Genomics Data 6 (2015) 217–221

Contents lists available at ScienceDirect

Genomics Data

journal homepage: http://www.journals.elsevier.com/genomics-data/

Data in Brief

Global analysis of CPSF2-mediated alternative splicing: Integration of global iCLIP and transcriptome profiling data

Ashish Misra [a,b,*], Jianhong Ou [b], Lihua Julie Zhu [b,c,d], Michael R. Green [a,b,*]

[a] Howard Hughes Medical Institute, USA
[b] Department of Molecular, Cell and Cancer Biology, University of Massachusetts Medical School, Worcester, MA 01605, USA
[c] Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, MA 01605, USA
[d] Program in Molecular Medicine, University of Massachusetts Medical School, Worcester, MA 01605, USA

ARTICLE INFO

Article history:
Received 25 September 2015
Accepted 28 September 2015
Available online 3 October 2015

Keywords:
iCLIP
RNA-seq
Alternative splicing
CPSF
SYMPK

ABSTRACT

Alternative splicing is a key mechanism for generating proteome diversity, however the mechanisms regulating alternative splicing are poorly understood. Using a genome-wide RNA interference screening strategy, we identified cleavage and polyadenylation specificity factor (CPSF) and symplekin (SYMPK) as cofactors of the well-known splicing regulator RBFOX2. To determine the role of CPSF in alternative splicing on a genome-wide level, we performed paired-end RNA sequencing (RNA-seq) to compare splicing events in control cells and RBFOX2 or CPSF2 knockdown cells. We also performed individual-nucleotide resolution UV cross-linking and immunoprecipitation (iCLIP) to identify direct binding targets of RBFOX2 and CPSF2. Here, we describe the experimental design, and the quality control and data analyses that were performed on the dataset. The raw sequencing data have been deposited in NCBI's Gene Expression Omnibus and are accessible through GEO Series accession number GSE60392.

Specifications	
Organism/cell line/tissue	Human Flp-In™ 293 cell line
Sex	Female
Sequencer or array type	Illumina HiSeq2000
Data format	Raw (SRA) and analyzed (for iCLIP — bed files; for RNA-seq — count files)
Experimental factors	Flp-In™ 293 cells were transduced with lentiviruses encoding non-silencing, RBFOX2 and CPSF2 short hairpin RNAs.
Experimental features	iCLIP and RNA-seq analysis of RBFOX2 and CPSF2
Consent	Not applicable
Sample source location	Not applicable

1. Direct link to deposited data

The iCLIP and RNA-seq data are available at the following GEO series: GSE60392, http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE60392.

Cell	Treatment	Feature	Replicates	GEO accession URL
Human Flp-In™ 293	Untreated	RBFOX2 iCLIP	2	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1477476
Human Flp-In™ 293	Untreated	CPSF2 iCLIP	2	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1477475
Human Flp-In™ 293	Non-silencing lentivirus knockdown	Paired-end RNA-seq	2	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1477603 http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1477604
Human Flp-In™ 293	RBFOX2 lentivirus knockdown	Paired-end RNA-seq	2	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1477601 http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1477602
Human Flp-In™ 293	CPSF2 lentivirus knockdown	Paired-end RNA-seq	2	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1477599 http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1477600

* Corresponding authors at: Department of Molecular, Cell and Cancer Biology, University of Massachusetts Medical School, Worcester, MA 01605, USA.
E-mail addresses: ashish.misra@umassmed.edu (A. Misra), michael.green@umassmed.edu (M.R. Green).

http://dx.doi.org/10.1016/j.gdata.2015.09.022

2. Experimental design, materials and methods

2.1. Cell lines

Flp-In-293 cells (Invitrogen) were used for the experiments. Cells were transduced with lentiviruses expressing shRNAs targeting RBFOX2 or CPSF2, or as a control a non-silencing (NS) shRNA, and then puromycin selected for 4 days. Cells were harvested 8–10 days later for downstream processing.

2.2. Individual-nucleotide resolution UV cross-linking and immunoprecipitation (iCLIP)

iCLIP experiments were carried out essentially as previously described [1] with the following modifications. Flp-In-293 cells were grown in 10 cm plates and then subjected to UV-C irradiation (200 mJ/cm^2, Stratalinker 2400). Upon removal of phosphate buffered saline (PBS), cells were scraped off and precipitated by centrifugation at 3000 rpm for 10 min. Pellets were resuspended in 1 ml lysis buffer (50 mM Tris–HCl pH 7.4, 100 mM NaCl, 1 mM $MgCl_2$, 0.1 mM $CaCl_2$, 1% NP-40, 0.1% SDS, 0.5% sodium deoxycholate). Samples were partially digested with RNase I by incubation with 10 μl RNase I (Ambion) (diluted 1:500 in lysis buffer) and 5 μl Turbo DNase (Ambion) for 5 min at 37 °C while shaking at 1100 rpm. Cells were centrifuged at 14,000 rpm for 15 min to collect the cross-linked lysate.

To prepare the beads for RNA purification, 50 μl of protein G-coated Dynabeads (Invitrogen) was washed twice with 1 ml lysis buffer and resuspended in 200 μl lysis buffer containing 3–5 μg of RBFOX2 (Bethyl Laboratories, A300-864A) or CPSF2 (Bethyl Laboratories, A301-580A) antibody. The beads were rotated at room temperature for 60 min, washed twice with lysis buffer, and added to the cross-linked lysate. After overnight incubation at 4 °C, the beads were washed four times with high-salt wash buffer (50 mM Tris–HCl pH 7.4, 1 M NaCl, 1 mM EDTA, 0.1% SDS, 0.5% sodium deoxycholate, 1% NP-40) and once with PNK wash buffer (20 mM Tris–HCl pH 7.4, 10 mM $MgCl_2$, 0.2% Tween-20). The beads were resuspended in 5 μl 10× PNK buffer, 44 μl H_2O and 1 μl PNK enzyme (NEB) and incubated at 37 °C for 20 min while shaking at 1100 rpm. Samples were then washed twice with high-salt wash buffer and once with 900 μl PNK wash buffer.

To ligate linkers to the 3′ end of RNAs, 4.5 μl of purified RNA was mixed with 1 μl of 10 μM miRCat33 Adaptor (Integrated DNA Technologies), 1.5 μl of RNA Ligase buffer, 0.5 μl of T4 RNA Ligase 2, Truncated K227Q (NEB), and 7.5 μl of PEG 8000, and incubated at 30 °C for 6 h. Samples were then end-labeled using PNK enzyme and run on a NuPAGE Novex 10% Bis–Tris Protein Gel (Invitrogen) with 1× MOPS running buffer (Invitrogen). Proteins and covalently bound RNAs were then transferred to a nitrocellulose membrane (Whatman) using a Novex wet transfer apparatus (Invitrogen). The nitrocellulose membrane was rinsed with 1× PBS, wrapped in plastic wrap and exposed to X-ray film.

To isolate the RNA–protein complex, the region of the nitrocellulose membrane corresponding to the complex size of interest was cut out. Cross-linked RNAs were then isolated by incubating the nitrocellulose membrane pieces with PK buffer (100 mM Tris–HCl pH 7.5, 50 mM NaCl, 10 mM EDTA) containing 2 mg/ml proteinase K (Roche) for 20 min at 37 °C, after which an equal volume of PK buffer containing 7 M urea was added to the tube and the samples were incubated for another 20 min at 37 °C. Subsequently, 600 μl of phenol/chloroform (Ambion) was added, and the mixture was incubated for 10 min at room temperature. Samples were then centrifuged for 10 min at 13,000 rpm at room temperature, and the aqueous phase was transferred into a new microtube and again centrifuged. 400 μl of supernatant was mixed with 0.5 μl GlycoBlue (Ambion), 40 μl 3 M sodium acetate pH 5.5, and 1 ml 100% ethanol and incubated overnight at −20 °C. RNAs were precipitated by centrifugation for 30 min at 15,000 rpm at 4 °C, washed with 500 μl 80% ethanol and resuspended in 30 μl H_2O.

To carry out reverse transcription, 30 μl purified RNA was mixed with 1 μl of either BC2 (5′-pGGCGATGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT-SP18-CTCG GCATTCCTGCTGAACCGCTCTTCCGATCT-CCTTGGCACCCGAGAATTCCA-3′) or BC3 (5′-pGGCGATGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT-SP18-CTCG GCATTCCTGCTGAACCGCTCTTCCGATCT-CCTTGGCACCCGAGAATTCCA-3′) pRT primer (10 μM concentration) and 1.5 μl 10 mM dNTP mix, and incubated at 65 °C for 5 min and then transferred to ice. Subsequently, 9 μl of 5× first strand synthesis buffer (Invitrogen), 2.25 μl 0.1 M DTT, and 1.25 μl Superscript III reverse transcriptase (Invitrogen) were added to a final volume of 45 μl. Reverse transcription was performed according to the manufacturer's protocol.

The cDNAs were size-separated on a 6% TBE urea gel (Invitrogen). Two size fractions were recovered from the gel by cutting the gel corresponding to a cDNA size of 120–155 nt and 150–225 nt. Gel fragments were mixed with 400 μl TE buffer, crushed with a 1 ml syringe plunger and centrifuged overnight at 1100 rpm at 37 °C. Gel pieces and the supernatant were separated using a Costar SpinX column (Corning Incorporated). 40 μl 3 M sodium acetate pH 5.5 and 0.5 μl glycogen were added to the supernatant and centrifuged at 13,000 rpm at −20 °C overnight to precipitate the reverse transcription product. Pellets were dried at room temperature and resuspended in 10 μl H_2O.

The cDNA samples were then circularized by mixing with 2.0 μl 10× CircLigase buffer (Epicentre), 1 μl 50 mM $MnCl_2$ and 1 μl CircLigase II (Epicentre) and incubated for 3–4 h at 60 °C. For high-throughput sequencing, the circularized cDNAs were PCR-amplified. Briefly, cDNAs were mixed with 1 μM each of PE1.0 (5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3′) and PE2.0 (5′-CAAGC AGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT) primers (10 μM concentration) and KAPA 2× Ready mix (Kapa Biosystems), and PCR amplification was performed as per the manufacturer's instructions. The PCR product was gel eluted and subjected to sequencing on a HiSeq 2000 sequencing system (Illumina).

2.3. High-throughput sequencing and bioinformatic analysis of iCLIP data

Two biological replicates of the RBFOX2 and CPSF2 iCLIP libraries were sequenced with a single-end length of 100 bp. Only reads containing the 5′ adapter were considered for analysis. The 3′ and 5′ adaptor sequences were trimmed from these reads. Furthermore, ribosomal RNA reads were filtered out using SortMeRNA(v1.9). The RNA database used by SortMeRNA was downloaded from http://www.arb-silva.de/ [2]. Homopolymer sequences >10 bp were removed followed by sequence quality assessment using fastQC (v 0.10.1) [http://www.bioinformatics.babraham.ac.uk/projects/fastqc/] (Fig. 1). After clean up, the read lengths were in the range of 21–72 bp. The fastQC score for all the bases was >30, indicating that the reads were of high quality. Quality assessment was also performed on raw reads before clean up and similar quality was observed. Following cleanup, the reads were mapped against human reference genome (GRCh37/hg19, Feb. 2009) using bowtie (v1.0.0) [3] with the following parameter setting: "-q -e 100 -l 20 -m 1 –best –strata". Finally, the iCLIP-seq reads were collapsed to remove PCR artifacts using picard tools (v1.94). Reproducibility of the biological replicates was assessed using Pearson correlation analysis (RBFOX2 $r = 0.825$, p-value $< 2.2e-16$; CPSF2 $r = 0.824$, p-value $< 2.2e-16$) (see Fig. S2D in [4]).

2.4. High-throughput sequencing and bioinformatic analysis of RNA-seq data

RNA-seq samples were prepared using a TruSeq RNA Library Prep Kit v2 (Illumina) according to the manufacturer's instructions. Sequencing was done using an Illumina HiSeq 2000 with a paired-end length of 100 bp for duplicated NS, RBFOX2- knockdown and CPSF2-knockdown

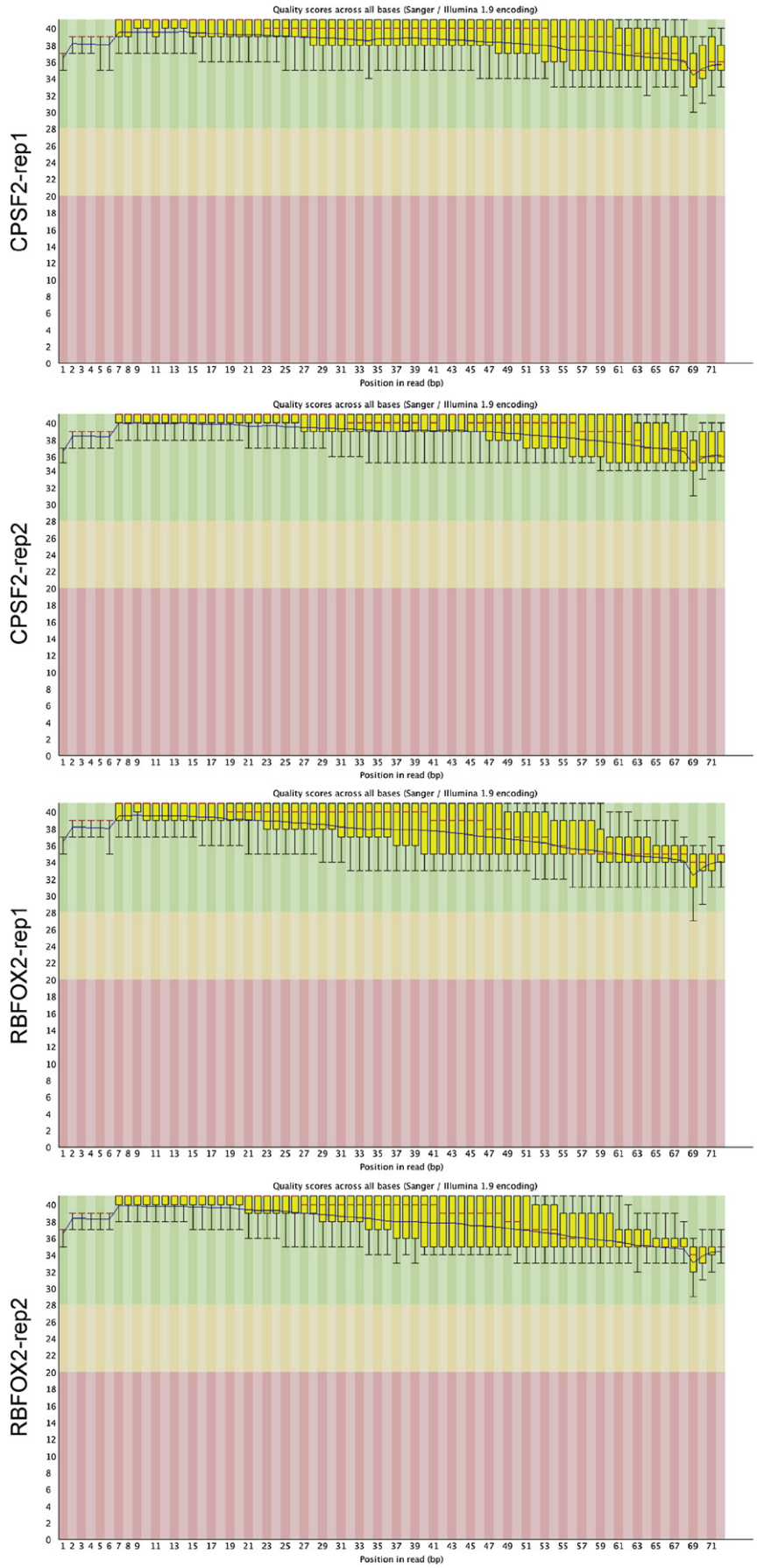

Fig. 1. Graphs representing per base quality using Phred score for the iCLIP data after cleanup. The figure is generated by fastQC.

samples. The quality of the sequencing reads was assessed using fastQC (v 0.10.1). The fastQC score for all the bases was >30, indicating that the reads were of high quality. The reads were aligned against human reference genome (GRCh37/hg19, Feb. 2009) using TopHat (v2.0.9, bowtie2/2.1.0) [3] with the following parameter setting: "-G [ucsc_hg19_knownGene] –mate-inner-dist 50 –b2-very-sensitive". The ucsc_hg19_knownGene annotation file was downloaded from UCSC table browser and quality of alignment was assessed using SAMStat (v 1.09) [5]. A pie chart describing the quality of sequence alignment distribution showed that >97% of the reads have MAPQ ≥ 30, indicating the high mapping quality of the reads (Fig. 2A). Gene expression level (FPKM) and differential gene expression analysis were performed using Cufflinks (v2.1.1) [6]. Python script, provided by the DEXSeq (v1.10.6) package [7], was used for exon level read count estimation with the following parameter setting: "-p yes -r pos -s no" to count the number of reads. Pearson correlation analysis of gene expression levels was performed to evaluate the reproducibility between biological replicates (NS: $r = 0.991$, p-value $< 2.2e-16$; RBFOX2: $r = 0.991$, p-value $< 2.2e-16$; CPSF2: $r = 0.992$, p-value $< 2.2e-16$) (Fig. S2A in [4]) Pearson correlation analysis of exon level expression also demonstrated high reproducibility between biological replicates (NS: $r = 0.997$, p-value $< 2.2e-16$; RBFOX2: $r = 0.999$, p-value $< 2.2e-16$; CPSF2: $r = 0.989$, p-value $< 2.2e-16$) (Fig. 2B). Multidimensional scaling (MDS) plot was generated to visualize the similarity of gene expression between biological replicates and dissimilarity among NS, RBFOX2 and CPSF2 knockdown samples using cummeRbund package (v 2.8.2) [8]. The results of the MDS plot showed that the biological replicates clustered closely while there was a clear segregation among NS, RBFOX2 and CPSF2 knockdown samples, indicating that

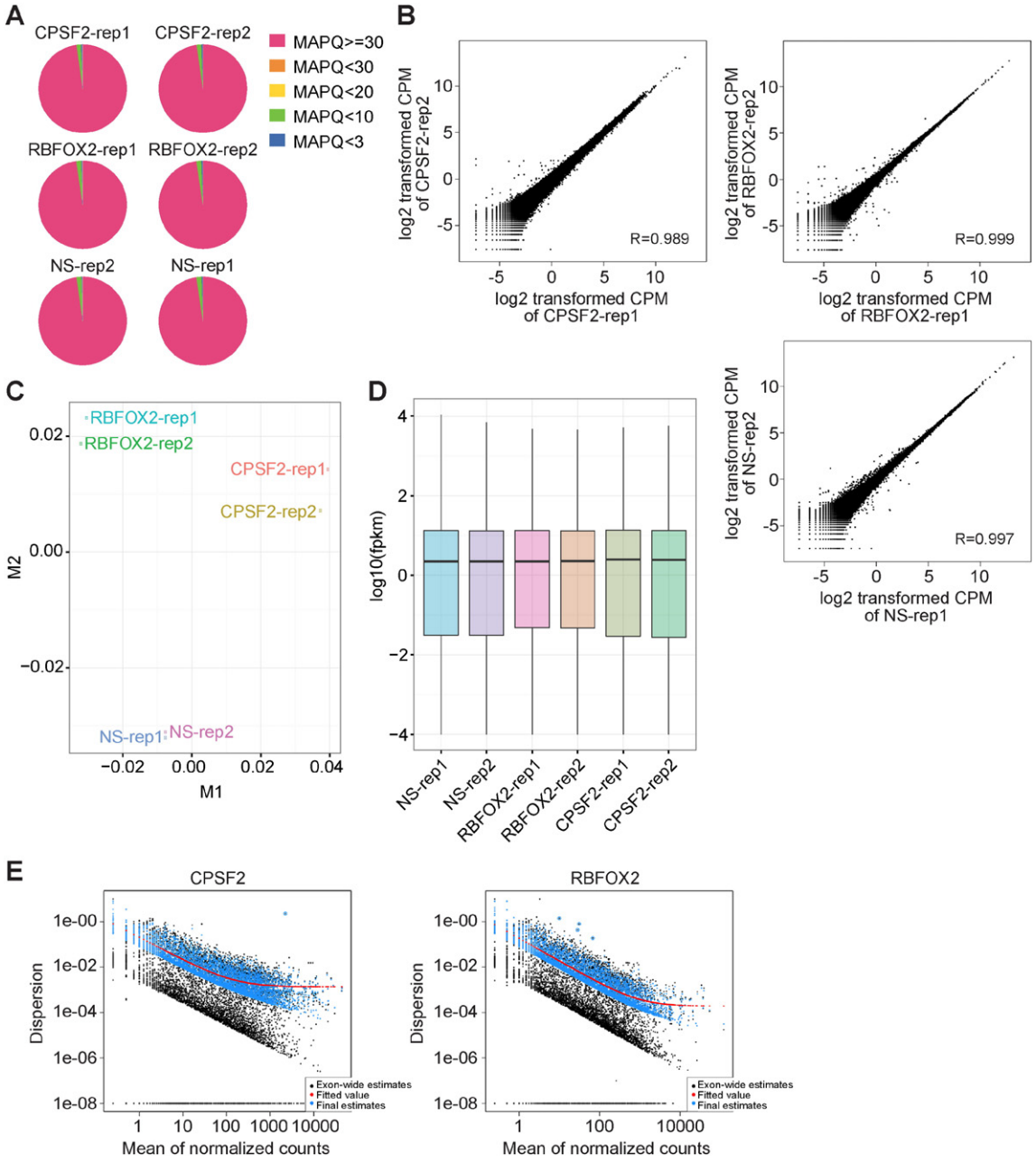

Fig. 2. Quality assessment of RNA-seq data and alignment quality. A. Pie chart obtained with SAMstat depicting the distribution of sequence alignment quality. B. Scatter plots of RNA-seq data to assess consistency of exon level expression between biological replicates for all three knockdown samples. C. Multi-dimensional scaling (MDS) plot of RNA-seq data for all six samples. D. Boxplot of the exon expression levels (log10 transformed FKPM) for all six samples. E. Dispersion plot generated using DEXseq from randomly subsampled exons.

biological replicates were similar to each other and different knockdown groups had different expression profiles (Fig. 2C). To visualize the distribution of gene expression level for each sample, a boxplot was generated for each of the samples using log10 transformed FPKM values from Cufflinks (Fig. 2D). The quartiles and overall range were consistent between biological replicates, indicating that the data were reproducible and of high quality.

Alternative splicing events were identified using the DEXSeq (v1.10.6) package, which models the exon counts as the negative binomial distribution and uses generalized linear models to test differential exon-usage using variance estimated from biological replicates [6]. Sequence depths were used to normalize the exon counts for each sample followed by dispersion estimation before detecting alternative splicing events. Dispersion plots showed that most of the final exon level dispersions follow the fitted line as expected (Fig. 2E). Exons with a false discovery rate < 0.05, log2 fold change greater than 1.2, and a minimum of 10 counts for at least one of the samples were considered differentially expressed. Alternative splicing events with chromosome coordinates overlapping with constitutive exons (Ensemble Release version 75) by at least 1 bp were filtered out using the Bioconductor GenomicRanges (v 1.18.1) package. Exons shorter than 12 bp or belonging to differentially expressed genes were filtered out from the list. DEXSeq creates a set of disjoint exon fragments (pseudo exon) from the Ensemble exon annotation file. Alternative spliced pseudo exons were removed from the list.

3. Discussion

Here, we described RBFOX2 and CPSF2 iCLIP datasets and their effects on alternative splicing at internal exons and introns. Our results show that CPSF along with SYMPK is involved in promoting alternative splicing at internal introns and exons and shed light on a new role of mRNA 3′ end formation factors [4].

Conflict of interest

The authors declare no conflict of interest.

Acknowledgments

We thank Sara Deibler for editorial assistance. This work was supported by a grant from the National Institutes of Health (R01GM035490) to M.R.G. M.R.G. is an investigator of the Howard Hughes Medical Institute.

References

[1] J. Konig, K. Zarnack, G. Rot, T. Curk, M. Kayikci, B. Zupan, D.J. Turner, N.M. Luscombe, J. Ule, iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nat. Struct. Mol. Biol. 17 (2010) 909–915.
[2] E. Kopylova, L. Noe, H. Touzet, SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics 28 (2012) 3211–3217.
[3] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, S.L. Salzberg, TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14 (2013) R36.
[4] A. Misra, J. Ou, L.J. Zhu, M.R. Green, Global promotion of alternative internal exon usage by mRNA 3′-end formation factors. Mol. Cell 58 (2015) 819–831.
[5] T. Lassmann, Y. Hayashizaki, C.O. Daub, SAMStat: monitoring biases in next generation sequencing data. Bioinformatics 27 (2011) 130–131.
[6] C. Trapnell, D.G. Hendrickson, M. Sauvageau, L. Goff, J.L. Rinn, L. Pachter, Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat. Biotechnol. 31 (2013) 46–53.
[7] S. Anders, P.T. Pyl, W. Huber, HTSeq — a python framework to work with high-throughput sequencing data. Bioinformatics 31 (2014) 166–169.
[8] L. Goff, C. Trapnell, D. Kelley, cummeRbund: Analysis, Exploration, Manipulation, and Visualization of Cufflinks High-throughput Sequencing Data. R Package Version 2.8.2. 2013.