Next-generation sequencing-based 5’ rapid amplification of cDNA ends for alternative promoters

Bambarendage P.U. Perera
Louisiana State University

Joomyeong Kim
Louisiana State University

Follow this and additional works at: https://repository.lsu.edu/biosci_pubs

Recommended Citation
Notes & tips

Next-generation sequencing-based 5′ rapid amplification of cDNA ends for alternative promoters

Bambarendage P.U. Perera, Joomyeong Kim*

Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, USA

ARTICLE INFO

Article history:
Received 26 October 2015
Received in revised form
3 November 2015
Accepted 9 November 2015
Available online 23 November 2015

Keywords:
5′ RACE
NGS
Alternative promoters
Peg3

ABSTRACT

Mammalian genomes contain many unknown alternative first exons and promoters. Thus, we have modified the existing 5′RACE (5′ rapid amplification of cDNA ends) approach into a next-generation sequencing (NGS)-based new protocol that can identify these alternative promoters. This protocol has incorporated two main ideas: (i) 5′RACE starting from the known second exons of genes and (ii) NGS-based sequencing of the subsequent cDNA products. This protocol also provides a bioinformatics strategy that processes the sequence reads from NGS runs. This protocol has successfully identified several alternative promoters for an imprinted gene, Peg3. Overall, this NGS-based 5′RACE protocol is a sensitive and reliable method for detecting low-abundant transcripts and promoters.

In eukaryotic cells, gene transcription is regulated through several cis-regulatory DNA elements such as promoters, enhancers, silencers, and insulators [1]. Identifying promoters is one of the initial steps for characterizing transcriptional regulation for any given gene. In model organisms, the promoters of the majority of genes have been identified through large-scale sequencing efforts [2]. According to recent results, however, transcripts are quite often detected from the upstream regions of many genes that are outside of the previously defined transcribed regions. This indicates that additional transcription start sites and alternative promoters may exist for many genes [3]. In the current study, thus, we have modified and adapted the existing 5′RACE (5′ rapid amplification of cDNA ends)1 [4] approach into a next-generation sequencing (NGS)-based new protocol that can identify the alternative promoters for these transcripts. The main strength of this new protocol is its sensitivity; this can detect the 5′ ends of mRNA that are of very low abundance. The main ideas and detailed steps for this protocol are described below.

Two main ideas have been incorporated into this new protocol. First, the initial cDNA for 5′RACE needs to be prepared with the reverse transcription (RT) reaction using gene-specific oligonucleotide primers (Fig. 1A). These gene-specific primers also need to be derived from the known second exons, but not from the known first exons, for genes. The known first exons are not a good choice because these exons do not have splicing acceptor sites at their 5′ ends. Instead, they have transcription start sites; thus, they cannot be connected to any upstream exons. By contrast, the known second exons have proper splicing acceptor sites; thus, they should be ideal regions for anchoring the upstream alternative exons that have not been identified so far. Second, potential unknown alternative first exons most likely represent the sequences that belong to very low-abundant transcripts given all of the sequencing efforts by the genomic community. Therefore, this new protocol uses NGS-based approaches to identify the sequences derived from these very minor transcripts. NGS-based sequencing should allow the identification of these low-abundant transcripts given their sequencing capacity (million reads per library). This new protocol, therefore, includes several steps that can easily convert polymerase chain reaction (PCR) products into NGS libraries.

The detailed steps for the new protocol are as follows (Fig. 1A; see also Supplementary Material 1 in the online supplementary material). In step 1, the total RNA isolated from tissue samples is reverse-transcribed with gene-specific primers (dark arrow in Fig. 1B). The subsequent cDNA needs to be purified through phenol/chloroform extraction followed by ethanol precipitation. In step 2, the cDNA is modified through the tailing reaction using deoxyguanosine triphosphate (dTTP) and terminal deoxynucleotidyl transferase (TdT). In step 3, a pool of the G-tailed cDNA is targeted

Abbreviations: 5′RACE, 5′ rapid amplification of cDNA ends; cDNA, complementary DNA; NGS, next-generation sequencing; RT, reverse transcription; mRNA, messenger RNA; PCR, polymerase chain reaction; dGTP, deoxyguanosine triphosphate; TdT, terminal deoxynucleotidyl transferase; Peg3, paternally expressed gene 3; UCSC, University of California, Santa Cruz.

* Corresponding author.
E-mail address: jkim@lsu.edu (J. Kim).

© 2015 Elsevier Inc. All rights reserved.

http://dx.doi.org/10.1016/j.ab.2015.11.006
0003-2697/© 2015 Elsevier Inc. All rights reserved.
and ligated to barcoded adapters as shown by the green boxes. A set of several 5 primer: an internal gene-specific primers; two gene-specific primers (gray arrows) and two primers targeting the G-tailed portion of cDNA (XC and X; blue arrows). A small amount of PCR products from each PCR needs to be separated on a 2% agarose gel to monitor the proper amplification of cDNA. In step 4, the amplified product from the second PCR is modified with the end repair reaction and later with the ligation reaction to be analyzed with NGS-based sequencing. The current protocol has been tested multiple times using an NGS platform (PGM2, Ion Torrent), which requires two specific primers: X and the c primer used for the second PCR of the nested scheme (Fig. 1B). This sorting process can be executed through several Unix command lines, which has been included as Supplementary Material 1. These filtered reads are further processed to identify alternative exons using a PERL script, which can detect and count the number of raw reads that have different combinations of exon joining with the initial exon (E2 in Fig. 1B). In step 7, each group of the raw reads displaying different exon joining can be mapped to the genome sequence of a given species using the UCSC (University of California, Santa Cruz) genome browser, which will then identify new alternative exons. This bioinformatics pipeline along with one PERL script has been included as Supplementary Material 2.

The feasibility of this protocol has been tested through identifying alternative promoters for an imprinted gene called PEG3 (paternally expressed gene 3) (Fig. 2). This gene is composed of 9 exons that are spread throughout 25-kb genomic regions in both human and mouse, and it is highly expressed in the brain [5–7]. Interestingly, the 200-kb upstream region of this gene is well conserved among all of the mammals without any additional open reading frames (ORFs); thus, this region has been suspected to harbor other unknown cis-regulatory elements [8]. This new protocol, therefore, has been applied to test whether this region contains any unknown alternative promoters. First, the total RNA from mouse and human brains was individually used for generating 5’RACE cDNA libraries, and later these libraries were sequenced using an NGS platform. Each library derived on average several hundred thousand reads, and these raw reads were further analyzed with the bioinformatics strategy described previously. According to the results, the majority of cDNA sequences were indeed derived from the transcripts with the known exon combination (E1–E2): 93% for mouse and 74% for human PEG3 (Fig. 2B). However, this protocol also detected some minor transcripts from both mouse and human, and these transcripts were very low in abundance based on their representations in the libraries: 0.33% for mouse and 0.13% for human PEG3. Detailed inspection further revealed that these minor transcripts skip the known first exons but are connected to previously unknown alternative exons (U1 in Fig. 2A). It also revealed that the new U1 exons of PEG3 are not conserved between mouse and human genomic regions. Individual RT–PCR analyses indeed confirmed the presence of these minor transcripts for mouse and human PEG3 (Fig. 2C). Until now, there has never been any clue suggesting the presence of alternative promoters for the PEG3 locus, although this locus has been intensively studied for more than two decades. Thus, this new protocol
Fig. 2. Alternative transcripts and promoters identified for mouse and human PEG3. A Genomic structure of PEG3. The black and blue boxes represent the 9 previously known exons and the newly identified alternative exons of PEG3, respectively. The bent black arrow indicates the transcriptional direction of PEG3. The E1–E2 primer combination targets the transcript containing the known first and second exons of PEG3. The mouse U1–E2 transcript harbors the newly identified alternative first exon but skips the known first exon of PEG3. The human U1–E2 transcript starts from the newly identified alternative first exon, which also skips the known first exon of PEG3. B Relative frequency of alternative transcripts for mouse and human PEG3. The table summarizes the relative representation (%) of the transcripts with E1 (known first exons) and U1 (alternative first exons) of PEG3. The relative representations (%) are calculated through dividing the number of reads containing the E1, mU1 (mouse), and hU1 (human) exons by the total number of reads for each library. The total number of reads for each library represents the sorted sequences that still contain both the G-tail and gene-specific exons. C RT–PCR analyses for alternative exons. The mouse U1–E2 primer combination detects the transcript with the newly found alternative first exon connected to the second exon of PEG3. The top and bottom panels represent the RT–PCR products from mouse and human brains, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Acknowledgments

We thank Drs. Scott Herke and Suman Lee for their help with NGS preparation and sequencing runs. We also thank the members of the JooKim lab for their discussion of the manuscript. This work was supported by the National Institutes of Health (R01-GM066225 and R01-GM097074).

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jb.2015.11.006.

References