

# PACdb: PolyA Cleavage Site and 3'-UTR Database

J. Michael Brockman<sup>1,2,4</sup>, Priyam Singh<sup>1,2,4</sup>, Donglin Liu<sup>1</sup>, Sean Quinlan<sup>1,2</sup>,  
Jesse Salisbury<sup>1,3</sup>, Joel H. Graber<sup>1</sup>

**Keywords:** polyadenylation, 3' untranslated region, 3'-UTR, mRNA processing

## 1 Introduction.

The addition of a polyadenylate tail to the 3' end of most eukaryotic mRNA molecules has been shown to affect mRNA localization, stability, and translational efficiency [1]. Alternative 3'-processing has been seen under varying cellular conditions and/or tissue types [2]. Use of alternate 3'-processing sites can result in different post-transcriptional gene regulation or even post-transcriptional gene silencing in the case of cleavage within the coding sequence [3].

The "PolyA Cleavage Site and 3'-UTR Database" (PACdb) is a web-accessible database that catalogs putative 3'-processing sites and 3'-UTR sequences for multiple organisms. While other resources exist to identify 3' processing sites and/or 3'-UTR sequence [4,5], but they have different focus, scope, or methodologies than PACdb. We have identified sites primarily via EST-genome alignments, enabling delineation of both the specificity and heterogeneity of 3'-processing events within or across genomes. The database currently contains putative processing sites for a diverse set of organisms including human, mouse, rat, dog, chicken, zebrafish, fugu, fruitfly (*D. melanogaster*), mosquito, nematode (*C. elegans*), *A. thaliana*, rice (*O. japonica*), and baker's yeast.

## 2 Implementation and Availability.

We use an automated process to determine putative 3'-processing sites. This process has four main parts: EST filtering, EST-genome alignment, mapping putative sites to genes, and putative site confidence assessment. We intentionally retain potentially false processing sites in PACdb to allow for comparison with gene expression measurements such as SAGE [6]. When the automated process to determine putative 3'-processing sites is complete, the data is fed into three separate but connected databases which store information on ESTs, EST-genome alignment, and putative 3'-processing sites.

Data can be accessed at our website (<http://harlequin.jax.org/pacdb>) using online forms or programmatically using a simple CGI-based API. When viewing the 3'-processing sites for a specific gene, PACdb displays a graphical representation of the gene structure, the aligned ESTs, and the putative 3'-processing sites (Figure 1). A-rich regions (denoted 'A') or potential restriction enzyme sites (denoted 'R') could imply false processing sites. Optional coloring of the aligned ESTs indicates tissue groups. The results from any web-based form can be output in XML, tab delimited text, HTML Table, or FASTA when appropriate. Other web-based databases can easily link to PACdb using the CGI-based API.

---

<sup>1</sup> Jackson Laboratory, 600 Main St, Bar Harbor, ME 04609, USA.

<sup>2</sup> Bioinformatics Program, Boston University, Boston, MA 02215, USA.

<sup>3</sup> Functional Genomics Program, University of Maine, Orono, ME 04469, USA.

<sup>4</sup> These authors contributed equally to this work

<sup>5</sup> To whom correspondence should be addressed. E-mail: [jhgraber@jax.org](mailto:jhgraber@jax.org)

### 3 Figures and tables.

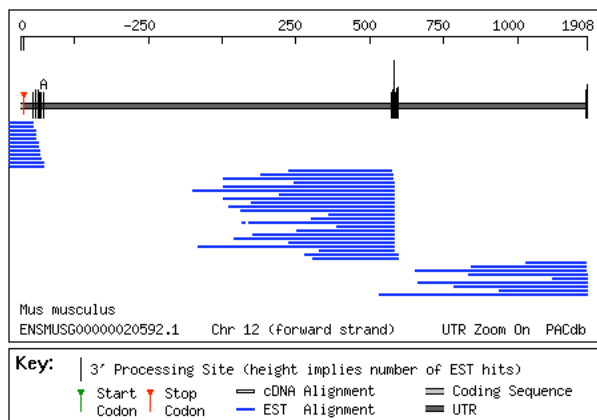


Figure 1: Mouse cleavage and polyadenylation specific factor 5 gene.

Organism	ESTs	Genes in PACdb	Genes with one site	Genes with 2+ sites	Putative 3' Processing Sites	Singleton ESTs	Unmapped Sites
Chicken	44126	9417	3056	6361	33348	28806	4084
Dog	25337	2162	629	1533	7599	5176	2585
Rat	219848	16347	1760	14587	92054	59472	18055
Zebrafish	12812	1270	395	875	5491	4246	1470
<i>D. melanogaster</i>	38800	na	na	na	25624	21155	na
Fugu	2974	na	na	na	2151	1858	na
Mouse	296441	15458	1910	13548	100690	66102	na
<i>Arabidopsis</i>	138662	14735	3465	11270	67543	47983	na

Table 1: Current data in PACdb. In progress: human, *C. elegans*, mosquito, rice, yeast. Unavailable data: "na".

### 4 References.

- [2] Edwalds-Gilbert, G., Veraldi, K.L., and Milcarek, C. 1997. Alter-native poly(A) site selection in complex transcription units: means to an end? *Nucleic Acids Res.*, 25:2547-2561.
- [4] Mignone, F, Grillo, G, Licciulli, F, Iacono, M, Liuni, S, Kersey, PJ, Duarte, J, Saccone, C, Pesole, G. 2005. UTRdb and UTRsite: a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. *Nucleic Acids Res.*, 33:141-146.
- [3] van Hoof, A., Frischmeyer, P.A., Dietz, H.C., Parker, R. 2002. Exosome-mediated recognition and degradation of mRNAs lacking a termination codon. *Science*, 295:2262-2264.
- [6] Velculescu, V.E., Zhang, L., Vogelstein, B., Kinzler, K.W. 1995. Serial analysis of gene expression. *Science*, 270:484-487.
- [5] Zhang, H, Hu, J, Recce, M, Tian, B. 2005. PolyA\_DB: a database for mammalian mRNA polyadenylation. *Nucleic Acids Res.*, 33:116-120.
- [1] Zhao, J, Hyman, L., and Moore, C. 1999. Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelation-ships with other steps in mRNA synthesis. *Microbiol. Mol. Biol. Rev.*, 63:405-445.