Miladi[1*], Alexander Junge[2] , Fabrizio Costa[1], Stefan Seemann[2],
Jan Gorodkin[2], Rolf Backofen[1]
[1] Bioinformatics Group, University of Freiburg, Germany,
[2] Center for non-coding RNA in Technology and Health, University of Copenhagen, Frederiksberg C, Denmark

*To whom correspondence should be addressed: miladim@informatik.uni-freiburg.de

BACKGROUND

Clustering RNA molecules is an established approach to classify and functionally annotate non-coding RNAs (ncRNAs). The GraphClust pipeline [1] performs a graph kernel-based clustering to identify ncRNAs related in terms of sequence and local secondary structure.

We present *RNAalignClust*, a novel algorithm to cluster a set of multiple sequence alignments aiming to discover families of ncRNAs. Each multiple alignment contains a group of RNAs derived from a common ancestor. This evolutionary relationship allows to infer conserved sequence and structure commonalities in each group. Previous approaches designed to cluster unaligned RNA sequences do not utilize this information.

RNAalignClust furthermore extends *GraphClust* by an extended graph kernel that scores both global and local structural similarities and also by producing the final result in a single non-iterative clustering.

RESULTS

Given a set of multiple alignments of evolutionary related sequences, *RNAalignClust* computes (sub)optimal structures for each alignment using highly conserved base pairs as constraints. Similar structural alignments are then identified using a fast nearest neighbors search approach based on hashing and graph kernel similarity notions.

To benchmark our algorithm, ncRNA families stored in the Rfam database were split into several subalignments and recovered using *RNAalignClust*.

CONCLUSIONS

RNAalignClust is a novel approach to identify families of ncRNAs related in terms of both sequence and structure from a given set of multiple alignments. It is able to take alternative secondary structures into account while still achieving a linear-time performance. This enables the analysis of complete genomes and transcriptomes within hours of computation time.

REFERENCES

1. Heyne S, Costa F, Rose D, Backofen R. GraphClust: alignment-free structural clustering of local RNA secondary structures, Bioinformatics2012, 28: i224–i232.

Specialized Reference Transcriptomes for Sailfishe

Stephen Mount[1*], Julien Buchbinder[1], Mary Same[1], Michael Kleyman[1],
Rob Patro[2] and Carl Kingsford[2]
[1]Department of Cell Biology and Molecular Genetics and Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland, USA. [2]Lane Center for Computational Biology, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania,USA
*To whom correspondence should be addressed: smount@umd.edu

BACKGROUND

Sailfish is a computational method for quantifying the abundance of transcripts (previously-annotated RNA isoforms) from RNA-seq data that works by assigning k-mers from reads to specific transcripts. Because Sailfish does not map reads, which is a time-consuming step in all current methods, it provides quantification estimates often more than 20 times faster than those methods. We find [1] using both simulated and real data that this can be accomplished without loss of accuracy.

A Sailfish index is built from a particular set of reference transcripts and needs to be rebuilt only when that reference or the value of k changes. The quantification phase of Sailfish applies an expectation-maximization (EM) procedure to determine maximum likelihood estimates for the relative abundance of each transcript in the reference, measured in Reads Per Kilobase per Million mapped reads (RPKM), Transcripts Per Million (TPM) and K-mers per Kilobase per Million mapped k-mers (KPKM).

Sailfish is free and open-source software and is available at www.cs.cmu.edu/~ckingsf/software/sailfish (short url: ongen.us/SFish).

RESULTS

The speed of Sailfish makes routine reanalysis of data from archived experiments possible in the face of new genome annotations such as newly discovered isoforms, RNA editing sites, polymorphisms, etc.; facilitates the inclusion of repeated or foreign transcripts (e.g. retrotransposons and parasites); and makes it practical to revisit data in order to address specific questions (such as whether or not the data support an isoform hypothesized to exist).

We are exploring the application of specialized reference transcriptomes to a number of data sets.
1) **Tiled reference transcriptomes** can be used in the case where genome sequence is available, but lacks annotation, or has unreliable annotation. Tiles that overlap by k-1 place every genome k-mer in one and only one tile. Sailfish can use such a tiling as the reference transcriptome.
2) **Transcript segment annotation files** that substitute individual exons, junctions of length 2k-2 and alternatively spliced intervals for full transcripts. Preliminary analysis of a Drosophila data set (the time course of a viral infection) shows qualitatively similar results to use of a full transcription annotation file. The use of transcript segments should facilitate the detection and quantification of novel or aberrant isoforms.

We will also present further work exploring alternative choices of k in specialized contexts to trade-off between the ambiguity of the kmer origin and sensitivity to polymorphisms and to sequencing errors. k=20 is sufficient to uniquely identify 99.914% of unique (2k-2)mers in the Drosophila genome, and 99.516% of unique (2k-2)mers in the human transcriptome.

CONCLUSIONS

The speed of Sailfish allows re-analysis of RNAseq data with customized reference transcriptomes.
The default value of k (20) is sufficient to uniquely specify the vast majority of transcript segments (2k-2).

REFERENCES

1. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nature Biotechnology (2014).

The eukaryotic RNA-protein interaction code

Quaid Morris [1,2,3,4,5*], Debashish Ray[1], Shankar Vembu[1], Kate B. Cook[3], Hilal Kazan[2%], Xiao Li[3], Matthew T. Weirauch[1#], Hamed S. Najafabadi[1,4], and Timothy R. Hughes[1,3]

[1]Donnelly Centre, University of Toronto, 160 College St, M5S 3E1, [2]Department of Computer Science, University of Toronto, [3]Department of Molecular Genetics, University of Toronto, [4]Department of Electrical and Computer Engineering, University of Toronto, [5]Visiting researcher, Gene Regulation, Stem Cells and Cancer group in the Centre for Genomic Regulation, Barcelona, Catalonia, Spain

[%]Present address: Dept. Computer Science, Antalya International University, Turkey

[#]Present address: Center for Autoimmune Genomics and Etiology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio

*To whom correspondence should be addressed: quaid.morris@utoronto.ca

BACKGROUND

Between 30-60% of the variability in human protein levels is explained by cytoplasmic regulatory processes and is likely controlled by cis-regulatory elements residing in mature mRNA sequence. As such, regulation by RNA-binding proteins (RBPs) and regulatory ncRNAs in mammals contributes as much to gene expression levels as transcription factors but the search space for their cis-regulatory elements is 10-100x smaller. If the RBP sequence binding preferences were known, computational modeling of these regulatory processes would be extremely feasible.

RESULTS

We have recently biochemically-measured RNA binding preferences for more than 200 RBPs and, because RBP RNA-specificity is highly conserved, we were able to infer motifs for nearly 5,000 more RBPs by homology1. The motifs are available here: http://cisbp-rna.ccbr.utoronto.ca/. Our measured or inferred motifs include nearly half of the known or predicted sequence-specific RBP complement of humans, 30% of the predicted metazoan RBPs and 10% of all predicted eukaryotes RBPs. We have also developed computational methods to infer and summarize RBP sequence and structure binding preferences from these in vitro binding data as well as increasingly available in vivo binding data.

CONCLUSIONS

I will discuss our efforts to define RNA-binding specificities for the rest of the eukaryotic RBP complement, and to use these motif models to infer RBP function and building regulatory models for post-transcriptional regulation in metazoa.

REFERENCES

1. Ray et al, Nature. 2013 Jul 11;499(7457):172-7

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

Courtney E. French[1*], Aparna Chhibber[2], Eric R. Gamazon[3], Sook Wah Yee[2], Xiang Qin[4], Elizabeth Theusch[5], Amy Webb[6], Scott Weiss[7], Marisa W. Medina[5], Ronald M. Krauss[5], 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] University of Chicago, Chicago, IL [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/Harvard Medical School, Boston,MA

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

BACKGROUND

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

RESULTS

Using 90 samples from 5 major tissues (liver, kidney, adipose, heart, and lymphoblastoid cell lines [LCLs]) of pharmacologic importance, some with extensive pharmacogenomic phenotyping, we performed RNA sequencing. The data were analyzed for expression quantification, splice junction analysis, and transcript reconstruction. We utilized the JuncBASE pipeline developed by members of our consortium to identify and classify splicing events. In samples from heart, kidney, liver and adipose tissues, similar numbers of transcripts and genes were detected; however, notable differences in expression levels of important pharmacogenes (see http://www.pharmgkb.org/search/annotatedGene/) across the various tissues were observed (FDR < 0.05). For example, CYP enzymes (e.g., CYP2C19 and CYP2D6) were highly expressed in the liver with low expression in other tissues. Other important drug metabolizing enzymes such as DPYD and TPMT showed more balanced gene expression patterns across the tissues. We 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 individual samples of the same tissue. For example, we found evidence of a novel alternative last exon for SLC22A7, a gene involved in transport of various drugs, which is variably spliced in liver between individuals. In addition, given the importance of LCLs as a pre-clinical model for human genetic studies, we systematically investigated differential expression and splicing between LCLs and the other tissues. LCLs do not express many genes and splice variants present in the physiological tissues, but do express splice variants of genes of pharmacological interest that are not observed in these primary tissues.

CONCLUSIONS

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

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

Anna Desai[1], **Courtney E. French[1*]**, and Steven. E. Brenner[1]
[1] University of California, Berkeley, CA
*To whom correspondence should be addressed: cfrench@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, 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 and PTB / nPTB are regulated in the same manner.

RESULTS

After an extensive literature search, we generated a splicing factor regulatory network that encompasses all current knowledge of splice factor regulatory interactions including the known extent of NMD regulation coupled with alternative splicing. The currently available data shows that majority of the SR proteins and a few hnRNP splicing factors are known to be regulated via alternative splicing coupled with NMD. Since all the SR proteins and many hnRNP splicing factors produce isoforms degraded by NMD, we expect that this mode of regulation is prevalent among all splicing factors. In addition, CLIP-seq data reveals yet more extensive splicing factor-mRNA interactions, providing an additional hint that many more splicing factors might be regulated via alternative splicing coupled with NMD by other splice factors themselves.

Comparison of D. melanogaster and C. elegans Developmental Stages, Tissues, and Cells by modENCODE RNA-Seq data

Jingyi Jessica Li[1,2*], Haiyan Huang[2], Peter J. Bickel[2], and Steven E. Brenner[3]
[1] current address: Department of Statistics, UCLA, [2] Department of Statistics, UC Berkeley, [3] Department of Plant and Microbial Biology, UC Berkeley

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

BACKGROUND

Drosophila melanogaster and *Caenorhabditis elegans* are model systems for studying molecular, cellular and developmental processes in animals. As morphologically different and evolutionarily distant organisms separated by as much as 600 million years in evolution, *D. melanogaster* and *C. elegans* have striking differences in cell differentiation and whole-organism developmental biology. Despite these differences, many individual conserved mechanisms have been observed in *D. melanogaster* and *C. elegans*. Indeed, the conservation of embryonic development in animal species has been a unifying concept since von Baer's observations in the 19th-century, and the conservation of developmental genes between animals has been long studied in evolutionary developmental biology. However, genome wide analyses that have systematically characterized the conservation in gene expression during development 8are lacking.

RESULTS

Here we report a statistical study to discover transcriptome similarity of developmental stages from *D. melanogaster* and *C. elegans* using modENCODE [1-2] RNA-Seq data. We focus on "stage-associated genes" that capture specific transcriptional activities in each stage, and use them to map pairwise stages within and between the two species by a hypergeometric test. Within each species, temporally adjacent stages exhibit high transcriptome similarity as expected. Additionally, fly female adults and worm adults are mapped with fly and worm embryos respectively due to maternal gene expression. Between fly and worm, an unexpected strong collinearity is observed in the time course from early embryos to late larvae. Moreover, a second parallel pattern is found between fly prepupae through adults and worm late embryos through adults, consistent with the second large wave of cell proliferation and differentiation in the fly life cycle. The results also indicate a partially duplicated developmental program in fly.

CONCLUSIONS

Our results constitute the first comprehensive comparison between *D. melanogaster* and *C. elegans* developmental time courses, and provide new insights into similarities in the development of these two species. We use an analogous approach to compare tissues and cells from fly and worm and also compare them with the developmental stages. Findings include strong transcriptome similarity of fly cell lines relative to dissected tissues, clustering of fly adult tissues by origin regardless of sex and age, and clustering of worm tissues and dissected cells by developmental stage. Gene Ontology analysis provides strong functional support for our mapping results, and gives a uniquely detailed annotation of the biological functions enriched in different stages, as well as tissues and cells. Finally, we show that commonly used correlation analyses are not able to detect many of the mappings found by our method.

REFERENCES

1. Celniker SE et al. 2009. Unlocking the secrets of the genome. Nature **459**(7249): 927-930.
2. Gerstein MB et al. 2010. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science **330**(6012): 1775-1787.

Systems Level Analysis of Alternative Pre-mRNA Splicing: The roles of RNA structure and RNA Chaperone Proteins

Yeon Lee[1]*, Ming Hammond[2], and Don Rio[1,3]
[1] University of California, Berkeley, California Institute for Quantitative Biosciences, Berkeley, CA, 94720, [2] University of California, Berkeley, Chemistry, Berkeley, CA, 94720 and [3] University of California, Berkeley, Molecular and Cell Biology, Berkeley, CA, 94720

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

BACKGROUND

Alternative pre-mRNA splicing is one of the major mechanisms utilized by metazoans to regulate gene expression and to increase the functional diversity of the eukaryotic proteomes. In humans, ~95% of multi-exon genes are alternatively spliced and these RNA processing events have implications for health and disease since disease gene mutations that affect the splicing process result in human genetic disorders.

Alternative pre-mRNA splicing is regulated both by RNA-binding proteins that interact with pre-mRNAs and by RNA secondary structures. Recent studies have indicated that RNA secondary structures can play an important, and previously underappreciated, role in the regulation of alternative splicing. RNA chaperone proteins are known to alter RNA secondary structure in vitro through RNA-RNA annealing or unwinding and by RNP remodeling. These RNA chaperones aid in RNA folding, but have also been shown to be involved in splicing and transcription *in vivo*.

Although much progress has been made in understanding different alternative splicing mechanisms on an individual transcript or gene basis, much remains to be learned including how RNA structure and/or RNA chaperones affects alternative splicing on a global transcriptome-wide level. We aim to systemically link *cis*-regulatory elements in pre-mRNAs to RNA structural features and protein binding sites that control alternative pre-mRNA splicing *in vivo*.

RESULTS

Transcriptome-wide binding maps of two RNA chaperone proteins, hnRNP A1 and the p68/DDX5 RNA helicase were determined by iCLIP-seq[1]. RNA in complex with RNA binding proteins from UV-irradiated cells were subjected to immunoprecipitation. Protein-RNA covalent complexes from corresponding bands were recovered following autoradiography and SDS-PAGE, isolated RNA were ligated to an RNA-adapter at the 3' end, RT-PCR amplified, and sequenced on the Illumina HiSeq.

The siRNA-mediated k/down condition was optimized in both K562 and GM12878 cell lines, and strand-specific RNA-seq libraries were prepared from these cells. Once the raw reads were received, these reads were mapped against human genome using Tophat[2]. Different splicing events, including alternative 3' or 5' splice site, mutually exclusive exons, retained introns, or skipped exons (or cassette exons), were then analyzed by MISO[3] for either scr siRNA control samples or hnRNPA1 or p68/DDX5 k/down samples.

Over 175,000 binding sites of hnRNP A1 was mapped and 3600 statistically significant differential splicing events were detected upon hnRNP A1 k/down in K562 cells. Additionally, over 5900 differential splicing events were detected uponp68/DDX5 k/down.

CONCLUSIONS

We aim to systemically link *cis*-regulatory elements in pre-mRNAs to RNA structural features and protein binding sites that control alternative pre-mRNA splicing *in vivo*. The RNA binding and chaperone activities of two RNA chaperone proteins, hnRNP A1 and the p68/DDX5 will be compared to transcriptome-wide changes in RNA structure using chemical probing information and alternative splicing patterns with RNAi-knockdowns of these factors using high-throughput cDNA sequence analyses.

REFERENCES

1. Julian Konig, K.Z., Gregor Rot, Tomaz Curk, Melis Kaylkci, Blaz Zupan, Daniel J. Turner, Nicholas M. Luscombe, Jernej Ule, *iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution.* Nat Struct Mol Biol. , 2010. **17**(7): p. 909-915.
2. Trapnell, C., L. Pachter, and S. Salzberg, *TopHat: discovering splice junctions with RNA-seq.* Bioinformatics, 2009. **25**(9): p. 1105-1111.
3. Yarden Katz, E.T.W., Edoardo M Airoldi, Christopher B Burge, *Analysis and design of RNA seqencing experiments for identifying isoform regulation.* Nature Methods, 2010. **7**: p. 1009-1015.

The Evolution and Distribution of Alu Elements in Long Noncoding RNAs and mRNAs

Eugene Z Kim[1], Adam R Wespiser[1], and Daniel R Caffrey[1*]

[1] Department of Medicine, University of Massachusetts Medical School, Worcester USA

*To whom correspondence should be addressed: daniel.caffrey@gmail.com

BACKGROUND

The primate-specific Alu element is the most abundant mobile element in the human genome. Approximately 75% of the human genome is transcribed and many of these spliced transcripts contain Alu elements. The majority of exonized Alu elements are located in long noncoding RNAs (lncRNAs) and the untranslated regions of mRNA. Recently, some exonized Alu elements have been shown to perform important molecular functions in both lncRNAs and mRNAs.

RESULTS

To further assess the potential for Alu elements to be repurposed as functional RNA domains, we investigated distribution and evolution of Alu elements in spliced transcripts. Our analysis revealed that the presence of Alu elements was less common in mRNA and the major classes of lncRNA (lincRNAs, antisense RNAs, and processed RNAs) than their corresponding controls. However, the percentage of Alu-containing RNAs encoded at some chromosomes was significantly greater than their controls. For example, there was a significant enrichment of Alu elements in all RNA types encoded by chromosome 19, and the enrichment of Alu-containing mRNA was associated with intrachromosomal gene duplications. Unlike mRNA, which had equal proportions of reverse and forward-oriented Alu elements that were primarily near the 3'-end, lncRNAs had a preference for reverse-oriented Alu elements that were not biased toward the 3'-end. Indeed, the enrichment of reverse-oriented Alu elements often corresponded to complete internal exons in lncRNAs. Both mRNAs and lncRNAs contained Alu elements that were primarily exonized into non-random domains that corresponded to complete dimeric Alu elements, left monomers, and right monomers. In particular, the proportion of dimeric Alu elements was significantly greater in mRNA and some classes of lncRNA relative to their non-transcribed controls. Furthermore, dimeric Alu elements belonging to the major Alu subfamilies were generally under greater evolutionary constraint in mRNAs and lncRNAs relative to their controls.

CONCLUSIONS

In summary, the non-random exonization of specific Alu domains into transcripts and their evolutionary constraints suggest that they are capable of forming stable RNA structures that potentially act as modular functional domains.

Boosting RNA-Seq analysis of alternative splicing via high-precision exon junction detection

Alberto Gatto[1*], Fátima Sánchez-Cabo[2], Carlos Torroja[2] and Enrique Lara-Pezzi[1]
[1] Cardiovascular Development and Repair Department, Centro Nacional de Investigaciones Cardiovasculares, Madrid, Spain [2] Bioinformatics Unit, Centro Nacional de Investigaciones Cardiovasculares, Madrid, Spain
*To whom correspondence should be addressed: agatto@cnic.es

BACKGROUND

Analysis of alternative splicing (AS) is one of the most challenging applications of RNA-Seq. The basic and most critical step in any analysis workflow is read mapping, an especially difficult task for reads spanning (possibly unannotated) exon-exon junctions. In spite of the large number of available tools, recent benchmarking efforts by the RGASP3 consortium showed that splice junctions false discovery rates, annotation usage and multi-mapping reads remain critical issues in spliced read alignment [1]. To address this problem, we aimed at dissecting systematically the impact of design and method on the mapping, detection and quantification of junction-spanning reads from RNA-Seq data. Based on the results, we propose a pipeline to bridge the gap between the superior mapping accuracy of transcriptome-first alignment and the low false-positive rates of intron-centric approaches.

RESULTS

Using synthetic data sets and publicly available RNA-Seq experiments, we evaluated the splice-site level mapping, detection and quantification performance from a selection of alignment solutions: TopHat2, GSNAP, STAR, OLego and SOAPsplice. These aligners offer different approaches to mapping (exon-first, seed-and-extend and multi-seed), indexing (hash tables and FM-index), annotation use (transcript-based, intron-based and ab initio) and de novo splice-site prediction [2]. We show that transcript-based aligners achieve better mapping and quantification accuracy but introduce a high number of spurious gapped alignments, while intron-centric approaches attain superior detection precision at the expense of mapping and quantification performance. We propose a simple pipeline to fulfil both prerogatives by constraining the splice junction detection problem at the post-processing stage. Our strategy couples TopHat2 to a novel method (FineSplice) that enables to identify unreliable gapped alignments and filter out false-positive junctions via semi-supervised logistic regression. This approach allows capitalising on the higher mapping accuracy of transcriptome-first aligners while reducing the false discovery rate up to 10-fold at small loss in sensitivity, and improves quantification by rescuing multi-mapping reads with a unique location after filtering [3].

CONCLUSIONS

We present pros and cons of different strategies for spliced read mapping and introduce a new method, FineSplice, aiming to close the gap between the superior mapping accuracy of transcript-based solutions and the better precision of intron-centric approaches. Our pipeline conjugates an efficient mapping solution, TopHat2, with a novel anomaly detection scheme that allows effective elimination of spurious junction hits arising from artifactual alignments. FineSplice enhances precision without sacrificing sensitivity, and improves read count estimates by rescuing multiple mapping reads. This ultimately provides an effective way to accurately detect and quantify junction-spanning reads. We show in practice how this can benefit the analysis of alternative splicing events and avoid improper estimates of exon inclusion levels.

REFERENCES

1. Engström, P.G., Steijger, T., Sipos, B., Grant, G.R., Kahles, A., The RGASP Consortium, Rätsch, G., Goldman, N., Hubbard, T.J., Harrow, J., et al. (2013) Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat Meth*, **10**, 1185–1191.
2. Alamancos, G.P., Agirre, E. and Eyras, E. (2014) Methods to Study Splicing from High-Throughput RNA Sequencing Data. In *Spliceosomal Pre-mRNA Splicing*, Methods in Molecular Biology.Vol. 1126, pp. 357–397.
3. Gatto, A., Torroja-Fungairiño, C., Mazzarotto, F., Cook, S.A., Barton, P.J.R., Sánchez-Cabo, F. and Lara-Pezzi, E. (2014) FineSplice, enhanced splice junction detection and quantification: a novel pipeline based on the assessment of diverse RNA-Seq alignment solutions. *Nucl. Acids Res.*, **42**, e71–e71.

Understanding the structure-function relationship in group II introns through computational modeling and design

Srinivas Somarowthu[1], Michal Legiewicz[1], Kevin Keating[1] and Anna Marie Pyle[1,2,3]

[1]Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, CT, 06511, USA
[2]Department of Chemistry, Yale University, New Haven, CT, 06511, USA
[3]Howard Hughes Medical Institute, Chevy Chase, Maryland, 20815, USA
*To whom correspondence should be addressed: anna.pyle@yale.edu

BACKGROUND

Group II introns are large RNAs that acts as self-splicing ribozymes and retroelements, found in bacteria, archaea and eukaryotes [1]. Understanding the structure-function relationship of group II introns is of great interest due to their evolutionary relationship with the eukaryotic spliceosome and their role in gene expression and genomic organization in many organisms. Group II introns are structurally grouped into three distinct classes (Group IIA, IIB, and IIC) [2]. The group IIB introns are of particular interest because they are believed to be closely related to the spliceosome, they are structurally complex and display exceptional sequence-specificity. Currently, there is no structure available for a group IIB intron; however, there are crystal structures available for group IIC intron [3]. Group IIC introns are structurally primitive and lack many domains of group II introns. However, they represent a starting point to visualize the catalytic core of group II introns and also serve as a template to model other group II introns through homology modeling.

RESULTS

Here, we modeled the 3D structure of the ai5γ group IIB intron using a combination of homology and *de novo* modeling methods [4]. Modeling was performed in three steps. In the first step, the core region common in group IIB and IIC introns was generated through homology modeling using a crystal structure of the group IIC intron from *Oceanobacillus iheyensis* (PDBID: 3IGI) [3] as the template. Next, all additional domains specific to ai5γ were modeled de novo using MC-Sym and docked onto the core structure. Finally, the backbone of the model was refined with the newly developed plugin in RCrane and the whole structure was energy minimized with AMBER. The resulting model was validated experimentally using RNA structure probing methods. The model provides major insights into the mechanism and regulation of splicing, such as the position of the branch-site before and after the second step of splicing, and the location of subdomains that control target specificity.

CONCLUSIONS

In conclusion, we show that it is possible to model large RNAs even from remote homologs. We present the 3D model of ai5γ IIB intron that provides structural insights in to group II introns and explains functional differences between group IIC introns and more structurally evolved group IIB introns. Further, the model now serves as a guide to design the structural motifs and long-range interactions that are not present in group IIC intron.

REFERENCES

1. Ferat, J.L. and F. Michel, *Group II self-splicing introns in bacteria.* Nature, 1993. **364**(6435): p. 358-61
2. Toor, N., G. Hausner, and S. Zimmerly, *Coevolution of group II intron RNA structures with their intron-encoded reverse transcriptases.* RNA, 2001. **7**(8): p. 1142-52.
3. Toor, N., et al., *Crystal structure of a self-spliced group II intron.* Science, 2008. **320**(5872): p. 77-82.
4. Somarowthu, S., et al., *Visualizing the ai5gamma group IIB intron.* Nucleic Acids Res, 2014. **42**(3): p1947-58.

The isoform with the most conserved protein features is the major isoform and it is the major protein isoform in the cell.

Iakes Ezkurdia[1], Jose Manuel Rodriguez[2], Alfonso Valencia[2] and Michael Tress[2*]

[1] Centro Nacional de Investigaciones Oncolgicas, Madrid, Spain [2] Spanish National Cancer Research Centre (CNIO), Madrid, Spain

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

BACKGROUND

Studies repeatedly show the expression of a wide range of alternatively spliced transcripts in eukaryotic cells, but the equivalent diversity of cellular proteins remains elusive. Although a number of studies have addressed the identification of isoforms at the protein level, the variety of methods and criteria used for splice isoform identification have generated confusing and contradictory results.

RESULTS

Here we analysed the peptides detected in 7 large-scale human proteomics experiments with a rigorous strategy for peptide search and identification. We map peptides to almost 60% of the protein coding genes in the human genome, and find convincing evidence for alternatively spliced proteins in 232 genes. Many of these alternative protein isoforms are only subtly different from the main isoform, and few of the splicing events associated with these alternative isoforms would disrupt structural or functional domains. For the vast majority of the genes identified in the proteomics experiments the peptide evidence mapped to a single protein isoform. We found that the dominant isoform we detected in the proteomics experiments was also almost always the isoform with the most conserved protein features.

Parallel experiments carried out with mouse proteomic datasets confirmed a similar pattern, with the prevalence of a single dominant isoform and the conservation of functional domains for those few alternative splicing events detected at the protein level. Many of the alternative isoforms that we identified in mouse were also found in the human experiments, and almost half of these were generated by the splicing of mutually exclusive homologous exons.

CONCLUSIONS

Our results suggest that the expression of alternative splice isoforms is subject to some level of cellular control, that most genes have a single dominant protein isoform and that this isoform can be determined from an analysis of the structure, function and conservation of the annotated splice variants.

Promoter sequences influence cytoplasmic localization and translation of mRNAs during nutrient limitation

Brian M. Zid[1*] and Erin K. O'Shea[1,2]

[1] Faculty of Arts and Sciences Center for Systems Biology, Harvard University, Cambridge, MA 02138, [2] Howard Hughes Medical Institute, Harvard University, Cambridge, MA 02138

*To whom correspondence should be addressed: zid@fas.harvard.edu

BACKGROUND

A universal feature of the response to stress and nutrient limitation is transcriptional upregulation of genes encoding proteins important for survival. Interestingly, under many of these conditions overall protein synthesis levels are reduced, thereby dampening the stress response at the level of protein expression[1]. For example, during glucose starvation in yeast, translation is rapidly and reversibly repressed, yet transcription of many stress- and glucose-repressed genes is increased[2,3].

RESULTS

Using ribosome profiling and microscopy, we found that this transcriptionally upregulated gene set consists of two classes: (1) one producing mRNAs that are translated during glucose limitation and are diffusely localized in the cytoplasm – this class includes many heat shock protein mRNAs; and (2) another producing mRNAs that are not efficiently translated during glucose limitation and are concentrated in foci that co-localize with P bodies and stress granules – this class is enriched for glucose metabolism mRNAs. Remarkably, the information specifying differential localization and protein production of these two classes of mRNAs is encoded in the promoter sequence – promoter responsiveness to heat shock factor (Hsf1) specifies diffuse cytoplasmic localization and higher protein production upon glucose starvation, whereas different promoter elements upstream of genes encoding glucose metabolism mRNAs that produce less protein direct these mRNAs to RNA granules under glucose starvation.

CONCLUSIONS

Thus, promoter sequences and transcription factor binding can influence not only mRNA levels, but also subcellular localization of mRNAs and the efficiency with which they are translated, enabling cells to tailor protein production to environmental conditions.

REFERENCES

1. Simpson, C. E., & Ashe, M. P. (2012). Adaptation to stress in yeast: to translate or not? Biochemical Society Transactions, 40(4), 794–9. doi:10.1042/BST20120078
2. Arribere, J. A., Doudna, J. A., & Gilbert, W. V. (2011). Reconsidering Movement of Eukaryotic mRNAs between Polysomes and P Bodies. Molecular Cell, 44(5), 745–758. doi:10.1016/j.molcel.2011.09.019
3. Ashe, M. P., Long, S. K. De, & Sachs, A. B. (2000). Glucose Depletion Rapidly Inhibits Translation Initiation in Yeast. Molecular Biology of the Cell, 11(March), 833– 848.

Charting the landscape of functional and 'ambient' splicing in immune-cell activations

Manikandan Narayanan[1], Andrew Martins[1], Zachary Benet[2], and John Tsang[1*]

[1] Systems Genomics and Bioinformatics Unit, [2] Signaling Systems Unit, Laboratory of Systems Biology, National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health (NIH)

* To whom correspondence should be addressed: john.tsang@nih.gov

BACKGROUND

Functionally important alternative splicing events are highly prevalent across tissues or cell types (e.g., [1]), and it remains intriguing to figure the extent to which cells utilize splicing over expression regulation in responding to distinct stimuli. Low-frequency splicing events at evolutionarily non-conserved splice sites have also been observed in deep sequencing data (e.g., [2]). While many of them likely originate from background or 'ambient' splicing with little functional impact (e.g., [2]), it is less clear whether/how the frequency of these events are regulated across different conditions or pathways. In particular, an integrative assessment of the cellular regulation of both functional and 'ambient' splicing in distinct environments have not been well examined; and immune cells such as macrophages with their ability to mount dramatic responses to environmental signals are excellent models to explore this question.

RESULTS

We generated deep RNAseq data from resting vs. inflammatory RAW 264.7 mouse macrophage cells and employed appropriate statistical analyses to define, quantify and compare functional/ambient splicing events between the two conditions. We discovered a class of regulated functional splicing events among signaling, membrane and vesicle associated genes, wherein the overall transcriptional output at a gene locus was not regulated between two conditions, but the proportional shunting of splicing along different isoforms of the gene was. Quantifying ambient splicing showed it to be enriched at surprisingly low levels among metabolic pathway genes. More interestingly, certain pathways showed significantly altered ambient splicing levels under immune activation. We are exploring these observations to enquire if functional and ambient splicing could be driven by shared or distinct splicing regulators.

CONCLUSIONS

Our analysis delineated the extent to which resting macrophages respond to inflammatory stimulus by up/down-regulating specific genes versus specific isoforms of already expressed genes, and revealed surprising condition/ pathway-specific distribution of low-frequency ambient splicing events.

REFERENCES

1. Ergun A, Doran G, Costello JC, Paik HH, Collins JJ, Mathis D, Benoist C, and ImmGen Consortium. Differential splicing across immune system lineages. PNAS **110** (2013).
2. Pickrell JK, Pai AA, Gilad Y, and Pritchard JK. Noisy splicing drives mRNA isoform diversity in human cells. PLoS Genetics **6** (2010).

This research was supported by the Intramural Research Program of the NIH/NIAID. We thank Dr. Iain Fraser for valuable inputs, all members of Dr. John Tsang's group for feedback, and Patrick Burr and NIAID HPC/BCBB team for help with Illumina sequencing run and data preprocessing pipelines.

Differential patterns of RISC loading of microRNAs: Implications in aging and neurodegeneration

Andrey Grigoriev[1*], and Nancy Bonini[2*]

[1] Department of Biology, Center for Computational and Integrative Biology, Rutgers University, Camden, NJ 08102, USA, [2] Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA

*To whom correspondence should be addressed: andrey.grigoriev@rutgers.edu; nbonini@sas.upenn.edu

BACKGROUND

Acting via RNA-induced silencing complexes (RISCs) with Argonaute/Ago proteins, small interfering RNAs (siRNAs) and microRNAs (miRNAs) posttranscriptionally regulate gene expression. Traditionally, involvement of small RNAs in development/aging is measured by changes in their abundance with time.

RESULTS

A recent exploration of the diversity of miRNA in Drosophila performed by our labs revealed a striking diversity of isoform length patterns and loading into different RISC complexes with age [1].When we analyzed not only the abundance but also miRNA isoform length, modifications and RISC partitioning, a complex picture of intricately linked changes emerged. We observed an increase in the abundance of the longest isoform of certain miRNAs with age and protection of these isoforms at the 3'end by methylation. Although most miRNAs are loaded into Ago1-RISC and remain unmodified, a subset are loaded into Ago2-RISC and become 2'-O-methylated and protected from 3'-to-5' exonuclease trimming with age [2,3]. Thus aging appears to significantly change the partitioning of miRNAs into the different RISC complexes affecting the function of the latter.

To address the biological significance of this shift in loading of miRNAs, we examined Hen1 and Ago2 mutants lacking 2'-O-methylation. These animals showed dramatic brain degeneration and reduced lifespan, suggesting a potential impact of this increase of 2'-O-methylation of small RNAs on aging and neurodegenerative processes. Analysis of microarray data on aging flies further revealed different regulation patterns for genes affected by miRNAs loaded into Ago1-RISC and Ago2-RISC, with genes preferentially downregulated by the latter complex.

CONCLUSIONS

This novel role of miRNAs in aging also highlights how much remains to be understood in the rapidly evolving view of the miRNA function and biogenesis. Might an aging organism be adjusting the efficiency of target gene expression for ongoing or upcoming age-associated stresses by shifting miRNAs towards Ago2-RISC from Ago1-RISC? Or could the accumulation of miRNAs in the Ago2-RISC be a by-product of the increased stability of their 2'-O-methylated isoforms? This phenomenon can also be viewed in the context of an age-associated increase of transposon expression, neuronal decline and shorter lifespan noted in Ago2 mutants [4]. Does the accumulation of miRNAs in Ago2-RISC impact the efficiency of these and other biological phenomena that depend on Ago2-RISC, for example, the action of siRNAs and endo-siRNAs in the aging organism?

REFERENCES

1. Abe, M., Naqvi, A., Hendriks, G. J., Feltzin, V., Zhu, Y., Grigoriev, A., & Bonini, N. M. Impact of age-associated increase in 2′-O-methylation of miRNAs on aging and neurodegeneration in Drosophila. Genes & Dev, **28** (2014).
2. Czech B, Zhou R, Erlich Y, Brennecke J, Binari R, Villalta C, Gordon A, Perrimon N, Hannon GJ. 2009. Hierarchical rules for Argonaute loading in Drosophila. Mol Cell **36** (2009).
3. Ghildiyal M, Xu J, Seitz H, Weng Z, Zamore PD. 2010. Sorting of Drosophila small silencing RNAs partitions microRNA* strands into the RNA interference pathway. RNA **16** (2010).
4. Li W, Prazak L, Chatterjee N, Gruninger S, Krug L, Theodorou D, Dubnau J. Activation of transposable elements during aging and neuronal decline in Drosophila. Nat Neurosci **16** (2013).

GenTrAn: a new tool for de-novo transposon structural variant detection from single-end deep-sequencing data

Reazur Rahman[1*], Yuliya Sytnikova[2], and Nelson C. Lau[2]
[1]Brandeis University, Department of Biology, and Rosenstiel Center for Biomedical Research, [2]Brandeis University, Department of Biology, and Rosenstiel Center for Biomedical Research
*To whom correspondence should be addressed: reazur@brandeis.edu

BACKGROUND

Tansposons are major structural variants (SVs) in animal genomes. In cancer and human biology, there is a need to determine new transposon SVs beyond the tremendous load of existing transposons (>45% of the human genome). Most current efforts to discover transposon SVs rely on Paired-End (PE) reads from genome deep-sequencing, but the greater costs of PE reads compared to Single-End (SE) reads (the standard form of genome deep-sequencing) motivated us to develop a new bioinformatics tool called GenTrAn (Genome Transposon Analyzer).

RESULTS

GenTrAn discovers de-novo transposon SVs with high sensitivity and specificity, by scanning SE read libraries with a hybrid approach of split-read mapping and then filtering with various quality criteria. Importantly, the transposon SV sites that GenTrAn identifies display target site duplications indicative of a recent transposition event, and point to precise genomic coordinates that enable discrimination of SVs that disrupt coding gene exons versus less-disruptive intronic insertions. We demonstrate the efficacy of our tool by discovering the genome-wide distributions of transposon SVs in four different Drosophila melanogaster cell lines.

CONCLUSIONS

GenTrAn showed that transposon SV landscapes can be surprisingly diverse even in a natural cell line, and these SVs tend to avoid coding exons, yet prefer to insert near genes in intergenic regions. In addition, GenTrAn can measure the allele ratio of transposon SVs and all predicted SVs were successfully validated by genomic PCR. GenTrAn's precision in transposon SV detection and feasibility to mine the more economical SE read libraries make this an attractive tool for genome diagnostics.

Finding functional potential in transposable element-derived exons

Adam Frankish[*1], Jonathan Mudge[1], Jose M. Gonzalez[1], Barbara Uszczynska[2], Dimitri Pervouchine[2], Roderic Guigó[2] and Jennifer Harrow[1]

[1] Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK.
, [2] Centre for Genomic Regulation, Barcelona. Catalonia. Spain

*To whom correspondence should be addressed: af2@sanger.ac.uk

BACKGROUND

The GENCODE geneset comprises genome-wide manual annotation of protein-coding, long non-coding and pseudogene loci supplemented by automatic annotation from Ensembl. Manual annotation is both informed by and QC checked against computational predictions to ensure high sensitivity and specificity and gene models are validated by RTPCR_seq. Following complete first pass manual annotation of the entire human genome we extended analysis of transposable element (TE)-derived exons performed on the 1% of the genome included in the ENCODE pilot project (Mudge et al 2011) to identify all human exons overlapping TEs.

RESULTS

Our initial survey identified more than 7500 TE-derived exons in GENCODEv19. We will describe the careful filtering, manual curation and checking of these exons to identify a set of 100 TE-derived exons with strong conservation in vertebrate species outside the primate clade, the remainder being assigned to either primate or human-specific sets. While conservation of an exon in non-primate species is strongly suggestive of functionality, for those exons that have emerged more recently, an alternative proxy for functionality is required. We will present the analysis of the expression level and inclusion rates of TE-derived exons in 18 ENCODE cell-lines that have led us to identify more than 200 putatively functional examples displaying high expression and inclusion rates or very strong tissue specificity. We will discuss the functional impact of exon inclusion, specifically whether transcripts including TE-derived exons encode a full-length CDS or are likely to be targeted by the nonsense-mediated decay pathway. We will also illustrate the relative contribution of different repeat families to conserved and lineage-specific functionality and assess the repeat contribution to the transcript at the RNA and amino acid sequence levels.

CONCLUSIONS

We present a comprehensive analysis of TE-derived exons in the human genome and identify more than 300 such exons, both constitutive and alternatively spliced, which putatively contribute to the functionality of the transcripts in which they are included.

REFERENCES

1. Mudge JM, Frankish A, Fernandez-Banet J, Alioto T, Derrien T, Howald C, Reymond A, Guigó R, Hubbard T, Harrow J. The origins, evolution, and functional potential of alternative splicing in vertebrates. Mol Biol Evol. 2011 Oct;28(10): 2949-59.

Integrative analysis of hnRNP L-regulated mRNA splicing and expression in T cells

Brian S. Cole[*], Ganesh Shankarling, and Kristen W. Lynch[*]

Department of Biochemistry and Molecular Biophysics, Perelman School of Medicine,
University of Pennsylvania, Philadelphia, PA

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

BACKGROUND

HnRNP L plays a critical role in T cell development and function [1]. Recently we have described the transcriptome-wide association of hnRNP L with mRNAs in primary and cultured human T cells [2]. Among mRNAs associated with hnRNP L interaction sites, we discovered novel cases of hnNRP L-regulated alternative splicing, suggestive of a broad role for hnRNP L in regulating pre-mRNA processing in T cells. However, the transcriptome-wide impact of hnRNP L on transcript structure and abundance was unknown.

RESULTS

We have now completed a comprehensive analysis of changes in mRNA expression and splicing upon hnRNP L depletion in T cells, using a combination of RNA-Seq and RASL-Seq. Together these data allow us to generate a map relating hnRNP L binding and function. We find that hnRNP L regulates alternative splicing in hundreds of mRNAs, a subset of which are associated with hnRNP L interaction sites. Among hnRNP L-regulated cassette exons, we find cases of hnRNP L binding within, upstream, and downstream of both enhanced and repressed alternative exons, suggesting that location of hnRNP L binding does not alone determine regulatory outcome. To further understand the mechanism(s) of hnRNP L function we are analyzing the relationship of other gene features - such as splice site strength, exon/intron length and the presence of co-associated proteins - to the activity of hnRNP L. Additionally, we identify hundreds of mRNAs whose expression is altered by hnRNP L knockdown, yet these mRNAs are not enriched for hnRNP L occupancy or hnRNP L-regulated cassette exon splicing. We are therefore investigating if hnRNP L has additional roles in mRNA transcription, polyadenylation or mRNA stability.

CONCLUSIONS

We demonstrate that an integrative genomics approach provides insight into the relationship between hnRNP L-mRNA interaction and regulated pre-mRNA processing in T cells. We observe alternative splicing in hundreds of exons across the T cell transcriptome that are regulated by hnRNP L depletion using a combination of genomics approaches. Additionally, we quantify mRNA expression upon hnRNP L depletion. Finally, we relate hnRNP L binding information with transcripts whose splicing and/or expression is responsive to hnRNP L knockdown, setting the stage for multivariable analysis of features that may work together with hnRNP L binding to exert combinatorial control of mRNA splicing and expression.

REFERENCES

1. Gaudreau M.C., Heyd F., Bastien R., Wilhelm B., and Moroy, T.A. Alternative splicing controlled by heterogeneous nuclear ribonucleoprotein L regulates development, proliferation, and migration of thymic pre-T cells. Journal of Immunology. **188** (2012).
2. Shankarling G., Cole B.S., Mallory M.J., and Lynch K.W. Transcriptome-wide RNA interaction profiling reveals physical and functional targets of hnRNP L in human T cells. Molecular Cell Biology. **34** (2014).

Conserved splicing regulatory motifs in plants

Michael Hamilton[1], Anireddy SN Reddy[2], and Asa Ben-Hur[1*]

[1] Computer Science Dept, Colorado State University, Fort Collins, CO USA [2] Department of Biology, Colorado State University, Fort Collins CO USA

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

BACKGROUND

Intron retention is the predominant form of alternative splicing in plants, and is established to be a regulated phenomenon [1]. And yet, very little is known about the regulatory sequences in pre-mRNAs that regulate this event. To identify putative elements that contribute to intron excision or retention, we have developed a computational pipeline that uses all known intron-retention (IR) and intron excision events in a diverse selection of flowering plants (Arabidopsis, soybean, poplar, rice, brachypodium, and sorghum) to detect over-represented hexamers that are conserved across species. Intron retention events from the gene models of those species were augmented by IR events detected in a large collection of RNA-seq data.

RESULTS

After clustering the hexamers that showed a statistically significant over-representation into motifs, we have found 75 of them that are conserved across all the above species. Those motifs fall either in the flanking exons, or in the intron itself, associated with either retention or excision of an intron, i.e. function as either splicing enhancers or suppressors. The largest number of motifs was detected in the downstream exon (45), while very few (3) are present in the upstream exon. Many of the detected intronic suppressors exhibit a preference for the regions around the splice junctions, whereas splicing enhancers exhibited the opposite effect. A large fraction of the motifs are also known to regulate splicing in animals [2,3,4], and some were also detected in moss, which is a non-flowering plant.

CONCLUSIONS

In view of the high degree of conservation and similarity of our detected motifs with known splicing regulatory elements in animals, the elements we detected are likely to be functional.

REFERENCES

1. Reddy, A.S., Rogers, M.F., Richardson, D.N., Hamilton, M., and Ben-Hur, A. Deciphering the plant splicing code: experimental and computational approaches for predicting alternative splicing and splicing regulatory elements. Front Plant Sci, **3**:18 (2012).
2. Wang, Y., Ma, M., Xiao, X., and Wang, Z. Intronic splicing enhancers, cognate splicing factors and context-dependent regulation rules. Nat. Struct. Mol. Biol., **19**(10):1044-52 (2012).
3. Wang, Y., Xiao, X., Zhang, J., Choudhury, R., Robertson, A., Li, K., Ma, M., Burge, C.B., and Wang, Z. A complex network of factors with overlapping affinities represses splicing through intronic elements. Nat. Struct. Mol. Biol., **20**(1): 36-45 (2013).
4. Wang, Z., Rolish, M.E., Yeo, G., Tung, V., Mawson, M., and Burge, C.B. Systematic identification and analysis of exonic splicing silencers. Cell, **119**(6):831-45 (2004).

IdiffIR: Identifying differential intron retention from RNA-seq

Michael Hamilton[1], Anireddy SN Reddy[2], and Asa Ben-Hur[1*]

[1] Computer Science Dept, Colorado State University, Fort Collins, CO USA [2] Department of Biology, Colorado State University, Fort Collins CO USA

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

BACKGROUND

Deep sequencing of mRNA (RNA-seq) is a promising experimental technique for uncovering the regulation of eukaryotic splicing. Intron retention (IR) is the dominant form of alternative splicing in plants and occurs in large numbers in mammals even though it is not as common as other forms of alternative splicing. Its role as a regulatory mechanism and contribution to the enrichment of proteome diversity has been established. Therefore there is a need for an effective model that detects differential usage of introns.

RESULTS

We introduce iDiffIR, a method specifically designed for detecting differential IR events (see Figure 1 for an example) from high-throughput sequencing data. iDiffIR captures the notion of IR using an interpretable statistic, and can be used when no replicates are available thanks to the use of a novel measure of intragenic expression variability. We demonstrate the effectiveness of iDiffIR on publicly available RNA-seq datasets from human and Arabidopsis by showing that our method is able to detect hundreds of differential IR events from weakly and strongly expressed intronic regions. The detected events exhibit significantly over-represented GO categories that are in agreement with the experimental conditions. We compare iDiffIR to MISO [1] and MATS [2], on real and simulated data and find that it detects more events than these methods, with a similar level of accuracy.

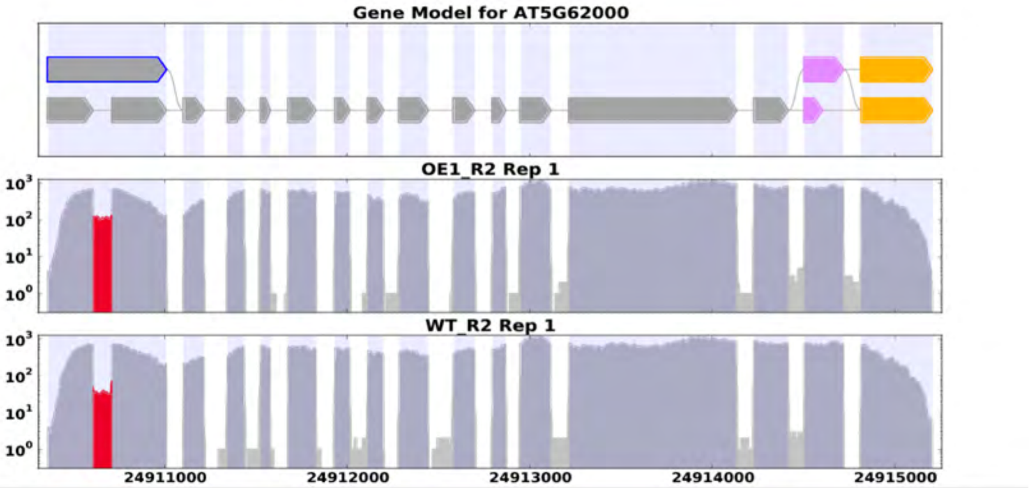

Figure 1: Example of a differential intron retention event. The top track displays the splice graph for the gene and the bottom two tracks show the read depths for the gene under two conditions (in log scale). Intronic reads for the event are highlighted in red.

CONCLUSIONS

We have shown the effectiveness of iDiffIR and its ability to detect differential IR. Our results further demonstrate the biological significance of intron retention in plants and animals. The iDiffIR software is available at http://combi.cs.colostate.edu/idiffir, and provides visualization tools that produce publication quality figures.

REFERENCES

1. Katz, Y., Wang, E.T., Airoldi, E.M., and Burge, C.B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat. Methods, **7**(12):1009-15 (2010).
2. Shen, S., Park, J.W., Huang, J., Dittmar, K.A., Lu, Z.X., Zhou, Q., Carstens, R.P., and Xing, Y. MATS: a Bayesian framework for

Splicing analysis using RNA-Seq and splicing code models - from in silico to in vivo

Jorge Vaquero-Garcia[1,2*], Alejandro Barrera[1,2], Matthew R. Gazzara[1,3], Juan González-Vallinas[1,2], Kristen W. Lynch[1,3], and Yoseph Barash[1,2]

[1] Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, 19104, USA, [2] Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, 19104, USA, [3] Department of Biochemistry and Biophysics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, 19104, USA

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

BACKGROUND

With the growing appreciation of RNA splicing's role in gene regulation, development, and disease, researchers from diverse fields find themselves investigating exons of interest. Commonly, researchers are interested in knowing if an exon is alternatively spliced, if it is differentially included in specific tissues or in developmental stages, and what regulatory elements control its inclusion. Answering these questions has been challenging. First, if relevant experimental data such as RNA-Seq is available, there is a need to both quantify and visualize splicing changes across experiments. If such experimental data does not exist or is scarce (e.g., low read coverage) one may still be interested in computational predictions of splicing changes. Second, there is a need to identify the underlying regulatory elements. In recent years, machine-learning techniques have been applied to high-throughput data to develop probabilistic splicing code models [1-3]. These models are able to predict directly from genomic sequence what would be the context-specific splicing outcome and identify important regulatory features. Recently, the web-tool AVISPA was developed as a user-friendly front end to these splicing code models [4].

RESULTS

Using Bridging integrator 1 (BIN1) as a case study, we first demonstrate how to visualize local splicing variations derived from RNA-Seq data using our newly developed visualization package VOILA. BIN1 is a nucleocytoplasmic adaptor protein known to be functionally regulated through alternative splicing in a tissue-specific manner. Specific *Bin1* isoforms have been associated with muscular diseases and cancers, making the study of its splicing regulation of wide interest. Next, we demonstrate how AVISPA, a tool deployed over a Galaxy server instance (http://avispa.biociophers.org), can be applied to analyze alternative splicing of *Bin1*. Applying AVISPA to *Bin1* exons, we show that many *Bin1* tissue-dependent isoforms are correctly predicted, along with many of its known regulators. We also demonstrate how AVISPA is used to generate high confidence novel regulatory hypotheses, and experimentally validate predicted regulators of *Bin1* alternative splicing [5].

CONCLUSIONS

The visualization package VOILA facilitates analysis of alternative splicing from RNA-Seq data while AVISPA enables researchers to perform *in silico* analysis of alternative splicing. This *in silico* analysis can then guide *in vivo* experiments and lead to novel findings of splicing regulatory *cis* elements and *trans* acting factors.

REFERENCES

1. Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, Blencowe BJ, Frey BJ: Deciphering the splicing code. *Nature* 2010, 465:53–59.
2. Xiong HY, Barash Y, Frey BJ: Bayesian prediction of tissue-regulated splicing using RNA sequence and cellular context. *Bioinformatics* 2011, 27:2554–2562.
3. Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, Slobodeniuc V, Kutter C, Watt S, Çolak R, Kim T, Misquitta-Ali CM, Wilson MD, Kim PM, Odom DT, Frey BJ, Blencowe BJ: The Evolutionary Landscape of Alternative Splicing in Vertebrate Species. *Science* 2012, 338:1587–1593.
4. Barash Y, Vaquero-Garcia J, Gonzalez-Vallinas J, Xiong H, Gao W, Lee L, Frey B: AVISPA: a web tool for the prediction and analysis of alternative splicing. *Genome Biology* 2013, 14:R114.
5. Gazzara MR, Vaquero-Garcia J, Lynch KW, Barash Y: In silico to in vivo splicing analysis using splicing code models. *Methods* 2013.

Consistent alternative splicing isoform switches across tumor samples provide novel signatures in cancer

E. Sebestyen, M. Zawisza, G.P. Alamancos, A. Pages, E. Eyras

Computational Genomics, Universitat Pompeu Fabra, Barcelona, Spain
Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain

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

BACKGROUND

Cancer genome projects are instrumental to uncover recurrent alterations in tumours [1]. Multiple studies have shown that the mutational and expression patterns can be successfully applied to classify and predict various tumor conditions [2]. However, alternative pre-mRNA splicing alterations, which bear major importance in terms of the understanding of cancer [3], have not been exhaustively studied yet in the context of recent cancer genomics efforts.

RESULTS

We have used RNA sequencing data from The Cancer Genome Atlas (TCGA) project for more than 1000 tumor and normal samples to characterize the alternative splicing of genes in 10 different cancer types. After careful quality assessment of the samples and data normalization, we have applied a new algorithm based on the consistent reversal of relative expression of splicing isoforms to define alternative splicing isoform switches. Cross-validation allows us to find predictive models based on isoform switch rules, which include known tumor suppressors (*TPM1*, *TSC2*, *NUMB*, *FBLN2* and *QKI)* and which correctly classifies nearly 100% of the samples in each cancer type as well as for the subtypes in breast cancer (luminal A, luminal B, Her2+ and basal), colon cancer (hypermutated and non-hypermutated), and lung squamous cell carcinoma (primitive, classical, secretory and basal). We further investigate how these isoform switches may remodel the network of protein interactions and predict a number of interactions that are disrupted in specific cancer types.

CONCLUSIONS

Our method provides an effective way to process and interpret large RNA sequencing datasets from cancer genomics projects. Additionally, we describe novel alternative splicing signatures that are predictive of tumors and that may contribute to cancer pathology.

REFERENCES

1. Stephens et al. (2012). The landscape of cancer genes and mutational processes in breast cancer. Nature, 486(7403), 400-4.
2. Hudson (2013). Genome variation and personalized cancer medicine. J Intern Med. 274(5): 440-50.
3. Bechara et al. (2013) RBM5, 6, and 10 differentially regulate NUMB alternative splicing to control cancer cell proliferation. Mol Cell. 52(5):720-33.

Epigenetic transmission of thousands of long, noncoding transcripts containing RNA-cached copies of whole chromosomes

Kelsi Lindblad[1*], John Bracht[2], April Williams[1], and Laura Landweber[2]
[1] Lewis-Sigler Institute for Integrative Genomics, Princeton University, [2] Department of Ecology and Evolutionary Biology, Princeton University

*To whom correspondence should be addressed: kalindbl@princeton.edu

BACKGROUND

The ciliate *Oxytricha trifallax* maintains two genomes, a germline genome that is active only during sexual conjugation and a somatic genome transcribed during asexual growth. The somatic genome derives from the germline through a process of extensive reduction and rearrangement, which eliminates about 95% of the germline sequence. The remaining segments fuse to form the set of ~16,000 gene-sized chromosomes in the somatic genome [1]. A class of 27nt piRNAs mark the sections of the germline genome retained during development [2], but they cannot provide enough information to decode the ~20% of the somatic genome segments whose precursors are in an encrypted order and require permutation or inversion to produce their somatic forms. Previous experiments found that long, non-coding (lnc) RNA "templates"—telomere-to-telomere copies of mature chromosomes—can program the rearrangement process [3]. Here we report the results of genome-wide sequencing efforts to capture this RNA-cached copy of the maternal somatic genome.

RESULTS

Whole-cell RNA was collected over a time course during somatic development and enriched for molecules containing two telomeric sequences, then sequenced using a standard paired-end Illumina protocol. This approahch identified lncRNA templates corresponding to the majority of somatich chromosomes (9,381 out of ~16,000). RT-PCR could identify many additional molecules that did not appear in the Illumina dataset, suggesting that all chromosomes are transcribed during development. Contrary to expectations, we also found that template levels fluctuate over time and distinct populations of templates appear in greater abundance at different stages during development.

CONCLUSIONS

Oxytricha produces an RNA cached copy of its somatic genome, which it transmits to the daughter cell during sexual conjugation. There, the long, noncoding RNAs guide DNA rearrangements in the developing somatic genome. The levels and identities of lncRNAs vary over the course of development. This genome-wide study identified over 9,000 lncRNAs representing complete, telomere-to-telomere copies of *Oxytricha's* chromosomes.

REFERENCES

1. Swart E.C. e*t al.* The Oxytricha trifallax macronuclear genome: a complex eukaryotic genome with 16,000 tiny chromosomes. PloS Biology **11**(1) 2013: e1001473.
2. Fang, W., Wang, X., Bracht J.R., Nowacki M., and Landweber L.F. Piwi-interacting RNAs protect DNA against loss during Oxytricha genome rearrangement. Cell **151**(6) 2012: 1243-55.
3. Nowacki M., Vijayan V., Zhou Y., Schotanus K., Doak T.G., and Landweber L.F. RNA-mediated epigenetic programming of a genome-rearrangement pathway. Nature **451**: 153-8.

The Huntington's disease gene regulates miRNA associated mRNA splicing to modulate neuronal differentiation

Sonia M. Vallabh[1, (*)], Ashok Ragavendran[1,(*)], Serkan Erdin[1,(*)], Nicole Solomos[1], Ihn Sik Seong[1], Jong-Min Lee[1,], James F. Gusella[1,2], Michael E. Talkowski[1,2], Marcy E. MacDonald[1,2,(#)] and Marta Biagioli[1,(#)]

[1] Center for Human Genetic Research, Massachusetts General Hospital,
Boston, Massachusetts, USA.

[2] Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA.

(*) These authors contributed equally to this work

(#) Correspondence: macdonam@helix.mgh.harvard.edu; biagioli@chgr.mgh.harvard.edu

BACKGROUND

Huntington's disease (HD) is devastating neurodegenerative disease caused by a CAG repeat expansion mutation in the *HTT* gene. The pathological process, that ultimately leads to neuronal cell loss in the caudate and putamen, as well as in the cortex and other brain regions, is induced in a dominant manner by a single copy of the mutant allele. Both the normal and the CAG-expanded mutant proteins are expressed from conception, implying effects of the mutant allele that may be related to its inherent normal activities in the regulation of neurogenesis as delineated *in vivo* and in embryonic stem cell (ESC)-derived cells by phenotypes produced by knock-out of the mouse homolog *Htt* (previously *Hdh*) [1, 2, 3].

RESULTS

By combined analysis of chromatin analysis of histone H3K4me3 histone mark with miRNA-seq, we identified miR124 as significantly upregulated in *Htt*-null embryonic stem cells-derived neuronal progenitors. The increased levels of this neuronal-specific miRNA, in turn, regulated the levels of expression of the *trans*-splicing regulators polypyrimidine tract binding protein (*Ptbp1*) and Ser/Arg repetitive matrix 4 (*Srrm4*), thereby leading to altered inclusion of exons defining specific isoforms of neuronal-specific target genes (i.e. *Bin1*, *Aplp2*, *Kif1b*).

CONCLUSIONS

These observations disclose a new *Htt* function through which the CAG-expanded *Htt* allele may alter regulation of miRNA and alternative mRNA splicing from the earliest stages of development thereby fundamentally changing the intrinsic potential of single neurons and, in turn, sensitizing them to neurodegeneration.

REFERENCES

1. White J. K. *et al*. Huntingtin is required for neurogenesis and is not impaired by the Huntington's disease CAG expansion. Nature Genet. 17, 404–410 (1997).
2. Auerbach W. *et al*. HD mutation causes progressive lethal neurological disease in mice expressing reduced levels of huntingtin. Hum. Mol. Genet. 10, 2515–2523 (2001).
3. Conforti P. *et al*. Lack of huntingtin promotes neural stem cells differentiation into glial cells while neurons expressing huntingtin with expanded polyglutamine tracts undergo cell death. Neurobiol. Dis. 2013 Feb;50:160-70.

Systematic analysis of transcription initiation in human reveals tissue-specific alternative promoter and putative distal 5’ exon usage

R. Taylor Raborn[1*], Daniel S. Standage[1], and Volker Brendel[1,2]

[1] Department of Biology and [2] School of Informatics and Computing, Indiana University, Bloomington, IN 47405

*To whom correspondence should be addressed: rtraborn@indiana.edu

BACKGROUND

The remarkable plasticity functional encoded within metazoan genomes is underpinned by precise and extensive regulation of gene expression via large numbers of control elements [1]. Deep-sequencing of transcriptomes reveals that the transcriptional landscape in metazoan genomes is more varied and complex than previously thought [2]. Interrogation of transcriptomes from diverse cell types within the ENCODE project revealed that transcription is pervasive in the human genome [3]. Expression of multiple transcript isoforms per gene was observed across all cell lines, with three-quarters of protein-coding genes shown to simultaneously express two or more major transcript isoforms [3]. Profiling of 5’ mRNAs end by ENCODE [3] and FANTOM5 [4] using CAGE provided evidence of alternative promoter usage, suggesting that this may be a widespread feature of mammalian gene expression regulation. Alternative promoter usage has not been studied extensively, especially with respect to the tissue-specificity and motif composition of promoters and associated gene function. To address this, we characterized alternative promoter usage and regulation in diverse human cell types using an integrative genomic approach. Given limited evidence of distal 5’ exon usage in metazoans [5],[6], we also investigated the novel possibility that some alternative promoter usage is related to alternative splicing

RESULTS

After retrieving and processing 5’ end profiling data from ENCODE (3.16x109 CAGE tags), derived from 38 distinct cell types, we identified and annotated putative promoters from TSSs in all conditions using our software package, TSRchitect. We found that alternative promoter usage was widespread in human, consistent with a recent finding [4]. Peaked promoters, which are associated with defined core promoter elements such as TATA and Inr [1], were more likely to participate in alternative promoter usage, as were genes expressed in more restricted set of cell types. We characterized promoters with tissue- or cell-type-specific usage patterns, and associated them with global chromatin marks from matched cell-types, where available. Considering the positions of alternative promoter usage, we found that a sizable fraction of genes (n=1274) utilize a putative alternative promoter 10kb or more away from the prominent promoter in at least one cell type. Hypothesizing that this population of alternative promoters reflect genes with distal 5’ exons, we evaluated the evidence that these represent bona fida alternative splice forms. We found that a fraction of these putative alternative promoters were likely enhancers (i.e. deriving from eRNAs [7]), but also discovered among these a subset of candidate distal 5’ exons that were supported by other RNA evidence.

This work is preliminary; efforts are currently underway to identify splice sites and promoter motifs from these candidate 5’ exons. Future work will include an experimental confirmation of these putative 5’ distal exon-associated splice forms and an analysis of the protein isoforms they are predicted to encode.

CONCLUSIONS

We find that alternative promoter usage is widespread in human, and varies considerably with cell type. A subset of the annotated alternative promoters were identified at large distances upstream of the respective protein-coding region; many of these are consistent with 5’ alternative splice forms, predicted to encode distinct protein isoforms.

REFERENCES

1. B. Lenhard, A. Sandelin, and P. Carninci, “Metazoan promoters: emerging characteristics and insights into transcriptional regulation.,” Nat Rev Genet, vol. 13, no. 4, pp. 233–245, Apr. 2012.
2. A. Jacquier, “The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs.,” Nat Rev Genet, vol. 10, no. 12, pp. 833–844, Dec. 2009.
3. S. Djebali, C. A. Davis, A. Merkel, et al., “Landscape of transcription in human cells,” Nature, vol. 489, no. 7414, pp. 101–108, Sep. 2012.
4. FANTOM Consortium and the RIKEN PMI and CLST (DGT), “A promoter-level mammalian expression atlas.,” Nature, vol. 507, no. 7493, pp. 462–470, Mar. 2014.
5. J. R. Manak, S. Dike, V. Sementchenko, et al., “Biological function of unannotated transcription during the early development of Drosophila melanogaster,” Nat Genet, vol. 38, no. 10, pp. 1151–1158, Oct. 2006.
6. F. Denoeud, P. Kapranov, C. Ucla, A. et al., “Prominent use of distal 5' transcription start sites and discovery of a large number of additional exons in ENCODE regions.,” Genome Research, vol. 17, no. 6, pp. 746–759, Jun. 2007.
7. R. Andersson, C. Gebhard, I. Miguel-Escalada, et al., “An atlas of active enhancers across human cell types and tissues,” Nature, vol. 507, no. 7493, pp. 455–461, Mar. 2014.

Identifying MMPs, RECK (canonical and splicing isoforms) and microRNAs dysregulated in Deep Endometriosis

Carreira, A. C.[1]*, Trombetta-Lima, M. T., Abrão, M. S[2] and Sogayar, M. C.[1]

[1]NUCEL (Cell and Molecular Therapy Center) and NETCEM (Center for Studies in Cell and Molecular Therapy)
School of Medicine - Chemistry Institute, Biochemistry Department, University of São Paulo, SP, Brazil, [2]
Department of Obstetrics and Gynecology School of Medicine, University of São Paulo, SP, Brazil

*To whom correspondence should be addressed: ancoc@iq.usp.br

BACKGROUND

Endometriosis is a benign gynecologic disorder characterized by the ectopic growth of misplaced endometrial cells, and affects about 10 to 15% of women of reproductive age. The most widely accepted etiologic theory of endometriosis proposes that viable endometrial tissue is refluxed through the fallopian tubes during menstruation and implants on peritoneal surface or pelvic organs. The ectopic endometrium degrades the extracellular matrix (ECM), through the activity of metalloproteinases (MMPs) and other molecules, and invades the surrounding tissue with corresponding cell proliferation and neoangiogenesis. In the last decade, numerous gene expression profiling studies have demonstrated that many genes are deregulated in endometriosis, including *MMPs*. The degradation of the ECM by the MMPs is closely regulated by TIMPs during normal physiological conditions. Additional MMPs regulators are: RECK, which negatively regulates MMPs in tumors, and SPARC, which upregulates MMPs, both of which have never been analyzed in deep infiltrating endometriosis. We have recently described three new *RECK* gene isoforms (*RECKB, RECKD, RECKI*), which are modulated in glioblastoma [1]. Recently, miRNAs were shown as *RECK* regulators. Aberrant miRNA expression is associated with several human diseases such as cancer, cardiovascular disorders, inflammatory diseases, and benign or malignant pathologies of the human female reproductive tract. miRNAs are naturally occurring posttranscriptional regulatory molecules that potentially play a role in endometriotic lesion development. Fully understanding of RECK and its isoforms and MMPs specific miRNA profile may contribute to refine novel strategies for endometriosis, since miRNAs are potential therapeutic targets for treating this disease.

RESULTS

To evaluate the expression profile of *MMPs, TIMPs, SPARC, RECK* and their isoforms, we have used a high-throughput qRT-PCR platform for tissue samples of eutopic and ectopic endometrium and bowel of five patients with deep infiltrating endometriosis [2]. As normal control, we included a pool of endometrium of healthy fertile women (laparoscopy for tubal ligation). We described an increase in *MMP2, MMP7, MMP14, SPARC, TIMP1, 2 and 3* in ectopic endometrium *vs* bowel tissue ($p<0.05$)*; MMP9* is increased in ectopic endometrium *vs* eutopic endometrium of woman with and without endometriosis; and *RECK* and its isoforms tend to increase their expression in ectopic endometrium *vs* bowel tissue. To evaluate the miRNA-mRNA correlation in deep endometriosis, we have evaluated the global miRNAs expression profile and focused on specific miRNAs regulators of RECK and its isoforms in these samples. Networks are under construction by integrating mRNA and miRNA profiles.

CONCLUSIONS

We were able to identify a specific expression pattern of *MMPs* and their regulators, *RECK* and its new isoforms, in deep endometriosis. The description of regulatory networks of miRNAs and their targets in deep endometriosis could provide new insights into the molecular pathology and contribute to identify novel biomarker for early stages of endometriosis.

REFERENCES

1. Trombetta-Lima, M., Winnischoffer, S.M. B., Carreira, A.C., Demasi, M.A.A., Colin, C., Oba-Shinjo, S. M., Marie, S.K.N., Sogayar, M.C. Abstract 2354: Characterization of three novel splice variants of the RECK tumor and metastasis suppressor gene. Cancer Research (2011).
2. Carreira, A.C., Trombetta-Lima, M., Baracat, E.C., Abrão, M.C, Sogayar, M.C. Involvement of MMPs regulators SPARC and RECK (canonical and splicing isoforms) in endometriosis. Proceedings of the 12th World Congress on Endometriosis, p. 228 (2014).

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 2014