Nucleic Acids Research, 2007, Vol. 35, Database issue ***D165–D168***
doi:10.1093/nar/gkl870

PolyA_DB 2: mRNA polyadenylation sites in vertebrate genes

Ju Youn Lee, Ijen Yeh, Ji Yeon Park and Bin Tian*

Department of Biochemistry and Molecular Biology, New Jersey Medical School, University of Medicine
and Dentistry of New Jersey, Newark, NJ 07101-1709, USA

Received September 15, 2006; Accepted October 6, 2006

ABSTRACT

Polyadenylation of nascent transcripts is one of the key mRNA processing events in eukaryotic cells. A large number of human and mouse genes have alternative polyadenylation sites, or poly(A) sites, leading to mRNA variants with different protein products and/or 3′-untranslated regions (3′-UTRs). PolyA_DB 2 contains poly(A) sites identified for genes in several vertebrate species, including human, mouse, rat, chicken and zebrafish, using alignments between cDNA/ESTs and genome sequences. Several new features have been added to the database since its last release, including syntenic genome regions for human poly(A) sites in seven other vertebrates and *cis*-element information adjacent to poly(A) sites. Trace sequences are used to provide additional evidence for poly(A/T) tails in cDNA/ESTs. The updated database is intended to broaden poly(A) site coverage in vertebrate genomes, and provide means to assess the authenticity of poly(A) sites identified by bioinformatics. The URL for this database is http://polya.umdnj.edu/PolyA_DB2.

INTRODUCTION

Polyadenylation of nascent transcripts is an essential step for all mRNAs in eukaryotic cells, with the exception of some histone transcripts (1,2). The process is coupled with other steps of mRNA processing, such as termination of transcription and splicing (3,4), and involves two reactions (5,6): an endonucleolytic cleavage of a nascent mRNA, followed by polymerization of an adenosine tail at the 3′ end of the mRNA to a length specific to the species, e.g. 200–250 nt in mammals and 70–90 nt in yeasts. The poly(A) tail is critical for many aspects of mRNA metabolism, including mRNA stability, translation and transport (7,8). A number of *cis*-regulatory elements, or *cis*-elements, have been found or suggested to play roles in polyadenylation [(9) and references therein]. The most prominent elements in mammals are the Polyadenylation Signal (PAS) located within ∼40 nt upstream of the cleavage site, including AAUAAA, AUUAAA, and their variants, and the U/GU-rich elements located within ∼40 nt downstream of the cleavage site. *Cis*-element information can be used to computationally predict poly(A) sites with high sensitivity and selectivity (10). Recently, polyadenylation has been implicated in the degradation of some nuclear RNAs in eukaryotic cells by the exosome (11). The polyadenylation reaction in this process appears to involve a distinct set of factors than those responsible for polyadenylation of nascent transcripts, and the resulting poly(A) tail is usually short and contains variable nucleotides (12).

Over half of all human genes have alternative poly(A) sites, with locations in internal exons, introns and 3′-most exons (13,14), leading to transcript variants with different open reading frames and/or variable 3′-untranslated regions (3′-UTRs). Thus, alternative polyadenylation significantly contributes to the overall complexity of mRNA pool in the cell. A growing number of alternative polyadenylation events have been shown to regulate gene function in a tissue-specific manner (15,16). Bioinformatic techniques using cDNA sequences and Expressed Sequenced Tags (ESTs) offer a systematic approach to identify poly(A) sites in genomes (13,14,17–20). PolyA_DB 2 is a database designed to catalog poly(A) sites in all genes of major vertebrate species. Poly(A) sites are identified by aligning cDNA/ESTs with genome sequences. Several types of information are available in the database and can be retrieved and presented in Views, including intron/exon structure, poly(A) site location, supporting cDNA/ESTs, ortholog groups across human, mouse, rat and chicken, PAS types and positions, locations of various *cis*-elements, cDNA library information, and syntenic sequences across eight vertebrates. In addition, Trace sequences are used, when available, to further support the poly(A/T) tails in cDNA/ESTs. These data not only provide comprehensive information for poly(A) sites in several vertebrate species, but also allow researchers to assess the authenticity of

*To whom correspondence should be addressed. Tel: +1 973 972 3615; Fax: +1 973 972 5594; Email: btian@umdnj.edu

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors poly(A) sites identified by bioinformatics. This database can be of great value to those interested in studying the mechanism of polyadenylation as well as gene regulation involving alternative polyadenylation.

METHODS

Identification of poly(A) sites in genomes using cDNA/ESTs and Trace sequences

We retrieved all cDNA/EST sequences listed in human, mouse, rat, chicken and zebrafish UniGene databases from NCBI (July and August 2005 versions), and aligned them with genome sequences downloaded from the UCSC Genome Bioinformatics Site (http://genome.ucsc.edu, hg17 for human, mm5 for mouse, rn3 for rat, galGal2 for chicken and danRer1 for zebrafish) using BLAT (21). Poly(A) sites were identified by parsing dangling ends of the alignments using the method described in (13). All internal priming candidates were discarded. Human, mouse, rat and zebrafish Trace sequences were downloaded from NCBI Trace Archive and were used to extend terminal poly(A/T) sequences in cDNA/ESTs as described in (22).

Grouping poly(A) sites according to locations and genes

We grouped together poly(A) sites belonging to the same gene using NCBI UniGene database. To eliminate antisense transcripts and other erroneous transcripts, we cleaned up UniGene Bins (or Clusters) as shown in Supplementry Figure 1. The cleaned UniGene Bins are called CLUBs (CLeaned UniGene Bin). This step was carried out first by selecting a representative sequence called initiator for the CLUB, followed by iteratively including cDNA/ESTs that have the same transcriptional orientation as the initiator and have sequence overlap with cDNA/ESTs already in the CLUB. Initiators were selected based on the order RefSeqs > other cDNAs > ESTs. Sequences included in a CLUB are called CLUB members. One UniGene Bin may have more than one CLUB. To maximize the number of supporting cDNA/EST sequences for a poly(A) site, the 3′ ends of sequences without poly(A/T) tails were compared with identified poly(A) sites. An cDNA/EST is considered to be supporting a poly(A) site if its 3′ end is near the poly(A) site within 24 nt. Transcripts with unknown transcriptional orientation are assigned as associated CLUB members if one of their sequence ends is near a poly(A) site within 24 nt, and the inferred transcriptional orientation based on the poly(A) site does not conflict with that of the CLUB. They were also included as supporting cDNA/ESTs. Poly(A) sites that are located within 24 nt from one another, due to heterogeneous cleavage, were iteratively clustered together in the 5′ to 3′ direction. The position of the middle cleavage site in a cluster is used to represent the cluster. Thus, in PolyA_DB 2, poly(A) site ID is composed of three parts, i.e. UniGene ID, CLUB number, and site number. For example, Hs.44402.1.46 is based on UniGene ID Hs.44402, CLUB number 1, and site number 46. Gene information was retrieved from NCBI Gene databases (August 2005 versions) and assigned to CLUBs based on the relationship between Gene IDs and RefSeq sequences in the CLUBs.

Annotation of *cis*-elements

We used position-specific scoring matrices (PSSMs) of previously identified 15 *cis*-elements to search poly(A) regions (9). For each matrix we derived all possible positive scores for matching a sequence with the same size. The 25th, 50th and 75th percentiles were used to benchmark other sequence matches. For all sequences surrounding human and mouse poly(A) sites, we compared their matching scores with the benchmarks. A match was considered ‘very strong’ if its score was above the 75th percentile; ‘strong’ for the 50th–75th percentile; ‘weak’ for the 25th–50th percentile; ‘very weak’ for below the 25th percentile; and ‘no match’ for negative scores.

Finding syntenic regions for human poly(A) sites

We used the eight-way genome alignments from the UCSC Genome Bioinformatics Site to obtain syntenic regions for human poly(A) sites. The eight-way genome alignment files contain genomes of *Homo sapiens* (human), *Pan troglodytes* (chimpanzee), *Canis familiaris* (dog), *Mus musculus* (mouse), *Rattus norvegicus* (rat), *Gallus gallus* (chicken), *Danio rerio* (zebrafish) and *Takifugu rubripes* (puffer fish). We first parsed out the alignment blocks overlapping the −300 to +300 nt region surrounding human poly(A) sites and identified corresponding positions in other genomes if they had sequences aligned with the region. We then retrieved genomic sequences from all aligning species, including human, and re-aligned the sequences with CLUSTALW (23). We also annotated all cDNA/ESTs from the aligning species whose sequence ends are located within 24 nt from human poly(A) sites in the alignments. This information can support conservation of poly(A) sites.

Database and website

Data in the database are stored in a relational database, implemented with MySQL. PHP and Perl are used for the web interface. Bioperl modules are used for graphical representation of sequences (24). Queries are based on Gene IDs, UniGene IDs, CLUB IDs and Site IDs. Large batch downloads are available upon request.

RESULTS AND DISCUSSION

As of September 2006, PolyA_DB 2 contains 54 686 human, 30 235 mouse, 26 602 rat, 6287 chicken, and 5830 zebrafish poly(A) sites (Table 1). Various types of information regarding poly(A) sites and their corresponding genes are available and are presented in Views. We have updated several Views reported in the last release (25), such as Gene View, cDNA/EST evidence View, Ortholog View, PAS View and Library View (see the online help file for details), and have added several new Views.

Site View

Site View contains information about individual poly(A) sites. Each poly(A) site has an ID as described above. Its corresponding Gene ID, Chromosome number, and position on the chromosome are indicated. Since each site may have several cleavage sites, the left-most and right-most cleavage

Table 1. Poly(A) sites and Genes in PolyA_DB 2

	Hs	*Mm*	*Rn*	*Gga*	*Dr*
No. of cDNA/ESTs used	3 106 770	1 966 056	388 581	264 625	176 740
No. of poly(A/T)-tailed cDNA/ESTs	530 199	189 619	144 364	16 761	21 650
No. of NCBI Genes	22 001	23 005	16 560	6742	6554
No. of UniGenes	39 105	28 622	32 945	18 194	12 184
No. of CLUBs[a]	39 181	28 678	32 982	18 194	12 190
No. of Poly(A) sites[b]	54 686	30 235	26 602	6287	5830

Hs, Homo sapiens; Mm, Mus musculus; Rn, Rattus norvegicus; Gga, Gallus gallus; Dr, Danio rerio.

[a] CLUBs, Cleaned UniGene Bins.

[b] Some Genes, UniGenes, or CLUBs do not have poly(A) sites identified by cDNA/ESTs.

locations, as well as number of cleavage sites are provided. The maximum length of poly(A/T) tail based on all supporting cDNA/ESTs, and additional tail length from Trace sequences are listed, which can help determine the authenticity of a poly(A) site (see below). In addition, supporting cDNA/ESTs and the genomic sequence −125 to +125 nt flanking each site are provided.

cis-Element View

The −125 to +125 nt region of a poly(A) site is searched for 15 elements identified by a bioinformatic method described in (9). Matches with *cis*-elements are divided into five categories as described in Methods.

Synteny View

For each human poly(A) site, a multiple genome alignment of eight vertebrate species is presented (see Methods for details).

PolyA_DB 2 is designed to provide higher poly(A) coverage in more species than the previous release (25). In addition, new features are included to help researchers assess the authenticity of poly(A) sites identified by bioinformatics. Major sources for false identification of poly(A) sites include (i) internal priming of poly(A)-rich sequences within mRNAs (26) and (ii) polyadenylation of mRNAs marked for degradation by the exosome (11,12). We have taken rigorous measures to eliminate internally primed cDNA/ESTs (22), and thus its effect should be minimal, if any. Polyadenylated mRNAs marked for degradation can potentially lead to false identification of poly(A) sites. Their poly(A) tails, however, appear to be shorter than those added by the polyadenylation process for nascent transcripts. Given that most oligo(dT) primers are ∼15–20 nt, long poly(A/T) tails in cDNA/ESTs, e.g. >30 nt, can be considered to be derived from poly(A) tails made by the polyadenylation process for nascent transcripts. In this regard, additional poly(A/T) sequences from the Trace database are highly valuable for the selectivity of poly(A) site identification. These sequences are usually removed from cDNA/ESTs due to low quality of sequence or low complexity of sequence. Other types of information can also provide circumstantial evidence as to the authenticity of a poly(A) site: (i) large number of cDNA/ESTs; (ii) several cleavage sites for a poly(A) site; (iii) presence of canonical PAS, i.e. AAUAAA or AUUAAA; (iv) presence of other *cis*-elements for polyadenylation and (v) high sequence conservation in syntenic regions in other species. All these are available in PolyA_DB 2.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Nick Proudfoot for discussion on the polyadenylation involved in RNA degradation. This work was supported by The Foundation of the University of Medicine and Dentistry of New Jersey. Funding to pay the Open Access publication charges for this article was provided by UMDNJ.

Conflict of interest statement. None declared.

REFERENCES

1. Edmonds,M. (2002) A history of poly A sequences: from formation to factors to function. *Prog. Nucleic Acid Res. Mol. Biol.*, **71**, 285–389.
2. Marzluff,W.F. (2005) Metazoan replication-dependent histone mRNAs: a distinct set of RNA polymerase II transcripts. *Curr. Opin. Cell Biol.*, **17**, 274–280.
3. Minvielle-Sebastia,L. and Keller,W. (1999) mRNA polyadenylation and its coupling to other RNA processing reactions and to transcription. *Curr. Opin. Cell Biol.*, **11**, 352–357.
4. Proudfoot,N. (2004) New perspectives on connecting messenger RNA 3′ end formation to transcription. *Curr. Opin. Cell Biol.*, **16**, 272–278.
5. Colgan,D.F. and Manley,J.L. (1997) Mechanism and regulation of mRNA polyadenylation. *Genes Dev.*, **11**, 2755–2766.
6. Zhao,J., Hyman,L. and Moore,C. (1999) Formation of mRNA 3′ ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. *Microbiol. Mol. Biol. Rev.*, **63**, 405–445.
7. Mangus,D.A., Evans,M.C. and Jacobson,A. (2003) Poly(A)-binding proteins: multifunctional scaffolds for the post-transcriptional control of gene expression. *Genome Biol.*, **4**, 223.
8. Wickens,M., Anderson,P. and Jackson,R.J. (1997) Life and death in the cytoplasm: messages from the 3′ end. *Curr. Opin. Genet. Dev.*, **7**, 220–232.
9. Hu,J., Lutz,C.S., Wilusz,J. and Tian,B. (2005) Bioinformatic identification of candidate *cis*-regulatory elements involved in human mRNA polyadenylation. *RNA*, **11**, 1485–1493.
10. Cheng,Y., Miura,R.M. and Tian,B. (2006) Prediction of mRNA polyadenylation sites by support vector machine. *Bioinformatics*, **22**, 2320–2325.
11. Houseley,J., LaCava,J. and Tollervey,D. (2006) RNA-quality control by the exosome. *Nature Rev. Mol. Cell Biol.*, **7**, 529–539.
12. West,S., Gromak,N., Norbury,C.J. and Proudfoot,N.J. (2006) Adenylation and exosome-mediated degradation of cotranscriptionally cleaved pre-messenger RNA in human cells. *Mol. Cell*, **21**, 437–443.
13. Tian,B., Hu,J., Zhang,H. and Lutz,C.S. (2005) A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res.*, **33**, 201–212.
14. Yan,J. and Marr,T.G. (2005) Computational analysis of 3′-ends of ESTs shows four classes of alternative polyadenylation in human, mouse, and rat. *Genome Res.*, **15**, 369–375.
15. Zhang,H., Lee,J.Y. and Tian,B. (2005) Biased alternative polyadenylation in human tissues. *Genome Biol.*, **6**, R100.

16. Edwalds-Gilbert,G., Veraldi,K.L. and Milcarek,C. (1997) Alternative poly(A) site selection in complex transcription units: means to an end? *Nucleic Acids Res.*, **25**, 2547–2561.
17. Gautheret,D., Poirot,O., Lopez,F., Audic,S. and Claverie,J.M. (1998) Alternate polyadenylation in human mRNAs: a large-scale analysis by EST clustering. *Genome Res.*, **8**, 524–530.
18. Iseli,C., Stevenson,B.J., de Souza,S.J., Samaia,H.B., Camargo,A.A., Buetow,K.H., Strausberg,R.L., Simpson,A.J., Bucher,P. and Jongeneel,C.V. (2002) Long-range heterogeneity at the 3′ ends of human mRNAs. *Genome Res.*, **12**, 1068–1074.
19. Brockman,J.M., Singh,P., Liu,D., Quinlan,S., Salisbury,J. and Graber,J.H. (2005) PACdb: PolyA cleavage site and 3′-UTR Database. *Bioinformatics*, **21**, 3691–3693.
20. Loke,J.C., Stahlberg,E.A., Strenski,D.G., Haas,B.J., Wood,P.C. and Li,Q.Q. (2005) Compilation of mRNA polyadenylation signals in Arabidopsis revealed a new signal element and potential secondary structures. *Plant Physiol.*, **138**, 1457–1468.
21. Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
22. Lee,J.Y., Park,J.Y. and Tian,B. (2006) Identification of mRNA polyadenylation sites in genomes using cDNA sequences, expressed sequence tags, and Trace. *Methods Mol. Biol.*, in press.
23. Chenna,R., Sugawara,H., Koike,T., Lopez,R., Gibson,T.J., Higgins,D.G. and Thompson,J.D. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.*, **31**, 3497–3500.
24. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G., Korf,I., Lapp,H. *et al.* (2002) The Bioperl toolkit: perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.
25. Zhang,H., Hu,J., Recce,M. and Tian,B. (2005) PolyA_DB: a database for mammalian mRNA polyadenylation. *Nucleic Acids Res.*, **33**, D116–D120.
26. Nam,D.K., Lee,S., Zhou,G., Cao,X., Wang,C., Clark,T., Chen,J., Rowley,J.D. and Wang,S.M. (2002) Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription. *Proc. Natl Acad. Sci. USA*, **99**, 6152–6156.