

Quality of *Daphnia* cDNA libraries: pilot project for the production of full-length cDNA sequences for the *Daphnia* genome annotation and for microarray probes

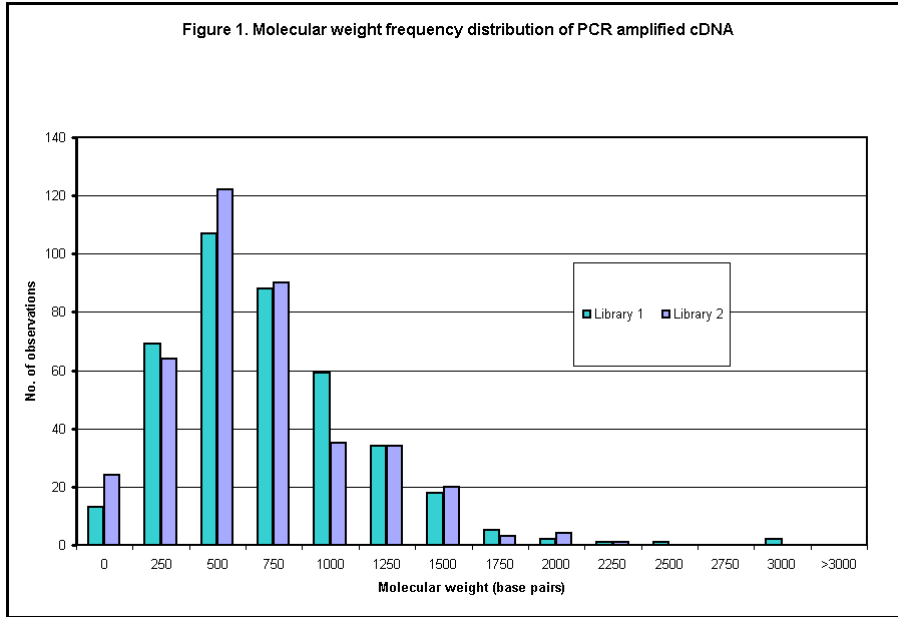
Darren Bauer and Kelley Thomas – HCGS, University of New Hampshire
Elizabeth Bohuski and John Colbourne – CGB, Indiana University

The *Daphnia* genome sequencing project is now underway at the Joint Genome Institute. Although gene prediction algorithms are increasingly growing more sophisticated and accurate in producing genome annotations, there is currently no better resource for discovering genic regions than full-length cDNA sequences. For the purpose of both helping in the annotation of the genome and creating a Unigene (encyclopedic) collection of cDNA for printing onto microarrays, our colleagues at HCGS are generating full-length cDNA libraries from cultures of this same isolate (log50) exposed to seven separate ecological challenges (hypoxia, starvation, exposure to chaoborus, fish exposure, bacteria, UV and heavy metals). To ensure that cDNA libraries will adequately and consistently serve these functions, the present pilot project aimed to determine the quality of cDNA libraries constructed using the Creator SMART (Clontech) system, by harvesting tissues of the surrogate clone (log52) under standard laboratory growing conditions.

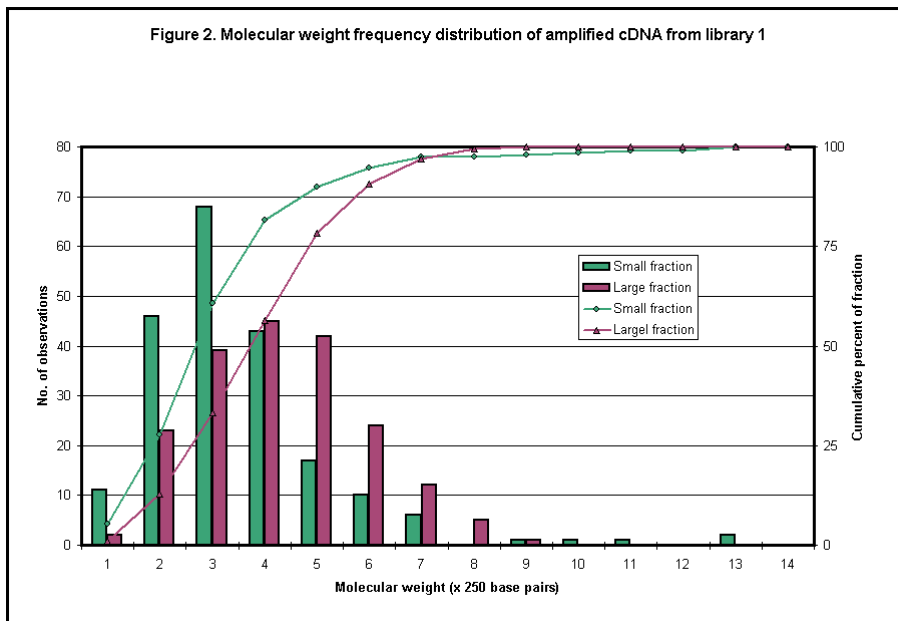
Two libraries were created from separate harvests of log52 clones. This situation permitted an evaluation of the consistency in the protocol – from isolating RNA to creating frozen glycerol stocks of bacterial isolates with full-length cDNA. Library 1 was transferred from agar to 384 well plates by hand. Library 2 was picked robotically. To control normal bias in the ligations toward smaller fragments, reactions were performed on four enriched full-length double stranded cDNA size fractions. Transformants were then pooled into two size classes. From each frozen library, 384 colonies were chosen for sequencing. Half were from the large size class and half from the small class. Bacterial vectors were amplified, plasmids were purified by standard plasmid-prep protocols and quantified by spectrophotometer. Product sizes were measured by PCR amplification of cDNA inserts followed by agarose gel electrophoresis against standards and visualized using a Kodak 440cf imaging station. Sequencing reactions were performed by priming at the 5' end of cDNA using ABI chemistry and the 3700 sequencer. Vector sequence and poor quality sequences were trimmed and contigs were assembled from redundant reads using the SeqManII software (DNASTAR package). Homologies with Genbank entries were discovered using Blastx against the non-redundant (nr) protein database. Those sequences with expectation-values better than 1×10^{-27} were further examined for the presence of an annotated ATG start codon at the 5' end of the ORF. This last step was accomplished using NCBI's ORF finder service (<http://www.ncbi.nlm.nih.gov/gorf/>). Only those sequences whose Methionine aligned (including gaps) with the first amino acid of complete sequences were considered full-length transcripts.

1) Sizing of PCR amplified cDNA

Of the 384 samples from each library, library 1 yielded 399 amplified products while library 2 yielded 398 PCR fragments. These irregular results stem from the presence of 2-3 cDNA vectors within the same location of a 384 well plate (robot and human are evenly matched here). Three samples from library 1 failed to amplify. Four PCR amplifications failed from library 2. Size distributions of the products from both libraries are shown in Figure 1. Mean sizes of cDNA fragments (minus 260 plasmid nucleotides) for libraries 1 and 2 are 847 bp and 803 bp respectively. No significant differences are found.



As expected, the mean size distributions were significantly different between the two size classes obtained from RNA fractionation ($p < 0.0001$). Results are shown for library 1 (Figure 2); mean size of the small fraction was 754 bp, whereas the mean size of large fraction was 945 bp.



This study is the first to estimate the size of *Daphnia*'s cDNA sequences based on PCR amplification of putatively full-length cDNA and suggests that they may be substantially shorter than those of other model organisms like *Drosophila* (mean size = 2 kb).

2) Sequencing of cDNA

Of the 768 sequences, 619 were informative. Only four plasmids were void of inserts and the spoiled reactions (19%) were the result of capillary failures. The average length of the trimmed sequences was 529

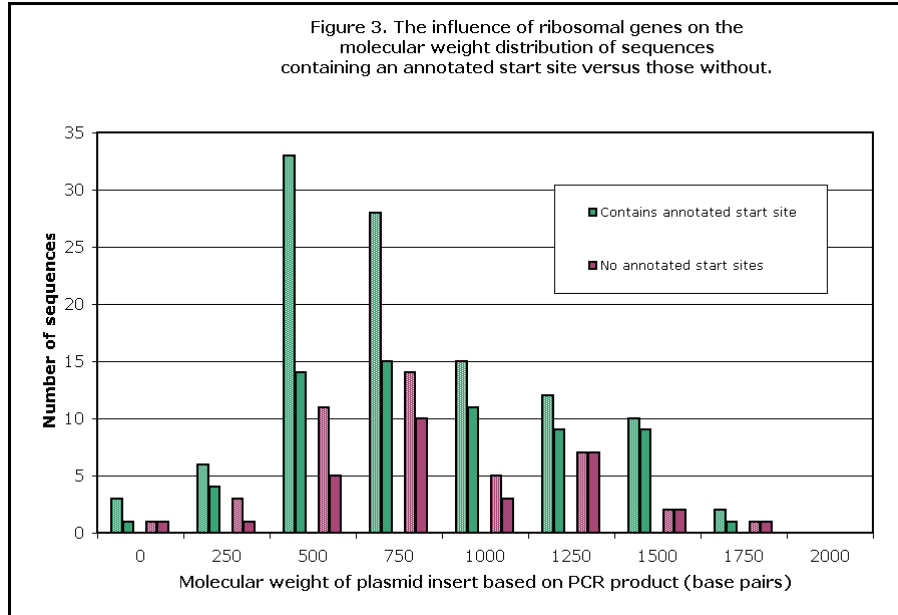
bp; 150 bp of vector sequence were initially trimmed while an addition 128 bp were lost by failing quality thresholds on the eletropherograms.

Unique sequences comprised 68% of the total, reflecting the relative abundances of specific cDNA within the non-normalized libraries. This number would likely diminish with greater sequencing effort. Of the unique sequences, ~50% matched Genbank entries with e-values $< 1 \times 10^{-10}$, and the scores for 204 sequences were better than 1×10^{-27} . A sufficient number and diversity of sequences were gathered to subsequently determine what fraction of cDNAs are full-length. In total, 170 were investigated for the presence of a translational start site.

3) NCBI ORF finding

Of the 170 investigated sequences, 64% contained unambiguous ORFs with an annotated ATG start site at their 5' end, 26% did not contain an ATG that aligned with the start sites of corresponding database sequences. Of the remaining sequences, 4% were likely full-length because gapped alignments of the amino acids suggested poor evolutionary conservation at the N-terminus of the proteins, and 6% were unresolved because alignments failed altogether.

Our estimate, that roughly 64-68% of cDNA are full-length, may be upwardly biased given that the most conserved proteins within our non-normalized libraries are ribosomal and that these seldom have long transcripts, which challenge the performance of the reverse transcriptase. Indeed, the maximum length of investigated cDNA for ORFs was < 2 kb, whereas the maximum length of PCR amplified inserts was nearly 3.5 kb. However, of the investigated sequences, 58 were predicted to be ribosomal genes and only 40% contained the annotated start site. When comparing the number of sequences with and without annotated start sites, sorted by the molecular weight of cDNA (Figure 3), no association was found between the proportion of full-length transcripts and size, including ribosomal genes (first bar of every pair) and excluding these genes (second of the pair).



A separate investigation was achieved, by calculating the proportions of redundant sequences that shared nucleotides within the first 50 bases of the longest cDNA within contigs. Of 233 sequences forming 81 separate contigs (Table 1), 31 (21%) failed to share the first 50 bp of assumed full-length cDNA. In 10% of cases, these sequences matched with nucleotides at another 50 bp downstream.

Table 1. Statistics related to sequences within 81 contigs

Contig	Contig Length	Num. Seq. in Contig	Sharing First 50 bp	Sharing First 100 bp	% With Equal Length
1	613	7	7	7	100
2	478	7	5	6	71
3	1006	7	6	6	86
4	704	7	7	7	100
5	848	7	7	7	100
6	689	7	6	7	86
7	694	6	6	6	100
8	780	5	5	5	100
9	843	4	4	4	100
10	791	4	4	4	100
11	960	4	3	3	75
12	753	4	4	4	100
13	789	4	4	4	100
14	776	4	4	4	100
15	674	3	3	3	100
16	539	3	3	3	100
17	718	3	3	3	100
18	398	3	3	3	100
19	847	3	2	2	67
20	616	3	1	1	33
21	716	3	3	3	100
22	537	3	2	3	67
23	493	3	3	3	100
24	723	3	1	2	33
25	719	3	3	3	100
26	1277	3	1	1	33
27	622	3	3	3	100
28	829	3	3	3	100
29	775	3	2	2	67
30	783	3	3	3	100
31	704	3	3	3	100
32	801	3	2	2	67
33	571	3	3	3	100
34	677	3	3	3	100
35	771	3	2	2	67
36	682	3	3	3	100
37	544	2	2	2	100
38	386	2	2	2	100
39	445	2	2	2	100
40	514	2	2	2	100
41	408	2	2	2	100
42	490	2	1	1	50
43	479	2	2	2	100
44	785	2	1	1	50
45	378	2	1	2	50
46	459	2	2	2	100
47	556	2	1	2	50
48	505	2	2	2	100
49	658	2	1	1	50
50	524	2	1	2	50
51	571	2	2	2	100
52	803	2	2	2	100
53	602	2	2	2	100

54	579	2	2	2	100
55	574	2	2	2	100
56	754	2	2	2	100
57	810	2	1	1	50
58	1149	2	1	1	50
59	422	2	1	1	50
60	575	2	2	2	100
61	746	2	2	2	100
62	507	2	2	2	100
63	348	2	1	2	50
64	701	2	2	2	100
65	609	2	1	2	50
66	729	2	2	2	100
67	778	2	2	2	100
68	777	2	2	2	100
69	775	2	2	2	100
70	720	2	2	2	100
71	606	2	2	2	100
72	731	2	1	1	50
73	790	2	2	2	100
74	634	2	2	2	100
75	797	2	1	1	50
76	739	2	2	2	100
77	557	2	1	1	50
78	537	2	2	2	100
79	259	2	2	2	100
80	598	2	2	2	100
81	940	2	1	1	50
SUM	-	233	202	211	-

4) Summary

This pilot project suggested that we should expect quality cDNA libraries using the Clontech system, yielding a diversity of gene transcripts that average ~800 bp in length with a maximum size >3 kb. Whether their below average size compared to *Drosophila* is a biological reality or simply an artifact from having characterized a non-normalized library is still unclear. However, our initial (costly) strategy to sequence 40,000-50,000 oligonucleotide fingerprinted cDNA from both ends may need revisiting, given that many full-length sequences could be read by priming only from the 5' end.

The estimated fraction of full-length cDNA from ORF investigated sequences is roughly 68%. This figure assumes that our sequences represent a random subset. These results are comparable to those from similarly constructed (yet non-redundant) libraries of human skeletal muscle cDNA (77%) (Biotechniques 30:892-897) and of oral squamous cell carcinoma cDNA (61%) (International Journal of Molecular Medicine 12:429-435). Although these fractions are (as always) below the reported maximum efficiency for the alternative CAP-trapper method used by RIKEN (90%) (Gene 263:93-102), this oligo-trapping technology is sufficient for our purposes. There is seemingly no size bias in the generation of cDNA, and on condition that all transcripts are equally likely to produce full-length cDNA, the oligonucleotide fingerprinting should identify those molecules among the redundant copies that are most likely full-length.