The transcriptome of *Anguillicoloides crassus*. Part A: Pilot-sequencing

Emanuel G Heitlinger

Abstract

In preparation of high-throughput transcriptome sequencing of the swim-bladder nematode *Anguillicoloides crassus* expressed sequence tags (ESTs) were generated using traditional Sanger-technology. In total 944 reads from adult *A. crassus* (5 libraries from 4 cDNA preparations) and 288 reads from liver-tissue of the host species *Anguilla japonica* (3 libraries from 3 cDNA preparations) were sequenced. 125 of the nematode and 35 of the host reads were of sufficient quality for further processing. The low number of high quality ESTs can be explained by a high degree of rRNA contamination present in all libraries. Additionally 10 sequences originating from host-contamination had to be removed from the *A. crassus* data-set for submission to NCBI-dbEST, reducing it to 115 ESTs. Nevertheless the stringent quality trimming and processing of raw reads, as summarized in the present document, make the remaining ESTs a valuable resource for comparison with future 454-sequencing-data.

Introduction

After sampling adult and larval *A. crassus* from the wild in Taiwan and Germany the generation of cDNA libraries from small amounts of starting material was made difficult to achieve for two reasons:

- The larvae are of small size and only available in limited numbers from the wild.
- Adult worms consist predominantly of ingested host blood.

Therefore an amplification protocol for cDNA had to be used. As low amounts of starting material and the use of these protocols could produce unwanted contamination by amplification artifacts obtained cDNA libraries were analyzed using traditional Sanger-sequencing.

Material and methods

During sampling in Taiwan and Germany, single adult *A. crassus* were preserved in RNAlater(Quiagen, Hilden, Germany), after their sex had been determined. Total RNA was extracted from single, whole worms using the RNeasy kit (Quiagen, Hilden, Germany), following the manufacturers protocol. Alternatively parts of the liver of the host species *Anguilla japonica*, which also had been preserved in RNAlater were used for RNA extraction, following the same protocol.

The Evrogen MINT cDNA synthesis kit (Evrogen, Moscow, Russia) was then used to amplify mRNA transcripts according to the manufacturers protocol. It uses an adapter sequence at 3' the end of a a poly dT-primer for first strand synthesis and adds a second adapter complementary to the bases at the 5' end of the transcripts by terminal transferase activity and template switching. Using these adapters it is possible to specifically amplify mRNA enriched for full-length transcripts. The obtained cDNA preparations were undirectionally cloned into TOPO2PCR-vectors (Invitrogen, Carlsbad, USA) and TOP10 chemically competent cells (Invitrogen, Carlsbad, USA) were transformed with this construct. The cells were plated on LB-medium-agarose containing Kanamycin (5mg/ml), xGal (5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside) and IPTG (Isopropyl- β-D-1-thiogalactopyranosid). After 24h of incubation at 36 °C cells were picked into 96-well micro-liter-plates containing liquid LB-medium and Kanamycin (5mg/ml) and incubated for another 24h. Subsequently 2ml of the cells were used as template for amplification of the insert by PCR using the primers

Forward M13F(GTAAAACGACGGCCAGT) and

Reverse M13R(GGCAGGAAACAGCTATGACC)

in a concentration of 10μM. The protocol for PCR cycling is shown in table 1.

Inital denaturation	94 °C	5min	
Denaturation	94 °C	30s	
Annealing	54 °C	45s	35 cycles
Elongation	72 °C	2min	
Filnal Elongation	72 °C	10min	

Table 1: PCR protocol for insert amplification

Amplification products were controlled on gel and cleaned using SAP (Shrimp Alkaline Phosphatase) and ExoI (Exonuclease I). Sequencing reactions were performed using the BigDye-Terminator kit and PCR-primers

(forward or reverse) in a concentration of 3.5μM and sequenced on an ABI 3730 DNA Analyzer (Applied Biosystems, Foster City, California, USA). For *A. crassus* the following libraries were prepared:

Ac_197F: Female from Taiwanese aquaculture
Ac_106F: Female from Taiwanese aquaculture
Ac_M175: Male from Taiwanese aquaculture
Ac_FM: Female from Taiwanese aquaculture
Ac_EH1: Same cDNA preparation as Ac_FM, but sequenced by students in a practical

For *Anguilla japomica* the following three libraries:

Aj_Li1: liver of an eel from aquaculture
Aj_Li2: liver of an eel from aquaculture
Aj_Li3: liver of an eel from aquaculture

The original sequencing-chromatographs ("trace-files") were renamed according to the NERC environmental genomics scheme. "Ac" was used as project-identifier for *Anguillicoloides crassus*, "Aj" for *Anguilla japonica*. In *Anguillicoloides* sequences information on the sequencing primer (forward or reverse PCR primer; *Anguilla japonica* sequences were all sequenced using the forward PCR primer) was temporarily stored in the middle "library"-field, resulting in names of the following form:

```
Ac_[\d|\w]{2,4}(f|r)_\d\d\w\d\d
Aj_[\d|\w]{2,4}_\d\d\w\d\d
```

The last field indicates the plate number (two digits), the row (one letter) and the column (two digits) of the corresponding clone. For first quality trimming trace2seq, a tool derived from trace2dbEST (both part of PartiGene [1]) was used, briefly it performs quality trimming using phred[2] and trimming of vector sequences using cross-match[3]. The adapters used by the MINT kit were trimmed by supplying them in the vector-file used for trimming along with the TOPO2PCR-vector. After processing with trace2seq additional quality trimming was performed on the produced sequence-files using a custom script. This trimming was intended to remove artificial sequences produced when the sequencing reaction starts at the 3' end of the transcript at the poly-A tail. These sequences typically consist of numerous homo-polymer-runs throughout their length caused by "slippage" of the reaction. The basic perl regular expression used for this was:

```
/(.*A{5,}|T{5,}|G{5,}|C{5,}.*){$lengthfac,}/g
```

Where `$lengthfac` was set to the length of the sequence devided by 70 and

rounded to the next integer. So only one homo-polymer-run of more then 5 bases was allowed per 105 bases. Results of this screening were checked by blasting the sequences excluded as artificial against nempep4, a nematode rRNA database and a fish-protein database. Two sequences which were identified as false positives (hitting proteins in nempep4) were moved manually to the sequences still categorized as good. These were screened against *Anguillicoloides* rRNA or fish rRNA using cross-match[3] with standard parameters for screening. Finally GS content was tabulated for the sequences intended for submission and screening statistics were calculated.

After this step sequences were screened for host contamination by a comparison of BLAST searches against nempep4 and a fish protein database. Sequences producing better bit scores againt fish proteins than nematode proteins were removed.

Only the trace-files corresponding to the sequences still regarded as good after this step were processed with trace2dbEST. Additionally to the processing of traces already included in trace2seq sequences were preliminary annotated using BLAST versus the NCBI-NR non-redundant protein database and a EST-submission-file was produced. This file parsed for the information on the sequencing primer (stored in the library-field) and the corresponding primer-entries in the file were replaced.

Further analysis and plotting was carried out using R[4] and included in a LaTeX document using the R-package Sweave[5].

Results

Initial quality screening (see table 2).

The initial quality screening of *Anguillicoloides*-sequences revealed a high number of sequences that had to be discarded due to failed sequencing reactions (sequences beeing too short after quality trimming by trace2seq) in the libraries prepared by students. For sequences of *Anguilla japonica* and the other libraries from *A. crassus* failed sequencing reactions were less common.

In the next screening-step for *A. crassus* 123 (13.03%) and for *Anguilla japonica* 68 (23.61%) of the sequences were excluded because of homopolymer-runs considered as artificial.

rRNA screening (see also table 2)

The further screening of *Anguillicoloides* sequences revealed a high amount of rRNA contamination(see fig.1) ranging from 62.69% to 77.42%. One rea-

	short	poly	rRNA	good*
Aj_Li2(n=96)	2	28	49	17
Aj_Li3(n=96)	7	15	68	6
Aj_Li1(n=96)	2	25	57	12
Aj_total(n=288)	11	68	174	35
Ac_197F(n=96)	4	17	58	17
Ac_106F(n=96)	25	9	48	14
Ac_FM(n=96)	12	29	35	20
Ac_M175(n=116)	30	19	42	25
Ac_EH1(n=540)	297	49	145	49
Ac_total(n=944)	368	123	328	125

Table 2: Total numbers for screening statistics of A. crassus. *before screening for host contamination

son to sequence the libraries from the eels host was to elucidate whether this contamination was nematode or species-typical (e.g caused by poly-dT primers binding to A-rich rRNA regions), or caused by shortcomings in the preparation.

Even higher amounts of rRNA were found in these host-libraries (see fig. 1), ranging from 74.24% to 91.89% . This contamination in libraries from both species was mainly responsible for a low ammount of sequences beeing of sufficient quality for submission to NCBI-dbEST.

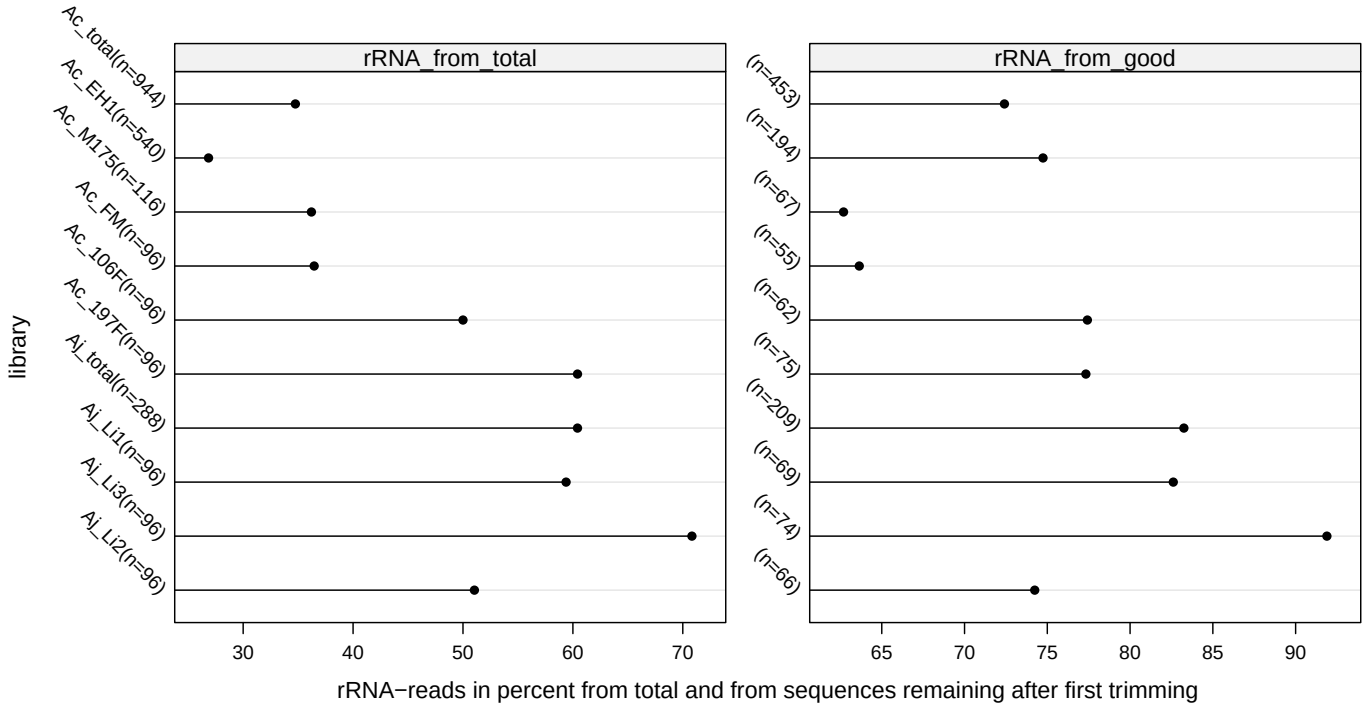

Figure 1: Proportion of rRNA contamination in different libraries for *A. crassus*.

Screening for host-contamination)

The GC-content of *A. crassus* ESTs gave first pointers on possible contamination still left in the sequences after these steps. *A. crassus* had a lower mean GC-content ($p<0.001$) (38.49 ± 9.35; mean $\pm$ sd) than *Anguilla japonica* (45.99 ± 8; mean $\pm$ sd). The probability density function of GC-contents ~~pointed on~~ some unusual sequences containing as high as 65% GC (see fig. 2) in the *A. crassus* set. Further inspection of preliminary annotations from BLAST searches against NR, showed that the first sequence with a annotation relating it to a nematode protein was Ac_FMf_08A08 (the sequence with the 8th highest GC-content), for which similarity to "collagen col-34" from *Brugia malayi* was inferred.

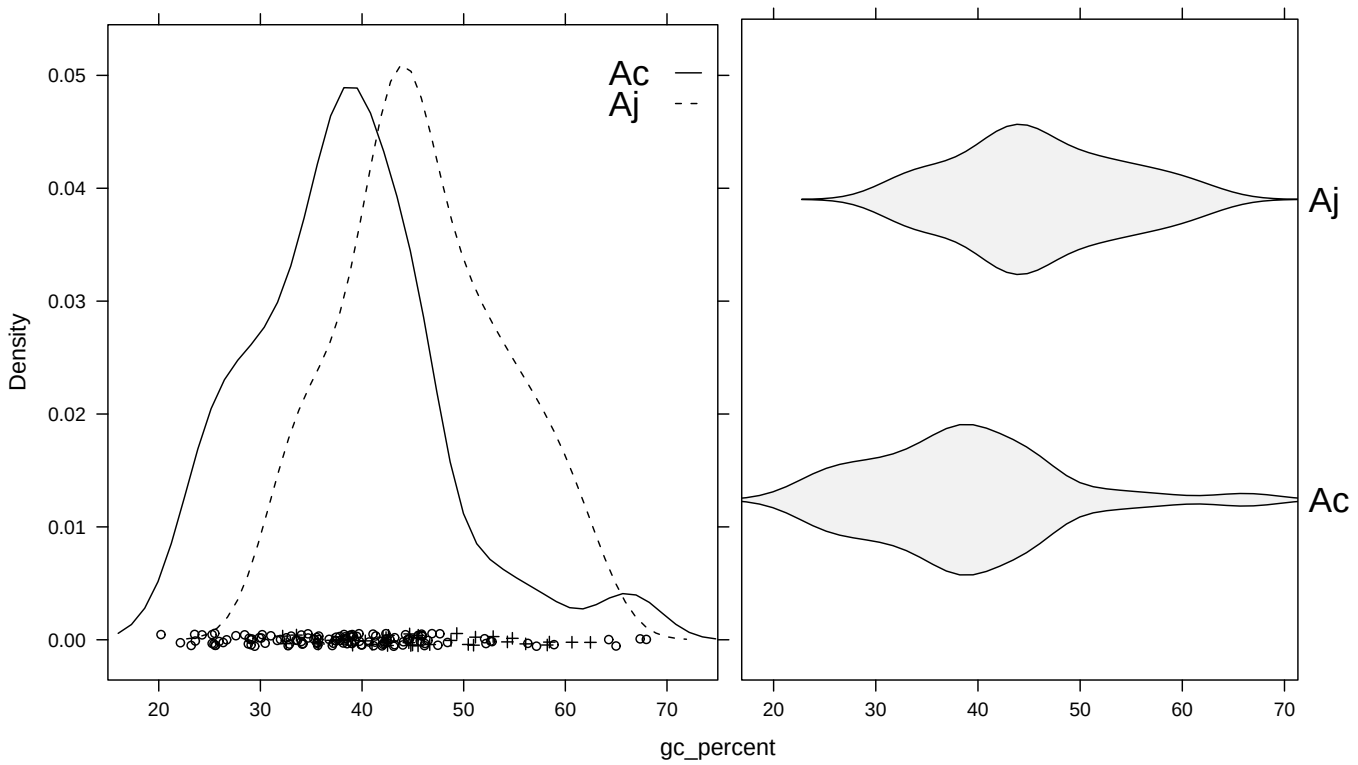

Figure 2: GC-content of sequences from *Anguilla japonica* and *A.crassus*.

Manual inspection of the remaining annotations made clear that another filtering for host sequences was necessary (3 sequences had best BLAST hits to proteins from non-teleost vertebrates, 7 from teleosts). A comparison of BLAST results for these sequences versus nempep4 and a fishprotein database ~~(derived from NCBI non-redundant),~~ showed that they were more likly to originate from host contamination than from *A. crassus* (see table 3). They were excluded from the submission file, reducing the number of sequences for submission to 113.

Sequences (forward and reverse) from Ac_EH1_01D10 which possibly are derieved from bacterial ~~contaminarion~~ were left in the submission-file, because it could not be fully excluded that they originated from symbiotic bacteria.

Anoter sequence for ~~wich~~ a xenobiotic origin seems ~~pessible~~ is Ac_FM_08D01 annotated as "putative senescence-associated protein" from the plant *Pisum sativum*.

EST	fishpep annotation	evalue
Ac_106F_01A06	PREDICTED: similar to FAT tumor suppressor homolog 4 [Danio rerio]	1.00E-017
Ac_EH1_01C10	Ferritin, middle subunit [Salmo salar]	2.00E-088
Ac_EH1_009C03	hypothetical protein LOC567037 [Danio rerio]	7.00E-037
Ac_EH1_01A02	muscle-specific beta 1 integrin binding protein 2 [Epinephelus coioides]	2.00E-091
Ac_EH1_01A07	RecName: Full=Hemoglobin anodic subunit beta; AltName: Full=Hemoglobin anodic beta chain	2.00E-076
Ac_M175_01B06	lrrc15 [Danio rerio]	2.00E-016
Ac_FM_08F03	PREDICTED: similar to stromal antigen 2 isoform 1 [Danio rerio]	2.00E-057
Ac_M175_01H02	Peroxiredoxin-1 [Salmo salar]	1.00E-101
Ac_EH1_005B07	cyclin G1 [Poecilia reticulata]	9.00E-019
Ac_EH1_005B07	Cyclin G1 [Danio rerio]	1.00E-011

Table 3: Sequences excluded because of inferred "host-like" annotation

Discussion

A total of 944 sequencing-reads from *A. crassus* (4 different libraries) and 288 reads from liver tissue (3 libraries) of the host species *Anguilla japonica* resulted in 115 nematode- and 35 host quality ESTs submittable to NCBI dbEST. The low proportion of high quality ESTs is mainly a result of a high degree of rRNA contamination present in all libraries and of the fact that 40% of the sequencing rections for *Anguillicoloides* were prepared by unexperienced undergraduate students in the framework of a practical.

The high proportion of rRNA in the host-ESTs clearly shows, that shortcomings in the preparation of cDNA rather than sequence-specific difficulties are responsible for this contamination.

A minor fraction of reads had to be excluded because of sequencing-reactions starting at a poly-A tail producing artificial homopolymer-runs. The basic regular expression used for this purpose produced good results identifying only two sequences, which had a hit to protein annotated by blast similarity within nempep4 wrongly (false positives). Another ~~evitable~~ false positive had a hit to a nempep4 protein transcribed by ESTscan[6] within ~~PArtigene.~~ Visual inspection of this sequence led to the conclusion,

~~that~~ not the classification as artificial inferred here is likely to be wrong, but rather nembase4 contains an artificial low-complexity sequence leading to a blast-hit. Nevertheless the method for filtering of these seqences could still be improved further by designing a more ~~sofisticated~~ algorithm, that e.g. assigns different penalty-scores to homopolymer-runs of different length and excludes sequences of a certain penalty-score per length.

After exclusion of low quality and rRNA sequences the remaining sequences from *A. crassus* had to be examined for host contamination.The present analysis showed that examination of GC-content is a powerful tool for this purpose in *A. crassus* transcriptome sequences.The 7 sequences with the highest GC-contents contained six sequences that had to be removed from the submission file because of an annotation making a host origin likely. In total 10 sequences had to be excluded.

This amount of host-contamination is rather high compared to other nematode EST studies. Nevertheless given the large proportion of host blood ingested by *A. crassus* compared to the small nematode itself, such a contamination is not too suprising.

References

[1] Parkinson J, Anthony A, Wasmuth J, Schmid R, Hedley A, Blaxter M: **PartiGene–constructing partial genomes**. *Bioinformatics* 2004, **20**(9):1398–1404.

[2] Ewing B, Hillier L, Wendl MC, Green P: **Base-Calling of automated sequencer traces using Phred. I. Accuracy Assessment**. *Genome Res.* 1998, **8**(3):175–185, [[http://genome.cshlp.org/cgi/content/abstract/8/3/175]].

[3] Green P: *PHRAP documentation.* 1994, [[http://www.phrap.org]].

[4] R Development Core Team: *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria 2009, [[http://www.R-project.org]].

[5] Leisch F: **Sweave: Dynamic Generation of Statistical Reports Using Literate Data Analysis**. In *Compstat 2002 — Proceedings in Computational Statistics*. Edited by Härdle W, Rönz B, Physica Verlag, Heidelberg 2002:575–580, [[http://www.stat.uni-muenchen.de/ leisch/Sweave]]. [ISBN 3-7908-1517-9].

[6] Iseli C, Jongeneel C, Bucher P: **ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences.** *Proc Int Conf Intell Syst Mol Biol* 1999, :138–148.