The transcriptome of the invasive eel swimbladder nematode parasite Anguillicola crassus

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

[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 EH9 3JT, UK [3]The GenePool Sequencing Service, The Ashworth laboratories, The University of Edinburgh, King's Buildings Campus, Edinburgh EH9 3JT, UK

Email: Emanuel 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 economically and ecologically important parasitic nematode of eels. The native range of *A. crassus* is in Asia, where it infects *Anguilla japonica*, the japanese eel. *A. crassus* was introduced into European eels, *Anguilla angulla*, 30 years ago. The parasite is more pathogenic in its new host than in its native one, and is thought to threaten the endangered *An. anguilla* across its range. The molecular bases for the increased pathogenicity of the nematodes in their new hosts is not known.

Results: A reference transcriptome was assembled for *A. crassus* from Roche 454 pyrosequencing data. Raw reads (756,363 total) from nematodes from *An. japonica* and *An. anguilla* hosts were filtered for likely host contaminats and ribosomal RNAs. The remaining 353,055 reads were assembled into 11,372 contigs of a high confidence assembly (highCA) (spanning 6.6 Mb) and 21,153 singletons and contigs of a lower confidence (lowCA) assembly (spanning an additional 6.2 Mb). Roughly 55% of the highCA derived contigs could be annotated with domain- or protein sequence similarity match-derived functional information. Sequences conserved only in nematodes, or unique to *A. crassus* were more likely to have secretory signal peptides. Thousands of high quality single nucleotide polymorphisms were identified, and coding polymorphism was correlated with differential expression between individual nematodes. Transcripts identified as being under

positive selection were enriched in peptidases. Enzymes involved in energy metabolism were enriched in the set of genes differentially expressed between European and Asian *A. crassus*.

Conclusions: The reference transcriptome of *A. crassus* is of high quality, and will serve as a basis for future work on the invasion biology of this important parasite. The polymorphisms identified will provide a key tool set for analysis of population structure and identification of genes likely to be involved in increased pathogenicity in European eel hosts. The identification of peptidases under positive selection is a first step in this programme.

Background

The nematode *Anguillicola crassus* Kuwahara, Niimi et Itagaki, 1974 is a native parasite of the Japanese eel *Anguilla japonica* [1]. Adults localise to the swim bladder where they feed on blood [2]. Larvae are transmitted via crustacean intermediate hosts [3]. Originally endemic to East- Asian populations of *An. 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*. *A. crassus* was recorded for the first time in Europe in North-West Germany in 1982 [4], where it was most likely introduced through the live-eel trade [5, 6], *A. crassus* has subsequently spread rapidly through populations of its newly acquired host [7], and currently is found in all *An. anguilla* populations except those in Iceland [8]. *A. crassus* can thus be regarded as a model for the introduction and spread of invasive parasite [9].

In *An. anguilla*, prevalence and mean intensity of infection by *A. crassus* are higher than in *An. japonica* [10, 11]. In *An. anguilla* infections, the adult nematodes are larger, and have an earlier onset of reproduction and a greater egg output [12]. *An. anguilla* shows increased pathology, including thickening and inflammation of the swim bladder wall [13]. It has been suggested that the life history modifications and changed virulence observed in *A. crassus* in the new host are due to an inadequate immune response in *An. anguilla* [14]. *An. japonica* is capable of killing histotropic larvae of the parasite after vaccination [15] or under high infection pressure [16], but this does not happen in *A. anguilla*.

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 filariasis and ascariasis) as well as prominent veterinary parasites. Molecular phylogenetic analyses place

Anguillicola in a clade of spirurine nematodes (Spirurina B of [20]) that have a 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 filariasis and ascariasis), and thus *A. crassus* may be used as an outgroup taxon to understand the evolution of parasitic phenotypes in these species.

The differences in the biology of *A. crassus* in *An. japonica* (coevolved) and *An. anguilla* (recently captured) eel hosts likely results from differential interactions between host genetics and parasite genetics. While genetic differences between the host species are expected, it is not known what part, if any, genetic differentiation between the invading European and endemic Asian parasites plays. European *A. crassus* are less genetically variable than parasites taken from Asian hosts [21], reflecting the derived nature of the invading populations and the likely population bottlenecks this entailed. 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 host 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 [22] is particularly suited to transcriptomics of previously unstudied species [23].

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 within the nematode.

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 [16]. Nematodes from *An. anguilla* were sampled from 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.

Raw sequencing reads are archived under study-accession number SRP010313 in the NCBI Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/Traces/sra) [24]. All samples were sequenced using the FLX Titanium chemistry, except for the Taiwanese female sample T1, 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 MINT adapter sequences using cross-match [25] (with parameters -minscore 20 -minmatch 10). Seqclean [26] 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 [27], 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 [23]: two assemblies were generated, one using Newbler v2.6 [22] (with parameters -cdna -urt), the other using Mira v3.2.1 [28] (with parameters –job=denovo,est,accurate,454). The resulting two assemblies were combined into one using Cap3 [29] at default settings and contigs were labeled by whether they derived from both assemblies (high confidence assembly; highCA), or one assembly only (lowCA; for a detailed analysis of the assembly categories see the supporting Methods file). The superset of highCA and lowCA contigs and the remaining unassembled reads defines the set of tentatively unique genes (TUGs).

Post-assembly classification and taxonomic assignment of contigs

We rescreened the assembled assembly 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 [30, 31]. 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. Sequences with a best hit to non-Metazoans or to Chordata within Metazoa were excluded from further analysis.

Protein prediction and annotation

Protein translations were predicted from the contigs using prot4EST (version 3.0b) [32]. Proteins were predicted either by joining single high scoring segment pairs (HSPs) from a BLAST search of uniref100 [33], or by ESTscan [34], 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) [35], using the annotated sequences available in uniref100 [33]. 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 [36] was used to predict signal peptide cleavage sites and signal anchor signatures for the *A. crassus*-transcriptome and for the proteomes of the two model-nematodes. InterProScan [37] (command line utility iprscan version 4.6 with options -cli -format raw -iprlookup -seqtype p -goterms) was used to obtain domain 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 [38] to WormBase v. 220 through the R-package biomaRt [39].

Single nucleotide polymorphism analysis

We mapped the raw reads to 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) [40]. From the ssaha2 output, pileup -files were produced using samtools [41], discarding reads mapping to multiple regions. VarScan [42] (pileup2snp) was used with default parameters on pileup -files to output lists of single nucleotide polymorphisms (SNPs) and their locations.
In the 10,496 SNPs thus defined, the ratio of transitions (ti; 6,908) to transversion (tv; 3,588) was 1.93. From the prot4EST predictions, 7,189 of the SNPs were predicted to be inside an ORF, with 2,322 at codon first positions, 1,832 at second positions and 3,035 at third positions. As expected, ti/tv inside ORFs (2.39) was higher than outside ORFs (1.25). The ratio of synonymous polymorphisms per synonymous site to non-synonymous polymorphisms per non-synonymous site in this unfiltered SNP set (dn/ds) was 0.45, rather high compared to other analyses. Roche 454 sequences have well-known systematic errors associated with homopolymeric nucleotide sequences [58], and the effect of exclusion of SNPs in, or close to, homopolymer regions was explored. When SNPs were discarded using different size thresholds for homopolymer runs and proximity thresholds, the ti/tv and in dn/ds ratios changed (Additional Figure 1). Based on this SNPs associated 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 were discarded. There was 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 (Additional Figure 2).
For enrichment analysis of GO-terms we used the R-package GOstats [43].
Using Samtools [41] (mpileup -u) and Vcftools [44] (view -gcv) we genotyped individual libraries for each of the master list of 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 [45] to calculate internal relatedness [46], homozygosity by locus [47] and standardised heterozygosity [48] from these data.
We confirmed the significance of heterozygote-heterozygote correlation by analysing 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 for genotyping using the R- package Rsamtoools [49].
LowCA contigs and contigs with less than 32 reads over all libraries were excluded from analysis. Libraries E1 and L2 had very low overall counts and thus we excluded these libraries from analysis. The statistic of

Audic and Claverie [50] as implemented in ideg6 [51] was used to contrast single libraries. Differential expression between libraries from male versus female nematodes was accepted for genes that differed in expression values between all the female libraries (E2, T1 and T2; see Table 1) versus the male (M) library ($p < 0.01$), but had no differential expression within any of the female libraries at the same threshold. Differential expression between libraries from nematodes of European *An. anguilla* and Taiwanese *An. japonica* origin was accepted for genes that differed in expression values between library E2 and both libraries T1 and T2 ($p < 0.01$), but showed no differences between T1 and T2.

Over-representation analyses

The R-package annotationDbi [52] was used to obtain a full list of associations (along with higher-level terms) from annot8r annotations prior to analysis of GO term over-representation in gene sets selected on the basis of dn/ds or expression values. The R-package topGO [53] was used to traverse the annotation graph and analyse each node term for over-representation in the focal gene set compared to an appropriate universal gene set (all contigs with dn/ds values or all contigs analysed for gene expression) with the "classic" method and Fisher's exact test. Terms for which an offspring term was already in the table and no additional counts supported overrepresentation were removed. Mann-Whitney u-tests were used to test the inuence of factors on dn/ds values. To investigate multiple contrasts between groups (factors) Nemenyi-Damico-Wolfe-Dunn tests were used, and for overrepresentation of one group (factor) in other groups (factors) Fisher's exact test was used.

General coding methods

The bulk of analysis (unless otherwise described) presented was carried out in R [54] using custom scripts. For visualisation we used the R-packages ggplot2 [55] and VennDiagram [56]. We used a method provided in the R-packages Sweave [57] and Weaver [58] for "reproducible research" combining R and LaTeXcode in a single file. All intermediate data files needed to compile the present manuscript from data can be downloaded from xxx ***.

Results

Sampling A. crassus

One female *A. crassus* and one male *A. crassus* were sampled from an *An. japonica* aquaculture with high infection loads in Taiwan, and an additional female was sampled from an *An. japonica* caught in a stream

with low infection pressure adjacent to the aquaculture. A female nematode and pool of L2 larval stages were sampled from *An. anguilla* in the river Rhine, and one female from *A. anguilla* from a lake in Poland. All adult nematodes were replete with host blood. To assist in downstream filtering of host from nematode reads, we also sampled RNA from the liver of an uninfected taiwanese *An. japonica*.

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 (spanning 169,863,104 bases). From An. japonica liver RNA 159,370 raw reads were generated, and 135,072 retained after basic quality screening. These eel reads were assembled into 10,639 contigs.

The *A. crassus* reads were screened for contamination by host sequence by comparison to our assembled *An. japonica* 454 transcriptome and publicly accessible *An. anguilla* sequence data, and 30,071 reads removed. By comparison to *A. crassus* small subunit ribosomal RNA (sequenced previously) and large subunit ribosomal RNA (assembled from our reads in preliminary analyses), 181,783 were tagged and removed. The L2 larval library proved to have contributions for other cobionts of the eel, and 5,286 reads were removed because they matched closely to cercozoan (likely parasite) ribosomal RNA genes.

Assembly and post-assembly screening

The remaining 353,055 reads (spanning 100,491,819 bases) were assembled using the combined assembler strategy [23], employing Roche 454 GSassembler (version 2.6) and MIRA (version 3.21) [28]. In this coassembly, 13,851 contigs were supported by both assembly algorithms, 3,745 contigs were supported by only one of the assembly algorithms and 22,591 singletons were not assembled by either program (Table 2). Contigs supported by both assemblers were longer, and were more likely to have a significant similarity to previously sequenced protein coding genes than contigs assembled by only one of the algorithms, or the remaining unassembled singletons. These constitute the high credibility assembly (highCA), while those with evidence from only one assembler and the singletons are the low credibility assembly (lowCA). These datasets were 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 was prioritised, while in the complete assembly (highCA plus lowCA) completeness was prioritised. The 40,187 sequences (contig consensuses and singletons) in the complete assembly are referred to as tentatively unique genes (TUGs).

We screened the complete assembly for remaining host contamination, and identified 3,441 TUGs that had significant, higher similarity to eel (and chordate; EMBLBank Chordata proteins) than to nematode sequences [31]. Given the identification of cercozoan ribosomal RNAs in the L2 library, we also screened the complete assembly for contamination with other transcriptomes.

1,153 TUGs were found with highest significant similarity to Eukaryota outside of the kingdoms Metazoa, Fungi and Viridiplantae. These contigs matched genes from a wide range of protists from Apicomplexa (mainly Sarcocystidae, 28 hits and Cryptosporidiidae 10 hits), Bacillariophyta (diatoms, mainly Phaeodactylaceae, 41 hits), Phaeophyceae (brown algae, mainly Ectocarpaceae, 180 hits), Stramenopiles (Albuginaceae, 63 hits), Kinetoplasitda (Trypanosomatidae, 26 hits) and Heterolobosea (Vahlkampfidae, 38 hits). Additionally 298 TUGs had best, significant matches to genes from fungi (e.g Ajellomycetaceae, 53 hits) and 585 TUGs had best, significant matches to genes from plants. Outside the Eukaryota there were significant best matches to Bacteria (825 TUGs; mostly to members of the Proteobacteria), Archaea (8 TUGs) and viruses (9 TUGs). No TUGs had significant, best matches to Wolbachia or related Bacteria known as symbionts of nematodes and arthropods. All TUGs with highest similarity to sequences deriving from taxa outside Metazoa were excluded. The final, screened *A. crassus* assembly has 32,525 TUGs, spanning 12,733,095 bases (of which 11,372 are highCA-derived, and span 6,575,121 bases). All analyses reported below are based on this filtered dataset.

Annotation

For 32,418 screened TUGs a protein was predicted using prot4EST [32] (Table 2). An apparently full-length open reading frame (ORF) was obtained in 353 TUGs, while for 2,683 the 5' ends and for 8,283 the 3' ends were complete. In 13,383 TUGs the corrected sequence with the imputed ORF was slightly changed compared to the raw sequence. One third of the TUGs had significant similarity to proteins from other nematodes:

9,556 TUGs matched *C. elegans* proteins, 9,664 TUGs matched *B. malayi*, and 11,620 TUGs had matches in NEMPEP4 [30, 31]. Comparison to the UniProt reference identified 11,115 TUGs with significant similarities. We used annot8r [35] to assign GO terms to 6,511 TUGs, EC numbers for 2,460 TUGs and KEGG pathway annotations for 3,846 TUGs (Table 2). Additionally 5,125 highCA derived contigs were annotated with GO terms through InterProScan [37].

Nearly one third (6,989) of the *A. crassus* TUGs were annotated with at least one identifier, and 1,831 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 Spirurid filarial nematode *B. malayi* and the complete proteome of *C. elegans* (Figure 2). The occurrence of GO terms in the annotation of the partial transcriptome of *A. crassus* was more similar to that for the proteome of *B. malayi* (0.95; Spearman correlation coefficient) than to the that of the proteome of *C. elegans* (0.9).

Despite the lack of completeness at the 5' end suggested by peptide prediction, just over 3% of the TUGs were predicted to be secreted (920 with 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 of *B. malayi* (742 signal peptide cleavage sites and 41 with transmembrane anchor) than for the proteome of *C. elegans* (4,273 signal peptide cleavage sites and 154 with transmembrane anchor).

By comparison to RNAi phenotypes for *C. elegans* genes [59, 60] likely to be orthologous to *A. crassus* TUGs, 6,029 TUGs were inferred to be essential (RNAi lethal phenotype in *C. elegans*).

To explore the phylogenetic conservation of *A. crassus* TUGs, they were classified as conserved across kingdoms, conserved in Metazoa, conserved in Nematoda, conserved in Spirurina or novel to *A. crassus* by comparing them to custom database subsets using BLAST (Table 3). Using a relatively strict cutoff, a quarter of the highCA derived contigs were conserved across kingdoms, and 10% were apparently restricted to Nematoda. Nearly half of the highCA contigs were novel to *A. crassus*.

Similar patterns were observed for conservation assessed at different stringency, and when assessed across all TUGs, except that a higher proportion of all TUGs were apparently unique to *A. crassus*.

Proteins predicted to be resticted 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 3).

The proportion of lethal RNAi phenotypes was significantly higher for *C. elegans* presumed orthologs of TUGs conserved across kingdoms (97.23%) than for orthologs of TUGs not conserved across kingdoms (94.59%; $p<0.001$, Fisher's exact test).

Identification and analysis of single nucleotide polymorphisms

Single nucleotide polymorphisms (SNPs) were called using VARScan [42] on the 1,100,522 bases of TUGs that had coverage of more then 8-fold available. SNPs predicted to have more than 2 alleles, or that mapped to an undetermined (N) base were excluded, as were SNP likely to be due to base calling errors close to homopolymer tracts and SNP calls resulting from apparent rare variants.

Our filtered SNP dataset includes 5,113 SNPs, with 4.65 SNPs per kb of contig sequence. There were 7.95 synonymous SNPs per 1000 synonymous bases and 2.44 non-synonymous SNPs per 1000 non synonymous bases. A mean dn/ds of 0.244 was calculated for the 765 TUGs (all highCA derived contigs) containing at least one synonymous SNP. Positive selection can be inferred from high dn/ds ratios. Over-represented GO ontology terms associated with TUGs with dn/ds higher than 0.5 were identified (Table 4; Additional Figures 4). Within the molecular function category, "peptidase activity" was the most significantly overrepresented term. The twelve TUGs annotated as peptidases each had unique orthologs in *C. elegans* and *B. malayi*. Other overrepresented terms abundant over categories identified subunits of the respiratory chain: "heme-copper terminal oxidase activity" and "cytochrome-c oxidase activity" in molecular function and "mitochondrion" in cellular compartment (Table 4 and Additional Figures 4). Contigs identified as novel to clade III and novel in *A. crassus* had a significantly higher dn/ds than other contigs (Additional Figure 3).

Signal peptide containing proteins have been shown to have higher rates of evolution than cytosolic proteins in a number of nematode species. *A. crassus* TUGs predicted to contain signal peptide cleavage sites showed a non-significant trend towards higher dn/ds values than TUGs without signal peptide cleavage sites ($p = 0.22$; two sided Mann-Whitney-test).

Orthologs of *C. elegans* transcripts with lethal RNAi phenotype are expected to evolve under stronger selective constraints and 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 ($p=0.815$, two-sided U-test).

The genotypes of single adult nematodes were called using Samtools [41] and Vcftools [44], and 199 informative sites (where two alleles were found in at least one assured genotype at least in one of the nematodes) were identified in 152 contigs. Internal relatedness [46], homozygosity by loci [47] and standardised heterozygosity [48] all identified the Taiwanese nematode from aquaculture (sample T1) as the most and the European nematode from Poland (sample E2) as the least heterozygous individuals. The genome-wide representativeness of these 199 SNP markers for the whole genome in population genetic studies was confirmed using heterozygosity-heterozygosity correlation [45]: mean internal relatedness = 0.78, lower bound of 95% confidence intervals from 1000 bootstrap replicates (cil) = 0.444; mean homozygosity by loci = 0.86, cil = 0.596; standardised heterozygosity = 0.87, cil= 0.632.

Differential gene expression

Gene expression was inferred by the unique mapping of 252,388 (71.49%) of the raw reads to the fullest assembly (including the all assembled contigs as a "filter"; see Table 1). In analysis, non-*A. crassus* contigs, and all contigs with fewer than 32 reads overall were excluded. Thus 658 TUGs were analysed for differential expression using ideg6 658 for normalisation and the statistic of Audic and Claverie [50] for detection of differences. Of these TUGs, 54 showed expression predominantly in the male library, 56 TUGs were more highly represented in the female library, 56 TUGs were primarily expressed in the libraries from Taiwan, and 22 TUGs were overrepresented in the European library.

Overrepresentation of of GO-terms differentially expressed between the male and female libraries highlighted especially ribosomal proteins oxidoreductases and collagen processing enzymes as enriched (Table 6a and Additional Figures 4). Ribosomal proteins were all overexpressed in the male library, oxidoreductases and collagen processing enzymes were overexpressed female libraries.

Analysis of overrepresentation of of GO terms associated with TUGs differentially expressed between male and female libraries identified ribosomal proteins, oxidoreductases and collagen processing enzyme terms (Table 6a and Additional Figures 4). The ribosomal proteins were all overexpressed in the male library, while the oxidoreductases and collagen processing enzymes were overexpressed female libraries. Similar analysis of overrepresentation of of GO terms associated with the TUGs differentially expressed between European nematodes and Asian nematodes identified several terms of catalytic activity especially related to metabolism (Table 6b; Additional Figures 1). TUGs annotated as acyltransferase were upregulated in the European libraries. However, the expression patterns for other TUGs with overrepresented terms connected to metabolism did not show concerted up or down-regulation. Thus for the term "steroid biosynthetic process", 2 TUGs were downregulated and 3 contigs upregulated in European nematodes. No enrichment of of signal peptide positive TUGs, of TUG conservation categories, or TUGs with *C. elegans* orthologs with lethal or non-lethal RNAi-phenotypes was identified. Significantly elevated dn/ds was found for TUGs differentially expressed in European versus Asian nematodes (Fisher's exact test p=0.007; also both up- or downregulated were significant). TUGs overexpressed in the female libraries showed elevated levels of dn/ds (Fisher's exact test p=0.041), but contrast male overexpressed genes showed decreased levels of dn/ds (Fisher's exact test p=0.014).

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 will enable a broad spectrum of molecular research on this ecologically important and evolutionarily interesting parasite. As *A. crassus* lives in close association with its host, we used exhaustive filtering to remove all host-derived, and host-associated organism-derived contamination from the raw and assembled data.

We generated a transcriptome dataset from the definitive host *An. japonica* as part of this filtering process. In addition to eel-derived transcripts, we also removed data apparently derived from protists, particularly cercozoans, that may have been co-parasites of the eels sampled.

Similar taxonomic screening of NGS transcrioptome data has been shown to be important previously [61], particularly in rejection of hypotheses of horizontal gene transfer into the focal species [62]. We were not able to use base frequency- and codon usage-based screening to identify contaminant data [63, 64] because contaminant sequences in our data derived from multiple genomes.

We used a combined assembly approach [23] to generate a transcriptome estimate that had lower redundancy and higher completeness. Projects using single assemblers often report substantially greater numbers of contigs for datasets of similar size (see e.g. [65]). The 3' bias in the assembly likely derivesd from the use of oligod(T) in mRNA capture and cDNA synthesis and bias is near-ubiquitous in deep transcriptome sequencing projects (e.g. [66]). The final *A. crassus* TUG assembly (32,418 contig consensuses) spans 12.7 Mb, and thus likely covers most of the expected span of the transcriptome (the *C. elegans* transcriptome spans 30 Mb, and the *B. malayi* transcriptome 14 Mb).

Comparison between free-living and parasitic nematode species can be used to identify genes that may underpin adaptations for parastism [67, 68]. Annotations were derived for a 30% of all, and over 50% of the highCA *A. crassus* TUGs using sequence similarity to known proteins. Domain annotations were derived for 45% of the highCA TUGs using InterProScan [37].

Comparison with the complete proteomes of *B. malayi* and *C. elegans* showed a remarkable degree of congruence in annotation spectrum in the two parasitic nematodes. This implies that the *A. crassus* transcriptome is a representative partial genome [69]. Using a taxonomically-stratified analysis of BLAST similarities, we identified more *A. crassus* TUGs that apparently arose in the common ancestor of Nematoda than arose in the last common ancestor of the Spirina (Clade III). As *A. crassus* is part of a lineage that arises basally in Spirurina (Clade III), the lack of genes associated with the all-parasitic nematodes of Clade III may be due to phylogenetic distance obscuring relationships, particularly if the

genes underpinning parasitism are, as would be expected, rapidly evolving. TUGs predicted to be part of gene families that arose in the last common ancestor of Nematoda or to be novel to *A. crassus* contained the highest proportion of genes predicted to have secretory signal peptides. This confirms observations made in a *Nippostrongylus brasiliensis* [70], where secreted and surface proteins were less conserved. Analysis of dn/ds (see below) across conservation categories favors the hypothesis of rapid evolution in proteins with more restricted phylogenetic origins.

Transcriptome data were generated from multiple individual *A. crassus* of Taiwanese and European origin. We identified SNPs both within and between populations, but noted aberrant patterns in the ratio of transitions to transversions (ti/tv) and the ratio of non-synonymous SNPs per non-synonymous site to synonymous SNPs per synonymous site (dn/ds). Screening of SNPs in or adjacent to homopolymer regions, removing "noise" associated with common homopolymer errors [71], improved overall measurements of SNP quality, increased the ti/tv ratio to more closely resemble canonical dataset, and resulted in a reduced, credible dn/ds ratio distribution. The corrected ti/tv value of 1.925 (1.25 outside and 2.39 inside ORFs) is in good agreement with the overall ti/tv of *Homo sapiens* (2.16 [72]) or *Drosophila melanogaster* (2.07 [73]). The mean dn/ds ratio decreased with removal of SNPs adjacent to homopolymer regions from 0.45 to 0.244. While interpretation of dn/ds ratios within populations is not unproblematic [74], the assumption of negative (purifying) selection on most protein coding genes makes lower mean values seem more plausible. We applied a threshold value for the minority allele of 7% for exclusion of SNPs, as approximately 10 haploid equivalents were sampled (5 individual nematodes plus a negligible contribution from the L2 library and offspring within the adult female nematodes). This screening reduced non-synonymous SNPs in high coverage TUGs and removed the dependence of dn/ds on coverage, and removed the need to control for sampling biased by depth (i.e. coverage; see [75] and [76]).

The final dn/ds estimates seem plausible, as *D. melanogaster* female reproductive tract transcripts have dn/ds of 0.15 [77] and a 454 transcriptomic analysis of the parasitic nematode *Ancylostoma canium* reported dn/ds of 0.3 [78]. We used a dn/ds threshold for inference of positive selection on coding sequence of 0.5 has been suggested as threshold for assuming positive selection [77] and identified 46 TUGs that may be under positive selection. Twelve of these TUGs were annotated as peptidases, and the GO term peptidases was significantly overrepresented in the set of positively selected TUGs. These twelve peptidases are deeply conserved, as all have unique orghtologue pairs in *B. malayi* and *C. elgans*. Peptidases have previously been proposed to have acquired prominent roles in host-parasite interactions, and an *A. crassus* trypsin-like proteinase may be utilised by the tissue-dwelling L3 stage to penetrate host tissue and an

aspartyl proteinase may be a blood meal digestive enzyme in adults [2].

The twelve proteinases under positive selection could be targets of adaptive immunity developed against *A. crassus* [15, 79], which is often only elicited against subtypes of larvae [80].

A set of 199 high-credibility SNPs with high information content for population genetic studies was identified by genotyping individual nematodes. The low number of SNPs inferred reflects both the variance in allele contribution introduced transcriptomic data and the stringency of software targeted at higher throughput genome sequence data [81]. Nevertheless, levels of genome-wide heterozygosity found for the five adult nematodes examined are in agreement with existing microsattelite data that show reduced heterozygosity in European populations of *A. crassus* [21]. The polish female nematode was the most highly inbred, while the nematode from the wild *An. japonica* in Taiwan was the most highly outbred.

While the experimental design was not ideal for identification of differential expression between conditions (due to low replication) we used methods developed for comparison of cDNA libraries [50] to infer differential gene expression according to the origin of the sequencing libraries. This approach is widely used with 454 transcriptome data (e.g. [78]).

We can only tentatively infer differential expression of a gene under different conditions (sex, origin) based on identification of significantly differential expression between libraries. Genes over-expressed in the male *A. crassus* included major sperm proteins [83], and, surprisingly, a suite of ribosomal proteins. Collagen processing enzymes are were overexpressed in the female nematodes in line with modulation of collagen synthesis in nematode larval development, and the ovoviviparity of this species [82]. Acetyl-CoA acetyltransferase was identified as overexpressed in European nematodes compared to the Asian one. Acetyl-CoA acetyltransferases act in fatty-acid-oxidation in peroxisomes and mitochondria [83]. Together with a change in steroid metabolism and the enrichment of mitochondrially localised enzymes these suggest changes in the energy metabolism of *A. crassus* from different origins. Possible explanations could include a change to more or less aerobic processes in nematodes in Europe due to their bigger size and/or increased availability of nutrients. TUGs overexpressed in the female libraries showed elevated levels of dn/ds but genes overexpressed in males had decreased levels of dn/ds. The first finding is unexpected, as genes overexpressed in female libraries will also include TUGs related to larval development (such as the collagen modifying enzymes discussed above), and these larval transcripts in turn are expected to be under purifying selection because of pleiotropic effects of genes in early development [84]. The second contrasts with findings that male specific traits and transcripts often show hallmarks of positive selection [85, 86]. In *Ancylostoma caninum* however, female-specific transcripts showed an enrichment of "parasitism genes" [78]

and a possible explanation would be a similar enrichment of positively selected parasitism related genes in our dataset. For males the decreased dn/ds may be explained by the by the high number of ribosomal protein-encoding TUGs, which all show very low levels of dn/ds. That these TUGs were found to be differentially expressed remains puzzling.

Some male-overexpressed TUGs, such as that encoding major sperm protein, showed elevated dn/ds. It is unlikely that correlation of differential expression with positive selection results from mapping artifacts, as all the ribosomal protein encoding TUGs identified overexpressed in males have very low dn/ds.

Genes differentially expressed according to the geographic origin of the nematodes showed significantly elevated levels of dn/ds. We interpret tis as reflecting a correlation between sequence evolution and phenotypic modification in different host environments or correlation between sequence evolution and evolution of gene expression. Whether expression of these genes is modified in different hosts or evolved rapidly in the contemporary divergence between European and Asian populations of *A. crassus*, is one focus of ongoing work building on the reference transcriptome presented here. For such an analysis it will be important to disentangle the influence of the host and the nematode population in a common garden, co-inoculation experiment.

Conclusions

The *A. crassus* transcriptome provides a basis for a new era of molecular research on this ecologically important species. It will aid not only analysis of the invasive biology of this parasite, assisting in identifying the origins of invading opopulations as well as the adaptations that may be being selected in the new European host, but also in the investigation of the acquisition of parasitism in the great clade of animal parasites, Spirurina. In particular, positive selection of proteinases and differences in energy metabolism between European and Asian *A. crassus* constitute a candidate phenotype relevant for phenotypic modification or contemporary divergent evolution as well as for the long term evolution of parasitism.

Competing interests

The authors declare no competing interests.

Authors contributions

EH, HT 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. EH and MB interpreted results and prepared the manuscript. All authors have read and approved the final manuscript.

Acknowledgements

This work has been made possible through a grant provided to EH by Volkswagen Foundation, "Förderinitiative Evolutionsbiologie". The GenePool Genomics facility is core funded by The School of Biological Sciences, University of Edinburgh, the Darwin Trust of Edinburgh, the UK Natural Environment Research Council and the UK Medical Research Council. We are grateful to Karim Gharbi for oversight of the project within the GenePool. Sujai Kumar and Graham Thomas gave essential analytic and informatic support.

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. 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].
3. 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.
4. Neumann W: **Schwimblasenparasit Anguillicola bei Aalen**. *Fischer und Teichwirt* 1985, :322.
5. Koops H, Hartmann F: **Anguillicola-infestations in Germany and in German eel imports**. *Journal of Applied Ichthyology* 1989, **5**:41–45, [http://onlinelibrary.wiley.com/doi/10.1111/j.1439-0426.1989.tb00568.x/abstract].
6. 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].
7. 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].
8. 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.
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 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].
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, 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. 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].

14. 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].

15. 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].

16. 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].

17. 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].

18. 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].

19. 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].

20. Laetsch DR, Heitlinger EG, Taraschewski H, Nadler SA, Blaxter M: **The phylogenetics of Anguillicolidae (Nematoda: Anguillicolidea), swimbladder parasites of eels**. *BMC Evolutionary Biology* 2012, **12**(60), [http://www.biomedcentral.com/1471-2148/12/60].

21. 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].

22. 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].

23. 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].

24. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Federhen S, Feolo M, Fingerman IM, Geer LY, Helmberg W, Kapustin Y, Krasnov S, Landsman D, Lipman DJ, Lu Z, Madden TL, Madej T, Maglott DR, Marchler-Bauer A, Miller V, Karsch-Mizrachi I, Ostell J, Panchenko A, Phan L, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Slotta D, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Wang Y, Wilbur WJ, Yaschenko E, Ye J: **Database resources of the National Center for Biotechnology Information**. *Nucleic Acids Res.* 2012, **40**:13–25, [http://www.ncbi.nlm.nih.gov/pubmed/22140104].

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

26. 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].

27. 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].

28. 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].

29. 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].

30. 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].

31. 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].

32. 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].

33. 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].

34. 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].

35. 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].

36. 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].

37. 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].

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

39. 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].

40. 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].

41. 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].

42. 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].

43. 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].

44. Danecek, P and Auton,†A and Abecasis, G and Albers CA and Banks, E and DePristo, MA and Handsaker RE and Lunter G and Marth GT and Sherry ST and McVean GT and Durbin T and the 1000 Genomes Project: **The variant call format and VCFtools**. *Bioinformatics* 2011, **27**:2156–2158, [http://www.ncbi.nlm.nih.gov/pubmed/21653522].

45. 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].

46. 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].

47. 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].

48. 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].

49. 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].

50. Audic S, Claverie JM: **The significance of digital gene expression profiles**. *Genome Res.* 1997, **7**:986–995, [http://www.ncbi.nlm.nih.gov/pubmed/9331369].

51. Romualdi C, Bortoluzzi S, D'Alessi F, Danieli GA: **IDEG6: a web tool for detection of differentially expressed genes in multiple tag sampling experiments**. *Physiol. Genomics* 2003, **12**:159–162, [http://www.ncbi.nlm.nih.gov/pubmed/12429865].

52. Pages H, Carlson M, Falcon S, Li N: *AnnotationDbi: Annotation Database Interface.* [R package version 1.16.10].

53. Alexa A, Rahnenfuhrer J: *topGO: topGO: Enrichment analysis for Gene Ontology* 2010. [R package version 2.6.0].

54. 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].

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

56. 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].

57. 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].

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

59. Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, Kanapin A, Le Bot N, Moreno S, Sohrmann M, Welchman DP, Zipperlen P, Ahringer J: **Systematic functional analysis of the Caenorhabditis elegans genome using RNAi**. *Nature* 2003, **421**(6920):231–237, [http://www.ncbi.nlm.nih.gov/pubmed/12529635].

60. Harris TW, Antoshechkin I, Bieri T, Blasiar D, Chan J, Chen WJ, De La Cruz N, Davis P, Duesbury M, Fang R, Fernandes J, Han M, Kishore R, Lee R, Muller HM, Nakamura C, Ozersky P, Petcherski A, Rangarajan A, Rogers A, Schindelman G, Schwarz EM, Tuli MA, Van Auken K, Wang D, Wang X, Williams G, Yook K, Durbin R, Stein LD, Spieth J, Sternberg PW: **WormBase: a comprehensive resource for nematode research**. *Nucleic Acids Res.* 2010, **38**(Database issue):D463–467, [http://www.ncbi.nlm.nih.gov/pubmed/19910365].

61. 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].

62. 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].

63. 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].

64. 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].

65. 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].

66. 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].

67. Parkinson J, Mitreva M, Whitton C, Thomson M, Daub J, Martin J, Schmid R, Hall N, Barrell B, RH W, McCarter J, Blaxter M: **A transcriptomic analysis of the phylum Nematoda**. *Nat Genet* 2004, **36**(12):1259–1267, [http://dx.doi.org/10.1038/ng1472].

68. 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].

69. 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].

70. 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].

71. 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].

72. 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].

73. 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].

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

75. 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].

76. 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].

77. 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].

78. 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].

79. 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].

80. 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.

81. Danecek, P and others: **The variant call format and VCFtools**. *Bioinformatics* 2011, **27**:2156–2158, [http://www.ncbi.nlm.nih.gov/pubmed/21653522].

82. Johnstone IL: **Cuticle collagen genes. Expression in Caenorhabditis elegans**. *Trends Genet.* 2000, **16**:21–27, [http://www.ncbi.nlm.nih.gov/pubmed/10637627].

83. Middleton B: **The oxoacyl-coenzyme A thiolases of animal tissues**. *Biochem. J.* 1973, **132**:717–730, [http://www.ncbi.nlm.nih.gov/pubmed/4721607].

84. Cutter AD, Ward S: **Sexual and temporal dynamics of molecular evolution in C. elegans development**. *Mol. Biol. Evol.* 2005, **22**:178–188, [http://www.ncbi.nlm.nih.gov/pubmed/15371532].

85. Eberhard WG: **Evolutionary conflicts of interest: are female sexual decisions different?** *Am. Nat.* 2005, **165 Suppl 5**:19–25, [http://www.ncbi.nlm.nih.gov/pubmed/15795858].

86. 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].

Figures

Figure 1 - Annotation of the Anguilicolla crassus transcriptome

Number of annotated sequences in the transcriptome of *A. crssus* for all TUGs (a) and for highCA derived contigs (b). Annotations with Gene Ontology (GO), Enzyme Commission (EC) and Kyoto Encyclopedia of Genes and Genomes (KEGG) terms were inferred for predicted proteins using annot8r (version 1.1.1) [35]. For highCA contigs additional domain-based annotations obtained with InterProScan [37] are also enumerated.

Figure 2 - Comparing high level GO-slim annotations

Comparing high level GO-slim annotations obtained through annot8r (version 1.1.1) [35] for *A. crassus* to those for the model-nematodes *C. elegans* and *B. malayi* infered using the same pipeline. For GO categories molecular function, cellular compartment and biological process the number of terms in high level GO-slim categories is given. In the two parasitic nematodes a higher degree of congruence in annotation spectrum is observed (0.95; Spearman correlation coefficient) than in comarison to the complete proteome of C. elegans (0.90).

Figure 3 - Enrichment of Signal-positives for categories of evolutionary conservations

Categories of evolutionary conservation recorded using the taxonomy of BLAST-matches at two different bitscore thresholds (50 or 80) are compared for the occurence of signal peptide cleavage sites and signal anchor signatures, predicted using SignalP V4.0 [36]. Contigs were categorised as conserved, novel in the kingdom Metazoa, the phylum Nematoda or nematode clade III *sensu* [17] (Spirurina). TUGs without any hit at a given threshold were categorised as novel in *A. crassus* (Ac).

The highest proportions of genes predicted to have secretory signal peptides are observed in TUGs predicted to be part of gene families that arose in the last common ancestor of Nematoda or to be novel to *A. crassus*.

Tables

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

Summary statistics for the two sequencing libraries of female adults and L2-larvae (L2) of *A. crssus* from *An. anguilla* in Europe (E1 and E2) and one male and two female nematodes from *An. japonica* in Taiwan (T1 and T2). The number of raw sequencing reads and of reads discarded due to low quality or length in Seqclean [26] is given. Additionally the number of reads screened because of similarity to *A. crassus*-rRNA, eel- mRNA and eel- rRNA is given. In the library obtained from L2 larvae hits to Cercozoan rRNA were found and discarded these are addtionally enumerated and the number of reads regarded valid afterthis screening is enumerated and their span (in bases) is given.

Libraries and sequencing read from those are then characterised according to their mapping to the final assembly: The number of reads mapping (using ssaha2 [40]) at all or mapping uniquely to the assembly, the number of reads mapping to the part of the assembly considered to be of *A. crassus* origin (see post-assembly screening) and the number of reads mapping to the highCA-derived part of the assembly (and also of *A. crassus* origin) are given. Finally the number of reads mapping to contigs with overall read counts of more than 32, as considered in gene-expression analysis, is given.

sequencing library	E1	E2	L2	M	T1	T2
lifecycle stage	adult female	adult female	L2 larvae	adult male	adult female	adult female
source population	Europe Rhine	Europe Poland	Europe Rhine	Asia cultured	Asia cultured	Asia wild
raw reads	209325	111746	112718	106726	99482	116366
low quality reads	92744	10903	15653	15484	7947	27683
A. crassus rRNA reads	76403	11213	30654	31351	24929	7233
eel-host mRNA reads	4835	3613	1220	1187	7475	11741
eel-host rRNA reads	13112	69	1603	418	514	38
Cercozoa reads (rRNA)	0	0	5286	0	0	0
valid reads	22231	85948	58302	58286	58617	69671
span of valid reads (in bases)	7167338	24046225	16661548	17424408	14443123	20749177
reads mapping (uniquely)	12023	65398	39690	36782	42529	55966
reads mapping to *A. crassus*-contigs	8359	61073	12917	31673	37306	50445
reads mapping highCA contigs	5883	48009	8475	18998	28970	41963
reads mapping to contigs with count >32	3595	34115	1602	10543	21413	22909

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 numbers of total contigs, number of contigs screened due to similarity to eel- mRNA, non-eukaryote or Chordata sequences in NCBI-nr or NCBI-nt and of contigs remaining after this screening, additionally the total length of remaining contigs is given.

The mean per base coverage of these contigs (as obtained from mapping of reads to contigs using ssaha2 [40]) is given without (non-unique), and with discarding of reads mapping to multiple locations.

The number protein predictions derived with prot4Est [32] using BLAST similarities, ESTscan or the longest ORF is enumerated. The number of contigs considered complete and complete at 3' end and at 5' end is given. The number of contigs annotated with GO-terms, KEGG-pathways, EC-numbers, with BLAST-hit to nematode in nr and with BLAST-hit to nematode or non-nematode (eukaryote non chordate) sequence in NCBI-nr is given.

	lowCA	highCA	combined
total contigs	26336	13851	40187
contigs hitting rRNA	829	59	888
contigs hitting eel-mRNA or Chordata	2419	1022	3441
non-eukaryote contigs	1935	1398	3333
contigs remaining	21153	11372	32525
total span of remaining contigs (in bases)	6157974	6575121	12733095
non-unique mean base coverage of contigs	14.665	10.979	12.840
unique mean base coverage of contigs	2.443	6.838	4.624
protein predictions by BLAST similarity	4357	5664	10021
protein predictions by ESTscan	8324	3597	11921
protein predictions by longest ORF	8352	2085	10437
contigs without protein prediction	93	14	107
contigs with complete 3' end	5909	2714	8623
contig with complete 5' end	1484	1270	2754
full length contigs	104	185	289
contigs with GO- annotation	2636	3875	6511
contigs with EC- annotation	967	1493	2460
contigs with KEGG- annotation	1609	2237	3846
contigs with InerProScan- annotation	0	7557	7557
contigs with BLAST hit to nematode	4869	5821	10690
contigs with any BLAST hit	5107	6008	11115

Table 3 - Evolutionary conservation and novelty

Recording the taxonomy of all BLAST-matches at two different bitscore thresholds (50 or 80) contigs were categorised as conserved, novel in the kingdom Metazoa, the phylum Nematoda or nematode clade III *sensu* [17] (Spirurina). TUGs without any hit at a given threshold were categorised as novel in *A. crassus* (Ac).

The number of all TUGs and highCA contigs respectively by conservation- category. Numbers for conservation would be obtained as the cumulative sum of lower taxonomy- level novelty.

	conserved	Metazoa	Nematoda	Clade3	Ac
50 all	5604	1715	2173	1485	21548
80 all	3506	1383	2015	1525	24096
50 highCA	3479	876	1010	601	5406
80 highCA	2457	833	1084	716	6282

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

The annotation graph for the GO- ontolgy was traversed using the R-package topGO [53] and over-representation was analysed at each node term comparing the abundance of positively selected (dn/ds > 0.5 genes) to an universal gene set of all contigs with dn/ds values. Terms for which an offspring term was already in the table and no additional counts supported overrepresentation were removed.
Significantly ($p<0.05$) over-represented GO-terms in contigs putatively under positive selection are tabulated. Horizontal lines separate categories of the GO-ontology. First category is molecular function, second biological process, last cellular compartment. P-values (p.value) for over-representation are given along with the number of positively selected contigs (Significant; dn/ds > 0.5) and the number of universal contigs (Annotated) and the description of the GO-term (Term). For a graph representation of induced GO-terms see also Additional Figures 4.

GO.ID	Term	Annotated	Significant	Expected	p.value
GO:0008233	peptidase activity	43	13	6.08	0.0034
GO:0015179	L-amino acid transmembrane transporter activity	2	2	0.28	0.0198
GO:0043021	ribonucleoprotein complex binding	6	3	0.85	0.0396
GO:0070011	peptidase activity, acting on L-amino acid peptides	35	9	4.95	0.0442
GO:0004175	endopeptidase activity	25	7	3.54	0.0488
GO:0042594	response to starvation	15	7	2.13	0.0022
GO:0009083	branched chain family amino acid catabolic process	3	3	0.43	0.0027
GO:0006914	autophagy	12	6	1.70	0.0031
GO:0009063	cellular amino acid catabolic process	10	5	1.42	0.0071
GO:0009267	cellular response to starvation	7	4	0.99	0.0093
GO:0006520	cellular amino acid metabolic process	44	12	6.24	0.0128
GO:0006915	apoptotic process	78	18	11.06	0.0147
GO:0009308	amine metabolic process	57	14	8.08	0.0189

GO:0005997	xylulose metabolic process	2	2	0.28	0.0199
GO:0006739	NADP metabolic process	2	2	0.28	0.0199
GO:0007616	long-term memory	2	2	0.28	0.0199
GO:0009744	response to sucrose stimulus	2	2	0.28	0.0199
GO:0010172	embryonic body morphogenesis	2	2	0.28	0.0199
GO:0015807	L-amino acid transport	2	2	0.28	0.0199
GO:0050885	neuromuscular process controlling balance	2	2	0.28	0.0199
GO:0007281	germ cell development	17	6	2.41	0.0226
GO:0090068	positive regulation of cell cycle process	17	6	2.41	0.0226
GO:0042981	regulation of apoptotic process	64	15	9.07	0.0232
GO:0051329	interphase of mitotic cell cycle	23	7	3.26	0.0320
GO:0044106	cellular amine metabolic process	55	13	7.80	0.0325
GO:0031571	mitotic cell cycle G1/S transition DNA damage checkpoint	14	5	1.98	0.0355
GO:0010564	regulation of cell cycle process	34	9	4.82	0.0377
GO:0006401	RNA catabolic process	6	3	0.85	0.0398
GO:0010638	positive regulation of organelle organization	6	3	0.85	0.0398
GO:0009056	catabolic process	149	28	21.12	0.0398
GO:0008219	cell death	93	19	13.18	0.0441
GO:0007154	cell communication	144	27	20.41	0.0455
GO:0051726	regulation of cell cycle	52	12	7.37	0.0474
GO:0030330	DNA damage response, signal transduction by p53 class mediator	15	5	2.13	0.0475
GO:0033238	regulation of cellular amine metabolic process	15	5	2.13	0.0475
GO:0030532	small nuclear ribonucleoprotein complex	7	4	0.99	0.0093
GO:0005739	mitochondrion	137	28	19.38	0.0113
GO:0005682	U5 snRNP	2	2	0.28	0.0198
GO:0015030	Cajal body	2	2	0.28	0.0198
GO:0046540	U4/U6 x U5 tri-snRNP complex	2	2	0.28	0.0198
GO:0016607	nuclear speck	6	3	0.85	0.0396

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

Genotyping using Samtools [41] and Vcftools [44] resulted in the identification of 199 SNPs infromative at the individual level. Different measurements were obtained to asses genome-wide heterozygosity using the R- package Rhh [45]: measurements for relative heterozygosity (number of homozygous sites/ number of heterozygous sites), internal relatedness [46], homozygosity by loci [47] and standardised heterozygosity [48] are given with the number of SNPs informative for the respective library. All these measurements are pointing to sample T1 (Taiwanese worm from aquaculture) as the most heterozygous and sample E2 (the European worm from Poland) as the least heterozygous individual. Heterozygote-heterozygote correlation [45] confirmed the genome-wide significance of these markers.

	relative heterozygosity	internal relatedness	homozygosity by loci	standardised heterozygosity	informative SNPs
T2	0.45	-0.73	0.59	1.00	121
T1	0.93	-0.95	0.34	1.62	136
M	0.37	-0.73	0.66	0.84	92
E1	0.38	-0.83	0.60	0.91	65
E2	0.18	-0.35	0.82	0.50	140

Table 6 - Over- representation of GO-terms differentially expressed

The annotation graph for the GO- ontolgy was traversed using the R-package topGO [53] and over-representation was analysed at each node term comparing the abundance of differentially expressed genes to an universal gene set of all contigs tested for overexpression. Terms for which an offspring term was already in the table and no additional counts supported overrepresentation were removed. Significantly ($p<0.05$) over-represented GO-terms in contigs differentially expressed between male and female worms (a) or between European and Asian origin (b). Horizontal lines separate categories of the GO-ontology. First category is molecular function, second biological process, last cellular compartment. P-values (p.value) for over-representation are given along with the number of differentially expressed contigs (Significant) and the number of universal contigs with this annotation (Annotated) and the description of the GO-term (Term). For a graph of incuced GO-terms see also Addional Figures 4.

a)

GO.ID	Term	Annotated	Significant	Expected	p.value
GO:0005198	structural molecule activity	52	18	8.39	0.00024

GO:0016706	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxyge...	3	3	0.48	0.00400
GO:0004656	procollagen-proline 4-dioxygenase activity	2	2	0.32	0.02562
GO:0048731	system development	147	35	23.00	0.00021
GO:0034621	cellular macromolecular complex subunit organization	73	22	11.42	0.00024
GO:0034641	cellular nitrogen compound metabolic process	161	37	25.19	0.00024
GO:0071822	protein complex subunit organization	71	21	11.11	0.00050
GO:0043933	macromolecular complex subunit organization	82	23	12.83	0.00057
GO:0032774	RNA biosynthetic process	72	21	11.26	0.00063
GO:0000022	mitotic spindle elongation	20	9	3.13	0.00122
GO:0006139	nucleobase-containing compound metabolic process	141	32	22.06	0.00189
GO:0048856	anatomical structure development	189	39	29.57	0.00237
GO:0071841	cellular component organization or biogenesis at cellular level	140	31	21.90	0.00418
GO:0071842	cellular component organization at cellular level	136	30	21.28	0.00575
GO:0090304	nucleic acid metabolic process	107	25	16.74	0.00673
GO:0040007	growth	139	30	21.75	0.00867
GO:0050789	regulation of biological process	199	39	31.13	0.00928
GO:0016070	RNA metabolic process	98	23	15.33	0.00988
GO:0007275	multicellular organismal development	222	42	34.73	0.01108
GO:0009791	post-embryonic development	117	26	18.30	0.01201
GO:0042157	lipoprotein metabolic process	7	4	1.10	0.01306
GO:0040010	positive regulation of growth rate	62	16	9.70	0.01505
GO:0018996	molting cycle, collagen and cuticulin-based cuticle	23	8	3.60	0.01557
GO:0042274	ribosomal small subunit biogenesis	11	5	1.72	0.01706
GO:0022414	reproductive process	109	24	17.05	0.02003
GO:0032501	multicellular organismal process	242	44	37.86	0.02062
GO:0071840	cellular component organization or biogenesis	172	34	26.91	0.02135
GO:0010467	gene expression	116	25	18.15	0.02279
GO:0009416	response to light stimulus	8	4	1.25	0.02310
GO:0008543	fibroblast growth factor receptor signaling pathway	2	2	0.31	0.02407

GO:0018401	peptidyl-proline hydroxylation to 4-hydroxy-L-proline	2	2	0.31	0.02407
GO:0046887	positive regulation of hormone secretion	2	2	0.31	0.02407
GO:0071570	cement gland development	2	2	0.31	0.02407
GO:0016043	cellular component organization	168	33	26.28	0.02835
GO:0009792	embryo development ending in birth or egg hatching	124	26	19.40	0.02873
GO:0009152	purine ribonucleotide biosynthetic process	5	3	0.78	0.02876
GO:0000279	M phase	45	12	7.04	0.02921
GO:0002164	larval development	107	23	16.74	0.03246
GO:0065007	biological regulation	218	40	34.10	0.03745
GO:0042254	ribosome biogenesis	22	7	3.44	0.03929
GO:0048518	positive regulation of biological process	127	26	19.87	0.04015
GO:0022613	ribonucleoprotein complex biogenesis	27	8	4.22	0.04202
GO:0007010	cytoskeleton organization	58	14	9.07	0.04305
GO:0000003	reproduction	141	28	22.06	0.04750
GO:0044267	cellular protein metabolic process	135	27	21.12	0.04864
GO:0005634	nucleus	163	38	25.71	0.00010
GO:0030529	ribonucleoprotein complex	64	20	10.09	0.00034
GO:0043232	intracellular non-membrane-bounded organelle	116	28	18.30	0.00187
GO:0044444	cytoplasmic part	260	48	41.01	0.00194
GO:0043231	intracellular membrane-bounded organelle	253	47	39.91	0.00294
GO:0005829	cytosol	151	33	23.82	0.00359
GO:0031981	nuclear lumen	68	18	10.73	0.00725
GO:0005618	cell wall	18	7	2.84	0.01279
GO:0043229	intracellular organelle	272	48	42.90	0.01372
GO:0070013	intracellular organelle lumen	94	22	14.83	0.01377
GO:0044446	intracellular organelle part	195	38	30.76	0.01470
GO:0009536	plastid	28	9	4.42	0.01871
GO:0045169	fusome	2	2	0.32	0.02446
GO:0070732	spindle envelope	2	2	0.32	0.02446
GO:0022627	cytosolic small ribosomal subunit	16	6	2.52	0.02606
GO:0005791	rough endoplasmic reticulum	5	3	0.79	0.02939
GO:0009507	chloroplast	26	8	4.10	0.03508
GO:0005773	vacuole	46	12	7.26	0.03660
GO:0005811	lipid particle	31	9	4.89	0.03690

b)

GO.ID	Term	Annotated	Significant	Expected	p.value
GO:0016453	C-acetyltransferase activity	3	3	0.37	0.0018
GO:0003824	catalytic activity	158	27	19.50	0.0079
GO:0016746	transferase activity, transferring acyl groups	8	4	0.99	0.0097
GO:0003682	chromatin binding	2	2	0.25	0.0149
GO:0003985	acetyl-CoA C-acetyltransferase activity	2	2	0.25	0.0149
GO:0008061	chitin binding	2	2	0.25	0.0149
GO:0003713	transcription coactivator activity	6	3	0.74	0.0268
GO:0005543	phospholipid binding	6	3	0.74	0.0268
GO:0004090	carbonyl reductase (NADPH) activity	3	2	0.37	0.0412
GO:0008289	lipid binding	12	4	1.48	0.0473
GO:0016853	isomerase activity	12	4	1.48	0.0473
GO:0016126	sterol biosynthetic process	5	4	0.60	0.00081
GO:0044281	small molecule metabolic process	106	22	12.68	0.00090
GO:0048732	gland development	9	5	1.08	0.00169
GO:0006694	steroid biosynthetic process	10	5	1.20	0.00307
GO:0006338	chromatin remodeling	4	3	0.48	0.00586
GO:0006695	cholesterol biosynthetic process	4	3	0.48	0.00586
GO:0042180	cellular ketone metabolic process	57	13	6.82	0.00800
GO:0023051	regulation of signaling	29	8	3.47	0.01318
GO:0001822	kidney development	2	2	0.24	0.01399
GO:0006611	protein export from nucleus	2	2	0.24	0.01399
GO:0009953	dorsal/ventral pattern formation	2	2	0.24	0.01399
GO:0048581	negative regulation of post-embryonic development	2	2	0.24	0.01399
GO:0051124	synaptic growth at neuromuscular junction	2	2	0.24	0.01399
GO:0070050	neuron homeostasis	2	2	0.24	0.01399
GO:0019752	carboxylic acid metabolic process	54	12	6.46	0.01417
GO:0008152	metabolic process	268	37	32.06	0.01595
GO:0019219	regulation of nucleobase-containing compound metabolic process	42	10	5.02	0.01617
GO:0006355	regulation of transcription, DNA-dependent	30	8	3.59	0.01637
GO:0010033	response to organic substance	62	13	7.42	0.01729
GO:0019953	sexual reproduction	44	10	5.26	0.02265
GO:0048747	muscle fiber development	6	3	0.72	0.02461
GO:0032787	monocarboxylic acid metabolic process	21	6	2.51	0.02763

GO:0051171	regulation of nitrogen compound metabolic process	52	11	6.22	0.02827
GO:0048545	response to steroid hormone stimulus	16	5	1.91	0.03065
GO:0048609	multicellular organismal reproductive process	60	12	7.18	0.03331
GO:0050794	regulation of cellular process	152	24	18.18	0.03448
GO:0009966	regulation of signal transduction	22	6	2.63	0.03462
GO:0065008	regulation of biological quality	82	15	9.81	0.03600
GO:0009308	amine metabolic process	41	9	4.90	0.03874
GO:0002026	regulation of the force of heart contraction	3	2	0.36	0.03877
GO:0007595	lactation	3	2	0.36	0.03877
GO:0030518	intracellular steroid hormone receptor signaling pathway	3	2	0.36	0.03877
GO:0034612	response to tumor necrosis factor	3	2	0.36	0.03877
GO:0035071	salivary gland cell autophagic cell death	3	2	0.36	0.03877
GO:0035220	wing disc development	3	2	0.36	0.03877
GO:0043628	ncRNA 3'-end processing	3	2	0.36	0.03877
GO:0045540	regulation of cholesterol biosynthetic process	3	2	0.36	0.03877
GO:0051091	positive regulation of sequence-specific DNA binding transcription factor activity	3	2	0.36	0.03877
GO:0051289	protein homotetramerization	3	2	0.36	0.03877
GO:0002165	instar larval or pupal development	7	3	0.84	0.03951
GO:0007589	body fluid secretion	7	3	0.84	0.03951
GO:0048872	homeostasis of number of cells	7	3	0.84	0.03951
GO:0060047	heart contraction	7	3	0.84	0.03951
GO:0006066	alcohol metabolic process	35	8	4.19	0.04124
GO:0007165	signal transduction	69	13	8.25	0.04239
GO:0006357	regulation of transcription from RNA polymerase II promoter	12	4	1.44	0.04276
GO:0006351	transcription, DNA-dependent	42	9	5.02	0.04489
GO:0007276	gamete generation	42	9	5.02	0.04489
GO:0031967	organelle envelope	47	12	5.49	0.0031
GO:0005740	mitochondrial envelope	29	8	3.38	0.0112
GO:0005643	nuclear pore	2	2	0.23	0.0133
GO:0005739	mitochondrion	93	17	10.85	0.0173
GO:0031966	mitochondrial membrane	28	7	3.27	0.0312
GO:0005902	microvillus	3	2	0.35	0.0369

Additional Files

Additional text

The additional text describes the assembly process and evaluation of assembly quality in further detail. This text contains 6 figures and 4 tables.

Additional tables

Additional table 1 (a + b): All data computed on the contig level, (a) including sequences (raw, coding, imputed, protein), or (b) listing only the metadata.

Additional table 2: High quality SNPs. The contig, the base relative to the start of the contig (base), the reference base-call (from), the alternative base-call (to), the number of reads supporting the reference (nfrom) and the alternative (nto), the percentage of the alternate allele (perc), whether the SNP is in the region of an ORF (inORF), the position in the Frame (inFrame) and the effect of the SNP (effect; synonymous, non-synonymous or nonsense) are given.

Additional table 3 (a + b): Contigs differentially expressed between male and female worms (a) and European and Asian worms (b). Normalised counts and the natural logarithm of fold changes are given.

Additional figures

Additional Figure 1: 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).

Additional Figure 2: Overabundance of SNPs at codon-position two (a) and of non-synonymous SNPs (c) 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) absence of such a correlation afterwards (R^2 <0.001, p =0.192).

Additional Figure 3: Box-plots for dn/ds in TUGs according to different categories of evolutionary conservation. Significant comparisons are sequences novel in Metazoa vs. novel in *A. crassus* (0.009 and 0.002; p-value for bitscore of 50 and 80, Nemenyi-Damico-Wolfe-Dunn test), in Nematoda vs. in *A. crassus* (0.03 and 0.009). Sequences novel in cladeIII failed to show significantly elevated dn/ds, despite higher median values due to the low number of contigs of this category with a dn/ds obtained.

Additional Figures 4: Subgraphs of the GO-ontology categories induced by the top 10 terms identified as enriched in different sets of genes. Boxes indicate the 10 most significant terms. Box colour represents the relative significance, ranging from dark red (most significant) to light yellow (least significant). In each node the category-identifier, a (eventually truncated) description of the term, the significance for enrichment and the number of DE / total number of annotated gene is given. Black arrows indicate a is "is-a" relationship. GO-ontology category and the set of genes analysed for the enrichment are indicated in each figure.