Published in final edited form as:

Insect Mol Biol. 2010 March ; 19(0 2): 23–31. doi:10.1111/j.1365-2583.2009.00946.x.

A full-length cDNA resource for the pea aphid, *Acyrtosiphon pisum*

Shuji Shigenobu^{1,2,3,*}, Stephen Richards⁴, Andrew G. Cree⁴, Mizue Morioka⁵, Takema Fukatsu⁶, Toshiaki Kudo^{7,†}, Shin-ya Miyagishima⁸, Richard A. Gibbs⁴, David L. Stern¹, and Atsushi Nakabachi^{8,*}

¹HHMI and Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ 08544, USA

²PRESTO, JST, 4-1-8 Honcho Kawaguchi, Saitama, 332-0012, Japan

³Okazaki Institute for Integrative Bioscience, National Institutes of Natural Sciences, Higashiyama, Myodaiji, Okazaki 444-8787, Japan

⁴Human Genome Sequencing Center, Baylor College of Medicine, 1 Baylor Plaza, Houston, TX, 77030

⁵Department of Biological Sciences, Graduate School of Science, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

⁶National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba 305-8566, Japan

⁷Discovery Research Institute, RIKEN, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan.

⁸Advanced Science Institute, RIKEN, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan.

Abstract

Large collections of full-length cDNAs are important resources for genome annotation and functional genomics. We report the creation of a collection of 50,599 full-length cDNA clones from the pea aphid, *Acyrtosiphon pisum*. Sequencing from 5' and 3' ends of the clones generated 97,828 high-quality expressed sequence tags (ESTs), representing approximately 9,000 genes. These sequences were imported to AphidBase and are shown to play crucial roles in both automatic gene prediction and manual annotation. Our detailed analyses demonstrated that the full-length cDNAs can further improve gene models and can even identify novel genes that are not included in the current version of the official gene set. This full-length cDNA collection can be utilized for a wide variety of functional studies, serving as a community resource for the study of the functional genomics of the pea aphid.

*Joint correspondence author: Shuji Shigenobu Okazaki Institute for Integrative Bioscience, National Institute for Basic Biology Higashiyama, Myodaiji, Okazaki 444-8787, Japan Phone: +81-564-59-5875, Fax: +81-564-59-5879, shige@nibb.ac.jp Atsushi Nakabachi Advanced Science Institute, RIKEN, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan Phone: +81-48-467-9332, Fax: +81-48-462-9329, bachi@riken.jp.

[†]Present address: Faculty of Fisheries, Nagasaki University, 1-14 Bunkyo-machi, Nagasaki-shi 852-8521, Japan

Keywords

full-length cDNA; aphid; functional genomics; EST analysis

Introduction

A large collection of full-length cDNAs is a powerful tool for accurate annotation of genomic sequences. Full-length cDNA clones carry the complete protein-coding sequences (CDSs) as well as 5'- and 3'-untranslated regions (UTRs), which dramatically improve the accuracy of gene predictions (Brent, 2008; Stanke et al., 2008). Full-length cDNAs are also useful for cataloging non-coding RNAs such as large intervening non-coding RNAs (lincRNAs) (Guttman et al., 2009), which are relatively difficult to identify by an alignment-based sequence search due to the reduced sequence conservation among species in comparison to protein-coding genes. In addition to these informatics aspects, full-length cDNA libraries also facilitate functional gene assays.

With the advent of the 464 Mb draft genome sequence released by the International Aphid Genomics Consortium (IAGC), the pea aphid, *Acyrtosiphon pisum*, is becoming a powerful genomic model for understanding insect-plant interactions, symbiosis, virus vectoring, and the evolution of complex life cycles and polyphenism (International Aphid Genomics Consortium, in review). Various genomic resources have been developed for the pea aphid with the goal of improving the genome annotation. For example, 40,904 expressed sequence tag (EST) sequences have already been obtained from different tissues and developmental stages (Sabater-Muñoz et al., 2006; Nakabachi et al., 2005), and cDNA microarrays have been fabricated from the EST clones (Wilson et al., 2006). Despite these efforts, the EST data currently available for the pea aphid are still insufficient in their coverage of the transcriptome. We present here a new collection of 50,599 pea aphid cDNA clones created from a normalized full-length cDNA library, which we characterized by generating 5'- and 3'-ESTs. We evaluate the impact of this library on the accuracy of gene model annotation.

Results and discussion

Construction of the pea aphid full-length cDNA library

We constructed a full-length cDNA library from whole bodies of parthenogenetic aphid females by using the CAP-trapper method (Table 1). This technology selectively captures the 5'-cap structure of mRNAs and dramatically enriches full-length cDNAs (Carninci et al., 1996). The CAP-trapper method is superior to other “full-length” methods such as oligo-capping and CapFinder (Clontech, now referred to as SMART) in terms of both the ability to clone long cDNAs and the percentage of full-length clones in the resulting libraries (Sugahara et al., 2001). This approach has another advantage in that it allows the removal of contaminant mRNA from bacterial symbionts (Nakabachi et al., 2005). To increase the gene discovery rate, the library was normalized. Examination of 96 randomly selected clones revealed that the DNA inserts ranged from 0.3 to 7.4 kb in size, with an average length of 2.1kb (Fig. 1).

EST sequencing

In total, 50,599 clones were sequenced from both ends, yielding 49,991 and 47,837 high-quality ESTs from the 5'- and 3'-ends, respectively, with an average length of 710 bp. Of the 96,828 ESTs, 95,861 (99%) were mapped in the Acyr 1.0 assembly of draft genome sequences of *Acyrtosiphon pisum*. Of the 46,229 clones having valid sequence data both from 5'- and 3'-ESTs, 40,841 clones (88%) showed "well-paired mapping", where both 5'- and 3'-ESTs mapping to the same scaffold with appropriate separation distance and opposite orientations (Table 2). This agreement of the ESTs with the genome assembly shows the quality of the cDNA library as well as the integrity of the genome assembly. Using BLASTN, we compared each EST sequence to the predicted transcripts of IAGC official gene models. The results of the genomic mapping and gene model assignments are summarized in Table 1-3; detailed data can be found on the IAGC Collaboration Wiki (Analysis file 1-4 at <https://dgc.cgb.indiana.edu/display/aphid/Full-length+cDNA+library>).

Comparison with 46,296 public pea aphid ESTs deposited in GenBank/EMBL/DDBJ prior to this project revealed that half of our new ESTs did not overlap with the old ESTs, indicating a high potential gene discovery rate. Indeed, among 8,437 genes identified in our full-length cDNA library (i.e., those that mapped to gene models, Table 3), 5,234 genes were not represented among the old ESTs. In addition, owing to the fact that the ESTs were generated from full-length cDNAs, the total length of the genomic coverage by all pea aphid ESTs increased by 13.0M bp up to 18.6M bp. We also identified novel genes that are not modeled in the official gene set (discussed below). In summary, the ESTs generated from the full-length cDNA library have remarkably extended the transcriptomic information available for the pea aphid.

Visual inspection of a subset of the EST mapping results with local GBrowse showed that in most cases our full-length cDNA ESTs cover CDS regions. A typical example is shown in Figure 2. The 14-3-3epsilon locus encompassed by 31 full-length cDNA clones showed that 87% (27/31) of 5'-ESTs begin upstream of the start codon and 100% of 3'-ESTs begin downstream of the stop codon. In contrast, most ESTs derived from non-full-length cDNA libraries reported prior to this study were mapped to the middle of the genes. The EST alignments of our cDNA library with the genome sequence also have consistent relationships with gene boundaries: the majority of the 5'-ESTs begin at about -400 bp from the start codon, while the 3'-EST start positions are distributed across several distinct preferential sites, indicating that the aphid 14-3-3epsilon gene has multiple alternative poly-A addition sites. To assess the "full-lengthness" of the library in a systematic way, we first determined what fraction of 5'-ESTs has a long open reading frame with no start codon, which gives us a rough estimation of the proportion of clones with partial 5'-ends. We searched all 5'-ESTs for open reading frames more than 300 bp with no start codons and found 814 (1.6%) incidents. Next, we compared our full-length cDNA ESTs with all of the pea aphid mRNA sequences deposited in GenBank that were annotated as containing complete coding sequences by manual curations. Among the 89 curated genes, 88 had corresponding clones in our EST library (1,400 clones). The comparison showed that 99.5% (1259 of 1275) and 99.9% (1102 of 1103) of the corresponding ESTs contained 5' UTR and 3' UTR, respectively, indicating that almost all of the clones contain complete coding

sequences. Notably, in most cases (5'UTR: 90.7%, 3'UTR: 80%), UTRs observed in our clones were significantly longer (>50bp) than those from the GenBank records. These results indicate that our library is highly enriched with cDNAs containing complete coding sequences along with more accurate UTR lengths.

We estimated the total number of genes represented in our full-length cDNA collection by two different approaches. First, we assembled 5'-ESTs and 3'-ESTs separately with CAP3 (Huang and Madan, 1999), resulting in 9,128 and 9,468 contigs, respectively. These totals are considered to be roughly equivalent to the number of represented genes, or a slight overestimate due to the alternative transcripts. Second, we compared our EST sequences with IAGC gene models (Acypi 1.0) and found that they matched 8,437 predicted genes (Table 3; see below for detail). With 248 novel genes that we identified (see below), the total number 8,635 should be close to the number of represented genes, or a slight underestimate due to gaps in the current genome assembly. Taken together, we estimate that our cDNA clone collection represents approximately 9,000 pea aphid genes.

Although only one-pass sequencing was performed from both ends for each clone, the paired sequences of 5'- and 3'-ESTs were sufficient to recover the complete insert sequence for 10,920 clones (Fig. S1). We termed these "full-insert sequences" (FISs) and mapped them onto the pea aphid genome to identify 3,040 unique loci. The longest clone was chosen as a representative for each locus and deposited in GenBank/EMBL/DBJ (accession numbers: AK339784-AK343184). Transcript data from FISs are considered to be stronger experimental evidence than that from ESTs; for this reason they should be used to update the official gene models in the public databases.

Estimation of gene length of the pea aphid from the full-length cDNA

Paired end sequences of clones from full-length cDNA libraries facilitate the detection of gene boundaries, because the start site of 5'-ESTs and 3'-ESTs mark transcription start sites and poly-A addition sites, respectively. Taking advantage of the "well-paired mapping" clones (Table 2), whose ESTs were clustered into 7,342 unique loci on the genome, we inferred the distribution of aphid gene length, which is defined as the span including exons and introns (Fig. 3). The median of gene length of *A. pisum* was 5.5 kb, while that of *Drosophila melanogaster* was 1.9 kb.

Contribution of full-length cDNAs to IAGC genome annotation

The collection of full-length cDNA ESTs played an important role in the automatic gene prediction and the manual annotation effort organized by the IAGC (IAGC, in review). cDNA evidence is most helpful for improving *de novo* gene finding (Stanke et al., 2008). Our ESTs were loaded into several gene prediction programs, such as NCBI Gnomon, Augustus and Maker, and contributed to the generation of the IAGC official gene set (IAGC, in review). Among the 10,249 RefSeq gene models, which are high quality evidence-based gene models presented by NCBI (Pruitt et al., 2009), 7,379 genes (72%) are supported by our full-length cDNA ESTs (78,623 ESTs). Our ESTs also support 1,058 non-RefSeq (*ab initio*) gene models (Table 3). In addition, the EST sequences were utilized in the manual annotation process. The EST and FIS sequences were imported to AphidBase,

where they could be browsed with GBrowse or Apollo to allow curators to evaluate and edit gene models [Legeai et al., in review, companion paper]. In particular, our ESTs were useful in determining gene boundaries. For example, XP_001949396.1 was initially predicted to be a single chimeric protein consisting of an unusual fusion of an angiotensin-converting peptidase with a homeodomain, but three 3'-ESTs were found between the sequences corresponding to the two domains, which revised the model to split into two genes (T. Murphy and J. Carolan, personal communication).

Further improvements of gene models

Although the full-length cDNA ESTs have already contributed to the construction of IAGC gene models, our ESTs have the potential to further improve these models. Here, we address five types of improvement: refinement of gene boundaries including annotation of UTRs, annotation of splicing variants, detection of non-coding genes, improvement of genome assembly and identification of novel genes.

Since the alignments of our ESTs with the pea aphid genome clearly delineate gene boundaries as shown above, they can be used to correct gene boundary errors, in which gene models are mistakenly merged or split. We found 247 loci (502 models) where our full-length cDNA sequences bridge two or more consecutive annotated genes. An example is shown in Figure 4A; these mergeable gene models are listed on the IAGC Collaboration Wiki (Analysis file 5 at <https://dgc.cgb.indiana.edu/display/aphid/Full-length+cDNA+library>). Conversely, we found 58 cases in which two non-overlapping full-length cDNAs were included in the same gene annotation, raising the possibility that the prediction had erroneously merged two genes (Fig. 4B, Analysis file 6 at IAGC Collaboration Wiki)

Ideally, both 5'- and 3'-ESTs for each clone should overlap with a single gene model, but there are a number of cases where one of them overlaps a gene model and the other is located outside of it. An example, 14-3-3epsilon, in which some 3'-ESTs are located outside the gene model, is shown in Figure 2. Of 40,841 clones examined, there were 6,809 clones (2298 loci) in which only the 5'-ESTs overlapped with the gene models, while the corresponding 3'-ESTs mapped outside of the gene models. Similarly, there were 565 clones (323 loci) in which only the 3'-ESTs overlapped with the gene models, while the 5'-ESTs mapped outside of the gene models. A future reconstruction of the pea aphid gene models should take pair-end mapping into consideration. A large part of these non-overlapping ESTs appear to correspond to UTRs, because they lack long open reading frames. This indicates that the UTR portions of many gene models may need to be extended.

In their current version, only 0.5% of the gene models are annotated with alternative transcripts. We evaluated the ability of our full-length cDNA resource to identify alternative splicing events, focusing on the 10,920 FIS clones. Out of 3,040 loci, 218 (7.1%) exhibited multiple alternative splicings. Examples are shown in Fig. S2.

We queried all our ESTs against non-coding RNA sequences in the Rfam database using BLASTN. We identified three precursor transcripts for the microRNA *miR-iab-4* (316K23:FF299755; 378E1:FF305321,FF307802; 536A13: FF334324), which is a highly conserved microRNA encoded in the Hox cluster (Shigenobu et al., in press). A

comprehensive survey of microRNA in the pea aphid genome is reported separately by Legeai et al (Legeai et al, in review).

Because of the incomplete nature of the draft genome sequence (IAGC in review), the current genomic assembly contains so many gaps that the gene models built from this sequence data inherit the assembly problems. Our full-length cDNA ESTs are useful to detect such problems, to correct the gene model errors derived from the assembly problems and even to improve the genome assembly. An example is shown in Figure S3, where an EST detected the erroneous a sequencing gap in the assembly and also includes sequence for an exon that likely falls in the other gap. We infer a substantial population of the full-length cDNA ESTs is in the similar situation from our observation that 7% of the ESTs did not completely align (< 90% of their entire length) to the best-hit scaffold sequence.

The full-length cDNA resource was also used to detect novel pea aphid genes that were overlooked during previous annotations. We detected 248 genomic loci mapped by our cDNAs, where neither RefSeq nor *ab initio* gene models had previously been predicted (Table S1). It remains to be elucidated whether these are protein-coding genes or non-coding genes; however, 31 of them contained CDSs longer than 300 bp and some showed similarities to proteins of other species. For example, NV135 and NV3 appear to be homologs of beta-1,4-galactosyltransferase and translocon-associated complex TRAP, respectively.

Future functional assays

For gene discovery and gene modeling of sequenced species, new transcriptomics technologies such as RNA-seq and tiling microarrays are replacing EST analysis, because these new technologies have advantages in cost and time efficiency over conventional EST analysis (Wang et al., 2009). However, large-scale collection of isolated cDNA clones, which can be obtained through EST projects but not by RNA-seq or tiling microarray experiments, still have great benefits, because these cDNA clones can be utilized for a variety of functional assays. One instant benefit is that we can access full-length clones, omitting laborious cloning procedures such as repetitive library screenings or rapid amplification of cDNA ends (RACE), which need to follow analyses of conventional partial-length cDNAs. Large-scale cDNA collections also enable “-omics” approaches to elucidate gene function, including large-scale *in situ* hybridization, yeast two-hybrid analysis and RNAi screening. Thus, this 50K full-length cDNA collection represents an important community resource for understanding the functional genomics of the pea aphid.

Experimental Procedures

Construction of a normalized full-length cDNA library

Total RNA was extracted from whole bodies of nymphs and adult winged and wingless parthenogenetic females of LSR1, the pea aphid strain that was used for genome sequencing. After isolating mRNA, a normalized full-length cDNA library was constructed by using the CAP-trapper method (Carninci and Hayashizaki, 1999; Carninci et al., 1996) at DNAFORM (Yokohama, Japan). The oligo(dT) primer for the first-strand cDNA synthesis was: 5’-

GAGAGAGAGAAGGATCCAAACGTGCTTTTTTTTTTTTTTTTTTTVN -3'. Double-stranded linkers used for the second-strand cDNA synthesis were prepared with the GN5 linker and N6 linker (molar ratio of N6:GN5=1:4) (Shibata et al., 2001). Normalization was performed using the hybridization method (Carninci et al., 2000). Second-strand cDNA was digested with Bam HI and XhoI, and ligated to a lambda FLC-III vector, which carries two loxP sites (Carninci et al., 2001). After amplification in C600 cells, the phage DNA was converted into plasmids with Cre recombinase. The plasmid library was electroporated into DH10B cells.

Sequence data sets

The 1.0 release of the *Acyrtosiphon pisum* genome assembly (EQ110872 – EQ133570) was used as the basis for bioinformatic analysis. The NCBI Gnomon gene model set, version 1, was first used as a reference gene set. The results were checked with GLEAN consensus gene models (Acypi 1.0), which was released by IAGC as an official gene model at a late stage of the project. To estimate the proportion of clones that contained complete CDSs and UTRs, we used all pea aphid mRNA sequences in GenBank which had been manually annotated as containing complete CDSs. The 89 sequences used are equivalent to ACYPI000001 – ACYPI000097. FlyBase Release 5.4 provided the genome sequence and gene models for *Drosophila melanogaster* used in this study. The pea aphid ESTs reported before this study were compiled from the NCBI UniGene repository as of November 26, 2007.

EST sequencing

DNA was isolated using a standard alkaline lysis procedure in an automated 384 well format. cDNA clones were end sequenced from both the 5' and 3' ends using 1/64th dilution AB Big Dye terminator chemistry. Reactions were run on ABI 3730 capillary sequence machines (Applied Biosystems, Foster city CA) using the 36 run module. Reads were vector trimmed, screened for bacterial contamination and sequence quality by a custom Perl script. Reads with greater than 100 bp of contiguous high quality (>Q20) sequence were submitted to dbEST (NCBI). The accession numbers are EX601480 – EX654440 and FF291997 – FF339412. The chromatograms of these sequences are also deposited at NCBI Trace Archive.

Mapping ESTs and cDNAs to the genome

ESTs and FISs were softmasked using RepeatMasker and then mapped to the genome using Exonerate 2.0.0 (Slater and Birney, 2005), using the est2genome model and a custom DNA substitution matrix (match: +5, mismatch: -6). Other parameters were as follows: score threshold = 300, DNA HSP threshold score = 140, gap open penalty = -12 and gap extend penalty = -4. These genomic alignments and reference gene models were visualized by GBrowse (Stein et al., 2002). We configured the color and the glyph of the GBrowse track to facilitate recognition of EST pairs on the screen.

Clustering of ESTs

Sequence-based clustering was carried out with CAP3 (Huang and Madan, 1999), using the following parameters: overlap length cutoff = 40bp, overlap identity cutoff = 94% and maximal overhang percent length = 25. Location-based clustering was carried out using the Exonerate genomic mapping data with a custom Ruby script which facilitated scanning and grouping of overlapping exons among ESTs.

Full-insert sequence generation and analysis

For each clone, the sequences of the 5'- and 3'-EST pair were assembled by CAP3, considering base call quality (phred score) and orientation consistency. The resultant FIS sequences were aligned to the pea aphid genome and then subjected to location-based clustering as described above resulting in 3,040 groups. The longest sequences were chosen as representative and submitted to GenBank/EMBL/DDBJ.

To identify alternative splicing events, for each group, the member FISs were further divided into subgroups by sequence-based CAP3 clustering. We generated a virtual cDNA sequence for this purpose using the matching pea aphid genomic sequence, because the FIS sequences are based on the assembly of single-pass reads and may contain sequencing errors. Note that this procedure does not distinguish the alternative transcription start sites or alternative polyA addition sites, and it may miss small differences between alternative transcripts.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

S.S. and A.N. thank Prof. Nancy A. Moran and the late Prof. Hajime Ishikawa for their supports for the full-length cDNA library construction. S.S. thanks Dr. Makoto Suzuki (DNAFORM Inc.) for the characterization of the cDNA library. S. S. also thanks T. Murphy (NCBI) for the careful curation of our ESTs and the helpful comments. This work was supported in part by Research Fellowship of the Japan Society for the Promotion of Science for Young Scientists to A.N..

References

- Brent MR. Steady progress and recent breakthroughs in the accuracy of automated genome annotation. *Nat Rev Genet.* 2008; 9:62–73. [PubMed: 18087260]
- Carninci P, Hayashizaki Y. High-efficiency full-length cDNA cloning. *Meth Enzymol.* 1999; 303:19–44. [PubMed: 10349636]
- Carninci P, Kvam C, Kitamura A, Ohsumi T, Okazaki Y, Itoh M, Kamiya M, Shibata K, Sasaki N, Izawa M, Muramatsu M, Hayashizaki Y, Schneider C. High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics.* 1996; 37:327–336. [PubMed: 8938445]
- Carninci P, Shibata Y, Hayatsu N, Itoh M, Shiraki T, Hirozane T, Watahiki A, Shibata K, Konno H, Muramatsu M, Hayashizaki Y. Balanced-size and long-size cloning of full-length, cap-trapped cDNAs into vectors of the novel lambda-FLC family allows enhanced gene discovery rate and functional analysis. *Genomics.* 2001; 77:79–90. [PubMed: 11543636]
- Carninci P, Shibata Y, Hayatsu N, Sugahara Y, Shibata K, Itoh M, Konno H, Okazaki Y, Muramatsu M, Hayashizaki Y. Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-

length cDNA libraries for rapid discovery of new genes. *Genome Res.* 2000; 10:1617–1630. [PubMed: 11042159]

Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, Cabili MN, Jaenisch R, Mikkelsen TS, Jacks T, Hacohen N, Bernstein BE, Kellis M, Regev A, Rinn JL, Lander ES. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature.* 2009; 458:223–227. [PubMed: 19182780]

Huang X, Madan A. CAP3: A DNA sequence assembly program. *Genome Res.* 1999; 9:868–877. [PubMed: 10508846]

International Aphid Genomics Consortium (IAGC). Genome sequence of the pea aphid *Acyrtosiphon pisum*. *PLoS Biology.* in review.

Legeai F, Shigenobu S, Gauthier J, Colbourne J, Rispe C, Collin O, Richards R, Wilson A, Tagu D. AphidBase: A centralized bioinformatic resource for annotation of the pea aphid genome. *Insect Mol Biol.* in review.

Legeai F, Rizk G, Walsh T, Edwards O, Gordon K, Lavenier D, Leterme N, Méreau A, Nicolas J, Tagu D, Jaubert-Possamai I. S. Bioinformatic prediction, deep sequencing of microRNAs and their role in phenotypic plasticity in the pea aphid, *Acyrtosiphon pisum*. *Insect Mol Biol.* in review.

Nakabachi A, Shigenobu S, Sakazume N, Shiraki T, Hayashizaki Y, Carninci P, Ishikawa H, Kudo T, Fukatsu T. Transcriptome analysis of the aphid bacteriocyte, the symbiotic host cell that harbors an endocellular mutualistic bacterium, *Buchnera*. *Proc Natl Acad Sci USA.* 2005; 102:5477–5482. [PubMed: 15800043]

Pruitt KD, Tatusova T, Klimke W, Maglott DR. NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res.* 2009; 37:D32–36. [PubMed: 18927115]

Sabater-Muñoz B, Legeai F, Rispe C, Bonhomme J, Dearden PK, Dossat C, Duclert A, Gauthier JP, Ducray DG, Hunter W, Dang P, Kambhampati S, Martinez-Torres D, Cortes T, Moya A, Nakabachi A, Philippe C, Prunier-Leterme N, Rahbé Y, Simon JC, Stern DL, Wincker P, Tagu D. Large-scale gene discovery in the pea aphid *Acyrtosiphon pisum* (Hemiptera). *Genome Biol.* 2006; 7:R21. [PubMed: 16542494]

Shibata Y, Carninci P, Watahiki A, Shiraki T, Konno H, Muramatsu M, Hayashizaki Y. Cloning full-length, cap-trapper-selected cDNAs by using the single-strand linker ligation method. *BioTechniques.* 2001; 30:1250–1254. [PubMed: 11414214]

Shigenobu, et al. Comprehensive survey of developmental genes in the pea aphid, *Acyrtosiphon pisum*: frequent lineage-specific duplications and losses of developmental genes. *Insect Mol Biol.* in review.

Slater GS, Birney E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics.* 2005; 6:31. [PubMed: 15713233]

Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics.* 2008; 24:637–644. [PubMed: 18218656]

Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S. The generic genome browser: a building block for a model organism system database. *Genome Res.* 2002; 12:1599–1610. [PubMed: 12368253]

Sugahara Y, Carninci P, Itoh M, Shibata K, Konno H, Endo T, Muramatsu M, Hayashizaki Y. Comparative evaluation of 5'-end-sequence quality of clones in CAP trapper and other full-length-cDNA libraries. *Gene.* 2001; 263(1-2):93–102. [PubMed: 11223247]

Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet.* 2009; 10:57–63. [PubMed: 19015660]

Wilson AC, Dunbar HE, Davis GK, Hunter WB, Stern DL, Moran NA. A dual-genome microarray for the pea aphid, *Acyrtosiphon pisum*, and its obligate bacterial symbiont, *Buchnera aphidicola*. *BMC Genomics.* 2006; 7:50. [PubMed: 16536873]

Figure 1. Size distribution of isolated cDNA inserts

Insert size of 96 randomly-selected clones were examined by colony PCR followed by agarose gel electrophoresis. Average insert size was 2.1 kb, with a range of 0.3 to 7.4kb.

Figure 2. Graphical representation of full-length cDNA ESTs on the genome sequence
ESTs from the full-length cDNA library were aligned to the 14-3-3epsilon locus on the genomic scaffold EQ122825. For each clone, the members of the EST pair are linked with a dotted line if they do not overlap. 5'-ESTs and 3'-ESTs are colored dark blue and light blue, respectively. For comparison, RefSeq gene model (CDS: brown, UTR: gray), ESTs of non-full-length cDNA (steel blue), EST contigs generated by sequence-based assembling (green) and *Drosophila melanogaster* orthologs (orange) are shown. Sequence orientations are indicated by arrows.

Figure 3. Distribution of gene lengths for *A. pisum* and *D. melanogaster*

Gene length is defined as the span including exons and introns. The gene length distribution for *A. pisum* was calculated from 7,432 loci for which one or more EST pairs from full-length cDNA clones were available. For *D. melanogaster*, all transcripts annotated in FlyBase (Release 5.4) were used for the statistics.

Figure 4. Examples of erroneous gene models detected by the full-length cDNA ESTs
 (A) illustrates the locus of PTEN (tumor suppressor gene) as an example in which the gene model was mistakenly split into two different models. The 5'- and 3'-members of the EST pair bridge these mistakenly split gene models. (B) illustrates XM_001952817, a case in which the gene model mistakenly contains two different genes. The ESTs clearly indicate a gene boundary between the 5th and 6th exons. Color coding of gene models and ESTs is the same as Figure 2.

Table 1

Characterization of the full-length cDNA library and derived ESTs

Library	
RNA source	<i>Acyrtosiphon pisum</i> , LSR1 strain whole body, parthenogenetic female
Construction method	CAP-trapper, normalized
Average insert length	2.1 kb (n = 96)
Cloning vector	pFLC-III
Sequenced clones	50,599
ESTs	
5'-EST	49,991
3'-EST	47,837
Average high quality read length	710 bp

Table 2

Summary of genomic mapping

Mapping type	#clones	Definition
Clones with both 5'-EST and 3'-EST	(46,229)	Valid sequences were obtained from both 5'- and 3'-ends
_Well-paired mapping	40,841	Both 5'-EST and 3'-EST were mapped on the same scaffold with appropriate separation distance * and opposite orientations.
_Neither mapped	63	Neither 5'-EST nor 3'-EST were mapped.
_Either mapped	705	Either 5'-EST or 3'-EST was mapped.
_Different scaffold	4,536	5'-EST and 3'-EST were mapped onto different scaffolds.
Problematic	84	Both ESTs were mapped on the same scaffold, but there was a problem in the orientation consistency or separation length.
Clones with either 5'-EST or 3'-EST	(4,370)	A valid sequence was obtained from only the 5'- or 3'-EST
_Mapped	4,234	
_Unmapped	136	

* Separation distance < 250kb were passed. EST pairs with distance more than 100kb were double-checked manually using GBrowse.

Table 3Comparison with *A. pisum* gene models

Hit category	#genes	#ESTs	
		5'-ESTs	3'-ESTs
RefSeq	7,390	44,161	34,462
<i>ab initio</i> *	1,058	2,741	1,975
no hit to models		3,089	11,400
_potential connection with model **	NA	984	8,372
novel gene ***	248+	432+	432+
_unmapped ****	NA	200	767

* *ab initio*: non-Refseq NCBI Gnomon gene models

** This category represents the EST that does not match any gene models, but the other end of the clone does match. Such “no hit” EST should connect with the gene models that the counterpart overlaps. The matching failure of one of the pair may be due to the incomplete gene model.

*** Novel genes predicted in this study (Table S1) are counted. Note that this prediction was restricted to the clones with “well-paired mapping” ESTs and there must be more genes that are not undocumented in the reference gene models.

**** ESTs that do not map to the genome.