Genomics 94 (2009) 169–176

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Critical evaluation of the FANTOM3 non-coding RNA transcripts

Karl J.V. Nordström [1,2,3], Majd A.I. Mirza [1,2,3,4,8], Markus Sällman Almén [2,5], David E. Gloriam [2,6,9], Robert Fredriksson [2,6], Helgi B. Schiöth *,[2,7]

Department of Neuroscience, Uppsala University, BMC Box 593, 751 24 Uppsala, Sweden

ARTICLE INFO

Article history:
Received 6 November 2007
Accepted 26 May 2009
Available online 6 June 2009

Keywords:
RIKEN
FANTOM3
ncRNA
Non-coding RNA
snoRNA
EST

ABSTRACT

We studied the genomic positions of 38,129 putative ncRNAs from the RIKEN dataset in relation to protein-coding genes. We found that the dataset has 41% sense, 6% antisense, 24% intronic and 29% intergenic transcripts. Interestingly, 17,678 (47%) of the FANTOM3 transcripts were found to potentially be internally primed from longer transcripts. The highest fraction of these transcripts was found among the intronic transcripts and as many as 77% or 6929 intronic transcripts were both internally primed and unspliced. We defined a filtered subset of 8535 transcripts that did not overlap with protein-coding genes, did not contain ORFs longer than 100 residues and were not internally primed. This dataset contains 53% of the FANTOM3 transcripts associated to known ncRNA in RNAdb and expands previous similar efforts with 6523 novel transcripts. This bioinformatic filtering of the FANTOM3 non-coding dataset has generated a lead dataset of transcripts without signs of being artefacts, providing a suitable dataset for investigation with hybridization-based techniques.

Background

The high number of functional non-coding RNA (ncRNA) genes discovered lately has changed our view of how we look at the cell as a functional unit. RNA was, with few exceptions (tRNA and rRNA), thought to have no function on its own, but rather only work as an intermediate messenger (mRNA) in the generation of proteins from DNA. Recently, many biologically functional molecules of RNA and several new subclasses have been added to the RNA family [1]. Moreover, several new specific functions have been attributed to these new actors. These functions include genomic imprinting [2], regulation of mRNA stability [3], transcription co-activation, repression, post-transcriptional regulation [4–6] and regulation of insulin secretion [7]. ncRNAs are involved in neural stem cell differentiation and development [8,9] and X-chromosome activation and inactivation [10,11]. Moreover, ncRNAs function as repressors of NFAT protein [12], as a response to heat shock and cell stress [13,14], are responsible for the proper formation of the photoreceptors in the retina of mice [15] and export toxic compounds from eukaryotic cells [16]. The list is expanding rapidly as many additional, recently discovered ncRNAs have been cloned and characterized [17].

The view on the transcriptional level in eukaryotes has changed. An enduring thought was that only small portions of the genomes of higher multi-cellular organisms are transcribed, whereas the genomes of lower microorganisms are nearly fully transcribed. However, recent reports demonstrate that also the dominating portions of the genomes of higher multi-cellular organisms are transcribed and the majority of these newly discovered genes are not protein-coding [18]. Still, with more than a decade of qualified guesses [19–24], including serious attempts to estimate the total number of protein-coding sequences in human and mouse [20,23,25], there are large uncertainties about the total number of transcribed elements in the mammalian genomes and the exact proportions of coding vs. non-coding genes have been extremely hard to determine [26]. Although many prediction algorithms for non-coding RNAs have been published [27–29], it is difficult to know whether the resulting datasets of non-coding transcripts really are of biological relevance i.e. have a physiological function. Still, the number of putative non-coding RNA genes in sequence databases has grown enormously in the last few years [30–33]. The new sequences

* Corresponding author.
E-mail addresses: karl.nordstrom@neuro.uu.se (K.J.V. Nordström), majd.mirza@medsci.uu.se (M.A.I. Mirza), markus.sallman-almen@neuro.uu.se (M.S. Almén), david.gloriam@neuro.uu.se (D.E. Gloriam), robert.fredriksson@neuro.uu.se (R. Fredriksson), helgi.schioth@neuro.uu.se (H.B. Schiöth).

[1] Equal contribution.

[2] All authors read and approved the final manuscript.

[3] KJVN and MAIM participated in the design of the study, carried out the analysis, filtration and drafted the manuscript.

[4] MAIM also did the RT-PCR.

[5] MSA contributed with the conservation study and helped to draft the manuscript.

[6] DEG and RF participated in the design of the study and helped to draft the manuscript.

[7] HBS conceived the study, participated in its design and coordination and helped to draft the manuscript.

[8] Current adress: Department of Medical Sciences, Uppsala University Hospital, 751 85 Uppsala, Sweden.

[9] Current adress: Department of Medicinal Chemistry, Copenhagen University, Universitetsparken 2, 2100 Copenhagen, Denmark.

0888-7543/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2009.05.012

have been found and characterized using different methods including large scale cDNA sequencing [32,34,35], large scale gene expression profiling, molecular cloning and tiling arrays [36–39].

One of the most important efforts, with the objective to identify all transcribed mRNAs in the mouse genome is the RIKEN cDNA project [34,35]. Their latest release, FANTOM3, comprises 102,801 cDNAs of which 38,129 have been classified as potential non-coding RNA genes [32]. A dedicated database, FANTOMDB, was created to store sequence information about the RIKEN full-length cDNA clones, annotation information and additional description [40]. Another database, RNAdb, produced by Pang et al., store information about small functional RNAs (microRNA, snoRNA) as well as putative ncRNAs [30,33]. It is still unknown if all of the above transcripts are functional elements and also to which extent the content and quality of these large datasets have been influenced by limitations in the experimental procedures. The RIKEN cDNA libraries are constructed with oligo-dT priming, advanced techniques for cap-trapping and aggressive normalization methods. By amplifying rare transcripts and defining the polyA-tail as the 3′ end, it is possible to extract partially degraded introns or pre-mRNAs if they contain a longer stretch of adenines. Therefore it has been argued that many of the cDNAs in the RIKEN libraries, assumed to be non-coding RNAs in fact are non-functional cDNAs with a low level of conservation [41]. On the other hand, it has been shown that some ncRNAs are highly conserved even between such distant species as chicken and pufferfish [32]. It has also been argued that the FANTOM3 dataset contains a number of functional transcripts with a regulated expression [42–44].

Expressed sequence tags (ESTs) comprise an important source of information in the identification of FANTOM3 cDNA clones [32]. These have been collected since 1991 [45] and currently, the largest EST-database, dbEST hosted by NCBI, contains more than 59 million entries whereof the two most represented species, human and mouse, contribute with almost 8 and 5 million, respectively. ESTs are have also proven a valuable source of information for identification of new genes [32,46].

Previous studies have used both bioinformatic tools and elaborative methods in order to study the many putative non-coding transcripts that are now known [28,42,44]. Numata et al. studied the subset of putative ncRNAs in FANTOM2 with respect to protein homology and EST support [47]. The number of putative ncRNAs has more than doubled between FANTOM2 and FANTOM3. Here, we updated and extended their analysis by conducting a rigorous investigation of the genomic locations of the transcripts.

In order to gain a better insight into the issue regarding the functionality of the many putative non-coding transcripts, we performed a careful and detailed analysis of the non-coding sequences of the FANTOM3 dataset. We divided the FANTOM3 non-coding transcripts into several subgroups based on their genomic positions relative to protein-coding genes. Specifically, we studied the position of putative ncRNAs giving the proportions of sense, antisense, intronic and intergenic transcripts in order to delineate which transcripts were associated with proteins and thus more likely to be artefacts. Further, we identified orthologs, if present, in rat and human for each transcript. We used this information to filter a dataset of almost 40,000 non-coding transcripts generating a subset of transcripts without obvious signs of being artefacts and at the same time is suitable to further investigation with hybridization-based techniques.

Results

Genomic mapping of the dataset

38,010 (99.7%) of the 38,129 FANTOM3 non-coding transcripts were successfully mapped to the mouse genome assembly with BLAT. The genomic mapping revealed that as much as, 30,572 (80.2%), of the transcripts were not spliced. The spliced transcripts had 2.9 exons in average. Clustering of overlapping transcripts resulted in 30,705 independent clusters with on average 1.2 transcripts in each.

Genomic positions of FANTOM3 non-coding transcripts in relation to protein-coding genes

The integrity of the FANTOM3 non-coding transcript dataset was further assessed by determining the genomic structures. To this end, we matched the positions, lengths and orientations of the FANTOM3 non-coding transcripts on the genome with those of exons, introns and UTRs of known protein-coding genes. Representative examples are displayed in Fig. 1 and the overall outcome, with the FANTOM3 non-coding transcripts separated into different spatial categories, is presented in Table 1. Interestingly, we found that 10,507 (27.6%) transcripts had an overlap with protein-coding genes and of these 3228 (8.5%) FANTOM3 transcripts were complete subsequences (had no unique sequence). Furthermore, EST sequences and EST clusters provided alternate and extended versions of the many protein-coding genes. This information indicated an additional 2701 (7.1%) FANTOM3 non-coding transcripts, which did not overlap protein entries directly, to be part of a splice variant or a longer version of a UTR. From both manual and computational analyses, it was clear that many additional transcripts were positioned just outside the borders of annotated protein-coding genes and were therefore likely to be part of UTRs although hitherto no protein or EST sequence data has been generated that shows this (no overlap). We therefore considered all FANTOM3 non-coding transcripts that did not overlap or was linked by ESTs or EST clusters to a protein-coding gene, within 5 kbp of the boundaries of protein-coding genes on the same strand to be putative parts of UTRs (2508 transcripts). The remaining dataset consists of intronic (9022 or 23.7%), antisense (2263 or 5.9%), and intergenic (11,009 or 28.9%) transcripts. Of the intronic transcripts, 185 transcripts were also located antisense to another protein-coding gene.

Assessments of protein-coding gene characteristics

We searched the FANTOM3 transcripts for characteristic features of protein-coding genes. By definition, ncRNAs do not code for proteins and should therefore not contain (long) open reading frames (ORFs). We investigated the FANTOM3 non-coding dataset and found that 5698 (14.9%) of all transcripts contain an ORF that was at least 100

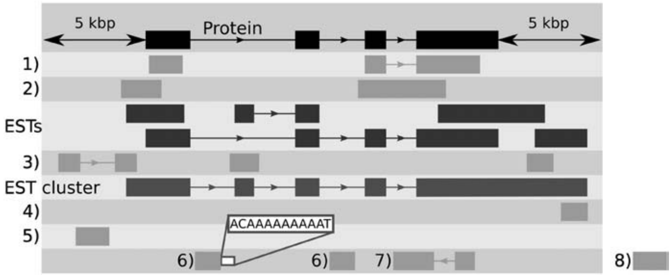

Fig. 1. Genomic structures of FANTOM3 non-coding transcripts in relation to protein-coding genes. A schematic picture of how a transcript can be located relative to a protein-coding gene. The top-most line depicts a protein with a 5 kbp padding upstream and downstream to include a putative UTR. The row denoted EST cluster shows the resulting cluster when the ESTs on the row denoted ESTs are clustered. 1) Two transcripts located completely within the exons of the protein. 2) These two transcripts are partially overlapping the protein. The rightmost transcript fulfils our pre-mRNA criteria. 3) This line contains three transcripts that overlap ESTs also overlapped by the protein. 4) In this, one transcript is linked to the protein by the EST cluster on the row above. 5) The transcript on this row is associated to the protein as a putative UTR. 6) On the last row, there are two intronic transcripts, of which, the leftmost has an adenine-rich region located downstream and might be internally primed. 7) A transcript on the other strand overlapping the last exon of the protein-coding gene. 8) Transcripts not classified to any of the above categories were considered intergenic.