BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 12 2013, pages 1565–1567
doi:10.1093/bioinformatics/btt171

Sequence analysis

Advance Access publication April 4, 2013

Exome-based analysis for RNA epigenome sequencing data

Jia Meng[1,2,3,*], Xiaodong Cui[4], Manjeet K. Rao[5,6], Yidong Chen[6,7] and Yufei Huang[4,7,*]

[1]Picower Institute for Learning and Memory, [2]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, [3]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, MA 02139, [4]Department of Electrical and Computer Engineering, University of Texas at San Antonio, San, Antonio, TX 78249, [5]Department of Cellular & Structural Biology, [6]Greehey Children's Cancer Research Institute and [7]Department of Epidemiology and Biostatistics, University of Texas Health Science Center at San Antonio, San Antonio, TX 78229, USA

Associate Editor: Inanc Birol

ABSTRACT

Motivation: Fragmented RNA immunoprecipitation combined with RNA sequencing enabled the unbiased study of RNA epigenome at a near single-base resolution; however, unique features of this new type of data call for novel computational techniques.

Result: Through examining the connections of RNA epigenome sequencing data with two well-studied data types, ChIP-Seq and RNA-Seq, we unveiled the salient characteristics of this new data type. The computational strategies were discussed accordingly, and a novel data processing pipeline was proposed that combines several existing tools with a newly developed exome-based approach 'exomePeak' for detecting, representing and visualizing the post-transcriptional RNA modification sites on the transcriptome.

Availability: The MATLAB package 'exomePeak' and additional details are available at http://compgenomics.utsa.edu/exomePeak/.

Contact: yufei.huang@utsa.edu or jmeng@mit.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 8, 2012; revised on April 3, 2013; accepted on April 5, 2013

1 INTRODUCTION

Despite the unprecedented prosperity in epigenetics studies of DNA and histone modifications with next-generation sequencing data (Bernstein *et al*., 2007; Thurman *et al*., 2012), the RNA epigenetics remains a largely uncharted territory (He, 2010) and has not benefitted as much from the advancement in sequencing technology until lately. Two recent studies of transcriptome-wide mRNA m^6A methylation (Dominissini *et al*., 2012; Meyer *et al*., 2012) proposed a new powerful protocol (differently named as 'm^6A-Seq' and 'MeRIP-Seq'), where mRNA is fragmented before the immunoprecipitation with anti-m^6A antibody, and the immunoprecipitated and input control fragments are then sequenced for reconstructing transcriptome-wide m^6A methylation sites. This protocol in theory enabled the transcriptome-wide unbiased study of a repertoire of >100 known post-transcriptional RNA modifications (Cantara *et al*., 2011) at a near single-base resolution, provided the corresponding antibody is available. To unify the nomenclatures, we call this protocol 'FRIP-Seq', which stands for 'Fragmented RNA ImmunoPrecipitation Sequencing'. FRIP-Seq is different from RIP-Seq (Zhao *et al*., 2010), where *full-length* RNA is subjected to immunoprecipitation for detection of protein–RNA interaction. In contrast, the immunoprecipitation of *fragmented* RNA in FRIP-Seq enables the precise prediction of RNA modification sites on the transcriptome.

From a technological perspective, FRIP-Seq can be considered a marriage of two relatively well-studied techniques: ChIP-Seq (Kidder *et al*., 2011) and RNA-Seq (Garber *et al*., 2011). This new technique brings a host of new computational challenges yet to be adequately addressed. Next, we discuss briefly the best practice for FRIP-Seq data analysis.

Mapping and Filtering Short Reads: As FRIP-Seq sequences mRNA, spliced aligners that allow reads to span exon–exon junctions should be implemented. One important issue is the widespread repetitive elements (Treangen and Salzberg, 2012) in a broad range of species (~50% of the human genome) that can lead to multi-reads (reads that could be mapped to multiple genomic locations) and ambiguities in alignment. Of the various existing strategies, the simplest yet effective way is to exclude the multi-reads completely from the analysis. (See Supplementary Material for detailed discussion.)

Fragment Length and Shifting Size: Currently, the most popular RNA sequencing protocol (unstranded RNA library and single-end sequencing) produces two shifted peaks on the '+' and '-' strands with a distance equal to the fragment length, and this pattern is also observed in FRIP-Seq. To correctly predict the precise methylation sites, reads need to be shifted by half of the fragment length or extended to the full length towards the 3′-end. In case that the fragment length is unknown, it may be estimated from the bimodal pattern (Zhang *et al*., 2008) or the cross-strand correlation (Kharchenko *et al*., 2008).

Peak Calling, Sequencing Bias and Control Sample: The detection of interaction sites has been formulated as the peak detection problem in ChIP-Seq (Micsinai *et al*., 2012; Wilbanks and Facciotti, 2010). Different from the mild sequencing bias in ChIP-Seq owing to nucleosome loss around the transcription starting sites, FRIP-Seq suffers from the depletion at both 5′- and 3′-end as a result of RNA fragmentation, considerable variations of expression levels for different genes, and most importantly, the positional bias on the gene locus caused by different isoform transcripts. Although ChIP-Seq peak calling can be conducted in the absence of a control sample by

*To whom correspondence should be addressed.

estimating the background from the neighborhood genomic regions, FRIP-Seq peak calling does require a paired input control sample of preferably fragmented RNAs before immunoprecipitation as supposed to an immunoglobulin G control sample.

Peak Annotation, Gene and Isoform Transcripts: The association between detected RNA modification sites and the specific mRNA transcripts can be highly problematic. Recent study shows that with an average of 10–12 isoforms per gene, most genes tend to express multiple isoforms simultaneously (Djebali *et al*., 2012). Since the sequencing read length is mostly <100bp, isoform quantification can be difficult, not to mention calling peaks for each individual isoform transcript. Nevertheless, an mRNA methylation site may be uniquely associated with a transcript when the site spans across the nearest exon(s) that uniquely belong to that transcript.

2 METHODS

We describe in the following a pipeline for the analysis of FRIP-seq data that combines several existing tools with a novel exome-based peak calling approach (Fig. 1a). In this pipeline, short reads are first aligned to the genome assembly with Tophat (Trapnell *et al*., 2009); then, the resulting BAM files are filtered, sorted and indexed with Samtools (Li *et al*., 2009); in the last, RNA methylation sites are predicted with 'exomePeak', a MATLAB package that outputs exome-based peaks (the genomic locus of RNA modification sites) in the BED format to facilitate the visualization in Integrative Genomics Viewer (IGV) (Robinson *et al*., 2011). See the Supplementary Material for an example.

The proposed exome-based approach can be considered as projecting transcriptome onto the genome, effectively avoiding the transcriptome heterogeneity. The advantages of exome-based analysis are mainly in two folds: first, working on exome avoids the isoform-related ambiguity, hence making all the computational operations (such as shifting, extension, smoothing, testing) straightforward. Second, it helps annotate the detected methylation sites. Despite the increasing attention on studying isoforms (Pearson, 2006), the gene convention is still at the center of most annotations and descriptions. A detected RNA modification site (on a particular transcript) needs to be associated first with a gene name before its function can be predicted. As a result, the developed 'exomePeak' package calls peaks on pooled exons of a specific gene so that the called peaks are automatically associated with that gene and hence its related functions.

More specifically, exomePeak first extracts and connects all the exons of a specific gene and then detects peaks using a sliding window with a conditional test that compares the means of two Poisson distributions (Przyborowski and Wilenski, 1940). The peak enrichment is considered for the IP versus control sample and for locus versus the entire gene in the IP sample. The former indicates a peak enriched in the IP sample, whereas the latter shows a locus-specific enrichment. The two *P*-values are then combined by the Fisher's or Stouffer's method. ExomePeak also swaps the IP and control samples to calculate False Discovery Rate (FDR). (See the Supplementary Material for detailed discussion.)

3 RESULTS

Our pipeline was applied to the data from (Meyer *et al*., 2012) and found at least one RNA m^{6}A site on 9218 genes in HEK293T cells at a significance level of P-value $= 10^{-5}$ (See the Supplementary Material for more details). We then retained

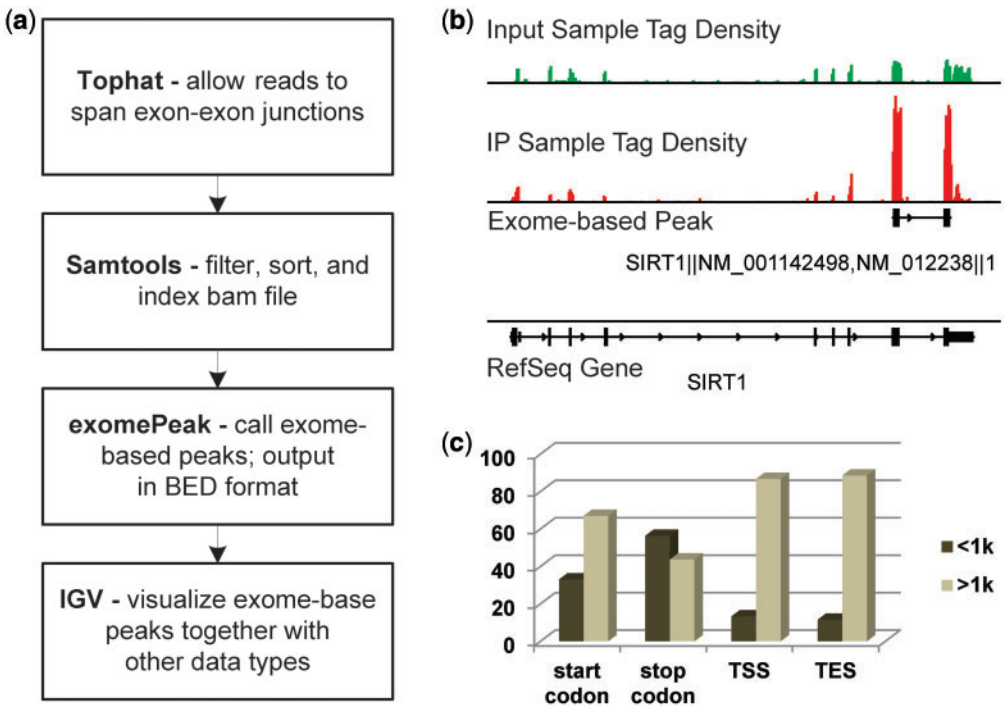

Fig. 1. Exome-based analysis of RNA epigenome sequencing data. (**a**) The pipeline for FRIP-Seq processing. (**b**) An m^{6}A site is found near the stop codon of Sirt1, which has two RefSeq isoform transcripts. The single RNA peak was split into two genomic loci when projected into the genome. (**c**) For the 3721 protein-coding genes with at least one m^{6}A site, at least 10 kb exons, and only one RefSeq transcript, 56.3% (5195 of 9228) m^{6}A sites fall within 1 kb distance to its stop codon on mRNA

only the protein-coding genes with coding length at least 10kb long and single RefSeq isoform, and then calculated the geometric distances between each detected m^6A sites and four transcript landmarks of the respective gene, including the start codon, stop codon, transcription starting site and transcription ending site. Result shows that the m^6A sites are enriched within 1 kb distance around the stop codons (Fig. 1c), consistent with previous studies (Dominissini *et al*., 2012).

Funding: The William and Ella Medical Research Foundation grant to MKR; National Institutes of Health (NIH-NCI P30CA54174) to YC; National Science Foundation (CCF-0546345) to YH; Qatar National Research Fund (09-874-3-235) to YC and YH. We thank computational support provided by the UTSA Computational Systems Biology Core Facility (NIH RCMI grant 5G12RR013646-12).

Conflict of Interest: none declared.

REFERENCES

Bernstein,B.E. *et al*. (2007) The mammalian epigenome. *Cell*, **128**, 669–681.

Cantara,W.A. *et al*. (2011) The RNA modification database, RNAMDB: 2011 update. *Nucleic Acids Res*., **39**, D195–D201.

Djebali,S. *et al*. (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–108.

Dominissini,D. *et al*. (2012) Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature*, **485**, 201–206.

Garber,M. *et al*. (2011) Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Methods*, **8**, 469–477.

He,C. (2010) Grand challenge commentary: RNA epigenetics? . *Nat. Chem. Biol*., **6**, 863–865.

Kharchenko,P.V. *et al*. (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol*., **26**, 1351–1359.

Kidder,B.L. *et al*. (2011) ChIP-Seq: technical considerations for obtaining high-quality data. *Nat. Immunol*., **12**, 918–922.

Li,H. *et al*. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Meyer,K.D. *et al*. (2012) Comprehensive analysis of mRNA methylation reveals enrichment in 3′ UTRs and near stop codons. *Cell*, **149**, 1635–1646.

Micsinai,M. *et al*. (2012) Picking ChIP-seq peak detectors for analyzing chromatin modification experiments. *Nucleic Acids Res*., **40**, e70.

Pearson,H. (2006) What is a gene? *Nature*, **441**, 398–401.

Przyborowski,J. and Wilenski,H. (1940) Homogeneity of results in testing samples from Poisson series: with an application to testing clover seed for dodder. *Biometrika*, **31**, 313–323.

Robinson,J.T. *et al*. (2011) Integrative genomics viewer. *Nat. Biotechnol*., **29**, 24–26.

Thurman,R.E. *et al*. (2012) The accessible chromatin landscape of the human genome. *Nature*, **489**, 75–82.

Trapnell,C. *et al*. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.

Treangen,T.J. and Salzberg,S.L. (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet*., **13**, 36–46.

Wilbanks,E.G. and Facciotti,M.T. (2010) Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS One*, **5**, e11471.

Zhang,Y. *et al*. (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol*., **9**, R137.

Zhao,J. *et al*. (2010) Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol. Cell*, **40**, 939–953.