**The transcriptome of the swim bladder nematode *Anguillicola crassus*: Resources for an alien parasite**

tt.10.11

**I have edited this rtf version, but it is quite different from the pdf so... If I am to participate in the process of editing, you will have to work on a way of making it accessible to me. MB 20111019**

**Authors**

Emanuel

Horst

someone from GenePool 454

Stephen Bridgett

Mark

**Affiliations**

**Keywords**

**Declaration of conflicts**

**Abstract**

**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.

**Methods**

***depending on where this gets sent, the methods come here or at the end, and are complete or summarised. i suggest writing a complete methods ssection here and trimming it/moving it as required****

*** these sections copied from results are methods, not results

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

To make this evaluation more specific we only sequences hitting whith the over 50% of their length with an identity >70% to label the TUGs as “contamination” accordig to the respective database-hit.

Two kinds sequences were labeled “valid” *A. crassus* sequences: Valid due to evidence of nematode origin (hitting nempep4), and valid due to absence of evidence of host-origin.

To elucidate possible xenobiotic sequences (comprising unexpected off-target data) we searched our complete dataset against both NCBI-nt and NCBI-nr. Figure (phylum_plots) gives an overview of the number of top-hits to different phyla sorted by kingdom.

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

In addition to the protein-sequence used for annotation we obtained a corrected nucleotide-sequence using the nucleotide equivalent of the protein-prediction. This was necessary as prot4Est joins high-scoring pairs from blast-searces introducing gaps (or unknown fill-bases respectively) relative to the raw nucleotide sequence if needed to obtain the correct frame for the ORF. A predicted ORF was defined at complete if it was starting with an “atg”-start-codon, having at least 3 non-coding bases 5’ of the start-codon and having a “tag” (2122 observations), “taa” (4989) or “tga” (3026) stop-codon 3’ of the ORF.

To further screen those SNPs we utilized observations made in combination with protein prediction, i.e. with the inferred open-reading frame. As noted above SNPs at unknown fill-bases had to be disregarded, but the imputation of a continuous corrected ORF had the advantage of making the coordinates for the ORF straight forward to use.

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

To asses the effect of a mutation at a given base we used an idea of Mark Blaxter (unpublished, some information at [26]), to classify every base as to its “response” to mutation. We used this classification to compute the number of synonymous and non-synonymous sites in a given TUG. In total 8847926 non-synonymous and 2739223 sites were found in ORFs, of these 858749.67 non-synonymous sites and 259935.33 synonymous sites were covered at least 8-fold, and thus available for SNP-calling.

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

Based on a slight overrepresentation of SNPs on first- and second-positions in the ORF and on a little bigger than average excess of non-synonymous polymorphisms in SNPs with a percentage of the minority allele smaller than 7% (see figure snp_pos_eff_plots.png a and b) we examined exclusion of SNPs at this percentage threshold.

Evaluation of an exclusion of SNPs based on coverage at the SNP-site, did not seem necessary (see figure snp_pos_eff_plots.png c and d).

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

Table 7 gives on overview of how the basic SNP-statistics described above changed with screening of the candidate SNPs. The change of ti/tv back to lower values, when SNPs were screened based on a 7% coverage threshold, left the benefit of this screening-step questionable.

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

However, calculating dn/ds on a per contig base, the screening based on percentage threshold of the minority-allele showed it's benefits: figure dn_ds_scales.png b shows how for the unscreened SNP-set dn/ds scaled with the coverage of a contig. This correlation was not longer present if the percentage screening was used (see figure dn_ds_scales.png d, the linear model had a slightly negative non-significant slope). For both the screened and unscreened sets of SNPs there was a significant slope for the number of SNPs in a contig predicting dn/ds (see figure dn_ds_scales.png a and c).

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

**Nematode samples, RNA extraction, cDNA synthesis and Sequencing**

*A. crassus* from *An. 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 **** For further details see the supplementary methods. The two assemblies were combined into one using Cap3 at default settings [62] and contigs labeled by whether they derived from both assemblies ort 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 [63, 25]. 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, 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) was 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 {reference].

For deepSAGE NlaIII-tag-sequencing, total RNA was prepared as described above from a SEX OF? 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. ****is this valid for two distinct methods of count generation where one should be reads per kilobase of transcript per million mapped and the orther should be counts per transcript???*** Spearman correlation coefficients were calculated for raw counts.

**General coding methods**

The bulk of analyses (unless otherwise cited) presented were carried out in R [71] using custom scripts. We used methods 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].

## Results

### Sampling *A. crassus*

****Describe the samples taken - lifecycle stage, population source and host source and reference table 1. Mention the issues of host contamination, and thus mention the eel transcriptome set.***

### Sequencing, trimming and pre-assembly screening

A total of 756,363 raw sequencing reads were generated for *A. crassus* (Table 1). These were trimmed for base call quality, and filtered by length to give 585,949 high-quality reads (XX bases). We then screened these reads for contamination by host (30,071 matched previously sequenced eel genes or our own *An. anguilla* 454 transcriptome), and for sequences derived from ribosomal RNAs of the nematode (202,823 reads matched large or small subunit nuclear ribosomal RNA sequences of . *crassus*) (Table 1). In addition to fish mRNAs, we identified (and removed) a few reads in the library derived from the L2 nematodes that had significant similarity to cercozoan (likely parasite) ribosomal RNA genes (Table 1). ***Mention eel data****

**** Table 1 should show for each library: source population, source stage, source host, number raw reads, number filtered for lowQ, num filtered for host, num filtered for rRNA, num filtered for cercozoa, remaining number, remaining span; add a column for the eel 454 data (obviously not reporting filtering for 'contamination') ****

### Transcriptome Assembly

We assembled the remaining 352,955 *** minus the cercozoan ones*** reads using the combined assembler strategy [21] and Roche 454 GSassembler (version xyz) and MIRA (version XYZ) [MiRA reference]. From this we derived 13,851 contigs that were supported by both assembly algorithms, 3,745 contigs only supported by one of the assembly algorithms and 22,591 singletons that were not assembled by either approach (****Table 2 or Figure 1****). 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 are of medium credibility (medCA), 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 medCA 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).

****I havent edited the next bit because I dont know what the difference between per-base coverage and per-base unique coverage **** The mean per-base coverage of the TUGs was 12.84 and the mean per-base unique coverage was 4.62. The mean per-base coverage of the quality contigs was lower with 10.98, but the unique coverage higher with 6.84. This indicates a higher amount of redundancy in the full assebmly set compared to the good-quality set. The distribution of mean per base coverages for single contigs or TUGs is given in Figure (coverage_per_contig.png)

Figure (coverage_plots) shows the relationship between, read-number in the assemblies and coverage. ***coverage is read number per contig? should it be (sum read length)/(contig length) - I cant comment as the units are not defined****

We screened the complete assembly for residual host contamination, and identified 1,816 contigs that had higher, significant similarity to eel (and fish) sequences (our 454 ESTs and EMBLBank fish 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 ***describe what was found***

***Now we want Table 2 describing the three datasets, with columns highCA, medCA, lowCA, complete assembly, and rows initial contig numbers, # fish contigs, # other contigs, # remaining contigs, span of remaining contigs, number of reads total in contigs, mean base coverage of contigs, number protein predictions derived in p4e, number complete at 5', number complete at 3", number with GO, number with KEGG, number with EC, number with any blast to nematode, number with any blast to anything***

Our assembly thus has 38371 [what is the number? - given as 31893 , 36166 later] TUGs spanning XY Mb (of which XYZ are highCA- or medCA-derived, and span XY Mb) that are likely to derive from of *A. crassus*.

**Protein prediction**

[was this done for the fish and other biont contigs? if so remove their numbers from this section]

For 39625 TUGs a protein was predicted using prot4EST [reference] (Table 2). The full open reading frame was obtained for in 414 TUGs, while for 3304 the 5' end and for 10178 the 3' end was complete. In 15988 TUGs the corrected sequence with the imputed ORF was slightly changed compared to the raw sequence.

**Annotation**

We used annot8r [reference] to assign gene ontology (GO) terms for 9569 TUGs, Enzyme Commission (EC) numbers for 3741 TUGs and Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathway annotations for 5990 TUGs (Table 2). Nearly one third (10274) of the *A. crassus* TUGs were annotated with at least one identifier, and 2854 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 *Brugia malayi*. The partial transcriptome of *A. crassus* shows a remarkably similar distribution **** this is not a scientific statement... yes they are similar but is it remarkable? iis it sgnificantly different? is C. elegans similar? what does this tell us about the completeness of our data?**** of GO-slim terms to the full proteome *B. malayi* (figure go_bm_com.png). We inferred presence of signal peptide cleavage sites in the predicted protein sequence using SignalP [28] predicted 1060 signal peptide cleavage sites and 83 signal peptides with a transmembrane signature. ***and compared this proportion with Brugia, or ...***

**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****

**Identification of single nucleotide polymorphisms**

We called single nucleotide polymorphisms (SNPs) on the 1,412,806 bases of the highCA and medCA contigs that had coverage of more then 8-fold available using which software [reference]. 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 XXXX 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 medCA 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 category, “amino acid transmembrane transporter activity” and “peptidase activity” 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.

****what about Taiwan versus Europe??****

**Differential gene expession**

***I have edited some of this but I am very unclear what is being tested and why: that nematodes differ in gene expression? that males differ from females? There are also issues from the methods in how you calculated read and thus expression values. The low correlation is so ringed around with caveats including low sequencing depth per sample that I doubt the usefulness of these analyses. This also places the usefulness of the single NlaIII sequence set in doubt. Its unclear from the text how this was analysed - were all tags mapping to a transcript counted, or only those mapping to the 3'end-most NlaIII site? if all, there are issues with pseudocounts... I think this needs to be more clearly contextualised, and this will in turn improve the analyses****

We used both the Roche 454 transcriptome data and a deepSAGE sample from an European nematode to estimate expression levels of each transcript. We werre able to map 97% (341,285) of the Roche 454 reads to the complete assembly ***I am assuming that this excludes rubbish***, 75% (264440) of which mapped uniquely. We generated 6.2 million NlaIII deepSAGE tags, and were able to map 82% (5.1 million) of these to the complete assembly. (559824 unique) ***WHAT DOES THIS MEAN?**** For each TUG, we estimated expression by summing the number of Roche 454 reads per sequencing library and the number of NlaIII tags. Correlation coefficients of estimated expression levels between the Roche 454 libraries, and between Roche 454 data and deepSAGE data, were generally low. However analyses limited to the highCA and medCA TUGs had improved correlation coefficients both between libraries and between Roche 454 data and deepSAGE tags (Table 9). Correlations between library T2 and other Roche 454 libraries, as well as with deepSAGE counts were lower than between other libraries. We note that this library was sequenced with the FLX standard rather than Titanium chemistry.

Despite the lack of replicates for male nematode we were able to identify 25 TUGs that were significantly over-expressed in male nematodes. All these TUGs were nearly exclusively expressed in males *** biology????****. For the L2 library we identified many apparently differentially expressed TUGs. These were enriched in sequences that may derive from non-nematode, non-eel sources: only 57 had best hits to Metazoa and only 6 to Nematoda. No TUGs showed significant differential expression between nematodes isolated from *An. anguilla* and those isolated from *An. japonica*, even if the data analysed were limited to highCA and medCA TUGs (though we noted that he lowest adjusted p-values for the highCA plus medCA data were around 0.4, while on the complete assembly only adjusted p-values above 0.8 could be obtained, suggesting a lack of power in the data rather than lack of signal).

**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.

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

edited to here

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

Author contributions

Acknowledgments

***Stephen should be an autor, no? **** We thank Stephen Bridgett from gene-pool sequencing service for general help with raw data and for cross-contamination screening of libraries.

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\s\do5(p)arsing\s\do5(g)roup]].

[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\s\do5(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]].

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.