1 Methods

1.1 Experimental infection of eels

An. anguilla were obtained from the Albe-Fishfarm in Haren-Rütenbrock, Germany. *An. japonica* were caught at the glass eel stage in the estuary of Kaoping River, Taiwan by professional fishermen and kept at a water temperature of 26°C until they reached a size of > 35 cm.

The absence of infections with *A. crassus* in both eel species was confirmed by dissection of 10 individuals of each species.

After an acclimatisation period of 4 weeks (*An. anguilla*) or when they reached a size of > 35cm (*An. japonica*) eels were infected using a stomach tube as described in [**?**]. During the infection period water temperature was held constant at 20°C. Eels were kept in 160-litre tanks in groups of 5-10 individuals and continuously provided with fresh, oxygenated water and once every two days with commercial fish pellets (Dan-Ex 2848, Dana Feed A/S Ltd, Horsens, Denmark) *at libitum*.

L2 larvae used for the infection were collected from the swimbladders of wild yellow and silver eels from the River Rhine near Karlsruhe and from Lake Müggelsee near Berlin in Germany. Taiwanese larvae were obtained from eels from an aquaculture adjacent to Kao Ping River in south Taiwan and from a second aquaculture in Yunlin county, approximately 150 km further north on the west coast of Taiwan. They were stored at 4°C for no longer than 2 weeks before copepods were infected. Mixed species samples of uninfected copepods were collected from a small pond near Karlsruhe, known to be free of eels (and *Anguillicola*). They were infected individually in wells of microtiter plates at an intensity of roughly 10 L2-larvae per copepod. One week after infection they were placed in bigger tanks. Twice a week yeast was provided as food and at 21 dpi infective L3 were harvested with using a tissue potter as described by [**?**]. 50 L3 for infection of individual eel were suspended in 100 μl RPMI-1640 medium (Quiagen, Hilden, Germany) and eels were infected as described above.

55-57 days post infection (dpi) eels were euthanized and dissected. The swimbladder was opened and after determination of the sex of adult worms under a binocular microscope (Semi 2000, Zeiss, Germany), adult *A. crassus* were immediately immersed in RNAlater (Quiagen, Hilden, Germany).

1.2 RNA extraction and preparation of sequencing libraries

RNA was extracted from 12 individual female worms and for 12 pools of male worms using the RNeasy kit (Quiagen, Hilden, Germany) (see table XXX).

3 individual female worms from each experimental group were chosen randomly to give in total twelve females. Additionally, from three individual male worms, and from 9 pools of male worms RNA was extracted (see tabled XXX). Pools consisted of worms from one infected eel individual each. All replicates were derived from infections of different eel individuals, with one exception from this form of statistical independence: from *An. japonica* European male worms

as well as a female worm had to be prepared from the same eel individuals. It was impossible to extract enough RNA from all but the biggest male worms especially of the Japanese eel/European worm combination, leaving no other choice. Because of the small size of male worms it was generally not possible to randomly choose individuals. Preparation of sufficient amounts of RNA was only achieved in pools of the biggest individuals. All male worms were thus chosen for preparation based on large size, even when pools of worms were used.

The paired-end TruSeqTM RNA sample preparation kit (Illumina) was followed to build sequencing libraries with insert sizes of roughly 270 bp for paired-end sequencing from cDNA libraries: briefly, poly-T oligo attached magnetic beads were used for purification of mRNA and to simultaneously fragment the RNA. The RNA was then primed with random hexamer primers for cDNA synthesis and reverse transcribed into first strand cDNA using reverse transcriptase. The cDNA was cleaned from the second strand reaction, overhangs were repaired to form blunt ends, a single "A"-nucleotide was added at the 3' end and paired end sequencing adapters ware ligated with a complementary "T"-overhang.

In this step multiple differently indexed paired-end adapters were used to enable multiplexing of the 24 different sequencing libraries in 3 pools of 8 samples each. These three pools all contained one random replicate each for each treatment combination ensuring complete statistical independence of replicates from sequencing-lane effects.

Molecules having adapter sequences were enriched in the mix using PCR and the libraries were controled for quality and quantity on the BioAnalyzer (Agilent). Clusters were generated by bridge amplification. The resulting clusters were sequenced on the Genome Analyzer IIX in combination with the paired-end module. The first read was sequenced using using the first primer Rd1 SP. The original template strand was then used to regenerate the complementary strand, the original strand was removed and complementary strand acted as a template for the second read, sequenced primed by the second sequencing primer Rd2 SP.

1.3 Quantitative PCR validation

!!!

1.4 Mapping and normalisation of read counts

All sequencing reads were mapped to the fullest 454 assembly described in !!!454paper!! and we excluded TUGs inferred as host or cobiont origin as filter, using BWA [**?**] (version 0.5.9-r16; BWA aln and BWA sampe with default options) and processed with samtools [**?**] (version 0.1.18; samtools view -uS -q 1) to only allow uniquely mapping reads. All reads mapping to host and cobiont off-target data were removed during downstream evaluation.

Counts were summed for technical replicates and counts to lowCA derived contigs (see !!!454paper!!!) were disregarded as well as spurious read counts to

contigs with less than 32 mapping reads in total (see however 1.6 for how these counts were used in further tests of reference fragmentation).

The remaining counts were normalised using DESeq (version 1.6.1) (i.e. the normalisation factor was estimated by the median of scaled counts, similar to the weighted trimmed mean of the log expression ratios used later in edgeR). All tables summarising read-counts are based on these normalised counts. We obtained "variance stabilised data" in an expression matrix for each gene and library using the "blind" option in a calculation not informed (and biased by) the model-design. These data were used in all gene-centring heatmap and multivariate visualisations. Additionally this matrix was transposed to get sample-to-sample distances.

1.5 Statistical analysis with generalised linear models (GLMs)

The R-package edgeR (version 2.4.1) [**?**] was used to build negative binomial generalised linear models, as these specialised GLMs outperformed GLMs in DESeq in speed and reliability of convergence. Modeled were based on a negative binomial distribution and the dispersion parameter for each transcript was approximated with a trend depending on the overall level of expression. In the maximal fitted model expression was regressed on worm-sex, host-species and parasite population, including all their interactions. The full model thus contained terms $S_i + H_j + P_k + (SH)_{ij} + (SP)_{ik} + (HP)_{jk} + (SHP)_{ijk} + \varepsilon$, where ε is the residual variance, S_i is the effect of the ith sex (male or female), H_j is the effect of the ith host species (*An. anguilla* or *An. japonica*), P_k is the effect of the kth population (European or Asian), $(SH)_{ij}$ is the sex-by-species interaction and similarly for the other interactions.

The hierarchical nature of generalised linear models was respected considering (removing) all interaction effects of a main-term (e.g. $(SP)_{ik}$, $(SH)_{ij}$ and $(SHP)_{ijk}$) when analysing models for the significance of that term (e.g. S_i). Resulting p-values were corrected for multiple testing using the method of Benjamini and Hochberg [**?**] and differential expression was inferred at a false discovery rate (FDR) of 5% (adjusted p-value of 0.05).

1.6 Count-collapsing for orthologs from two model-species

In order to test the influence of deficiencies (i.e. fragmentation) of the assembly on mapping we summed read counts over orthologous sequence in *C. elegans* and *B. malayi*. Differential expression for these orthologous-counts was analysed the same way as for contigs. Contigs were filtered based on inference from orthologous counts merging the two orthologous evaluations and the contig evaluation. Differential expression was accepted at a FDR of 5% for the contig evaluation and 10% for both of the two orthologous evaluations.

1.7 Multivariate confirmation of linear models

I used the R package vegan (version 2.0-2) to perform constrained redundancy analysis on contigs identified as significant in GLMs before. For each set of contigs (different for sex, eel-host or -worm-population) the appropriate constrained component was used. The proportion of the variance explained by the constrained component was recorded and the constrained component was tested for significance using a permutation test implemented in vegan.

1.8 Gene ontology enrichment analysis

Prior to analysis of GO term overrepresentation (based on dn/ds or expression values) we used the R-package annotationDbi [**?**] to obtain a full list of associations (also with higher-level terms) from annot8r-annotations. We then used the R-package topGO [**?**] to traverse the annotation-graph and analyse each node in the annotation for overrepresentation of the associated term 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.

1.9 Clustering analysis

The R package HeatmapPlus was used on variance stabilised expression values to visualise hierarchical clusters similar to the method of [**?**]. The results were displayed along with annotations stored in a Bioconductor eSet-class object.

2 General coding methods

The bulk of analysis (unless otherwise cited) presented in this paper was carried out in R [**?**] using custom scripts. We used a method provided in the R packages Sweave [**?**] and Weaver [**?**] for "reproducible research" combining R and LaTeXcode in a single file. All intermediate data files needed to compile the present paper are provided at For general visualisation we used the R packages ggplot2 [**?**] and VennDiagram [**?**].

3 Results

Dissection of eels 55-57 after infection (dpi) showed higher recovery of the present day sympatric European worms in *An. anguilla* and higher recovery of Taiwanese worms in *An. japonica*, compared to the allopatric host parasite combinations.

!!!Figure!!! Recovery of worms in coinoculation experiment. Mean numbers of worms recovered after 55-57 dpi for sample sizes given as n=x. Error-bars indicate the standard error (s.e.) of the mean. Recovered lifecycle stages of

the parasite are listed separately as L3-larvae (l3), L4-larvae (l4), adult females (adult.f) and adult males (adult.m).

In the sympatric host parasite pairs roughly eight or nine adult worms and the same number of larval stged could be recovered per eel, resultin in roughly 30% recovery as a proportion of the 50 administrated larvae. In the transplanted host parasite combinations only two or three adult worms were recovered on average and also the number of larval stages was not higher (recovery below 10

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.5000	1.1109	8.55	0.0000
host.spec.AJ	-8.0789	1.7472	-4.62	0.0000
worm.pop.T	-5.2222	1.3689	-3.81	0.0002
host.spec.AJ:worm.pop.T	11.7345	2.2010	5.33	0.0000

Table 1: Linear model for recovery of adult worms. The estimate gives the mean of the distribution of adult worms for the factor values in the rows. The intercept is set to "Aa. R" (*An. anguilla* and the European populations) further rows give variations for each factor. Std. Error is the standard error of this value. Additionally the probability of a t-value as small or smaller than the observed t-value are given. The signature of local adaptation is visible in the highly significant interaction term.

4 Sample preparation and sequencing

From three biological replicates data was obtained for each of 6 experimental groups: The two worm populations in each of the two eel species and for each of the two sexes of worms. This resulted in a total of 24 independent sequencing experiments.

!!!Table!!! Summary of RNA preparation. A summary of 24 samples sequenced. The label of the RNA preparation follows a convention based on the eel species (host; first two letter of label, AA for *An. anguilla* AJ for *An. japonica*), worm population (population - R for European, T for Taiwanese) and sex of worm(s) in preparation (F for female, M for male; last letter in label). The European samples were from two locations: river Rhine (R,) and Müggelsee near Berlin (B), the Taiwanese samples were from from Kao Ping River (K) and Yunlin county (Y). Additionally the intensity of infection (number of adult worms found in the infected eel; intensity) and the number of worms pooled in the preparation (only male worms are pooled for RNA extraction, individual female worms were used). Finally RNA-concentration in the preparation (conc in prep) is given in μg per ml.

Sequencing was performed in three multiplexed pools of eight libraries each. Each pool of eight was sequenced on two lanes, giving in total six lanes of data and two technical replicates for each library. Sequencing resulted in a total of

263,668,952 raw sequencing read-pairs, each read having a length of 51 bases and 270 bases mean insert size between the read pairs.

5 Examination of data-quality

Reads were mapped against the fullest pyrosequencing-assembly (!!! 454 paper !!!) using BWA [?]. Of the 263,668,952 raw read-pairs 173,602,387 mapped uniquely to the assembly and were counted on a per-library base.

The technical replicates demonstrated very low differences as inferred from a clustering analysis using variance stabilised data and transposed euclidean distances between samples (see figure).

!!!Figure!!! Distances between RNA-seq read-count for different samples. Euclidean distance (square distance between the two count vectors) for variance stabilised read-counts for all libraries including technical replicates; Red indicates low distance (high similarity), blue high distance (low similarity). a) Data before screening and summation of technical replicates. All technical replicates are clustered very closely, the distance between an outlier female sample (AJ_T26F) is high. b) Same illustration after summation of technical replicates and screening. Distance between outlier-sample and other female samples is reduced.

158,232,523 read-pairs were left after removal of hits to contigs for which non-*A. crassus* origin had been inferred in the analysis of the 454-transcriptome assembly.

!!!Table!!! Mapping Summary. Mapping is summarised for all 24 libraries. Rows indicate different libraries (worms or worm-pools as indicated in !!!! raw.reads gives the number of read-pairs sequenced, raw.mapped the number of reads mapping uniquely with their best hit, tax.mapped the number of reads after subtraction of reads to putative eel-host derived contigs and screened after subtraction of all reads mapping not to the highCA-derived assembly or to contigs with overall counts less than 32.

After another screening for spurious read-counts to low coverage transcripts and to transcripts of low reliability (lowCA in the 454-assembly; !!!454paper!!!) 137,477,156 read-pairs were left for further analysis. Distribution of these read-pairs over libraries showed roughly 2.7-fold differences, with a mean of 5,728,215 reads and a range from 3,422,526 read-pairs for library AJ_R3M to 9,453,468 read-pairs for library AA_R8F (see table XXX).

!!!Figure!!!Principle coordinate plot for expression in RNA-seq libraries. Distance between sample-pairs is the root-mean-square deviation (Euclidean distance) for the most differentially expressed (DE) genes. Distances can be interpreted as the log2-fold-change of the genes with the biggest changes, i.e. the log2-fold-change for the genes that distinguish the samples.

These reads mapped to 7,520 contigs from our 454 assembly, making them the basis for all further investigations.

In addition to hierarchical cluster analysis, also principal component analysis grouped libraries according to the sex of worms (the largest effect), but was

unable to identify libraries with expression correlated due to other experimental factors (eel host or parasite population; see figure b). Between-sample distance confirmed the hierarchical library clustering. Sex of the worms defined the overall distances between libraries, host- or population-differences were not visible in an overall effect in the top differentially expressed (DE) genes (see figure). Amoung male samples distance was smaller than among female samples.

6 Orthologous screening for expression differences

For the 7,520 contigs with expression values 4,382 *C. elegans*-orthologs and 4,292 *B. malayi*-orthologs were determined based on the annotation of our pyrosequencing-assembly (see cite !!!454-paper!!!). This resulted in 3,596 contigs with an expression measurement, having a measurement also for both corresponding orthologs (or group of orthologs) in both model-species and thus being available for analysis.

For all further evaluations the congruence of the basic contig-based statistics with orthologous-confirmed (OC) statistics is considered.

7 Expression differences in generalised linear models

Generalised linear models (GLMs) were used as implemented in the R-package edgeR. Using these models we obtained 2,588 contigs (34% of total) DE between male and female worms at a false discovery rate (FDR) of 5%. 1,101 (31% of total orthologous available) of these contigs of were confirmed by contigs in the orthologous evaluation. 1,425 (556 OC) of these were upregulated in male worms 1,163 (545 OC) in female worms.

At the same threshold, 55 contigs (0.7% of total; 9, 0.25% OC) showed significant differential response to the host-species. 38 (5 OC) were upregulated in *An. japonica*, 17 (4 OC) in *An. anguilla*.

68 contigs (0.9% of total; 15, 0.42% OC) showed differences according to the population of the worm. 39 (11 OC) of these were upregulated in the Taiwanese population, 29 (4 OC) in the European populations.

An important observation in these models is the prevalence of co-occurring significance of simple main effects. Expression changes overlapping for two main effects mean a significant difference in expression according to both factors. These differences are in the same direction for a combination of the factors. Most contigs DE according to the main effects of host-species or worm-population were also DE according to the sex of the worm. There was also a number of contigs differing for all three predictors in the same way. No contigs were observed DE in both the host-species and worm-population in the same direction but not according to worm-sex. From the 68 contigs DE in different *A. crassus*-

populations, 38 were also DE according to worm sex and 16 according to all three main effects (see figure).

In addition, interaction-effects were also observed. In these interactions a difference according to both focal factors in different directions for factor combinations is indicated. For interactions between host-species and parasite-population (eel/pop), for example, this mirrors the result of adult recovery i.e. a differential regulation according to sympatric host-species/parasite-population combinations as found in nature: 7 contigs (0 OC) showed differential expression according to the worm-sex/eel-species interaction, 12 (3 OC) to worm-sex/parasite-population, 13 (2 OC) to host-species/parasite-population, 1 (0 OC) contig showed significance for the 3-way interaction (see figure). It should be noted, that conclusions drawn from of simple main effects do not necessarily hold for contigs with significant interaction effects (e.g. significantly higher expression in European population can then mean higher values only in one of the host-species).

!!!Figure!!! Venn diagram of contigs significant for different terms in edgeR GLMs. Overlap between differences in simple main effects are given as black numbers in the Venn-Diagram. Numbers outside the circles in the lower left corner indicated non-significant contigs. The number of significant contigs for interaction effects are indicated in red for comparison. In (a) values for all contigs are given in (b) for ortholog-confirmed (OC) contigs.

In summary, a low amount of overlap in main effects between populations and host-species compared to the other main-effect overlaps and in relation a higher proportion of interaction effects between these two conditions was observed.

8 Confirmation of contig categories through principal component analysis

I performed constrained redundancy analysis for the effects of eel-host and worm-population. This technique, similarly to principal components analysis, can partition the variance into orthogonal components, and additionally constrain one of the components to the factor of interest. I found that 7% of the variance in contigs DE between eel-hosts and 11% of the variance in contigs DE between worm-population explained by the corresponding factor. In both evaluations more than 50% of the remaining variance could be explained by a single principal component, to which sex contributed over 99% (loading) (see figure and). When only OC-DE contigs were considered the explained variance for difference between eel-host dropped to 3.3% and the explained variance for differences between worm-population was raised to 23%, while the sex-effect explained 70% and 50% of the variance (see figure and). Significance of the constrained component evaluated by a permutation-test could be established at a $p < 0.05$ threshold for all but the OC eel-host DE subset.

!!!!Figure!!! Constrained redundancy analysis for host-DE contigs. Eel-host differences are displayed as constrained component on the x-axis, the sex con-

tributed ¿99% (loading) to the principal component on the y-axis. (a) Host differences partition the variance in samples in like expected for all contigs, the constrained component showed significance. (b) For OC contigs the constrained component fails to to partition the variance as expected, the component showed no significance for this subset of the data.

!!!Figure!!! Constrained redundancy analysis for population-DE contigs. Population differences are displayed as constrained component on the x-axis, the principal component on the y-axis corresponds to the sex of the worm. Host differences partition the variance in samples like expected for all contigs (a) as well as for OC-contigs (b). The constrained component showed significance in both subsets.

9 Biological processes associated with DE contigs

I employed tests for overrepresentation of categories in gene-ontology (GO). These tests respect the structure of the ontology and also consider overrepresentation of higher level (ancestor-) terms. Summarising annotations at higher levels it is therefore possible to conceive higher-order responses to the conditions investigated.

For the differences between male and female worms enriched annotations can be summarised into three broad categories: Terms overrepresented due to spermatogenesis (e.g. PP1-phosphatase and ester hydrolase are important for spermatogenesis in *C. elegans* [**?**, **?**]) embryo development (many obvious terms) and terms for other processes more related to metabolic differences between males and females (such as oxidoreductase activity; see table XXX but also additional figures and).

For the lower number of contigs DE between host-species inference of higher order terms was obviously only possible to a limited extent and in part also unnecessary, because annotations can be interpreted at face value. However, annotations for contigs DE between eel-hosts highlighted redundant terms associated with "antigen processing and presentation" proteins which are in mammals usually involved in antigen processing and cleavage of the invariant chain of the MHCII complex. These terms led to Contig566 and Contig26 and their *B. malayi*-orthologs "aspartic protease BmAsp-1, identical" and "eukaryotic aspartyl protease family protein". In blood feeding helminths these enzymes are in contrast usually involved in early cleavage events during the digestion of host haemoglobin [**?**].

!!!Figure!!!GO biological process graph for enriched terms in DE according to worm-population. Subgraph of the GO-ontology biological process category induced by the top 10 terms identified as enriched in DE genes between different parasite populations. 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 (eventu-

ally 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.

For contigs DE between worm populations despite the limited number of DE contigs, enrichment analysis identified "oxidoreductase activity" as an informative significantly enriched higher level term (see figure). The biological processes "response to metal ion" and "mitochondiral electron transport" (see figure) confirmed an evaluation linking these mainly to enzymes used in respiratory processes and highlighted additionally enzymes from lipid metabolism (especially β-oxidation of fatty acids) related to respriration and the availability of oxygen.

10 Clustering analysis

For the remainder of the text I will concentrate on these differences of the European and Taiwanese populations and mention the other differences only as far as they are related to this focal factor. In figure however, graphical analyses of the same type are presented for other factors.

Clustering analysis uses distance measurements between samples as well as genes (or transcripts) to highlight patterns of similarity. The classical distance measure used in hierarchical clustering throughout this document is Euclidean distance. Grouping of genes regulated in parallel in combination with annotation, the status of cellular processes can support notions based on single genes.

Hierarchical clustering analyses of genes DE between populations confirmed the results of principal component based multivariate analysis. The main factor grouping libraries was the sex of the worm. A sub-grouping of samples fully according to European and Taiwanese populations was only observed for male worms. In female worms other unmeasured co-factors were preventing a clustering fully according to this factor. In male worm however, library clustering even followed a pattern of similar expression in according to the second factor of eel-host. These statements are true for both the full set of contigs (see figure) and OC contigs (see table XXX).

Clustering of genes revealed three co-regulated groups in the full set of contigs and the OC set. The first of gene-clusters (top in **??** and **??**) was in sex-subgroups mainly following an expression pattern differing between populations. The second gene-group was much larger in the full set than in the OC set of contigs (middle in **??**). It was only very weakly reacting to any other factor but sex and was very sparsely annotated (therefore this group was much smaller in the OC set **??**). The third gene-group found again in both the full and OC contigs (bottom in **??** and **??**) was reacting on both the host and population factor in a converse way. Contigs in this cluster were mainly found to be significant for interaction effects.

!!!Figure!!! Clustering of expression values for contigs DE between populations. A heatmap of variance/mean stabilised expression values. Deprograms are based on hierarchical clustering. Green indicates expression be-

low the mean, red above the mean. Experimental conditions are indicated by black bars for groups of samples (columns) below the plot. Presence GO-term annotation for contigs (rows) are given as black bars right to the plot: isOxidoreductase = GO:0016491, oxidoreductase activity; isMitochondrial = GO:0005739, mitochondrion; isELDevelopment = GO:0002164, larval development or GO:0009791, post-embryonic development; isResponsetoStim = GO:0050896, response to stimulus; isPhosphatase = GO:0016791, phosphatase; isMembrane = GO:0016020, membrane; isAntigenProc = GO:0002478, antigen processing and presentation of exogenous peptide antigen; isEndosome = GO:0005768, endosome; isProtLipComp = GO:0032994, protein-lipid complex. Grey bars indicate no annotation available.

Consolidating the clusters with annotation and annotation-enrichment, the first cluster of genes was very well annotated and contained mostly catalytic enzymes involved in oxidation and reduction, the bottom cluster contained more unannotated genes and structural (cuticular collagen) genes.

11 Single gene differences

Tables on single transcript values of OC contigs DE between eel-hosts and populations can be found in additional tables XXX and XXX. Obviously for some contigs differences significant in the model are rendered inaccessibly by comparing simple mean values because of superposed interaction effects or overwhelming general effects of worm sex.

Cytochrome C oxidase subunit 2 (COXII) shows the clearest of all expression patterns for any of the observed genes. It differed significantly only between populations (showed no reaction an any other factor) and was on average over 1,000-fold stronger expressed in the Taiwanese population. At face values differed for every single individual (of the 12 investigated in each populations) at least 20-fold (highest normalised expression was 350 counts in a European worm, lowest normalised expression in any Taiwanese worm was 7,500 counts). Counts summed for orthologs were also significant only for this factor and showed over 10-fold stronger expression in the same direction. This accounts to the fact, that misassembled contigs containing fragments of COXII were only adding experimental noise.

12 Discussion

12.1 Recovery and adaptation

With some reservations discussed below observation of higher recovery of adult worms from sympatric *A. crassus-Anguilla* spp. host-parasite combinations imply local adaptation of different worm populations to host species: Roughly one third of the applied European worms were recovered form their sympatric host *An. anguilla* but only little over 10% for European worms in *An. japonica*. This pattern of recovery was precisely inverted for the Taiwanese population

of *A. crassus*, for which recovery was thus roughly 30% in the sympatric *An. japonica* and only 10% in *An. anguilla*.

Data for the European eel are in agreement with [**?**] and !!!Weclawski!!!, who did however not find lower recovery of the European populatio in the Japanese eel.

An ideally suited phenotype to infer local adaptation in a common garden experiment would be one with or resulting from direct fitness consequences, a so called fitness component. Fitness is defined as the differential contribution to the next generation, therefore such a fitness component would ideally be a measurement on a single individual, and individual life time reproductive success would be an ideal measurement. However, techniques to measure such individual life time reproductive success have not been established in *A. crassus*.

The recovery of certain developmental stages of worms is only a proxy, interpretable as a fitness component. It is a composite measurement of the speed of development from previous lifecycle stages (or speed of migration towards the swimbladder) and of survival. While survival is surely an important component of fitness, it is not completely clear whether fast development and/or migration to the swimbladder are. It is possible that under certain conditions slower development could lead to higher fitness, if it would, for example allow development without attracting the attention of the immune system.

12.2 Divergence not modification of expression

I decided on a study design using pools of individuals for one sex (males) and single individuals for the other. A study on *Fundulus heteroclitus* revealed that approximately 18% of the transcripts are differentially expressed between individual fish from the same population, grown under controlled environmental conditions [**?**]. And it thus not surprising that between individual variation in female samples was leading to higher variance of these female samples compared to pooled male samples in my study.

This interindividual variation in gene expression under a particular environmental condition is generally agreed to be closely linked to a genetic basis [**?**]. For example in a cross between two parental strains of yeast the genetic component of variation was estimated from haploid segregants to be 84% [**?**]. The genetic component was found to be the main factor determining expression level variability between two strains, sexes and ages of *Drosophila melanogaster* for 267 (7%) from 3,931 genes and at least 25% of the transcriptome were estimated to be affected mainly by genotypic factors in any of the groups [**?**]. Variation in the regulation of gene expression is thought to constitute a major source of evolutionary novelty [**?**].

A second study from the line of research on *Fundulus heteroclitus* [**?**] used genetic relatedness as inferred from phylogenetics to separate variation in gene expression in a common experimental environment into a neutral component and a selected component, this way removing variation most likely accounted for by the shared neutral evolutionary history. My case of *A. crassus* is potentially simpler: the investigated European populations are direct descendants

and thus a subset of a Taiwanese source population. In fact I studied two European and two Taiwanese populations as a few hundred kilometers between the geographical origins of the two different locations in Germany and Taiwan probably constitute a barrier to gene flow in a parasite with an aquatic intermediate host. However, I treated worms from both European and Taiwanese populations as replicates (and use the terminology of one European and one Asian population throughout the text) with the rationale of increasing variance for random genetic differences and raising the bar for potentially adaptive differences to be detected.

Gene expression values constitute nothing more than a molecular phenotype. This phenotype is not necessarily a fitness component.

Given the sampling of only twelve Taiwanese worms the question could be raised, whether these constitute a representative sample of the true source population, of which a subpopulation was funding European populations. A microsatellite study indicated gene flow even between populations of *A. crassus* separated by thousands of kilometers in Asia (Japan and Taiwan) [**?**]. Given the high interconnectivity of Taiwanese water systems used for aquaculture both by man build structural links and anthropogenic exchange of fish, a sampling from two Taiwanese populations similarly neutrally diverged from the true European funding population seems very unlikely. The worms sampled from Taiwan can thus be regarded a sample of the metapopulation appropriate for finding differences in relation to the source of the introduction.

Of no surprise was the abundance of differential expression between male and female worms in roughly one third of the genes. A large number of genes are known to be sex specific, regulating ovulation and spermatogenesis throughout the metazoa and especially in nematodes [**?**]. On top of these sex specific genes there are large numbers of genes differently expressed due to differences in metabolism between males and females. Estimates for *Drosophila* based on similar sample sizes to those used in my study range between one and two thirds of the transcriptome showing sex biased expression [?]. In the liver transcriptome of *Mus musculus*, even 70% of transcripts have been shown to differ between sexes [**?**] (note however that this study used 169 female and 165 male mice to guarantee the finding of even the most subtle differences). Given the scale of these differences in other species my estimate of roughly one third of the transcripts in *A. crassus* showing differential expression according to the sex of the worms implies conservative thresholds used in the statistical analysis and moderate power for detection of differences.

Nearly the same proportion (roughly 30%) of contigs was confirmed through summation and analysis of contigs for orthologs in *B. malayi* and *C. elegans*. Development of this orthologous confirmation method was necessitated by the possibly fragmented and chimeric transcriptome assembly. This introduces stringent conditions for the detection of significance, as p-value correction for multiple testing is employed during each analysis (once for raw counts and twice for orthologous counts). Although the underlying tests are not independent, the false discovery rate of 5% for raw contigs can be expected to be immensely lowered by applying a FDR of 10% twice.

In addition biological implications could produce false negatives in such an evaluation: All genes duplicated in *A. crassus* (a) and following antithetic expression patterns will be evaluated negatively, as will duplicated genes in any of the model species (b) following such a pattern. However, there is no other choice then applying these stringent conditions to screen for artefacts producing the same patterns based on mapping to fragmented (a) or chimeric (b) reference contigs. I think that an evaluation based on this scrutinised confidence in an assembly previously computed from 454 data is even more appropriate then an analysis solely based on counts collapsed for orthologs excluding only possible fragmentation artefacts (as used e.g. in [**?**]).

In general, my statistical analysis aimed to minimise false positives (type I error) at expenses of possible false negatives (type II error) and is thus not fully suited to address the proportions of differentially regulated genes.

Nevertheless it is surprising that less than 1% of transcripts were detected differentially expressed between worms in different host species and less then 0.3% were confirmed with the orthologous summation method. This was an unexpected finding, as the differences in the immune response of the host species have a big influence on other phenotypes of worms [**?**]. In addition to the low number of genes, multifactorial analysis revealed that below 10% of the variance could be explained by host species effect, even in significantly differential regulated genes for this factor.

Although these differences between worms in different host species were the most marginal of any of the factors, it is possible to connect some (at least two) of the genes to a prominent physiological difference: the digestion of haemoglobin. Two different aspartic proteases (both confirmed through orthologs, one of them differing for all three main effects, the other for an interaction of worm sex and host species) known to be involved in the first steps of digestion of haemoglobin from other nematodes [?] were overexpressed in worms in *An. anguilla*. This expression phenotype could potentially be linked to the often observed phenotype of bigger size of *A. crassus* in this host [?], as the main contribution to this increase in size is the larger volume of host blood taken up by the parasite. Accordingly the parasite probably digests haemoglobin at a higher rate.

Close to 1% of contigs were significantly different in expression between European and Asian *A. crassus*, making this difference significant for a higher number of contigs than the host differences. For this contrast the proportion of orthologous confirmation was lower than for sex differences but higher than for host species differences. Additionally multivariate analysis of all differently expressed transcripts for worm population revealed that the variance contributed by the population factor was higher than 10% for all significant contigs or even 20% for orthologous confirmed contigs.

The benefit of also allowing contrasting significant differences in interaction terms highlights the power of the GLM-approach.

Another important finding was the large overlap in contigs expressed differentially depending on worm sex and worm population. Such an overlap is expected if genes expressed differentially according to sex are evolving faster towards a differential expression according to other factors. Faster evolution

of reproductive (and especially male specific) traits has been shown in many species at a phenotypic and at a sequence level [**?**]. In *Drosophila*, male reproductive proteins have been shown to evolve at elevated levels and under positive selection [**?**]. Moreover, gene expression should evolve at a higher rate in sex specific genes. Indeed the transcriptomes of *Drosophila* species show that interspecific expression divergence is sex dependent and the action of sex dependent natural selection during species divergence has been inferred from this [**?**, **?**].

Taken together, my findings strongly support a stronger influence of genetic differences between European and Asian populations of *A. crassus* than of the modification in the different host species on gene expression. When additive and interaction effects are considered, the influence of host species even vanishes almost completely in favour of a combination of effects combining parasite population and sex of the worms.

12.3 Functions of genes with genetically fixed expression differences

From a functional perspective, genes identified to differ between populations can be categorised as important in general metabolic processes instead of specific host parasite interactions. This constitutes a negative evaluation of one of my *a priori* hypotheses based on finding parasite specific genes, identified as vaccine candidates in a number of nematodes, within the genes modified or diverged in my study (**??**). However, more direct host parasite interactions are expected in tissue dwelling larval stages (L3 and L4) and in fact most immunomodulators are expressed predominantly in these stages [**?**]. Adults of *A. crassus* could thus be the wrong lifecycle stages to detect such expression differences, if they existed.

12.3.1 Metabolism

Instead enzymes and enzyme subunits important for aerobic respiration are especially expressed at lower levels in European *A. crassus*. In fact, most transcripts significantly differing between populations were annotated as "oxidoreductase" in gene ontology (GO). Downregulation of cytochrome C oxidase subunit 2 (COXII) in the European population of *A. crassus* was the most persistent finding. This downregulation was confirmed by the low expression of the same contig in the European libraries compared to higher expression in all three libraries from Taiwanese worms in pyrosequencing. Cytochrome C oxidase subunits 1-3 are are essential components of respiratory chain complex IV, the cytochrome c oxidase. They are encoded in the mitochondrial genome and coordinate catalytic heme and copper cofactors [**?**].

In fact, not only enrichment analysis highlighted oxidoreductases, but expression values of COXII clustered with other enzymes related to the state of energy metabolism: two lecitin:cholesterol acyltransferase transcripts are putative recently duplicated genes. They showed slightly divergent protein sequences but hit the same orthologs in *C. elegans* and *B. malayi*. They also

shared very similar expression profiles. Expression of different cholesterol acyltransferases has been shown to vary in response to the presence of heme and anaerobiosis in yeast [**?**]. 3-hydroxyacyl-CoA dehydrogenase (involved fatty-acid β-oxidation [**?**]), malate/L-lactate dehydrogenase (from the anaerobic glycolytic pathway or the Krebs-cycle [**?**]) and aspartyl proteases (involved in the digestion of host haemoglobin in helminths [?]) completed this particular cluster.

These patterns can be interpreted as a biological confirmation of the at face values for single genes, especially for COXII. In addition the differential reaction of metabolic genes to different factors (genetic vs. modification) invites speculation on a causal structure behind these correlations. The expressions of metabolic enzymes are interpretable as a change to use a more anaerobic metabolism in the European population of *A. crassus*. In one possible scenario, in European worms one of the subunits of core enzymes of the respiratory chain (probably COXII) would have evolved a genetically fixed lower level of expression. This model follows the logic that the most differential expressed gene could be the driver of observed change. Other enzymes related to aerobic energy metabolism directly or indirectly via the redox state of cells (e.g. lipid metabolism) and only partially controlled by feedback mechanisms from oxidative phosphorylation and the citric acid cycle would show similar patterns of altered expression in European worms. However, the expression of these indirectly and also by additional environmental factors controlled genes would be perturbed when worms are applied back to their Asian hosts. Also in the two sexes differences in size and metabolism would be perturbing the pleiotropic effects of the persistent core change.

Such a scenario also provokes speculation about the adaptive value of such a change in a core metabolic process: aerobic respiration is a potential source for oxidative stress providing a steady source of reactive oxygen species (ROS) as electrons are leaking from the respiratory chain as superoxide anions. It is well established that such ROS production is especially harmful to blood feeding parasites, as free inorganic iron, as well as heme, have the potential to generate additional ROS [**?**]. Anaerobic metabolism is thus thought to occur in many haematophagous parasites as a counter measure against oxidative stress from haemoglobin catabolism [**?**]. It could thus be hypothesised that the bigger size and the larger amount of eel blood ingested leading to a higher rate of haemoglobin digestion provided the selective pressure to reduce aerobic respiration. Additionally helminths can simply get too large to maintain oxygen diffusion to mitochondriae in the absence of a cardiovascular system. As yet proton pumping electron transport constitutes the most profitable energy providing process, the mitochondriae of facultatively anaerobic helminths produce a proton gradient for the use of ATPase with the help of terminal electron acceptors other than O_2 [**?**]. Such an alternate electron sink is fumarate used in many helminths in a process called malat dismutation [**?**].

An interesting implication is that such metabolic differences could potentially be visible ultrastructurally. Indeed in my own diploma thesis [**?**] I identified two different kinds of mitochondriae, one with standard christae like morphology, the other with unusual sacculus like morphology in *A. crassus*. Ad-

ditionally I observed less electron dense inclusions (probably lipid reserves) in bigger worms and more glycogen granulae. The fact that such lipids are less usable under anaerobic conditions led me to the hypothesis that bigger worms are using less aerobic processes. Reanalysing this data and probably obtaining new data with additional histochemical staining methods could be a way to put gene expression into a physiological perspective. Furthermore, a biochemical examination of isolated mitochondriae could highlight changes in the mitochondrial respiratory chain under *in vitro* conditions [**?**]. Such direct measurements of COX enzyme activity (using well established assays [**?**]) would be desirable to establish even the validity of the first logical step in these adaptive speculations that underexpression of COXII is leading to decreased enzyme activity. It would be counterintuitive to expect higher enzyme activity when COXII mRNA levels are low, but, for example, in *Schistosoma mansoni* COXI over expression in praziquantel resistant strains is leading rather to decreased enzyme activity [**?**].

The sensitivity to perturbation of mitochondrial genes for respiratory chain complexes in nematode parasites is underlined by their upregulation after depletion of Wolbachia from filarial nematodes [**?**, **?**]. Wolbachia are obligate endosymbiont bacteria of some clade III nematodes, they are supplying heme to non-haematophagous parasites in the absence of an intrinsic pathway for heme synthesis [**?**] (which is absent also in free living *C. elegans* [**?**]). While my sequence analysis suggests the absence of wolbachial symbionts in *A. crassus*, such studies support a central role of host or endosymbiont derived heme for respiratory processes and suggest a propensity for evolutionary change in related processes (in Filaria even acquisition of an endosymbiont).

Assuming a genetically fixed lower expression of COXII in European *A. crassus* as a driver for other metabolic differences does not imply a simple regulation of the expression itself, or a genetically simple change underlying the changed expression phenotype. Regulation of the mitochondrially encoded genes has been extensively integrated into the regulatory network of eukaryotic cells and is controlled by and interacting with nuclear transcription factors [**?**].

Intriguingly overexpression of respiratory chain enzymes was limited to cytochrome c oxidase transcripts (COXII and to lesser extent also COXI and COXIII). Mitochondrial transcription produces multiple polycistronic unmatured transcripts, which are cleaved and modified in their expression post-transcriptionally. Cleavage occurs at t-RNA sequences interspersed between protein coding genes and can be imperfect to leave some transcripts polycistronic in a matured state. Nevertheless, due to posttranscriptional modification individual transcripts can be expressed uncoordinated, even when expressed on the same unmatured polycistronic transcript [**?**]. The addition of poly-A tails, for example, is vital for stability of mature transcripts in metazoans. The mitochondrial genome contains only very little untranscribed sequence, is polyploid (once homoplasmic, essentially maternally inherited like haploid) and transmitted completely linked, with very scarce recombination events [?].

Cis-regulatory change in a control region would thus be very easily detectable in my transcriptome data. Even if the sequence variation leading to the observed expression phenotypes would locate to the untranscribed hypervariable

mitochondrial control region (in D-Loop associated promoters), selection on such a variant would render the whole mitochondrial genome inadequate for phylogenetic analysis, as a variant sweeping to fixation would have removed polymorphism from the complete mitochondrial genome due to the prefect linkage [**?**]. If a sweep would be presently ongoing, high levels of heteroplasmy would be found in single individuals [**?**]. Such a pattern has not been found in populations of *A. crassus* in Europe when COXI was used as a marker [?, **?**] (see also figure) and is also not visible from preliminary analysis of polymorphism in mitochondrial genes in my RNA-seq data.

Functional constraints are also expected regarding the mechanism by which the expression of COXII could evolve. Most infective L3 larvae of parasitic nematodes rely on aerobic respiration [**?**]. Dixenous parasites like *A. crassus* migrate through tissues of definitive hosts, where oxygen is readily available, after leaving the haeomocoel of the intermediate host. Enzyme subunits building a functioning aerobic respiratory chain are thus likely to be expressed at earlier lifecycle stages of *A. crassus* and elevated anaerobiosis is expected to be restricted to the adult stages.

These considerations make sole or predominant cis-regulatory change in mitochondrial DNA unlikely to explain the divergent expression phenotypes. Still identification of the genetic architecture, for example sequence variation in a transcription factor, a co-factor or a protein modifying mitochondrial transcripts, may be possible (to a limited extent even in the present RNA-seq data).

RNAi screens in *C. elegans* for increased lifespan focus on genes leading to lower oxygen consumption and altered mitochondrial morphology and function [**?**]. Such candidate genes will provide an additional link back to functional considerations once screening for genomic regions with signature of selection will highlight candidate loci.

12.3.2 Collagens

A second group of genes differentially expressed in populations of *A. crassus* emerged from both cluster and enrichment analyses. Two transcripts in this cluster were significant for interaction effects between host species and parasite-population, they were annotated as collagens. For both genes this meant an "adjusted" (to avoid the suggestive "adapted") expression difference leading to a lower expression in sympatric host species/parasite population pairs. Cuticle collagens are a large multigene family (Interpro lists 164 entries for "Nematode cuticle collagen, N-terminal" for *C. elegans* and 51 for *B. malayi*), containing extensive repeat regions: roughly 50% Gly-X-Y residues, often Gly-Pro-Hpy. In the genome of *B. malayi* 82 genes encoding collagen repeats have been found [?]. It was thus very important to have orthologous confirmation for these two contigs, as misassembly could have easily lead false positives here.

The two collagens were clustered with a third contig sharing a collagen annotation (failing to be significant for the interaction term probably because of low overall expression) and a contig annotated as "Matrixin" (a metallo-proteinase assumed to be involved in remodelling of the extracellular matrix [**?**]) and a

ABC-transporter family protein.

Functional speculations are more difficult for collagen than for the respiratory chain enzymes. The cuticle constitutes an exoskeleton and a barrier between the worm and its host environment. Synthesis of most collagens is believed to occur at negligible levels in adult male worms and is rather constrained to discrete temporal periods in larval development, the moults [**?**]. The differential expression could thus be due to changes in larval development or due to alternations in the low level, steady renewal of the adult cuticle and remodelling of the extracellular matrix of hypodermis cells. Some considerations would favour of the second explanation: in *C. elegans* genes expressed after reproductive maturity evolve faster than genes expressed earlier in development [?]. This suggests a model of elevated pleiotropic effects in genes expressed at earlier stages of development and hence more conserved expression patterns in larval stages. Independent of these considerations, both the primary assembly and the constant remodelling of the cuticle involve complex post-translational processes hardly accessible at the transcriptomic level: a zipper-like nucleation/growth mechanism leads to the folding of a triple helix of and heterotrimers and homotrimers [?]. If and how differential expression of two particular collagens interferes with this process requests further research. As for the metabolic differences, differential expression patterns could be reflected in morphology. One approach would be to measure thickness and density of the cuticle of worms from coinoculation experiments.

13 Outlook

The presented project on the divergence of gene expression obviously constitutes work in progress. The observed differences in subunits of respiratory chain enzymes, especially in COXII, necessitate and permit confirmation by reverse transcription quantitative PCR (RTqPCR) for these transcripts. Such evaluations of a single gene (or few genes) will be possible on many individual specimen of *A. crassus* from both Europe and Taiwan to further test the significance of the observed differences. Therefore, in addition to the validation of expression values for sequenced samples, many of the worms from the presented coinoculation experiment yielding lower amounts of RNA inadequate for sequencing will be used to further establish the divergence in gene expression. Additionally sampling of worms from their present day sympatric hosts is possible for genes differing only for populations unconditional on eel host species. Moreover, if selection in Europe would have acted on standing variation, one would expect to find worms expressing for example COXII at low levels also in the Taiwanese source populations, at least in low frequency.

An assembly of the mitochondrial genome of *A. crassus* from preliminary genome-sequencing data (discussed below) and the identification of the polycistronic unmatured and, if present, matured transcripts (similar to [?]), will further inform and validate the analysis of the expression of mitochondrial genes. Additionally, disentangling assembly artefacts complicating mapping from real

nuclear or even mitochondrial [**?**] pseudogenes of mitochondrial genes will help increasing the power of expression analysis and furthermore permit the analysis of interaction of such pseudogenes with the expression of functional genes.

Multiple starting points also exist for further functional examination of metabolic change, as mentioned throughout the text. However, the search for ultimate causes for evolutionary change *sensu* [**?**] will potentially be even more rewarding.

I will expand the RNA-seq analysis presented here to study allele-specific expression and the association between gene expression and sequence variants. This kind of quantitative expression trait locus (eQTL) analysis is possible as both sequence and expression information are available from the present RNA-seq data. Both simple cis-acting variation in promoter or enhancer regions, as well as trans-acting variation can theoretically be detected [**?**]. To detect trans-acting variants, however, might be impossible with the (for population studies) relative low number of sequenced individuals, as it relies on statistical associations requiring broad sampling. Yet, cis-acting variation, more readily detectable as allele-specific variation, is unlikely to explain variations in mitochondrial gene expressions for the reasons discussed above.

Therefore, large scale meta-population wide sampling must not be limited to an evaluation of the divergent gene-expression phenotypes, but has to further elucidate the population genetic relationships between Taiwanese and European worms. A future research program will thus need to employ population-scale sampling of genotype data, densely spread across the genome. Genotyping of many European *A. crassus* from different populations and comparison with many individual genomes from different Asian populations will enable tests for selection: based on the fact that around selected variants nucleotide diversity is reduced by hitchhiking of neutral variation in so called selective sweeps [**?**], a punctual increase of population differentiation measured by the fixation index F_{st} [**?**] in regions linked to selected variants can be measured. Other well established population genetic measurements include Tajima's D, a measure based on the allele frequency spectrum [**?**]. When these methods are applied on a genome wide scale the neutral null-expectation to separate a loss in variability based on selection from neutral loss due to demography is given by the diversity across all regions of the genome. A microsatellite study [?] as well as my own evaluations (based on pyrosequencing see **??**) and RNA-seq (data not shown) indicate only a moderate genetic bottleneck caused by the introduction of *A. crassus* to Europe and thus the necessary neutral diversity as a background for these tests will be present.

Furthermore statistical models need to be parameterised by divergence time to disentangle the influence of demography and selection (i.e. to estimate the effective population size). Reliable estimates for divergence time are readily available for the introduction of *A. crassus* to Europe: 60 to 90 generations. As for such a short period linkage to putatively selected variants will not be broken down in large blocks, marker density is of minor concern, but priority should be given to the breadth (many individuals from many populations) of sampling.

One methods enabling such population wide genotyping emerging from NGS

technology is the sequencing of restriction-site associated DNA (RAD) markers. Preparation of RAD libraries involves digestion of genomic DNA with a restriction enzyme. Individually tagged adaptors can then be ligated to the fragments and individual samples can be pooled. The choice of restriction enzyme is important to optimise the number of restriction sites (depth of sampling the genome) relative to the number of individual samples being investigated [**?**]. In the case of *A. crassus* this optimisation also concerns the minimisation of restriction sites in host-genome, as present in unavoidable contamination.

The *de novo* assembly of a reference genome for *A. crassus* will enable the search for such an optimal restriction enzyme. Preliminary data has been generated for a female individual of the Polish population on one lane of the Illumina HiSeq machine, giving 110 million 100 bases long paired-end reads, in total over 10 gigabases of sequence data.

A preliminary assembly yielded a mean coverage of below 15-fold, for the *A. crassus* derived contigs. This coverage is surprisingly low given the large amount of input-data and I will need to construct improved assemblies informed by the analysis of this preliminary assembly. A seemingly trivial but nevertheless important prerequisite for any high-throughput genomic sequencing project on a parasite was the confirmation that genomic DNA could be obtained sufficiently clean from other xenobiont DNA.

!!!Figure!!! GC-content and coverage for a preliminary genome assembly. A preliminary assembly of roughly 10 Gb sequence data in over 110 million reads. The analysis of GC-content and coverage identifies host-contamination at higher GC, but lower coverage. Coverage and GC-content separate two distinct data-sources: a lower GC/higher coverage nematode subset and a higher GC/lower coverage eel subset (confirmed by BLAST [**?**]). For this sequencing library only 10-20% of the reads are lost to eel-host derived off-target data. The preliminary assembly was provided by Sujai Kumar from Mark Blaxter's lab.

It has been possible to isolate roughly 1μg of genomic DNA from a big individual worm. Only ca. 20% of the DNA were derived from the genome of the eel-host (see figure). As only 300 ng of DNA material (with low amounts of contamination with host-blood) are needed for RAD-sequencing, this can be achieved in most big specimen of *A. crassus*.

For both reference genome assembly and annotation and for the future genome-scans I will continue to collaborate with Mark Blaxter's laboratory at the University of Edinburgh. This group is actively developing methods especially for RAD-sequencing and applying them to questions in evolutionary model-species [**?**].

Another useful strategy enabled by RAD-sequencing is the construction of a physical genetic map in families of *A. crassus* (backcross is impossible). In addition to the population scale approaches outlined above mapping of gene-expression quantitative trait loci (eQTL) in mapping crosses between the two divergent expression-phenotypes constitutes a promising route for the investigation of genomic variants underlying the divergent expression-phenotypes. Once transcripts can be anchored on genomic contigs and linkage groups can be constructed to build a physical map of the genome, a readout for hybrid F2 indi-

viduals could even be transcriptomic data (RNA-seq) providing both genotype and expression-phenotype.

A prime example for a research program on the evolution of ecologically important traits is provided by the Stickleback *Gasterosteus aculeatus*: QTL-mapping has been performed to fine-map the loss of lateral plates in freshwater populations [?] and parallel adaptation has been investigated using population genomics [**?**]. Both approaches used RAD-sequencing. The sophistication and depth of insight available in such an evolutionary model species is underlined by research on adaptive reduction of pelvic structures, an evolutionary trajectory shown to be favoured by the localisation of the underlying change in an instable region of the genome [**?**].

The hope to develop a similar research program based on the present humble thesis seems presumptuous. Nevertheless, making full use of the advances in sequencing technology it might be possible to rapidly gain insight into the genomic organisation underlying contemporary evolutionary change. The present RNA-seq data will be crucial in achieving this goal, as it will be used to link expression phenotypes with genomic sequence. An evolutionary leap in a core metabolic process seems possible.

The ability to evolve via such a leap could even be an evolutionary old trait retained in *A. crassus* allowing it to colonise new hosts. Therefore, comparative genomics relating population genetic processes in *A. crassus* to putatively adaptive change during the acquisition of new host by other *Anguillicola* species in evolutionary time constitutes another route of research. If such a link between microevolutionary processes in *A. crassus* and the evolution of *Anguillicola*-species would exist, it would provide general insight in the evolution of parasitic phenotypes.