High-throughput verification of transcriptional starting sites by Deep-RACE

Olivarius, Signe; Plessy, Charles; Carninci, Piero

Published in:
BioTechniques

DOI:
10.2144/000113066

Publication date:
2009

Document Version
Publisher's PDF, also known as Version of record

High-throughput verification of transcriptional starting sites by Deep-RACE

Signe Olivarius*, Charles Plessy, and Piero Carninci
Functional Genomics Technology Team, Omics Science Center, RIKEN Yokohama Institute

BioTechniques 46:130-132 (February 2009) doi 10.2144/000113066
Keywords: promoters; transcription start sites; high-throughput; RACE; short reads sequencing

*S.O.'s present address is Molecular Evolution Group, Department of Biology, University of Copenhagen, Copenhagen, Denmark.

We present a high-throughput method for investigating the transcriptional starting sites of genes of interest, which we named Deep-RACE (Deep–rapid amplification of cDNA ends). Taking advantage of the latest sequencing technology, it allows the parallel analysis of multiple genes and is free of time-consuming cloning steps. In comparison to the sequencing of RACE PCR products, our approach is more precise and more cost-effective even for batches as small as 17.

Evidence that the genome is pervasively transcribed into hundreds of thousands different RNAs (1,2) necessitates methods for independent high-throughput verification of transcriptional starting sites (TSSs) determined by genome-wide approaches. 5′-RACE PCR is a well-established and widely used method to specifically amplify the 5′ end of a transcript, facilitating mapping of the TSS and the approximate location of promoter elements (3). Conventionally, this mapping is done by cloning the 5′-RACE PCR product into a bacterial vector and sequencing a few clones by classic electrophoresis-based Sanger sequencing. A great challenge for transcriptional analysis in general is the recently recognized prevalence of mechanisms that expand the potential transcript repertoire, such as alternative splicing, alternative promoter usage, and multiple TSSs (1,4). In order for a 5′-RACE PCR assay to reflect such diversity, sequencing a fairly large number of clones is necessary, and when more than one transcript is to be analyzed by 5′-RACE PCR this procedure becomes comprehensive and time-consuming, not to mention costly. To address this drawback, we developed a simple assay in which 5′-RACE PCR products are subjected directly to high-throughput sequencing by the newly developed flow cell sequencing technologies (5) that are currently revolutionizing DNA sequencing by enabling the sequencing of hundreds of millions bases in a single sequencing run. Based on clonal amplification of millions of single surface-immobilized DNA fragments (6,7), flow cell sequencing technologies have been recognized for their aptness for ultra-high throughput approaches such as whole-genome sequencing. However, since it eliminates the need for cloning steps and facilitates quantitative assessment of transcription, the high-throughput sequencing approach is also a very appealing option for single-gene analysis such as with 5′-RACE PCR.

To explore the possibility of combining 5′-RACE PCR with high-throughput sequencing, we selected 17 genes or genomic loci displaying expression in Hep G2 cells (Catalog no. HB-8065, ATCC, Manassas, VA, USA) according to the ENcyclopedia of DNA Elements (ENCODE) Project (2) and cap-analysis gene expression (CAGE) data (1). The RNA ligase-mediated (RLM) RACE PCR approach (8,3) for specifically amplifying the 5′ end of cDNA from full-length, capped transcripts was adapted to perform high-throughput 5′-end sequencing on total RNA extracted from HepG2 cells (Figure 1). Normally, flow cell sequencing requires a sample preparation step in which sequencing adapters are ligated to both ends of the DNA fragments. Omitting this step, the primers for the inner PCR were designed with ~20-bp-long derivatives of specific adapter sequences of the Illumina Genome Analyzer (Tokyo, Japan) attached to the 5′ ends (Figure 1). Following nested PCR, the inner 5′-RACE PCR reactions for all 17 genes were pooled and purified using a standard PCR purification kit (QIAGEN, Tokyo, Japan), and the resulting mixture of PCR products

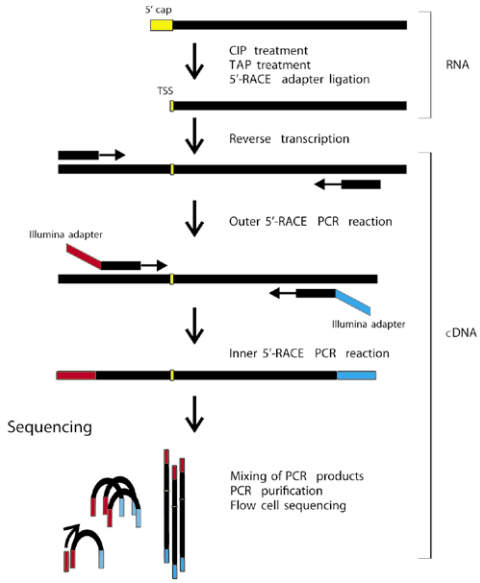

Figure 1. Schematic representation of the 5′-RACE PCR procedure optimized for the high-throughput Illumina Genome Analyzer. Total RNA from Hep G2 hepatocellular carcinoma cells was treated with calf intestine phosphatase (CIP) to remove free 5′ phosphates and with tobacco acid pyrophosphatase (TAP) to detach the cap of full-length transcripts, following the manufacturer's instructions of the FirstChoice RLM-RACE Kit (Applied Biosystems, Tokyo, Japan). The 5′ RACE adapter included in the kit was ligated to decapped molecules, and reverse transcription with random decamers was performed. Outer PCR reactions were carried out using a common adapter-specific forward primer included in the kit and custom gene-specific reverse primers. For the inner nested PCR reactions, in which the outer PCR reactions were used as templates, a common adapter-specific forward primer with the sequence **AATGATACGGCGACCACCG**AACACTGCGTTTGCTGGCTTTGATG was constructed (bold letters designate an Illumina-specific adapter sequence). The gene-specific inner reverse primers were designed with an Illumina-specific adapter sequence, **CAAGCAGAAGACGGCATACGA**, attached to the 5′ end. The resulting PCR products, designed to be approximately 100–300 bp in length, were PCR-purified and subjected to high-throughput sequencing-by-synthesis in an Illumina genome analyzer with a sequencing primer of sequence GACCACCGAACACTGCGTTTGCTGGCTTTGATG. Sequences were deposited in the NCBI Short Read Archive under accession no. SRA003626.

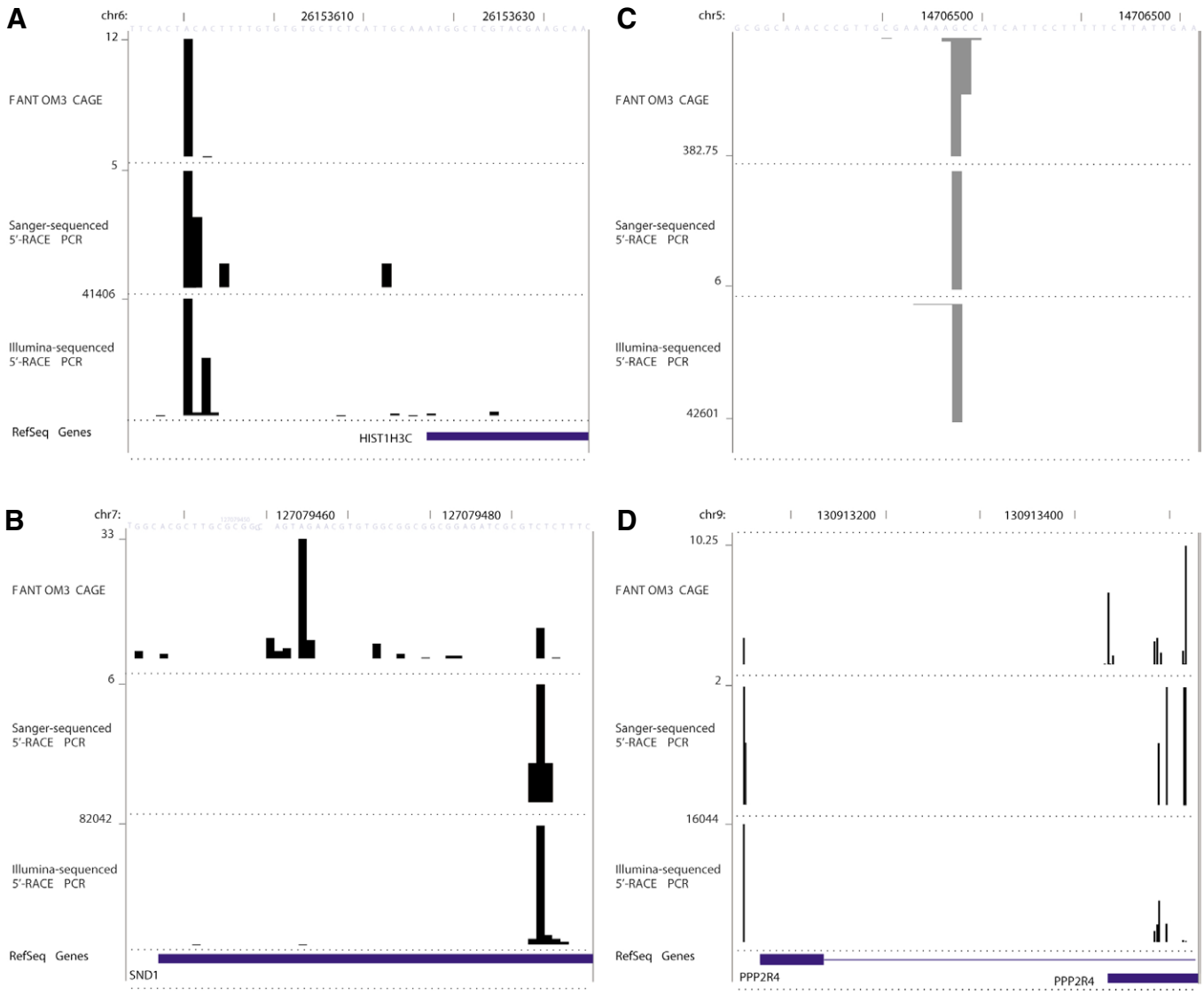

Figure 2. Transcription start sites for four human transcribed regions analyzed by Deep-RACE, classical RACE-PCR, and cap analysis of gene expression (CAGE). The TSSs for H3 histone family member C (HIST1H3C) (A), staphylococcal nuclease domain containing 1 (SND1) (B), an unannotated transcript on chromosome 5 (C), and protein phosphatase 2A, regulatory subunit B′ (PPP2R4) (D) were deduced from a conventional Sanger sequencing–based 5′-RACE PCR assay and from the high-throughput Illumina sequencing approach and compared with FANTOM3 CAGE library H22B (1). Each bar, depicting the number of sequences starting at a certain position, corresponds to a TSS. The maximum number of sequences comprising a TSS is indicated at the left. The position respective to the RefSeq gene model is indicated as a blue bar. The absolute coordinates on the human chromosomes (NCBI build 36.1) are indicated in the upper part of each panel. Note that in C, the bars are upside-down to reflect that transcription occurs on the minus strand.

was loaded in a single flow cell channel in the high-throughput sequencer Genome Analyzer. A total of 2,145,126 sequence reads were obtained and 1,280,189 of them could be mapped on 88,554 distinct TSSs of the human genome using the "nexalign" program as described by T. Lassmann et al. (manuscript in preparation). We used the Galaxy web service (galaxy.psu.edu; Reference 9) to identify the genomic intervals where most reads align. TSSs with a read count higher than 500 were selected and clustered, and the clusters were then extended by 100 bp in the 5′ and 3′ directions. This yielded 26 genomic intervals that we used to filter our original data and rescue TSSs with read counts lower than 500. Eighteen of the intervals corresponded to our loci of interest (one has two promoters), and the other corresponded to repeated regions of the genome. Each gene is covered by an average of more than 74,000 sequences. Even the gene with lowest coverage has 3,195 tags, suggesting that up to a few hundreds of different, non-overlapping transcriptional starting sites or genes could be investigated with a good chance of success.

To compare this Deep-RACE assay with the classic 5′-RACE PCR procedure based on Sanger sequencing, 9 of the 17 outer 5′-RACE PCR reactions were used as templates in inner 5′-RACE PCR reactions using primers without Illumina-specific adapter sequences. The resulting PCR products were cloned in TOPO2.1 vectors (Invitrogen), and 192 colonies (19–24 per gene) were picked and used as templates in PCR reactions using vector-specific primers. Afterwards, the colony PCR products were subjected to capillary electrophoresis sequencing in an ABI 3700 sequencer (Applied Biosystems, Tokyo, Japan). On average, 12 usable sequences per gene were obtained by the Sanger sequencing approach. The sequence reads were aligned to the genome

using the BLAT server of the University of California-Santa Cruz (UCSC) genome browser (genome.ucsc.edu; Reference 10). As expected, the positions of the main TSSs deduced from the two methods correspond well (Figure 2), and where a TSS flexibility can be observed, this is reflected in both assays (Figure 2, A and D). The average number of TSSs per gene is 5, according to the Sanger sequencing assay, while there are twice as many when the same genes are analyzed by high-throughput sequencing. Thus, although both methods appear to be suitable for qualitative assessment of TSSs, classic Sanger sequencing comprises risks of neglecting rarely used TSSs. Such rare TSSs have been discovered in the past using high-throughput methods like CAGE (1), and while their biological function is still mostly unexplored, they are valuable signals that can only be detected by deep sequencing. To assess the capacity of Deep-RACE to detect rare TSSs, we compared our results to the mapping of 811,195 CAGE tags from the HepG2 library H22B of the FANTOM3 project (fantom.gsc.riken.jp; Reference 1). CAGE libraries detect comparable amounts of TSSs, except in some cases where there is a global discrepancy of the results (Figure 2B). Qualitative differences between the three techniques compared here can be explained by the specificities of their methodologies. For instance, the CAGE library uses random priming, whereas 5′ RACE relies on a gene-specific primer, so the TSS of isoforms lacking sequences complementary to gene-specific primer will not be visible with RACE. Sequencing methods can also explain some discrepancies: while Deep-RACE does not depend on cloning, classical RACE methods contain some steps where the DNA has to be transfected in bacterial vectors. Sequences easier to clone can be favored, which would explain the occurrences of TSSs detected in classical RACE but not in the high-throughput CAGE and Deep-RACE methods.

In summary, we have shown that it is possible to combine transcriptional analysis by 5′-RACE PCR with high-throughput sequencing, thereby dramatically increasing the quantity of sequence data compared with a conventional 5′-RACE PCR assay while also saving considerable amounts of time and effort. By subjecting a mixture of RACE PCR products directly to sequencing in a second generation sequencer, an elaborate cloning procedure is omitted, and moreover, construction of primers with sequencing adapters removes the need for an overnight ligation step prior to sequencing. The simplicity of the method facilitates parallel TSS analysis of potentially hundreds of genes, and since millions of sequences are obtained in a single run, highly reliable assessments of TSS positions and relative usage should be gained even when a large number of genes are analyzed simultaneously. Finally, since the sequencing reaction occupies only a single flow cell channel regardless of the number of 5′-RACE PCR reactions, high-throughput sequencing of 5′-RACE PCR products is likely to be economical compared with Sanger sequencing. By sequencing only 17 genes with Deep-RACE, the cost is equivalent with the Sanger method (in terms of reagents alone, and not taking into account the cost of labor), suggesting that high-throughput sequencing will be very cost-effective for large-scale TSS verification. In comparison to global approaches, such as whole transcriptome shotgun sequencing (11,12), Deep-RACE needs at least one order of magnitude less sequencing, produces tags only from templates that are full-length in 5′, and works directly on total RNA. Also, Deep-RACE results are easy to analyze with graphical tools such as Galaxy (9). Hence, we predict that the presented method will be the tool of choice for parallel TSS analysis of multiple genes.

Acknowledgments

This work was supported by a grant from the 6th Framework of the European Union commission to the Neuro Functional Genomics consortium, a RIKEN presidential grant for collaborative research, and a grant-in-aid ("A Systematic Systems Biology Approach to Neural Plasticity") from the Japanese Society for the Promotion of Science. We thank our sequencing team for their support.

The authors declare no competing interests.

References

1. **Carninci, P., A. Sandelin, B. Lenhard, S. Katayama, K. Shimokawa, J. Ponjavic, C.A. Semple, M.S. Taylor, et al.** 2006. Genome-wide analysis of mammalian promoter architecture and evolution. Nat. Genet. *38*:626-635.
2. **ENCODE Project Consortium, E. Birney, J.A. Stamatoyannopoulos, A. Dutta, R. Guigó, T.R. Gingeras, E.H. Margulies, Z. Weng, et al.** 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447*:799-816.
3. **Frohman, M.A.** 1994. On beyond classic RACE (rapid amplification of cDNA ends). PCR Methods Appl. *4*:S40-S58.
4. **Irimia, M., J.L. Rukov, D. Penny, and S.W. Roy.** 2007. Functional and evolutionary analysis of alternatively spliced genes is consistent with an early eukaryotic origin of alternative splicing. BMC Evol. Biol. *7*:188.
5. **Holt, R.A. and S.J. Jones.** 2008. The new paradigm of flow cell sequencing. Genome Res. *18*:839-846.
6. **Aksyonov, S.A., M. Bittner, L.B. Bloom, L.J. Reha-Krantz, I.R. Gould, M.A. Hayes, U.A. Kiernan, E.E. Niederkofler, et al.** 2006. Multiplexed DNA sequencing-by-synthesis. Anal. Biochem. *348*:127-138.
7. **Ju, J., D.H. Kim, L. Bi, Q. Meng, X. Bai, Z. Li, X. Li, M.S. Marma, et al.** 2006. Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. Proc. Natl. Acad. Sci. USA *103*:19635-19649.
8. **Fromont-Racine, M., E. Bertrand, R. Pictet, and T. Grange.** 1993. A highly sensitive method for mapping the 5′ termini of mRNAs. Nucleic Acids Res. *21*:1683-1684.
9. **Giardine, B., C. Riemer, R.C. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang, D. Blankenberg, et al.** 2005. Galaxy: A platform for interactive large-scale genome analysis. Genome Res. *15*:1451-1455.
10. **Kent, W.J.** 2002. BLAT—The BLAST-like alignment tool. Genome Res. *12*:656-664.
11. **Mortazavi, A., B.A. Williams, K. McCue, L. Schaeffer, and B. Wold.** 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods *5*:621-628.
12. **Cloonan, N., A.R. Forrest, G. Kolle, B.B. Gardiner, G.J. Faulkner, M.K. Brown, D.F. Taylor, A.L. Steptoe, et al.** 2008. Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat. Methods *5*:613-619.

Received 9 September 2008; accepted 12 November 2008.

Address general correspondence to Piero Carninci, Functional Genomics Technology Team, Omics Science Center, RIKEN Yokohama Institute, 1-7-22 Suehiro-chô, Tsurumi-ku, Yokohama, 230-0045 Kanagawa, Japan, e-mail: rgscerg@gsc.riken.jp; and correspondence regarding experiments to Charles Plessy, Functional Genomics Technology Team, Omics Science Center, RIKEN Yokohama Institute, 1-7-22 Suehiro-chô, Tsurumi-ku, Yokohama, 230-0045 Kanagawa, Japan. e-mail: plessy@riken.jp