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

Emanuel G Heitlinger*[1,2] 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

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

*Corresponding author

Abstract

Background:

Results:

Conclusions: Yeh!

Background

Anguillicola crassus Kuwahara, Niimi et Itagaki, 1974 [1] is a nematode feeding on blood in the swimbladder of freshwater eels of the genus *Anguilla*. 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-range to Europe and the European eel (*Anguilla anguilla*). Soon after it had been recorded for the first time in 1982 in North-West Germany [2], to where it was most likely introduced by live-eel trade [3, 4], *A. crassus* rapidly spread throughout populations of its newly acquired host (for a review see [5]). At the present day it is found in all but the northernmost population of the European eel in Iceland [6].

The impact of *A. crassus* on the European eel has been a major focus of research during the past decades. High prevalence of the parasite of above 70% (e.g. [7]), as well as high intesities of infections were reported, throughout the newly colonized area [8]. Based on a broad base of work on its epidemiology *A. crassus* can be regarded as a model for parasite introduction and spread [9].

As in the natural host in Asia prevalences and intesities are lower [10], high epidemiological parameters were attributed to the inadequate immune-response of the European Eel [11]. Interestingly the differences in the two host also affect the size and life-history of the worm: In European eels the nematodes are bigger and develop and reproduce faster [12]. While the Japanese eel is capable of killing larvae of the parasite after vaccination [13] or under high infection pressure [14], only pathological effects such a thikening of the swimbladder wall [15] have been found in the European eel.

The genus *Anguillicola* holds a phylogenetic position basal to the Spirurina (clade III *sensu* Blaxter [16]), one of 5 major clades of nematodes [17, 18]. The Spirurina exclusively exhibit a parasitic lifestyle and comprise improtant human pathogens as well as prominent parasites of livestock (e.g. the Filaroidea and Ascarididae). This phylogenetic position makes the Anguillicoloidae an interesting system in the endeavour to understand the emergence of parasitism in Spirurina and as an "outgroup" for functional studies of parasitism in this clade. Some functionally interesting genes in this respect are thought to be under diversifying selection in an arms-race between host and parasite [19].

Recent advances in sequencing technology (often termed Next Generation Sequencing; NGS), provide the opprotunity for rapid and cost-effective generation of genome-scale data. The longer read length of 454-sequencing [20] compared to other NGS technologies, allows *de novo* assembly of Expressed Sequence Tags (ESTs) in organisms lacking previouse genomic or transcriptomic data (for a comprehensive list of studies using this approach before Oct 2010 see [21]). Such transcriptomic datasetes are still less expensive than genomic data-sets in terms sequencing costs and analytical needs.

The difference of the immune attack on *A. crassus* in the two diffent hosts provides an opportunity to investigate the parasite's response to different "immune environments" on a transcriptomic basis.

In this study we report assembly of raw reads from cDNA libraries of L2 larvae, four female adults and one male adult of *A. crassus* into contigs (continuous sequence, representing a mRNA-transcript) of a reference transcriptome. We describe screening for xenobiotics and host-contamination, prediction of protein sequences and functional annotation of these contigs. Furthermore combining ORF prediction and identification of SNPs allowed inferrence of diversifying selection.

As shorter read-length but higher throughput of the Illumina-Solexa platform provides superior means for

gene expression analyis [22], the transcriptome assembly presented here is planned to be used as a reference in a future study using gene-expression tags (SuperSAGE [23]) or RNA-seq [24]. To test the suitability of the presented transcriptome we mapped a pilot-set of expression-tags. This analysis is used as an additional estimate for the correctness of the hypotesis that each contig in the assembly constitutes.

Results

Sequencing, trimming and pre-assembly screening

In total 756363 raw sequencing reads were trimmed and quality screened. After this trimming 585949 sequences were regarded valid, table 1 gives an overview of trimming statistics for different libraries. Notably the variablity in the number of sequences excluded due to low quality and short quality-sequence in `seqclean` was high indicating different properties of read-quality in the different libraries. The 585949 reads regarded valid after this trimming were further screened for rRNA and host-contamination before assembly.

In total 232894 reads were discarded prior to assembly, 202823 due to `blast` hits (e-value cut of 1e-5, hit longer than 80% of the read-length and identity bigger 95%) to rRNA, 30071 to host-databases (same thresholds). Figure (rRNA_plots.png) and table 2 give an overview of this pre-assebmly rRNA and host-screening in different libraries. During annotoation of the rRNA-database we constructed for *A. crassus* (assembled from reads screened against general rRNA-databases) we found some contigs with high similarity 18S sequences from cercozoan and other flaggelate protists over the majority of their length. Table 3 gives an overview of the hits in NCBI-nt for these 11 contigs assembled from rRNA sequences. The origin of the rRNA-reads assembled in these contigs (in screening-database assembly, not the main assebmly, where they were screened out) showed that they originated nearly exclusively from library L2. Containing the highest number of raw reads, library E1 showed the highest amount of low-quality and short sequences in trimming and a high amount of rRNA and host-contamination, making it the library with the lowest number of sequences used for assembly.

Assembly

For assembly we used a method proposed by [21]: We combined two assemblies merging their contigs. For a detailed description of the assembly method see the supplementary methods file. In total 13851 contigs were regarded good quality, as they were supported by two assembly algorithms (in the remainder of the manuscript “the assembly” referres to this “second-order”-assembly and quality contigs referres to the

good-quality contig-category from this). In additon an expanded dataset consisting of the afore mentioned assembly plus the (3745) contigs only supported by one of the "first-order" assemblies and (22591) singletons was obtained (this is reffered to as the "fullest assembly", the sequences in this dataset are referred to as Tentative Unique Genes; TUGs). In the methods supplementary file we show that these datasets are considered the most parsiomonous (having the smallest size) for their quality (covering the largest amount of sequence in reference transcritomes). In the high quality assembly parsimony and low redundancy is prioritzsed, in the full assembly completeness.

The mean per-base coverage of the TUGs was 12.84 and the mean per-base unique coverage was 4.62. The mean per-base coveage of the quality contigs was lower with 10.98, but the unique coverage higher with 6.84. This indicates a higher amount of redundancy in the full assebmly set compared to the good-quality set. The distribution of mean per base coverages for single contigs or TUGs is given in Figure (coverage_per_contig.png)

Figure (coverage_plots) shows the realtionship between, read-number in the assemblies and coverage.

Marking possible host- and xenobiotic contamination

To elucidate the taxonimic origin of the assembled sequences, we used extensive `blast`-queries.

Host-contamination approach

In this evalutation we labeled all sequences hitting (e-value cut-off of 1e-5) the same databases used for screening prior to assembly (eel-mRNA, eelrRNA, *A. crassus*rRNA) with the respective database-hit, if the sequence did not produce a better hit (based on bit-score comparison) against the nucleotide version of nempep4 [25].

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

Two kinds sequences were labled "valid" *A. crassus* sequences: Valid due to evidence of nematode origing (hitting nempep4), and valid due to missingness of evidence of host-origin.

Table 3 gives the numbers of TUGs labled as possible host or rRNA contamination. The 38371 TUGs regarded to be of *A. crassus*-origin give a positive indication of the origin of sequence data. However TUGs hitting *A. crassus*-rRNA still have a high coverage, indicating a noticable loss of raw-read data (e.g. used in mapping) to additional rRNA-coverage. In contrast TUGs labeled to be of host-origin had a slighly lower coverage than TUGs labled valid, indicating an only moderate amount of raw-data being "lost" to

this kind of off-target data.

Off target data approach

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

For evey study of the biology of *A. crassus* based on the presented dataset we advise to use the taxonomic evaluation and host-contamination screening at latest at the stage of interpretation of results to avoid missleading conclusions based on differences in off target data.

For example a study excluding putative host-derived sequence prior to analysis, would have 36166 TUGs or 12796 quality-contigs left for analysis, based on choosing only sequence not labled "Chordata" (from nt- and nr-searches) and labled "valid" or "valid-nempep" (from the evaluation of host-contamination). A second analysis aiming to exclude all data suspicious to be not of *A. crassus*-origin would choose 31893 TUGs or 10999 quality-contigs, by using sequence labled "Nematoda" or "no hit" (from nt- and nr-searches) and "valid" or "valid-nempep". We use such restiricted dataset like in the second example in later analyses to verify that results are not induced by a possible contamination backround.

Evolutionary conservation

Based on taxonomically classified `blast`-results described above we also determined evolutionary at 2 thresholds (bitscore of 50 and 80; roughly equivalent with an e-value of e-6 and e-10 respectively) at three taxonomic levels: We categorized a sequence if it had a hit to "nematoda" and a second phylum as conserved across the phylum level and to "metazoa" and a second kindom as conserved across the kingdom level. As a third taxonomic level we defined the Spirurina (Clade III *sensu* [16]). Conversely from these data "conserved", "novel in metazoa", "novel in nematoda" and "novel in clade III" categories could be computed. See table 5 for a summary of these categories for the two thresholds.

Protein prediction

For in total 39625 TUGs a protein was predicted. A summary of protein prediction using the different methods implemented in prot4EST is given in table 6.

The full open reading frame was obtained for in total 414 TUGs, while for 3304 the 5' end and for 10178 the 3' end was complete. A predicted ORF was defined at complete if it was starting with an

"atg"-start-codon, having at least 3 non-coding bases 5' of the start-codon and having a "tag" (2122 observations), "taa" (4989) or "tga" (3026) stop-codon 3' of the ORF.

In additon to the protein-sequence used for annotation we obtained a corrected nucleotide-sequence using the nucleotide equivalent of the protein-prediction. This was necessary as `prot4Est` joins high-scoring pairs from `blast`-searces introducing gaps (or unkonwn fill-bases respectively) relative to the raw nucleotide sequence if needed to obtain the correct frame for the ORF. In 15988 TUGs the corrected sequence with the imputed ORF was slightly changed compared to the raw sequence.

SNP analysis

In order to improve SNP-calling we screened a set of 13518 candidate-SNPs predicted on in total 15718866 bases (1412806 bases with a minimum coverage of more then 8-fold available for SNP-calling). We first excluded SNPs predicted to be present as more than 2 alleles and to undetermined ("N") base in the reference retaining 13042 SNPs. The ratio of transitions (8663) to transversion (4379) in this set was 1.98 (ti/tv).

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

8309 of the raw SNPs were predicted to be inside an ORF, 2713 at a first positon in a codon, 2172 on a second base and 3424 on a third base in a codon.

As expected ti/tv inside ORFs was with 2.43 higher than outside ORFs with 1.42.

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

For the raw set of candidate-SNPs the ratio of synonymous polymorphisms per synonymous site to non-synonymous polymorphisms per non-synonymous site (dn/ds) was 0.45.

To improve SNP-calling with respect to the well-known homopolymer issues in 454-sequencing [27], we analysed the effect of exclusion of SNPs in, or close to, homopolymer regions. We observed changes in ti/tv and in dn/ds when SNPs were discareded due to different sizes thresholds for homoploymer-runs and

different proximity thresholds (see figure snp_ex_parameter_plots.png). Based on this we decided to exclude SNPs with a homopolymer-run as long as or longer than 4 bases inside a window of 11 bases (5 to bases to the right, 5 to the left) around the SNP.

After these screening steps based on a sequence-features we investigated the effect of data volume and mapping on the remaining candidate SNPs. Based on a slight overrepresentation of SNPs on first- and second-positions in the ORF and on a little bigger than average excess of non-synonymous polymorphisms in SNPs with a percentage of the minority allele smaller than 7% (see figure snp_pos_eff_plots.png a and b) we examined exclusion of SNPs at this percentage threshold. Evaluation of an exclusion of SNPs based on coverage at the SNP-site, did not seem necessary (see figure snp_pos_eff_plots.png c and d).

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

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

Figure dens_dn_ds.png gives the disribution of per-contig dn/ds for the fully screened set of SNPs. The final numbers of SNPs per kilo-base was 4.44, 7.87 synonymous SNPs per 1000 synonymous bases and 2.43 non-synonymous SNPs per 1000 non-synonymous bases. For a total of 980 TUGs (858 high quality contigs) a value for dn/ds could be obtained, because at least one synonymous SNP was found.

It should be noted that the overall dn/ds value of 0.32 is different from the mean dn/ds value of contigs 0.233367082602961, as in the first the complete length of the sequence (also from TUGs without synonymous SNP predicted) can be included.

Annotation

```
[1] 1
```

We obtained annotations with GO-terms for 9569, with EC-numbers for 3741 and with KEGG-pathways for 5990 TUGs using the `Blast`-based program annot8r.

Figure (annotataionVenn.tiff) gives an overview of the 10274 TUGs annotations were found for with different methods.

We compared GO-annotations for high-level GO-slim terms to the annotations obtained the same way for all *B. malayi*-proteins: Our transcriptome for *A. crassus* shows a remarkably similar distribution of GO-terms to the full set of annotations for the genome of the related nematode (see figure go_bm_com.png).

In addition we inferred presence of signal peptide cleavage sites in the predicted protein sequence: `SignalP` [28] predicted 4544 signal peptides using neural networks and 3352 and 2425 signal peptides and signal anchors respectively using hidden Markov models [29].

Consolidated results

Signal-positives have higher dn/ds

TUGs predicted to contain signal peptide cleavage sites by the neural-networks method in `SignalP` have higher dn.ds values than TUGs without signal peptide cleavage sites ($p = 0.053$; two sided U-test; see also figure sigp_dn_ds.png a). TUGs predicted by the hmm method in `SignalP` to contain a signal peptide have higher dn/ds than TUGs predicted to contain no signal peptide or a signal anchor ($p = 0.178$; Two sided U-test; see also figure sigp_dn_ds.png b). Although the differences were only significant for the nn-predicted signal-peptides the results of the hmm prediction show that the signal-peptides that are cleaved and possibly secreted are making the difference. These differences were consistent if only tested against a restricted data-set surely from *A. crassus*.

Enrichment of GO-categories in high dn/ds

In enrichment analysis, where the "gene-universe" to test against has to be defined, possible off-target sequences have to be excluded prior to analysis. Testing was performed against a universe of only TUGs we were sure to be of *A. crassus*-origin. We defined TUGs with a dn.ds higher than 0.5 as positively selected and tested each node-term in the ontology for a over-representation in this set. Tables X give an overview of enriched terms for different parts of the ontology.

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

TODO: BP and CC over-representation.

To make sure these inferences were not biased by redundancy in the fullest data-set we repeated the

analysis using only the good-quality contigs as gene-universe. The results showed to be consistent with those obtained for the fullest data-set (data not shown).

Novel in cladeIII have elevated dn/ds

Figure conservation_dn_ds.png shows differences in dn/ds across the categories defined for evolutionary conservation. At a bitscore threshold of 80 sequences novel in clade III had a significantly higher dn/ds than other sequences ($p = 0.08$; two sided U-test). At a bitscore thershold of 50 results were non-significant but showed the same trend.

Sequences novel in nematodes are enriched for Signal-positives

Figure signal_novel.png gives the proportions of `SignalP`-predictions for each category of evolutionary conservation. Generally - across bit-score thresholds and `SignalP` prediction methods - sequences novel in nematodes contained the highest proportion of signal-positives followed by sequences novel in clade III or in metazoa. Conserved sequences contain the lowest proportion of signal-positives.

Differential expression

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

Comparison with tag-sequencing pilot data-set

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

Table 9a gives correlations coefficients between tag-counts and 454-libraries.

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

Correlations between library T2 and other 454-libraries, as well as with solexa-tag counts were lower than between other libraries.

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

Differential expression between male and female worms

Despite the lack of replicates for male worms we were able to identify 25 sequences being significantly over-expressed in male worms. In fact all these TUGs were nearly exclusively expressed in males.

Differential expression between adults and L2-larvae

For the L2-library we changed our approach and used gene-expression analysis rather to highlight the off-target data in this library. For this reason we used counts for the fullest assembly.
479 sequences being expressed exclusively in L2 library were strongly enriched in sequences being labeled as possible off-target data in taxonomic classification. From these sequences only 57 had best hits to metazoa and only 6 to nematoda.

Differential expression between worms from the European and Japanese eel

None of the TUGs in the present evaluation showed significant differential expression between worms from the European and Japanese Eel. Diagnostic plots provided by `DESeq` made clear, that both depth of sequencing and number of replicates have to be higher contrasting these conditions.
However, comparing expression-analysis on the full data-set to analysis limited to the high quality of reliable *A. crassus*-contigs it was clear that the quality-data-set reduces within-condition variance and results were closer to significance: The lowest adjusted p-values for the cleaned data-set were around 0.4, while on the full data-set only adjusted p-values above 0.8 could be obtained.

Secreted immunomodulatory molecules

We further highlight a list of genes with similarities to known immunumodulators. Should I???

Discussion

We are providing transcriptome-data for the parasite *A. crassus*, enabling a broad spectrum of molecular research on this ecologically and economically important species.
We emphasize the importance of screening for xenobiotics. We consider this aspect important in any deep transcriptome project. First the depth of sequencing is leading to the generation of large amounts of

off-target data from a "metatranscriptomic community" associated with a target organism. Second due to the abundance of laboratory contamination and the possibility of cross-contamination if libraries are sequenced only on a subset of a picotiter-plate (i.e. without the use of barcodes distinguishing between samples [30]) non-biological contamination can be introduced. However, in the context of a parasite (or an infected host) the screening for off-target data and contamination becomes even more important: Correct inference of biological origin for a given contig constitutes a prerequisite for most downstream analysis or the interpretation of results.

Cross-contamination from different compartments of a picolitre-plate was ruled out by our sequence provider, using Multiplex Indexes (MID) for one library and similarity searches to neighboring lanes for the other libraries.

For the remaining off-target and contamination problem we archived separation of sequences in two steps, one before assembly, one afterward. Both screening-steps had to rely solely on sequence comparison. The screening-step before assembly has to employ lower stringency as sequence comparisons on sequence as short as reads are less informative than on longer contig-sequence. In our case of *A. crassus*, neither of the two host species has genomic data available for use in similarity searches. A publicly available transcriptome-data-set for European eel [31] in addition to a unpublished data-set for the same species was augmented with a data-set generated from the Japanese eel sequenced for the purpose of screening *A. crassus*-sequences in the present project. The pre-assembly screening had the rationale of facilitating the assembly process reducing the amount of divergent sequence from two host-species and the amount of extensively covered rRNA sequence. In our sequencing we were not able to reproducibly alleviate the rRNA coverage. This has probably been due to the fact that extraction of total-RNA from worms filled with host blood resulted in low amounts of starting material, and amplification using standard kits did not allow to reproducibly alleviate rRNA abundance. As the same problems existed in preparation of liver tissue of the host species it seems likely that the blood of eels contains substances limiting the success of specific amplification protocols. In fact it is known that compounds like hemoglobin can inhibit PCR reactions [32] and reverse transcription [33].

Although raw reads with rRNA hits were screened out prior to assembly, it was still possible to gain insights from these off-target data, as we assembled and annotated screening databases. Some of the rRNA data especially from the L2 library showed high similarity to flagellate eukaryotes. It could be possibly derived from an unknown protist living in the swimbladder of eels (possibly as a commensal of *A. crassus*), from where the L2 larvae for RNA-preparation were washed out. This seems worth further investigation,

especially as it has been controversial whether encapsulated objects in the swimbladder of eels could be attributed solely to *A. crassus* [14] or to opportunist coinfections.

A second examination of sequence origin was obtained after assembly. It can employ higher stringency do to the longer sequence length in an assembled data-set and because we did not discard any off-target data, but marked TUGs for careful selection prior to special analysis or differential interpretation afterward. Such attempts to employ taxonomic screening were used before in a transcriptome projects for the garter snake [34] and a study on lake sturgeon even evaluated horizontal gene-transfer, when xenobiotic sequences were found [35] (with a negative result). A study describing a custom pipeline for transcriptome-assembly from pyrosequencing reads [36] suggested the use of `EST3` [37], to infer sequence-origin based on nucleotide frequency. However, we were not able to use this approach successfully, probably due to the fact that xenobiotic sequence in our data-set stems from multiple sources with different gc-content and codon-usage. Data from our L2 larvae library showed its anomaly later in gene-expression analysis, when off-target xenobiotic data was found to be responsible for the differences to other libraries. This makes this off-target data an interesting starting point for future investigation of the species community living in infected swimbladders of eels.

Compared to other NGS transcriptome sequencing projects, our assembly approach generated a smaller number of contigs. Projects using the `mira` assembler often report substantially more contigs for data-sets of similar size (see e.g. [38]), comparable to the `mira` sub-assembly in our approach. We demonstrated that our assembly approach generated a set less redundant and more complete sequences compared to the two "out of the box" approaches tested.

Protein prediction showed a trend towards completeness and elevated coverage on the 3' end of transcripts, as a result of RNA preparation using oligo-dT primers, an effect that seems to be ubiquitous in deep transcriptome sequencing projects (e.g. [39]). The low number of complete ORFs and especially the low number of start-codons (and corresponding Methionine amino acids) in predicted proteins seems cumbersome but cross-validation in combination with SNP-calling demonstrated the overall correctness of predicted proteins.

We were able to demonstrate, that screening of SNPs in or adjacent to homopolymer regions "improved" overall measurements on SNP-quality:

First the ratio of transitions to transversions (ti/tv) increased. Such an increase is explainable by the removal of "noise" associated with common homopolymer-errors [27]. Assuming that errors would be independent of transversion-transition bias erroneous SNPs would have a ti/tv of 0.5 and thereby lower the

overall value. Other explanations for these observations are hard to find so it can be concluded that removing noise from homopolymer sequencing-error ti/tv increases. The value of 2.38 (1.82 outside, 2.74 inside ORFs) is in good agreement with the overall ti/tv of humans (2.16 [40]) or *Drosophila* (2.07 [41]). The ratio of non-synonymous SNPs per non-synonymous site to synonymous SNPs per synonymous site (dn/ds) decreased with removal of SNPs adjacent to homopolymer regions from 0.45 to 0.32 after full screening. Similar to ti/tv it 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 [42], assuming negative (purifying) selection on most protein-coding genes lower values seem more plausible, also in comparison with other studies (see further text).

We used a threshold value for the minority allele of 7% for exclusion of SNPs, this corresponds to the ca. 10 "haploid equivalents" (5 individual worms plus an negligible amount of L2 larvae - in the L2 library and within the female adult worms - bearing possibly additional diversity). It is hard to explain, that ti/tv decreased in this filtering step, while dn/ds still further decreased.

The benefit of this screening was mainly a reduction of non-synonymous SNPs in high coverage contigs. When it was applied dn/ds did not scale with coverage. Working with an estimate of dn/ds independent of coverage, efforts to control for sampling a biased by sampling depth (i.e. coverage) like developed [43] and used [44] could be avoided.

When the whole of coding sequences are studied, of which only a small subset of sites can be under diversifying selection, dn/ds of 0.5 has been suggested as threshold for assuming diversifying selection [45] instead of the classical threshold of 1 [46]. In the transcripts from the female reproductive tract of *Drosophila* dn/ds was 0.15 [45] and in the 0.21 male reproductive tract [47] (although for ESTs specific to the male accessory gland were shown to have a higher dn/ds of 0.47). Pyrosequencing studies found dn/ds to be between 0.13 and 0.27 (depending on tissue type genes were mainly expressed in) in the Zebra finch transcriptome [48], 0.12 in the transcriptome of *Tigriopus californicus* [49] and 0.3 in the parasitic nematode *Ancylostoma canium* [50]. In comparison with these results even our estimate after screening seems high (although it should be noted, that the latter tree studies report a mean dn/ds over contigs - the *A. canium* doesn't make clear what exactly is reported - and therefore the value has to be compared to our mean dn/ds over contigs of 0.23) and further investigation using deeper sequencing of more individuals on the solexa GAII platform will be used to fully exclude the possibility of this result being induced by sequencing error. Moreover such an experiment should try to test that divergence between populations is leading to positive selection on only the possibly diverging European populations. For such a study the set

of SNPs found here are invaluable, as it can be used to define a gold standard set of SNPs found with both technologies.

We were able to obtain high-quality annotations for a large set of TUGs. Comparison with protein sequence derived from *B. malayi* showed a a remarkable degree of agreement regarding the occurrence of terms. This implies, that our transcriptome-data-set is a representative subset of a nematode-parasite genome.

Over-representation of GO-term in genes under diversifying selection (at a threshold of dn/ds>0.5, as established above) highlighted many interesting gene-products:

In the molecular function category two amino acid transmembrane transporters ("Contig5699" and "Contig866") - the only contigs with this annotation (or annotation, which is an offspring-term of this) and a dn/ds obtained - were found to have a dn/ds>0.5. Such transporter are thought to be important in the survival of parasites in a host [51].

Enrichment in the category "peptidase activity" highlighted twelve peptidases (from 43 with a dn/ds obtained). All twelve have orthologs in *B. malayi* and *C. elgans* and are conserved across kingdoms. Despite their conservation peptidases are thought to have have acquired new and prominent roles in host-parasite interaction compared to free living organisms: In *A. crassus* a trypsin-like proteinase has been identified thought to be utilized by the tissue-dwelling L3 stage to penetrate host tissue and an aspartyl proteinase thought to be a digestive enzyme in adults [52].

The under-representation of ribosomal proteins (term "structural constituent of ribosome") in disruptively selected contigs is in good agreement with the notion that ribosomal proteins are extremely conserved across kingdoms [53] and should be under under strong negative selection.

The additional prediction of signal sites for cleavage allowed interpretation and cross-validation of the results from SNP-calling: The detection of signal-peptides secretion using *in silico* analysis of ESTs has been used to highlight candidate genes for example in *Nippostrongylus brasiliensis* [54] and in a large scale analysis across all nematode [55] ESTs. Proteomic analysis in *B. malayi* [56, 57] and *Heligmosomoides polygyrus* [58] was able to find evidence for excretion for some of the protein-products and to highlight additional candidate genes.

We found an elevated dn/ds for signal-positives. These result could be explained follow the logic of signal-positives being more likely to be secreted to the host-parasite interface and proteins involved in host-parasite interaction being more likely to be under disruptive selection. Signal-positive TUGs with high dn/ds constitute another set of genes worth further examination in future studies.

TUGs predicted to be novel in the phylum nematoda contained the highest proportion of signal-positives.

A interpretation of this findings could be a confirmation of a study on *Nippostrongylus brasiliensis* [54], where signal positives were reported as less conserved. In the present study we did not aim to identify "novelty to *A. crassus*" as we believe in a deep sequencing project he absence of sequence similarity could be attributed to erroneous sequence instead of true novelty, and thereby blur analysis. However novelty in nematodes and to a lesser extend novelty in Spirurina seems to support the notion, that - if not diversified within nematoda to an extend leading to a complete loss of similarity, like suggested in the mentioned study - signal positives in nematodes could have taken a divergent evolutionary path from their orthologs in other phyla.

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

DESeq was able to report genes significantly differing in expression between male and female worms and between the L2 library and the all other worms. This was possible for male worms as well as for L2-larvae, were no replicated samples were obtained, due due the special features of this package [59]. However only over-expression in non-repeated samples can be detected, as obviously lack of expression in one sample can't validate

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

Conclusions

Methods

Worm samples, RNA extraction, cDNA synthesis and Sequencing

A. crassus from Japanese eels were sampled from Kao-Ping river and an adjacent aquaculture in Taiwan as described in [14]. Worms from the European eel 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, all worms were stored in RNA-later (Quiagen, Hilden, Germany) until extraction of RNA. RNA was extracted from:

- one worm from a cultured Taiwanese eel (sample T1)

- one worm from a wild Taiwanese eel (sample T2)
- one worm from an eel from the German sampling site (E1)
- one worm an eel from Polish sampling site (E2)

In addition RNA was extracted from L2-Larve from the German sampling site (sample L2) and a from a male worm from the Taiwanese aquaculture (sample M). 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 Japanese eel was prepared using the methods as described above for *A. crassus* samples.
A emulsion PCR was performed for each cDNA library according to the manufacturer's potocol (Roche/454 Life Sciences). For library E1 a Multiplex Index (MID) (Roche/454 Life Sciences) was used in preparation of the sequencing adapter. The libraries were sequenced in different runs of the Roche/454 Genome Sequencer FLX System: Library T2 on an eighth plate of the instrument usning standard "FLX-chemistry", the remaining libraries on a eighth of a plate using "FLX-Titanium-chemistry".

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 [60] (with parameters -minscore 20 and -minmatch 10). Seqclean [61] was used to screen poly-A-tails, low quality, repetitive and short (<100 bases) sequences. In addition all reads were `blast`ed (1e-5 -F F) against the following databases:

- a combined eel-mRNA database consisting of an assembly of sequences from the liver of the Japanese eel sequenced for this purpose (as described above), a sequence assembly of unpublished (sanger-) ESTs (made available to us by Gordon Cramb; University of St Andrews) and from EeelBase [31] a publically availble transcriptome database for the European eel.
- a eel-rRNA database from a rRNA screening of the above and assembly together with publically available rRNA-sequences.
- an *A. crassus* rRNA-database from screening of our dataset against nematode-rRNA, and assembly of these rRNA reads. This database notably also contained xenobiotic rRNA sequences.

Reads mapping to one of the databases with more than 80% of their length and 95% identity were removed from the dataset. Screenig and trimming information was written back into sff-format using `sfffile` (Roch/454).

We used an approach proposed by Kumar and Blaxter [21], combining assemblies from the `mira` [62] and `newbler` [20]. Briefly the two assemblies are combined into one using Cap3 [63] and only contigs supported by both assemblers are regarded good quaility. For further details see the supplementary methods.

Post assembly classification and taxonomic assesment

After assembly contigs were assesed a second time for host-contamination and other xenobiotics:

The contigs were `blast`ed (with a cut-off 1e-5) against the same databases used prior to assembly (Eel-mRNA, Eel-rRNA, *A. crassus*-rRNA and additionally against the nucleotide version of nempep4 [25,64], determining the best hit across databases. These best hits across databases were screened and only such hits involving more then 50% of the

Additionally `blast` (`blastn` e-value cut-off 1e-5) against NCBI-nt and (`blastx` e-value cut-off 1e-5) against NCBI-nt was used to determine taxon-membership of the top hit at the family, phylum and kingdom rank.

Protein prediction and annotation

Proteins were predicted using the `Prot4EST` (version 3.0b) [65]: First `blast` searches against a rRNA-database, a mitochondrial database and against uniref100 [66] were preformed. Then results were used to predict proteins directly (joining single high scoring pairs, and thereby intorducing gaps and ambiguous bases if needed). Secondly using the codon-usage from `blast`-predictions a simulated transcriptome was generated, reverse translating the *B. malayi* proteom, as training-data-set for `ESTscan`'s [67] hidden Markov models. If both `blast`-based prediction and `ESTscan` failed, simply the longest ORF is inferred.

`Blast`-based annotations were inferred using Annot8r (version 1.1.1) [68]: Searches were performed against all sequences in uniref100 [66] being annotated with GO-terms, EC-numbers and KEGG-parthways. Up to 10 (possibly contradictory) annotations based on a bitscore cut-off of 55 were obtained for each annotated database. For comparison annotations were obtained the same way for all *Brugia malayi* proteins in uniref100.

`SignalP V3.0` [28] was used to predict signal peptide cleavage sites and signal anchor signatures.

SNP analysis

As protein-prediction inferres gaps (e.g from sequencing errors) to predict the most likely protein, not only start- and end-coordinates of open reading frames (ORFs) had to be extracted from the output of `Prot4EST`. We did this in a custom `perl`-script using a `blast`-search with the nucleotide equivalent of the protein as query and the raw sequence as subject. We obtain the hit-coordinates as ORF-coordinates and imputed the `blast`-query as corrected ORF-sequencs.

We mapped the raw reads against the the complete unigene set, with the imputed sequences for those contigs with proteins predicted, using ssaha2 (with parameters -kmer 13 -skip 3 -seeds 6 -score 100 -cmatch 10 -ckmer 6 -output sam -best 1).

`pileup`-files were produced using `samtools` [69], discarding sequences mapping to multiple regions with the best hit. VarScan [70] (pileup2snp) was used with default parameters on `pileup`-files. This output was further screened as described in the results part of the manuscript.

Gene-expression analysis

For NlaIII-tag-sequencing total RNA was prepared as described above from a worm from the Polish sampling site. A sequence-tag libray was created following the protocol supplied by Illmina for this method. Briefly after synthesis of cDNA on oligo(dt)-beads, this cDNA is digested with the enzyme NlaIII (restriction site "CATG"). After ligation of an adaptor containing its restriction site the enzyme MmeI cuts 17 bases downstream of its binding site generating a sequence tag of in total 21 bases.

For 454 reads, read counts were obtained from the mapping to imputed sequence described above.

Tag-sequences were mapped using `BWA` [71]. And read counts extracted using `Samtools`.

The R-package DESeq [59] was used to normalize for library-size and analyse statistical significance of differential expression.

General coding methods

The bulk of analysis (unless otherwise cited) presented in this paper was carried out in R [72] using custom scripts. We used a method provided in the R-packages Sweave [73] and Weaver [74] for "reproducible research" combining R and TEXcode in a single file. All intermediate data files needed to compile the present manuscript from data-sources are provided upon request. For visualistaion we used the R-packages lattice [75] and ggplot2 [76].

Competing interests

The authors declare no competing interests.

Authors contributions

Acknowledgments

We thank Stephen Bridgett from gene-pool sequencing service for general help with raw data and for cross-contamination screening of libraries.

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

References

1. Kuwahara A, Niimi H, Itagaki H: **Studies on a nematode parasitic in the air bladder of the eel I. Descriptions of Anguillicola crassa sp. n. (Philometridea, Anguillicolidae)**. *Japanese Journal for Parasitology* 1974, **23**(5):275–279.
2. Neumann W: **Schwimblasenparasit Anguillicola bei Aalen**. *Fischer und Teichwirt* 1985, :322.
3. Koops H, Hartmann F: **Anguillicola-infestations in Germany and in German eel imports**. *Journal of Applied Ichthyology* 1989, **5**:41–45.
4. Koie M: **Swimbladder nematodes (Anguillicola spp.) and gill monogeneans (Pseudodactylogyrus spp.) parasitic on the European eel (Anguilla anguilla)**. *ICES J. Mar. Sci.* 1991, **47**(3):391–398, [http://icesjms.oxfordjournals.org/cgi/content/abstract/47/3/391].
5. Kirk RS: **The impact of Anguillicola crassus on European eels**. *Fisheries Management & Ecology* 2003, **10**(6):385–394, [http://dx.doi.org/10.1111/j.1365-2400.2003.00355.x].
6. Kristmundsson A, Helgason S: **Parasite communities of ecls Anguilla anguilla in freshwater and marine habitats in Iceland in comparison with other parasite communities of eels in Europe**. *Folia Parasitologica* 2007, **54**(2):141.
7. Würtz J, Knopf K, Taraschewski H: **Distribution and prevalence of Anguillicola crassus (Nematoda) in eels Anguilla anguilla of the rivers Rhine and Naab, Germany**. *Diseases of Aquatic Organisms* 1998, **32**(2):137–43, [http://www.ncbi.nlm.nih.gov/pubmed/9676253].
8. Lefebvre FS, Crivelli AJ: **Anguillicolosis: dynamics of the infection over two decades**. *Diseases of Aquatic Organisms* 2004, **62**(3):227–32, [http://www.ncbi.nlm.nih.gov/pubmed/15672878].
9. Taraschewski H: **Hosts and Parasites as Aliens**. *Journal of Helminthology* 2007, **80**(02):99–128, [http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=713884].
10. Münderle M, Taraschewski H, Klar B, Chang CW, Shiao JC, Shen KN, He JT, Lin SH, Tzeng WN: **Occurrence of Anguillicola crassus (Nematoda: Dracunculoidea) in Japanese eels Anguilla japonica from a river and an aquaculture unit in SW Taiwan**. *Diseases of Aquatic Organisms* 2006, **71**(2):101–8, [http://www.ncbi.nlm.nih.gov/pubmed/16956057].
11. Knopf K: **The swimbladder nematode Anguillicola crassus in the European eel Anguilla anguilla and the Japanese eel Anguilla japonica: differences in susceptibility and immunity between a recently colonized host and the original host**. *Journal of Helminthology* 2006, **80**(2):129–36, [http://www.ncbi.nlm.nih.gov/pubmed/16768856].
12. Knopf K, Mahnke M: **Differences in susceptibility of the European eel (Anguilla anguilla) and the Japanese eel (Anguilla japonica) to the swim-bladder nematode Anguillicola crassus**. *Parasitology* 2004, **129**(Pt 4):491–6, [http://www.ncbi.nlm.nih.gov/pubmed/15521638].
13. Knopf K, Lucius R: **Vaccination of eels (Anguilla japonica and Anguilla anguilla) against Anguillicola crassus with irradiated L3**. *Parasitology* 2008, **135**(5):633–40, [http://www.ncbi.nlm.nih.gov/pubmed/18302804].

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

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

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

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

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

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

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

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

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

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

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

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

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

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

28. Emanuelsson O, Brunak S, von Heijne G, Nielsen H: **Locating proteins in the cell using TargetP, SignalP and related tools**. *Nat Protoc* 2007, **2**:953–971.

29. Nielsen H, Krogh A: **Prediction of signal peptides and signal anchors by a hidden Markov model**. *Proc Int Conf Intell Syst Mol Biol* 1998, **6**:122–130.

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

31. 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.

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

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

34. 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.

35. 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.

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

37. 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].

38. 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.

39. 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.

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

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

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

43. 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.

44. 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.

45. 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.

46. 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.

47. 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.

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

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

50. 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.

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

52. 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].

53. 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.

54. 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].

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

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

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

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

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

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

61. 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].

62. 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].

63. 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].

64. 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].

65. 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].

66. 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.

67. Iseli C, Jongeneel C, Bucher P: **ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences.** *Proc Int Conf Intell Syst Mol Biol* 1999, :138–148, [http://www.ncbi.nlm.nih.gov/pubmed/10786296].

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

69. 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.

70. 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.

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

72. 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].

73. 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].

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

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

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

Figures

Tables

Table 1 - Trimming statistics

	raw_reads	short	lowq	dust	shortq	valid
T2	116366	46	24361	50	3226	88683
T1	99482	44	6589	13	1301	91535
L2	112718	55	10571	147	4880	97065
M	106726	44	10761	225	4454	91242
E1	209325	49	49798	15	42882	116581
E2	111746	163	8478	20	2242	100843
total	756363	401	110558	470	58985	585949

Table 2 - Pre-assembly screening statistics

	AcrRNA	eelmRNA	eelrRNA	valid
E1	76403	4835	13112	22231
E2	11213	3613	69	85948
L2	35940	1220	1603	58302
M	31351	1187	418	58286
T1	24929	7475	514	58617
T2	7233	11741	38	69671

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

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

Table 4 - Post-assembly host-screening

	AcrRNA	eelmRNA	eelrRNA	valid_nempep	valid_no_hit
number	604	1162	50	1254	37117
mean coverage	3.28	1.80	2.00	6.67	2.44

Table 5 - Evolutionary conservation

	conserved	novel.in.metazoa	novel.in.nematoda	novel.in.clade3
bit.threshold.50	7741	1720	1769	1523
bit.threshold.80	4715	1402	1686	1695

Table 6 - Protein prediction statistics

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

Table 7 - SNP summry statistics

			pos	in	codon		ti/tv		
	No.SNPs	in.ORF	1	2	3	overall	ins.orf	outs.orf	dn.ds
raw	13042	8309	2713	2172	3424	1.98	2.43	1.42	0.45
h.screened	9523	6395	2058	1604	2733	2.86	3.32	2.16	0.39
p.screened	6276	4226	1368	925	1933	2.38	2.74	1.82	0.32

Table 8 - GO-terms in positively selected

Count	Size	Term	direction
12	43	peptidase activity	Over
2	2	L-amino acid transmembrane transporter activity	Over
44	269	catalytic activity	Over
3	6	ribonucleoprotein binding	Over
7	26	ATPase activity	Over
21	113	hydrolase activity	Over
2	3	carboxylic acid transmembrane transporter activity	Over
2	3	testosterone dehydrogenase activity	Over
2	3	organic acid transmembrane transporter activity	Over
2	3	oxidoreduction-driven active transmembrane transporter activity	Over
2	3	ribonuclease activity	Over
2	3	amino acid transmembrane transporter activity	Over
2	3	testosterone dehydrogenase (NAD+) activity	Over
2	3	amine transmembrane transporter activity	Over
1	44	structural constituent of ribosome	Under
3	3	branched chain family amino acid metabolic process	Over
3	3	branched chain family amino acid catabolic process	Over
10	32	brain development	Over
7	19	positive regulation of cell cycle process	Over
4	7	spermatid differentiation	Over
4	7	spermatid development	Over
6	15	response to starvation	Over
3	4	positive regulation of mitosis	Over
3	4	positive regulation of nuclear division	Over
11	40	central nervous system development	Over
13	52	regulation of cell cycle	Over
11	42	cellular amino acid metabolic process	Over
5	12	autophagy	Over
10	37	regulation of cell cycle process	Over
8	27	interphase	Over
8	27	interphase of mitotic cell cycle	Over
2	2	pentose metabolic process	Over
2	2	xylulose metabolic process	Over
2	2	response to disaccharide stimulus	Over

2	2	embryonic body morphogenesis	Over
2	2	response to sucrose stimulus	Over
2	2	L-amino acid transport	Over
2	2	NADP metabolic process	Over
17	81	apoptosis	Over
16	76	regulation of molecular function	Over
27	151	catabolic process	Over
12	52	cellular amine metabolic process	Over
33	195	response to stress	Over
5	14	mitotic cell cycle G1/S transition DNA damage checkpoint	Over
5	14	regulation of cellular amine metabolic process	Over
10	41	reproductive structure development	Over
12	53	amine metabolic process	Over
8	30	muscle organ development	Over
7	25	regulation of catabolic process	Over
3	6	centrosome organization	Over
3	6	RNA catabolic process	Over
3	6	microtubule organizing center organization	Over
3	6	positive regulation of organelle organization	Over
5	15	signal transduction in response to DNA damage	Over
5	15	mitotic cell cycle G1/S transition checkpoint	Over
5	15	G1/S transition checkpoint	Over
5	15	DNA damage response, signal transduction by p53 class mediator	Over
5	15	G1/S transition of mitotic cell cycle	Over
5	15	regulation of G1/S transition of mitotic cell cycle	Over
16	80	cell cycle phase	Over
11	49	mRNA metabolic process	Over
18	94	nervous system development	Over
4	11	skeletal muscle organ development	Over
4	11	imaginal disc development	Over
14	69	regulation of apoptosis	Over
14	69	regulation of programmed cell death	Over
2	3	positive regulation of mitotic metaphase/anaphase transition	Over
2	3	negative regulation of reproductive process	Over
2	3	germ cell migration	Over
2	3	centrosome duplication	Over
2	3	centrosome separation	Over
2	3	protein tetramerization	Over
2	3	protein homotetramerization	Over
2	3	mitotic centrosome separation	Over
2	3	regulation of the force of heart contraction	Over
2	3	spliceosomal conformational changes to generate catalytic conformation	Over
2	3	nuclear mRNA cis splicing, via spliceosome	Over
2	3	amino acid transport	Over
19	103	cell cycle	Over
8	33	ATP synthesis coupled electron transport	Over
8	33	mitochondrial ATP synthesis coupled electron transport	Over
8	33	regulation of mitotic cell cycle	Over
17	90	programmed cell death	Over
3	7	regulation of neurotransmitter levels	Over
3	7	cellular response to starvation	Over

18	97	cell death	Over
18	97	death	Over
13	64	macromolecule catabolic process	Over
6	22	response to protein stimulus	Over
15	188	gene expression	Under
1	46	cellular protein complex disassembly	Under
1	46	macromolecular complex disassembly	Under
1	46	cellular macromolecular complex disassembly	Under
1	46	protein complex disassembly	Under
1	45	viral genome expression	Under
1	45	viral transcription	Under
1	45	pancreas development	Under
1	45	endocrine pancreas development	Under
1	45	endocrine system development	Under
4	79	translation	Under
1	44	translational termination	Under
8	118	transcription	Under
2	55	cellular component disassembly at cellular level	Under
2	55	cellular component disassembly	Under
21	225	biosynthetic process	Under
0	25	positive regulation of intracellular protein kinase cascade	Under
14	165	cellular macromolecule biosynthetic process	Under
0	24	oocyte differentiation	Under
21	222	cellular biosynthetic process	Under
1	37	oogenesis	Under
0	23	oocyte development	Under
0	23	cation transport	Under
2	47	viral infectious cycle	Under
2	47	viral reproductive process	Under
15	168	macromolecule biosynthetic process	Under
1	35	positive regulation of response to stimulus	Under
0	22	positive regulation of MAPKKK cascade	Under
4	7	small nuclear ribonucleoprotein complex	Over
2	2	Cajal body	Over
2	2	U4/U6 x U5 tri-snRNP complex	Over
2	2	U5 snRNP	Over
3	6	nuclear speck	Over
27	152	mitochondrion	Over
5	15	nuclear body	Over
2	3	clathrin sculpted vesicle	Over
2	3	basement membrane	Over
2	3	plant-type cell wall	Over
2	3	plasma membrane respiratory chain complex I	Over
2	3	plasma membrane respiratory chain	Over
0	29	large ribosomal subunit	Under
0	27	cytosolic large ribosomal subunit	Under
27	266	nucleus	Under
17	185	non-membrane-bounded organelle	Under
17	185	intracellular non-membrane-bounded organelle	Under
4	65	nucleolus	Under

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

Table 9 a - analysing all TUGs

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

Table 9 b - analysing good-category contigs only

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

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

	solexa.tags	E1	E2	L2	M	T1	T2	all.reads
solexa.tags	1.000	0.373	0.524	-0.123	0.438	0.411	0.168	0.430
E1	0.373	1.000	0.311	0.000	0.346	0.313	0.160	0.341
E2	0.524	0.311	1.000	-0.186	0.271	0.421	0.148	0.434
L2	-0.123	0.000	-0.186	1.000	-0.106	-0.140	-0.258	0.001
M	0.438	0.346	0.271	-0.106	1.000	0.249	0.030	0.384
T1	0.411	0.313	0.421	-0.140	0.249	1.000	0.151	0.397
T2	0.168	0.160	0.148	-0.258	0.030	0.151	1.000	0.505
all.reads	0.430	0.341	0.434	0.001	0.384	0.397	0.505	1.000

Additional Files

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