

SUPPLEMENT

Data in this supplement is as of June 24, 2005. Updated versions of these tables/figures are available online at: <http://harlequin.jax.org/pacdb/supplement.php>

PACdb: PolyA Cleavage Site and 3'-UTR Database

Table of Contents:

Supplement Introduction

Methods

Putative 3'-Processing Site Determination

Database Implementation

Data Analysis and Characterization

EST-Site Properties and Confidence Level Assignment

Putative 3' UTR Length and Intergenic Distance

Putative 3'-Processing Sites per Gene

Canonical Word Usage in the PolyA Signal (PAS)

Supplement Conclusion

Tables and Figures

Supplement Introduction

The data generated and stored in the PolyA Cleavage Site and 3'-UTR Database (PACdb), as well as future updates, are meant for use by the biological research community to better characterize polyadenylation and post-transcriptional gene regulation, and also to improve gene transcript prediction. In this online supplemental, we present details of our methods/implementation and also some analysis of the data across organisms in PACdb. We were unable to present these details in the Bioinformatics paper due to length restrictions on the article format. The analysis here demonstrates both the quality of the data in the database and also biological trends across organisms.

Methods

Putative 3'-Processing Site Determination

We have developed an automated process to determine putative 3'-processing sites consistently across organisms. This process has four main parts: EST pre-processing, EST-genome alignment, gene mapping, and putative site characterization. See Figure 1 for a flowchart of methods described below. The processing of cDNA sequences is done using the same system, with slight modification.

EST pre-processing includes screening EST sequences for: (1) mitochondrial, bacterial, or viral sequence contaminants, (2) non-messenger RNA (tRNA, etc), (3) low quality sequence, (4) organism-specific repetitive elements, and (5) vector or library-specific linker sequences. Cases 1 and 2 are removed from further analysis, while cases 3-5 are masked.

Initial alignments are made with masked EST or masked genomic sequence using BLAT (Kent, 2002). Alignments that do not completely cover the EST are realigned using unmasked sequences. All possible genomic alignments for each EST are scored based on coverage of the EST and quality of the alignment. We retain all alignments that score within 1% of the best

alignment as possible alternatives for further analysis. When assembled genomic sequence is unavailable, we extract the best alignments of the 5' and 3' most ends of an EST since the best alignment could span more than one contig sequence.

The selected EST alignments are then assessed for likely orientation using the presence of a trailing polyA or leading polyT tail on the EST, splicing consensus, EST annotation, or base composition bias if single exon. The predicted orientation is then used to map the EST to the nearest annotated gene. To avoid incorrect mapping due to annotation gaps, we determine an organism specific threshold for putative 3' UTR length to limit ESTs from mapping to a distant gene (see Figure 4).

Gene annotations for PACdb are primarily obtained from Ensembl (Hubbard, et al., 2002) at this time. However, where Ensembl doesn't have information (especially for plants), we obtain gene annotation from organism-specific websites. In future work, we will expand gene annotations to multiple sources.

Database Implementation

Data from EST analysis and 3'-processing determination is stored in three inter-connected databases. The first database, ESTdb (Figure 2a), is a relational representation of NCBI's dbEST flat files (Boguski et al., 1993). This database contains the tissue, sequence, and other information found for each EST in dbEST. We are also able to add EST annotations, such as quality assessments of both ESTs and whole EST Libraries.

The second database, Sequence-Genome Analysis Database (SeqGen, Figure 2b), primarily stores the EST-Genome alignments and the EST filtering information, but can store any information that can be represented as an alignment. For example, SeqGen also stores the filtering information on each EST. This is similar to an alignment because the filtering program identifies a certain region of the EST which matches a particular contaminant or feature.

The third database, PACdb (Figure 2c), interprets the data in both EST db and SeqGen to identify and assess putative 3'-processing sites. For each putative 3'-processing site, PACdb stores the location of the site, the gene that it is associated with (if any), and the EST(s) that support the site. By linking this information with each putative 3'-processing site, it also becomes easy to access EST alignment, tissue, and quality information which can help the user decide when or if a putative 3'-processing site is actively used.

Data Analysis and Characterization

PACdb currently stores putative 3'-processing site information for human, mouse, rat, dog, chicken, zebrafish, fugu, fruitfly (*D. melanogaster*), mosquito, nematode (*C. elegans*), *A. thaliana*, rice (japonica), and baker's yeast. Table 1 lists various general statistics about the data in PACdb by organism, including the number of ESTs (after filtering and alignment thresholding), the number of coding genes with at least one putative 3'-processing site, the number of putative 3'-processing (cleavage) sites, the number of coding genes with only a single site, the number of coding genes with two or more putative 3'-processing sites, the number of coding genes with internal (inside the coding sequence) putative 3'-processing sites, and the number of singleton ESTs.

Note that the number of ESTs and genes shown are likely less than the maximum number of ESTs and genes. This is because ESTs found to contain contaminants (internal vector, E. coli genomic sequence, etc) are removed from consideration, and because PACdb only catalogs genes that have at least one putative 3'-processing site. Also note that the putative 3'-processing sites for data shown in Table 1 are unclustered (as described in the next paragraph).

Clustering adjacent putative 3'-processing sites would reduce the number of singleton ESTs and putative 3'-processing sites (depending primarily on window size). However, clustering can lead to an artificial grouping of separate cleavage sites when a clustering window size is arbitrarily assigned. Heterogeneity in mRNA 3'-processing is illustrated in Figure 3 for some organisms in PACdb. As can be seen in this figure, the threshold separation for clustering neighboring sites will necessarily be organism-specific, and determined empirically from the data (Figure 3).

It is important to note that PACdb does not force clustering on the data, all putative 3'-processing sites are recorded as the evidence suggests. Currently we leave it up to the user to assess heterogeneity, but in the near future we plan to offer organism-specific clustered data as well as the "raw", unclustered data. Users will have the option of viewing both data simultaneously on a gene-specific basis using the PACdb web interface.

EST-Site Properties and Confidence Level Assignment

For each EST-Site pair, certain properties are measured and recorded in PACdb. Some of the properties that are measured include site-flanking genomic A-rich sequence (evidence of internal priming), restriction enzyme cuts (a source of false 3' ends when there is no tail), unique versus multiple 'good' genomic alignments for a single EST, the presence of a trailing polyA or leading polyT tail on an EST, and the number of ESTs that support a given site. See Table 2 for data on recorded properties across organisms in PACdb.

To calculate the "A-rich Score" for a given nucleotide region, the number of runs of consecutive A's is observed and scored. We chose this analysis as a simple method of reproducing the thermodynamic phenomena that control primer binding. Thresholds were determined empirically through comparison of genomic data with simulated data based on observed dinucleotide frequencies. The score is calculated by using the sum $\sum 2^n$, where "n" is the length of a single consecutive run of A's and the sum is over all runs of consecutive A's in the region. The higher the score, the more likely that this is a candidate for false priming where the oligo-dT primer used in EST creation hybridized to 'internal' mRNA sequence rather than to the polyA tail. Improved distinction between true 3'-ends and internally primed sequences is an ongoing effort. Additional EST-Site properties are recorded but not reported here. See the "Confidence Levels" page on the PACdb website for more details on each property and a list of properties that we plan to record in the future.

The properties mentioned above are used to rationally assign a "confidence level". This assignment is summarized in Table 3. Certain properties (gray in Table 3), e.g., probable internal priming, can lower the confidence level if they alone are satisfied. If none of these "auto-assignment" fields are satisfied, then the highest possible confidence level is assigned as long as all of the required fields (anything except D/M) are satisfied. A field is considered satisfied if the value is at or better than what is required for the column. For example, consider an EST-site pair with no 3' restriction enzyme matches, an A-rich score less than 8, 6 or more ESTs (all uniquely aligned to the genome) supporting the site, but no polyA/polyT tail. None of the auto-assignment fields are satisfied, so we start at "Very High". Everything is satisfied in the "Very

High” column except the tail requirement, so the EST-site pair gets assigned to a confidence level of “High”.

Confidence level assignment is useful in determining whether a putative 3'-processing site is actively used. However, lack of assignment as either “Very High” or “High” confidence does not necessarily negate a 3'-processing site. For instance, rarely expressed genes, or rarely selected 3'-processing sites would correspondingly be rarely observed in our EST data, decreasing the confidence assessment. Also, in the plant *A. thaliana*, nearly all EST tails were clipped before being submitted to the public repositories (see “PolyA Tail Evidence” in Table 2) and therefore cannot attain “Very High” confidence since that data has been lost. Because of these and other similar examples, PACdb catalogs all putative 3'-processing sites, but provides the EST-site properties and confidence levels to help users intelligently filter false sites.

Table 4 shows results of confidence level assignment across organisms in PACdb.

Putative 3' UTR Length and Intergenic Distance

During the automated 3'-processing site determination process, ESTs are assessed for their likely orientation with respect to their transcript, and they are assigned to the nearest annotated (coding) gene within a certain threshold. This threshold is organism-specific, and prevents the assignment of ESTs to distant genes when the actual case is that there is a missing gene annotation. This phenomenon can be observed in Figures 4a-b for mouse and plants. In these figures the putative 3'-processing sites were mapped to the nearest gene regardless of the distance. The resulting putative 3'-UTR length distribution shows a bimodal characteristic, which when plotted along with the distance between genes (stop codon to stop codon), clearly indicates missing gene annotations/predictions.

Missing gene annotations could arise due to rarely transcribed protein-coding genes, however, for deeply studied organisms such as mouse or human, this is less likely than the intriguing possibility of a polyadenylated non-coding RNA (ncRNA) gene. In PACdb, we catalog and assign a confidence level to all putative 3'-processing sites, regardless of whether or not a putative gene can be assigned. This data could be used to locate putative ncRNA genes, especially for higher confidence sites of well-annotated organisms. For less-studied organisms, these sites could equally well indicate a missed gene annotation. See Table 5 for the confidence level breakdown of 3'-processing sites with no gene association for a selection of organisms.

It is also interesting to compare the distributions of putative 3'-UTR length among various organisms. As expected, closely related organisms exhibit similar putative 3'-UTR lengths (Figure 5a). It is also interesting to note that the slope of the cumulative distributions is less sharp in higher animals than lower animals and plants (Figure 5b). This illustrates the expected result (based on previous studies) that 3'-UTRs are more variable and longer in higher organisms and also suggests increased post-transcriptional regulation by the action of 3'-UTR *cis*-elements in metazoans, as compared to plants.

When graphing the median 3'-UTR length (restricted to high confidence sites) versus genome size, an approximately logarithmic trend appears (Figure 6). This trend is also hinted at in Figure 4b, where rice, with the larger genome, has putative 3'-UTR length and intergenic distributions that are longer than the corresponding *A. thaliana* distributions. However, median 3'-UTR length is not simply dependent on the genome size, and is likely also related to organism complexity, gene content, and extent of post-transcriptional gene regulation.

Putative 3'-Processing Sites per Gene

When looking at the number of sites per gene across organisms, the organism-specific curves tend to group based on the number of ESTs or putative sites for that organism, and not by evolutionary relationship (Figure 7a). This is even true when limiting data to higher confidence, clustered sites (Figure 7b).

When looking at putative sites per gene versus ESTs per gene, the trend of more sites given more ESTs is strong (Figure 8a). However, when restricting data to higher confidence, clustered sites, this trend is greatly reduced (Figure 8b), though still present.

Canonical Word Usage in the PolyA Signal (PAS)

For polyadenylation to occur, the cleavage and polyadenylation machinery must recognize and bind the polyA signal (PAS) in the pre-mRNA. The core of the canonical PAS is the word, "AAUAAA", although there is evidence that variants on this word are also used. Table 6 shows data on usage of the canonical hexamer and all single nucleotide variants. Sequences were classified by the presence of AAUAAA (column 4), AAUAAA or AUUAAA (column 5), or any single base variant (column 6) situated between 10 and 60 nucleotide upstream of the putative 3'-processing site. All sites for this table are clustered according to organism-specific thresholds (see Figure 3) and are "Very High" confidence except that the number of supporting ESTs varies (column 2 in the table). Comparison with Table 1 makes it clear that as the depth of sampling is increased, presence of the canonical hexamer or its most common variant AUUAAA decreases. We interpret this as evidence of inclusion of both rarely used *bona fide* 3'-processing sites, as well as false positives. The requirement of greater numbers of supporting ESTs in organisms with larger data sets (such as human or mouse) effectively reduces the sampling of the rarer sites, and results in measured hexamer usage similar to organisms with smaller EST collections (e.g., dog or fugu). As previously noted, plants have a greatly reduced usage of the canonical hexamer (Graber et al., 1999; Rothnie, 1996; Li and Hunt, 1997).

Supplement Conclusion

PACdb currently contains putative 3'-processing information for 13 organisms: human, mouse, rat, dog, chicken, zebrafish, fugu, fruitfly (*D. melanogaster*), mosquito, nematode (*C. elegans*), *A. thaliana*, rice (japonica), and baker's yeast. PACdb catalogs all putative 3'-processing sites and does not throw out sites that are less than "Very High" quality and also doesn't force clustering based on a single arbitrarily chosen window. However, PACdb records EST-site properties which can help the user more intelligently filter out false 3'-processing sites.

Future work on PACdb will involve keeping up-to-date with new genome versions and additional ESTs, including additional organisms in PACdb (such as *Ciona intestinalis*, *Xenopus tropicalis*, cow, wheat, and maize), using multiple sources for gene annotations, additional web interface features, and offering clustered site data based on organism-specific windows as well as the unclustered data.

Tables and Figures

Scientific Name	Common Name	ESTs in PACdb	Putative Sites	Singleton ESTs	Genes in PACdb	One Site Genes	Multi-Site Genes	Genes with Internal Sites
<i>M. musculus</i>	Mouse	794819	468695	374576	18890	2393	16498	11121
<i>R. norvegicus</i>	Norway Rat	219813	92016	59437	16346	2783	13563	8790
<i>T. rubripes</i>	Fugu Puffer	4479	2950	2445	2112	1657	455	548
<i>A. thaliana</i>	Thalecress	178893	90272	65548	15495	4078	11417	2634
<i>O. sativa</i>	Rice	83274	48400	36685	14693	5508	9185	2480
<i>S. cerevisiae</i>	Yeast	1414	1187	1029	826	618	208	127
<i>C. elegans</i>	Nematode	78041	47719	37476	11259	4387	6872	8299
<i>A. gambiae</i>	Mosquito	9905	7643	6616	4115	2504	1611	1148
<i>D. melanogaster</i>	Fruit Fly	55019	34003	27426	10383	3309	7074	5533
<i>D. rerio</i>	Zebrafish	107952	53966	42374	12847	4722	8125	6613
<i>G. gallus</i>	Chicken	35895	28153	24342	8724	3523	5201	4713
<i>H. sapiens</i>	Human	1522036	468208	333132	20204	1321	18883	16962
<i>C. familiaris</i>	Dog	88828	29730	20260	9286	3511	5775	2371

Table 1: General statistics for organisms in PACdb, all confidence levels included in this data

Organism	ESTs	A-rich Score				Restriction Enzyme Match			Genomic Hits		PolyA Tail Evidence		
		S < 8	8 ≤ s < 16	16 ≤ s < 36	s > 36	No RE	Imperfect RE	Perfect RE	Unique	Multi Hit	No tail	Trailing PolyA	Leading PolyT
Mouse	845835	57.2%	26.2%	6.5%	10.2%	84.3%	13.2%	2.5%	90.8%	9.2%	77.1%	8.5%	14.2%
Rat	219813	55.3%	23.4%	6.5%	14.8%	88.1%	9.1%	2.8%	100%	0%	33%	2.2%	62.7%
<i>A. thaliana</i>	180176	53%	36.5%	7.1%	3.4%	98.2%	1.4%	0.4%	98.8%	1.2%	96.8%	1.7%	1.6%
Rice	87614	57.7%	33.8%	4.9%	3.5%	52.2%	39.4%	8.4%	93.1%	6.9%	32.9%	43.2%	23.8%
Zebrafish	107952	51.3%	31%	9.1%	8.6%	96.8%	2.7%	0.4%	100%	0%	65.5%	14.8%	19.2%
Fugu	4479	57.8%	28.6%	5.4%	8.2%	96.9%	2.9%	0.2%	100%	0%	35.3%	34.6%	29.9%
Fruitfly	55019	42.3%	36.8%	11.5%	9.4%	92.6%	3.5%	3.9%	100%	0%	91.6%	5.7%	2.2%
Nematode	78041	55.7%	32.9%	8.2%	3.2%	99.5%	0.2%	0.2%	100%	0%	96.7%	3.2%	0.1%
Mosquito	9905	45.7%	32%	9.7%	12.5%	95.4%	1.2%	3.4%	100%	0%	69.3%	17.7%	12.8%
Dog	88828	60.1%	21.9%	5.7%	12.3%	96.7%	2.8%	0.6%	100%	0%	88.1%	0.9%	10.1%
Chicken	35895	60.8%	24.9%	5.2%	9%	96.4%	3.1%	0.5%	100%	0%	67.9%	22.4%	9.6%
Human	1551223	57.8%	24.3%	7%	10.9%	93.3%	5.8%	0.9%	96.3%	3.7%	61.2%	10.1%	27%
Yeast	1427	Data Temporarily Unavailable											

Table 2: EST-site pair properties statistics for data in PACdb. Organisms with 100% unique genomic hits were processed with a slightly more restrictive alignment threshold. In the next round of data updates, these percentages will likely change.

Bit Property	Very High	High	Medium	Low	Very Low
Genomic A-rich score (false priming)	$s < 8$	$s < 8$	$8 \leq s < 16$	$16 \leq s < 36$	$s > 36$
EST has unique/multiple genomic hits	Unique	Unique	Unique	Multiple	D/M
Number of EST hits supporting site	≥ 6 ESTs	≥ 3 ESTs	≥ 1 EST	≥ 1 EST	≥ 1 EST
3' end restriction enzyme site match (and no polyA/T tail)	None	None	None	Imperfect match	Perfect match
PolyA/T tail	Present	D/M	D/M	D/M	D/M

Table 3: Confidence level calculation based on EST-site properties. If a gray field is satisfied that field alone determines the confidence level. Otherwise an EST-site pair is assigned the highest confidence level where all conditions are met ("D/M" = Doesn't matter)

Confidence Level	Mouse	Rat	Fugu	<i>A. thaliana</i>	Rice	Nematode	Mosquito	Fruitfly	Zebrafish	Chicken	Dog	Human
Very High	68670	43485	399	885	9376	409	319	815	11667	1841	4069	270650
High	97798	36423	1653	38984	25773	27759	3459	14851	21686	15586	13640	215622
Medium	181842	71836	1675	97232	33101	40721	3482	25961	42428	10001	50677	618115
Low	56166	11730	232	11941	9060	6410	929	5691	8663	1630	4680	120227
Very Low	97828	38311	395	7794	10304	2741	1414	6864	9402	3455	11118	245319

Table 4: EST-site pair confidence levels by organism for sites mapped to genes (Yeast confidence levels temporarily unavailable)

Confidence Level	Mouse	Rat	<i>A. thaliana</i>	Mosquito	Zebrafish	Chicken	Dog	Human
Very High	1585	612	41	7	1003	20	147	1895
High	130382	5489	8230	145	4424	1880	1242	28729
Medium	97444	5742	11458	79	6215	1040	1889	30543
Low	37446	1448	2534	26	1002	180	308	6757
Very Low	76674	4737	1077	45	1462	262	1058	13366

Table 5: Confidence levels for putative 3'-processing sites that are not associated with a gene (selected organisms only). Higher confidence sites are excellent candidates for putative ncRNA gene searching.

Organism	Minimum supporting ESTs	Sites	AAUAAA	AWUAAA	AAUAAA and 1 nt variants
human	1	7783	36.52%	51.32%	81.14%
	5	1540	56.23%	72.99%	93.38%
	10	738	60.84%	79.54%	95.12%
dog	1	1015	62.36%	78.13%	95.27%
	2	396	67.93%	84.09%	97.47%
rat	1	9038	50.23%	66.08%	88.29%
	5	2378	66.99%	83.6%	97.06%
mouse	1	7237	45.67%	60.56%	82.64%
	5	1611	66.36%	83.43%	96.83%
chicken	1	2344	63.05%	80.03%	93.6%
	2	1047	70.2%	86.25%	95.8%
zebrafish	1	2534	64.52%	82.08%	96.45%
	3	505	77.03%	90.89%	99.8%
fugu	1	820	70.98%	97.07%	97.07%
<i>D. melanogaster</i>	1	142	53.52%	66.2%	94.37%
mosquito	1	451	66.08%	69.62%	91.13%
nematode	1	325	41.85%	44.62%	86.15%
<i>A. thaliana</i>	1	656	8.54%	11.89%	52.29%
rice	1	8502	7.46%	9.74%	43.44%

Table 6: Canonical and variant polyadenylation signal word usage within 10 to 60 bases upstream of sites for organisms in PACdb except yeast. All sites are clustered according to organism-specific thresholds and are “Very High” confidence (no multi-hit ESTs, no internal primed candidates, evidence of a polyA tail required) except that the number of supporting ESTs varies per organism (e.g. 1, 5, and 10 for human; 1 and 2 for dog).

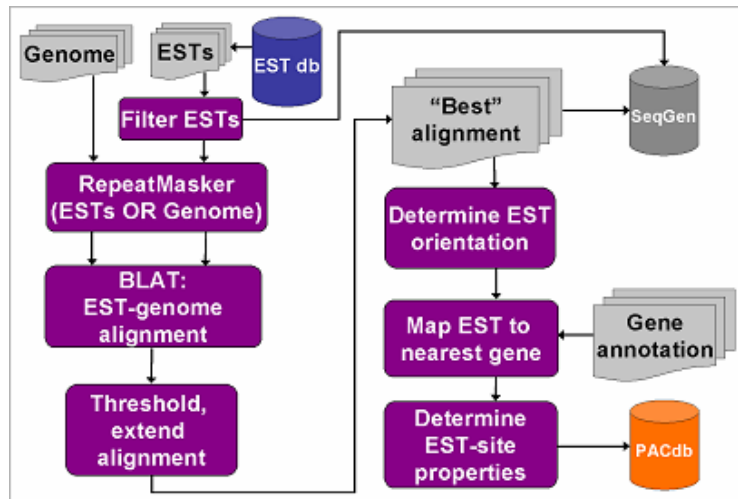


Figure 1: Flowchart of methods used to generate PACdb data. Here the “pages” represent text-based data, the rounded rectangles represent programmatic tasks, cylinders represent database servers, and arrows indicate the flow of data.

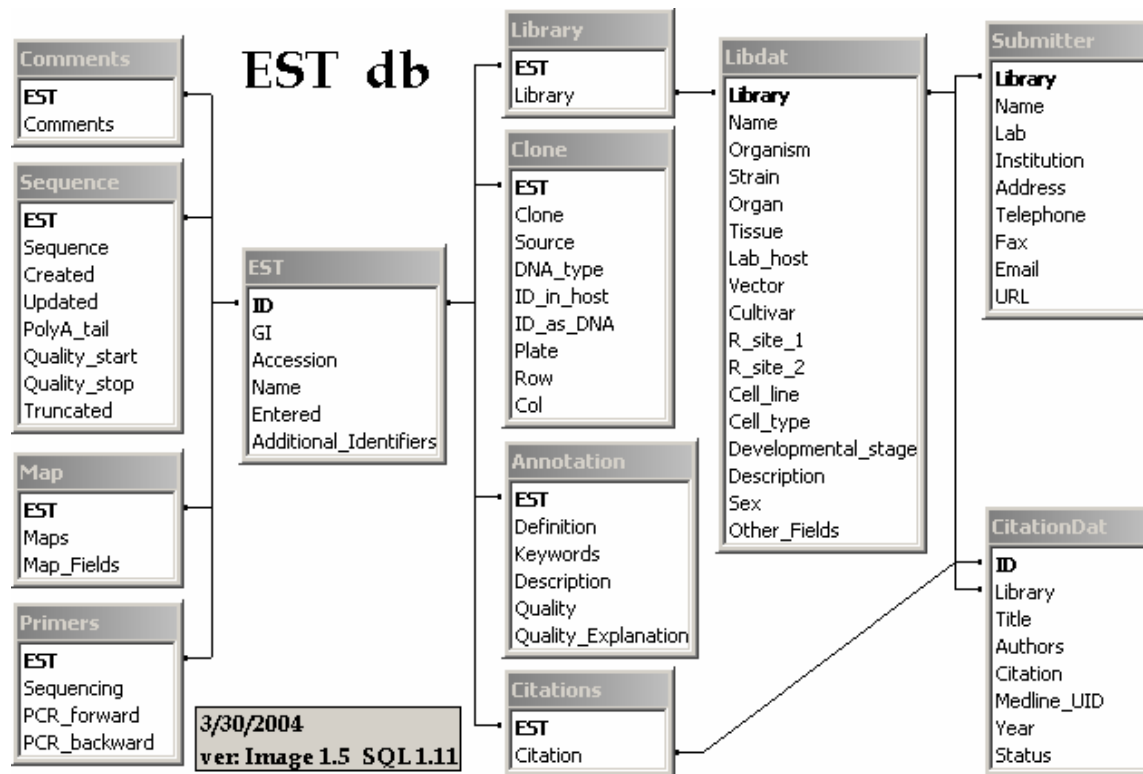


Figure 2a: EST db schema, based on NCBI's dbEST flat files

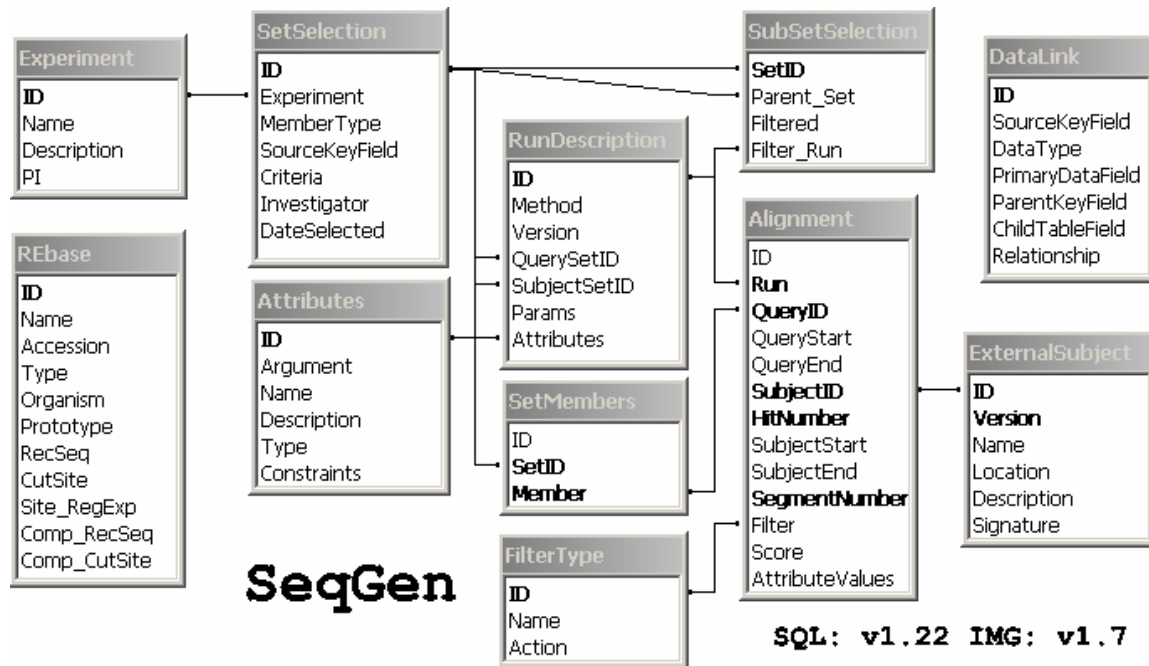


Figure 2b: SeqGen (Sequence-Genome Analysis Database) schema for holding EST/cDNA-Genome alignments as well as EST filtering results.

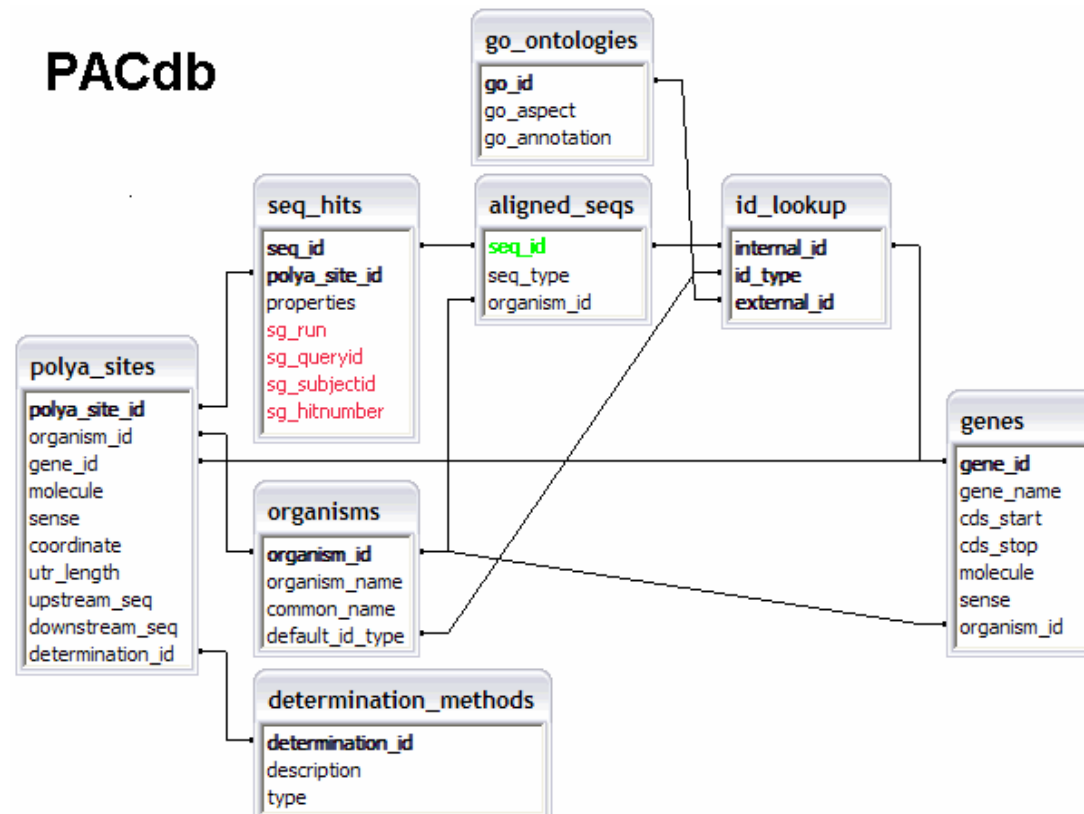


Figure 2c: PolyA Cleavage Site and 3'-UTR Database (PACdb) schema. Red fields link to SeqGen and green fields link to EST db.

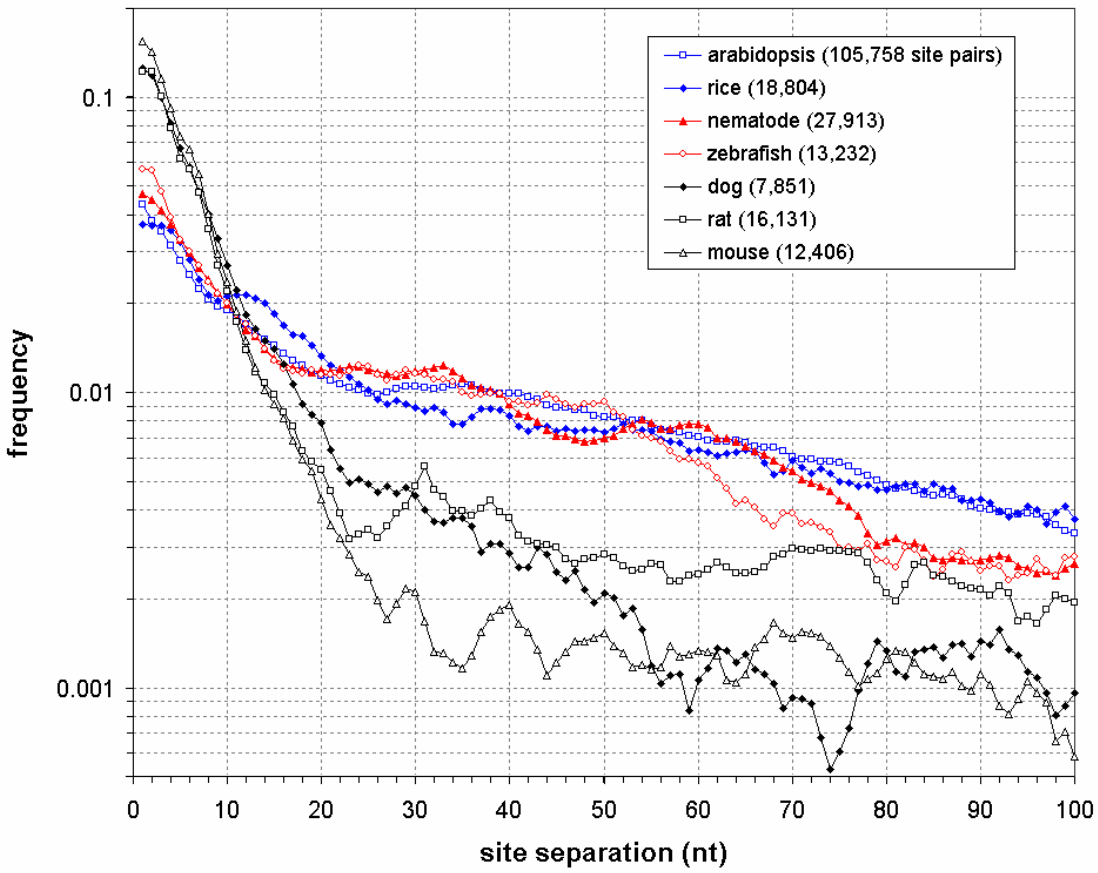


Figure 3: The distribution of distance between neighboring 3'-processing sites within a gene. All data used is "Very High" confidence except that the number of supporting ESTs is not restricted.

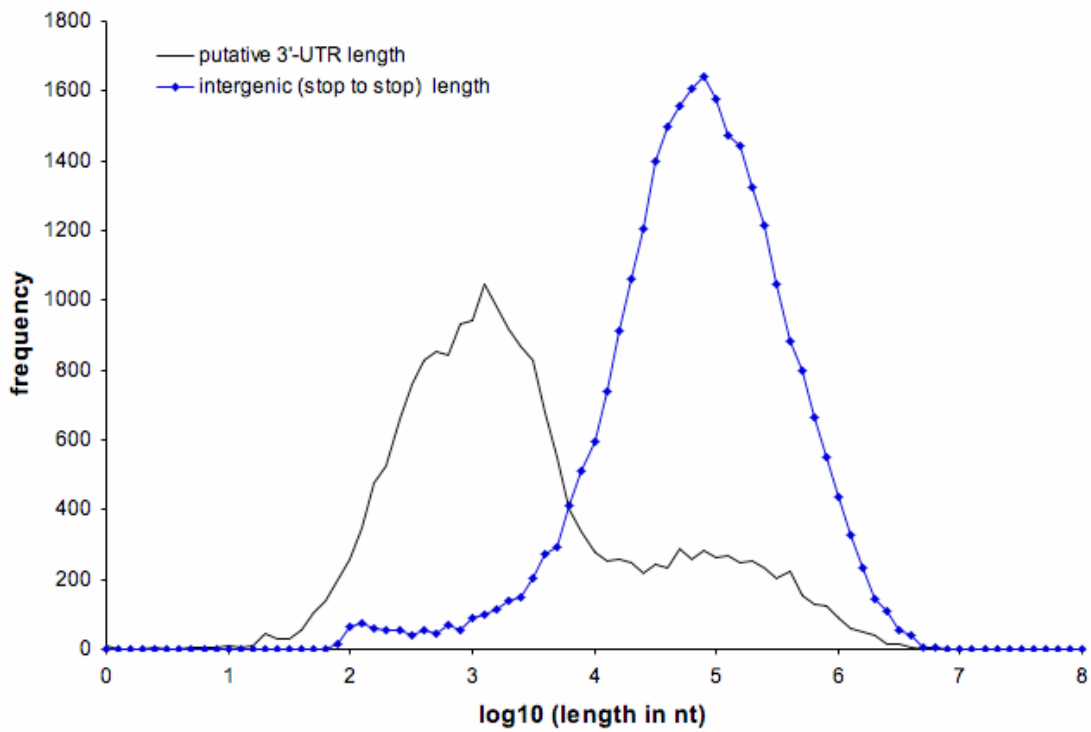


Figure 4a: *Mus musculus* putative 3' UTR and intergenic (stop codon to stop codon) distance distribution

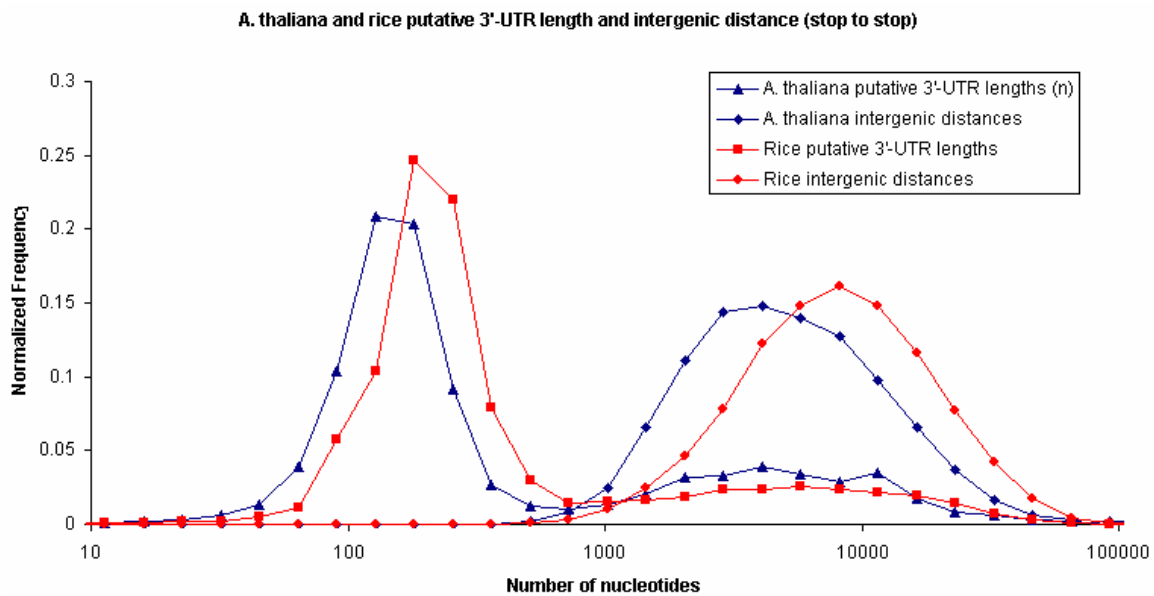


Figure 4b: Plant putative 3'-UTR length and intergenic (stop codon to stop codon) distance graphs

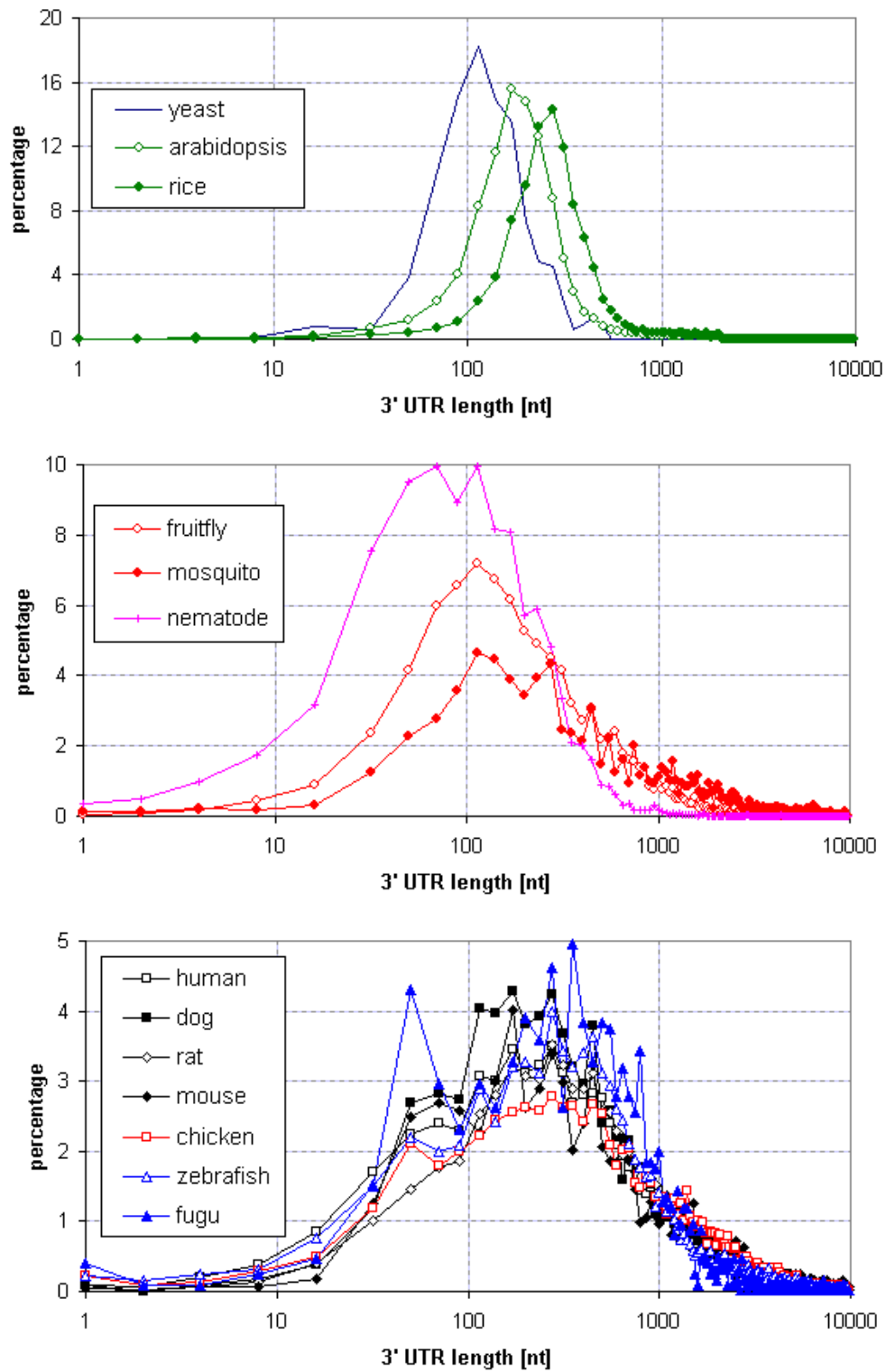


Figure 5a: Putative 3'-UTR length distribution for several organisms in PACdb.

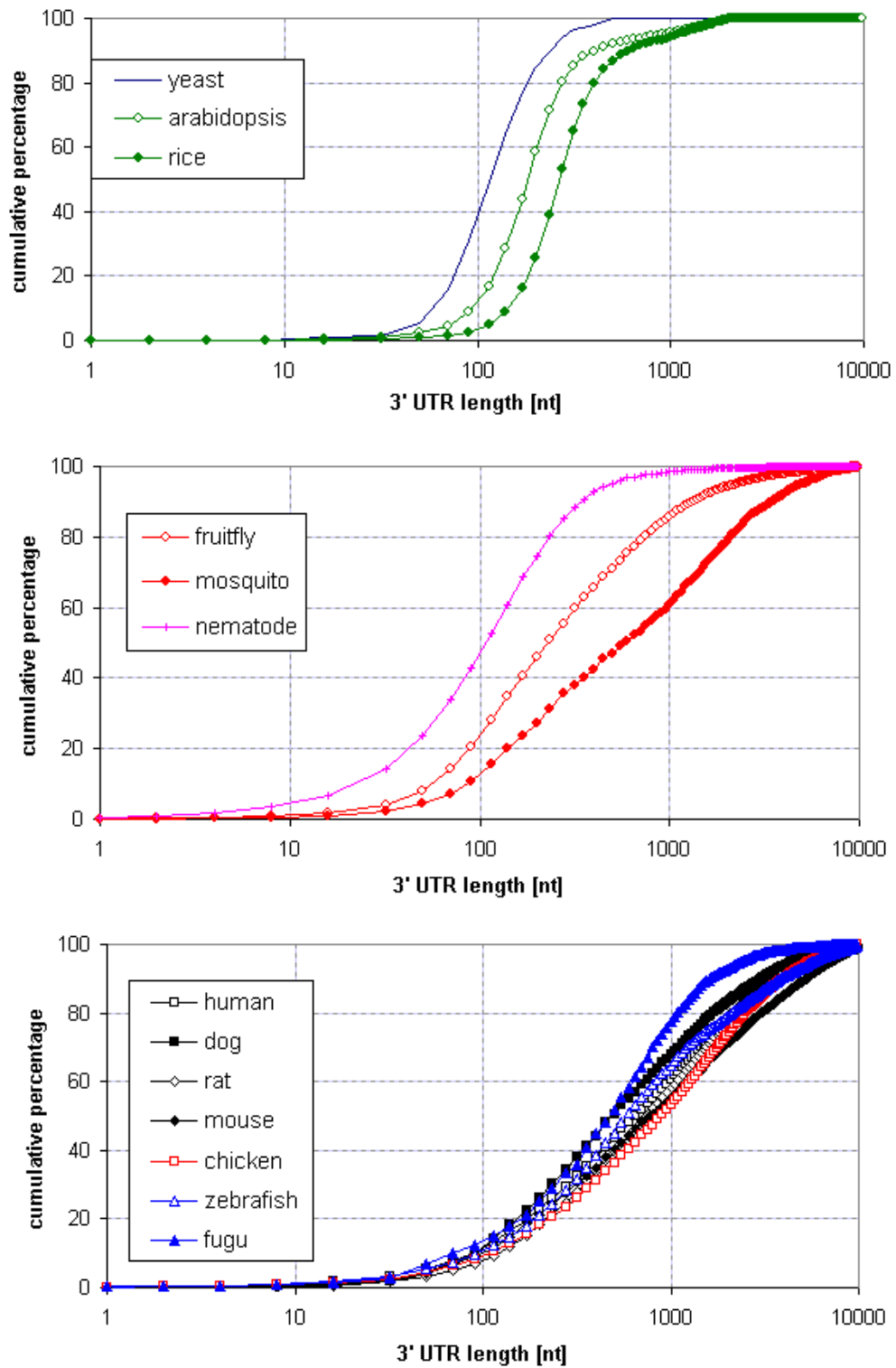


Figure 5b: Cumulative distribution of putative 3'-UTR lengths for several organisms in PACdb.

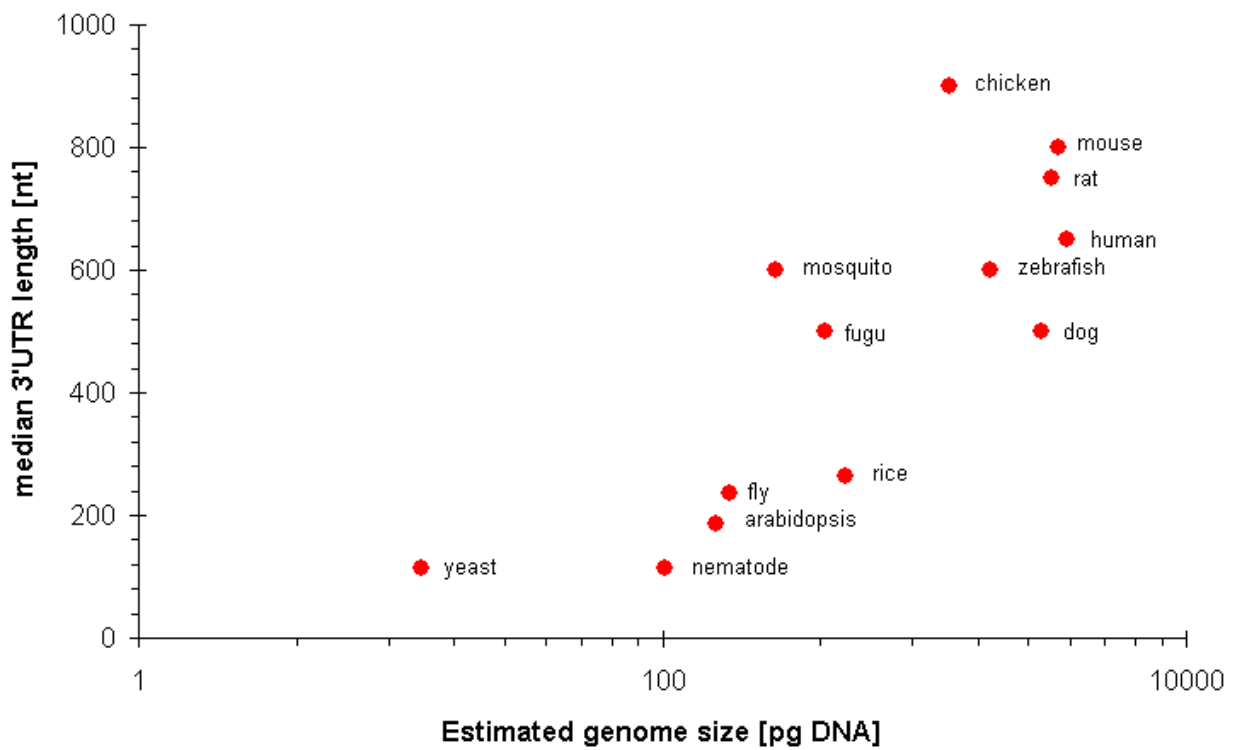


Figure 6: Median putative 3'-UTR length vs. estimated genome size for all organisms currently in PACdb. Genome sizes from Gregory, 2005; Bennet and Leitch, 2005; Kullman, *et al.*, 2005

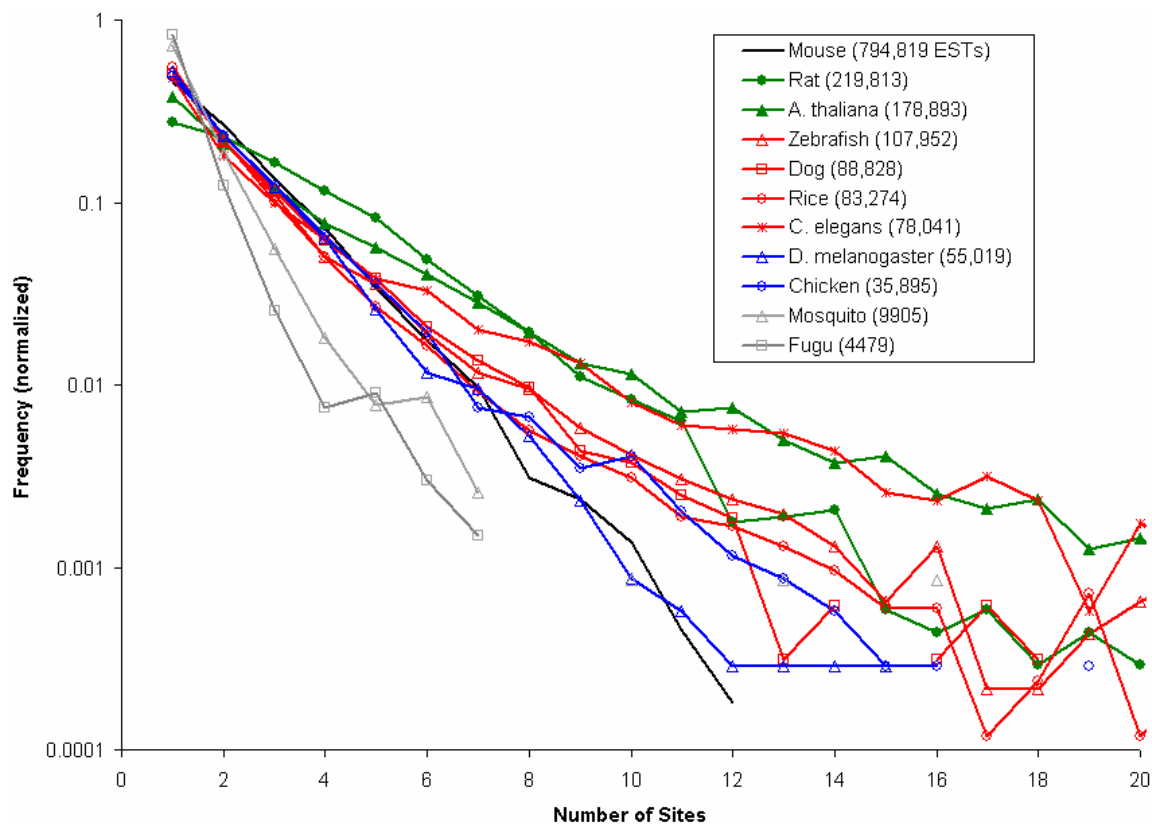
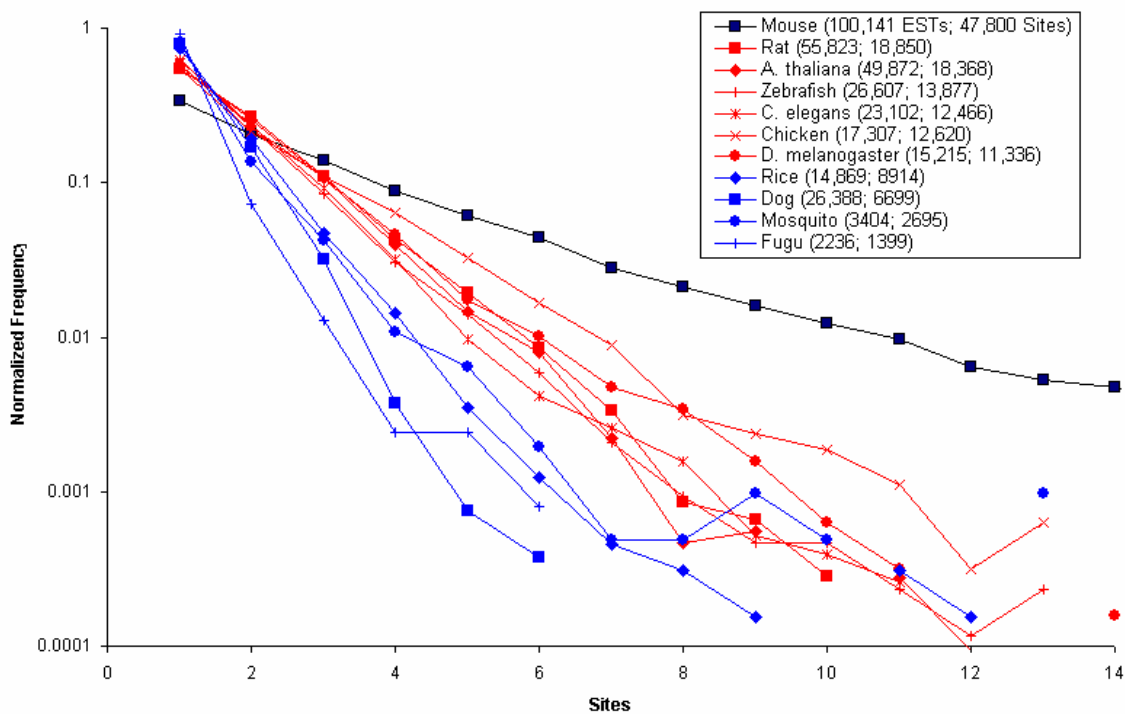


Figure 7a: Putative 3'-processing sites per gene across organisms in PACdb
 Condensed Sites per Gene, Hi Conf (No RE, No Int. Priming, No Multi-Hit)



7b: Putative "clustered, high confidence" 3'-processing sites per gene across organisms in PACdb

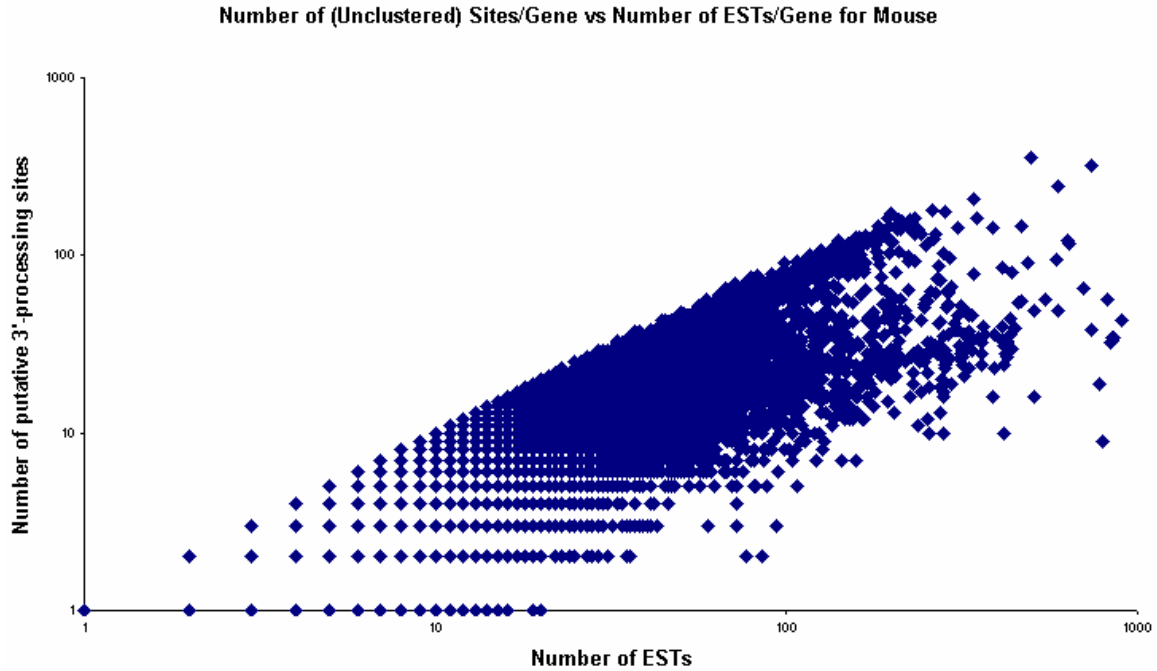


Figure 8a: Unique 3'-processing sites per gene vs ESTs per gene in mouse. Sites are unclustered and include all confidence levels.

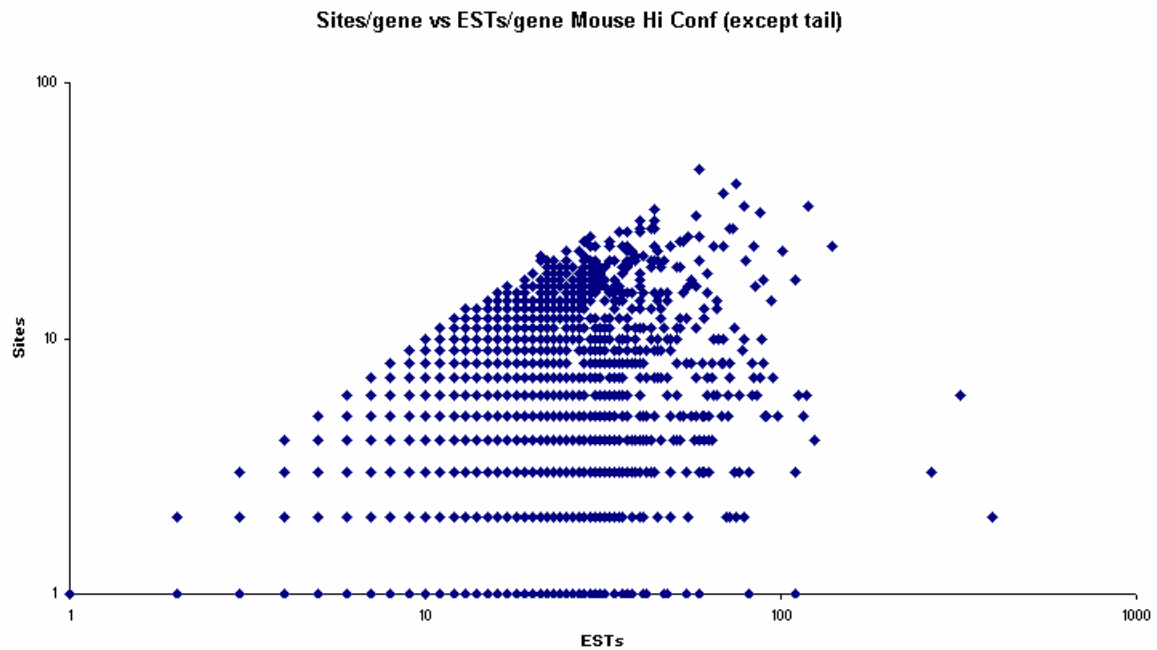


Figure 8b: 3'-processing sites per gene vs ESTs per gene in mouse, restricted to high confidence, clustered sites.

Supplement References

- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389-3402.
- Bennett M.D., Leitch I.J. (2005) Nuclear DNA amounts in angiosperms - progress, problems and prospects. *Annals of Botany*, **95**, 45-90.
- Boguski, M.S., Lowe T.M., Tolstoshev C.M. (1993) dbEST--database for "expressed sequence tags". *Nat. Genet.*, **4**, 332-333.
- Graber, J.H., Cantor, C.R., Mohr, S.C., and Smith, T.F. (1999) *In silico* detection of control signals: mRNA 3'-end-processing sequences in diverse species. *Proc Nat Acad Sci USA*, **96**, 14055-14060.
- Gregory, T.R. (2005). Animal Genome Size Database. <http://www.genomesize.com>.
- Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., et al. (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38-41.
- Kent, W.J. (2002) BLAT - the BLAST-like alignment tool. *Genome Res.*, **12**, 656-664.
- Kullman, B., Tamm, H. Kullman, K. (2005) Fungal Genome Size Database. <http://www.zbi.ee/fungal-genomesize>.
- Li, Q. and Hunt, A.G. (1997) The polyadenylation of RNA in plants. *Plant Physiol*, **115**, 321-325.
- Rothnie, H.M. (1996) Plant mRNA 3'-end formation. *Plant Mol Biol*, **32**, 43-61.