Minireview

From transcription start site to cell biology

Philipp Kapranov

Address: Helicos BioSciences Corporation, One Kendall Square Building 700, Cambridge, MA 02139, USA. Email: philippk08@gmail.com

Published: 20 April 2009

Genome ***Biology*** 2009, **10:**216 (doi:10.1186/gb-2009-10-4-217)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2009/10/4/217

Abstract

The regulation of transcription is a complex process. Recent novel insights concerning the *in vivo* regulation and expression of protein-coding and non-coding RNAs have added previously unimagined levels of complexity to these processes.

Knowledge of the exact position of a 5’ transcriptional start site (TSS) of an RNA molecule is crucial for the identification of the regulatory regions that immediately flank it. Traditionally, the most reliable method of identifying a TSS is to map a nucleotide to which a 5’ cap structure is added in the RNA. Over the past few years this approach has been used in a number of genome-wide surveys aimed at unbiased identification of TSSs (see [1,2] and references therein). These surveys identified many more sites where 5’ ends of capped RNAs could be mapped than those TSSs belonging to annotated genes. At the same time, large amounts of unannotated transcription had been detected in mammalian genomes [2-4] and numerous transcription factor binding sites found outside annotated promoter regions [5,6]. In addition, multiple start sites are often found for annotated, protein-coding genes very far from their ‘official’ start sites [2,7,8].

Three papers published recently in *Nature Genetics* by members of the FANTOM (Functional Annotation of Mouse) consortium [9-11] reveal yet further complexity of transcription initiation in animal genomes. Taft *et al.* [9] describe a new class of short RNAs made at promoters, while Faulkner *et al.* [10] show that repetitive elements can be a rich source of novel promoters. A study from the FANTOM consortium and the RIKEN Omics Science Center [11] shows how information on the precise positions of TSSs can be used to characterize global gene regulatory networks operating during cell differentiation.

How to identify a transcription start site

The critical issue in mapping a true site of transcription initiation is to be able to distinguish it from a 5’ end generated by RNA cleavage or degradation and from a 5’ end generated by incomplete copying of RNA into cDNA. The conventional hallmark of TSSs in most eukaryotes is addition of a 7-methyl guanosine cap structure to the 5’-triphosphate of the first base transcribed by RNA polymerase II. This unique feature of the transcription initiation nucleotide is the basis of several methods aiming to enrich and identify capped messages and subsequently to map the exact positions in the genome of the nucleotides to which the cap is added. The main methods used are cap analysis of gene expression (CAGE) [12], oligo-capping [13] and robust analysis of 5’-transcript ends (5’-RATE) [14]. CAGE is the most commonly used and exploits the 2’,3’-diol structure of the cap nucleotide, which is only present in only one other place on an RNA molecule besides the cap - its extreme 3' end. The diol structure is susceptible to a specific chemical oxidation which can be followed by biotinylation, enabling selection of capped messages by immunoprecipitation with streptavidin. The enriched capped RNA fraction is then converted into cDNAs that span the entire lengths of the capped RNA molecules. Oligo-capping and 5’-RATE take advantage of the fact that the 5’ cap is resistant to phosphatase treatment, which removes mono-, di- or triphosphates from cleaved or degraded RNA. Subsequent removal of the cap using tobacco acid pyrophosphatase leaves a 5’-monophosphate, which is amenable to ligation with a specific linker nucleotide that marks the position of the native 5’ end of RNA and can later be used to select and sequence the 5’ ends of capped cDNAs [13,14].

Full-length cDNAs generated by the techniques described above can be further converted into short DNA tags derived from their 5' ends [12,13,15], which are very suitable for next-generation sequencing [16]. The combination of cap-selection and next-generation sequencing can generate sequence information about the exact positions of cap-addition sites for millions of RNA molecules [4,15,17], thus making it possible to obtain digital information about the number of transcriptional initiation events occurring at any genomic position. This information can be used to infer the positions, as well as the relative strengths, of different promoter elements [15], as exemplified in the recent articles from the FANTOM consortium [9-11]. It can also be correlated with information on the positions of other annotated genomic elements, such as repetitive elements [10] or short RNAs [9,18], to identify any association between these elements and transcription initiation.

Complex transcriptional activity around TSSs

The immediate vicinity of a TSS is active ground for the production of a number of RNAs other than those destined to become full-length, protein-coding mRNAs. These RNAs can be transcribed from both DNA strands [19,20] and tend to be either short [19,18,21] or short-lived and are quickly degraded by the exosomal complex [22,23]. Working with the *Drosophila*, human and chicken genomes, Taft *et al.* [9] have now added a new class of promoter-related small RNAs, dubbed ‘tiny RNAs’, which map within -60 to +120 nucleotides around a TSS, with a peak density at 10-30 nucleotides downstream of the TSS. The size of the tiny RNAs, whose length distribution peaks at 18 nucleotides, distinguishes them from the larger promoter-associated short RNAs (PASRs) [19] and other RNAs generated at or near a promoter [21,22]. The tiny RNAs can be mapped mainly to the sense strand of the longer transcript and, like PASRs, they tend to be found in the promoters of expressed genes and associated with active chromatin marks [9].

An important question is whether any of the non-coding RNAs found at or near promoters and TSSs have any biological function, or whether they simply represent byproducts of stalled polymerases or the degradation of longer mRNAs. Several lines of evidence argue against the latter two explanations. First, the observation by Taft *et al.* [9] in *Drosophila* that only a fraction of tiny RNAs associate with promoters that show evidence of stalled RNA polymerase argues against abortive transcription as their sole source. Taft *et al.* [9] also establish that production of tiny RNAs and PASRs at promoters is common in organisms as diverse as humans and flies, and that their relative positions in the genome tend to be syntenically conserved between between humans and chickens, similarly to PASRs that are syntenically conserved between humans and mice [19]. Third, synthetic single-stranded PASR RNA sequences transfected into human cells can affect the expression of the genes with which they associate [18]. Fourth, small RNAs are found associated with 5’ ends of RNAs generated both by transcriptional initiation and by cleavage [18]. In both cases, the 5’ ends of these small RNAs are modified by the addition of the cap, a modification known to protect RNAs against degradation [24], and this is inconsistent with their being mere degradation products on a path to complete removal from the cell.

Repetitive elements: parasites or building blocks of the genome?

Over the past few years, unbiased transcriptional surveys have revealed that a large fraction of the genome can be detected as stable transcripts [1,2,4]. However, these experiments, often microarray-based, typically avoided interrogating the repetitive element fraction of genomes as hybridization signals could not be assigned to a unique region. The advent of next-generation sequencing has made it possible to uniquely assign an RNA sequence to a particular repetitive element as long as there is some divergence from other copies of the element in the genome. Faulkner *et al.* [10] have now shown that a significant fraction of all CAGE tag clusters found in their study of human and mouse could be uniquely mapped to repetitive regions of the genome: 18.1% for mouse and 31.4% for human, represented by 44,264 and 275,185 clusters, respectively. Transcription within repetitive elements, specifically within retrotransposons, is apparently driven by their own promoters, which are surprisingly different from those previously characterized for these elements, and is highly tissue- and condition-specific. Faulkner *et al.* [10] find that overall, 35% of retrotransposon-associated TSSs show a restricted pattern of expression, compared to 17% of the other TSSs. Conversely, different tissues express different levels and types of repetitive elements, with human embryonic tissues having the highest levels of CAGE tags in these elements - 30% of all CAGE tags.

The big question raised by this study is whether the large contribution of repetitive elements, and retrotransposons in particular, to a cell's transcriptome translates into a major influence on its phenotype. In this respect, an important aspect of the study of Faulkner *et al.* [10] is the finding that retrotransposons might provide alternative or tissue-specific promoters for protein-coding genes. In fact, 15,518 (in mouse) and 117,165 (in human) of the putative novel TSSs within retrotransposons were identified as being associated with protein-coding transcripts, and the activity of 154 mouse and 579 human putative retrotransposon promoters was confirmed from existing expressed sequence tag (EST) data. Also, when Faulkner *et al.* [10] profiled 24 annotated protein-coding genes with suspected alternative retrotransposon promoters by rapid-amplification of cDNA ends (RACE), eight were indeed found to have sequences associating them with these promoters. Taken together, these results show that repetitive elements could in fact drive the production of a wide array of novel isoforms of protein-coding genes whose regulation and coding potential could be different from the isoforms annotated so far. It will be interesting to see how many of these putative protein-coding transcripts initiating within repetitive elements are actually translated.

This question could be phrased as part of a more general question: what is the complexity of polypeptides made in human cells, given the apparently high transcriptional complexity of RNAs made from a protein-coding locus? Analysis of available EST data has shown that, on average, a protein-coding locus can produce 5.7 different isoforms [25]. Furthermore, unbiased profiling of every protein-coding locus within the ENCODE regions has revealed that around 90% of them have either a novel internal exon or a novel TSS that is used in at least one tissue tested, and that most of the novel isoforms are tissue-specific [8]. It is not known, however, what fraction of these novel transcripts is actually translated and what fraction of such novel proteins would be functional.

Global regulation of the transcriptome

Precise knowledge of the TSSs used in a given biological condition is indispensable for understanding how that transcription is regulated. This is made abundantly clear by the study from the FANTOM Consortium and the Riken Omics Science Center [11], which modeled the transcriptional regulatory networks of a differentiating human cell. The authors used information on the genomic positions of the regulatory regions for each transcript and changes in transcript copy number during differentiation. Promoters were identified as regions flanking clusters of CAGE tags representing putative TSSs. For each promoter, known motifs for transcription factor binding sites were identified and this information was linked to changes in expression levels of the downstream transcript to infer the activity of the relevant transcription factors. From this, the authors identified 30 motifs whose activity explained most of the observed variation in gene expression; many of these motifs correspond to known regulators of the differentiation of macrophages - the particular cell type under study. The main conclusion reached is that a large number of different transcriptional regulators are required for differentiation, as opposed to the model in which the process is controlled by a small number of ‘master regulators’.

A similar strategy could be applied to identify transcription factors involved in regulation of other developmental or disease systems. The information on the expression levels of transcripts linked to individual TSSs is particularly important, as the study described above [11] shows that empirical mapping of TSSs can explain expression data better than existing annotated TSSs can.

A caveat that must, however, be applied to techniques that use an RNA cap to identify TSSs, is the recent discovery that CAGE tags could represent 5' ends of RNAs generated by cleavage and subsequent re-capping [18], and that cytoplasmic enzyme complexes can add caps to 5'-monophosphate RNA molecules generated by ribonuclease cleavage [26]. This means that mere knowledge of the position of a capped nucleotide is not sufficient to define a TSS. Additional information, such as the distribution of putative initiation sites within a promoter region [27], chromatin hallmarks associated with active promotors, the presence of RNA polymerase II initiation complexes and transcription factors [2,28] and appropriate sequence content [29], will be required to prove that a true initiation site has been identified and to re-evaluate the number of TSSs in human and other genomes.

Acknowledgements

I wish to thank Tom Gingeras, Erica Dumais and Jackie Dumais for suggestions and comments on this article.

References

1. Carninci P, Yasuda J, Hayashizaki Y: **Multifaceted mammalian transcriptome.** *Curr Opin Cell Biol* 2008, **20:**274-280.
2. ENCODE Project Consortium, Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P, Boyle PJ, Cao H, Carter NP, Clelland GK, Davis S, Day N, Dhami P, Dillon SC, Dorschner MO, Fiegler H, *et al.*: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.** *Nature* 2007, **447:**799-816.
3. Kapranov P, Willingham AT, Gingeras TR: **Genome-wide transcription and the implications for genomic organization.** *Nat Rev Genet* 2007, **8:**413-423.
4. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermüller J, Hofacker IL, Bell I, Cheung E, Drenkow J, Dumais E, Patel S, Helt G, Ganesh M, Ghosh S, Piccolboni A, Sementchenko V, Tammana H, Gingeras TR: **RNA maps reveal new RNA classes and a possible function for pervasive transcription.** *Science* 2007, **316:**1484-1488.
5. Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA, Kampa D, Piccolboni A, Sementchenko V, Cheng J, Williams AJ, Wheeler R, Wong B, Drenkow J, Yamanaka M, Patel S, Brubaker S, Tammana H, Helt G, Struhl K, Gingeras TR: **Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs.** *Cell* 2004, **116:**499-509.
6. Martone R, Euskirchen G, Bertone P, Hartman S, Royce TE, Luscombe NM, Rinn JL, Nelson FK, Miller P, Gerstein M, Weissman S, Snyder M: **Distribution of NF-kappaB-binding sites across human chromosome 22.** *Proc Natl Acad Sci USA* 2003, **100:**12247-12252.
7. Djebali S, Kapranov P, Foissac S, Lagarde J, Reymond A, Ucla C, Wyss C, Drenkow J, Dumais E, Murray RR, Lin C, Szeto D, Denoeud F, Calvo M, Frankish A, Harrow J, Makrythanasis P, Vidal M, Salehi-Ashtiani K, Antonarakis SE, Gingeras TR, Guigó R: **Efficient targeted transcript discovery via array-based normalization of RACE libraries.** *Nat Methods* 2008, **5:**629-635.
8. Denoeud F, Kapranov P, Ucla C, Frankish A, Castelo R, Drenkow J, Lagarde J, Alioto T, Manzano C, Chrast J, Dike S, Wyss C, Henrichsen CN, Holroyd N, Dickson MC, Taylor R, Hance Z, Foissac S, Myers RM, Rogers J, Hubbard T, Harrow J, Guigó R, Gingeras TR, Antonarakis SE, Reymond A: **Prominent use of distal 5' transcription start sites and discovery of a large number of additional exons in ENCODE regions.** *Genome Res* 2007, **17:**746-759.
9. Taft RJ, Glazov EA, Cloonan N, Simons C, Stephen S, Faulkner GJ, Lassmann T, Forrest AR, Grimmond SM, Schroder K, Irvine K, Arakawa T, Nakamura M, Kubosaki A, Hayashida K, Kawazu C, Murata M, Nishiyori H, Fukuda S, Kawai J, Daub CO, Hume DA, Suzuki H, Orlando V, Carninci P, Hayashizaki Y, Mattick JS: **Tiny RNAs associated with transcription start sites in animals.** *Nat Genet* 2009, [Epub ahead of print].

10. Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, Schroder K, Cloonan N, Steptoe AL, Lassmann T, Waki K, Hornig N, Arakawa T, Takahashi H, Kawai J, Forrest AR, Suzuki H, Hayashizaki Y, Hume DA, Orlando V, Grimmond SM, Carninci P: **The regulated retrotransposon transcriptome of mammalian cells.** *Nat Genet* 2009, [Epub ahead of print].
11. The FANTOM Consortium and the Riken Omics Science Center: **The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line.** *Nat Genet* 2009, [Epub ahead of print].
12. Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T, Fukuda S, Sasaki D, Podhajska A, Harbers M, Kawai J, Carninci P, Hayashizaki Y: **Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage.** *Proc Natl Acad Sci USA* 2003, **100:**15776-15781.
13. Hashimoto S, Suzuki Y, Kasai Y, Morohoshi K, Yamada T, Sese J, Morishita S, Sugano S, Matsushima K: **5'-end SAGE for the analysis of transcriptional start sites.** *Nat Biotechnol* 2004, **22:**1146-1149.
14. Gowda M, Li H, Alessi J, Chen F, Pratt R, Wang GL: **Robust analysis of 5'-transcript ends (5'-RATE): a novel technique for transcriptome analysis and genome annotation.** *Nucleic Acids Res* 2006, **34:**e126.
15. de Hoon M, Hayashizaki Y: **Deep cap analysis gene expression (CAGE): genome-wide identification of promoters, quantification of their expression, and network inference.** *Biotechniques* 2008, **44:**627-632.
16. Mardis ER: **The impact of next-generation sequencing technology on genetics.** *Trends Genet* 2008, **24:**133-141.
17. Tsuchihara K, Suzuki Y, Wakaguri H, Irie T, Tanimoto K, Hashimoto SI, Matsushima K, Mizushima-Sugano J, Yamashita R, Nakai K, Bentley D, Esumi H, Sugano S: **Massive transcriptional start site analysis of human genes in hypoxia cells.** *Nucleic Acids Res* 2009, [Epub ahead of print].
18. Affymetrix ENCODE Transcriptome Project; Cold Spring Harbor Laboratory ENCODE Transcriptome Project: **Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs.** *Nature* 2009, **457:**1028-1032
19. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermüller J, Hofacker IL, Bell I, Cheung E, Drenkow J, Dumais E, Patel S, Helt G, Ganesh M, Ghosh S, Piccolboni A, Sementchenko V, Tammana H, Gingeras TR: **RNA maps reveal new RNA classes and a possible function for pervasive transcription.** *Science* 2007, **316:**1484-1488.
20. Core LJ, Waterfall JJ, Lis JT: **Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters.** *Science* 2008, **322:**1845-1848.
21. Seila AC, Calabrese JM, Levine SS, Yeo GW, Rahl PB, Flynn RA, Young RA, Sharp PA: **Divergent transcription from active promoters.** *Science* 2008, **322:**1849-1851.
22. Davis CA, Ares M, Jr.: **Accumulation of unstable promoter-associated transcripts upon loss of the nuclear exosome subunit Rrp6p in *Saccharomyces cerevisiae*.** *Proc Natl Acad Sci USA* 2006, **103:**3262-3267.
23. Preker P, Nielsen J, Kammler S, Lykke-Andersen S, Christensen MS, Mapendano CK, Schierup MH, Jensen TH: **RNA exosome depletion reveals transcription upstream of active human promoters.** *Science* 2008, **322:**1851-1854.
24. Cougot N, van Dijk E, Babajko S, Seraphin B: **'Cap-tabolism'.** *Trends Biochem Sci* 2004, **29:**436-444.
25. Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, Lagarde J, Gilbert JG, Storey R, Swarbreck D, Rossier C, Ucla C, Hubbard T, Antonarakis SE, Guigo R: **GENCODE: producing a reference annotation for ENCODE.** *Genome Biol* 2006, **7 Suppl 1:**S4.1-9.
26. Otsuka Y, Kedersha NL, Schoenberg DR: **Identification of a cytoplasmic complex that adds a cap onto 5'-monophosphate RNA.** *Mol Cell Biol* 2009, **29:**2155-2167.
27. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engström PG, Frith MC, Forrest AR, Alkema WB, Tan SL, Plessy C, Kodzius R, Ravasi T, Kasukawa T, Fukuda S, Kanamori-Katayama M, Kitazume Y, Kawaji H, Kai C, Nakamura M, Konno H, Nakano K, Mottagui-Tabar S, Arner P, Chesi A, Gustincich S, Persichetti F, *et al.*: **Genome-wide analysis of mammalian promoter architecture and evolution.** *Nat Genet* 2006, **38:**626-635.
28. Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, Richmond TA, Wu Y, Green RD, Ren B: **A high-resolution map of active promoters in the human genome.** *Nature* 2005, **436:**876-880.
29. Megraw M, Pereira F, Jensen ST, Ohler U, Hatzigeorgiou AG: **A transcription factor affinity-based code for mammalian transcription initiation.** *Genome Res* 2009, **19:**644-656.