The transcriptome of the swimbladder-nematode Anguillicola crassus: Resources for an alien parasite

Emanuel G Heitlinger*[1,2] Stephen Bridgett[3] Anna Montazam[3] Horst Taraschewski[1] and Mark Blaxter[2]

[1]Department of Ecology and Parasitology, Zoological Institute 1, University of Karlsruhe,Kornblumenstrasse 13, Karlsruhe, Germany [2]Institute of Evolutionary Biology, The Ashworth laboratories, The University of Edinburgh, King's Buildings Campus, Edinburgh, UK [3]The GenePool Sequencing Service, The Ashworth laboratories, The University of Edinburgh, King's Buildings Campus, Edinburgh, UK

Email: Emanuel G Heitlinger*- emanuelheitlinger@gmail.com; Stephen Bridgett- sbridget@staffmail.ed.ac.uk; Anna Montazam- Anna.Montazam@ed.ac.uk; Horst Taraschewski- dc20@rz.uni-karlsruhe.de; Mark Blaxter - mark.blaxter@ed.ac.uk;

*Corresponding author

Abstract

Background:

Results:

Conclusions: Yeh!

Background

The nematode Anguillicola crassus Kuwahara, Niimi et Itagaki, 1974 [1] is a parasite of freshwater eels of the genus Anguilla, and adults localise to the swim bladder where they feed on blood. Larvae are transmitted via crustacean intermediate hosts. Originally endemic to East-Asian populations of the Japanese eel (Anguilla japonica), A. crassus has attracted interest due to recent anthropogenic expansion of its geographic and host ranges to Europe and the European eel (Anguilla anguilla). Recorded for the first time in 1982 in North-West Germany [2], where it was most likely introduced through the live-eel trade [3, 4], A. crassus has spread rapidly through populations of its newly acquired host [5]. At the present

day it is found in all An. anguilla populations except those in Iceland [6]. A. crassus can be regarded as a model for invasive parasite introduction and spread [9].

A. crassus has a major impact on An. anguilla populations. In its natural host in Asia infection prevalence and mean intensity of infection are lower than in Europe [10], where high prevalence (above 70% [7]) and high infection intesities have been reported throughout the newly colonized area [8]. The virulence of A .crassus in this new host has been attributed to an inadequate immune response in An. anguilla [11]. While the An. japonica is capable of killing larvae of the parasite after vaccination [13] or under high infection pressure [14], responses in An. anguilla have hallmarks of pathology, including thickening of the swim bladder wall [15]. Interestingly host also affects the adult size and life-history of the nematodes: In European eels the nematodes are bigger and develop and reproduce faster [12].

The genus Anguillicola is placed in the nematode suborder Spirurina (clade III sensu [16]) [17, 18]. The Spirurina are exclusively parasitic and include important human pathogens (the causative agents of filariases and ascariasis) as well as prominent veterinary parasites. Molecular phylogenetic analyses place Anguillicola in a clade of spirurine nematodes (Spirurina B of [Laetsch et al submitted]) that have an freshwater or marine intermediate host, but infect a wide range of carnivorous definitive hosts. Spirurina B is sister to the main Spirurina C, including the agents of filariases and ascariasis), and thus A. crassus may be used as an outgroup taxon to understand the evolution of parasitic phenotypes in these species.

Recent advances in sequencing technology (often termed Next Generation Sequencing; NGS), provide the opportunity for rapid and cost-effective generation of genome-scale data. The Roche 454 platform [20] offers longer reads than other NGS technologies, and thus is suited to de novo assembly of genome-scale data in previously understudied species. Roche 454 data has particular application in transcriptomics [21].

The difference in the biology of A. crassus in An. japonica (coevolved) and An. anguilla (recently captured) eel hosts likely results from an interaction between different host and parasite responses, underpinned by definitive differences in host genetics, and possible genetic differentiation between the invading European and endemic Asian parasites. As part of aprogramme to understand the invasiveness of A. crassus in An. anguilla, we are investigating differences in gene expression and genetic distinction between invading European and endemic Asian A. crassus exposed to the two different host species. Here we report on the generation of a reference transcriptome for A. crassus based on Roche 454 data, and explore patterns of gene expression and diversity.

Results

Sampling A. crassus

One female worm and one male worm were sampled from an aquaculture with height infection loads in Taiwan. An additonal female worm was sampled from a stream with low infection pressure adjacent to the aquaculture. All these worms were parsitising endemic *An. japonica*. A female worm and pool of L2 larval stages were sampled from *An. anguilla* in the river Rhein, one female worm from a lake in Poland. All adult worms were filled with large amounts of host-bood, therefore we anticipated abundand host-contamination in sequencing and decided to sequence a liver sample of an unifected *An. japonica* for screening.

Sequencing, trimming and pre-assembly screening

A total of 756363 raw sequencing reads were generated for *A. crassus* (Table 1). These were trimmed for base call quality, and filtered by length to give 585949 high-quality reads (spanning 100491819 bases). In the eel data-set from 159370 raw reads 135072 were assembled after basic quality screening.
We then screened the *A. crassus* reads for contamination by host (30071 matched previously sequenced eel genes in our own *An. anguilla* 454 transcriptome, which was partitioned in 10639 mRNA and 53 rRNA TUGs after the nematode (181783 reads matched large or small subunit nuclear or mitochondrial ribosomal RNA sequences of A. crassus) (Table 1). In addition to fish mRNAs, we identified (and removed) 5286 reads in the library derived from the L2 nematodes that had significant similarity to cercozoan (likely parasite) ribosomal RNA genes (Table 1).

Assembly

We assembled the remaining 353055 reads (spanning 100491819 bases) using the combined assembler strategy [21] and Roche 454 GSassembler (version 2.6) and MIRA (version 3.21) [61]. From this we derived 13851 contigs that were supported by both assembly algorithms, 3745 contigs only supported by one of the assembly algorithms and 22591 singletons that were not assembled by either approach (Table 2). When scored by matches to known genes, the contigs supported by both assemblers are of the highest credibility, and this set is thus termed the high credibility assembly (highCA). Those with evidence from only one assembler and the singletons are of lower credibility (lowCA). These datasets are the most parsimonious (having the smallest size) for their quality (covering the largest amount of sequence in reference transcritomes). In the highCA parsimony and low redundancy is prioritized, while in the complete assembly (highCA plus lowCA) completeness is proiritized. The 40187 sequences (contig consensuses and

singletons) in the complete assembly are referred to below as tentatively unique genes (TUGs).
We screened the complete assembly for residual host contamination, and identified 40187 TUGs that had higher, significant similarity to eel (and chordate) sequences (our 454 ESTs and EMBLBank Chordata proteins) than to nematode sequences [25].
Given our prior identification of cercozoan ribosomal RNAs, we also screened the complete assembly for contamination with other transcriptomes, and found 365 TUGs with hits to fungi (e.g Ajellomycetaceae, 53 hits), 672 TUGs whith hits to plants and 2002 hits to Protists (e.g. Trypanosomatidae, 26 hits and Vahlkampfiidae, 38 hits), Bacteria (mostly Proteobacteria, 484 hits) and Viruses (see also additional figure phylum_plots.png).
No hits were found to Wolbachia or related Bacteria known as symbionts of Ecdyosozoans.
Our assembly thus has 32518 TUGs, spanning 154052 bases (of which 11371 are highCA-derived, and span 154052 bases) that are likely to derive from of *A. crassus*.

Protein prediction

```
[1] 1
```

```
[1] 1
```

For 32411 TUGs a protein was predicted using prot4EST [64] (Table 2). The full open reading frame was obtained in 353 TUGs, while while for 2683 the 5' end and for 8283 the 3' end was complete. In 13379 TUGs the corrected sequence with the imputed ORF was slightly changed compared to the raw sequence.

Annotation

We obtained basic annotations with orthologous sequences from *C. elgeans* for 9553 TUGs, from *B. malayi* for 9662 TUGs, from nempep [25, 63] for 10494 TUGs and whith uni-prot proteins for 10539 TUGs.
We used annot8r [67] to assign gene ontology (GO) terms for 6509 TUGs, Enzyme Commission (EC) numbers for 2458 TUGs and Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathway annotations for 3844 TUGs (Table 2). Additionally 2912 highCA derived contigs were anntotated with GO terms through Interpro Scan [**?**]. Nearly one third (6987) of the *A. crassus* TUGs were annotated with at least one identifier, and 1829 had GO, EC and KEGG annotations (Figure annotataionVenn.tiff).
We compared our *A. crassus* GO annotations for high-level GO-slim terms to the annotations (obtained the same way) for the complete proteome of the filarial nematode *B. malayi* and the complete proteome of

C. elegans.

Correlation shows the occurence of terms for the partial transcriptome of *A. crassus* to be more similar to the proteome of *B. malayi* (0.95; Spearman correlation coefficient) than to the proteome of *C. elegans* (0.9). Also the tow model-nematode compared ot each other (0.91) are less similar in the occurence of therm.

We inferred presence of signal peptide cleavage sites in the predicted protein sequence using SignalP [28]. We predicted 920 signal peptide cleavage sites and 65 signal peptides with a transmembrane signature. Again these predictions are more similar to predictions using the same methods for the proteome *B. malayi* (742 signal peptide cleavage sites and 41 with transmembrane anchor) than for the proteome of *C. elegans* (4273 signal peptide cleavage sites and 154 with transmembrane anchor).

Evolutionary conservation

Evolutionary conservation The A. crassus TUGs were classified as conserved, conserved in Metazoa, conserved in Nematoda, conserved in Spirurina or novel to A. crassus by comparing them to public databases and using two BLAST bit-score cutoffs to define relatedness.

****Now describe the data, and reference table 3, add novel in A. crassus****

Identification of single nucleotide polymorphisms

We called single nucleotide polymorphisms (SNPs) on the 1,412,806 bases of the TUGs that had coverage of more then 8-fold available using VARScan [69]. We excluded SNPs predicted to have more than 2 alleles or that mapped to an undertermined (N) base in the reference, and retained 13042 SNPs. The ratio of transitions (ti; 8663) to transversion (tv; 4379) in this set was 1.98. Using the prot4EST prtedictions and the corrected sequences, 8309 of the SNPs were predicted to be inside an ORF, with 2713 at codon first positons, 2172 at second positions and 3424 at third positions. As expected ti/tv inside ORFs (2.43) was higher than outside ORFs (1.42). The ratio of synonymous polymorphisms per synonymous site to non-synonymous polymorphisms per non-synonymous site (dn/ds) was 0.45. We filtered these SNPs to exclude those that might be associated with analytical bias. As Roche 454 sequences have well-known systematic errors associated with homopolymeric nucleotide sequences [27], we analysed the effect of exclusion of SNPs in, or close to, homopolymer regions. We observed changes in ti/tv and in dn/ds when SNPs were discarded using different size thresholds for homoploymer runs and proximity thresholds (see figure snp_ex_parameter_plots.png). Based on this we decided to exclude SNPs with a homopolymer-run as long as or longer than 4 bases inside a window of 11 bases (5 to bases to the right, 5 to the left) around the

SNP. We also observed a relationship between TUG dn/ds and TUG coverage, associated with the presence of sites with low abundance minority alleles (less than 7% of the allele calls), suggesting that some of these may be errors. Removing low abundance minority allele SNPs from the set removed this effect. Our filtered SNP dataset includes 5112 SNPs. We retained 4.44 SNPs per kb of contig sequence, with 7.87 synonymous SNPs per 1000 synonymous bases and 2.43 non-synonymous SNPs per 1000 non-synonymous bases. A mean dn/ds of 0.32 was calculated for the 980 TUGs (858 highCA and lowCA contigs) containing at least one synonymous SNP.

Polymorphisms associated with biological processes

We consolidated our annotation and polymorphism analyses by examining correlations between nonsynonymous variability and particular classifications. Signal peptide containing proteins have been shown to have higher rates of evolution than cytosolic proteins in a number of nematode species. In A. crassus, TUGs predicted to contain signal peptide cleavage sites by the neural-networks method in SignalP had higher dn/ds values than TUGs without signal peptide cleavage sites ($p = 0.055$; one sided U-test; Figure sigp_dn_ds.png). Proteins predicted to be novel to the Nematoda were significantly enriched in signal peptide annotation
(Figure signal_novel.png ***why is nematoda on the outside? should order by breadth of conservation***).
Positive selection can be inferred from dn/ds analyses, and we defined TUGs with a dn/ds higher than 0.5 as positively selected. We identified over- and under-represented GO ontology terms associated with these putatively positively selected genes (Table XX) ***can you give significances in the table***. Within the molecular function amino acid transmembrane transporter peptidase were the most significant? overrepresented terms, while terms associated with ribosomal proteins and transcription were underrepresented. TODO: BP and CC over-representation.
These inferences remained valid when only the highCA and medCA contigs were analysed (Figure conservation_dn_ds.png). At a bit score threshold of 80 sequences novel in clade III had a significantly higher dn/ds than other sequences ($p = 0.038$; two sided U-test). At a bitscore thershold of 50 results were non-significant but showed the same trend.

Signal-positives have higher dn/ds

TUGs predicted to contain signal peptide cleavage sites by the neural-networks method in `SignalP` have higher dn/ds values than TUGs without signal peptide cleavage sites ($p = 0.074$; one sided U-test; see also

figure sigp_dn_ds.png).

Enrichment of GO-categories in high dn/ds

We defined TUGs with a dn.ds higher than 0.5 as positively selected and tested each node-term in the ontology for a over-representation in this set. Tables X give an overview of enriched terms for different parts of the ontology.

The terms "amino acid transmembrane transporter activity" and "peptidase activity" were among the overrepresented terms for "molecular function". Underrepresented were on the other hand terms associated with ribosomal proteins and transcription (visible in "molecular function" and "cellular compartment") .

TODO: BP and CC over-representation.

To make sure these inferences were not biased by redundancy in the fullest data-set we repeated the analysis using only the good-quality contigs as gene-universe. The results showed to be consistent with those obtained for the fullest data-set (data not shown).

Novel in cladeIII have elevated dn/ds

Figure conservation_dn_ds.png shows differences in dn/ds across the categories defined for evolutionary conservation. At a bitscore threshold of 80 sequences novel in clade III had a significantly higher dn/ds than other sequences ($p = 0.038$; two sided U-test). At a bitscore thershold of 50 results were non-significant but showed the same trend.

Sequences novel in nematodes are enriched for Signal-positives

Figure signal_novel.png gives the proportions of `SignalP`-predictions for each category of evolutionary conservation. Generally - across bit-score thresholds - sequences novel in nematodes contained the highest proportion of signal-positives.

Differential expression

Using methods developed for sequencing data, we analyzed gene-expression inferred from mapping. Of the 341285 reads mapping to the fullest assembly 264440 mapped uniquely (with their best hit) and were counted on a per library base.

Comparison with tag-sequencing pilot data-set

5096312 of 6201930 (559824 unique) NlaIII-tags mapped to the fullest assembly. Only 1105618 (317782 unique) tags did not map to any sequence in the fullest assembly.

Table 9a gives correlations coefficients between tag-counts and 454-libraries. Correlations-coefficients between 454-libraries were generally low, indicating a high proportion of noise or biological differences between samples. Correlation between expression-tags and 454-read counts were even lower. However when only analyzing counts to good-quality contigs, correlation coefficients improved both between libraries and between 454-libraries and solexa-tags (see table 9b). No further improvements were made, when counts were limited to contigs surely *A. crassus* (see table 9c). Correlations between library T2 and other 454-libraries, as well as with solexa-tag counts were lower than between other libraries.

To gain power in statistical analysis we limited the set of gene-objects analyzed for differential expression to the good-quality contigs.

Differential expression between male and female worms

Despite the lack of replicates for male worms we were able to identify 49 sequences being significantly over-expressed in male worms. In fact all these TUGs were nearly exclusively expressed in males.

Differential expression between adults and L2-larvae

For the L2-library we changed our approach and used gene-expression analysis rather to highlight the off-target data in this library. For this reason we used counts for the fullest assembly.

137 sequences being expressed exclusively in L2 library were strongly enriched in sequences being labeled as possible off-target data in taxonomic classification. From these sequences only 23 had best hits to metazoa and only 1 to nematoda.

Differential expression between worms from the European and Japanese eel

None of the TUGs in the present evaluation showed significant differential expression between worms from the European and Japanese Eel. Diagnostic plots provided by `DESeq` made clear, that both depth of sequencing and number of replicates have to be higher contrasting these conditions.

However, comparing expression-analysis on the full data-set to analysis limited to the high quality of reliable *A. crassus*-contigs it was clear that the quality-data-set reduces within-condition variance and

results were closer to significance: The lowest adjusted p-values for the cleaned data-set were around 0.4, while on the full data-set only adjusted p-values above 0.8 could be obtained.

Discussion

We have generated a de novo transcriptome for A. crassus an important invasive parasite that threatens wild stocks of the European eel An. anguilla. These data enable a broad spectrum of molecular research on this ecologically and economically important parasite. As A. crassus lives in close association with its host, we have used exhaustive filtering to attempt to remove all host-derived, and host-associated organism-derived contamination from the data. To do this we have also generated a transcriptome dataset from the definitive host An. japonica. The non-nematode, non-eel data identified, particularly in the L2 sample, showed highest identity to flagellate protists, which may have been parasitising the eel (or the nematode). Encapsulated objects observed in eel swim bladder walls [14] could be due solely to immune attrition of A. crassus larvae or to other coinfections.

A second examination of sequence origin was performed after assembly, employing higher stringency cutoffs. Similar taxonomic screening was used in a garter snake transcriptome project [33], and an analysis of lake sturgeon tested and rejected hypotheses of horizontal gene-transfer when xenobiont sequences were identified [34]. A custom pipeline for transcriptome assembly from pyrosequencing reads [35] proposed the use of EST3 [36] to infer sequence origin based simply on nucleotide frequency. We were not able to use this approach successfully, probably due to the fact that xenobiont sequences in our data set derive from multiple sources with different GC content and codon usage.

Compared to other NGS transcriptome sequencing projects [references???], the combined assembly approach generated a smaller number of contigs that had lower redundancy and higher completeness. Projects using the mira assembler often report substantially greater numbers of contigs for datasets of similar size (see e.g. [37]), comparable to the mira sub-assembly in our approach. The use of oligo(dT) to capture mRNAs probably explains the bias towards 3' end completeness and a relative lack of true initiation codons in our protein prediction. This bias is near-ubiquitous in deep transcriptome sequencing projects (e.g. [38]).

We generated transcriptome data from multiple A. crassus of Taiwanese and European origin, and identified SNPs both within and between populations. Screening of SNPs in or adjacent to homopolymer regions improved overall measurements of SNP quality. The ratio of transitions to transversions (ti/tv) increased. Such an increase is explained by the removal of "noise" associated with common homopolymer

errors [27]. The value of 2.38 (1.82 outside, 2.74 inside ORFs) is in good agreement with the overall ti/tv of humans (2.16 [39]) or Drosophila (2.07 [40]). The ratio of non-synonymous SNPs per non-synonymous site to synonymous SNPs per synonymous site (dn/ds) decreased with removal of SNPs adjacent to homopolymer regions from 0.45 to 0.32 after full screening. The most plausible explanation is the removal of error, as unbiased error would lead to a dn/ds of 1. While dn/ds is not unproblematic to interpret within populations [41], the assumption of negative (purifying) selection on most protein-coding genes makes lower mean values seem more plausible. We used a threshold value for the minority allele of 7% for exclusion of SNPs, based on an estimate that approximately 10 haploid equivalents were sampled (5 individual worms plus an negligible contribution from L2 larvae in the L2 library and within the female adult worms). The benefit of this screening was mainly a reduction of non-synonymous SNPs in high coverage contigs, and a removal of the dependence of dn/ds on coverage. Working with an estimate of dn/ds independent of coverage, efforts to control for sampling a biased by sampling depth (i.e. coverage; see [42] and [43]) could be avoided.

*** When the whole of coding sequences are studied, of which only a small subset of sites can be under diversifying selection, dn/ds of 0.5 has been suggested as threshold for assuming diversifying selection [44] instead of the classical threshold of 1 [45]. In the transcripts from the female reproductive tract of *Drosophila* dn/ds was 0.15 [44] and in the 0.21 male reproductive tract [46] (although for ESTs specific to the male accessory gland were shown to have a higher dn/ds of 0.47). Pyrosequencing studies found dn/ds to be between 0.13 and 0.27 (depending on tissue type genes were mainly expressed in) in the Zebra finch transcriptome [47], 0.12 in the transcriptome of *Tigriopus californicus* [48] and 0.3 in the parasitic nematode *Ancylostoma canium* [49]. In comparison with these results even our estimate after screening seems high (although it should be noted, that the latter tree studies report a mean dn/ds over contigs - the *A. canium* doesn't make clear what exactly is reported - and therefore the value has to be compared to our mean dn/ds over contigs of 0.23) and further investigation using deeper sequencing of more individuals on the solexa GAII platform will be used to fully exclude the possibility of this result being induced by sequencing error. Moreover such an experiment should try to test that divergence between populations is leading to positive selection on only the possibly diverging European populations. For such a study the set of SNPs found here are invaluable, as it can be used to define a gold standard set of SNPs found with both technologies.

We were able to obtain high-quality annotations for a large set of TUGs. Comparison with protein sequence derived from *B. malayi* showed a a remarkable degree of agreement regarding the occurrence of terms.

This implies, that our transcriptome-data-set is a representative subset of a nematode-parasite genome.
Over-representation of GO-term in genes under diversifying selection (at a threshold of dn/ds>0.5, as established above) highlighted many interesting gene-products:
In the molecular function category two amino acid transmembrane transporters ("Contig5699" and "Contig866") - the only contigs with this annotation (or annotation, which is an offspring-term of this) and a dn/ds obtained - were found to have a dn/ds>0.5. Such transporter are thought to be important in the survival of parasites in a host [50].
Enrichment in the category "peptidase activity" highlighted twelve peptidases (from 43 with a dn/ds obtained). All twelve have orthologs in *B. malayi* and *C. elgans* and are conserved across kingdoms. Despite their conservation peptidases are thought to have have acquired new and prominent roles in host-parasite interaction compared to free living organisms: In *A. crassus* a trypsin-like proteinase has been identified thought to be utilized by the tissue-dwelling L3 stage to penetrate host tissue and an aspartyl proteinase thought to be a digestive enzyme in adults [51].
The under-representation of ribosomal proteins (term "structural constituent of ribosome") in disruptively selected contigs is in good agreement with the notion that ribosomal proteins are extremely conserved across kingdoms [52] and should be under under strong negative selection.
The additional prediction of signal sites for cleavage allowed interpretation and cross-validation of the results from SNP-calling: The detection of signal-peptides secretion using *in silico* analysis of ESTs has been used to highlight candidate genes for example in *Nippostrongylus brasiliensis* [53] and in a large scale analysis across all nematode [54] ESTs. Proteomic analysis in *B. malayi* [55, 56] and *Heligmosomoides polygyrus* [57] was able to find evidence for excretion for some of the protein-products and to highlight additional candidate genes.
We found an elevated dn/ds for signal-positives. These result could be explained follow the logic of signal-positives being more likely to be secreted to the host-parasite interface and proteins involved in host-parasite interaction being more likely to be under disruptive selection. Signal-positive TUGs with high dn/ds constitute another set of genes worth further examination in future studies.
TUGs predicted to be novel in the phylum nematoda contained the highest proportion of signal-positives. A interpretation of this findings could be a confirmation of a study on *Nippostrongylus brasiliensis* [53], where signal positives were reported as less conserved. In the present study we did not aim to identify "novelty to *A. crassus*" as we believe in a deep sequencing project he absence of sequence similarity could be attributed to erroneous sequence instead of true novelty, and thereby blur analysis. However novelty in

nematodes and to a lesser extend novelty in Spirurina seems to support the notion, that - if not diversified within nematoda to an extend leading to a complete loss of similarity, like suggested in the mentioned study - signal positives in nematodes could have taken a divergent evolutionary path from their orthologs in other phyla.

It was within our expectation, that expression analysis failed to give conclusive results, as the present data-set is not fully adequate for this kind of analysis: First we did not include replicates for libraries of male adults as well as for L2-larvae. Second one of the replicates for female worms (library E1) resulted in a low amount of sequence mappable to protein-coding (non-rRNA) genes. However some of the results are still valuable:

DESeq was able to report genes significantly differing in expression between male and female worms and between the L2 library and the all other worms. This was possible for male worms as well as for L2-larvae, were no replicated samples were obtained, due due the special features of this package [58]. However only over-expression in non-repeated samples can be detected, as obviously lack of expression in one sample can't validate

Comparisons were lacking significance, as methods are designed for deeper sequencing and more importantly more replicates would be needed. Differences between the L2-library and other libraries were mainly due to off-target data, and TUGs solely found in the L2 library are ...

Conclusions

Methods

Nematode samples, RNA extraction, cDNA synthesis and Sequencing

A. crassus from *JAn. japonica* were sampled from Kao-Ping river and an adjacent aquaculture in Taiwan as described in [14]. Worms from An. anguilla were sampled in Sniardwy Lake, Poland (53.751959N, 21.730957E) and from the Linkenheimer Altrhein, Germany (49.0262N, 8.310556E). After determination of the sex of adult nematodes, they were stored in RNA-later (Quiagen, Hilden, Germany) until extraction of RNA. RNA was extracted from individual adult male and female nematodes and from a population of L2 larvae (Table 1). RNA was reverse transcribed and amplified into cDNA using the MINT-cDNA synthesis kit (Evrogen, Moscow, Russia). For host contamination screening a liver-sample from an uninfected A. japonica was also processed. Emulsion PCR was performed for each cDNA library according to the manufacturer's potocols (Roche/454 Life Sciences), and sequenced on a Roche 454 Genome Sequencer FLX. All samples were sequenced using the FLX Titanium chemistry, except for the taiwanese female

sample T2, which was sequenced using FLX standard chemistry, to generate between 99,000 and 209,000 raw reads. For the L2 larval library, which had a larger number of non-A. crassus, non-An. anguilla reads, we confirmed that these data were not laboratory contaminants by screening Roche 454 data produced on the same run in independent sequencing lanes.

Trimming, quality control and assembly

Raw sequences were extracted in fasta format (with the corresponding qualities files) using sffinfo (Roche/454) and screened for adapter sequences of the MINT-amplification-kit using cross-match [59] (with parameters -minscore 20 and -minmatch 10). Seqclean [60] was used to identify and remove poly-A-tails, low quality, repetitive and short (<100 base) sequences. All reads were compared to a set of screening databases using BLAST (expect value cutoff E<1e-5, low complexity filtering turned off: -F F). The databases used were (a) a host sequence database comprising an assembly of the An. japonica Roche 454 data, an unpublished assembly of An. anguilla Sanger dideoxy sequencesd expressed sequence tags (made available to us by Gordon Cramb, University of St Andrews) and transcripts from from EeelBase [30] a publically availble transcriptome database for the European eel; (b) a database of ribosomal RNA (rRNA) sequences from eel species derived from our Roche 454 data and EMBL-Bank; and (c) a database of rRNA sequences identified in our A. crassus data by comparing the reads to known nematode rRNAs from EMBL-Bank. This last database notably also contained xenobiont rRNA sequences. Reads with matches to one of these databases over more than 80% of their length and with greater than 95% identity were removed from the dataset. Screening and trimming information was written back into sff-format using sfffile (Roche 454). The filtered and trimmed data were assembled using the combined assembly approach [21], combining assemblies from the mira [61] and newbler [20]. ****Give the details here and we will trim the text later **** . The two assemblies were combined into one using Cap3 [62] at default settings and contigs labeled by whether they derived from both assemblies or one assembly only.

Post-assembly classification and taxonomic assignment of contigs

After assembly contigs were assessed a second time for host and other contamination by comparing them (using BLAST) to the three databases defined above, and also to nembase4, a nematode transcriptome database derived from whole genome sequencing and EST assemblies [25, 63]. For each contig, the highest-scoring match was recorded as long as it spanned more than 50% of the contig. We also compared the contigs to the NCBI non-redundant nucleotide (NCBI-nt) and protein (NCBI-nr) databases, recording

the taxonomy of all best matches with expect values better than 1e-05.

Protein prediction and annotation

Protein translations were predicted from the contigs using prot4EST (version 3.0b) [64]. Proteins were predicted either by joining single high scoring segment pairs (HSPs) from a BLAST search of uniref100 [65], or by ESTscan [66], using a training data the *Brugia malayi* complete proteome back-translated using a codon usage table derived from the BLAST HSPs, or, if the first two methods failed, simply the longest ORF in the contig. For contigs where the proein prediction required insertion or deletion of bases in the original sequence, we also imputed an edited sequence for each affected contig. Annotations with Gene Ontology (GO), Enzyme Commission (EC) and Kyoto Encyclopaedia of Genes and Genomes (KEGG) terms were inferred for these proteins using Annot8r (version 1.1.1) [67], using the annotated sequences available in uniref100 [65]. Up to 10 annotations based on a BLAST similarity bitscore cut-off of 55 were obtained for each annotation set. The complete *B. malayi* proteome (as present in uniref100) and the complete *C. elegans* proteome (as present in wormbase v.220) were also annotated in the same way. SignalP V4.0 [28] was used to predict signal peptide cleavage sites and signal anchor signatures.

Single nucleotide polymorphism analysis

We mapped the raw reads against the the complete set of contigs, replacing imputed sequences for originals where relevant, using ssaha2 (with parameters -kmer 13 -skip 3 -seeds 6 -score 100 -cmatch 10 -ckmer 6 -output sam -best 1). From the ssaha2 output, pileup-files were produced using samtools [68], discarding reads mapping to multiple regions. VarScan [69] (pileup2snp) was used with default parameters on pileup-files to output lists of single nucleotide polymorphisms (SNPs) and their locations.

Gene-expression analysis

For Roche 454 data, read counts for each transcript were obtained from the mapping to imputed sequence performed for SNP analyses. Tag-sequences were mapped using BWA [70]. And read counts extracted using Samtools [68]. For deepSAGE NlaIII-tag-sequencing, total RNA was prepared as described above from a female nematode from the Polish sampling site. A deepSAGE library was constructed following the protocol supplied by Illumina. Briefly after synthesis of cDNA on oligo(dT)-beads, cDNA was digested with the NlaIII (recognition site CATG), and the oligo(dT)-anchored 3' ends of mRNAs retained. After ligation of an adaptor containing an MmeI restriction site, the type II enzyme MmeI was used to cut 17

bases from the 3' end fragment, genrating a 21 base tag, expected to be unique for most mRNAs. The R-package DESeq [58] was used to normalize for library size and analyse statistical significance of differential expression of both Roche 454 and deepSAGE data. Spearman correlation coefficients were calculated for raw counts.

General coding methods

The bulk of analysis (unless otherwise cited) presented in this paper was carried out in R [71] using custom scripts. We used a method provided in the R-packages Sweave [72] and Weaver [73] for "reproducible research" combining R and TEXcode in a single file. All intermediate data files needed to compile the present manuscript from data-sources are provided upon request. For visualistaion we used the R-packages lattice [74] and ggplot2 [75].

Competing interests

The authors declare no competing interests.

Authors contributions

Acknowledgments

The work of EGH is funded by Volkswagen Foundation, "Förderinitiative Evolutionsbiologie".

References

1. Kuwahara A, Niimi H, Itagaki H: **Studies on a nematode parasitic in the air bladder of the eel I. Descriptions of Anguillicola crassa sp. n. (Philometridea, Anguillicolidae)**. *Japanese Journal for Parasitology* 1974, **23**(5):275–279.
2. Neumann W: **Schwimblasenparasit Anguillicola bei Aalen**. *Fischer und Teichwirt* 1985, :322.
3. Koops H, Hartmann F: **Anguillicola-infestations in Germany and in German eel imports**. *Journal of Applied Ichthyology* 1989, **5**:41–45.
4. Koie M: **Swimbladder nematodes (Anguillicola spp.) and gill monogeneans (Pseudodactylogyrus spp.) parasitic on the European eel (Anguilla anguilla)**. *ICES J. Mar. Sci.* 1991, **47**(3):391–398, [http://icesjms.oxfordjournals.org/cgi/content/abstract/47/3/391].
5. Kirk RS: **The impact of Anguillicola crassus on European eels**. *Fisheries Management & Ecology* 2003, **10**(6):385–394, [http://dx.doi.org/10.1111/j.1365-2400.2003.00355.x].
6. Kristmundsson A, Helgason S: **Parasite communities of ecls Anguilla anguilla in freshwater and marine habitats in Iceland in comparison with other parasite communities of eels in Europe**. *Folia Parasitologica* 2007, **54**(2):141.
7. Würtz J, Knopf K, Taraschewski H: **Distribution and prevalence of Anguillicola crassus (Nematoda) in eels Anguilla anguilla of the rivers Rhine and Naab, Germany**. *Diseases of Aquatic Organisms* 1998, **32**(2):137–43, [http://www.ncbi.nlm.nih.gov/pubmed/9676253].
8. Lefebvre FS, Crivelli AJ: **Anguillicolosis: dynamics of the infection over two decades**. *Diseases of Aquatic Organisms* 2004, **62**(3):227–32, [http://www.ncbi.nlm.nih.gov/pubmed/15672878].

9. Taraschewski H: **Hosts and Parasites as Aliens**. *Journal of Helminthology* 2007, **80**(02):99–128, [http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=713884].

10. Münderle M, Taraschewski H, Klar B, Chang CW, Shiao JC, Shen KN, He JT, Lin SH, Tzeng WN: **Occurrence of Anguillicola crassus (Nematoda: Dracunculoidea) in Japanese eels Anguilla japonica from a river and an aquaculture unit in SW Taiwan**. *Diseases of Aquatic Organisms* 2006, **71**(2):101–8, [http://www.ncbi.nlm.nih.gov/pubmed/16956057].

11. Knopf K: **The swimbladder nematode Anguillicola crassus in the European eel Anguilla anguilla and the Japanese eel Anguilla japonica: differences in susceptibility and immunity between a recently colonized host and the original host**. *Journal of Helminthology* 2006, **80**(2):129–36, [http://www.ncbi.nlm.nih.gov/pubmed/16768856].

12. Knopf K, Mahnke M: **Differences in susceptibility of the European eel (Anguilla anguilla) and the Japanese eel (Anguilla japonica) to the swim-bladder nematode Anguillicola crassus**. *Parasitology* 2004, **129**(Pt 4):491–6, [http://www.ncbi.nlm.nih.gov/pubmed/15521638].

13. Knopf K, Lucius R: **Vaccination of eels (Anguilla japonica and Anguilla anguilla) against Anguillicola crassus with irradiated L3**. *Parasitology* 2008, **135**(5):633–40, [http://www.ncbi.nlm.nih.gov/pubmed/18302804].

14. Heitlinger E, Laetsch D, Weclawski U, Han YS, Taraschewski H: **Massive encapsulation of larval Anguillicoloides crassus in the intestinal wall of Japanese eels**. *Parasites and Vectors* 2009, **2**:48, [http://www.parasitesandvectors.com/content/2/1/48].

15. Würtz J, Taraschewski H: **Histopathological changes in the swimbladder wall of the European eel Anguilla anguilla due to infections with Anguillicola crassus**. *Diseases of Aquatic Organisms* 2000, **39**(2):121–34, [http://www.ncbi.nlm.nih.gov/pubmed/10715817].

16. Blaxter ML, Ley PD, Garey JR, Liu LX, Scheldeman P, Vierstraete A, Vanfleteren JR, Mackey LY, Dorris M, Frisse LM, Vida JT, Thomas WK: **A molecular evolutionary framework for the phylum Nematoda**. *Nature* 1998, **392**(6671):71–75, [http://dx.doi.org/10.1038/32160].

17. NADLER S, CARRENO R, MEJ?A-MADRID H, ULLBERG J, PAGAN C, HOUSTON R, HUGOT J: **Molecular Phylogeny of Clade III Nematodes Reveals Multiple Origins of Tissue Parasitism**. *Parasitology* 2007, **134**(10):1421–1442, [http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=1279744].

18. Wijová M, Moravec F, Horák A, Lukes J: **Evolutionary relationships of Spirurina (Nematoda: Chromadorea: Rhabditida) with special emphasis on dracunculoid nematodes inferred from SSU rRNA gene sequences**. *International Journal for Parasitology* 2006, **36**(9):1067–75, [http://www.ncbi.nlm.nih.gov/pubmed/16753171].

19. Zang X, Maizels RM: **Serine proteinase inhibitors from nematodes and the arms race between host and pathogen**. *Trends in Biochemical Sciences* 2001, **26**(3):191–197, [http://www.sciencedirect.com/science/article/B6TCV-42H1RTN-T/2/0a8af31e701aab88f214aad50e50bdca].

20. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM: **Genome sequencing in microfabricated high-density picolitre reactors**. *Nature* 2005, **437**:376–380, [http://dx.doi.org/10.1038/nature03959].

21. Kumar S, Blaxter ML: **Comparing de novo assemblers for 454 transcriptome data**. *BMC Genomics* 2010, **11**:571, [http://dx.doi.org/10.1186/1471-2164-11-571].

22. Malone JH, Oliver B: **Microarrays, deep sequencing and the true measure of the transcriptome**. *BMC Biol.* 2011, **9**:34.

23. Matsumura H, Yoshida K, Luo S, Kruger DH, Kahl G, Schroth GP, Terauchi R: **High-throughput SuperSAGE**. *Methods Mol. Biol.* 2011, **687**:135–146.

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

25. Elsworth B, Wasmuth J, Blaxter M: **NEMBASE4: The nematode transcriptome resource**. *Int. J. Parasitol.* 2011, **41**:881–894.

26. Blaxter M: *Base Ontology: An idea from Mark Blaxter* 2010, [http://genepool.bio.ed.ac.uk/nextgenbug/resources/gff_parsing_group].

27. Balzer S, Malde K, Jonassen I: **Systematic exploration of error sources in pyrosequencing flowgram data**. *Bioinformatics* 2011, **27**:i304–309.

28. Petersen TN, Brunak S, von Heijne G, Nielsen H: **SignalP 4.0: discriminating signal peptides from transmembrane regions**. *Nat. Methods* 2011, **8**:785–786.

29. Lennon NJ, Lintner RE, Anderson S, Alvarez P, Barry A, Brockman W, Daza R, Erlich RL, Giannoukos G, Green L, Hollinger A, Hoover CA, Jaffe DB, Juhn F, McCarthy D, Perrin D, Ponchner K, Powers TL, Rizzolo K, Robbins D, Ryan E, Russ C, Sparrow T, Stalker J, Steelman S, Weiand M, Zimmer A, Henn MR, Nusbaum C, Nicol R: **A scalable, fully automated process for construction of sequence-ready barcoded libraries for 454**. *Genome Biol.* 2010, **11**:R15.

30. Coppe A, Pujolar JM, Maes GE, Larsen PF, Hansen MM, Bernatchez L, Zane L, Bortoluzzi S: **Sequencing, de novo annotation and analysis of the first Anguilla anguilla transcriptome: EeelBase opens new perspectives for the study of the critically endangered European eel**. *BMC Genomics* 2010, **11**:635.

31. Wilson IG: **Inhibition and facilitation of nucleic acid amplification**. *Appl. Environ. Microbiol.* 1997, **63**:3741–3751.

32. Valasek MA, Repa JJ: **The power of real-time PCR**. *Adv Physiol Educ* 2005, **29**:151–159.

33. Schwartz TS, Tae H, Yang Y, Mockaitis K, Van Hemert JL, Proulx SR, Choi JH, Bronikowski AM: **A garter snake transcriptome: pyrosequencing, de novo assembly, and sex-specific differences**. *BMC Genomics* 2010, **11**:694.

34. Hale MC, Jackson JR, Dewoody JA: **Discovery and evaluation of candidate sex-determining genes and xenobiotics in the gonads of lake sturgeon (Acipenser fulvescens)**. *Genetica* 2010, **138**:745–756.

35. Papanicolaou A, Stierli R, Ffrench-Constant RH, Heckel DG: **Next generation transcriptomes for next generation genomes using est2assembly**. *BMC Bioinformatics* 2009, **10**:447.

36. Emmersen J, Rudd S, Mewes HW, Tetko IV: **Separation of sequences from host-pathogen interface using triplet nucleotide frequencies**. *Fungal Genet. Biol.* 2007, **44**:231–241, [http://dx.doi.org/10.1016/j.fgb.2006.11.010].

37. Gregory R, Darby AC, Irving H, Coulibaly MB, Hughes M, Koekemoer LL, Coetzee M, Ranson H, Hemingway J, Hall N, Wondji CS: **A De Novo Expression Profiling of Anopheles funestus, Malaria Vector in Africa, Using 454 Pyrosequencing**. *PLoS ONE* 2011, **6**:e17418.

38. Kunstner A, Wolf JB, Backstrom N, Whitney O, Balakrishnan CN, Day L, Edwards SV, Janes DE, Schlinger BA, Wilson RK, Jarvis ED, Warren WC, Ellegren H: **Comparative genomics based on massive parallel transcriptome sequencing reveals patterns of substitution and selection across 10 bird species**. *Mol. Ecol.* 2010, **19 Suppl 1**:266–276.

39. Yang H, Chen X, Wong WH: **Completely phased genome sequencing through chromosome sorting**. *Proc. Natl. Acad. Sci. U.S.A.* 2011, **108**:12–17.

40. Adey A, Morrison H, Asan X, Xun X, Kitzman J, Turner E, Stackhouse B, MacKenzie A, Caruccio N, Zhang X, Shendure J, Turner E, Stackhouse B, MacKenzie A, Caruccio N, Zhang X, Shendure J: **Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition**. *Genome Biol.* 2010, **11**(12):R119.

41. Kryazhimskiy S, Plotkin JB: **The population genetics of dN/dS**. *PLoS Genet.* 2008, **4**:e1000304.

42. Novaes E, Drost DR, Farmerie WG, Pappas GJ, Grattapaglia D, Sederoff RR, Kirst M: **High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome**. *BMC Genomics* 2008, **9**:312.

43. O'Neil ST, Dzurisin JD, Carmichael RD, Lobo NF, Emrich SJ, Hellmann JJ: **Population-level transcriptome sequencing of nonmodel organisms Erynnis propertius and Papilio zelicaon**. *BMC Genomics* 2010, **11**:310.

44. Swanson WJ, Wong A, Wolfner MF, Aquadro CF: **Evolutionary expressed sequence tag analysis of Drosophila female reproductive tracts identifies genes subjected to positive selection**. *Genetics* 2004, **168**:1457–1465.

45. Miyata T, Yasunaga T: **Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application**. *J. Mol. Evol.* 1980, **16**:23–36.

46. Swanson WJ, Clark AG, Waldrip-Dail HM, Wolfner MF, Aquadro CF: **Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in Drosophila**. *Proc. Natl. Acad. Sci. U.S.A.* 2001, **98**:7375–7379.

47. Ekblom R, Balakrishnan CN, Burke T, Slate J: **Digital gene expression analysis of the zebra finch genome**. *BMC Genomics* 2010, **11**:219.

48. Barreto FS, Moy GW, Burton RS: **Interpopulation patterns of divergence and selection across the transcriptome of the copepod Tigriopus californicus**. *Mol. Ecol.* 2011, **20**:560–572.

49. Wang Z, Abubucker S, Martin J, Wilson RK, Hawdon J, Mitreva M: **Characterizing Ancylostoma caninum transcriptome and exploring nematode parasitic adaptation**. *BMC Genomics* 2010, **11**:307.

50. Camicia F, Paredes R, Chalar C, Galanti N, Kamenetzky L, Gutierrez A, Rosenzvit MC: **Sequencing, bioinformatic characterization and expression pattern of a putative amino acid transporter from the parasitic cestode Echinococcus granulosus**. *Gene* 2008, **411**:1–9.

51. Polzer M, Taraschewski H: **Identification and characterization of the proteolytic enzymes in the developmental stages of the eel-pathogenic nematode Anguillicola crassus**. *Parasitology Research* 1993, **79**:24–7, [http://www.ncbi.nlm.nih.gov/pubmed/7682326].

52. Veuthey AL, Bittar G: **Phylogenetic relationships of fungi, plantae, and animalia inferred from homologous comparison of ribosomal proteins**. *J. Mol. Evol.* 1998, **47**:81–92.

53. Harcus Y, Parkinson J, Fernandez C, Daub J, Selkirk M, Blaxter M, Maizels R: **Signal sequence analysis of expressed sequence tags from the nematode Nippostrongylus brasiliensis and the evolution of secreted proteins in parasites**. *Genome Biology* 2004, **5**(6):R39, [http://genomebiology.com/2004/5/6/R39].

54. Nagaraj SH, Gasser RB, Ranganathan S: **Needles in the EST Haystack: Large-Scale Identification and Analysis of Excretory-Secretory (ES) Proteins in Parasitic Nematodes Using Expressed Sequence Tags (ESTs)**. *PLoS Neglected Tropical Diseases* 2008, **2**(9):e301.

55. Bennuru S, Semnani R, Meng Z, Ribeiro JM, Veenstra TD, Nutman TB: **Brugia malayi excreted/secreted proteins at the host/parasite interface: stage- and gender-specific proteomic profiling**. *PLoS Negl Trop Dis* 2009, **3**:e410.

56. Moreno Y, Geary TG: **Stage- and gender-specific proteomic analysis of Brugia malayi excretory-secretory products**. *PLoS Negl Trop Dis* 2008, **2**:e326.

57. Hewitson JP, Harcus Y, Murray J, van Agtmaal M, Filbey KJ, Grainger JR, Bridgett S, Blaxter ML, Ashton PD, Ashford DA, Curwen RS, Wilson RA, Dowle AA, Maizels RM: **Proteomic analysis of secretory products from the model gastrointestinal nematode Heligmosomoides polygyrus reveals dominance of Venom Allergen-Like (VAL) proteins**. *J Proteomics* 2011, **74**:1573–1594.

58. Anders S, Huber W: **Differential expression analysis for sequence count data**. *Genome Biol.* 2010, **11**:R106.

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

60. Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F, Parvizi B, Tsai J, Quackenbush J: **TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets**. *Bioinformatics* 2003, **19**:651–652, [http://www.ncbi.nlm.nih.gov/pubmed/12651724].

61. Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Muller WE, Wetter T, Suhai S: **Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs**. *Genome Res.* 2004, **14**:1147–1159, [http://www.ncbi.nlm.nih.gov/pmc/articles/PMC419793].

62. Huang X, Madan A: **CAP3: A DNA sequence assembly program**. *Genome Res.* 1999, **9**:868–877, [http://www.ncbi.nlm.nih.gov/pmc/articles/PMC310812].

63. Parkinson J, Whitton C, Schmid R, Thomson M, Blaxter M: **NEMBASE: a resource for parasitic nematode ESTs**. *Nucl. Acids Res.* 2004, **32**(suppl_1):D427–430, [http://nar.oxfordjournals.org/cgi/content/abstract/32/suppl_1/D427].

64. Wasmuth J, Blaxter M: **prot4EST: Translating Expressed Sequence Tags from neglected genomes**. *BMC Bioinformatics* 2004, **5**:187, [http://www.biomedcentral.com/1471-2105/5/187].

65. Bairoch A, Bougueleret L, Altairac S, Amendolia V, Auchincloss A, Argoud-Puy G, Axelsen K, Baratin D, Blatter MC, Boeckmann B, Bolleman J, Bollondi L, Boutet E, Quintaje SB, Breuza L, Bridge A, deCastro E, Ciapina L, Coral D, Coudert E, Cusin I, Delbard G, Dornevil D, Roggli PD, Duvaud S, Estreicher A, Famiglietti L, Feuermann M, Gehant S, Farriol-Mathis N, Ferro S, Gasteiger E, Gateau A, Gerritsen V, Gos A, Gruaz-Gumowski N, Hinz U, Hulo C, Hulo N, James J, Jimenez S, Jungo F, Junker V, Kappler T, Keller G, Lachaize C, Lane-Guermonprez L, Langendijk-Genevaux P, Lara V, Lemercier P, Le Saux V, Lieberherr D, Lima TdeO, Mangold V, Martin X, Masson P, Michoud K, Moinat M, Morgat A, Mottaz A, Paesano S, Pedruzzi I, Phan I, Pilbout S, Pillet V, Poux S, Pozzato M, Redaschi N, Reynaud S, Rivoire C, Roechert B, Schneider M, Sigrist C, Sonesson K, Staehli S, Stutz A, Sundaram S, Tognolli M, Verbregue L, Veuthey AL, Yip L, Zuletta L, Apweiler R, Alam-Faruque Y, Antunes R, Barrell D, Binns D, Bower L, Browne P, Chan WM, Dimmer E, Eberhardt R, Fedotov A, Foulger R, Garavelli J, Golin R, Horne A, Huntley R, Jacobsen J, Kleen M, Kersey P, Laiho K, Leinonen R, Legge D, Lin Q, Magrane M, Martin MJ, O'Donovan C, Orchard S, O'Rourke J, Patient S, Pruess M, Sitnov A, Stanley E, Corbett M, di Martino G, Donnelly M, Luo J, van Rensburg P, Wu C, Arighi C, Arminski L, Barker W, Chen Y, Hu ZZ, Hua HK, Huang H, Mazumder R, McGarvey P, Natale DA, Nikolskaya A, Petrova N, Suzek BE, Vasudevan S, Vinayaka CR, Yeh LS, Zhang J: **The Universal Protein Resource (UniProt) 2009**. *Nucleic Acids Res.* 2009, **37**:D169–174.

66. 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, [http://www.ncbi.nlm.nih.gov/pubmed/10786296].

67. Schmid R, Blaxter ML: **annot8r: GO, EC and KEGG annotation of EST datasets**. *BMC Bioinformatics* 2008, **9**:180, [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2324097].

68. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis GR, Durbin R: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25**(16):2078–2079.

69. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, Ding L: **VarScan: variant detection in massively parallel sequencing of individual and pooled samples**. *Bioinformatics* 2009, **25**:2283–2285.

70. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2009, **25**:1754–1760.

71. 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].

72. 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].

73. Falcon S: **Caching code chunks in dynamic documents**. *Computational Statistics* 2009, **24**(2):255–261, [http://www.springerlink.com/content/55411257n1473414].

74. Sarkar D: *Lattice: Multivariate Data Visualization with R*. New York: Springer 2008, [http://lmdvr.r-forge.r-project.org]. [ISBN 978-0-387-75968-5].

75. Wickham H: *ggplot2: elegant graphics for data analysis*. Springer New York 2009, [http://had.co.nz/ggplot2/book].

Figures

Tables

Table 1 - Sampling, trimming and pre-assembly screening

library	life.st	source.p	raw.reads	lowqal	AcrRNA	eelmRNA	eelrRNA	Cercozoa	valid	valid.span
E1	adult f	Europe R	209325	92744	76403	4835	13112	0	22231	7167338
E2	adult f	Europe P	111746	10903	11213	3613	69	0	85948	24046225
L2	L2 lavae	Europe R	112718	15653	30654	1220	1603	5286	58302	16661548
M	adult m	Asia C	106726	15484	31351	1187	418	0	58286	17424408
T1	adult f	Asia C	99482	7947	24929	7475	514	0	58617	14443123
T2	adult f	Asia W	116366	27683	7233	11741	38	0	69671	20749177
Eel	liver	Taiwan	159370	24298					135072	34482916

Table 2 - assembly classification and summarized contig stats

	lowCA	highCA
total.contigs	26278	13850
rRNA.contigs	555	34
fish.contigs	1775	708
xeno.contigs	2330	742
remaining.contigs	21094	11370
remaining.span	8086686	7971248
non.u.cov	14.665	10.979
cov	2	7
p4e.BLAST-similarity	4355	5663
p4e.ESTScan	8275	3596
p4e.LongestORF	8345	2085
p4e.no-prediction	93	14
full.3p	5905	2714
full.5p	1483	1270
full.l	104	185
GO	2635	3874
EC	966	1492
KEGG	1607	2236
IPR	0	7557
nem.blast	4864	5820
any.blast	5102	6007

Tabel describing summary statistics for contigs from different assembly-categories given in colums as highCA = high credibility assembly; lowCA = low credibility assembly, CA = complete assembly. Rows indicate summary statistics: total.contigs = numbers of total contigs, fish.contigs = number of contigs hitting eel-mRNA or Chordata in NCBI-nr or NCBI-nt (screened out), xeno.contigs = number of contigs with best hit (NCBI-nr and NCBI-nt) to non-eukaryote (screened out), remaining.contigs = number of contigs remaining after this screening, remaining.span = total length of remaining contigs, non.u.cov = non-unique mean base coverage of contigs, cov = unique mean base coverage of contigs, p4e.X = number protein predictions derived in p4e, where X describes the method of prediction (see Methods),

full.3p = number of contigs complete at 3', full.5p = number of contigs complete at 5', GO = number of contigs with GO-anntotation, KEGG = number of contigs with KEGG-anntotation, EC = number of contigs with EC-anntotation, nem.blast = number of contigs with blast-hit to nematode in nr, any.blast = number of contigs with blast-hit to non-nematode (eukaryote non chordate) sequence in NCBI-nr.

Table 3 - `Blast`-hits to protozoan rRNA in pre-assembly screening

sequence.identifier	sequence.identitiy	hsp.length
gi|299836113|gb|GU290110.1|	99.67	599
gi|261259658|emb|FN393299.1|	98.42	310
gi|219524834|gb|EU709197.1|	98.84	515
gi|225216791|gb|FJ176706.1|	90.02	610
gi|238617605|gb|FJ973380.1|	96.26	985
gi|323320595|gb|HQ918172.1|	97.36	606
gi|269993998|dbj|AB520736.1|	100	555
gi|224996440|gb|FJ654272.1|	98.63	145
gi|161015540|gb|EF577167.1|	99.87	759
gi|294831542|dbj|AB526843.1|	96.35	657
gi|225216791|gb|FJ176706.1|	97.68	257

Table 4 - Post-assembly host-screening

	AcrRNA	eelmRNA	eelrRNA	valid_nempep	valid_no_hit
number	953	1159	47	1211	36817
mean coverage	4.14	1.80	1.99	6.79	2.41

Table 5 - Evolutionary conservation

	conserved	novel.in.clade3	novel.in.metazoa	novel.in.Ac
bit.threshold.50	7741	1523	1720	23377
bit.threshold.80	4715	1695	1402	23377

Table 6 - Protein prediction statistics

	p4e->BLAST-similarity	p4e->ESTScan	p4e->LongestORF	no-prediction
plus strand	9701	8005	6393	562
minus strand	4813	5368	5345	0

Table 7 - SNP summry statistics

			pos	in	codon		ti/tv		
	No.SNPs	in.ORF	1	2	3	overall	ins.orf	outs.orf	dn.ds
raw	10458	7153	2310	1819	3024	1.93	2.41	1.25	0.42
h.screened	7514	5425	1710	1322	2393	2.87	3.35	2.01	0.36
p.screened	5112	3628	1149	771	1708	2.41	2.77	1.78	0.3

Table 8 - GO-terms in positively selected

Count	Size	Term	direction
13	45	peptidase activity	Over
7	18	heme-copper terminal oxidase activity	Over
7	18	oxidoreductase activity, acting on a heme group of donors	Over
7	18	oxidoreductase activity, acting on a heme group of donors, oxygen as acceptor	Over
7	18	cytochrome-c oxidase activity	Over
49	283	catalytic activity	Over
13	52	transmembrane transporter activity	Over
9	31	monovalent inorganic cation transmembrane transporter activity	Over
2	2	L-amino acid transmembrane transporter activity	Over
9	33	inorganic cation transmembrane transporter activity	Over
23	117	hydrolase activity	Over
8	29	hydrogen ion transmembrane transporter activity	Over
3	6	ribonucleoprotein binding	Over
13	58	transporter activity	Over
11	47	substrate-specific transmembrane transporter activity	Over
16	77	oxidoreductase activity	Over
1	53	structural constituent of ribosome	Under
7	93	RNA binding	Under
2	44	transition metal ion binding	Under
0	20	protein binding transcription factor activity	Under
0	20	transcription factor binding transcription factor activity	Under
0	20	transcription cofactor activity	Under
13	37	brain development	Over
14	45	central nervous system development	Over
6	12	response to electrical stimulus	Over
3	3	branched chain family amino acid metabolic process	Over
3	3	branched chain family amino acid catabolic process	Over
11	36	ATP synthesis coupled electron transport	Over
11	36	mitochondrial ATP synthesis coupled electron transport	Over
7	18	mitochondrial electron transport, cytochrome c to oxygen	Over
22	101	nervous system development	Over
11	38	oxidative phosphorylation	Over
6	15	response to starvation	Over
12	45	cellular amino acid metabolic process	Over
7	20	positive regulation of cell cycle process	Over
14	58	amine metabolic process	Over
4	8	positive regulation of organelle organization	Over
4	8	spermatid development	Over
4	8	spermatid differentiation	Over
5	12	hindbrain development	Over
5	12	cerebellum development	Over
5	12	metencephalon development	Over
5	12	response to methylmercury	Over
5	12	autophagy	Over
36	203	response to stress	Over
2	2	embryonic body morphogenesis	Over
2	2	xylulose metabolic process	Over
2	2	L-amino acid transport	Over
2	2	neuromuscular process controlling balance	Over
2	2	response to sucrose stimulus	Over

2	2	NADP metabolic process	Over
2	2	response to disaccharide stimulus	Over
2	2	pentose metabolic process	Over
15	66	behavior	Over
8	27	interphase	Over
8	27	interphase of mitotic cell cycle	Over
11	43	electron transport chain	Over
11	43	respiratory electron transport chain	Over
29	156	catabolic process	Over
3	5	positive regulation of mitosis	Over
3	5	positive regulation of nuclear division	Over
13	56	cellular amine metabolic process	Over
20	99	aging	Over
10	39	regulation of cell cycle process	Over
17	81	apoptosis	Over
16	75	regulation of molecular function	Over
13	57	regulation of cell cycle	Over
5	14	mitotic cell cycle G1/S transition DNA damage checkpoint	Over
5	14	sleep	Over
4	10	cellular amino acid catabolic process	Over
10	41	reproductive structure development	Over
3	6	microtubule organizing center organization	Over
3	6	RNA catabolic process	Over
3	6	centrosome organization	Over
8	30	muscle organ development	Over
11	47	cellular respiration	Over
13	59	energy derivation by oxidation of organic compounds	Over
7	25	regulation of catabolic process	Over
5	15	signal transduction in response to DNA damage	Over
5	15	G1/S transition of mitotic cell cycle	Over
5	15	regulation of G1/S transition of mitotic cell cycle	Over
5	15	mitotic cell cycle G1/S transition checkpoint	Over
5	15	G1/S transition checkpoint	Over
5	15	DNA damage response, signal transduction by p53 class mediator	Over
5	15	regulation of cellular amine metabolic process	Over
6	20	response to copper ion	Over
24	131	cellular catabolic process	Over
4	11	imaginal disc development	Over
4	11	amine catabolic process	Over
4	11	skeletal muscle organ development	Over
11	49	mRNA metabolic process	Over
2	3	nuclear mRNA cis splicing, via spliceosome	Over
2	3	germ cell migration	Over
2	3	positive regulation of mitotic metaphase/anaphase transition	Over
2	3	mitotic centrosome separation	Over
2	3	oligosaccharide catabolic process	Over
2	3	spliceosomal conformational changes to generate catalytic conformation	Over
2	3	amino acid transport	Over
2	3	negative regulation of reproductive process	Over
2	3	centrosome duplication	Over
2	3	centrosome separation	Over

2	3	protein tetramerization	Over
2	3	protein homotetramerization	Over
15	201	gene expression	Under
1	57	cellular protein complex disassembly	Under
1	57	macromolecular complex disassembly	Under
1	57	protein complex disassembly	Under
1	57	cellular macromolecular complex disassembly	Under
1	55	pancreas development	Under
1	55	endocrine pancreas development	Under
1	55	endocrine system development	Under
1	55	viral genome expression	Under
1	55	viral transcription	Under
8	131	transcription	Under
1	54	translational termination	Under
4	89	translation	Under
2	66	cellular component disassembly	Under
2	66	cellular component disassembly at cellular level	Under
14	178	cellular macromolecule biosynthetic process	Under
22	243	biosynthetic process	Under
22	240	cellular biosynthetic process	Under
15	181	macromolecule biosynthetic process	Under
2	57	viral reproductive process	Under
2	57	viral infectious cycle	Under
0	26	positive regulation of intracellular protein kinase cascade	Under
1	38	positive regulation of response to stimulus	Under
0	24	oocyte differentiation	Under
0	23	oocyte development	Under
0	23	cation transport	Under
0	22	positive regulation of MAPKKK cascade	Under
24	234	growth	Under
4	7	small nuclear ribonucleoprotein complex	Over
31	164	mitochondrion	Over
2	2	Cajal body	Over
2	2	U5 snRNP	Over
2	2	U4/U6 x U5 tri-snRNP complex	Over
17	80	mitochondrial part	Over
3	6	nuclear speck	Over
5	15	nuclear body	Over
14	65	mitochondrial membrane	Over
14	66	mitochondrial envelope	Over
2	3	clathrin sculpted vesicle	Over
2	3	plasma membrane respiratory chain complex I	Over
2	3	plasma membrane respiratory chain	Over
2	3	basement membrane	Over
2	3	plant-type cell wall	Over
0	37	large ribosomal subunit	Under
0	35	cytosolic large ribosomal subunit	Under
28	280	nucleus	Under
19	201	non-membrane-bounded organelle	Under
19	201	intracellular non-membrane-bounded organelle	Under
4	71	nucleolus	Under

3	60	cytosolic ribosome	Under
1	38	plastid	Under
4	68	cytosolic part	Under
1	36	chloroplast	Under
5	73	ribosome	Under

Table 9 - Correlation between read-counts in 454-libraries and solexa-tags

Table 9 a - analysing all TUGs

	solexa.tags	E1	E2	L2	M	T1	T2	all.reads
solexa.tags	1.000	0.257	0.356	-0.165	0.320	0.233	0.127	0.315
E1	0.257	1.000	0.154	-0.076	0.229	0.145	0.071	0.254
E2	0.356	0.154	1.000	-0.246	0.126	0.134	0.090	0.295
L2	-0.165	-0.076	-0.246	1.000	-0.181	-0.237	-0.266	0.127
M	0.320	0.229	0.126	-0.181	1.000	0.077	0.016	0.278
T1	0.233	0.145	0.134	-0.237	0.077	1.000	0.029	0.210
T2	0.127	0.071	0.090	-0.266	0.016	0.029	1.000	0.350
all.reads	0.315	0.254	0.295	0.127	0.278	0.210	0.350	1.000

Table 9 b - analysing good-category contigs only

	solexa.tags	E1	E2	L2	M	T1	T2	all.reads
solexa.tags	1.000	0.371	0.528	-0.196	0.450	0.393	0.199	0.385
E1	0.371	1.000	0.324	-0.064	0.366	0.307	0.172	0.312
E2	0.528	0.324	1.000	-0.280	0.324	0.411	0.197	0.373
L2	-0.196	-0.064	-0.280	1.000	-0.191	-0.242	-0.358	0.084
M	0.450	0.366	0.324	-0.191	1.000	0.264	0.083	0.347
T1	0.393	0.307	0.411	-0.242	0.264	1.000	0.156	0.324
T2	0.199	0.172	0.197	-0.358	0.083	0.156	1.000	0.437
all.reads	0.385	0.312	0.373	0.084	0.347	0.324	0.437	1.000

Table 9 c - analysing good-category contigs surely from A. crassus only

	solexa.tags	E1	E2	L2	M	T1	T2	all.reads
solexa.tags	1.000	0.378	0.528	-0.129	0.445	0.417	0.174	0.433
E1	0.378	1.000	0.318	-0.001	0.352	0.318	0.164	0.343
E2	0.528	0.318	1.000	-0.194	0.284	0.429	0.157	0.435
L2	-0.129	-0.001	-0.194	1.000	-0.110	-0.145	-0.268	0.004
M	0.445	0.352	0.284	-0.110	1.000	0.259	0.040	0.386
T1	0.417	0.318	0.429	-0.145	0.259	1.000	0.158	0.399
T2	0.174	0.164	0.157	-0.268	0.040	0.158	1.000	0.506
all.reads	0.433	0.343	0.435	0.004	0.386	0.399	0.506	1.000

Additional Files

File A_crassus_contigs_full.csv lists all data computed on the contig level, including sequences (raw, coding, imputed). File A_crassus_contigs_readable.csv lists only the metadata not including sequences.