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: *Anguillicola crassus* is an ecologically, economically and evolutionary interesting nematode. It has been introduced from Asia, where it parasitises the Japanese eel *Angilla japonica*, to Europe 30 years ago. Today it infects stocks of the endangered, commercially exploited European eel *Anguilla anguilla*, permitting and necessitating research in a newly established host-parasite system. Furthermore phylogenetics places *A. crassus* at a key position for the emergence of parasitism, basal to one of the major clades of parasitic nematodes.

Results: After extensive screening of 756.363 raw pyrosequencing reads, we assembled 353.055 into 11.371 contigs spanning 7.971.550 bases and additionally obtained 21.147 singleton and lower quality contigs spanning 8.095.986 bases. We obtained annotations for ca. 60% of the contigs and 40% of the tentatively unique genes (TUGs) confirming the high quality of especially the contigs. We found an overabundance of predicted signal peptide cleavage sites in sequence conserved in Nematoda and novel in *A. crassus*. We identified 5112 high quality Single nucleotide polymorphisms (SNPs) and suggest 199 of them as most suitable markers for population-genetic studies. GO-term analysis identified enrichment of peptidases and subunits of the respiratory chain to be enriched for transcritps under positive selection. Comparing male and female as well as Asian and European *A. crassus* we developed a method for future work with this transcriptome as a reference in mapping

experiments.

Conclusions: The transcriptome of *A. crassus* is a basis for molecular research on this important species. It furthermore provides unique insights into the evolution of parasitism in the Spirurina.

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 [2]. 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 [3], where it was most likely introduced through the live-eel trade [4, 5], *A. crassus* has spread rapidly through populations of its newly acquired host [6]. At the present day it is found in all *An. anguilla* populations except those in Iceland [7]. *A. crassus* can be regarded as a model for invasive parasite introduction and spread [8].

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 [9], where high prevalence (above 70% [10]) and high infection intesities have been reported throughout the newly colonized area [11]. The virulence of A .crassus in this new host has been attributed to an inadequate immune response in *An. anguilla* [12]. 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 [16].

The genus Anguillicola is placed in the nematode suborder Spirurina (clade III sensu [17]) [18, 19]. 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 a programme 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

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 *An. japonica* was also processed. Emulsion PCR was performed for each cDNA library according to the manufacturer's protocols (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-*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` [22] (with parameters `-minscore 20 -minmatch 10`). `Seqclean` [23] 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, a unpublished assembly of *An. anguilla* Sanger dideoxy sequenced expressed sequence tags (made available to us by Gordon Cramb, University of St Andrews) and transcripts from EeelBase [24] a publicly available 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]: Two assemblies were generated, one using `Newbler v2.6` [20] (with parameters `-cdna -urt`), the other using `Mira v3.2.1` [25] (with parameters `-job=denovo,est,accurate,454`). The resulting two assemblies were combined into one using `Cap3` [26] at default settings and contigs were labeled by whether they derived from both assemblies or one assembly only (for a detailed analysis of the assembly categories see the supporting Methods file).

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 [27, 28]. 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. TUGs with a best hit to non-Metazoans and to Chordata within Metazoa were additionally excluded from further analysis.

Protein prediction and annotation

Protein translations were predicted from the contigs using `prot4EST` (version 3.0b) [29]. Proteins were predicted either by joining single high scoring segment pairs (HSPs) from a `BLAST` search of uniref100 [30], or by `ESTscan` [31], using as 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 protein 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 Encyclopedia of Genes and Genomes (KEGG) terms were inferred for these proteins using `Annot8r` (version 1.1.1) [32], using the annotated sequences available in uniref100 [30]. 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` [33] was used to predict signal peptide cleavage sites and signal anchor signatures for the *A. crassus*-transcriptome and similarly again for the proteomes of the tow model-worms. Additionally `InterProScan` [34] (command line utility `iprscan` (version 4.6) with options `-cli -format raw -iprlookup -seqtype p -goterms`) was used to obtain domain based annotations for the high credibility assembly (highCA) derived contigs.

We recorded the presence of a lethal rnai-phenotype in the *C. elegans* ortholog of each TUG using the biomart-interface [35] to wormbase v. 220 through the R-package `biomaRt` [36].

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` [37] (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` [38], discarding reads mapping to multiple regions. `VarScan` [39] (`pileup2snp`) was used with default parameters on pileup-files to output lists of single nucleotide polymorphisms (SNPs) and their locations. For enrichment analysis of GO-terms we used the R-package `GOstats` [40].

Using `Samtools` [38] (`mpileup -u`) and `Vcftools` [41] (`view -gcv`) we genotyped individual libraries for the list of previously found overall SNPs. Genotype-calls were accepted at a phred-scaled genotype quality threshold of 10. In addition to the relative heterozygosity (number of homozygous sites/number of heterozygous sites) we used the R package `Rhh` [42] to calculate internal relatedness [43], homozygosity by

loci [44] and standardized heterozygosity [45] from these data.

Using 1000 bootstrap replicates we confirmed the significance of heterozygote-heterozygote correlation by analyzing the mean and 95% confidence intervals from 1000 bootstrap replicates estimated for all measurements.

Gene-expression analysis

Read-counts were obtained from the `bam`-files generated also for genotyping using the R-package `Rsamtoools` [46]. TUGs with less than 48 reads over all libraries were excluded from analysis, as diagnostic plot (not shown) indicated a lack of statistical power for lower overall expression. We used the R-package `DESeq` [47] to assess statistical significance of differences in counts according to groups of libraries. Additionally we collapsed TUGs by their orthologous assignment in *C. elegans* and *B. malayi*. We used the sums of counts for these orthologous-groups to asses the influence of mapping to our potentially fragmented reference. For both model-nematodes fold-change and p-values were obtained the same way than for the contigs and merged with these.

General coding methods

The bulk of analysis (unless otherwise cited) presented in this paper was carried out in `R` [48] using custom scripts. We used a method provided in the R-packages `Sweave` [49] and `Weaver` [50] for "reproducible research" combining `R` and LaTeXcode in a single file. All intermediate data files needed to compile the present manuscript from data-sources are provided upon request. For visualisation we used the R-packages `ggplot2` [51] and `VennDiagram` [52].

Results

Sampling A. crassus

One female worm and one male worm were sampled from an aquaculture with height infection loads in Taiwan. An additional female worm was sampled from a stream with low infection pressure adjacent to the aquaculture. All these worms were parasitising endemic *An. japonica*. A female worm and pool of L2 larval stages were sampled from *An. anguilla* in the river Rhine, one female worm from a lake in Poland. All adult worms were filled with large amounts of host-blood, therefore we anticipated abundant host-contamination in sequencing data and decided to sequence a liver sample of an uninfected *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 169863104 bases). In the eel dataset 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 or our own *An. japonica* 454 transcriptome, which had been assembled into 10639 mRNA contigs. (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) [25]. 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 transcriptomes). In the highCA parsimony and low redundancy is prioritized, while in the complete assembly (highCA plus lowCA) completeness is prioritized. 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 3441 TUGs that had higher, significant similarity to eel (and chordate) sequences (our 454 ESTs and EMBLBank Chordata proteins) than to nematode sequences [28].

Given our prior identification of cercozoan ribosomal RNAs, we also screened the complete assembly for contamination with other transcriptomes.

1153 TUGs were found mapping to Eukaryota outside of the kingdoms Metazoa, Fungi and Viridiplantae. These hits included a wide range of Protists ranging from Apicomplexa (mainly Sarcocystidae, 28 hits and Cryptosporidiidae 10 hits) over Bacillariophyta (diatoms, mainly Phaeodactylaceae, 41 hits) and

Phaeophyceae (brown algae, mainly Ectocarpaceae, 180 hits) and Stramenopiles (Albuginaceae, 63 hits) to Kinetoplasitda (Trypanosomatidae, 26 hits) and Heterolobosea (Vahlkampfiidae, 38 hits).
Additionally we found 298 TUGs with hits to fungi (e.g Ajellomycetaceae, 53 hits) and 585 TUGs with hits to plants.
Hits outside the Eukaryota were mainly to Bacteria (825 hits) and within those mostly to members of the Proteobacteria (484 hits). No hits were found to Wolbachia or related Bacteria known as symbionts of nematodes and arthropods. 9 TUGs were hitting sequence from Viruses and 8 from Archaea.
We excluded all TUGs with best hits outside Metazoa and 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

For 32411 TUGs a protein was predicted using prot4EST [29] (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. elegans* for 9554 TUGs, from *B. malayi* for 9662 TUGs, from nempep [27, 28] for 11617 TUGs and with uniprot proteins for 11113 TUGs.
We used annot8r [32] to assign gene ontology (GO) terms for 6509 TUGs, Enzyme Commission (EC) numbers for 2458 TUGs and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway annotations for 3844 TUGs (Table 2). Additionally 5125 highCA derived contigs were annotated with GO terms through `InterProScan` [34]. 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 1).
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* (Figure 2).
Correlation shows the occurrence 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 to each other (0.91) are less similar in the occurrence of terms than the two parasites.

We inferred presence of signal peptide cleavage sites in the predicted protein sequence using `SignalP` [33]. 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).

We inferred the presence of a lethal rnai phenotype in the orthologous annotation of *C. elegans*. For 257 TUGs a non-lethal phenotype was inferred for 6029 TUGs a lethal phenotype.

Evolutionary conservation

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 (Table 3).

Roughly a third and a quarter of the higCA derived contigs were categorized as conserved across kingdoms at a bitscore threshold of 50 and 80, respectively. Roughly half or 3/5 of the these contigs were identified as novel in *A. crassus*.

The remaining higCA contigs spread across intermediate relatedness-levels. More sequences were categorised as novel at the phylum level (Nematoda) compared to kingdom and clade III level and the number of contigs at intermediate relatedness-levels was roughly consistent for the two bitscore thresholds. The latter points about intermediate conservation levels were also true, when all TUGs were analysed. The numbers of TUGs categorised at these intermediate levels roughly doubled. In contrast, the proportion of additional conserved lowCA TUGs is small compared to additional TUGs categorised as novel in *A. crassus*, mirroring the higher amount of erroneous sequence.

Proteins predicted to be novel to Nematoda and novel in *A. crassus* were significantly enriched in signal peptide annotation compared to conserved proteins, proteins novel in Metazoa and novel in clade III (Fisher's exact test $p<0.001$; Figure 5) .

The proportion of lethal rnai phenotypes was significantly higher for orthologs of conserved TUGs (97.23%) than for orthologs of TUGs not conserved (94.65%) across kingdoms ($p<0.001$, Fisher's exact test).

Identification of single nucleotide polymorphisms

We called single nucleotide polymorphisms (SNPs) on the 1099419 bases of the TUGs that had coverage of more then 8-fold available using `VARScan` [39]. We excluded SNPs predicted to have more than 2 alleles or

that mapped to an undetermined (N) base in the reference, and retained 10458 SNPs. The ratio of transitions (ti; 6890) to transversion (tv; 3568) in this set was 1.93 . Using the prot4EST predictions and the corrected sequences, 7153 of the SNPs were predicted to be inside an ORF, with 2310 at codon first positions, 1819 at second positions and 3024 at third positions. As expected ti/tv inside ORFs (2.41) was higher than outside ORFs (1.25). The ratio of synonymous polymorphisms per synonymous site to non-synonymous polymorphisms per non-synonymous site (dn/ds) was 0.42. 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 [53], 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 homopolymer runs and proximity thresholds (see Figure 3). 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 (Figure 4). Our filtered SNP dataset includes 5112 SNPs. We retained 4.65 SNPs per kb of contig sequence, with 8.37 synonymous SNPs per 1000 synonymous bases and 2.4 non-synonymous SNPs per 1000 non-synonymous bases. A mean dn/ds of 0.231 was calculated for the 859 TUGs (762 highCA-derived 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 in SignalP showed a trend towards higher dn/ds values than TUGs without signal peptide cleavage sites ($p = 0.187$; two sided Mann-Whitney-test)

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 4). Within the molecular function category, "peptidase activity" was the most significantly overrepresented term and had 13 TUGs supporting the overrepresentation. The

highlighted 13 peptidases annotated with eleven unique orthologs in *C. elegans* and *B. malayi*. The term "structural constituent of ribosome" was underrepresented.

While the biological process and cellular compartment categories provide less information for a nematode (highlighting e.g. brain or pancreas development), underrepresented terms in both were connected to ribosomal proteins, validating the analysis for the molecular function category.

Other overrepresented terms abundant over categories pointed to subunits of the respiratory chain e.g. "heme-copper terminal oxidase activity" and "cytochrome-c oxidase activity" in molecular function and "mitochondrion" in cellular compartment.

At both bitscore thresholds contigs novel in clade III and novel in *A. crassus* had a significantly higher dn/ds than other contigs (novel.in.metazoa - novel.in.Ac, 0.005 and 0.015; novel.in.nematoda - novel.in.Ac, 0.005 and 0.002; novel.in.nematoda - novel.in.clade3, 0.207 and 0.045; comparison, p-value from bitscore of 50 and p-value from bitscore of 80, Nemenyi-Damico-Wolfe-Dunn test, given only for significant comparisons; Figure 5).

Orthologs of *C. elegans* transcripts with lethal rnai phenotype are expected to evolve under stronger selective constraints. Indeed the values of dn/ds showed a non-significant trend towards lower values in TUGs with orthologs with a lethal phenotype compared to a non-lethal phenotypes (p=0.138, two-sided U-test).

SNP markers for single worms

We used `Samtools` [38] and `Vcftools` [41] to call genotypes in single worms (adult sequencing libraries). This resulted in 199 informative sites in 152 contigs, where two alleles were found in at least one assured genotype at least in one of the worms.

Internal relatedness [43], homozygosity by loci [44] and standardised heterozygosity [45] were all highlighting the Taiwanese worm from the wild population (sample T1) as the most and the European worm from Poland (sample E2) as the least heterozygous individual. The other worms had intermediate values between these two extremes.

We confirmed the genome-wide significance of these estimates using heterozygosity-heterozygosity correlation [42]. These tests confirmed the representativeness of the 199 SNP-markers for the whole genome in population genetic studies ($\mu = 0.78$, ci_l=0.444; $\mu = 0.86$ and $ci_l = 0.596$; $\mu = 0.87$ and ci_l= 0.632; mean and lower bound of 95% confidence intervals from 1000 bootstrap replicates for internal relatedness, homozygosity by loci and standardised heterozygosity). Using a higher number of genotyped individuals

these markers would allow to asses the amount of inbreeding in populations of *A. crassus*..

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 [54], and an analysis of lake sturgeon tested and rejected hypotheses of horizontal gene-transfer when xenobiont sequences were identified [55]. A custom pipeline for transcriptome assembly from pyrosequencing reads [56] proposed the use of EST3 [57] 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 [58], 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. [59]), 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. [60]).

***START: I moved the chunck below up (from a paragraph you did not correct last time) to have annotation before SNPs

We were able to obtain high-quality annotations for a large set of TUGs: For 40% of the complete assembly and 60% of our highCA assembly `BLAST`-based annotations could be obtained. 45% of the contigs in the highCA assembly were additionally decorated with domain-based annotations through `InterProScan` [34]. Comparison with complete protein sequence from the genomes of *B. malayi* and *C. elegans* showed a

remarkable degree of agreement regarding the occurrence of terms in the two parasitic worms. This agreement was higher than with the free living nematode *C. elegans* and even the two genome-sequencing-derived proteomes showed less agreement with each other than the filarial parasite with our dataset. This implies that our transcriptome is truly a representative partial genome [61] of a parasitic nematode.

Analysis of conservation identified more sequence novel in Nematode than in the eukaryote kingdom or in clade III this is in agreement with prevalence of genic novelty in the Nematoda [62]. Furthermore the basal position of *A. crassus* in clade III could be leading to most novelty in the clade not being shared with *A. crassus*.

TUGs predicted to be novel in the phylum Nematoda and novel to *A. crassus* contained the highest proportion of signal-positives. This confirms observations made in a study on *Nippostrongylus brasiliensis* [63], where signal positives were reported as less conserved. Interestingly enrichment of signal sequence bearing TUGs in our dataset was constrained to sequences novel in nematodes and *A. crassus* (i.e. not to the level of clade III). This may be explained, whit two different hypotheses involving the basal position of *A. crassus*: First the signal positives shared with all nematodes could be conserved molecules not excreted by parasites. A different class of secreted/excreted molecules with prominent role in host parasite interactions would not have arisen early in the evolution of parasitism in clade III - or be too fast-evolving - and thus be detected as specific to deeper sub-clades (i.e. to *A. crassus* in our dataset). A second explanation would be, that orthologs of excreted parasite-specific genes could be among those shared with other nematodes and the fewer shared with clade III implying a predisposition to parasitism outside of the Spirurina or even the convergent evolution of secreted molecules in other parasitic nematodes. However analysis of dn/ds (see below) across conservation categories favours the first hypothesis, as it identifies a higher amount of positive selection in TUGs novel to clade III and *A. crassus* than to nematodes.

***End: I moved the above chunck up to have annotation before SNPs

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 [53]. The value of 1.93105381165919 (1.25 outside, 2.41 inside ORFs) is in good agreement with the overall ti/tv of humans (2.16 [64]) or *Drosophila* (2.07 [65]). 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 0.42 to 0.231 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 [66], 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 biased by depth (i.e. coverage; see [67] and [58]) could be avoided.

*** you corrected up to her last time

Also in comparison with published intra-species values of dn/ds our final estimate of seems plausible: in transcripts from the female reproductive tract of *Drosophila* dn/ds was 0.15 [68] and 0.21 in the male reproductive tract [69] (although for ESTs specific to the male accessory gland were shown to have a higher dn/ds of 0.47). A pyrosequencing study in the parasitic nematode *Ancylostoma canium* [70] reported dn/ds of 0.3.

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 [68] instead of the classical threshold of 1 [71]. The use of this threshold for positive selection led to the identification of over-represented of GO-term highlighting very interesting transcripts:

13 peptidases under positive selection (from 43 with a dn/ds obtained) meant an enrichment in the category. All 13 have different orthologs in *B. malayi* and *C. elgans* and are conserved across kingdoms. Despite their conservation peptidases are thought to 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 utilised by the tissue-dwelling L3 stage to penetrate host tissue and an aspartyl proteinase thought to be a digestive enzyme in adults [72]. The 13 proteinases under positive selection could be the targets of the adaptive immunity developed against *A. crassus* [13, 73], which is often only elicited against subtypes of larvae [74].

The under-representation of ribosomal proteins (term “structural constituent of ribosome”) in positive selected contigs is in good agreement with the notion that ribosomal proteins are extremely conserved across kingdoms [75] and should be under under strong negative selection.

Genotyping of individual worms identified a set of 199 SNPs with highest credibility and a high information content for population-genetic studies. Levels of genome-wide heterozygosity found for the 5 adult worms examined in our study are in agreement with microsattelite data [76] showing reduced heterozygosity in European populations of *A. crassus*.

We were able to use the DESeq [47] to report transcripts significantly differing in expression between male and female worms. This was possible for male worms, despite the fact that no replicated samples were obtained. However only over-expression in the non-repeated samples could be detected, as obviously lack of expression in one sample can't statistically validate under-expression. Genes over-expressed in male *A. crassus* comprise major sperm proteins well known for their high expression in nematode sperm [77].

Conclusions

The *A. crassus* transcriptome provides a basis of molecular research on this important species. It further provides insight in the evolution of parasitism complementing the catalogue of available transcriptomic data with a member of the Spirurina phylogenetically distant to so far sequenced parasites in this clade.

Competing interests

The authors declare no competing interests.

Authors contributions

EGH and MB conceived and designed the experiments. EH carried out bioinformatic analyses. SB assisted in bioinformatic analyses. AM prepared sequencing libraries. HT provided close supervision throughout. EGH and MB interpreted results and prepared the manuscript. All authors have read and approved the final manuscript.

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. De Charleroy D, Grisez L, Thomas K, Belpaire C, Ollevier F: **The life cycle of Anguillicola crassus**. *Diseases of Aquatic Organisms* 1990, **8**(2):77–84.

3. Neumann W: **Schwimblasenparasit Anguillicola bei Aalen**. *Fischer und Teichwirt* 1985, :322.

4. Koops H, Hartmann F: **Anguillicola-infestations in Germany and in German eel imports**. *Journal of Applied Ichthyology* 1989, **5**:41–45.

5. 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].

6. 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].

7. 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.

8. 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].

9. Münderle M, Taraschewski G, 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].

10. 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].

11. 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].

12. 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].

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. 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].

16. Blaxter M, De Ley P, Garey J, X Liu L, Scheldeman P, Vierstraete A, Vanfleteren J, Mackey L, Dorris M, Frisse L, Vida J, Thomas W: **A molecular evolutionary framework for the phylum Nematoda**. *Nature* 1998, **392**(6671):71–75, [http://dx.doi.org/10.1038/32160].

17. Nadler SA, Carreno RA, Meja-Madrid H, Ullberg J, C 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. 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].

20. 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].

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

22. 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].

23. 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, [http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3012609].

24. 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].

25. 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].

26. 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].

27. Elsworth B, Wasmuth J, Blaxter M: **NEMBASE4: The nematode transcriptome resource**. *Int. J. Parasitol.* 2011, **41**:881–894, [http://www.ncbi.nlm.nih.gov/pubmed/21550347].

28. 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].

29. 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, [http://www.ncbi.nlm.nih.gov/pubmed/18836194].

30. Iseli C, Jongeneel CV, 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].

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

32. Petersen TN, Brunak S, von Heijne G, Nielsen H: **SignalP 4.0: discriminating signal peptides from transmembrane regions**. *Nat. Methods* 2011, **8**:785–786, [http://www.ncbi.nlm.nih.gov/pubmed/21959131].

33. Zdobnov EM, Apweiler R: **InterProScan–an integration platform for the signature-recognition methods in InterPro**. *Bioinformatics* 2001, **17**:847–848, [http://www.ncbi.nlm.nih.gov/pubmed/11590104].

34. Kasprzyk A: **BioMart: driving a paradigm change in biological data management**. *Database (Oxford)* 2011, **2011**:bar049, [http://www.ncbi.nlm.nih.gov/pubmed/22083790].

35. Durinck S, Spellman PT, Birney E, Huber W: **Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt**. *Nat Protoc* 2009, **4**:1184–1191, [http://www.ncbi.nlm.nih.gov/pubmed/19617889].

36. Ning Z, Cox AJ, Mullikin JC: **SSAHA: a fast search method for large DNA databases**. *Genome Res.* 2001, **11**:1725–1729, [http://www.ncbi.nlm.nih.gov/pmc/articles/PMC311141].

37. 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, [http://dx.doi.org/10.1093/bioinformatics/btp352].

38. 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, [http://www.ncbi.nlm.nih.gov/pubmed/19542151].

39. Falcon S, Gentleman R: **Using GOstats to test gene lists for GO term association**. *Bioinformatics* 2007, **23**:257–258, [http://www.ncbi.nlm.nih.gov/pubmed/17098774].

40. Danecek P, Auton becasis G, CA A, Banks E, DePristo M, RE H, G L, GT M, ST S, GT M, T D, 1000 Genomes Project: **The variant call format and VCFtools**. *Bioinformatics* 2011, **27**:2156–2158, [http://www.ncbi.nlm.nih.gov/pubmed/21653522].

41. Alho JS, Valimaki K, Merila J: **Rhh: an R extension for estimating multilocus heterozygosity and heterozygosity-heterozygosity correlation**. *Mol Ecol Resour* 2010, **10**:720–722, [http://www.ncbi.nlm.nih.gov/pubmed/21565077].

42. Amos W, Wilmer JW, Fullard K, Burg TM, Croxall JP, Bloch D, Coulson T: **The influence of parental relatedness on reproductive success**. *Proc. Biol. Sci.* 2001, **268**:2021–2027, [http://www.ncbi.nlm.nih.gov/pubmed/11571049].

43. Aparicio JM, Ortego J, Cordero PJ: **What should we weigh to estimate heterozygosity, alleles or loci?** *Mol. Ecol.* 2006, **15**:4659–4665, [http://www.ncbi.nlm.nih.gov/pubmed/17107491].

44. ColtMan W, G PJ, A SJ, ton JM P: **Parasite-mediated selection against inbred Soay sheep in a free-living, island population**. *Evolution* 1999, **81**:1259–1267, [http://www.jstor.org/stable/2640828].

45. Morgan M, Pagès H: *Rsamtools: Import aligned BAM file format sequences into R / Bioconductor*[http://bioconductor.org/packages/release/bioc/html/Rsamtools.html]. [R package version 1.4.3].

46. Anders S, Huber W: **Differential expression analysis for sequence count data**. *Genome Biol.* 2010, **11**:R106, [http://www.ncbi.nlm.nih.gov/pubmed/20979621].

47. 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].

48. 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].

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

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

51. Chen H, Boutros PC: **VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R**. *BMC Bioinformatics* 2011, **12**:35, [http://www.ncbi.nlm.nih.gov/pubmed/21269502].

52. Balzer S, Malde K, Jonassen I: **Systematic exploration of error sources in pyrosequencing flowgram data**. *Bioinformatics* 2011, **27**:i304–309, [http://www.ncbi.nlm.nih.gov/pubmed/21685085].

53. 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, [http://www.ncbi.nlm.nih.gov/pubmed/21138572].

54. 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, [http://www.ncbi.nlm.nih.gov/pubmed/20386959].

55. Papanicolaou A, Stierli R, Ffrench-Constant RH, Heckel DG: **Next generation transcriptomes for next generation genomes using est2assembly**. *BMC Bioinformatics* 2009, **10**:447, [http://www.ncbi.nlm.nih.gov/pubmed/20034392].

56. 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].

57. 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, [http://www.ncbi.nlm.nih.gov/pubmed/20478048].

58. 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, [http://www.ncbi.nlm.nih.gov/pubmed/21364769].

59. 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, [http://www.ncbi.nlm.nih.gov/pubmed/20331785].

60. Parkinson J, Anthony A, Wasmuth J, Schmid R, Hedley A, Blaxter M: **PartiGene–constructing partial genomes**. *Bioinformatics* 2004, **20**(9):1398–1404, [http://bioinformatics.oxfordjournals.org/cgi/content/abstract/20/9/1398].

61. Wasmuth J, Schmid R, Hedley A, Blaxter M: **On the Extent and Origins of Genic Novelty in the Phylum Nematoda**. *PLoS Neglected Tropical Diseases* 2008, **2**(7):e258, [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2432500].

62. 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].

63. Yang H, Chen X, Wong WH: **Completely phased genome sequencing through chromosome sorting**. *Proc. Natl. Acad. Sci. U.S.A.* 2011, **108**:12–17, [http://www.ncbi.nlm.nih.gov/pubmed/21169219].

64. Adey A, Morrison H, Asan X, Xun X, Kitzman 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, [http://genomebiology.com/content/11/12/R119].

65. Kryazhimskiy S, Plotkin JB: **The population genetics of dN/dS**. *PLoS Genet.* 2008, **4**:e1000304, [http://www.ncbi.nlm.nih.gov/pubmed/19081788].

66. 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, [http://www.ncbi.nlm.nih.gov/pubmed/18590545].

67. 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, [http://www.ncbi.nlm.nih.gov/pubmed/15579698].

68. 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, [http://www.ncbi.nlm.nih.gov/pubmed/11404480].

69. 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, [http://www.ncbi.nlm.nih.gov/pubmed/20470405].

70. 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, [http://www.ncbi.nlm.nih.gov/pubmed/6449605].

71. 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].

72. Knopf K, Madriles Helm A, Lucius R, Bleiss W, Taraschewski H: **Migratory response of European eel (Anguilla anguilla) phagocytes to the eel swimbladder nematode Anguillicola crassus**. *Parasitology Research* 2008, **102**(6):1311–6, [http://www.ncbi.nlm.nih.gov/pubmed/18311570].

73. Molnár K: **Formation of parasitic nodules in the swimbladder and intestinal walls of the eel Anguilla anguilla due to infections with larval stages of Anguillicola crassus**. *Diseases of Aquatic Organisms* 1994, **20**(3):163–170.

74. 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, [http://www.ncbi.nlm.nih.gov/pubmed/9664699].

75. Wielgoss S, Taraschewski H, Meyer A, Wirth T: **Population structure of the parasitic nematode Anguillicola crassus, an invader of declining North Atlantic eel stocks**. *Molecular Ecology* 2008, **17**(15):3478–95, [http://www.ncbi.nlm.nih.gov/pubmed/18727770].

76. Scott AL: **Nematode sperm**. *Parasitol. Today (Regul. Ed.)* 1996, **12**:425–430, [http://www.ncbi.nlm.nih.gov/pubmed/15275275].

Figures

Figure 1 - Number of contigs annotated with different methods

Number of annotations obtained for Gene Ontology (GO), Enzyme Commission (EC) and Kyoto Encyclopedia of Genes and Genomes (KEGG) terms through `Annot8r` [32] for all TUGs (a) and for higCA derived contigs (b). The latter includes additional domain-based annotations obtained with `InterProScan` [34].

Figure 2 - Comparing high level GO-slim annotations

For Gene Ontology (GO) categories molecular function, cellular compartment and biological process the number of terms in high level GO-slim categories is given as obtained through `Annot8r` [32].

Figure 3 - Changes in ti/tv and dn/ds due to exclusion of homopolymer-runs

When SNPs in or adjacent to homopolymeric regions are removed changes in ti/tv and dn/ds are observed: As the overall number of SNPs is reduced both ratios change to more plausible values. Note the reversed axis for dn/ds to plot these lower values to the right. For homopolymer length > 3 a linear trend for the total number of SNPs and the two measurements is observed. A width of 11 for the screening window provides most plausible values (suggesting specificity) while still incorporating a high number of SNPs (sensitivity).

Figure 4 - SNP calling and SNP categories

Overabundance of SNPs at (a) codon-position two and of (c) non-synonymous SNPs for low percentages of the minority allele. (b) Significant positive correlation of coverage and dn/ds before removing these SNPs at a threshold of 7% ($p < 0.001$, $R^2 = 0.015$) and (d) afterwards (R^2 <0.001, p =0.211).

Figure 5 - Positive selection and evolutionary conservation

Box-plots for dn/ds in TUGs according to different categories of evolutionary conservation. Significant comparisons are novel.in.metazoa - novel.in.Ac (0.005 and 0.015), novel.in.nematoda - novel.in.Ac (0.005 and 0.002), novel.in.nematoda - novel.in.clade3 (0.207 and 0.045; p-value for bitscore of 50 and 80, Nemenyi-Damico-Wolfe-Dunn test).

Figure 6 - Enrichment of Signal-positives for categories of evolutionary conservations

Proportions of `SignalP`-predictions for each category of evolutionary conservation. Generally - across bit-score thresholds - TUGS novel in nematodes and in *A. crassus* have the highest proportion of signal-positives.

Tables

Table 1 - Sampling, trimming and pre-assembly screening, library statistics

For libraries two sequencing libraries from European eels (E1 and E2) one form L2-larvae (L2), one from male (M) and two from Eels in Taiwan (T1 and T2) the following statistics are given. life.st = lifecycle stage: f for female m for male. source.p = source population: R for Rhine, P for Poland, C for cultured, W for wild. raw.reads = raw number of sequencing reads obtained. lowqal = number of reads discarded due to low quality or length in *Seqclean* [23]. AcrRNA = number of reads hitting *A. crassus*-rRNA (screened). eelmRNA = number of reads hitting eel transcriptome-sequences (screened). eelrRNA = number of reads hitting eel-rRNA genes (screened). Cercozoa = number of reads hitting cercozoan rRNA (screened). valid = number of reads valid after screening (assembled). valid.span = number of bases valid (assembled). mapping.unique = number of reads mapping uniquely to the assembly. mapping.Ac = number of reads mapping to the part of the assembly considered *A. crassus* origin (see post-assembly screening). mapping.MN = number of reads mapping to the highCA-derived part of the assembly (and also *A. crassus* origin).

library	E1	E2	L2	M	T1	T2
life.st	adult f	adult f	L2 lavae	adult m	adult f	adult f
source.p	Europe R	Europe P	Europe R	Asia C	Asia C	Asia W
raw.reads	209325	111746	112718	106726	99482	116366
lowqal	92744	10903	15653	15484	7947	27683
AcrRNA	76403	11213	30654	31351	24929	7233
eelmRNA	4835	3613	1220	1187	7475	11741
eelrRNA	13112	69	1603	418	514	38
Cercozoa	0	0	5286	0	0	0
valid	22231	85948	58302	58286	58617	69671
valid.span	7167338	24046225	16661548	17424408	14443123	20749177
mapping.unique	12023	65398	39690	36782	42529	55966
mapping.Ac	8359	61070	12917	31656	37158	50018
mapping.MN	5883	48006	8475	18986	28823	41545

Table 2 - assembly classification and contig statistics

Summary statistics for contigs from different assembly-categories given in columns as highCA = high credibility assembly; lowCA = low credibility assembly, combined = 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-annotation, KEGG = number of contigs with KEGG-annotation, EC = number of contigs with EC-annotation, 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.

	lowCA	highCA	combined
total.contigs	26336	13851	40187
rRNA.contigs	835	60	895
fish.contigs	2419	1022	3441
xeno.contigs	1935	1398	3333
remaining.contigs	21147	11371	32518
remaining.span	8095986	7971550	16067536
non.u.cov	14.665	10.979	12.840
cov	2.443	6.838	4.624
p4e.BLAST-similarity	4356	5663	10019
p4e.ESTScan	8324	3597	11921
p4e.LongestORF	8347	2085	10432
p4e.no-prediction	93	14	107
full.3p	5906	2714	8620
full.5p	1484	1270	2754
full.l	104	185	289
GO	2635	3874	6509
EC	966	1492	2458
KEGG	1608	2236	3844
IPR	0	7557	7557
nem.blast	4868	5820	10688
any.blast	5106	6007	11113

Table 3 - Evolutionary conservation and novelty

The kingdom Metazoa (novel.in.metazoa), the phylum Nematoda(novel.in.nematoda) and clade III (Spirurina; novel.in.spirurina) were assessed for occurrences of `BLAST`-hits at two different bitscore thresholds (50 = bit.50 and 80 = bit.80). TUGs without any hit at a given threshold were categorized as novel in *A. crssus* (novel.in.Ac). Both novelty and conservation can be derived from this (numbers for conservation would be the cumulative sum of lower-level novelty).

	conserved	novel.in.metazoa	novel.in.nematoda	novel.in.clade3	novel.in.Ac
bit.50.all	5604	1713	2173	1485	21543
bit.80.all	3506	1382	2014	1525	24091
bit.50.highCA	3479	875	1010	601	5406
bit.80.highCA	2457	832	1084	716	6282

Table 4 - Over- and under-representation of GO-terms in positively selected

GO-terms over-represented in contigs putatively under positive selection. Horizontal lines separate categories of the GO-ontology. First category is molecular function, second biological process, last cellular compartment. P values (classic) for over-representation are given along with the number of positively selected contigs (Significant; dn/ds > 0.5) and the number of contigs with this annotation for which a dn/ds was obtained (Annotated) and the description of the GO-term (Term).

GO.ID	Term	Annotated	Significant	Expected	classic
GO:0008233	peptidase activity	43	12	5.26	0.0028
GO:0015179	L-amino acid transmembrane transporter activity	2	2	0.24	0.0147
GO:0016787	hydrolase activity	110	20	13.45	0.0262
GO:0043021	ribonucleoprotein binding	6	3	0.73	0.0266
GO:0005102	receptor binding	26	7	3.18	0.0288
GO:0046982	protein heterodimerization activity	16	5	1.96	0.0348
GO:0004129	cytochrome-c oxidase activity	3	2	0.37	0.0407
GO:0004540	ribonuclease activity	3	2	0.37	0.0407
GO:0005275	amine transmembrane transporter activity	3	2	0.37	0.0407
GO:0005342	organic acid transmembrane transporter activity	3	2	0.37	0.0407
GO:0009081	branched chain family amino acid metabolic process	3	3	0.36	0.0017
GO:0009083	branched chain family amino acid catabolic process	3	3	0.36	0.0017
GO:0042594	response to starvation	15	6	1.82	0.0052
GO:0006914	autophagy	12	5	1.45	0.0090
GO:0006520	cellular amino acid metabolic process	44	11	5.33	0.0102
GO:0007281	germ cell development	17	6	2.06	0.0105
GO:0090068	positive regulation of cell cycle process	17	6	2.06	0.0105
GO:0009308	amine metabolic process	57	13	6.90	0.0118
GO:0051325	interphase	23	7	2.79	0.0139
GO:0051329	interphase of mitotic cell cycle	23	7	2.79	0.0139
GO:0030532	small nuclear ribonucleoprotein complex	7	4	0.84	0.005
GO:0005682	U5 snRNP	2	2	0.24	0.014
GO:0015030	Cajal body	2	2	0.24	0.014
GO:0046540	U4/U6 x U5 tri-snRNP complex	2	2	0.24	0.014
GO:0016607	nuclear speck	6	3	0.72	0.025
GO:0005739	mitochondrion	136	23	16.35	0.031
GO:0005604	basement membrane	3	2	0.36	0.039
GO:0060198	clathrin sculpted vesicle	3	2	0.36	0.039
GO:0016604	nuclear body	13	4	1.56	0.059
GO:0008021	synaptic vesicle	9	3	1.08	0.082

0.1 Table 5 - Measurements of multi-locus heterozygosity for single worms

Genotyping for a set of 199 SNPs, different measurements were obtained to asses genome-wide heterozygosity. Measurements for relative heterozygosity (rel.het; number of homozygous sites/ number of heterozygous sites), internal relatedness (int.rel; [43]), homozygosity by loci (ho.loci; [44]) and standardized

heterozygosity (std.het; [45]) are given. All these measurements are pointing to sample T1 (Taiwanese worm from a wild population) as the most heterozygous and sample E2 (the European worm from Poland) as the least heterozygous individual. Heterozygote-heterozygote correlation [42] confirmed the genome-wide significance of these markers.

	rel.het	int.rel	ho.loci	std.het
T2	0.45	-0.73	0.59	1.00
T1	0.93	-0.95	0.34	1.62
M	0.37	-0.73	0.66	0.84
E1	0.38	-0.83	0.60	0.91
E2	0.18	-0.35	0.82	0.50

0.2 Table 6 - Expression analysis for contigs and orthologous groups

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.