Divergence of an introduced population of the Swimbladder-nematode *Anguillicola crassus* - a transcriptomic perspective

Zur Erlangung des akademischen Grades eines
DOKTORS DER NATURWISSENSCHAFTEN
(Dr. rer. nat.)
an der Fakultät für Chemie und Biowissenschaften
des Karlsruher Institut für Technologie (KIT) - Universitätsbereich
vorgelegte
Dissertation
von

Emanuel Heitlinger

geboren in
Schwäbisch Gmünd

Dekan:

Referent: Prof. Dr. Horst Taraschewski

Korreferent: Prof. Mark Blaxter

Tag der mündlichen Prüfung:

Abstract

The ability to expand into new environments and niches, despite being highly adapted for survival in their habitual environment, is a fascinating feat of organisms. The propensity of *Anguillicola crassus* to capture new hosts can serve as a model for an extreme case of this, as parasites are thought to live in competition with and closely adapted to their hosts. Differential selection in such new environments leading to local adaptation is considered a driving force of divergence and thus for the origin of species and biotic diversity.

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

Gene regulatory networks, as a bridge between genotype and phenotype, are thought to play a central role both in the response to stress (e.g. from sofar unexperienced environmental stressors) and in the divergence and eventually establisment of reproductive barriers between populations.

In the present project the differences in gene-expression in *A. crassus* populations should be illuminated and genetic components of differences isolated. With this aim we conducted cross-innoculation experiments with both Asian and European host-species and parasite populations.

In our sequence assembly we identified twelve proteases under positive seleciton highlighting this group of enzymes as possible targets of an immune-attack on *A. crassus*.

In

Zusammenfassung

Die Fähigkeit sich in neuen Umgebungen und Nieschen auszubreiten, obwohl sie höchst angepasst an ihren angestammten Lebensraum sind, stellt eine faszinierende Eigenschaft von Lebenwesen dar. Der Wechsel der Wirtsart durch *Anguillicola crassus* kann als Modell für einen Extremfall dieses Vorganges gesehen werden, bei dem Parasiten neue Wirte besiedeln. Selektion in solch einer neuen Umgebung, die zu einer Anpassung führt gilt als eien treibende Kraft für Divergenz und so zum Entstehen neuer Arten und biologischer Vielfalt. Gen-regulatorische Netzwerke, als eine Brücke zwischen Genotyp and Phenotyp, haben eine zentrale Rolle sowohl in der Antwort auf Stress (etwa durch eine veränderte Umwelt) als auch in der Entwicklung von Barrieren für die Fortpflanzung.

Im hier vorgestellen Projekt sollen die Unterschiede im Transkriptom zweier Populationen von *A. crassus* untersucht werden. Der Parasit wurde vor 30 Jahren nach Europa eingeschleppt, wo er sich erfolgreich in einer neuen Wirtsart ausbreitet und etablierte.

To my grandmother Ruth my brother Roman and my wife Silvia

Acknowledgements

I would like to acknowledge the thousands of individuals who have coded for free software and open source projects. It is due to their efforts that code is shared, tested, challenged and improved. Sharing their intellecual property as a general good, they serve progress in science and technology.

Contents

List of Figures

List of Tables

Glossary

DNA Desoxy Ribonucleic Acid; a chemical molecule bearing the heritable genetic information in all life on earth

dpi Days post infection; In infection experiments, a point in time given in days after an individual has been infected

ORF Open Reading Frame; a region in a DNA-sequence begining with a start-codon and not containing a stop-condon. For example a region within a processed mRNA transcript being transcribed into a protein

SNP Single Nucleotide Polymorphism; variation occurring in a single nucleotide between two closely related homlogous sequences. Leading to for example to allelic differences within a population or even the homologous chromosomes in an individual

1

Introduction

1.1 The study organism: *Anguillicola crassus*

1.1.1 Ecological significance

Anguillicola crassus Kuwahara, Niimi and Ithakagi 1974 (1) is a swimbladder nematode naturally parasitising the Japanese eel (*Anguilla japonica*) indigenous to East-Asia. In the last 30 years anthropogenic expansions of its geographic- and host-range to new continents and host-species attracted interest of limnologists and ecologists. The newly acquired hosts are, like the native host, freshwater eels of the genus *Anguilla*, and the use of the definitive host seems to be limited to this genus (2). However the nematode displays a high versatility and plasticity in most other aspects of it's life, and this has been proposed as one of the reasons for its success invading new continents (3).

A. crassus colonized Europe in the early 1980ies and spread through almost all populations of the European eel (*Anguilla anguilla*) during the following decades (reviewed in (4)). This spread includes populations of the European eel in North Africa(5, 6). At the present day *A. crassus* is found in all but the northernmost population of the European eel in Iceland (7). It has to be noted however, that low water temperature (8) and salinity (9) limit the dispersal of *A. crassus* larvae and thus high epidemiological parameters are rather expected in freshwater and in southern latitudes.

Wielgoss et al. (10) studied the population structure of *A. crassus* using microsattelite markers and inferred details about the colonization process and history. Their data are in good agreement with previous knowledge about the history of introduction

and dispersal. Therefore the process of introduction and spread can be considered very well illuminated:

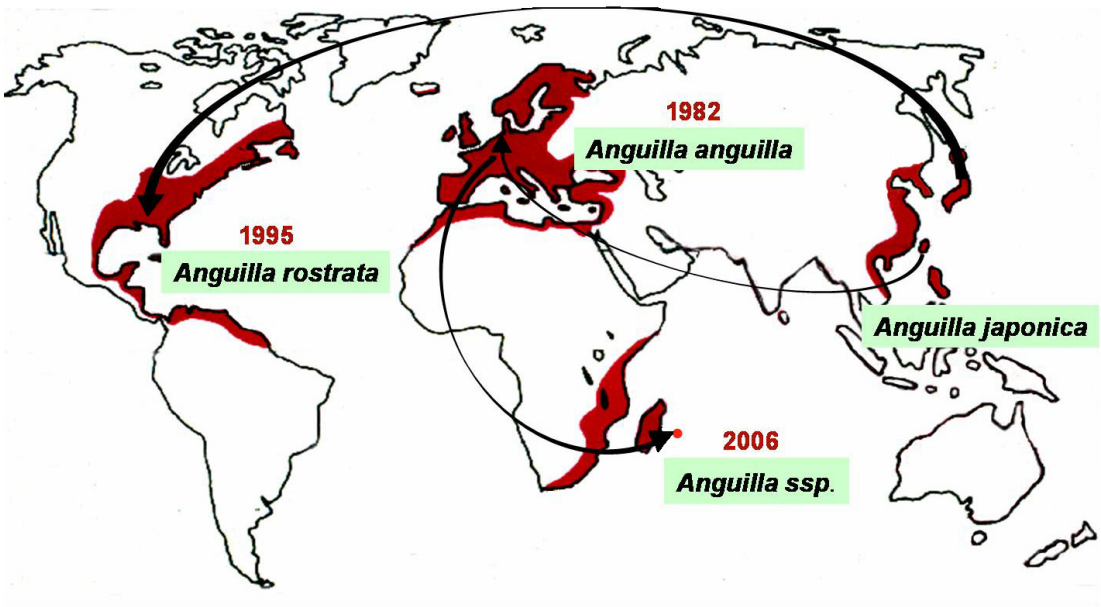

Figure 1.1: Transcontinental dispersal of *A. crassus:* - Invasions of different continents by different source-populations are illustrated using arrows. Red color indicates the range of the eel species targeted by the invasion. Modified form (11), based on data reviewed in (4) and newer findings in (10) and (12).

A. crassus was first recorded in 1882 in North-West Germany, and this record was published in a German fishery magazine in 1985 (13). The import of Japanese Eels from Taiwan to the harbor of Bremerhaven in 1980, was soon identified as most likely source of introduction (14). Taiwan as the most likely geographical source of the introduction was in turn also inferred from population structure using microsatellites. Furthermore, from the fact that genetic diversity is highest in northern regions of Germany and gradually declines to the south, Wielgoss et al. concluded a single introduction event to Germany as source for all populations of *A. crassus* in the comprehensive set of investigated populations of the European eel. This signal was persistent together with a punctual signal for anthropogenic mixing of eels and parasite populations due to restocking (15). However a recent study found additional haplotypes for Cytochrome C oxidase subunit I (COXI) in Turkey, and a second introduction to the Eastern Mediterranean seems possible. These Turkish haplotypes cluster with Taiwanese haplotypes

and the introduction source would be similar to the main introduction Laetsch et al. !!!!CITE (see also figure 1.5).

A second colonization of *A. crassus*, succeeded in North-America. Since the 1990s populations of the American eel (*Anguilla rostrata*) have been invaded as novel hosts (16, 17, 18). Wielgoss et al. identified Japan as the most likely source of this American population of *A. crassus* using microsatellite data. Laetsch et al. CITE!! showed that all sources populations for different introductions (even the introduction to the US from Japan) are from one of two separated clades of *A. crassus* endemic all over East Asia (see also figure 1.5).

Finally *A. crassus* has been detected in three indigenous species of freshwater eels on the island of Reunion near Madagascar (12).

Copepods and ostracods serve as intermediate hosts of *A. crassus* in Asia, as well as in the introduced ranges (19). In these hosts L2 larvae develop to L3 larvae infective for the final host. Once ingested by an eel they migrate through the intestinal wall and via the body cavity into the swimbladder wall (20), i.a. using a trypsin-like proteinase(21). In the swimbladder wall L3 larvae hatch to L4 larvae. After a final moult from the L4 stage to adults (via a short pre-adult stage) the parasites inhabit the lumen of the swimbladder, where they eventually mate. Eggs containing L2 larvae are released via the eel's *ductus pneumaticus* into it's intestine and finally into the water (22). The time needed for the completion of a typical life-cycle from egg to reproducing female is interesting to determine the number of generations European populations of *A. crassus* have spent in their newly acquired environment. Based on laboratory infections it can be estimated to vary between 70 and 120 days at water temperatures around 20°. Such an estimate is leading to 2 generations completed per year in Europe and a total of circa 60 generations since introduction.

High prevalence of the parasite of above 70% (e.g. (23, 24)), as well as high intensities of infections were reported, throughout the newly colonized area (25). In the natural host in Asia prevalence and intensities are lower than in Europe (26).

One of the possible differences between Asian and European population of *A. crassus* could be the widespread use of paratenic hosts in European waters (27, 28). Such a use of paratenic hosts has not been reported from the Asian range of the parasite and there are some speculation that the use and availability of paratenic hosts could be a factor explaining the success of invasion or even the higher epidemiological parameters

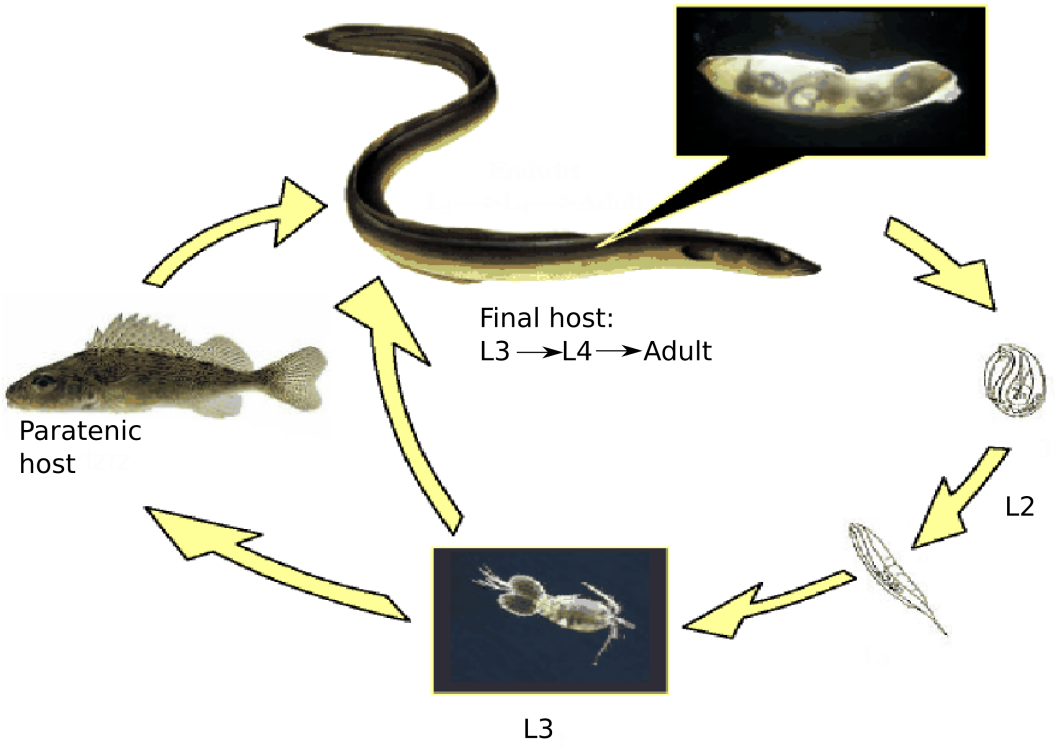

Figure 1.2: Life-cycle of *A. crassus* - Adult females deposit already hatched L2 in the lumen of the swimbladder. Larvae migrate through the *ductus pneumaticus* and the intestine into the open water. Copepods serve as intermediate host where infective L3-larvae develop. These can be transported and accumulated in paratenic hosts or directly ingested by an eel. They migrate through the eel's intestinal wall into the swimbladder wall. After the final molt to adults worms arrive in the lumen of the swimbladder, feed on blood and reproduce. Modified from (11).

in Europe compared to Asia. However the lack of evidence for the use of paratenic host in Asia is rater likely to be a a result of the lack of appropriate studies in Asian water systems, given the broad spectrum of paratenic hosts use by *A. crassus* (28, 29, 30), including even amphibians and larvae of aquatic insects (31).

Also the abundance of the final hosts *An. anguilla* and *An. japonica* itself could have an effect on epidemiological parameters (32). This parameter however is thought to be similar for each of two host-species in its endemic area (33), the density of the host-species however are in decline for the last decades both in Asia and Europe (34).

These factors are thus unlikely to explain the differences in epidemiological parameters and the differences in abundance and intensity of *A. crassus* infections in East Asia compared to Europe are commonly attributed to the different host-parasite relations in the final eel host permitting a differential survival of the larval and the adult parasites (35, 36).

The impact of *A. crassus* on the European eel has been a major focus of research during the past decades. Pathogenic effects on the eels can lead to mortality of eels, when combined with co-stressors (37).

Especially the changes in the tissue of the swimbladder wall have been shown to influence swimming behavior and it has been speculated that eel may fail to complete their spawning migration (38). While nobody would claim Anguillicolosis (the condition caused by *Anguillicola*) to be the main reason for the decline of eel stocks, it could very well be a cofactor (39) to the tragic main factor of overfishing of glass-eels (34).

Responses in *An. anguilla* have hallmarks of pathology, including thickening (40) and inflammation (41) of the swimbladder wall, infiltration with white blood cells and dilated blood vessels.

Data from experimental infections of *An. anguilla* with *A. crassus* suggest that in this host the parasite undergoes (under experimental conditions) a density-dependent regulation keeping the number of worms within a certain (high) range (42).

In contrast to the European eel, the Japanese eel is capable of killing larvae of the parasite after vaccination (43) or under high infection pressure (44): A high mortality of *A. crassus* larvae has been reported in the swimbladder wall of *An. japonica* (26) and under high infection pressure even more pronounced in the intestinal wall (44).

Furthermore it has been shown that the establishment of encapsulated larvae inside the intestinal wall is related to killing of larvae in the swimbladder wall: significant

numbers of encapsulated larvae in the intestinal wall were not observed when capsules in the swimbladder-wall were absent. No capsules in the intestinal wall have been found in single, non-repeated experimental infections of Japanese eels, while larvae are killed in the swimbladder wall. These observation shows that larvae are first encapsulated in the swimbladder wall and encapsulation inside the intestinal wall follows only repeated heavy infections. These features suggest a major role of acquired or infection induced immunity in the formation of capsules (44).

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 (35).

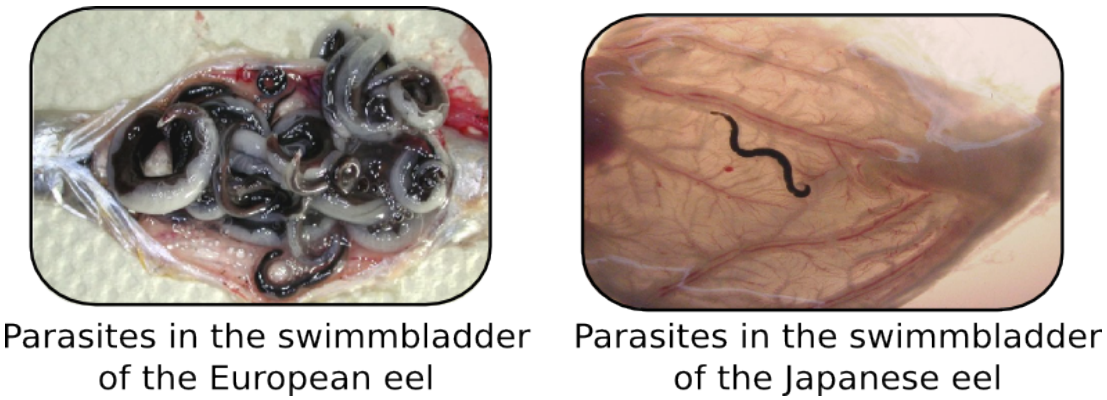

Figure 1.3: Difference between worms in the swimbladder of the European eel and the Japanese eel - Note the bigger size and higher number of worm in a typically infected European eel. In comparison in the Japanese eel worms are smaller and intensities of infection are much lower. The dark brown matter is ingested eel-blood visible through the transparent nematode body- and intestinal wall, the white matter are developing eggs and larvae in ovaries of female *A. crassus*.

1.1.2 Evolutionary significance

1.1.2.1 The eel-host

With a view on the potential co-evolution and especially adaptation of *An. anguilla* to *A. crassus* the catadromous reproduction of freshwater eels might play an important role. Individuals of both Atlantic species *An. anguilla* and *An. rostrata* migrate thousands of kilometers to reproduce in the area of the Sargasso sea (45). The Japanese

eel in its endemic area migrates to the west of the southern West Mariana Ridge (46). Eel larvae then migrate to their freshwater habitats with the help of oceanic currents. While hybrids between the two Atlantic eel species have only been reported from Iceland (47), European eels as a species are considered panmictic (48): Signals for population structure, interpreted as evidence against panmixia first (49), have been shown to be an artifact of temporal variation between cohorts of juvenile eels (47, 50, 51). Such panmixia reduces the effectiveness of selection. Uninfected populations participating in reproduction make rapid local adaptation to a prasite less likely.

Interestingly it has been shown, that individual genetic heterozygosity in *An. anguilla* is no predictor for *A. crassus* infestation (52). This is remarkable, as in a diverse spectrum of organisms such as plants, marine bivalves, fish or mammals correlations between heterozygosity and fitness-related traits and especially with parasite-infestation have been observed (53, 54). Variation at highly polymorphic loci is one of the cornerstones of host-adaptation (55). Once variation is present in a population, overdominance (or heterozygote superiority) can favor heterozygous individuals (56, 57). Matching parasite antigens and allowing to present them as an epitope, the MHC class II molecule for example has been demonstrated to be under diversifying selection in many vertebrate species. Stickleback display variable copy-numbers of a class IIb MHC gene and *A. crassus* using it a paratenic-host has been shown to select for variability and heterozygosity at these loci (58). Vice verca the vertebrate immune system and especially its memory component is thought to be driving positive selection on antigens of microorganisms (59).

Morphological and functional differences between the immune systems of teleost fishes and other vertebrates (especially mammals) are prevalent (60). The immune system of eels especially differs in many details. It lacks all but the M-class of antibodies and response to macro-parasites is carried out mainly by neutrophile rather than eosinophile granulocytes (61). However, the immune systems of mammals and fish also show some genetic, molecular and cellular similarity. While for example the Atlantic cod has lost genes for MHC II (62), this gene shows conservation in the adaptive immune system of jawed vertebrates (63) and its presence has been confirmed in transcriptome data for *An. anguilla* (64).

A decline of epidemiological parameters for European populations of *A. crassus* has been hypothesized based on data published over two decades. This decline however,

has not been confirmed in an explicit meta-analysis. If it would be present, possible explanations would include lower population density of the eel (likely (32)), an evolution of the eel host towards better resistance (rather unlikely; see above), and an evolution of *A. crassus* towards lower or at least altered virulence (part of the present investigation).

1.1.2.2 Interest in *A. crassus* based on its phylogeny

The genus *Anguillicola* comprises five morphospecies (65): In East Asia, in addition to *A. crassus*, *A. globiceps* Yamaguti, 1935 (66) parasitises *An. japonica*. *A. novaezelandiae* is endemic to New Zealand and South-Eastern Australia in *Anguilla australis* and *A. australiensis* Johnston et Mawson, 1940 (67) parasitises the long-fin eel *Anguilla reinhardtii* in North-Eastern Australia. Finally *A. papernai* is known from the African longfin eel *Anguilla mossambica* in Southern Africa and Madagascar.

In 2006 F. Moravec promoted the the former subgenus *Anguillicoloides*, comprising all species but *A. globiceps*, to the rank of a genus (68). This subdivision of the Anguillicolidae in two genera was revised based on the rejection of monophyly of the new genus *Anguillicoloides* and *"Anguillicoloides crassus"* was restored to *Anguillicola crassus* by CITE!! Laetsch. In the same study, *A. crassus* was identified as the basal species in the genus, analyzing the nuclear genes small ribosomal subunit (nSSU) and large ribosomal subunit (nLSU, see figure 1.4). An alternative phylogenetic hypothesis derived from mitochondrial cytochrome c oxidase subunit I (COX I) sequences would place *A. crassus* in a clade with the oceanic species and *A. globiceps* and *A. papernai* in a sister clade (see figure 1.5).

Neither of these phylogenetic hypotheses is compatible with the phylogeny of the eel-hosts without host-switching: Assuming the establishment of *Anguillicola* in an ancestral Indo- pacific host at least three host-switch events are needed, even to explain classical (non-recent, i.e. non-anthropogenic) host-parasite associations. Two of these host-capture events must have spanned the major splits in the eel phylogeny (69): Oceanic *Anguillicola* must have captured hosts transitioning between the clade of *An. reinhardtii* and *An. japonica* to the clade in which *An. australis* is found. Also the basal species of freshwater eels *An. mossambica* must have been captured in an host-capture event involving a phylogenetically distant host-species.

The recent anthropogenic host-switches of *A. crassus* from *An. japonica* to *An. anguilla* and *An. rostrata* constitute additional acquisitions of phylogenetically well

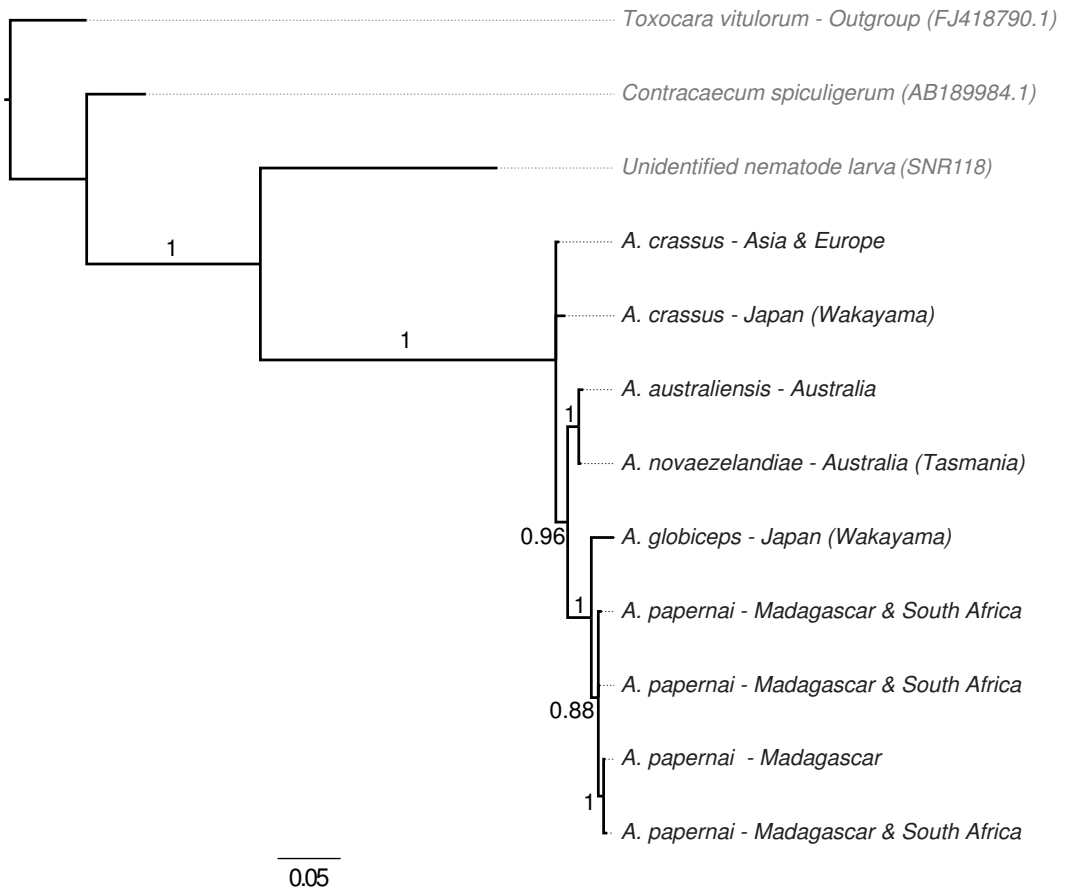

Figure 1.4: Phylogeny of the genus *Anguillicola* based nLSU - Phylogram inferred from nuclear large ribosomal subunit (nLSU) of *Anguillicola* and outgroups using Bayesian Inference. Labels on internal branches indicate Bayesian posterior probabilities. From Laetsch et al. CITE!!

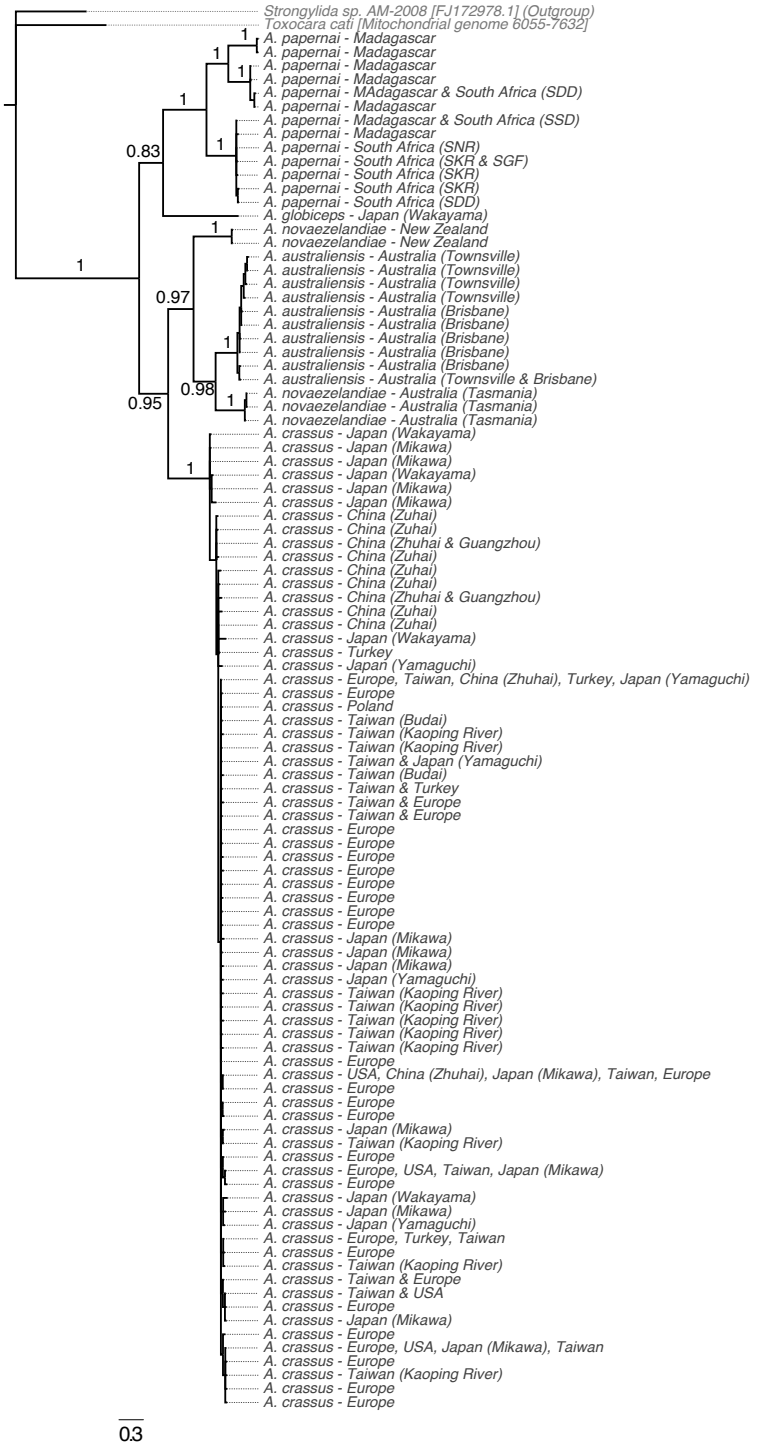

Figure 1.5: Phylogeny of the genus *Anguillicola* based on COXI - Phylogram inferred for *Anguillicola* and outgroups based on mitochondrial Cytochrome C oxidase subunit I (COXI) using Bayesian Inference. Labels on internal branches indicate Bayesian posterior probabilities. From Laetsch et al. CITE!!

separated hosts. This affinity for host-switching may be an evolutionary relict found only in one of the two clades (putative cryptic species) in which *A. crassus* can be devided !!CITE Laetsch.

The to date most likely phylogenetic hypothesis places the genus *Anguillicola* (the only genus in the family Anguillicolidae) at a basal position in the Spirurina (clade III *sensu* (70)), one of 5 major clades of nematodes (71, 72). The Spirurina exclusively exhibit a animal-parasitic lifestyle and comprise important human pathogens as well as prominent parasites of livestock (e.g. the Filaroidea and Ascarididae). The finer subdivision of the Spirurina into Spirurina A, and the Sister clades Spirurina B and C from Laetsch et al. can be seen in figure 1.6.

Within the Spirurina B an enormous phylogenetic diversity of the definitive hosts can be observed, ranging from fresh-water fish as hosts for the Anguillicolidae to cartilaginous fish for Echinocephalus, mammals parasitised by Gnathostoma and Linstowinema to reptiles as hosts for Tanqua. In addition to this diversity, a common characteristic of Spirurina B and C is a complex life-cycle involving freshwater or marine intermediate hosts. Application of parsimony principles thus favors a complex life history as the ancestral state for the Spirurina.

This phylogenetic position makes the Anguillicolidae an interesting system as outgroup taxa to understand the evolution of parasitic phenotypes in the Spirurina. In addition the recent anthropogenic expansion of *A. crassus* to new host species provides the opportunity to observe phenotypic modifications as well as early genetic divergence making it an ideal model for basal Spriurne Nematodes.

1.1.2.3 A taxonomy of common garden experiments and the divergence of *A. crassus* populations

Common-garden and transplant experiments are a method to separate genetic components (G) of phenotypic differences from environmental (E) influences, used for almost as long as scientists investigate evolution (73, 74).

The goal of a classical common garden experiment is the exclusion of environmental factors: By carefully choosing an universal environment (the garden) genetic differences between potentially diverged population of a species should be isolated and elucidated. This approach is equivalent to one-factorial design investigating only the genetic factor (G). However, an experimental design aiming to exclude environmental effects bears

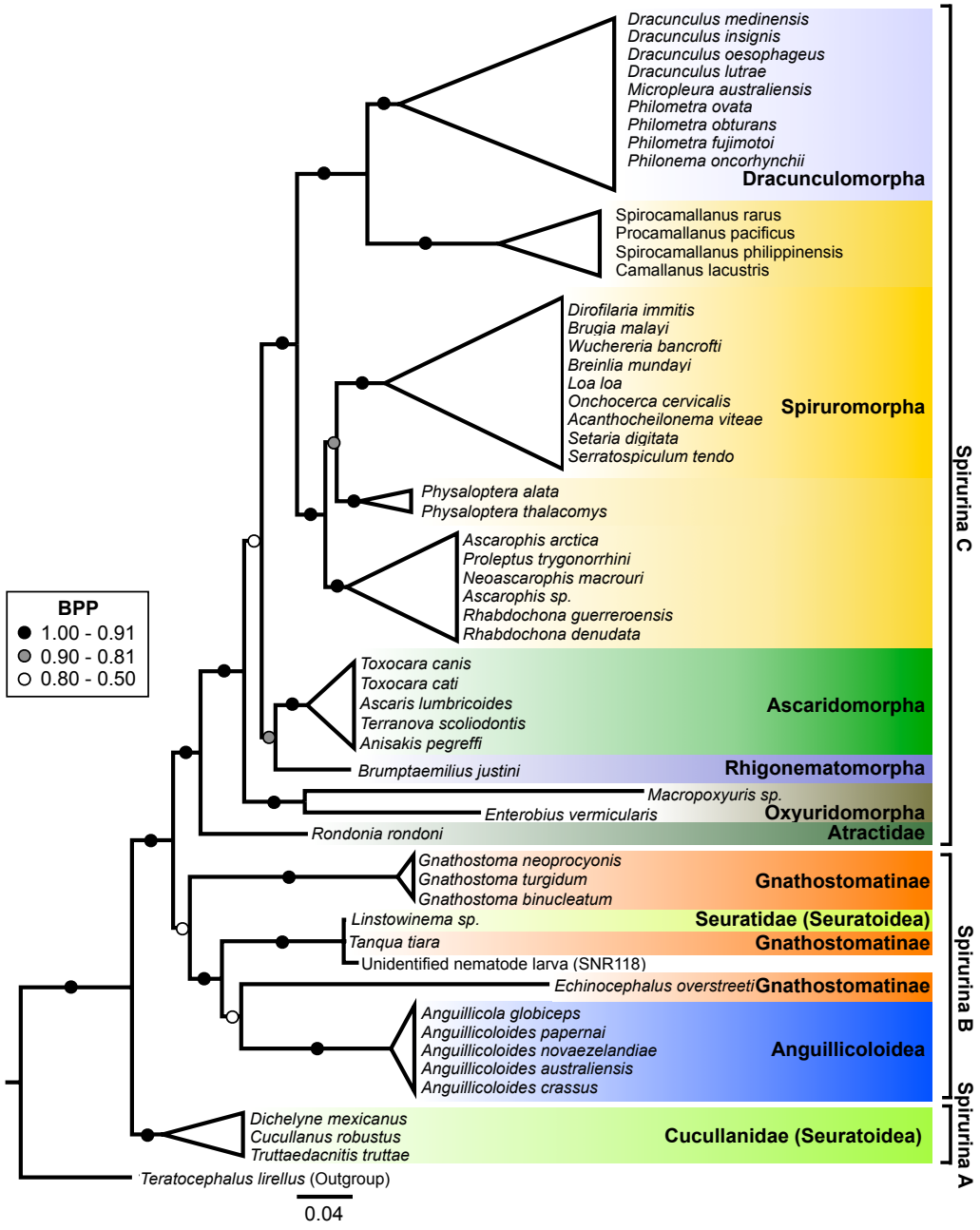

Figure 1.6: Phylogeny of nematode clade III based on nuclear small ribosomal subunit - Phylogram inferred from nuclear small ribosomal subunit for Spirurina using Bayesian Inference. Branches are collapsed to highlight major groups. Labels on internal branches indicate Bayesian posterior probabilities. From Laetsch et al. CITE!!

the risk of overlooking main effects of the genotype component blurred by genotype by environment (GxE) interactions. In other words: there are situations in which the differences in genotypes could be visible only under special environmental conditions.

This limitations of the common garden approach are addressed in transplant experiments. Representatives of each population are raised in the other population's natural environment. Explicitly including the environmental component this represents a tow-factorial design in which interactions between genotype and environment (GxE) can be incorporated into an analytical model.

In situations where host-parasite interactions should be studied the experimental design is complicated by one further genetic factor. When a common garden scenario is applied to different parasites infecting a hosts-species (or vice versa) such an experiment can be best described as "inoculation experiment under common garden conditions". Often only one of the interacting species can be regarded as the focal species. In the presented *A. crassus*:*Anguilla* project it is the parasite, as definitive genetic differences between the host-species are not in the focus. However using only one host-species the experiment would be equivalent to the analysis of the focal genotype, missing GxG interactions. This is addressed by a "reciprocal cross-inoculation experiment under common garden conditions" (75). The infection of both host-species with both parasite populations allows the incorporation of genotype by genotype (GxG) effects into an analytical model. This approach is chosen in the experiments presented in this thesis.

In a recent study also using this method (and inspiring the experimental design for my project) both European and Japanese eels were infected under laboratory conditions with worms from three geographic origins: Southern Germany, Poland and Taiwan.

In these experiments differences between the two European populations and and the Taiwanese population of worms manifested. These differences were especially (but not solely) visible in the early stages of the life-cycle:

In the European eel the number of L3 larvae from the Taiwanese population of worms was higher than from European worms. From the Taiwanese population less L4 larvae were observed at 25 dpi and the levels of this larval stage were stable during the infection, in contrast the numbers of L4 for the European populations decreased with time. Additionally up to 50 dpi there were less living adults observed for worm from the Taiwanese population and fewer dead adult worms were recorded for the Taiwanese population beginning from 50 dpi.

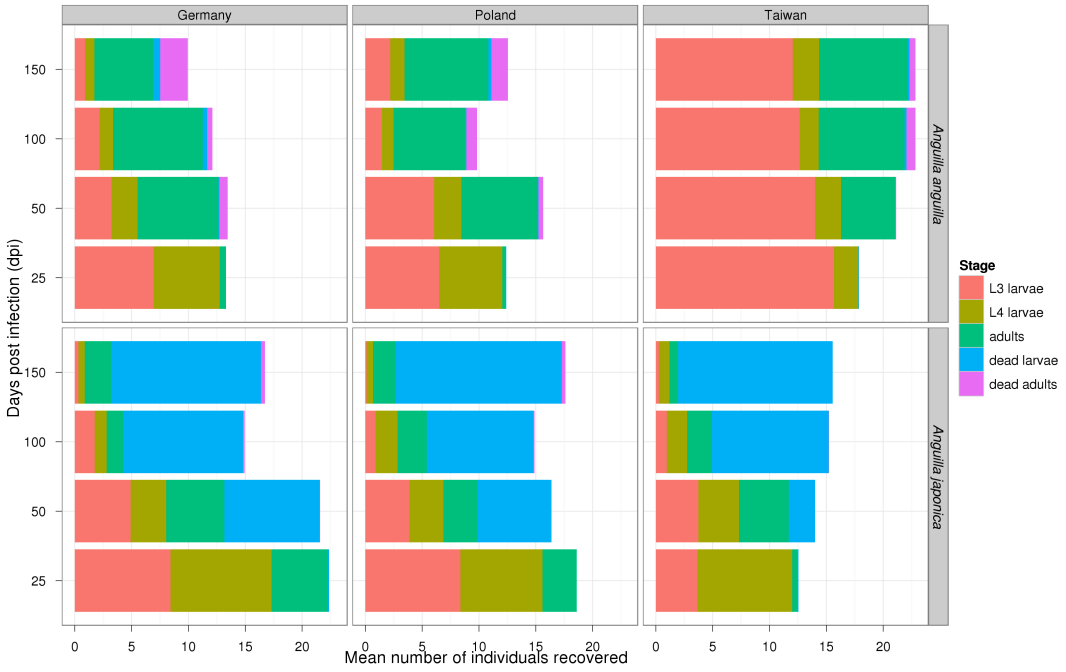

Figure 1.7: Differences in developmental speed - Three populations of *A. crassus* (panels in columns) were raised in two different hosts (panels in rows). Eels were dissected at 4 different time points post infection (dpi). Bars represent means of recovered individuals from three different life-cycle stages indicated by color. Differences between parasite-populations are pointed out in the main text. Data courtesy of Urszula Weclawski.

In the Japanese eel fewer L3 larvae at 25dpi were observed from the Taiwanese population compared to the European population of worms. Additionally more L4 larvae at this point in time and fewer living adults at 25 and 150 dpi, as well as fewer adults beginning from 50 dpi from worms of Taiwanese origin could be recovered compared to worms of European origin (Weclawski et al. unpublished; see figure 1.7).

These findings can be consolidated to the interpretation that an increase in the speed of development was observed in the European populations of *A. crassus* compared to the Taiwanese source population.

Measurements at different time-points are not easy to integrate into a more general interpretation of observed recovery of worms as fitness-components. Such fitness-components are usually thought to be a approximation to fitness (with life-time reproductive success as one of the closest approximations). Life history traits generally possess lower heritability and are under stronger selection (76). The inferred faster development of the European population of *A. crassus* can thus be regarded highly interesting as candidate-phenotype for adaptation. However the slightly delayed development of the Taiwanese population even in the natural host *An. japonica* would constitute an maladaptation (77) in one possible interpretation of these results.

The differences however are small in *An. japonica* and could possibly have a second explanation: GxG interactions could be hidden in *An. japonica* by GxGxE interactions. Such triple interactions could lead to superior fitness-components of the natural host-parasite genotype combination e.g. only at elevated water temperature or under other (even additional biotic) environmental conditions. An optimal experimental approach would thus be able to disentangle even GxGxE interactions and a design would be advantageous as it would explicitly include potential heterogeneity in the environment shaping GxGxE interaction as predicted theory of the geographic mosaic of coevolution (78). Such an experimental design a "reciprocal cross-inoculation under reciprocal transplant conditions" (79) is however impossible to implement in a mobile host-parasite system threatening biosafety as artificial secondary introductions are required for a transplant.

Nevertheless, the present experimental results provide a solid foundation for further research. They demonstrate divergence of the European population of *A. crassus*. Furthermore the loss of genetic diversity in the European population (10) seems to not have led to a decrease of fitness.

Interpretation of morphological characters in these studies proved difficult: Size of the worms seems to be mainly determined by the uptake of host-blood and is thus largely object to phenotypic modification, with a genetic component hard to detect. The approach taken in the study underlying this thesis builds on the above design but uses gene-expression levels as the phenotypic entity studied. This approach is enabled by recent advances in DNA-sequencing technology.

1.2 DNA sequencing

1.2.1 Two out of three: DNA sequencing and the central dogma of molecular biology

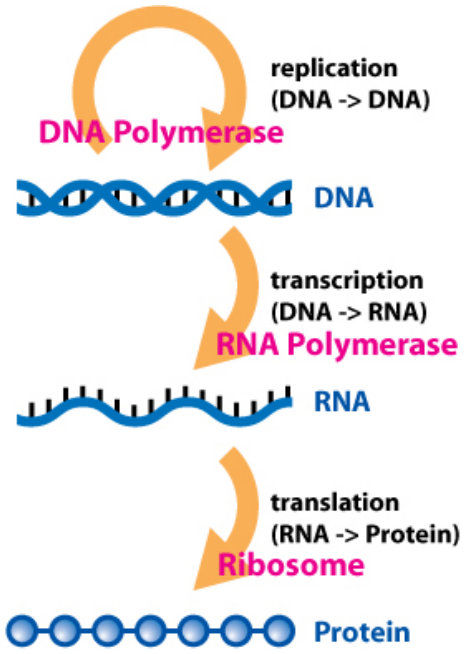

Figure 1.8: Major macromolecules bearing biological sequence information: - A schematic view on the flow of genetic information in a cellular life: Enzymes (red font) process macromolecules carrying genetic information from DNA to RNA, from RNA to protein. Picture from wikipedia.

Two kinds of macromolecules carry all the information evolution shaped over the course of the last 3.5 billion years from generation to generation: DNA and only in some viruses RNA. Proteins as the building blocks and functional molecules of life are transient manifestation of this information (80). In all cellular life genetic information flows from the replicating DNA to RNA in a process called transcription and from RNA to Protein in a process called translation (81) (see figure 1.8).

The relatively inert DNA is adapted to carry information over generations and to limit the number of mutation (also by evolving low error in polymerase) (82). The single stranded, more reactive RNA on the other hand can create secondary structures by base-pairing with itself or other macromolecules and is involved in numerous

cellular processes making use of this reactivity (83): microRNAs (miRNAs) regulate translation by binding mRNA, initiate degradation and thus decrease it's levels (84, 85), small nuclear RNAs (snRNAs) are (among other functions) part of the spliceosome (see below), small nucleolar RNAs (snoRNAs) direct a machinery to perform site-specific rRNA modification (86). In addition a variety of poorly understood other non-protein coding RNA (ncRNA) families exist (87). Together with proteins ribosomal RNAs (rRNAs) are building blocks of the ribosome, where translation takes place. Transfer RNAs (tRNAs) carry amino acids to the ribosome specific to their anti-codon sequence. There, at the ribosome, amino acids are incorporated into the polypeptide chain according to codon recognized in coding sequence (CDS) of a messenger RNA (mRNA) molecule and a protein is synthesized (88).

These mRNAs (like the untranslated RNAs above) have been transcribed from genomic DNA (see figure 1.9). Eukaryotic mRNAs have a special structure to prevent them from and regulate degradation and to allow interaction with non-coding RNA and with the ribosome during translation: The 5' CAP-structure and the 3' poly-A tail are added directly during transcription.

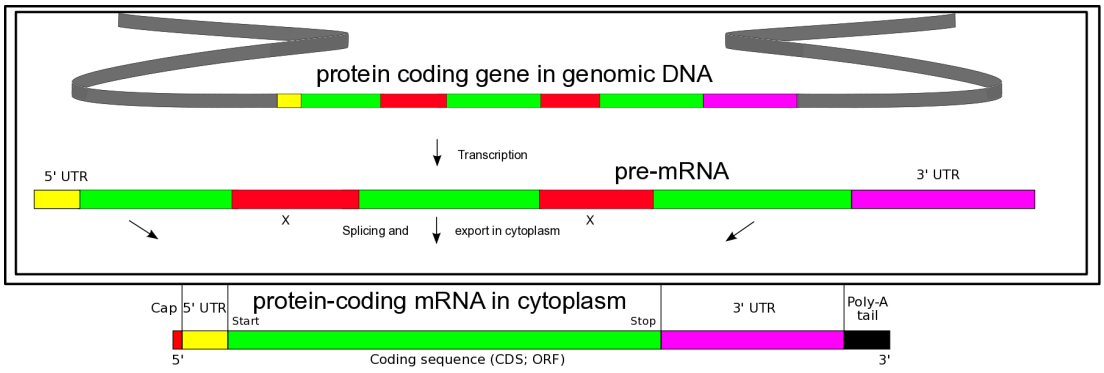

Figure 1.9: The structure of a protein coding gene and it's mRNA - A schematic view of posttranscriptional modifications in an eukaryotic gene. Introns are spliced, 5' and 3' structures are added and the mRNA molecule is exported into the cytoplasm. Note that the double stranded nature of the genomic DNA (gray) is not indicated in this comic and no indication of the enzymes unwinding genomic DNA for transcription is given.

Other post- or co-transcriptional modifications often include the excision of introns, non-coding regions found in genomic DNA. This excision is is directed by the spliceosome

containing snRNAs and proteins. In this splicing step alternative exons can be joined, skipped or even introns can be retained, increasing transcriptome and proteome diversification (89). Only after the processing of pre-mRNA to mature mRNA, the molecule is released into the cytoplasm where it eventually can be translated (see above).

The complete set of transcripts in a cell is called the transcriptome. The major goal of transcriptomics (the analysis of the transcriptome) is to asses quantity of transcripts for a specific treatment, genetic background, developmental stage or physiological condition. Intermediate goals in this process are the categorization of transcript into one of the diverse families above (mRNAs or ncRNAs and small RNAs) and the determination of the transcriptional and translational structure of genes (i.e. finding their start sites for both transcription from the genome and for translation into protein, 5' and 3' ends, splicing patterns and other post-transcriptional modifications) (90).

Transcriptome-projects and transcriptomic data have been invaluable to determine the structure of the genome (information gained from the transcriptme informs about genomic features) but they are also in the center of one of the major challenges in biology linking genotypes to phenotypes. The "expression" of the gene in an literal sense would be the phenotype visible for natural selection. It is known that posttranslational modification, the degradation and turnover of both mRNA and proteins have a strong influence on this gene-expression, and in this sense the global measurement protein expression (proteomics) would be one step closer towards a phenotype. Indeed increasingly proteomic information is used to complement genomics and transcriptomics (91). However overall levels of mRNA abundance correlate well with protein abundance (92). Measurements of protein levels are methodically more demanding than measurements of mRNA levels (see 1.2.2) and thus all estimates of gene-expression in this thesis are based on measurements of RNA-abundance and the term gene-expression is even used as a synonym for RNA abundance. All mentions of protein sequence in the results of this document are derived from computational prediction based on nucleotide sequence of mRNA.

All sequencing technologies for nucleic acid outlined below have in common, that they work on DNA not on RNA. Therefore transcriptome sequencing involves a step in which mRNA is reverse transcribed into complementary DNA (cDNA). The RNA-dependent DNA-polymerase (reverse transcriptase) used for this process is originally found in retroviruses.

1.2.2 The history and methods of high-throughput DNA-sequencing

For almost three decades the method developed by Frederick Sanger (93) was the only practical choice for determining the sequence of nucleic acid. Starting from denatured DNA, the method uses four different dideoxynucleotides (ddATP, ddCTP, ddGTP, ddTTPs) to terminate synthesis throughout the reaction (along the whole molecule) at the respective incorporation sites. The method first used radioactive labels attached to primers in four separate reactions for each of the ddNTP. The length of the partial DNA-sequences then had to be determined on a single-base resolution agarose gel. Later fluorescent labeling of ddNTPs allowed all four reactions to be performed together. Additionally modern machines use the chain-termination method combined with capillary gel electrophoresis (94) in a highly parallelized way.

Due to these advancements it was possible to tackle sequencing of bigger genomes, after phages in the first years of DNA sequencing (95): The bacterium *Escherichia coli* in 1997 (96), the baker's yeast *Saccharomyces cerevisiae* in 1996 (97), the nematode *Caenorhabditis elegans* in 1998 (98), the fruit fly *Drosophila melanogaster* in 2000 (99) and the mouse *Mus musculus* in 2002 (100) were the first organisms with sequenced genomes. For these laboratory model-organisms multi-national consortia financed and coordinated sequencing in multi-million dollar projects. This "first generation of genomics" culminated in the publication of the human genome in 2001 (101).

In parallel to the mentioned genome-projects transcriptome projects were conducted. Mapping ESTs to the genome identified coding regions in genomic sequences (102). First estimate of the number of genes in the human genome for example are based on extrapolation of the number of genes found in the early sequenced regions (103).

Costs and labor constrained genome-sequencing to the well established laboratory-model organisms mentioned above. In addition to the sequencing reaction itself, it was the need for cloning into DNA vectors for separation and amplification of DNA-fragments, that made costs and labor associated with this method prohibitive for a large scale application in non-model organisms.

1.2.3 DNA-sequencing in Nematodes

In 1998 *Caenorhabditis elegans* had become the first multicellular organism with a sequenced genome (98). Soon it was noted, that in addition to it's use as a general model

system for the metazoa and beyond, knowledge gained in this species has the potential to be even more valuable in the phylum nematoda (104). The breadth and detail of genomic information available for *C. elegans* to date is illustrated by a recent publication (105), useing transcriptomics to provide detailed annotation of the diverse functional genomic elements and their interactions at single base resolution. With this amount of data digested into usable information *C. elegans* continues be an invaluable resource in nematode genomics: With its 21,000 protein coding genes, over 5,000 RNA genes and 100.2 megabases (Mb) genome-size it still provides the rough expectations when new genome projects are started.

The genome sequence of *Caenorhabditis elegans* was soon complemented by the genome of *Caenorhabditis briggsae* (106), a second nematode from the genus *Caenorhabditis* sequenced a satellite system for comparative genomics inside this genus. As a second satellite model in clade V the necromenic *Pristionchus pacificus* (living in close association with beetles) has a published draft genome (107).

The first published genome of a parasitic nematode in the Spirurina was the draft genome of *Brugia malayi* (108) and as a second genome in the Spirurina recently the genome of *Ascaris suum* has been published (109).

Also in the remaining clades of the nematoda genome sequencing flourished: For the animal-parasite *Trichinella spiralis* from clade I (110), the plant parasites *Meloidogyne incognita* (111) and *Meloidogyne hapla* (112) as well as the the pinewood nematode *Bursaphelenchus xylophilus* (113) (a plant parasite using a beetle as an vector) from clade IV recently genome sequences have have been analyzed and published.

The current revolution in sequencing methodology (see 1.2.4) brings into sight many more sequenced nematode genomes (including that of *A. crassus*). The 959 nematode genomes initiative promotes such sequencing of nematode genomes and makes working-drafts of genome-assemblies available for analytical purposes on a `blast`-server (114).

Before the advent of NGS the lack of genomic information in many species of nematodes promoted the use of ESTs as a tool for gene-discovery. Partial genomes *sensu* (115) were successfully interrogated for a large array of genes interesting for various scientific communities. In nematode parasites of vertebrates, pathogenic factors were described as potential vaccine candidates (116). Change in expression of these molecules constitutes an *a priori* hypothesis to be tested for different populations and host-environments in *A. crassus*:

Cystein-proteinase inhibitors (cystatins) and serin proteinase inhibitors (serpins) are thought to interact with the antigen presentation in vertebrate hosts (116). Homologues of mammalian cytokines were identified, which are believed to interact with mammalian cytokine receptors to divert the immune response to a TH2-type response (117) (an anti-inflammatory, cellular response, thought to be non-effective against helmiths). Further molecules involved in host-parasite interaction identified in transcriptome-projects include abundant larval transcripts of *B. malayi* (Bm-ALT) (118) and venom like allergens (Bm-VLA) (119).

In some of these studies secreted proteins were in the center of interest. They could potentially be excreted by the nematode to allow movement and food-uptake but also to interact with the host's immune system. The detection of signal-peptides for secretion using *in silico* analysis of ESTs has been used to highlight candidate genes for example in *Nippostrongylus brasiliensis* (120), and across all nematode ESTs (121).

Over the years sequence information derived Sanger-sequencing derived EST-data and whole genome sequencing has been collected and updated into the nembase transcriptome databases (122, 123). The recent compendium nembase4 describes clustering of 679,480 raw ESTs in 233,295 clusters from 62 species (124). This database provides an invaluable source collection the abouve infromation for comparison, validation and hypothesis generation when new transcriptomes are analysed as in the present project.

Obviously NGS also leaves it's marks currently in nematode transcriptomics: NGS analysis on the transcriptomes of *Ancylostoma caninum* (125), *Pristionchus pacificus* (126), *Litomosoides sigmodontis* (127) and *Ascaris suum* (128) have been published and a recent review (129) lists 8 further datasets for other species already available in public repositories. Additionally for *Haemonchus contortus* pyrsequencing-transcriptome has been published (130) unnoticed by the above review, illustrating the explosive expansion of data and publications.

1.2.4 Advances in sequencing technology

Advances in sequencing technology (often termed "Next Generation Sequencing"; NGS), provide the opportunity for rapid and cost-effective generation of genome-scale DNA-sequence data. Labor and costs associated with DNA-sequences were drastically reduced during the last 5 years.

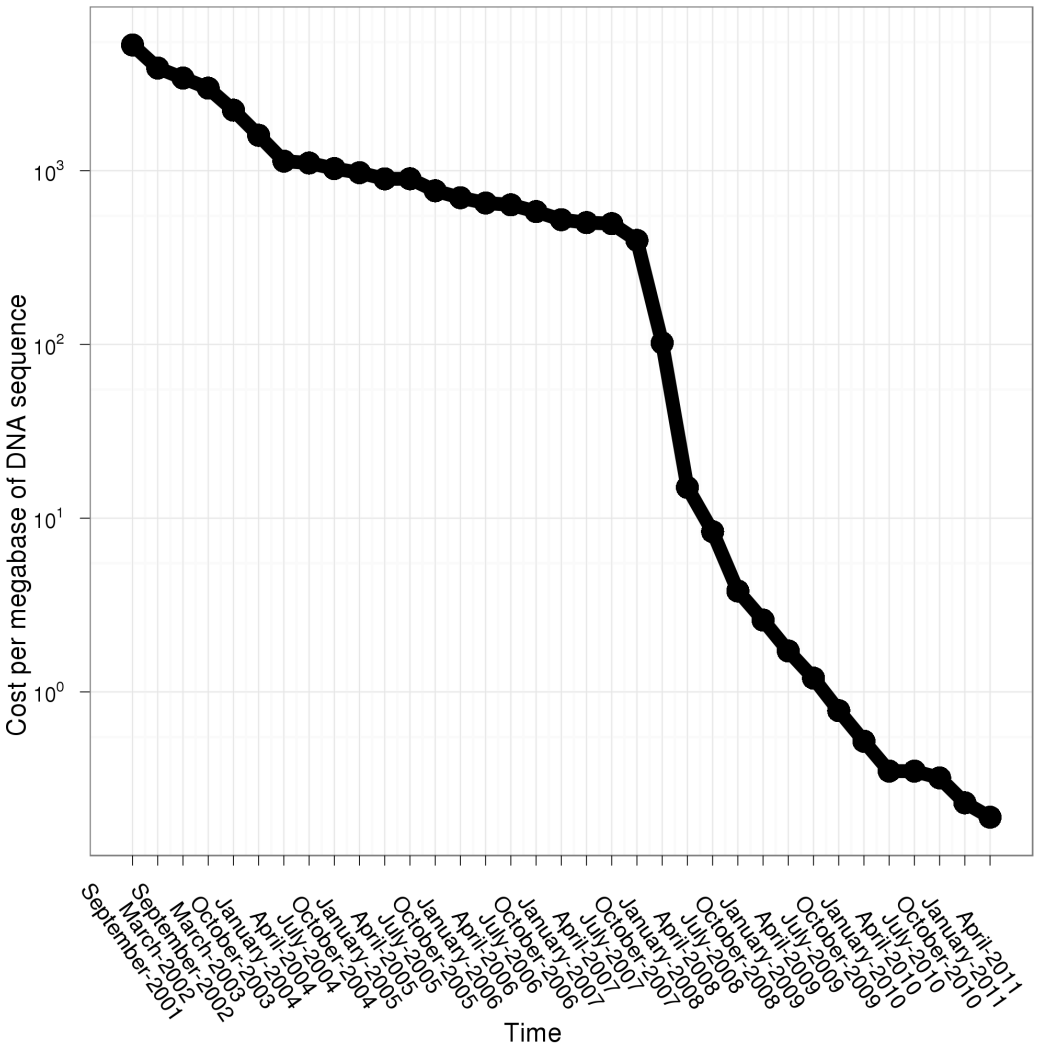

Figure 1.10: Falling sequencing costs - Sequencing costs falling due to advances in Solexa-sequencing: Due to improved read-length and data-volume on this platform per base sequencing-prices for many applications tumble into free fall. Data provided by National Human Genome Research Institute, NHGRI.

The technologies portrayed here and used in the work underlying this thesis can't work on single molecules and thus target molecules have to be amplified like in Sanger-sequencing. This amplification has to produce spatially separated templates. Immobilization on a solid surface to archive this clonal amplification is used in preparation of both pyrosequencing and for the illumina-platform (131). The detailed implementation of this solid-state amplification in each technology differs and will be explained in the corresponding sub-chapter.

One cumbersome aspect of the need for amplification is the high amount of DNA starting-material ($3 - 20\mu g$) required (131). Other disadvantages include, that mutations during clonal amplification in templates can disguise error as sequence variants. Nucleotide composition of the target may also introduce amplification bias and thus biased product yield (132). This in turn leads to underrepresentation of certain molecules. The last point can be detrimental in quantitative applications, such as RNA-seq (90). However, while alternative single molecule approaches exist (eg. (133, 134)) and can be applied to address the above stated the problems (135, 136), to date these technologies are in throughput and reliability not competitive for most real life applications.

1.2.4.1 Pyro-sequencing

Prior to pyrosequencing (or 454-sequencing; named by the company making it commercially available) an emulsion PCR is used to clonally amplify DNA molecules attached to beads (figure 1.11): After fragmentation by mechanical shearing or ultrasound (138) (see figure 1.11), DNA is ligated to adapters, denatured and single stranded molecules are attached to complementary sequence on a bead. Emulsion of beads in oil together with enzymes under conditions that favor one bead per water/enzyme droplet allows PCR in micro-scale reactions. This covers each bead with multiple copies of one target molecule. The beads are then distributed over the wells of a fiber-optic slide, the so called picolitre plate. A single bead per well is covered with enzymes on the surface of smaller beads. These enzymes are used in the actual pyrosequencing reaction originally developed by Pål Nyrén in the 1990s (139). The release of inorganic PPi as a result of nucleotide incorporation by polymerase starts a cascade of enzymatic reactions. The released PPi is converted to ATP by ATP sulfurylase, providing energy for luciferase to oxidize luciferin and to generate light. The added nucleotide is known as nucleotides are flushed over the plate one at a time. A high resolution camera records the emission

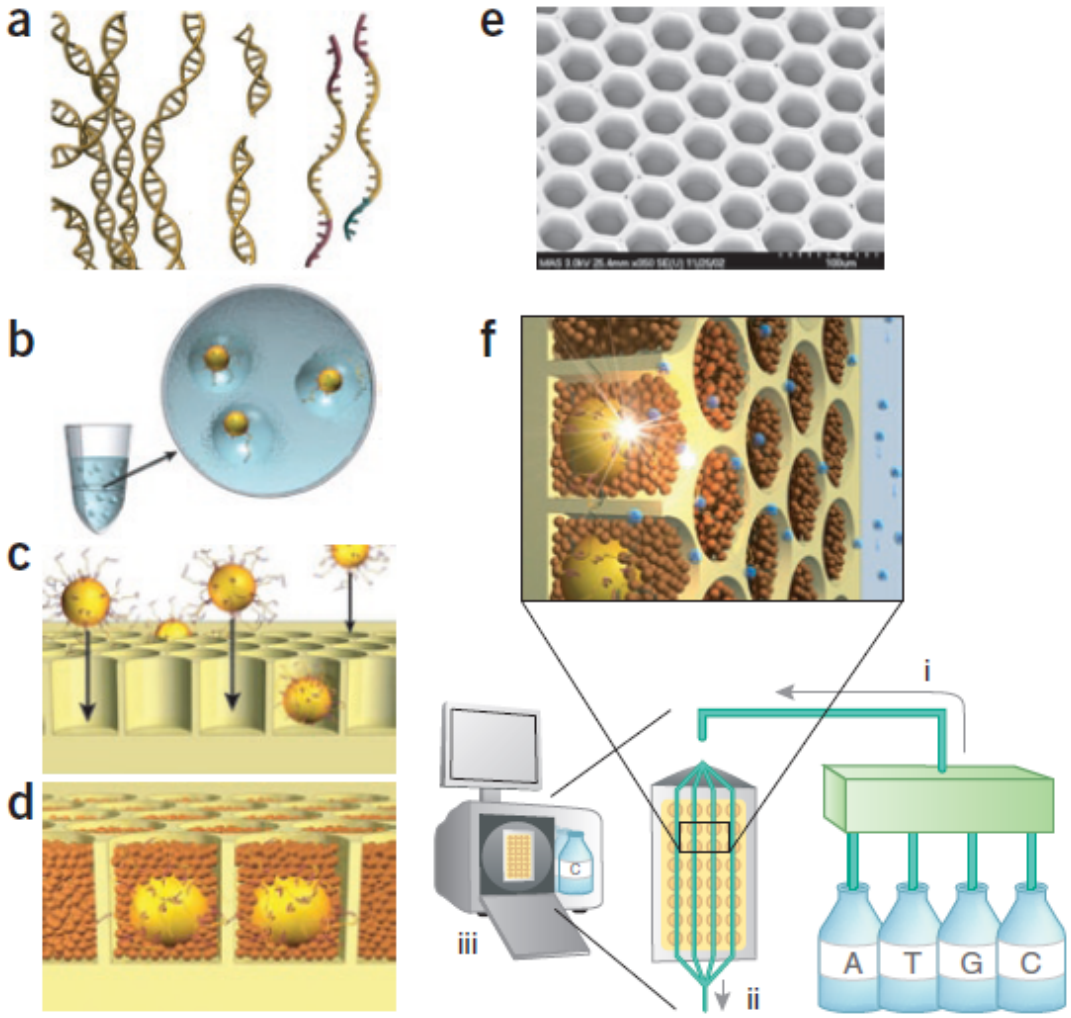

Figure 1.11: Schematic representation of pyrosequencing - (a) DNA (genomic or transcriptomic) is isolated, fragmented, ligated to adapters and denatured into single strands (b) Under conditions that favor one fragment per bead fragments are bound to beads. These beads are isolated and compartmentalized in the droplets of an emulsion and PCR (a mixture of reagents in oil). Within each droplet DNA is amplified, and beads are obtained which carrying millions of copies of a unique DNA template. (c) After denaturation of DNA, beads are deposited into wells of a fiber-optic slide (called picolitre plate). (d) Immobilized enzymes carried on smaller beads are added to each well and a solid phase pyrophosphate sequencing reaction is initiated. (e) A portion of a fiber-optic slide, in a scanning electron micrograph (prior to bead deposition) (f) Major subsystems of the 454 sequencing instrument: a fluidic assembly holding nucleotides separately (object i), the well-containing picolitre-plate in a flow cell (object ii), a CCD camera assembly and the user interface for instrument control (object iii) (137)

of light. The intensity of emitted light is proportional to the number of nucleotides incorporated.

The ability to distinguish length of homopolymeric runs of the same nucleotide decreases with the length of such homopolymer runs (140). Current "Titanium chemistry" is producing read of > 350 bases length, "FLX chemistry" (used up to 2009) was able to produce reads of roughly 250 bases length (141).

This longer read length of 454-sequencing (142) compared to other NGS technologies (see 1.2.4.2), allows *de novo* assembly of Expressed Sequence Tags (ESTs) in organisms lacking previous genomic or transcriptomic data (for a comprehensive list of studies using this approach before Oct 2010 see (127)).

1.2.4.2 Illumina-Solexa sequencing

Solexa illumina technology is to date (Dec. 2011) the most competitive commercial sequencing platforms enabling a broad spectrum of applications.

The Illumina-Solexa platform uses bridge-amplification to produce clonal copies of DNA molecules in clusters on a glass slide (figure 1.12): Fragmented, double-stranded DNA is therefore ligated to a pair of oligonucleotide-adapters in a forked configuration (the adapter-ends have non-complementary sequence). Two primers are used in an initial amplification and a double-stranded molecule with a different adapter on either end is produced. Denatured single-strands are then annealed to complementary adapters on the surface of a glass slide. Using the 3' end of the surface-bound oligonucleotide as a primer, a new strand is synthesized. Subsequently the adapter sequence at the 3' end of newly synthesized copied strand is bound to another surface-bound complementary oligonucleotide. This results in a bridge-structure and generation of a new priming-site for synthesis after denaturation. Multiple cycles of this kind of solid-state PCR result in growth of clusters on the surface of the glass-slide (143).

In the actual sequencing reaction these clusters are sequenced using a sequencing by synthesis technique: polymerase and all four nucleotides simultaneously are flushed over the class slide in successive cycles. To avoid incorporation of multiple nucleotides, "removable terminator"-nucleotides are used, which allow only incorporation of one nucleotide per strand pre cycle. These nucleotides are labeled each with a different removable fluorophore. Transient incorporation of a nucleotide is detected using a high

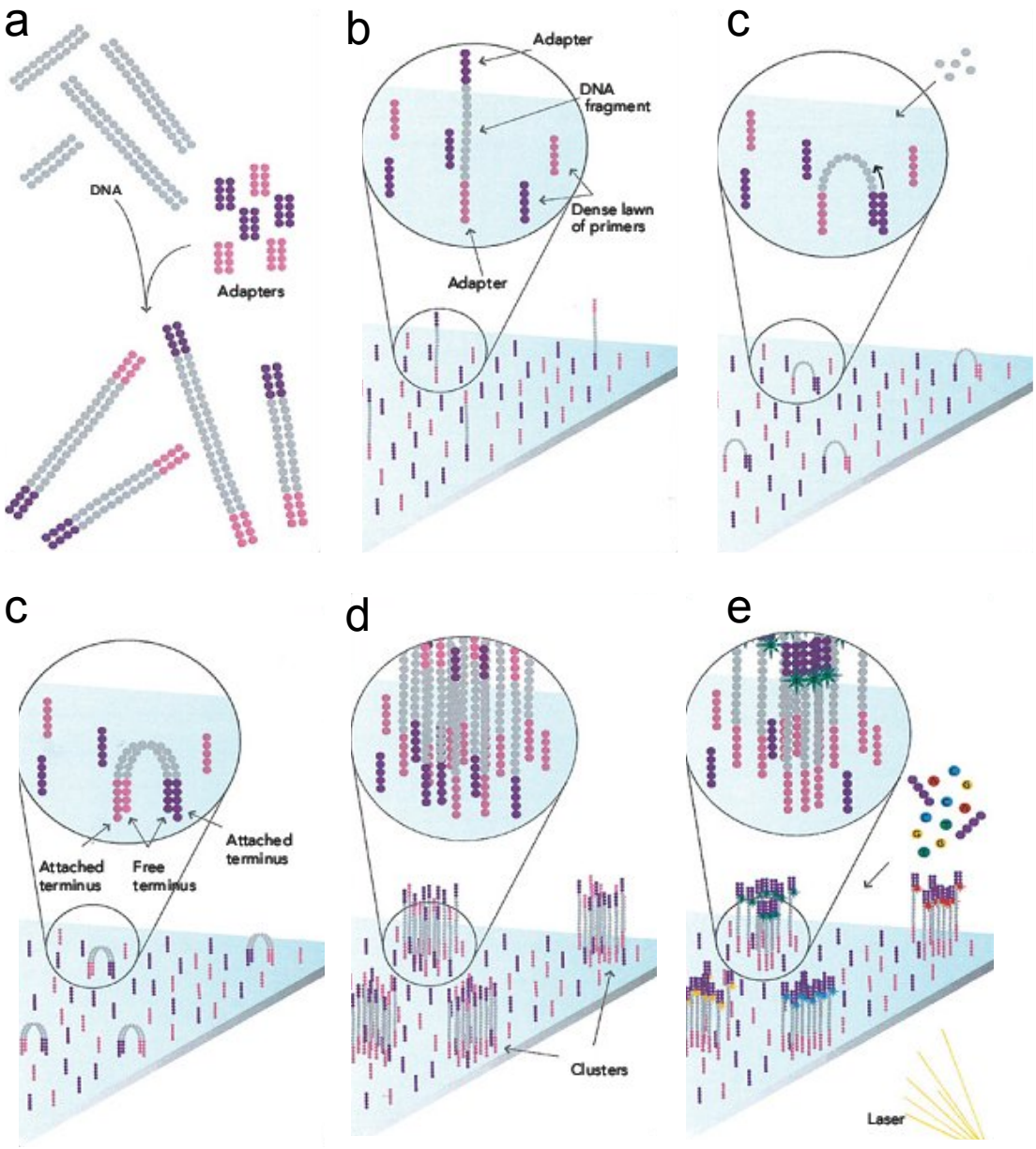

Figure 1.12: Schematic representation of illumina sequencing - (a) DNA (genomic or transcriptomic) is isolated, fragmented and ligated to adapters. (b) Single stranded fragments are bound to a glass-slide. (c-d) Solid-phase bridge amplification using unlabeled nucleotides, primers (binding the adapters) and polymerase leaves clusters of double stranded DNA distributed over the slide. (e) four labeled reversible terminators, primers (binding the adapters) and polymerase are added. Laser excitation an image of the emitted fluorescence is taken . Step (e) is repeated multiple times (=length of sequence). Modified from Seqanswers-forum

resolution camera after laser-induced excitation. The fluorophore is removed and next cycle initiated (143).

This leads to an error model different from 454 sequencing: Runs of homopolymeric sequence are not problematic, but due to the decreasing propensity of terminators for removal, sequencing quality decreases in from 5' to 3' direction.

An slight alternation of the above method, which is extremely useful to inform assembly, is paired-end sequencing: After the first sequencing (as above), the original template strand is used to regenerate the complementary strand. This complementary strand then acts as a template for the second sequencing producing complementary sequence from the other end of the molecule. Using template molecules of a certain size range, sequence information can be obtained spanning 200-500 bases (the possible span of a nucleotide bridge in bridge-amplification) (143).

Additionally recent increases in read length (from 35 bases in 2008 to over 100 bases in 2011) are beginning to allow *de novo* sequencing and assembly of large eukaryotic genomes (e.g. that of the giant panda (144)) and transcriptomes (145) (but see also 1.2.5 for methodical challenges). In the same period throughput also increased from 6,000,000 reads in 2008 to 20.000.000 reads in 2011 per lane of the instrument.

The high throughput of the Illumina-Solexa platform makes it also first choice for gene expression analysis (146): RNA-seq has revolutionized transcriptomics both in model and non-model organisms (90). SuperSAGE (147) using expression-tags provides the benefit of classical SAGE-analysis (148) with those of the ultra heigh throughput of Illumina-Solexa sequencing.

Although the sequencing reaction itself differs between platforms, the technologies described as above have in common that up to date they produce much more, but shorter reads than classical Sanger-sequencing.

This fostered use and development of new methods to assemble large-scale shotgun sequences, as higher coverage but shorter read-length (and also lower accuracy) are increasing the computational complexity of the assembly-problem (reviewed in (149)).

1.2.5 Computational methods in DNA-sequence analysis

In the context of computational tools another common characteristic of all DNA-sequencing methods has to be emphasized: Read-length is usually shorter than the length of the target molecule to be sequenced. This potential problem is solved by

oversampling the target molecule, producing overlapping sequence. The amount of redundancy of the overlap is termed coverage (e.g. 10-fold coverage means a base is sequenced 10 times redundantly) the method as such is referred to as shotgun-sequencing and has - shortly after sequencing chemistry - been described by Sanger (150). Soon computer programs were necessary to align sequences, to compute overlaps and consensus sequences (151) and the process of computationally reconstructing the target molecule was termed sequence-assembly (152). This reconstructed target molecules are termed contigs, derived from contiguous sequence. In an (hardly achieved) optimal genome-assembly a conigs would thus represent a chromosome, in an optimal transcriptome assembly there would be a contig for every transcript in the cell.

The first step in the overlap-consensus approach is to detect overlapping sequence in a series of pairwise alignments. Two classical approaches exist, the first being local "Smith-Waterman" alignment (153) the second "Needleman-Wunsch" global alignment (154). Of course these alignment methods have usages outside of sequence assembly in general sequence comparison, including protein sequence.

The program `Blast` (155), for example, enables large scale comparison of sequences against databases. It is based on a heuristic approximation of Smith-Waterman alignments: After a seeding step, in which small regions of similarity (protein) or perfect matches (nucleotide) are found, it uses local-alignments to extend regions of similarity and to form high-scoring segment pairs (HSPs). Using a sophisticated statistical procedure it reports two measurements used to asses the significance of matches: The e-value reports the number of hits as good or better than the present hit expected against the current database by chance. It is usually used to order hits from a search. The bit-score in contrast is normalized with respect to the scoring system and database and can thus be used to compare hits from different searches.

With the advent of next generation sequencing (see 1.2.4) even the heuristic approach of `Blast` or it's mapping equivalent `Blat` (156) was not ideally suited for the massive amounts of data. New kinds of alignment methods were needed to handle data volume, error structure and short read-length. Mapping describes a subset of the assembly problem and mapping programs confine themselfs to this sub-problem. In mapping only the positions (and the quality) of a match relative to an already sequenced longer contig are interrogated. `Ssaha2` (157) is able to speed up such sequence searches by orders of magnitude. It builds a hash table indexing k-tuples (k contiguous bases,

an approch implicitly also used in the seeding step of `Blast/Blat`). Then sorting of matching indices gives regions of high similarity without an alignment, but these region can then be aligned using a banded Smith-Waterman algorithm. `Burrows-Wheeler Aligner` (`BWA`) (158) builds a suffix array holding the starting positions of suffixes of a lexicographically ordered string. Then exact as well as inexact matches can be found and a gapped alignment can be generated.

For *de novo* assembly of genomes new algorithmic approaches involve consturction of a de Bruijn-graph. In most formulations of this new approach instead of nodes in the graph (sequences) edges (overlaps) are traversed. This way problematic repeats are joined and sub-sequences reused. The method uses a splitting of sequences in k-mers of defined length (edges in the de Bruijn-graph) and is thus optimal for very short reads (159).

On top of this complexity found in *de novo* assembly of genomes, transcriptome assembly has to deal with additional challenges resulting from the biology of the transcriptome (see 1.2.1): (a) The depths of reads obtained from cDNA for different transcripts differs dramatically, additionally target molecules may be covered uneven across their length. (b) In highly expressed transcripts more erronous bases are found in total. (c) Transcripts from adjacent loci can overlap and can be erronously fused to form chimeric transcripts. (d) Multiple real transcripts can exist per genomic locus, due to alternative splicing. (e) Additionally sequences that are repeated in different genes (domains) introduce ambiguity (160).

Using pyrosequencing instead of the solexa-platform problems (a) and (b) are less pronounced because of the overall lower coverage. Problems (c) and (e) can be better resolved because of the longer read-length. For the same reason the power for the resolution of alternate splicing isoforms (d) is enhanced (at least for high-coverage transcripts). Recent versions of `gsAssembler` (also called `Newbler`; Roche/454) provide an opportunity to asses alternative splicing (161).

The project presented here takes the approach of first using pyrosequencing to define a reference transcriptome and then mapping reads from the solexa-platform to this reference.

But also downsteam of the sequence assembly translation of the highly complex, potentially biased, multidimesional data into biological relevant knowledge provides computational challenges.

Inferrence of single nucleotide polymorphism (SNPs) requires statistical categorization in true polymorphisms and sequencing errors. Tools like `VarScan` (162) or `VCFtools` (163) combine alignment depth, quality of the base call in each sequence, quality of mapping to the reference and the base composition in the region into a statistical framework. `GigaBayes` (164) uses additionally an *a priori* expected polymorphism rate. Less attention is usually paid to indels (insertions or deletions), genomic rearrangements, copy number polymorphisms caused by local duplication and other structural variations. While these are common types of variation between genomes, they can be harder to detect (165).

Assesment of statistical sigificance of differences in read counts (from transcirptomic data; also called "digital transciptomics"), needs some special treatment in comparison to the well established methods for microarray-data (166). While both kinds of data need normalisation relative to overall transcript abundance measured (fluorescence or counts), sequencing derived read counts follow a negative binomial distribution (167) instead of a normal distribution for microarray data. To make allow testing for low numbers of replicates sofware commonly uses global estimates of variance to restrain and partly replace individual variance. State of the art methods using these approaches are implemented in the R-packages `DESeq` (168), `edgeR` (169) and `baySeq` (170).

The functional interpretation of results (from e.g SNP-calling or digital transcriptomics) needs a standardized vocabulary in a datastucture across species and databases. Gene ontolgy (GO) provides such an vocabulary of controlled terms. The terms are organized in an directed, acyclic graph. This means, that a hierarchical stucture links lower level "child"-terms (more specific) to higher level "parent"-terms (less specific) through a standardized set of directional relations. Back-links forming circles are not allowed (171, 172). E.g. "endopeptidase activity" "is a" "peptidase activity", not the other way round. The "is a" in the previous sentence is such a directional realtion and other possible links would be e.g. "part of" or "regulates".

1.2.6 Applications of NGS in ecology and evolution and gene-expression divergence

NGS technologies are are increasingly used in studies on organisms with ecological and evolutionary significance. Such ecological and evolutionary model organisms often lack reference genomes to guide the assembly-process.

Today, both theoretical arguments as well as field and laboratory data suggest that evolution, including divergence of populations, can occur very rapidly given the right selective pressure. Such situations provide us with the opportunity of examining how divergence and even speciation work at the molecular genetic level (173).

In *Drosophila* variation of gene-expression within a single species can be attributed more to trans-regulatory elements, while expression divergent between species is dominated by cis-regulatory differences (174). Furthermore sterility of hybrids between species of this genus has been shown to result from incompatibilities in gene-regulatory networks (175).

A study on trout in Lake Superior (176) used an approach similar to the approach in the work presented here: Fish, which show two different phenotypes were raised in a common environment, demonstrating the genetic fixation of the phenotypic trait. 454 sequencing was then used to measure the gene expression levels and successfully identified 40 genes from two biochemical pathways being differently expressed. However, in addition to showing divergent evolution of gene-expression, this study highlighted the limitations of 454 sequencing for gene-expression analysis.

In the seagrass *Zostera marina* norther and southern populations display different patterns of resilience of expression patterns after a heat wave.

A study on two phylogenetically distant mangrove species found convergent evolution of gene expression. From the fact, that closer relatives of the studied species with different ecological niches do not show the same similarities the study condluded an adaptation to the similar environment (177).

CHANGE http://www.pnas.org/content/103/14/5425.full: Gene expression has been hypothesized to be of adaptive importance (1), and heritable variation that affects fitness is necessary for evolution by natural selection. Although adaptive differences in expression have been identified in single-gene studies (2âĂŞ6; see ref. 7 for review), microarray approaches offer great promise to rigorously address this hypothesis because they assay many loci at once. Furthermore, it is generally agreed that much of variation in gene expression for a particular environmental condition has a genetic basis (8, 9) according to studies in yeast (10âĂŞ13), Drosophila (14âĂŞ18), mice (19), and humans (20âĂŞ22). CHANGE

Positive or diversifying selection on parasite proteins from the host-parasite interface can lead to a overabundance of non-synonymous changes (altering the protein sequence)

over synonymous polymorphisms e.g. in *Plasmodium* (178).

An additional feature of parasite gene-expression is the theoretically deduced need to express only a single allele of a polymorphic parasite infection locus: In parasites gene-expression is thought to evolve towards avoidance of co-expression: For polymorphism to be positively selected it requires the evolution of a regulator locus or the evolution of polymorphism is followed by the evolution of a regulator locus is (179).

Two virulence factor LbGAP in venom-producing tissues that the major virulence factor in the wasp *Leptopilina boulardi* differs only quantitatively. The regulation of gene expression might thus be major mechanism at the origin of intraspecific variation of virulence (180).

2

Aims of the project

2.1 Final aim

In a reciprocal transplant experiment genetic components of the differences in

2.2 Preliminary aims

In order to investigate transcriptomic response to environmental stimuli, the responding unit, the transcripts have to be established first. As extremely short reads providing ultra high thoroughput are hard to assemble *de-novo*, a reference was created first using 454 pyrosequencing technology providing longer read-length.

3

Pilot sequencing (Sanger method)

3.0.1 Overview

This chapter reports a small pilot-project investigating the RNA-extraction and cDNA preparation inn preparation of high-throughput transcriptome sequencing of the swim-bladder nematode *A. crassus*. Expressed sequence tags (ESTs) were generated using traditional Sanger-technology. This sequencing was used as an first assesment of the sequence diversity expected in deeper sequencing especially the expected coverage unwanted rRNA and host-derived sequences was investigated.

In total 945 reads from adult *A. crassus* (5 libraries from 4 cDNA preparations, including 541 sequences generated by students in a laboratory course) and 288 reads from liver-tissue of the host species *An. japonica* (3 libraries from 3 cDNA preparations) were sequenced.

3.0.2 Initial quality screening

The initial quality screening of *A. crassus*-sequences revealed a high number of sequences that had to be discarded due to failed sequencing reactions (sequences beeing too short after quality trimming by `trace2seq`) in the library prepared by students. For sequences of *Anguilla japonica* and the other libraries from *A. crassus* failed sequencing reactions were less common.

In the next screening-step for *A. crassus* 125 (13.23%) and for *Anguilla japonica* 64 (22.22%) of the sequences were excluded because of homopolymer-runs considered artificial. This resulted in 452 of the nematode and 195 of the host reads regarded of sufficient quality for further processing after base-calling and quality screening.

3.0.3 rRNA screening

The further screening of sequences revealed a high abundance of rRNA (see Figure 3.1) ranging from 71.67% to 91.67% of obtained sequences. High abundances of rRNA were also found in the libraries from host liver tissue (see table 3.1), ranging from 71.67% to 77.42%. This contamination in libraries from both species was mainly responsible for a low ammount of sequences beeing of sufficient quality for submission to NCBI-dbEST. At this point for the *A. japonica*-dataset 36 sequences were submitted to NCBI-dbEST under the Library Name "*Anguilla japonica* liver" and were assigned the accession LIBEST_027503.

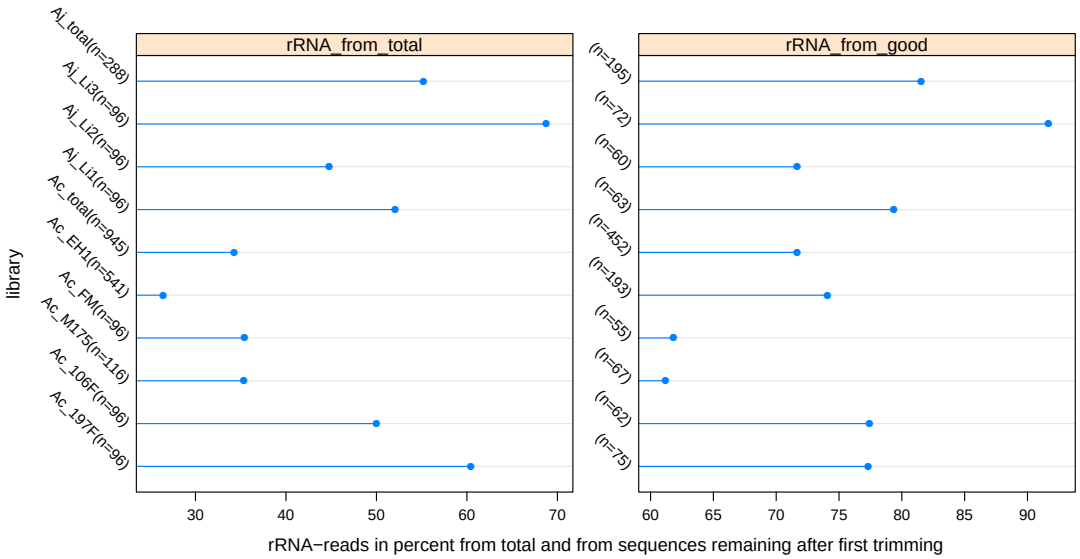

Figure 3.1: Proportion of rRNA in different libraries for *A. crassus* and *A. japonica* - rRNA abundance as proportion of the raw sequencing-reads (rRNA from total) and as proportion of the reads after quality screening (rRNA from good). Libraries starting with "Ac_" are from *A.crassus*, libraries starting with "Aj_" are from *A. japonica*.

3.0.4 Screening for host-contamination

For the *A. crassus*-dataset screening for host-sequences at this stage was regarded necessary based on the notion that a big proportion of the tissue prepared in RNA extraction consisted of eel-blood inside the gut of the worms (see also Figure 1.3). Additionally a

	short	poly	rRNA	fishpep	good
Ac_197F(n=96)	4	17	58	1	16
Ac_106F(n=96)	25	9	48	0	14
Ac_M175(n=116)	30	19	41	3	23
Ac_FM(n=96)	12	29	34	1	20
Ac_EH1(n=541)	297	51	143	8	42
Ac_total(n=945)	368	125	324	13	115
Aj_Li1(n=96)	10	23	50		13
Aj_Li2(n=96)	10	26	43		17
Aj_Li3(n=96)	9	15	66		6
Aj_total(n=288)	29	64	159		36

Table 3.1: Screening statistics for pilot sequencing - Number of ESTs discarded at each screening-step for single libraries and totals for species. Short, sequence to short in `trace2seq`; poly, sequences with artificial homopolymer-runs from poly-A tails; rRNA, with hits to rRNA databases; fishpep with better hits to host-protein-databases than to nematode protein databases; good, sequences regarded "valid" after all screening steps. Note that the 13 sequences in the *A. crassus*-dataset, for which fish-origin was inferred, were still submintted to NCBI-dbEST.

bimodal distribution of GC-content in the *A. crassus*-dataset was observed with one of the modes consistent with the mean GC-content of the ESTs from the Japanese eel.

Comparison of `Blast-` results for these sequences versus nempep4 and a fishprotein-database (derived from NCBI non-redundant), showed that 13 sequences were more likely to originate from host contamination than from *A. crassus*. These 13 sequences in the *A. crassus* data-set were submitted to NCBI-dbEST with a comment, that host origin had been inferred. This reduced the dataset esentially to 115 ESTs. However it has to be noted that these 13 ESTs are still accessible through the same library name "Adult *Anguillicola crassus*" and library-identifier LIBEST_027505 and are taxonomically attributed to *A. crassus* on NCBI-dbEST.

After screening of host-sequences the GC-content of *A. crassus* ESTs had a unimodal distribution (see Figure 3.2). *A. crassus* had a lower mean GC-content (37.32 ± 8.36 mean $\pm$ sd) than *Anguilla japonica* (45.79 ± 8.36 mean $\pm$ sd; two-sided t-test $p < 0.001$). The distribution of the GC-contents for sequences, for which host-origin was inferred was in agreement with the GC-distribution for host sequences.

`Blast`-annotations obtained (by similarity searches against NCBI-nr, bit-score thresh-

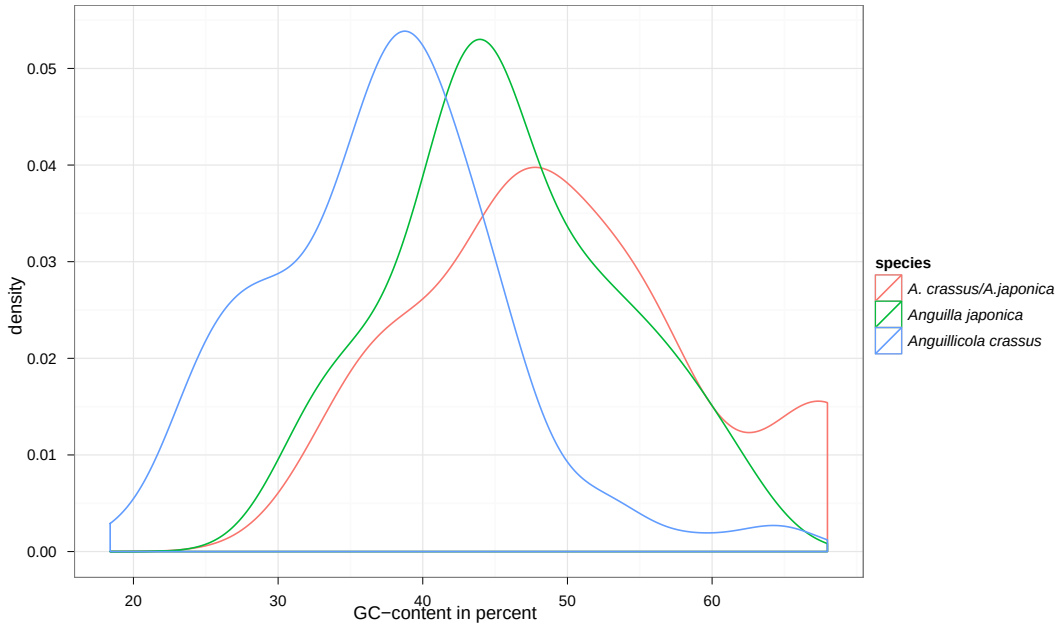

Figure 3.2: GC-content of sequences from *A. japonica* and *A. crassus* - The Japanese eel has a slighly higher GC-content than the parasite: This sequence characteristic is useful for seperation of sequences from the host-parasite interface, note the higer GC-content of the sequences from *A.crassus*, for which host origin was inferred from similarity searches (red line labeled *A.crassus/A.japonica*).

old of 55) for the sequences of putative host origin were also largely in agreement with the expectations for eel-blood: One sequence could be identified being higly similar to "Hemoglobin anodic subunit" from the European eel. Others were annotated with best hits to highly exressed houskeeping genes from fish or vertebrates (see table 3.2). Two sequences in the set had lower similarities only to proteins predicted from genome-sequences of Chordates, and one sequence of the 13 lacked any similarity to NCBI-nr above the treshold of 55 bits.

115 of the sumbitted sequences for "Adult *Anguillicola crassus*" (LIBEST_027505) were regarded "valid" i.e. not clearly host origin.

However it should be noted, that two ESTs (Ac_EH1f_01D10 and Ac_EH1r_01D10; forward and reverse read of the same clone) were annotated with "ref|ZP_05032178.1|; Exopolysaccharide synthesis, ExoD superfamily" from *Brevundimonas* sp. BAL3. The family Caulobacteraceae, comprises bacteria living in oligothroph freshwater and sequences are probably derived from a commensal, symbiont or pathogen of eels or swimbladder-nematodes. These off-target data was left in the submission file.

For 66 (58.4%) of the remaining 113 ESTs annotations were obtained from orthologous sequences. All these orthologous sequences were from other species in the phylum nematoda.

sequence	hit identifier	hit description	species	bit-score	e-value
Ac_EH1f_005B07	gb|AAQ97992.1|	cyclin G1	*Danio rerio*	67.0	9e-10
Ac_EH1f_01A02	gb|ACO10003.1|	Nicotinamide riboside kinase 2	*Osmerus mordax*	333	1e-89
Ac_EH1f_01C10	gb|ADF80517.1|	ferritin M subunit	*Sciaenops ocellatus*	328	5e-88
Ac_EH1r_004A04	ref|XP_003340320.1|	cytoplasmic 1-like actin	*Monodelphis domestica*	102	3e-20
Ac_EH1r_005B07	gb|ABN80454.1|	cyclin G1	*Poecilia reticulata*	90.5	8e-17
Ac_EH1r_009C03	ref|NP_001122208.1|	THAP domain containing protein 4	*Danio rerio*	176	1e-42
Ac_EH1r_01A07	sp|P80946.1|	Hemoglobin anodic subunit beta	*Anguilla anguilla*	283	1e-74
Ac_FMf_08F03	ref|XP_003226802.1|	cohesin subunit SA-2-like isoform 2	*Anolis carolinensis*	219	8e-56
Ac_M175_01H02	emb|CAQ87569.1|	NKEF-B protein	*Plecoglossus altivelis*	365	3e-99
Ac_197Ff_01E04	ref|XP_002121150.1|	CUB and sushi domain-containing protein 3	*Ciona intestinalis*	80.5	2e-13
Ac_EH1f_01D07	ref|XP_002606965.1|	hypothetical protein	*Branchiostoma floridae*	82.8	3e-14
Ac_M175_01B06	ref|XP_422710.2|	hypothetical protein	*Gallus gallus*	123	1e-26

Table 3.2: Annotaion of putative host-derived sequences in the *A. crassus*-dataset - Sequences excluded because of inferred host-origin comparing similarity to nematode- and fish-proteins. The annotation obtained against NCBI-nr are in agreement with this inference of host origin, as only best hits to vertebrate proteins are found.

4

Evaluation of an assembly strategy for pyrosequencing reads

4.1 Overview

This chapter reports on an important methodical detail of 5: the sequence-assembly. The quality of this sequence assembly constitues a fundamental foundation of the later chapters.

The pre-processed *A. crassus* data-set consisting of 100491819 bases in 353055 reads (58617 generated using "FLX-chemistry", 294438 using "Titanium-chemistry") was assembled following an approach proposed by (127): Two assemblies were generated, one using `newbler v2.6` (142), the other using `mira v3.2.1` (181). The resulting assemblies (referred to as first-order assemblies) were merged with `Cap3` (182) into a combined assembly (referred to as second-order assembly).

Summary statistics for the assemblies, demonstrating the superiority of the second-order assembly are reported as well as summary statistics for single contigs. These metadata on contigs are important in evalutation of downsteam results. As a perfect assembly with each contig represtenting a single full transcript is illusive and every contig constitutes an hypothesis, it becomes important to validate and question analyses based on as much information as possible.

4.2 The `newbler` first-order assembly

During transcriptome-assembly `newbler` can split individual reads spanning the breakpoints of alternate isoforms, to assemble e.g. the first portion of the reads in one contig,

the second portion in two different contigs. Later multiple so called isotigs would be constructed and reported, one for each putative transcript-variant. While this approach could be helpful for the detection of alternate isoforms, it also produces short contigs (especially at error-prone edges of high-coverage transcripts) when the building of isotigs fails. The read-status report and the assembly output in `ace`-format the program provides are including short contigs only used during the assembly-process, but not reported in the contigs-file used in transcriptome-assembly projects (`454Isotigs.fna`). Therefore to get all reads not included in contigs (i.e. a consistent definition of "singleton") it was necessary to add all reads appearing only in contigs not reported in the fasta-file to the reported singletons. The number of singletons increased in this step form the 26211 reported to 109052. We later also address the usefulness of `newbler's` report vs. the expanded singleton-category, but for the meantime we define singletons as all reads not present in a given assembly.

As mentioned above, the splitting of reads in the `newbler` assembly can give useful information on possible isoforms, however the the number of contigs `newbler` splitted one read into (in some cases mort than 100 contigs) seems artificially inflated (see figure 4.1). If information would correspond to real isoforms it should be about an order of magnitude lower. This fact emphasises the need for further processing of the contigs. The maximal number of read-splits in a given contig and it's usefulness will be discussed later in greater detail.

4.3 The `mira`-assembly and the second-order assembly

The `mira` assembly provided a second estimate of the transcriptome. In this assembly individual reads are not split. The number of reads not used in the `mira`-assembly was 65368.

To combine the two assemblies `cap3` was used with default parameters and including the quality information from first-order assemblies. The reminder of this text deals with the exploratory analysis of how information from both estimates of the transcriptome are integrated into the final second-order assembly.

Table 4.1 gives basic summary-statistics of the different assemblies. `mira` clearly produced the biggest assembly, both in terms of number of contigs and bases), the second-order assembly is slightly smaller size than the `newbler` assembly. The second-order assembly had on average longer contigs than both first-order assemblies and a higher weighted median contig size (N50).

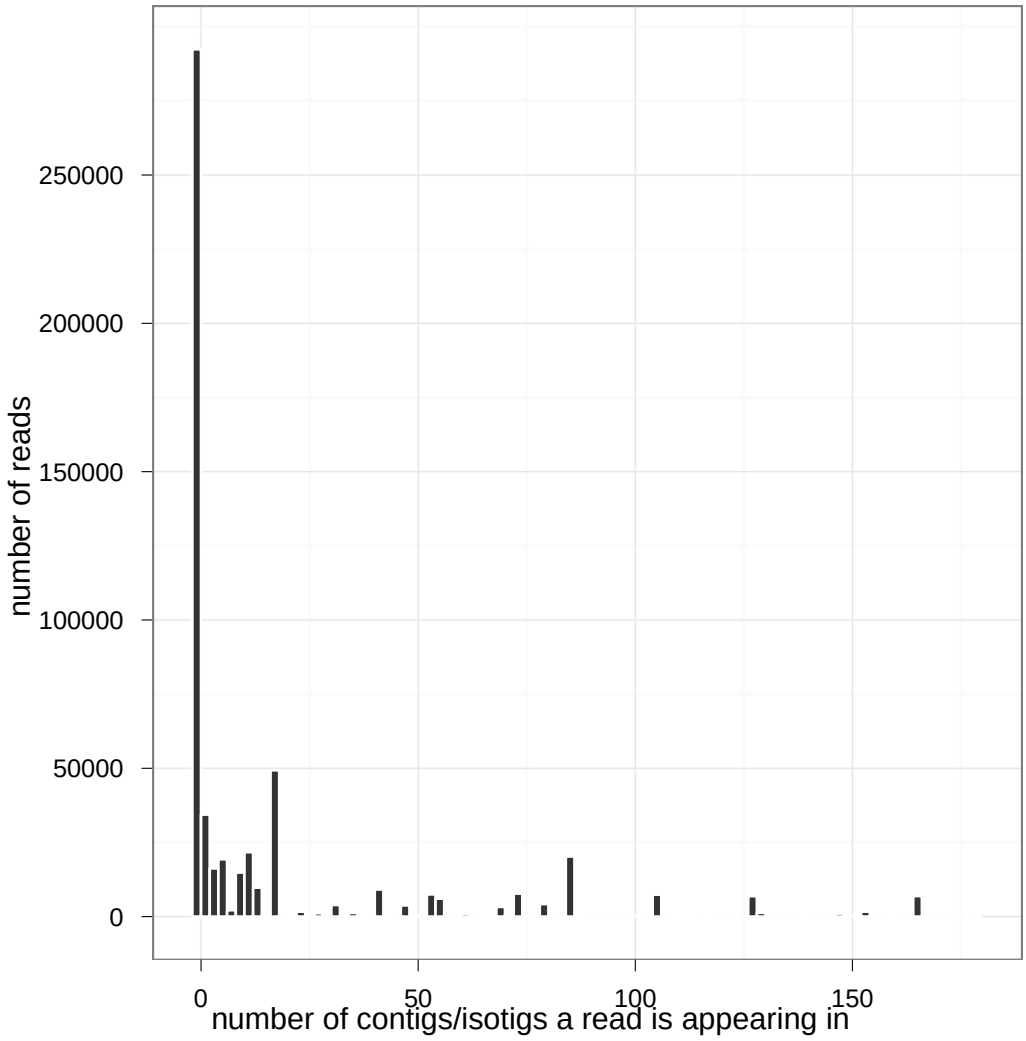

Figure 4.1: Number of contigs/isotigs splitted - A histogram of the number of contigs or isotigs `newbler` splited a single read into

	Newbler	Mira	Second-order(MN)
Max length	6300	6352	6377
Number of contigs	15934	22596	14064
Number of Bases	8085922	12010349	8139143
N50	579	579	662
Number of congtigs in N50	4301	6749	3899
non ATGC bases	375	29962	5245
Mean length	508	532	579

Table 4.1: Statistics for the first-order assemblies - Basic statistics for the first-order assemblies and the second-order assemly (for which only the most relieable category of contigs (MN) is shown see refsec:data-categ-second)

4.4 Data-categories in the second-order assembly

Three main categories of assembled sequence data can be distinguished in the second-order assembly, each one with different reliability and purpose in downstream applications: The first category of data obtained are the singletons of the final second-order assembly. It comprises raw sequencing reads that neither of the first-order assemblers used. It is therefore the intersecion of the `newbler`-singletons (as defined in 4.2) and the `mira`-singletons. 47669 reads fell in this category. A second category of sequence contains the first-order contigs, that could not be assembled in the second-order assembly (the singletons in the `cap3`-assembly; M_1 and N_1 in table 4.2). Furthermore second-order contigs in which first-order contigs from only one assembler are combined (M_n and N_n in table 4.2) also have to be included in this category. Sequences in this category should be considered only moderately reliable as they are supported by only one assembly algorithm.

Finally the category of contigs considered most reliable contains all second-order contigs with contribution from both first-order assemblies (MN in table 4.2).

For the last, most reliable (MN) category, reads contained in the assembly can be categorized depending on whether they entered the assembly via both or only via one first-order assembly.

Figure 4.2 gives a more detailed view of the fate of the reads `newbler` splited during first-order assembly. Interestingly most reads `newbler` splited ended in the high-quality category of the second order assembly only

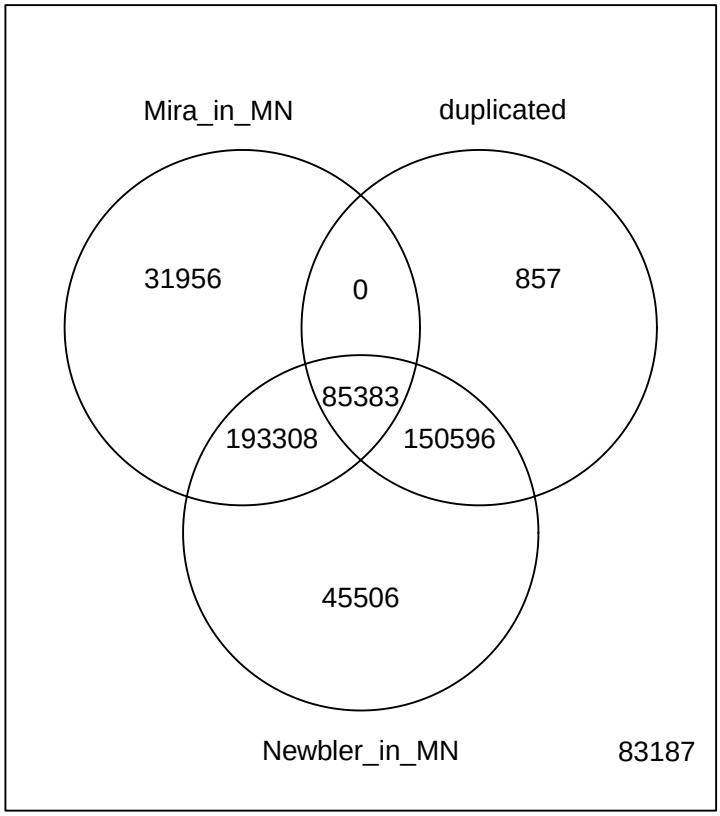

Figure 4.2: Origing of reads - Reads in the most reliable (MN) assembly-category are categorized by the way they entered the assembly: Although they are in a highly credible contig, reads can still have entered from only one first order assembly (Mira_in_MN or Newbler_in_MN). The intersection gives the reads which entered via both routes. The duplicated category gives the number of reads splitted by Newber and the intersection reads, which were splitted and entered the assembly.

	M_1	M_n	MN	N_n	N_1
Snd.o.con		164	13887	13	
Fst.o.con	2347	897	mira=19352/newbler=14410	40	1484
reads	42172	21153	one=269868/both=193308	1538	13100

Table 4.2: Number of reads in assemblies for first-order contigs (Fst.o.con) and second-order contigs (Snd.o.con) numbers for for different categories of contigs are given: M_1 and N_1 = first-order contigs not assembled in second-order assembly, from mira and newbler respectively; M_n and N_n = assembled in second-order contigs only with contigs from the same first-order assembly; MN = assembled in second-order contigs with first order contigs from both first order assemblies.

4.5 Contribution of first-order assemblies to second-order contigs

Looking at the contribution of contigs from each of the assemblies to one second-order contig in figure 4.3 a it becomes clear, that the `mira`-assembly had a high number of redundant contigs. These were assembled into the same contig by `newbler` and finally also in one second-order contig by `Cap3`.

A different picture emerges from the contribution of reads through each of the first-order assemblies (figure 4.3 b). Here for most second-order contigs many more reads are contributed through `newbler`-contigs. This is because `newbler` has more reads summed over all contigs caused by the duplication due to splitting of reads.

4.6 Evaluation of the assemblies

To further compare assemblies (`mira, newber` first-order assemblies including or excluding their singletons) and the second-order assembly (including different contigs-categories and singletons) we evaluated the number of bases or proteins their contigs and singletons (partially) cover in the related model-nematodes, *Caenorhabditis elegans* and *Brugia malayi*.

In addition, the size of the assembly can give an indication of redundancy or artificially assembled data: If it increases without improving the reference-coverage the dataset is likely to contain more redundant or artificial information, a more parsimonious assembly should be preferred.

The database-coverage for the two reference species can then be plottedf against the

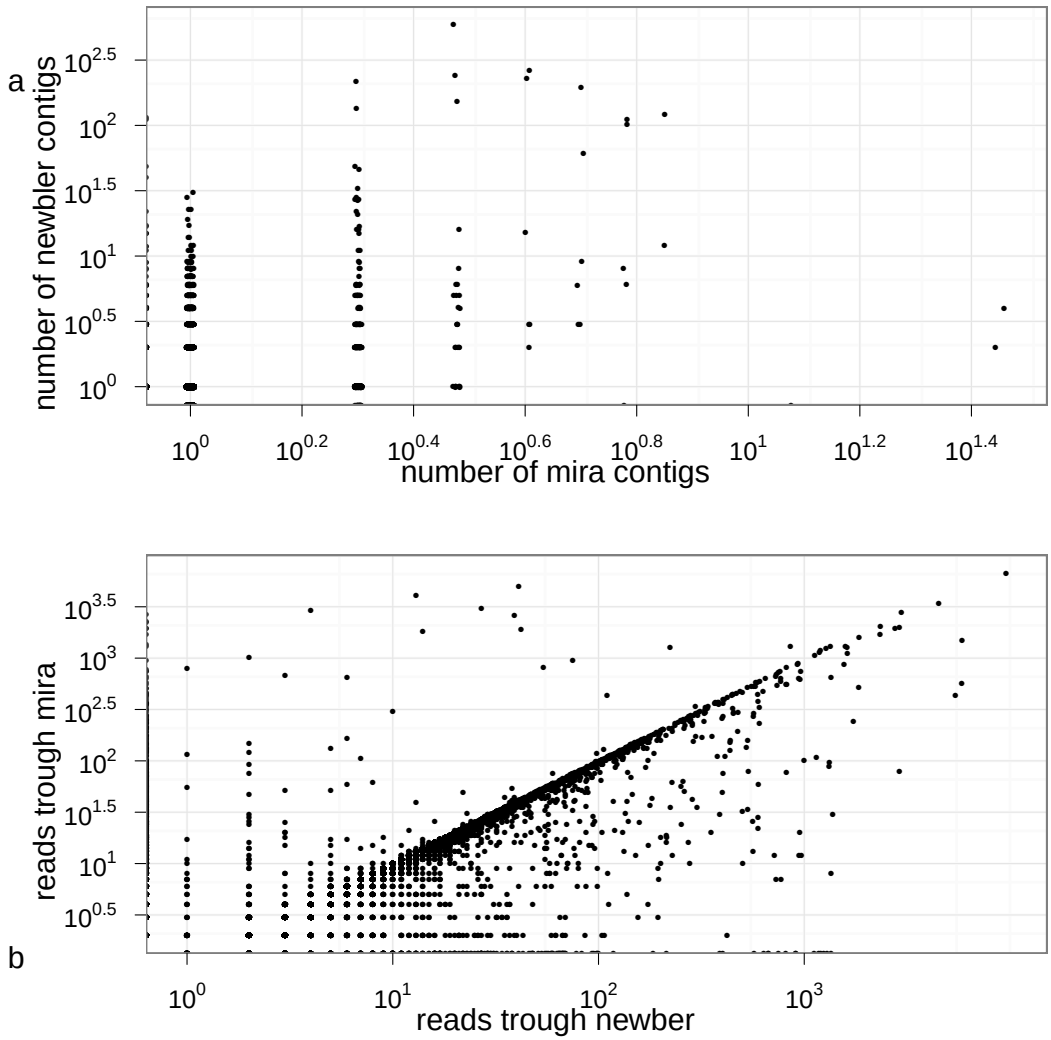

Figure 4.3: Contribution to second-order assembly - Number of first-order contigs from both first-order assemblies for each second order contig (a) number of reads through `newbler` and `mira` for each second-order contig (b)

size of the assembly-dataset to estimate the completeness conditional to the size of the assembly (figures 4.4, 4.5, 4.5) .

From the assemblies excluding singletons (in the lower left corner with lower size and database-coverage) the highly reliable contig-category of the second-order assembly produced the highest per-base coverage in both reference-species, with the `newbler` assembly on a second place and `mira` producing the lowest reference-coverage. When adding the contigs considered lower quality supported by only one assembler to the second-order assembly the reference-coverage increased moderately.

Including singletons the `mira` and `newber` assemblies were of increased size. A comparison of the `newbler's` reported singletons with all singletons addet to the `newbler`-assembly shows, that the reported singletons increased reference-coverage to the same amount than all singletons, while the non-reported singletons only increased the size of the assembly. It can be concluded, that the latter contain hardly any additional information but only error-prone or variant reads.

The second-order assembly including the intersection of first-order singletons performed similar to the `newbler` assembly for the number of bases coverd, but was larger in size. Adding the less reliable set of one-assembler supported second-order-contigs the assembly coverd the most bases in both references. When not the singleton of the second-order assembly (as defined in 4.2) but the intersection of `newbler's` "reported singletons" and `mira's` singletons were considered a very parsimonious assembly with high reference-coverage (termed fullest assembly; and labeled FU in the plots above) was obtained.

Considering the reference-database with any kind of coverage the second-order assembly performed less preferable. Excluding singletons it was covering similar numbers of database-proteins than the `newber`-assembly and and was outperformed by the `mira`-assembly, although the latter showed again to be least parsimonious. The same general picture emerged from this analysis when singletons were considered additionally. `newbler` and second-order assemblies coverd similar amounts of reference-data.

When database-proteins covered for at least to 80% of their length are considered the second-order assembly showed it's superiority: Both ex- and including singletons the second-order assembly outperformed the first-order assemblies. Moderate gains in reference coverage were made again for the addition of dubious single-assembler supported second-order contigs. We give most weight in our analysis to these results as in average longer correct contigs will allow finding the highest number of putative full-length genes.

Given this evaluation we defined a "minimal adequate" assembly as the subset of

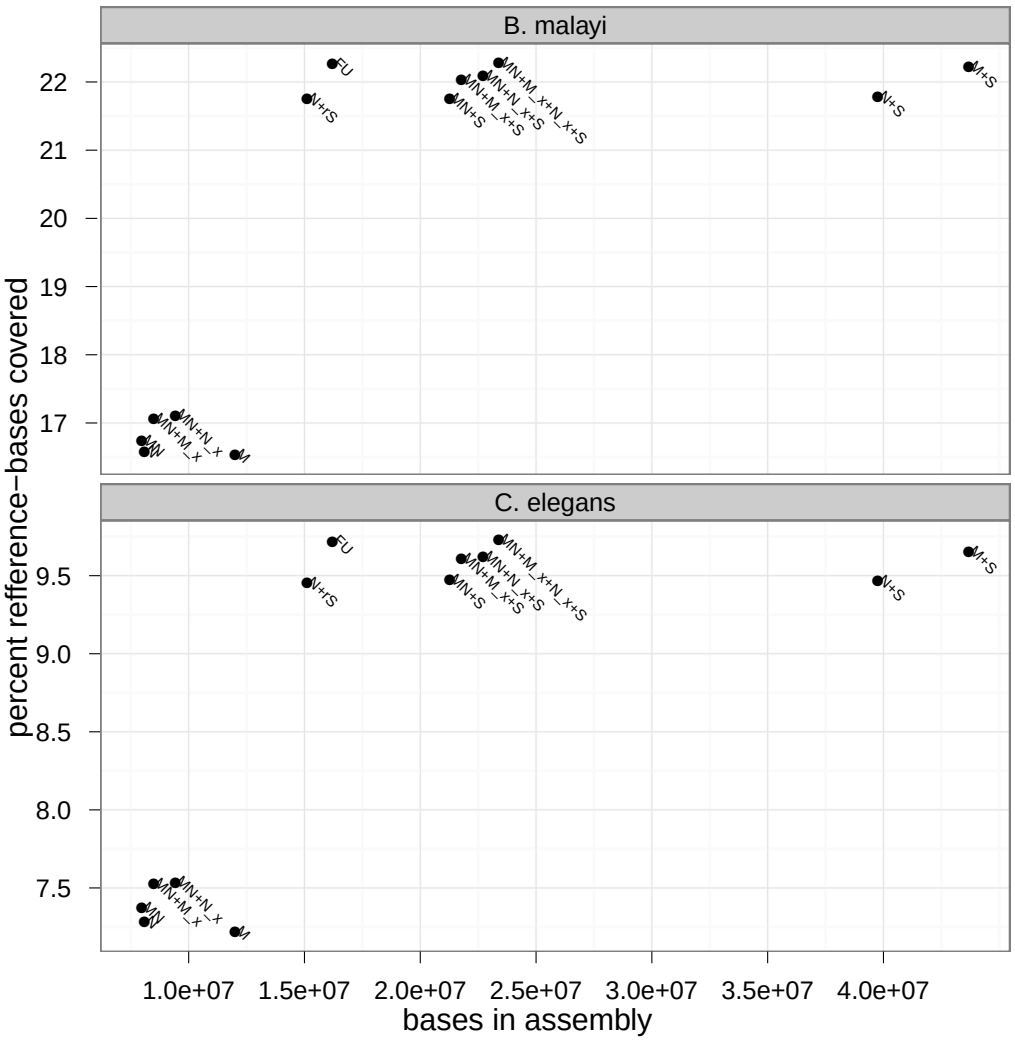

Figure 4.4: Base-content and reference-transcriptome coverage in percent of bases - for different assemblies and assembly-combinations; M = `mira`; N = `newbler`; $M + S$ = `mira` + singletons; $N + S$ = `newbler` plus singletons; $N + Sr$ = `newbler` plus singletons reported in readstatus.txt; MN = second-order contigs supported by both first-order; $MN + N\_x$ = second-order MN plus contigs only supported by `newbler`; $MN + M\_x$ = same for `mira`-first-order-contigs; $MN + M\_x + S$ and $MN + N\_x + S$ same with singletons; FU = second-order contigs supported by both or one assembler plus the intersection of `newbler` reported singletons and `mira`-singletons = the basis for the "fullest assembly" used in later analyses

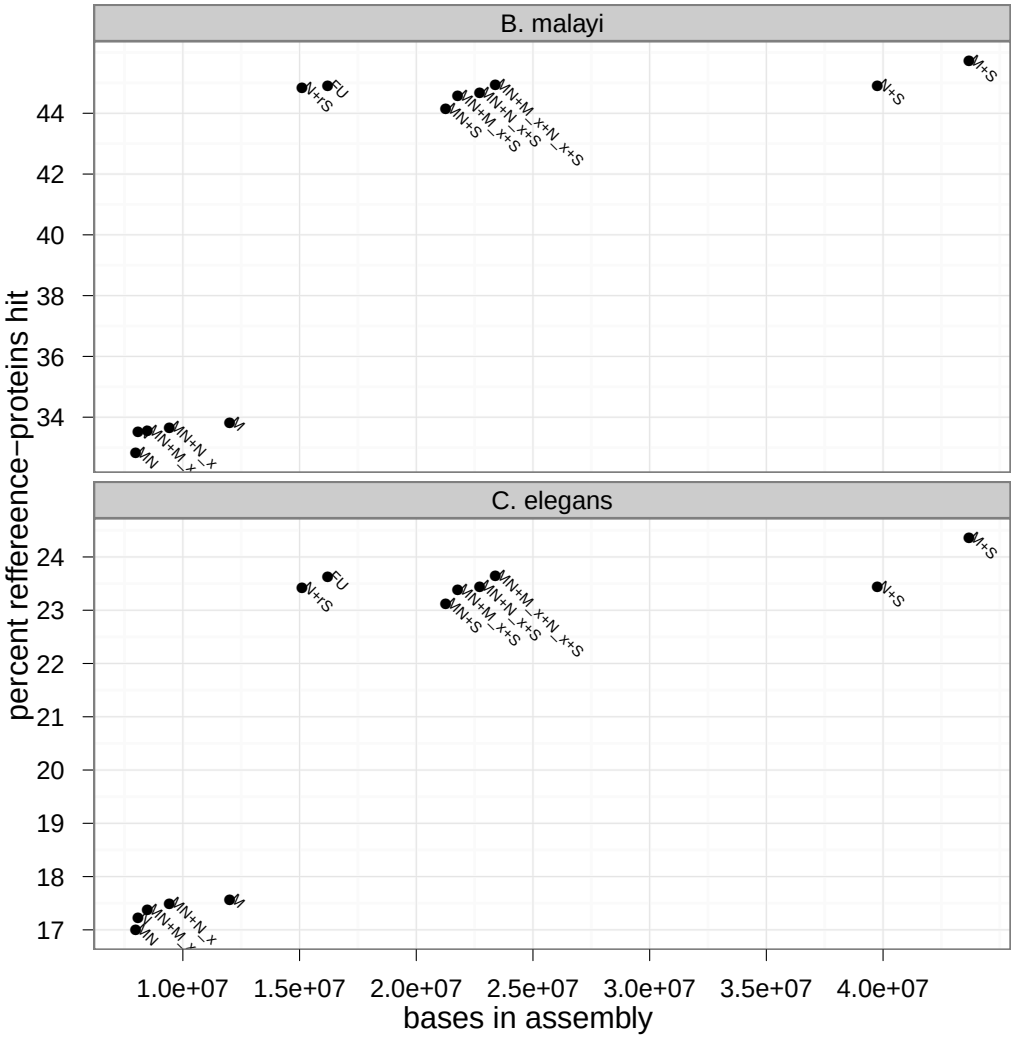

Figure 4.5: Base-content and reference-transcriptome coverage - in percent of proteins hit for different assemblies and assembly-combinations (for category-abrevations see figure 4.4)

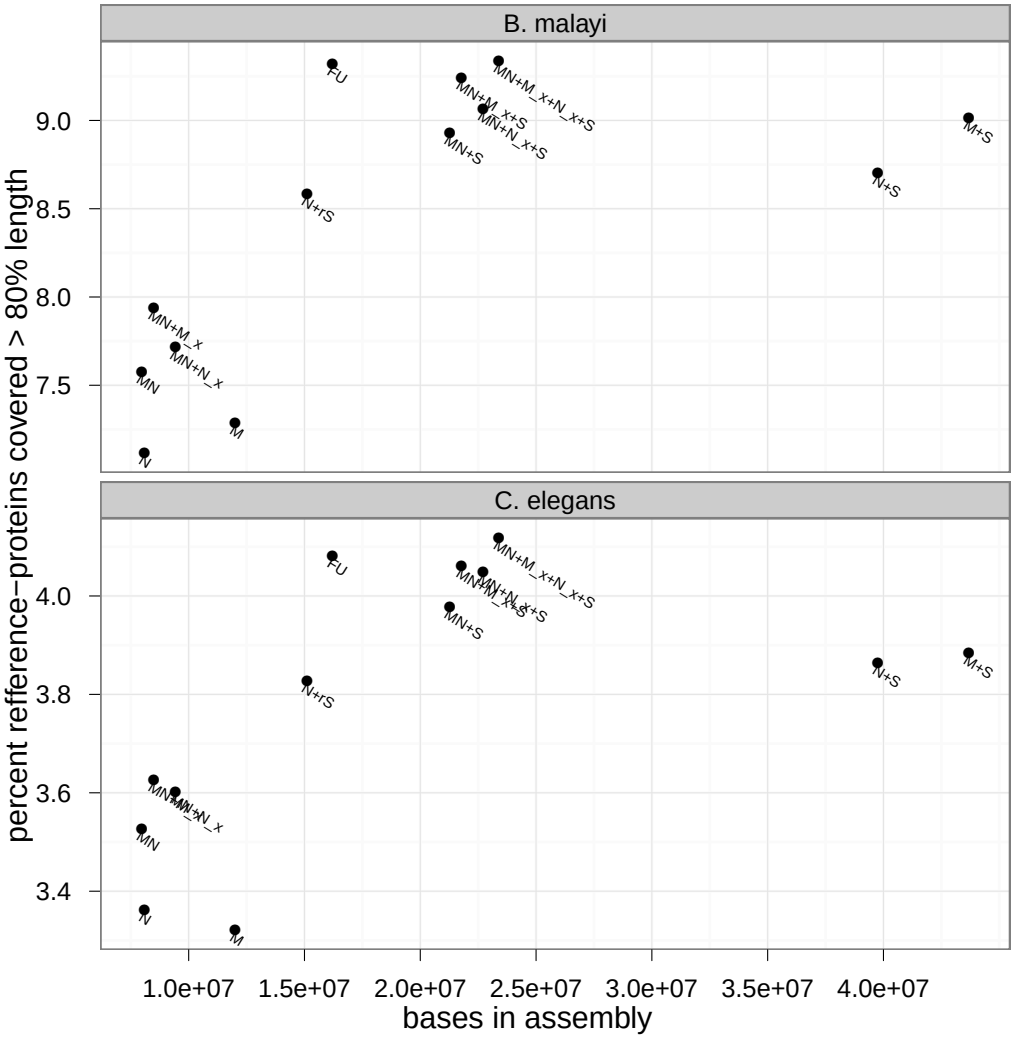

Figure 4.6: Base-content and reference-transcriptome coverage in percent of proteins coverd to at least 80% - of their length for different assemblies and assembly-combinations (for category-abrevations see figure 4.4)

contigs of the second-order assembly supported by both assemblers (labeled MN above).

Given the performance of the singletons `newbler` reported we defined a "fullest-assembly" as all second-order contigs (including those supported by only one assembler) plus the intersection of reported `newbler`-singletons and `mira` singletons.

4.7 Measurments on second-order assembly

Based on the following reads through the complicated assembly process, we calculated the following for each contig in the second-order assembly, to report it to for use in later analysis.

- number of `mira` and `newbler` first-order contigs
- number of reads through `mira` and reads through `newbler`
- number of reads being split by `newbler` in first-order assembly
- number of read-split events in the first-order assembly (equals the sum of reads multiplied by number of contigs a read has been split into)
- maximal number of first-order contigs a read in the contig has been split into during `newbler`-assembly
- the number of reads same-read-paires from the `newbler` and `mira` first order-assembly merged in a second order contig
- cluster-id of the contig: All contigs "connected" by sharing reads (similar to the graph clustering reported in (161)).
- number of other second order contigs containing the same read (size of the cluster)

4.7.1 Contig coverage

As well defined coverage-information is not readyly avaiable from the output of this combined assembly aproach (although we followed individual reads through the process) we inferred coverage by mapping the reads used for assembly against the fullest assembly using ssaha2 (157) with parameters (-kmer 13 -skip 3 -seeds 6 -score 100 -cmatch 10 -ckmer 6 -output sam -best 1). We converted the `sam`-ouput via a sorted `bam`-file to `pileup`-format using `samtools` (183).

For a second evalutation we excluded best-hits mapping to multiple contigs before converting the `sam`-file.

- mean per base coverage
- mean unique per base coverage

4.7.2 Example use of the contig-measurements

Based on these measurements the emergence of a given contig from the assembly process can be reconstructed. Table 4.3 gives an excerpt of the contig-measurements reported in additional-file `contig-data.csv`. The example contigs are all from large contig-clusters (cluster.size), where interpretation of the assembly history is complicated, but not impossible:

	Contig1047	Contig10719	Contig104	Contig13672
reads_through_Newbler	16	1351	0	14
reads_through_Mira	26	651	135	0
Newbler_contigs	1	5	0	2
Mira_contigs	1	9	4	0
category	MN	MN	M_n	N_n
num.new.split	8	1314	0	0
sum.new.split	16	2628	0	0
max.new.split	2	2	0	0
num.SndO.pair	13	644	0	0
cluster.id	CL62	CL6	CL176	CL235
cluster.size	24	18	5	5
coverage	4.200342	267.495458	41.003369	2.920755
uniq_coverage	4.248960	7.425507	2.568000	1.196078

Table 4.3: Example table for assembly-measurements - measurements on contigs (as given in additional file `contig-data.csv`), row-labels are explained in a detailed example in the main text

Contig1047 is in the well trusted MN category of contigs. It consists of only one contig from each first-order assembly (newbler_contigs and mira_contigs), each containing a set of reads of moderate size: 16 from `newbler` (reads_through_newbler) 26 from `mira` (reads_through_mira). 8 of the 16 reads `newbler` used in its one assembled contig were also assembled to a differnt `newbler`-contig (num.new.split). That each of the 8 reads was only appearing in one other `newbler`-contig is visible from the fact, that the number of split events is 16 (sum.new.split) and the maximal number of splits for one

read is 2 (max.new.split). 13 (num.SndO.pair) same-read-pairs from the tow different first-order assemblies were merged in this second-order contig, leaving 3 (16-13) reads in `newbler`-contigs and 13 (26-13) reads in `mira` contigs, which all could potentially have ended up in other contigs. The contig is in a cluster (CL62), which contains in total 24 contigs (cluster.size). It has to be addmitted that the whole graph-structure linking this 24 contigs can't be reconstructed from this contig summary data. On the other hand the summary data makes clear, from what source the links for cluster-affiliation have resulted: In this case from 3 and 13 unlinked read-paires from both first-order assemblies and 8 split-reads from `newbler`-fistr order contigs.

A comprehensive interpretation of the other example-contigs depicted is left to the reader. It should just be remarked, that in case of one-assembler supported contigs, all reads in that contig could potentially be represented in other contigs, making average cluster-size in these contigs bigger than in the MN category.

One of the most interesting measurement calculated for each contig is the cluster-membership and cluster-size. Such clusters can represent close paralogs, duplicated genes, isoforms from alternative splicing or allelic variants.

These measurements can be used in later analysis to e.g. reevaluate the likelihood of misassembly in a given set of contigs. An evaluation of other cluster members, when biologically interesting properties are inferred for a contig in a cluster is e.g. advised and will be demonstrated in later in the manuscript.

4.8 Finalising the fullest assembly set

In order to minimize the amount of sequence with artificially inferred isoform-breakpoints we used the unique-mapping-information described above to detect contigs and singletons not supported by any raw data (reads). Table 4.4 gives a summary of these unsupported data by contig-category. For all downstream-analysis we removed all well trusted MN-category contigs having no coverage at all and the contigs (and singletons) from other categories having no unique coverage.

	singletons	M_1	M_n	MN	N_1	N_n
coverage == 0	546	34	2	36	158	0
unique coverage == 0	584	48	2	42	210	3

Table 4.4: Final filtering of the assembly - Number of contigs with a coverage and unique-coverage of zero, inferred from mapping of raw reads, listed by contig-category

Thereby we reduced our dataset to 40187 tentative unique genes (TUGs), redifining the "fullest assembly" dataset. Based on the above evaluation we decided to treat the MN-category of contigs as high credibility assembly (highCA) and to subsume the M_n, N_n, M_1, N_1 and `Newbler's` reported singletons as additional low credibility assembly (lowCA).

5

Pyrosequencing of the *A. crassus* transcriptome

5.1 Overview

In this chater the transcriptome assembly of *A. crassus* is analysed in its biological context. It constitues a basis for molecular research on this important species and furthermore provides unique insights into the evolution of parasitism in the Spirurina.

After extensive screening of 756.363 raw pyrosequencing reads, we assembled 353.055 into 11.371 contigs spanning 7.971.550 bases and additionally obtained 21.147 singleton and lower quality contigs spanning 8.095.986 bases. We obtained annotations for ca. 60% of the contigs and 40% of the tentatively unique genes (TUGs) confirming the high quality of especially the contigs. We found an overabundance of predicted signal peptide cleavage sites in sequence conserved in Nematoda and novel in *A. crassus*. We identified 5112 high quality Single nucleotide polymorphisms (SNPs) and suggest 199 of them as most suitable markers for population-genetic studies. GO-term analysis identified 13 proteinases under positive selection. Comparing male and female as well as Asian and European *A. crassus* we developed a method for future work with this transcriptome as a reference in mapping experiments.

5.2 Sampling *A. crassus*

One female worm and one male worm were sampled from an aquaculture with height infection loads in Taiwan. An additional female worm was sampled from a stream with low infection pressure adjacent to the aquaculture. All these worms were parasitising

endemic *An. japonica*. A female worm and pool of L2 larval stages were sampled from *An. anguilla* in the river Rhine, one female worm from a lake in Poland. All adult worms were filled with large amounts of host-blood, therefore we anticipated abundant host-contamination in sequencing data and decided to sequence a liver sample of an uninfected *An. japonica* for screening.

5.3 Sequencing, trimming and pre-assembly screening

A total of 756363 raw sequencing reads were generated for *A. crassus* (see table 5.1). These were trimmed for base call quality, and filtered by length to give 585949 high-quality reads (spanning 169863104 bases). In the eel dataset from 159370 raw reads 135072 were assembled after basic quality screening.

We then screened the *A. crassus* reads for contamination by host (30071 matched previously sequenced eel genes or our own *An. japonica* 454 transcriptome, which had been assembled into 10639 mRNA contigs. (181783 reads matched large or small subunit nuclear or mitochondrial ribosomal RNA sequences of *A. crassus*) . In addition to fish mRNAs, we identified (and removed) 5286 reads in the library derived from the L2 nematodes that had significant similarity to cercozoan (likely parasite) ribosomal RNA genes (see table 5.1).

5.4 Assembly

We assembled the remaining 353055 reads (spanning 100491819 bases) using the combined assembler strategy (127) and Roche 454 GSassembler (version 2.6) and MIRA (version 3.21) (181). From this we derived 13851 contigs that were supported by both assembly algorithms, 3745 contigs only supported by one of the assembly algorithms and 22591 singletons that were not assembled by either approach (see table 5.2). When scored by matches to known genes, the contigs supported by both assemblers are of the highest credibility, and this set is thus termed the high credibility assembly (highCA). Those with evidence from only one assembler and the singletons are of lower credibility (lowCA). These datasets are the most parsimonious (having the smallest size) for their quality (covering the largest amount of sequence in reference transcriptomes). In the highCA parsimony and low redundancy is prioritized, while in the complete assembly (highCA plus lowCA) completeness is prioritized. The 40187 sequences (contig consensuses and singletons) in the complete assembly are referred to below as tentatively unique genes (TUGs).

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

Table 5.1: Statistics for different libraries For two sequencing libraries from European eels (E1 and E2) one form L2-larvae (L2), one from male (M) and two from Eels in Taiwan (T1 and T2) the following statistics are given. life.st = lifecycle stage: f for female m for male. source.p = source population: R for Rhine, P for Poland, C for cultured, W for wild. raw.reads = raw number of sequencing reads obtained. lowqal = number of reads discarded due to low quality or length in *Seqclean* (184). AcrRNA = number of reads hitting *A. crassus*-rRNA (screened). eelmRNA = number of reads hitting eel transcriptome-sequences (screened). eelrRNA = number of reads hitting eel-rRNA genes (screened). Cercozoa = number of reads hitting cercozoan rRNA (screened). valid = number of reads valid after screening (assembled). valid.span = number of bases valid (assembled). mapping.unique = number of reads mapping uniquely to the assembly. mapping.Ac = number of reads mapping to the part of the assembly considered *A. crassus* origin (see post-assembly screening). mapping.MN = number of reads mapping to the highCA-derived part of the assembly (and also *A. crassus* origin).

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

Table 5.2: Assembly classification and contig statistics - Summary statistics for contigs from different assembly-categories given in columns as highCA = high credibility assembly; lowCA = low credibility assembly, combined = complete assembly. Rows indicate summary statistics: total.contigs = numbers of total contigs, fish.contigs = number of contigs hitting eel-mRNA or Chordata in NCBI-nr or NCBI-nt (screened out), xeno.contigs = number of contigs with best hit (NCBI-nr and NCBI-nt) to non-eukaryote (screened out), remaining.contigs = number of contigs remaining after this screening, remaining.span = total length of remaining contigs, non.u.cov = non-unique mean base coverage of contigs, cov = unique mean base coverage of contigs, p4e."X" = number protein predictions derived in p4e, where "X" describes the method of prediction (see Methods), full.3p = number of contigs complete at 3', full.5p = number of contigs complete at 5', GO = number of contigs with GO-annotation, KEGG = number of contigs with KEGG-annotation, EC = number of contigs with EC-annotation, nem.blast = number of contigs with `BLAST`-hit to nematode in nr, any.blast = number of contigs with `BLAST`-hit to non-nematode (eukaryote non chordate) sequence in NCBI-nr.

We screened the complete assembly for residual host contamination, and identified 3441 TUGs that had higher, significant similarity to eel (and chordate) sequences (our 454 ESTs and EMBLBank Chordata proteins) than to nematode sequences (124).

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

1153 TUGs were found mapping to Eukaryota outside of the kingdoms Metazoa, Fungi and Viridiplantae. These hits included a wide range of Protists ranging from Apicomplexa (mainly Sarcocystidae, 28 hits and Cryptosporidiidae 10 hits) over Bacillariophyta (diatoms, mainly Phaeodactylaceae, 41 hits) and Phaeophyceae (brown algae, mainly Ectocarpaceae, 180 hits) and Stramenopiles (Albuginaceae, 63 hits) to Kinetoplasitda (Trypanosomatidae, 26 hits) and Heterolobosea (Vahlkampfiidae, 38 hits).

Additionally we found 298 TUGs with hits to fungi (e.g Ajellomycetaceae, 53 hits) and 585 TUGs with hits to plants.

Hits outside the Eukaryota were mainly to Bacteria (825 hits) and within those mostly to members of the Proteobacteria (484 hits). No hits were found to Wolbachia or related Bacteria known as symbionts of nematodes and arthropods. 9 TUGs were hitting sequence from Viruses and 8 from Archaea.

We excluded all TUGs with best hits outside Metazoa and our assembly thus has 32518 TUGs, spanning 154052 bases (of which 11371 are highCA-derived, and span 154052 bases) that are likely to derive from of *A. crassus*.

5.5 Protein prediction

For 32411 TUGs a protein was predicted using prot4EST (185) (see table 5.2). The full open reading frame was obtained in 353 TUGs, while while for 2683 the 5' end and for 8283 the 3' end was complete. In 13379 TUGs the corrected sequence with the imputed ORF was slightly changed compared to the raw sequence.

5.6 Annotation

We obtained basic annotations with orthologous sequences from *C. elegans* for 9554 TUGs, from *B. malayi* for 9662 TUGs, from nempep (122, 124) for 11617 TUGs and with uniprot proteins for 11113 TUGs.

We used annot8r (186) to assign gene ontology (GO) terms for 6509 TUGs, Enzyme Commission (EC) numbers for 2458 TUGs and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway annotations for 3844 TUGs (see table 5.2). Additionally

5125 highCA derived contigs were annotated with GO terms through `InterProScan` (187). Nearly one third (6987) of the *A. crassus* TUGs were annotated with at least one identifier, and 1829 had GO, EC and KEGG annotations (see figure 5.1).

We compared our *A. crassus* GO annotations for high-level GO-slim terms to the annotations (obtained the same way) for the complete proteome of the filarial nematode *B. malayi* and the complete proteome of *C. elegans* (see figure 5.2).

Correlation shows the occurrence of terms for the partial transcriptome of *A. crassus* to be more similar to the proteome of *B. malayi* (0.95; Spearman correlation coefficient) than to the proteome of *C. elegans* (0.9). Also the tow model-nematode compared to each other (0.91) are less similar in the occurrence of terms than the two parasites.

We inferred presence of signal peptide cleavage sites in the predicted protein sequence using `SignalP` (188). We predicted 920 signal peptide cleavage sites and 65 signal peptides with a transmembrane signature. Again these predictions are more similar to predictions using the same methods for the proteome *B. malayi* (742 signal peptide cleavage sites and 41 with transmembrane anchor) than for the proteome of *C. elegans* (4273 signal peptide cleavage sites and 154 with transmembrane anchor).

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

5.7 Evolutionary conservation

A. crassus TUGs were classified as conserved, conserved in Metazoa, conserved in Nematoda, conserved in Spirurina or novel to *A. crassus* by comparing them to public databases and using two `BLAST` bit-score cutoffs to define relatedness (see table 5.3).

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

The remaining higCA contigs spread across intermediate relatedness-levels. More sequences were categorised as novel at the phylum level (Nematoda) compared to kingdom and clade III level and the number of contigs at intermediate relatedness-levels was roughly consistent for the two bitscore thresholds.

The latter points about intermediate conservation levels were also true, when all TUGs were analysed. The numbers of TUGs categorised at these intermediate levels roughly doubled. In contrast, the proportion of additional conserved lowCA TUGs is small compared to additional TUGs categorised as novel in *A. crassus*, mirroring the

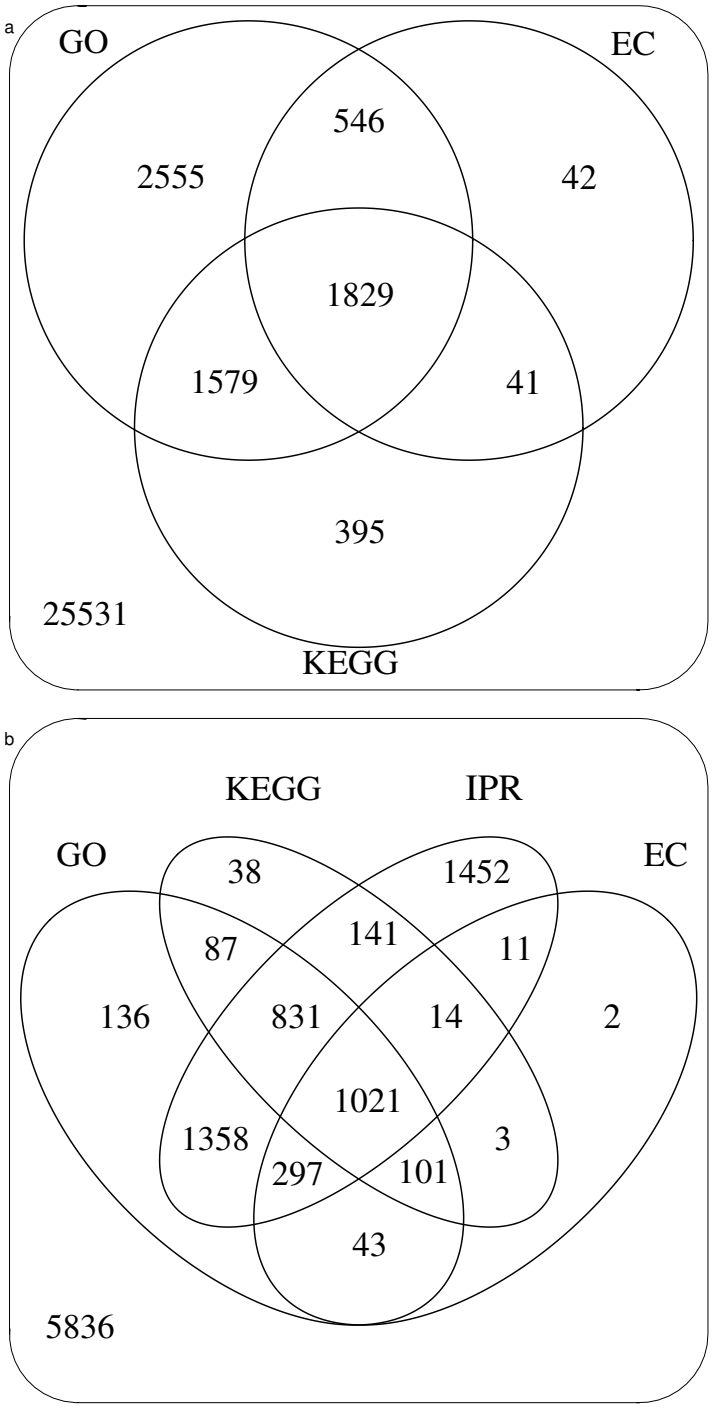

Figure 5.1: Annotation using different identifiers - Number of annotations obtained for Gene Ontology (GO), Enzyme Commission (EC) and Kyoto Encyclopedia of Genes and Genomes (KEGG) terms through `Annot8r` (186) for all TUGs (a) and for higCA derived contigs (b). The latter includes additional domain-based annotations obtained with `InterProScan` (187).

Figure 5.2: Cross taxa comparison of annotation - For Gene Ontology (GO) categories molecular function, cellular compartment and biological process the number of terms in high level GO-slim categories is given as obtained through `Annot8r` (186).

	conserved	novel.in.m	novel.in.n	novel.in.cl3	novel.in.Ac
bit.50.all	5604	1713	2173	1485	21543
bit.80.all	3506	1382	2014	1525	24091
bit.50.highCA	3479	875	1010	601	5406
bit.80.highCA	2457	832	1084	716	6282

Table 5.3: Evolutionary conservation and novelty - The kingdom Metazoa (novel.in.m), the phylum Nematoda (novel.in.n) and clade III (Spirurina; novel.in.cl3) were assessed for occurrences of `BLAST`-hits at two different bitscore thresholds (50 = bit.50 and 80 = bit.80). TUGs without any hit at a given threshold were categorized as novel in *A. crssus* (novel.in.Ac). Both novelty and conservation can be derived from this (numbers for conservation would be the cumulative sum of lower-level novelty).

higher amount of erroneous sequence.

Proteins predicted to be novel to Nematoda and novel in *A. crassus* were significantly enriched in signal peptide annotation compared to conserved proteins, proteins novel in Metazoa and novel in clade III (Fisher's exact test $p<0.001$; 5.3).

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

5.8 Identification of single nucleotide polymorphisms

We called single nucleotide polymorphisms (SNPs) on the 1099419 bases of the TUGs that had coverage of more then 8-fold available using `VARScan` (162). We excluded SNPs predicted to have more than 2 alleles or that mapped to an undetermined (N) base in the reference, and retained 10458 SNPs. The ratio of transitions (ti; 6890) to transversion (tv; 3568) in this set was 1.93 . Using the prot4EST predictions and the corrected sequences, 7153 of the SNPs were predicted to be inside an ORF, with 2310 at codon first positions, 1819 at second positions and 3024 at third positions. As expected ti/tv inside ORFs (2.41) was higher than outside ORFs (1.25). The ratio of synonymous polymorphisms per synonymous site to non-synonymous polymorphisms per non-synonymous site (dn/ds) was 0.42. We filtered these SNPs to exclude those that might be associated with analytical bias. As Roche 454 sequences have well-known systematic errors associated with homopolymeric nucleotide sequences (140), we analysed the effect of exclusion of SNPs in, or close to, homopolymer regions. We observed changes in ti/tv and in dn/ds when SNPs were discarded using different size thresholds for homopolymer runs and proximity thresholds (see figure 5.4).

Based on this we decided to exclude SNPs with a homopolymer-run as long as or longer than 4 bases inside a window of 11 bases (5 to bases to the right, 5 to the left) around the SNP. We also observed a relationship between TUG dn/ds and TUG coverage, associated with the presence of sites with low abundance minority alleles (less than 7% of the allele calls), suggesting that some of these may be errors. Removing low abundance minority allele SNPs from the set removed this effect (see figure 5.5). Our filtered SNP dataset includes 5112 SNPs. We retained 4.65 SNPs per kb of contig sequence, with 8.37 synonymous SNPs per 1000 synonymous bases and 2.4 non-synonymous SNPs per 1000 non-synonymous bases. A mean dn/ds of 0.231 was calculated for the 859 TUGs (762 highCA-derived contigs) containing at least one synonymous SNP.

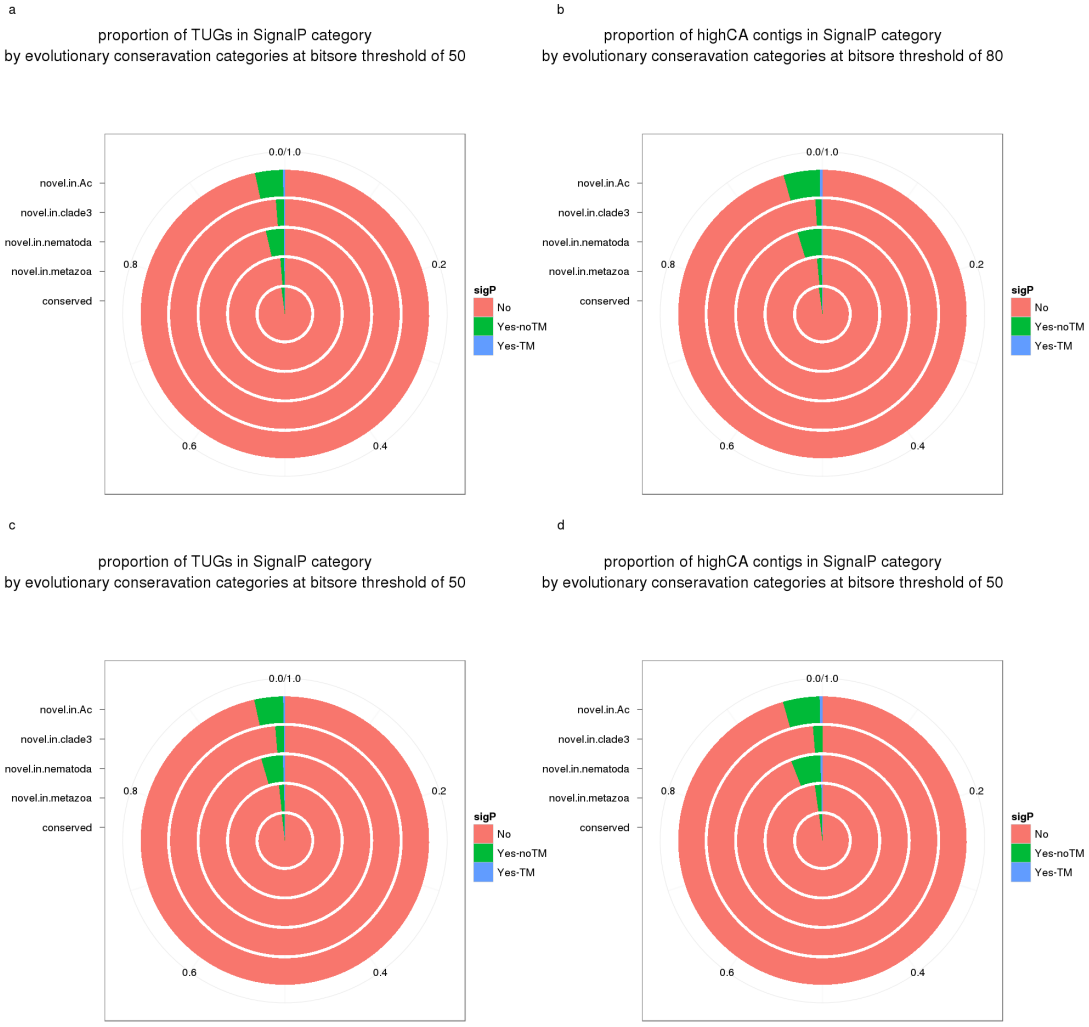

Figure 5.3: Enrichment of Signal-positives for categories of evolutionary conservations - Proportions of `SignalP`-predictions for each category of evolutionary conservation. Generally - across bit-score thresholds - TUGS novel in nematodes and in *A. crassus* have the highest proportion of signal-positives.

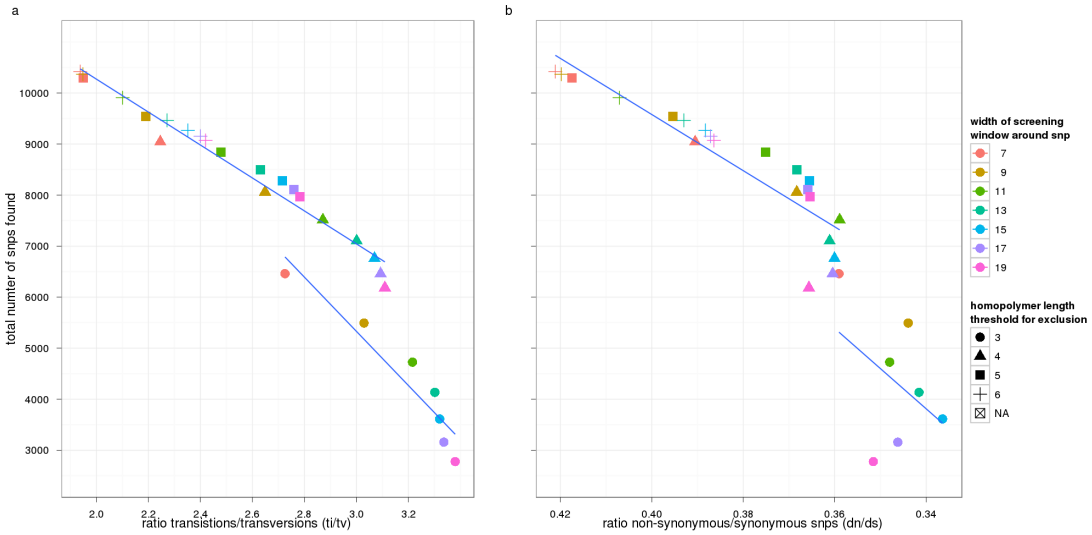

Figure 5.4: Homoplymer screening for SNP-calling - When SNPs in or adjacent to homopolymeric regions are removed changes in ti/tv and dn/ds are observed: As the overall number of SNPs is reduced both ratios change to more plausible values. Note the reversed axis for dn/ds to plot these lower values to the right. For homopolymer length $>$ 3 a linear trend for the total number of SNPs and the two measurements is observed. A width of 11 for the screening window provides most plausible values (suggesting specificity) while still incorporating a high number of SNPs (sensitivity).

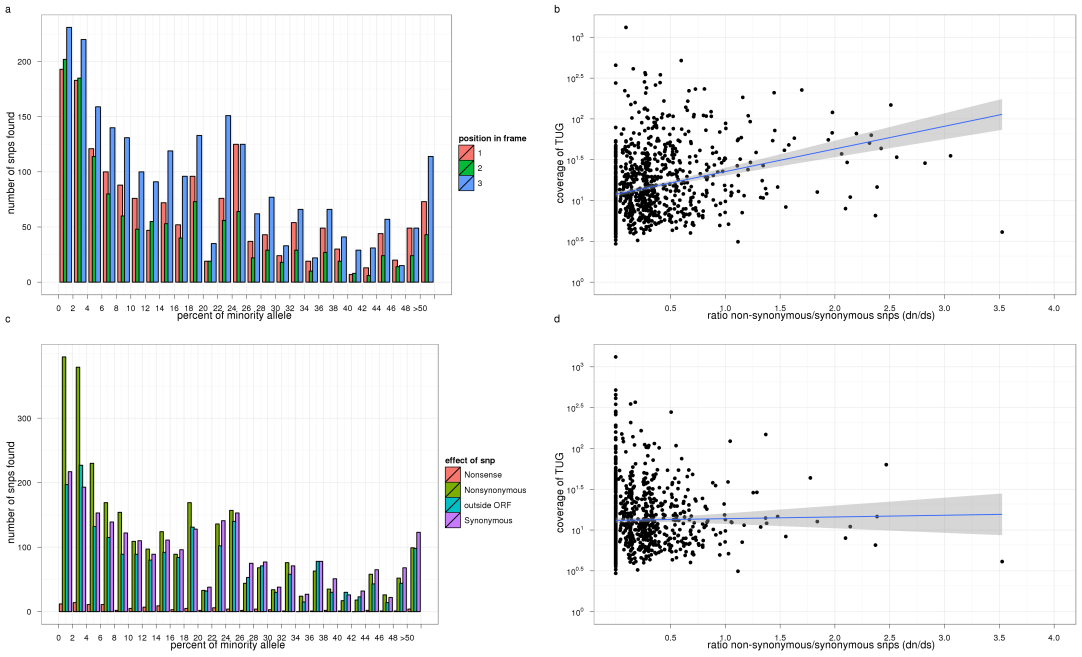

Figure 5.5: SNP calling and SNP categories - Overabundance of SNPs at (a) codon-position two and of (c) non-synonymous SNPs for low percentages of the minority allele. (b) Significant positive correlation of coverage and dn/ds before removing these SNPs at a threshold of 7% ($p < 0.001$, $R^2 = 0.015$) and (d) afterwards (R^2 <0.001, p =0.211).

5.9 Polymorphisms associated with biological processes

We consolidated our annotation and polymorphism analyses by examining correlations between nonsynonymous variability and particular classifications.

Signal peptide containing proteins have been shown to have higher rates of evolution than cytosolic proteins in a number of nematode species. In *A. crassus*, TUGs predicted to contain signal peptide cleavage sites in SignalP showed a trend towards higher dn/ds values than TUGs without signal peptide cleavage sites ($p = 0.074$; two sided Mann-Whitney-test).

Positive selection can be inferred from dn/ds analyses, and we defined TUGs with a dn/ds higher than 0.5 as positively selected. We identified over- and under-represented GO ontology terms associated with these putatively positively selected genes (see table 5.4). Within the molecular function category, "peptidase activity" was the most significantly overrepresented term and had 13 TUGs supporting the overrepresentation. The highlighted 13 peptidases annotated with eleven unique orthologs in *C. elegans* and *B. malayi*. The term "structural constituent of ribosome" was underrepresented.

Pvalue	Count	Size	Term	direction
0.00	13	45	peptidase activity	Over
0.01	7	18	heme-copper terminal oxidase activity	Over
0.01	7	18	oxidoreductase activity, acting on a heme group of donors	Over
0.01	7	18	oxidoreductase activity, acting on a heme group of donors, oxygen as acceptor	Over
0.01	7	18	cytochrome-c oxidase activity	Over
0.01	49	283	catalytic activity	Over
0.02	13	52	transmembrane transporter activity	Over
0.02	9	31	monovalent inorganic cation transmembrane transporter activity	Over
0.02	2	2	L-amino acid transmembrane transporter activity	Over
0.03	9	33	inorganic cation transmembrane transporter activity	Over
0.03	23	117	hydrolase activity	Over
0.03	8	29	hydrogen ion transmembrane transporter activity	Over
0.04	3	6	ribonucleoprotein binding	Over

0.04	13	58	transporter activity	Over
0.05	11	47	substrate-specific transmembrane transporter activity	Over
0.05	16	77	oxidoreductase activity	Over
0.00	1	53	structural constituent of ribosome	Under
0.03	7	93	RNA binding	Under
0.04	2	44	transition metal ion binding	Under
0.05	0	20	protein binding transcription factor activity	Under
0.05	0	20	transcription factor binding transcription factor activity	Under
0.05	0	20	transcription cofactor activity	Under
0.00	13	37	brain development	Over
0.00	14	45	central nervous system development	Over
0.00	6	12	response to electrical stimulus	Over
0.00	3	3	branched chain family amino acid metabolic process	Over
0.00	3	3	branched chain family amino acid catabolic process	Over
0.00	11	36	ATP synthesis coupled electron transport	Over
0.00	11	36	mitochondrial ATP synthesis coupled electron transport	Over
0.01	7	18	mitochondrial electron transport, cytochrome c to oxygen	Over
0.01	22	101	nervous system development	Over
0.01	11	38	oxidative phosphorylation	Over
0.01	6	15	response to starvation	Over
0.01	12	45	cellular amino acid metabolic process	Over
0.01	7	20	positive regulation of cell cycle process	Over
0.01	14	58	amine metabolic process	Over
0.01	4	8	positive regulation of organelle organization	Over
0.01	4	8	spermatid development	Over
0.01	4	8	spermatid differentiation	Over
0.01	5	12	hindbrain development	Over
0.01	5	12	cerebellum development	Over

0.01	5	12	metencephalon development	Over
0.01	5	12	response to methylmercury	Over
0.01	5	12	autophagy	Over
0.02	36	203	response to stress	Over
0.02	2	2	embryonic body morphogenesis	Over
0.02	2	2	xylulose metabolic process	Over
0.02	2	2	L-amino acid transport	Over
0.02	2	2	neuromuscular process controlling balance	Over
0.02	2	2	response to sucrose stimulus	Over
0.02	2	2	NADP metabolic process	Over
0.02	2	2	response to disaccharide stimulus	Over
0.02	2	2	pentose metabolic process	Over
0.02	15	66	behavior	Over
0.02	8	27	interphase	Over
0.02	8	27	interphase of mitotic cell cycle	Over
0.02	11	43	electron transport chain	Over
0.02	11	43	respiratory electron transport chain	Over
0.02	29	156	catabolic process	Over
0.02	3	5	positive regulation of mitosis	Over
0.02	3	5	positive regulation of nuclear division	Over
0.02	13	56	cellular amine metabolic process	Over
0.02	20	99	aging	Over
0.02	10	39	regulation of cell cycle process	Over
0.03	17	81	apoptosis	Over
0.03	16	75	regulation of molecular function	Over
0.03	13	57	regulation of cell cycle	Over
0.03	5	14	mitotic cell cycle G1/S transition DNA damage checkpoint	Over
0.03	5	14	sleep	Over
0.03	4	10	cellular amino acid catabolic process	Over
0.03	10	41	reproductive structure development	Over
0.03	3	6	microtubule organizing center organization	Over
0.03	3	6	RNA catabolic process	Over
0.03	3	6	centrosome organization	Over
0.03	8	30	muscle organ development	Over

0.04	11	47	cellular respiration	Over
0.04	13	59	energy derivation by oxidation of organic compounds	Over
0.04	7	25	regulation of catabolic process	Over
0.04	5	15	signal transduction in response to DNA damage	Over
0.04	5	15	G1/S transition of mitotic cell cycle	Over
0.04	5	15	regulation of G1/S transition of mitotic cell cycle	Over
0.04	5	15	mitotic cell cycle G1/S transition checkpoint	Over
0.04	5	15	G1/S transition checkpoint	Over
0.04	5	15	DNA damage response, signal transduction by p53 class mediator	Over
0.04	5	15	regulation of cellular amine metabolic process	Over
0.04	6	20	response to copper ion	Over
0.04	24	131	cellular catabolic process	Over
0.05	4	11	imaginal disc development	Over
0.05	4	11	amine catabolic process	Over
0.05	4	11	skeletal muscle organ development	Over
0.05	11	49	mRNA metabolic process	Over
0.05	2	3	nuclear mRNA cis splicing, via spliceosome	Over
0.05	2	3	germ cell migration	Over
0.05	2	3	positive regulation of mitotic metaphase/anaphase transition	Over
0.05	2	3	mitotic centrosome separation	Over
0.05	2	3	oligosaccharide catabolic process	Over
0.05	2	3	spliceosomal conformational changes to generate catalytic conformation	Over
0.05	2	3	amino acid transport	Over
0.05	2	3	negative regulation of reproductive process	Over
0.05	2	3	centrosome duplication	Over
0.05	2	3	centrosome separation	Over

0.05	2	3	protein tetramerization	Over
0.05	2	3	protein homotetramerization	Over
0.00	15	201	gene expression	Under
0.00	1	57	cellular protein complex disassembly	Under
0.00	1	57	macromolecular complex disassembly	Under
0.00	1	57	protein complex disassembly	Under
0.00	1	57	cellular macromolecular complex disassembly	Under
0.00	1	55	pancreas development	Under
0.00	1	55	endocrine pancreas development	Under
0.00	1	55	endocrine system development	Under
0.00	1	55	viral genome expression	Under
0.00	1	55	viral transcription	Under
0.00	8	131	transcription	Under
0.00	1	54	translational termination	Under
0.00	4	89	translation	Under
0.00	2	66	cellular component disassembly	Under
0.00	2	66	cellular component disassembly at cellular level	Under
0.01	14	178	cellular macromolecule biosynthetic process	Under
0.01	22	243	biosynthetic process	Under
0.01	22	240	cellular biosynthetic process	Under
0.01	15	181	macromolecule biosynthetic process	Under
0.01	2	57	viral reproductive process	Under
0.01	2	57	viral infectious cycle	Under
0.02	0	26	positive regulation of intracellular protein kinase cascade	Under
0.03	1	38	positive regulation of response to stimulus	Under
0.03	0	24	oocyte differentiation	Under
0.03	0	23	oocyte development	Under
0.03	0	23	cation transport	Under
0.04	0	22	positive regulation of MAPKKK cascade	Under
0.05	24	234	growth	Under
0.01	4	7	small nuclear ribonucleoprotein complex	Over
0.01	31	164	mitochondrion	Over

0.02	2	2	Cajal body	Over
0.02	2	2	U5 snRNP	Over
0.02	2	2	U4/U6 x U5 tri-snRNP complex	Over
0.03	17	80	mitochondrial part	Over
0.04	3	6	nuclear speck	Over
0.04	5	15	nuclear body	Over
0.04	14	65	mitochondrial membrane	Over
0.05	14	66	mitochondrial envelope	Over
0.05	2	3	clathrin sculpted vesicle	Over
0.05	2	3	plasma membrane respiratory chain complex I	Over
0.05	2	3	plasma membrane respiratory chain	Over
0.05	2	3	basement membrane	Over
0.05	2	3	plant-type cell wall	Over
0.00	0	37	large ribosomal subunit	Under
0.01	0	35	cytosolic large ribosomal subunit	Under
0.01	28	280	nucleus	Under
0.02	19	201	non-membrane-bounded organelle	Under
0.02	19	201	intracellular non-membrane-bounded organelle	Under
0.02	4	71	nucleolus	Under
0.02	3	60	cytosolic ribosome	Under
0.02	1	38	plastid	Under
0.03	4	68	cytosolic part	Under
0.03	1	36	chloroplast	Under
0.05	5	73	ribosome	Under

Table 5.4: Over- and under-representation of GO-terms in positively selected - GO-terms over- or under-represented (direction) in contigs putatively under positive selection. Horizontal lines separate categories of the GO-ontology. First category is molecular function, second biological process, last cellular compartment. P values (Pval) for over- or under-representation are given along with the number of positively selected contigs (Count; dn/ds > 0.5) and the number of contigs with this annotation for which a dn/ds was obtained (Size) and the description of the GO-term (Term).

While the biological process and cellular compartment categories provide less information for a nematode (highlighting e.g. brain or pancreas development), underrepre-

sented terms in both were connected to ribosomal proteins, validating the analysis for the molecular function category.

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

At both bitscore thresholds contigs novel in clade III and novel in *A. crassus* had a significantly higher dn/ds than other contigs (novel.in.metazoa - novel.in.Ac, 0.005 and 0.015; novel.in.nematoda - novel.in.Ac, 0.005 and 0.002; novel.in.nematoda - novel.in.clade3, 0.207 and 0.045; comparison, p-value from bitscore of 50 and p-value from bitscore of 80, Nemenyi-Damico-Wolfe-Dunn test, given only for significant comparisons; figure 5.6).

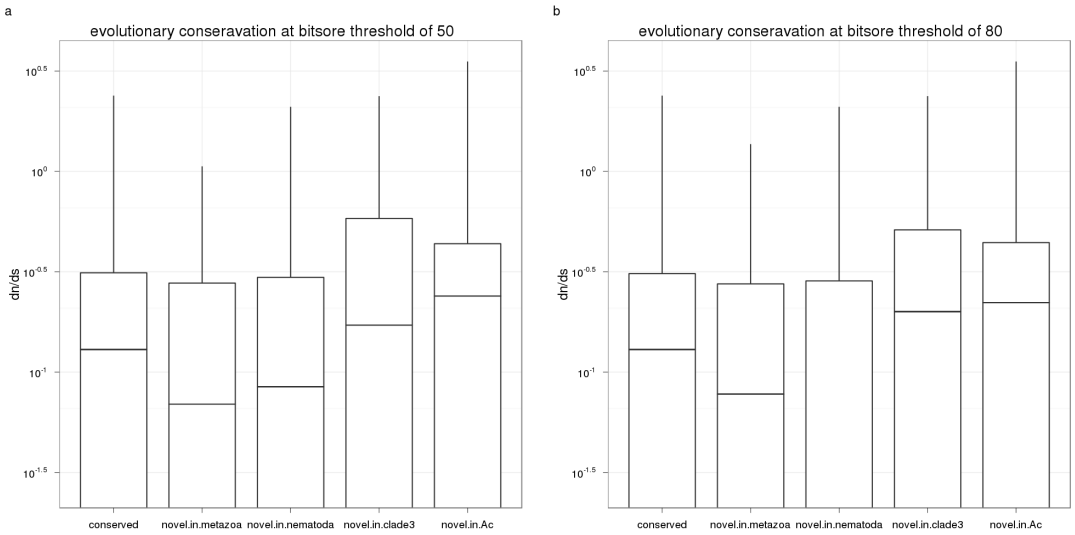

Figure 5.6: Positive selection and evolutionary conservation - Box-plots for dn/ds in TUGs according to different categories of evolutionary conservation. Significant comparisons are novel.in.metazoa - novel.in.Ac (0.005 and 0.015), novel.in.nematoda - novel.in.Ac (0.005 and 0.002), novel.in.nematoda - novel.in.clade3 (0.207 and 0.045; p-value for bitscore of 50 and 80, Nemenyi-Damico-Wolfe-Dunn test).

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

5.10 SNP markers for single worms

We used `Samtools`(183) and `Vcftools`(163) to call genotypes in single worms (adult sequencing libraries). This resulted in 199 informative sites in 152 contigs, where two alleles were found in at least one assured genotype at least in one of the worms.

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

Table 5.5: Measurements of multi-locus heterozygosity for single worms - Genotyping for a set of 199 SNPs, different measurements were obtained to asses genome-wide heterozygosity. Measurements for relative heterozygosity (rel.het; number of homozygous sites/ number of heterozygous sites), internal relatedness (int.rel; (189)), homozygosity by loci (ho.loci; (190)) and standardized heterozygosity (std.het; (191)) are given. All these measurements are pointing to sample T1 (Taiwanese worm from a wild population) as the most heterozygous and sample E2 (the European worm from Poland) as the least heterozygous individual. Heterozygote-heterozygote correlation (192) confirmed the genome-wide significance of these markers.

Internal relatedness (189), homozygosity by loci (190) and standardised heterozygosity (191) were all highlighting the Taiwanese worm from the wild population (sample T1) as the most and the European worm from Poland (sample E2) as the least heterozygous individual. The other worms had intermediate values between these two extremes (see table 5.5).

We confirmed the genome-wide significance of these estimates using heterozygosity-heterozygosity correlation (192). These tests confirmed the representativeness of the 199 SNP-markers for the whole genome in population genetic studies ($\mu = 0.78$, ci_l=0.444; $\mu = 0.86$ and $ci_l = 0.596$; $\mu = 0.87$ and ci_l= 0.632; mean and lower bound of 95% confidence intervals from 1000 bootstrap replicates for internal relatedness, homozygosity by loci and standardised heterozygosity). Using a higher number of genotyped individuals these markers would allow to asses the amount of inbreeding in populations of *A. crassus*.

5.11 Differential expression

Using methods developed for sequencing data in the R-package DESeq (168), we analysed gene-expression inferred from mapping. Of the 353055 reads 252388 (71.49%) mapped uniquely (with their best hit) to the fullest assembly (including the all assembled contigs as a "filter" later removing screened out sequences for analysis).

22 Contigs were clearly over-expressed at adjusted p-values < 0.001 in the male worm compared to the female worms. 7 of these were annotated as "Major sperm protein", all with different orthologs in *B. malayi* and with 5 different orthologs in *C. elegans*. Other annotation were "putative P40" for "Contig2545" and "Contig94", "MFP2" for "Contig437", "KH domain containing protein" for "Contig96", "Protein kinase domain containing protein" for "Contig96", "Protein-tyrosine phosphatase containing protein" for "Contig382" and "PDZ domain containing protein" for "Contig106". For all these contigs read-sums for orthologous groups in the two model nematodes also showed significant over-expression in the same direction.

An exception was "Contig194" ("Phosphoenolpyruvate carboxykinase"), for which the sums of orthologous-counts for both model-nematodes did not show a significant over-expression in the male.

Comparing libraries from European worms to Taiwanese worms, two contigs were observed at adjusted p-values < 0.001: "Contig5250" was only expressed in library E2 and was annotated as an ortholog of alpha tubulin in both *C. elegans* and *Brugia malayi*. This annotation was shared with 18 and 16 contigs respectively. Contrary to the identified contig, the sums of the read counts for this orthologous groups showed a non-significant trend towards higher expression in the the Asian libraries. "Contig13931"was only expressed in sample E1 and showed high similarity to *Ascaris lumbricoides* small nuclear RNAs.

Two contigs were identified as differentially expressed at adjusted p-values below 0.2 and were expressed highly in library E2: "Contig5320" shared annotation as ortholog of "Cuticle collagen 7" in both model-nematodes with 5 contigs. Counts for the orthologous groups showed a non-significant opposite trend of over-expression in the in the Taiwanese libraries. "Contig200" shared similarity only with one predicted protein from *C. elegans*.

This was followed by a set of 9 contigs over-expressed at exactly 20% false discovery rate. They were all found only in Taiwanese libraries and the ortholog-groups showed a tendency towards expression in the same direction. Three of these are worth further notice: "Contig5164" annotated as "galactoside binding lectin" was not only expressed in one of the Taiwaneselibraries, but in both. "Contig110" ("cysteine protease") had

adjusted p-values < 1 and "Contig6355" ("Trypsin family protein") nearly significant adjusted p-values < 0.4 for orthologous groups in *C. elegans* and *B. malayi*.

6

Transcriptomic divergence in a common garden experiment

6.1 Infection experiments

Dissection of eels 55-57 dpi showed higher recovery of European worms in *An. anguilla* and higher recovery of Taiwanes worms in *An. japonica*, compared to the other parasite populations. In other words, in host-parasite combinations of matching origin, parastites performed better.

In host-species/parastite-population pairs found in nature roughly eight or nine adult worms could be recoverd per eel. In the transplanted host/parasite combinations only two or three adult worms were recovered on average (see figure 6.1). In *An. anguilla* no differences in the recovery of larval stages was recorded. In *An. japonica* however roughly two individals more were recorded from both larval stages in the host/parasite combination found in nature.

These differences are highly significant especially for adult worms (see table 6.1) and are interpretable as a sign of local adaptation, as adult survival and recovery can be regarded a fitness component.

Recovery as a proportion of the 50 larvae eels were inocculated with, was thus roughly 30% for the adapted pairs compared to only roughly 10% in non-adapted host-parasite pairs.

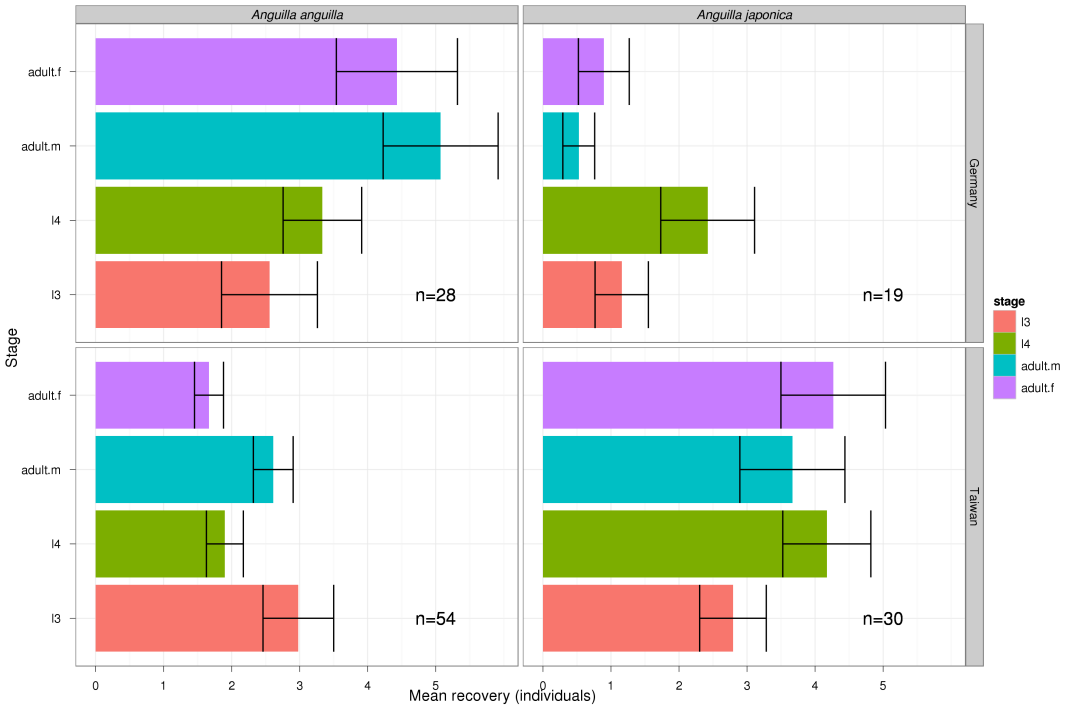

Figure 6.1: Recovery of worms in cross-infection experiment - Mean numbers of worms recovered after 55-57 dpi for sample sizes given as n=x. Error-bars indicate the standard error (s.e.) of the mean. Stages are L3-larvae (l3), L4-larvae (l4), adult females (adult.f) and adult males (adult.m).

	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 6.1: 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 population) further rows give variations for each factor. Std. Error is the standard error of this value. Additonally the probabilty of a t-value as small or smaller then the oberserved t-value are given. The signature of local adaptation is visible in the highly significant interaction term.

6.2 Sample preparation and sequencing

Three biological replicates were obtained from each of the two worm population in each of th two eel-host for each of the two sexes of worms. This resulted in a total 24 RNA-extractions prepared for sequencing: 3 individual female worms from each experimental group were choosen randomly, to give in total twelve females. Additionally from three individual male worms, and from 9 pools of male worms RNA was extracted (see table 6.2). Pools consisted of worms from one infected eel individual each. All worms or worm-pools were derived form infections of different eel individuals, with one small exception from this form of statistical independence: From eel AJ/R3 a male worms as well as a female worm had to be prepared. 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 choosen for preparation based on large size, even when pools of worms were used.

Sequencing was performed in three multiplexed pools of eight libraries each. The samples were partitioned into these pools spreading replicates over lanes in a blocking design to further guarantee statistical independence from sequencing-lane effects. 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.

6.3 Examination of data-quality

Reads were mapped against the fullest assembly (see 4.8) using `BWA` (158). Of the 263.668.952 raw read-pairs 173.602.387 mapped uniquely to the assembly and were counted on a per-library base.

Read counts for one library were originally obtained from two different technical replicate lanes of a flow-cell. The technical replicate of read-counts demonstrated very low differences as inferred from a clustering analysis using variance stabilized data and euclidian distances as implemented in `DESeq` (168) (see **??**).

158.232.523 read-pairs were left after removal of hits to contigs for which non-*A. crassus* origin had been inferred in the curration of the 454-transcriptome assembly.

After another screening for spurious read-counts to low covered transcripts and to transcripts of low relieability (lowCA in the 454-assembly; see 4.8) 137.477.156 read-pairs were left for further analysis. Distribution of these read-pairs over libraries showd roughly three fold differences, with a mean of 5.728.215 reads and a range from 3.422.526 read-pais for library AJ_R3M to 9.453.468 read-pairs for library AA_R8F (see 6.3).

These reads mapped to 7520 contigs from our 454 assembly, making them the basis for all further inverstigations. Analysis of between-sample distance confirmed the 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 6.3).

6.4 Orthologous-screened expression differences

For the 7520 contigs with analysed exprssion values 4382 *C. elegans*-orthologs and 4292 *B. malayi*-orthologs with read more than 32 read-counts over all libraries were determined based on the annotation of the assembly (see 5.6). This resulted in 3596 contigs with measured expression also having a measurement for a corresponding orhtolog (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-derived statistics is considered.

6.5 Expression differences in general linear models

Generalized linear models (GLMs) were used as implemented in the R-package `edgeR`. Using these models we obtained 2588 contigs (34% of total) de between male and female worms at a false discovery rate (FDR) of 5%. 1101 (31% of total orthologous available)

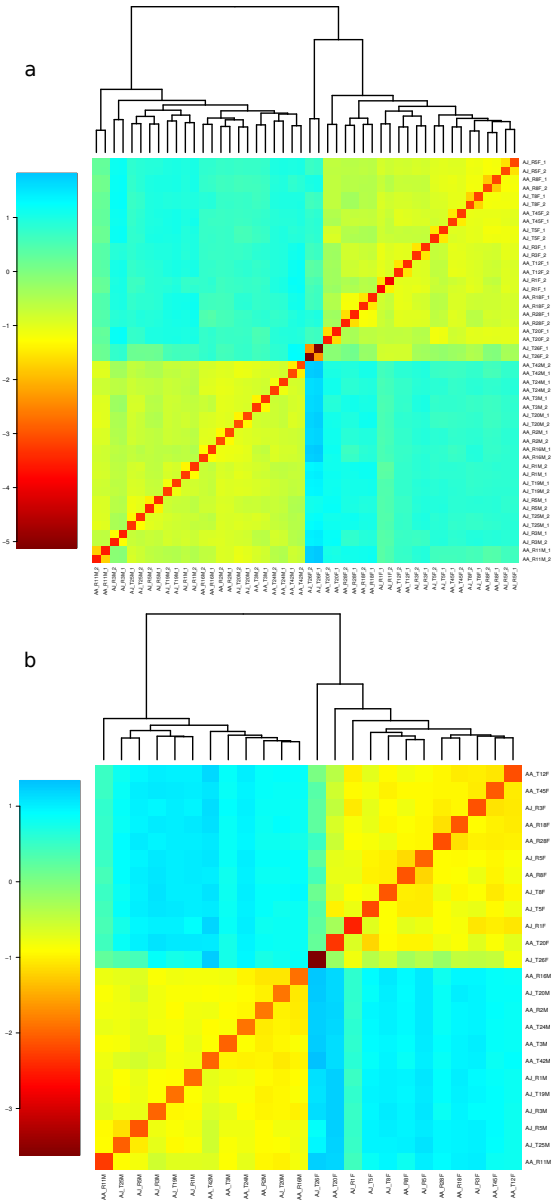

Figure 6.2: Distances between RNA-seq read-count for different samples - Euclidean distance (square distance between the two count vectors) for variance stabilized read-counts for all libraries including technical replicates; Red indicates low distance (high similarity), blue heigh distance (low similarity). a) Data before screening and summation of technical replicates. All technical replicates are clusterd 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.

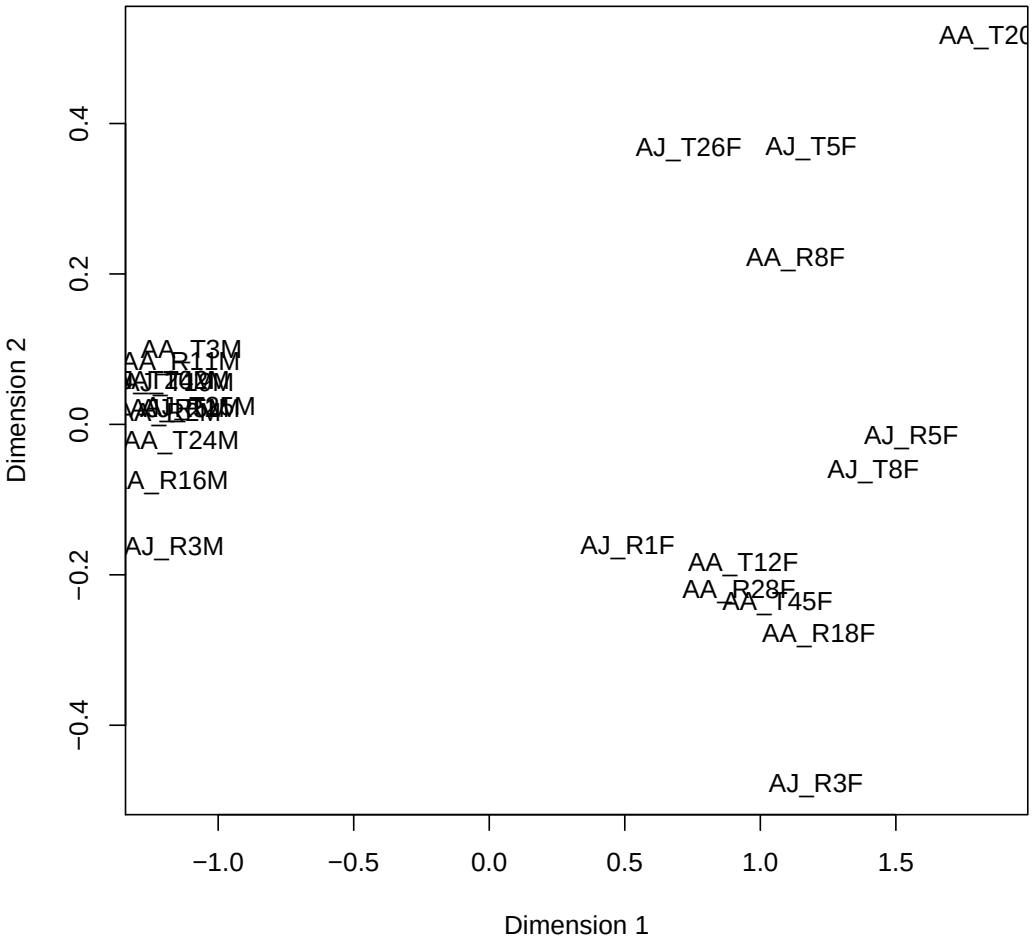

Figure 6.3: Principle coordinate plot for expression inRNA-seq libraries - Distance between sample-pairs is the root-mean-square deviation (Euclidean distance) for the most differentially expressed (DE) genes. Distances can be interpreted as the log2-fold-change of the genes with the biggest changes, i.e. the log2-fold-change for the genes that distinguish the samples.

of these contigs of were confirmed by contigs in the orthologous evaluation. 1425 (556 orthologous confirmed; OC) of these were upregulated in male worms 1163 (545 OC) upregulated 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 upregulated in *An. japonica*, 17 (4 OC) upregulated 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 population.

An important observation in these models is the prevalence of co-occuring 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-populations 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 3 main effects (see figure **??**).

The benefit of also allowing contrasting significant differences in interaction terms highlights the power of the GLM-approach. 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 host-species/parasite-population combinations found in nature.

And indeed also interaction-effects were observed: 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*parastie-population, 1 (0 OC) contig showed significance for the 3-way interaciton (see figure **??**).

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 amount of interaction effects between these two conditions was observed.

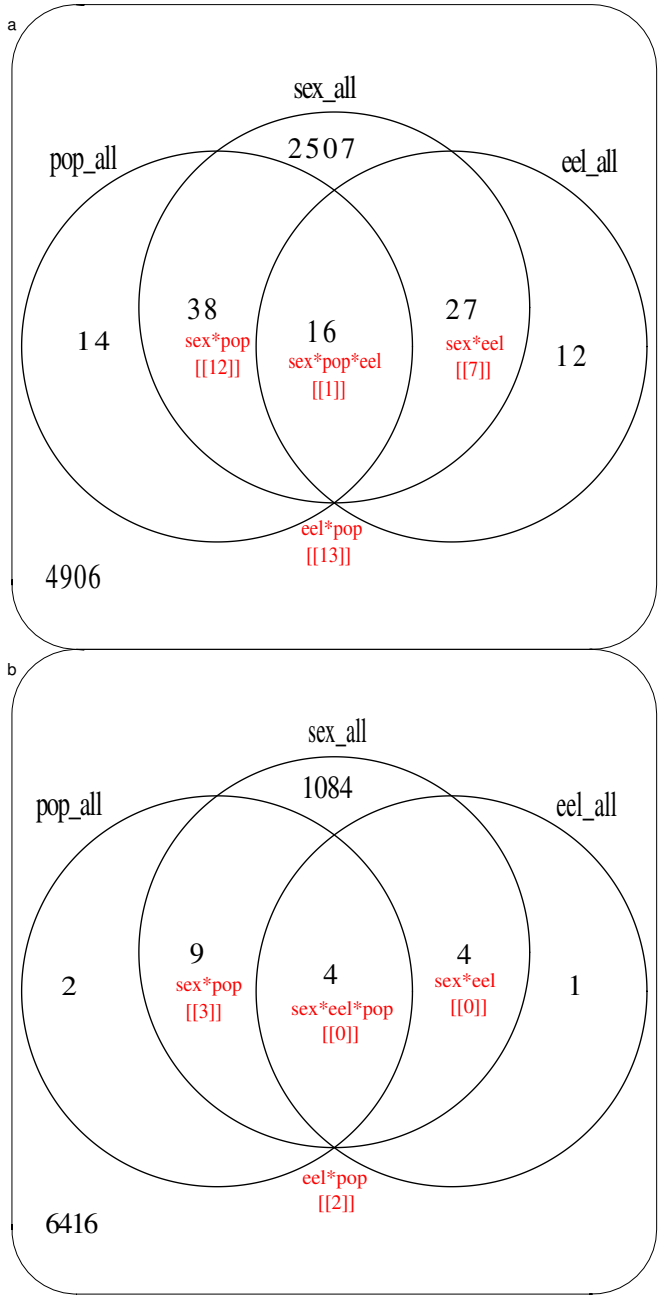

Figure 6.4: Venn diagramm of contigs significant for different terms in `edgeR`-GLMs - Overlap between differences in simple main effects are given as black numbers in the Venn-Diagramm. Numbers outside the circles in the lower left corner idicated non-significant contigs. The number of significant contigs for interaction effects are indicated in red for comparison. In (a) values for all contigs are given in (b) for orhtolog-confirmed (OC) contigs.

6.6 Confirmation of contig categories through multivariate clustering

We 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. We found 7% of the variance in contigs DE between eel-hosts and 11% of the variance in contigs DE between worm-population were explained by the corresponding factor. In both evaluations more than 50% of the remaining variance could be explained by a single principal component, which was mainly consisting of worm-sex (see figure 6.5 a and 6.6 a). When only OC-DE contigs were considered the explained variance for difference between eel-host droped 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 6.5 b and 6.6 b). Significance of the constrained component reported ba a premutation could be established at a $p = 0.05$ threshold for all but the OC eel-host DE subset.

6.7 Biological processes associated with DE contigs

We 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. Summarizing 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 the

6.8 Single gene differences

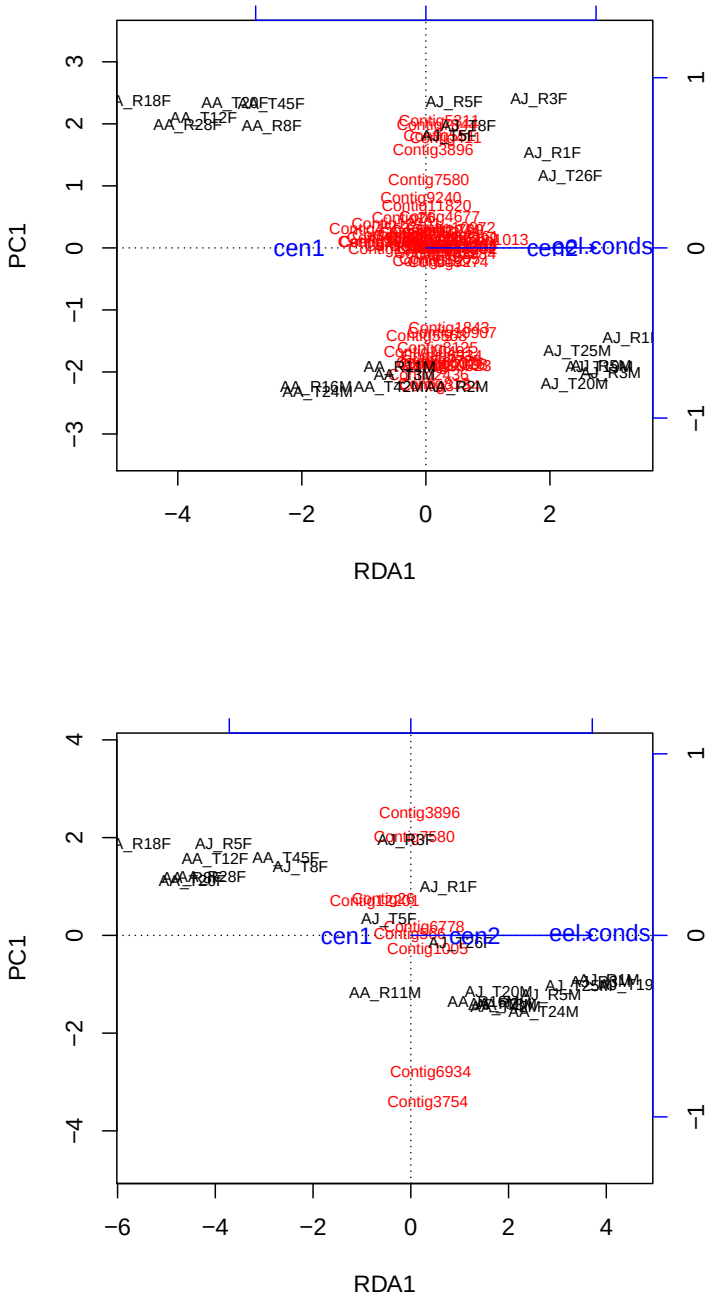

Figure 6.5: Constrained redundancy analysis for host-DE contigs - Eel-host differences are displayed as constrained component on the x-axis, the principal component on the y-axis corresponds to the sex of the worm. (a) Host differences patition the variance in samples in like expcted for all contigs, 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

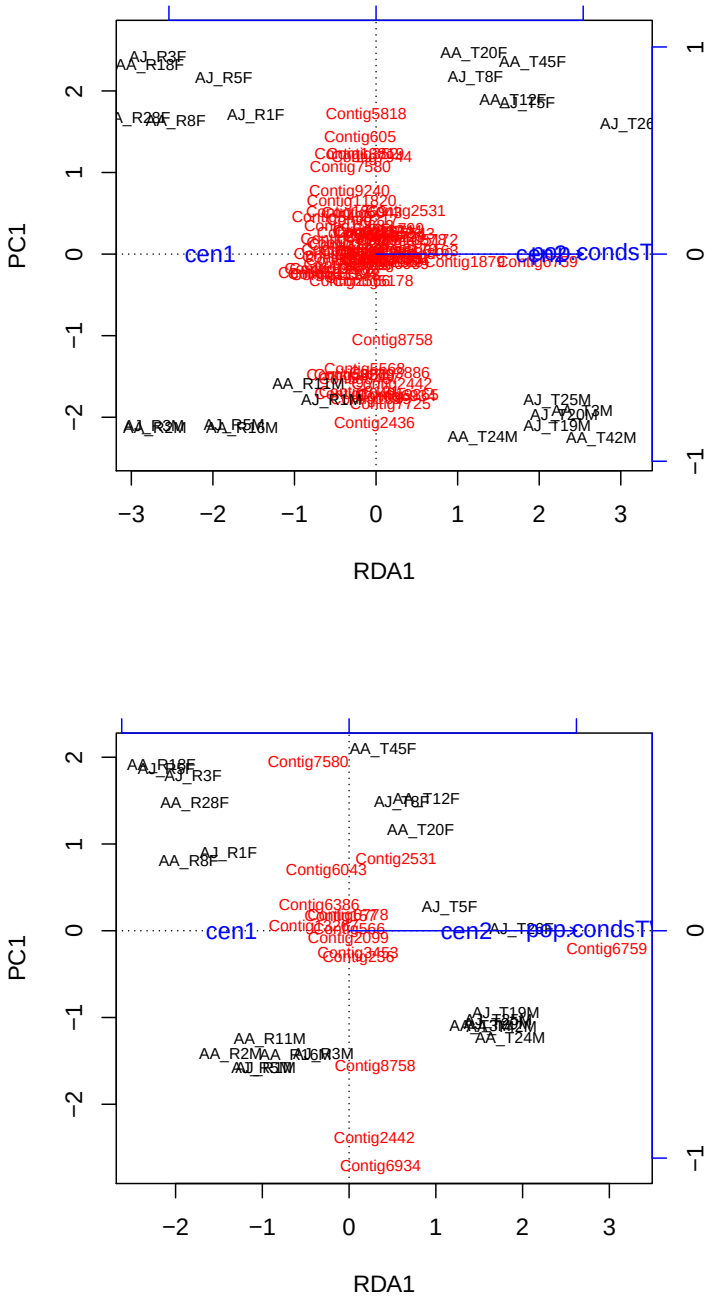

Figure 6.6: Constrained redundancy analysis for population-DE contigs - Population differences are displayed as constrained component on the x-axis, the principal component on the y-axis corresponds to the sex of the worm. Host differences patition the variance in samples like expcted for all contigs (a) as well as for OC-contigs (b). The constrained component showed significance in both subsets.

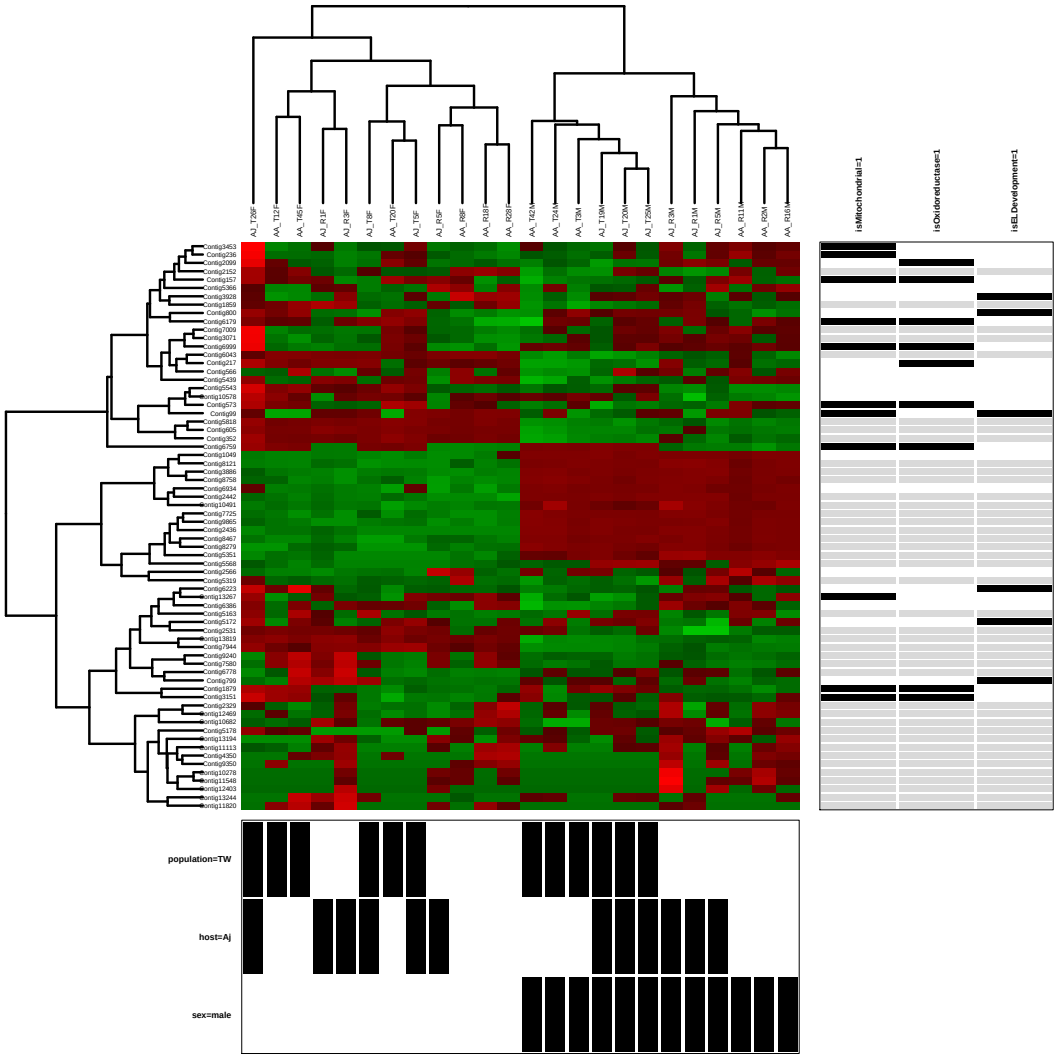

Figure 6.7: Variance/mean stabilized expression values for contigs different between populations - green indicates expression below the mean, red above the mean

label	sex	host	population	intensity	worms in prep	conc in prep
AA/T20F	female	AA	Taiwan	1	1	5.60
AA/T12F	female	AA	Taiwan	14	1	6.80
AA/T45F	female	AA	Taiwan	5	1	8.00
AA/T24M	male	AA	Taiwan	6	3	4.80
AA/T42M	male	AA	Taiwan	11	1	5.60
AA/T3M	male	AA	Taiwan	5	4	4.88
AA/R18F	female	AA	Europe	4	1	4.80
AA/R28F	female	AA	Europe	10	1	5.20
AA/R8F	female	AA	Europe	27	1	5.20
AA/R16M	male	AA	Europe	10	4	5.20
AA/R11M	male	AA	Europe	25	14	6.40
AA/R2M	male	AA	Europe	10	4	6.60
AJ/T8F	female	AJ	Taiwan	10	1	5.91
AJ/T5F	female	AJ	Taiwan	2	1	4.80
AJ/T26F	female	AJ	Taiwan	2	1	2.40
AJ/T25M	male	AJ	Taiwan	24	5	4.05
AJ/T19M	male	AJ	Taiwan	24	7	3.50
AJ/T20M	male	AJ	Taiwan	20	8	3.80
AJ/R1F	female	AJ	Europe	3	1	5.92
AJ/R3F	female	AJ	Europe	3	1	6.90
AJ/R5F	female	AJ	Europe	10	1	4.04
AJ/R1M	male	AJ	Europe	3	1	2.50
AJ/R3M	male	AJ	Europe	3	2	2.60
AJ/R5M	male	AJ	Europe	10	1	2.23

Table 6.2: A summary of all 24 samples prepared for RNA-seq - 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; letter after the slash) and sex of worm(s) in preparation (F for female, M for male; last letter in label). 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 (conc in prep) in the preparation is given in μg per ml

library	raw.reads	raw.mapped	tax.mapped	screened
AA_R11M	11986442	8628520	7868814	6889551
AA_R16M	10810349	6858585	6217540	5276284
AA_R18F	9227615	6552527	5933235	5200958
AA_R28F	10135670	6665381	6005399	5171806
AA_R2M	12469746	7628428	6929651	5906422
AA_R8F	15270570	11527867	10758535	9453468
AA_T12F	11299438	7842479	7195621	6332396
AA_T20F	11740839	7744179	7114349	6323422
AA_T24M	8552723	5254194	4662053	3969305
AA_T3M	11031751	6460836	5800042	4993726
AA_T42M	11573501	7567845	6787375	5694801
AA_T45F	10646847	7714472	7173709	6283585
AJ_R1F	9855005	6400558	5890748	5167912
AJ_R1M	10211903	5851063	5313544	4506254
AJ_R3F	9897937	6425201	5948079	5124077
AJ_R3M	8775211	4562324	4073621	3422526
AJ_R5F	11949105	8442537	7830247	6882280
AJ_R5M	11231532	7504494	6772010	5913016
AJ_T19M	9195576	4798404	4293123	3635843
AJ_T20M	10862591	6880937	6251674	5280529
AJ_T25M	11195315	7162880	6480185	5645097
AJ_T26F	11195335	7439917	6641973	6031374
AJ_T5F	10357569	7413685	6794507	6007930
AJ_T8F	14196382	10275074	9496489	8364594

Table 6.3: Maping summarized for all 24 libraries - Rows indicate different libraries (worms or worm-pools as indicated in 6.2) raw.reads gives the number of read-paires sequenced, raw.mapped the number of reads mapping uniquely with their best hit, tax.mapped the number of reads after substraction of reads to putatively eel-host derived contigs and screened after substraction of all reads mapping not to the higCA-derived assembly or to contigs with overall counts less than 32.

GO.ID	Term	Annotated	Significant
Molecular function			
GO:0042578	phosphoric ester hydrolase activity	99	59
GO:0016791	phosphatase activity	88	53
GO:0004721	phosphoprotein phosphatase activity	65	42
GO:0004722	protein serine/threonine phosphatase act...	34	24
GO:0005509	calcium ion binding	78	43
GO:0046873	metal ion transmembrane transporter acti...	32	21
GO:0003824	catalytic activity	1354	482
GO:0016614	oxidoreductase activity, acting on CH-OH...	46	27
GO:0016616	oxidoreductase activity, acting on the C...	42	25
GO:0017018	myosin phosphatase activity	10	9
Biological process			
GO:0050896	response to stimulus	1535	583
GO:0006470	protein dephosphorylation	63	41
GO:0007391	dorsal closure	32	25
GO:0016476	regulation of embryonic cell shape	13	13
GO:0001700	embryonic development via the syncytial ...	49	33
GO:0007392	initiation of dorsal closure	15	14
GO:0046664	dorsal closure, amnioserosa morphology c...	15	14
GO:0016311	dephosphorylation	86	49
GO:0042221	response to chemical stimulus	864	337
GO:0007394	dorsal closure, elongation of leading ed...	11	11
Cellular compartment			
GO:0031224	intrinsic to membrane	372	164
GO:0016021	integral to membrane	368	162
GO:0005576	extracellular region	250	115
GO:0031226	intrinsic to plasma membrane	176	86
GO:0005887	integral to plasma membrane	172	84
GO:0030054	cell junction	145	72
GO:0000267	cell fraction	435	179
GO:0016020	membrane	1154	417
GO:0000164	protein phosphatase type 1 complex	14	12
GO:0072357	PTW/PP1 phosphatase complex	14	12

Table 6.4: GO-terms enriched in DE between male and female worms - The top 10 enriched GO-categories are given for genes DE between the different male and female worms.

GO.ID	Term	Annotated	Significant	Expecte
Molecular function				
GO:0004190	aspartic-type endopeptidase activity	7	2	0.0
GO:0070001	aspartic-type peptidase activity	7	2	0.0
GO:0030248	cellulose binding	1	1	0.0
GO:0030600	feruloyl esterase activity	1	1	0.0
GO:0052689	carboxylic ester hydrolase activity	27	2	0.1
GO:0045505	dynein intermediate chain binding	2	1	0.0
GO:0016788	hydrolase activity, acting on ester bond...	193	4	0.9
GO:0016787	hydrolase activity	604	7	2.8
GO:0030235	nitric-oxide synthase regulator activity	3	1	0.0
GO:0044183	protein binding involved in protein fold...	3	1	0.0
Biological process				
GO:0002478	antigen processing and presentation of e...	7	2	0.0
GO:0019886	antigen processing and presentation of e...	7	2	0.0
GO:0019884	antigen processing and presentation of e...	8	2	0.0
GO:0002495	antigen processing and presentation of p...	9	2	0.0
GO:0002504	antigen processing and presentation of p...	9	2	0.0
GO:0048002	antigen processing and presentation of p...	13	2	0.0
GO:0019882	antigen processing and presentation	15	2	0.0
GO:0008219	cell death	406	7	2.1
GO:0016265	death	406	7	2.1
GO:0048102	autophagic cell death	19	2	0.1
Cellular compartment				
GO:0005768	endosome	109	4	0.4
GO:0043230	extracellular organelle	2	1	0.0
GO:0065010	extracellular membrane-bounded organelle	2	1	0.0
GO:0070062	extracellular vesicular exosome	2	1	0.0
GO:0043025	neuronal cell body	105	3	0.4
GO:0000323	lytic vacuole	106	3	0.4
GO:0044297	cell body	109	3	0.4
GO:0000328	fungal-type vacuole lumen	3	1	0.0
GO:0061200	clathrin sculpted gamma-aminobutyric aci...	3	1	0.0
GO:0061202	clathrin sculpted gamma-aminobutyric aci...	3	1	0.0

Table 6.5: GO-terms enriched in DE between eel-hosts - The top 10 enriched GO-categories are given for genes DE between the different eel-hosts.

GO.ID	Term	Annotated	Significant
Molecular function			
GO:0016491	oxidoreductase activity	189	9
GO:0004129	cytochrome-c oxidase activity	17	3
GO:0015002	heme-copper terminal oxidase activity	17	3
GO:0016676	oxidoreductase activity, acting on a hem...	17	3
GO:0016616	oxidoreductase activity, acting on the C...	42	4
GO:0004622	lysophospholipase activity	4	2
GO:0016675	oxidoreductase activity, acting on a hem...	19	3
GO:0016614	oxidoreductase activity, acting on CH-OH...	46	4
GO:0004607	phosphatidylcholine-sterol O-acyltransfe...	5	2
Biological process			
GO:0034186	apolipoprotein A-I binding	5	2
GO:0046688	response to copper ion	25	4
GO:0006123	mitochondrial electron transport, cytoch...	11	3
GO:0010035	response to inorganic substance	233	9
GO:0010038	response to metal ion	182	8
GO:0008202	steroid metabolic process	64	5
GO:0034370	triglyceride-rich lipoprotein particle r...	4	2
GO:0034372	very-low-density lipoprotein particle re...	4	2
GO:0009408	response to heat	76	5
GO:0009266	response to temperature stimulus	117	6
Cellular compartment			
GO:0034375	high-density lipoprotein particle remode...	5	2
GO:0034364	high-density lipoprotein particle	4	2
GO:0032994	protein-lipid complex	5	2
GO:0034358	plasma lipoprotein particle	5	2
GO:0031090	organelle membrane	505	11
GO:0044421	extracellular region part	174	6
GO:0005576	extracellular region	250	7
GO:0005739	mitochondrion	605	11
GO:0005743	mitochondrial inner membrane	162	5
GO:0031967	organelle envelope	313	7
GO:0031975	envelope	314	7

Table 6.6: GO-terms enriched in DE between wrom-populations - The top 10 enriched GO-categories are given for genes DE between the different worm populations.

	Aa:EU	Aa:TW	Aj:EU	Aj:TW
Contig1005.mean	518.35	630.47	1512.31	831.26
Cytochrome P450 family protein	1123.86	1204.98	2647.29	1620.76
ce.ortho.mean	557.65	662.20	1658.80	1004.08
Contig12201.mean	514.90	549.58	116.02	99.56
Lipase family protein	502.48	553.48	119.47	101.09
ce.ortho.mean1	501.19	549.00	119.20	99.67
Contig26.mean	11007.58	5406.06	3206.43	2541.48
Aspartic protease BmAsp-1, identical	12994.14	7671.50	4466.98	4926.97
ce.ortho.mean2	12670.54	7237.48	4206.98	4402.80
Contig3754.mean	490.23	901.35	922.95	663.19
MGC79044 protein, putative	660.74	1110.31	1180.48	884.49
ce.ortho.mean3	488.55	883.91	971.48	682.95
Contig3896.mean	123.17	85.71	109.09	60.18
Transcription factor AP-2 family protein	119.36	86.89	111.08	59.46
ce.ortho.mean4	119.08	85.79	111.17	58.87
Contig566.mean	642.74	484.47	337.05	691.06
Eukaryotic aspartyl protease family protein	651.38	496.17	377.95	733.26
ce.ortho.mean5	654.89	491.93	381.14	724.47
Contig6778.mean	39.00	768.10	1028.40	92.46
Nematode cuticle collagen N-terminal domain containing protein	621.79	1259.66	1508.45	447.50
ce.ortho.mean6	38.62	752.61	1056.15	95.26
Contig6934.mean	449.66	639.22	632.23	572.12
Serine/threonine-protein phosphatase	788.16	1133.91	1236.79	1041.83
ce.ortho.mean7	448.17	628.16	663.55	591.01
Contig7580.mean	240.34	1318.57	2215.65	38.30
Cuticular collagen Bmcol-2	286.57	1490.40	2531.07	227.23
ce.ortho.mean8	231.55	1298.61	2272.71	38.23

	Aa:EU	Aa:TW	Aj:EU	Aj:TW
Contig13267.mean	103.86	38.57	111.01	83.54
ABC transporter family protein	101.36	37.67	114.79	94.25
ce.ortho.mean	101.74	37.76	115.19	89.28
Contig157.mean	362.46	394.14	369.26	449.27
Probable 3-hydroxyacyl-CoA dehydrogenase B0272.3, putative	361.60	378.14	381.70	545.36
ce.ortho.mean1	362.40	367.51	380.95	504.83
Contig2099.mean	289.41	327.82	367.54	556.00
Malate/L-lactate dehydrogenase family protein	316.68	360.99	418.67	754.71
ce.ortho.mean2	319.36	357.47	421.73	699.56
Contig236.mean	266.65	164.76	183.18	840.76
Lecithin:cholesterol acyltransferase family protein	2797.98	2969.10	2306.91	6119.67
ce.ortho.mean3	2716.28	2886.46	2225.58	5278.32
Contig2442.mean	284.39	360.83	521.53	408.18
Putative uncharacterized protein	782.07	1102.11	1432.12	960.61
ce.ortho.mean4	797.22	1131.03	1448.22	970.06
Contig2531.mean	21.38	53.89	25.65	35.20
Cutical collagen 6, putative	20.78	52.54	26.07	37.82
ce.ortho.mean5	20.86	51.95	26.08	36.53
Contig3453.mean	269.89	209.33	277.53	1032.13
Lecithin:cholesterol acyltransferase family protein1	2797.98	2969.10	2306.91	6119.67
ce.ortho.mean6	2716.28	2886.46	2225.58	5278.32
Contig566.mean	642.74	484.47	337.05	691.06
Eukaryotic aspartyl protease family protein	651.38	496.17	377.95	733.26
ce.ortho.mean7	654.89	491.93	381.14	724.47
Contig6043.mean	1003.44	841.34	942.26	631.00
Putative uncharacterized protein1	977.73	834.03	964.85	670.11
ce.ortho.mean8	978.45	823.82	967.65	647.85
Contig6386.mean	68.17	31.29	68.01	48.09
Matrixin family protein	66.79	30.60	69.64	53.52
ce.ortho.mean9	72.76	36.38	72.47	55.31
Contig6759.mean	47.39	12737.30	115.48	28013.11
Cytochrome c oxidase subunit 2	5647.97	19163.28	9116.07	43335.23

7

Discussion

7.1 Overview

7.2 Sanger-method pilot-sequencing

In was not achieved to reproducibly alleviate the rRNA-levels in libraries prepared for sequencing. 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 reaction conditions did not allow specific amplification of mRNA from a rRNA background. 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 (193) and reverse transcription (194).

Nevertheless the stringent quality trimming and processing of raw reads, as summarized in 3, made the remaining ESTs a valuable resource for comparison with future 454-sequencing-data.

In fact all sequenced ESTs, for which host-orgin was inferrred were later found also in pyrosequencing: The observation hemoglobin and ferritin subunits form *An. anguilla* are expected, as fish erythrocytes contain a nucleus and still transcribe genes actively (195) and these are typical proteins for the functioning of red blood cells. The overservation of fish cyclin G1 (in Sanger and pyrosequencing) and cohesin (in Sanger sequencing) is remarkable as fish erythrocytes are thought to exhibit low rates of mitosis (196). Other obersvation of host-sequences like e.g. Leukocyte cell-derived chemotaxin 2 or natural killer cell-enhancing factor (NKEF)-B protein in pyrosequencing make an analysis of this fish-derived off-target data (form all sequencing technologies) a very

promising, it is howerver beyond the scope of the present thesis.

7.3 454-pyrosequencing

We have generated a de novo transcriptome for *A. crassus* an important invasive parasite that threatens wild stocks of the European eel *An. anguilla*. These data enable a broad spectrum of molecular research on this ecologically and economically important parasite. As *A. crassus* lives in close association with its host, we have used exhaustive filtering to attempt to remove all host-derived, and host-associated organism-derived contamination from the data. To do this we have also generated a transcriptome dataset from the definitive host *An. japonica*. The non-nematode, non-eel data identified, particularly in the L2 sample, showed highest identity to flagellate protists, which may have been parasitising the eel (or the nematode). Encapsulated objects observed in eel swim bladder walls (44) could be due solely to immune attrition of *A. crassus* larvae or to other coinfections.

A second examination of sequence origin was performed after assembly, employing higher stringency cutoffs. Similar taxonomic screening was used in a garter snake transcriptome project (161), and an analysis of lake sturgeon tested and rejected hypotheses of horizontal gene-transfer when xenobiont sequences were identified (197). A custom pipeline for transcriptome assembly from pyrosequencing reads (198) proposed the use of EST3 (199) to infer sequence origin based simply on nucleotide frequency. We were not able to use this approach successfully, probably due to the fact that xenobiont sequences in our data set derive from multiple sources with different GC content and codon usage.

Compared to other NGS transcriptome sequencing projects (200), the combined assembly approach (see 4.1) generated a smaller number of contigs that had lower redundancy and higher completeness. Projects using the mira assembler often report substantially greater numbers of contigs for datasets of similar size (see e.g. (201)), comparable to the mira sub-assembly in our approach. The use of oligo(dT) to capture mRNAs probably explains the bias towards 3' end completeness and a relative lack of true initiation codons in our protein prediction. This bias is near-ubiquitous in deep transcriptome sequencing projects (e.g. (202)).

We were able to obtain high-quality annotations for a large set of TUGs: For 40% of the complete assembly and 60% of our highCA assembly `BLAST`-based annotations could be obtained. 45% of the contigs in the highCA assembly were additionally decorated with domain-based annotations through `InterProScan` (187).

Comparison with complete protein sequence from the genomes of *B. malayi* and *C. elegans* showed a remarkable degree of agreement regarding the occurrence of terms in the two parasitic worms. This agreement was higher than with the free living nematode *C. elegans* and even the two genome-sequencing-derived proteomes showed less agreement with each other than the filarial parasite with our dataset. This implies that our transcriptome is truly a representative partial genome (115) of a parasitic nematode.

Analysis of conservation identified more sequence novel in Nematode than in the eukaryote kingdom or in clade III this is in agreement with prevalence of genic novelty in the Nematoda (123). Furthermore the basal position of *A. crassus* in clade III could be leading to most novelty in the clade not being shared with *A. crassus*.

TUGs predicted to be novel in the phylum Nematoda and novel to *A. crassus* contained the highest proportion of signal-positives. This confirms observations made in a study on *Nippostrongylus brasiliensis* (120), where signal positives were reported as less conserved. Interestingly enrichment of signal sequence bearing TUGs in our dataset was constrained to sequences novel in nematodes and *A. crassus* (i.e. not to the level of clade III). This may be explained, whit two different hypotheses involving the basal position of *A. crassus*: First the signal positives shared with all nematodes could be conserved molecules not excreted by parasites. A different class of secreted/excreted molecules with prominent role in host parasite interactions would not have arisen early in the evolution of parasitism in clade III - or be too fast-evolving - and thus be detected as specific to deeper sub-clades (i.e. to *A. crassus* in our dataset). A second explanation would be, that orthologs of excreted parasite-specific genes could be among those shared with other nematodes and the fewer shared with clade III implying a predisposition to parasitism outside of the Spirurina or even the convergent evolution of secreted molecules in other parasitic nematodes. However analysis of dn/ds (see below) across conservation categories favours the first hypothesis, as it identifies a higher amount of positive selection in TUGs novel to clade III and *A. crassus* than to nematodes.

We generated transcriptome data from multiple *A. crassus* of Taiwanese and European origin, and identified SNPs both within and between populations. Screening of SNPs in or adjacent to homopolymer regions improved overall measurements of SNP quality. The ratio of transitions to transversions (ti/tv) increased. Such an increase is explained by the removal of "noise" associated with common homopolymer errors (140). The value of 1.93 (1.25 outside, 2.41 inside ORFs) is in good agreement with the overall ti/tv of humans (2.16 (203)) or *Drosophila* (2.07 (204)). The ratio of non-synonymous SNPs per non-synonymous site to synonymous SNPs per synonymous site (dn/ds) decreased with removal of SNPs adjacent to homopolymer regions from 0.42 to 0.231 after

full screening. The most plausible explanation is the removal of error, as unbiased error would lead to a dn/ds of 1. While dn/ds is not unproblematic to interpret within populations (205), the assumption of negative (purifying) selection on most protein-coding genes makes lower mean values seem more plausible. We used a threshold value for the minority allele of 7% for exclusion of SNPs, based on an estimate that approximately 10 haploid equivalents were sampled (5 individual worms plus an negligible contribution from L2 larvae in the L2 library and within the female adult worms). The benefit of this screening was mainly a reduction of non-synonymous SNPs in high coverage contigs, and a removal of the dependence of dn/ds on coverage. Working with an estimate of dn/ds independent of coverage, efforts to control for sampling biased by depth (i.e. coverage; see (206) and (200)) could be avoided.

Also in comparison with published intra-species values of dn/ds our final estimate of seems plausible: in transcripts from the female reproductive tract of *Drosophila* dn/ds was 0.15 (207) and 0.21 in the male reproductive tract (208) (although for ESTs specific to the male accessory gland were shown to have a higher dn/ds of 0.47). A pyrosequencing study in the parasitic nematode *Ancylostoma canium* (125) reported dn/ds of 0.3.

When the whole of coding sequences are studied, of which only a small subset of sites can be under diversifying selection, dn/ds of 0.5 has been suggested as threshold for assuming diversifying selection (207) instead of the classical threshold of 1 (209). The use of this threshold for positive selection led to the identification of over-represented of GO-term highlighting very interesting transcripts:

13 peptidases under positive selection (from 43 with a dn/ds obtained) meant an enrichment in the category. All 13 have different orthologs in *B. malayi* and *C. elgans* and are conserved across kingdoms. Despite their conservation peptidases are thought to have acquired new and prominent roles in host-parasite interaction compared to free living organisms: In *A. crassus* a trypsin-like proteinase has been identified thought to be utilised by the tissue-dwelling L3 stage to penetrate host tissue and an aspartyl proteinase thought to be a digestive enzyme in adults (21). The 13 proteinases under positive selection could be the targets of the adaptive immunity developed against *A. crassus* (43, 210), which is often only elicited against subtypes of larvae (211).

The under-representation of ribosomal proteins (term "structural constituent of ribosome") in positive selected contigs is in good agreement with the notion that ribosomal proteins are extremely conserved across kingdoms (212) and should be under under strong negative selection.

Genotyping of individual worms identified a set of 199 SNPs with highest credibility

and a high information content for population-genetic studies. Levels of genome-wide heterozygosity found for the 5 adult worms examined in our study are in agreement with microsattelite data (10) showing reduced heterozygosity in European populations of *A. crassus*.

We were able to use the DESeq (168) to report transcripts significantly differing in expression between male and female worms. This was possible for male worms, despite the fact that no replicated samples were obtained. However only over-expression in the non-repeated samples could be detected, as obviously lack of expression in one sample can't statistically validate under-expression. Genes over-expressed in male *A. crassus* comprise major sperm proteins well known for their high expression in nematode sperm (213).

We developed a method to assess the possible influence of fragmentation of our reference-transcriptome on mapping. Using the annotation with orthologous sequences of *C. elegans* and *B. malayi* it is possible to highlight problems: For example a contig annotated as "Phosphoenolpyruvate carboxykinase" showed significant over-expression in the male. The 8 other contigs with this annotation however, were nearly identical for the middle region of their predicted proteins, had a high amount of homopolymeric regions in their nucleotide sequence and attracted the missing reads mapping from the other libraries. Our orthologous-collapsing method for read-counts following manual inspection, could thus exclude this contig as a false-positive.

In the comparison of European with Asian libraries most differentially expressed contigs could be debunked as false positives with this method. From the 6 remaining contigs one was annotated with only one ortholog ('Contig200"), 3 were lacking any annotation ("Contig4", "Contig5311" and "Contig40"). The only contig with differential expression fully validated by our method ("Contig6355") was annotated as ("Trypsin family protein"). The identification of another proteinase (it is not among those under positive selection) is highlighting the interesting nature of these molecules in *A. crassus*.

Analysis of the expression data made clear what data is required to analyse the differences in European vs. Asian nematodes: As variance is high in outbred individuals from natural populations both replication and depth of analysis have to be high. Furthermore it is important to disentangle the influence of the host and the nematode population e.g. in a co-inoculation experiment. The last approach would also allow the analysis of transcriptomes roughly synchronised for the time of development.

The *A. crassus* transcriptome provides a basis of molecular research on this important species. It further provides insight in the evolution of parasitism complementing the catalogue of available transcriptomic data with a member of the Spirurina phyloge-

netically distant to so far sequenced parasites in this clade.

7.4 Experimental infections

Whith some reservations discussed below our observation of higher recovery of adult worms from locally matching *A. crassus-Anguilla spp.* host-parastite combinations allow the inferrence of local adaptation of different worm populations to host-species.

It has to be emphasized that the observations made in common-garden experiments first and foremost have to be interpreted as phenotypes. An ideally suited phenotype to infer local adaptation would be one with obviouse direct fitness-consequences, a so called fitness-component. Such a fitness-component would ideally be a measurement on a single individual and individual life-time reporductive success would be such an ideal measurement. However the techniques to measure this individual life-time reporductive success have not been established in *A. crassus* and it would be very difficult to do so.

The recovery of certain developmental stages of worms is not an optimal fitness-component. It is a composite measurement the speed of development from previouse lifecycle stages (or speed of migration towards the swimbladder) and of survival. While suvival is surely an important component of the fitness, it is not so clear whether fast development and/or migration to the swimbladder are. It is immaginable that under certain requirements slower development could lead to higher fitness, if it would e.g. allow to develop without attracting the attention of the immune system.

Our results regarding recovery are also not in complete agreement with previouse findings by Weclawski et al. (unpublished; see 1.1.2.3). Similar to our sutdy they found a higher recovery of the European population of worms in the European eel but did not find the complementary result of lower recovery in the Japanese eel for this diverged population. They recorded recovery at slightly different timepoints after infection (25, 50 and 100 dpi).

A slight problem with recovery in the experiments is that this value is a mean measurement over many individuals.

While mRNA abundance (or gene-exprssion in its synonymous meaning) is often closely linked to a genetics basis, the observed value of expression in the expreient constituts a molecular phenotype. Partly this phenotype has a genetic foundation partly it is ifluenced by environmental factors (such as host genotype).

Nevertheless

One of the dangers of genomic data and

Additionally, whereas functional effects are often caused by selec- tion, functional

differences alone do not demonstrate the past or present action of selection.

CHANGE Almost 30 years ago, Gould and Lewontin76 launched a crusade against the adaptive paradigm that functional differences must be adaptive, that is, caused by natural selection. Although their arguments were controversial at the time, they have been highly influential on the community of evolution- ary biologists. As a new generation of biologists with a background in genomics, molecular biology or bioinformatics has taken leadership in the field of genomic evolutionary biology, the old lessons from Gould and Lewontin seem to have been forgotten. It is a desir- able addition to a story of selection to identify possible functional reasons why selection might be acting, but it will never be a method for identifying or verifying selection. CHANGE

adaptationist (214) Nielsen genomics (215)

The orthologous screening method develped in the analysis of differential expression for the possibly fragmented transcriptome assembly uses stringent conditions for the detection significance. Demanding

—

No surprise was the abundance of differential expression between male and female worms. A lage number of genes are known to be even sex-specific, regulating ovulation and spermatogenesis throughout the metazoa (216) and especially in nematodes (217). On top of these sex-specific genes there is large number of genes differently expressed due to differences in metablolism between males and females.

Sex specific genes, especially male specific genes show elevated rates of sequence evolution (208, 218).

A study on divergence between the transcriptome of *Drosophila* species show that interspecific expression divergence is sex dependent (219) and inferred the action of sex-dependent natural selection during species divergence.

in many species male-specific reproductive traits evolve faster than other traits (Eberhard 1985; Civetta and Singh 1998a).

Genes differentially expressed between males and females can have generall highly modifieable expression level.

—

A surprise was the low number of genes detected as differentially between the two host-species. The genetic differences in those host species leading to a different immune response have a big influence on other phenotypes of worms.

Genes identified to differ between populatins can be rather categorized as important in general metabolic processes instead of specific host-parasite intereaction.

Interpro lists 164 entries for *C. elegans* "Nematode cuticle collagen, N-terminal" (IPR002486) and 51 for *B. malayi* (IPR002486).

Aerobic respiration is potential source for oxidative stress providing a steady source of reactive oxigen 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 are generating additional ROS.

An increase in glycogenic enzymes was observed in *Aedes*

8

Materials & methods

8.1 Sampling of worms from wild eels (Sanger- and pyro-sequencing)

Cultured eels were acquired from an aquaculture directly adjacent to Kaoping river (22.6418N; 120.4440E) 15km stream upwards from it's estuary. Wild eel were bought from a fisherman, fishing in the estuary of Kao-Ping river (22.5074N; 120.4220E). All eels were transported to the Institute of Fisheries Science at the National Taiwan University in Taipei in aerated plastic bags, where they were sheltered until dissection.

Eels were decapitated, length (to the nearest 1.0mm) and weight (to the nearest 0.1g) were measured, and sex was determined by visual inspection of the gonads. The swimbladder was opened, adult worms were removed from the lumen with a forceps, their sex was determined, and they were counted. All adult *A. crassus* were preserved in RNAlater(Quiagen, Hilden, Germany) in individual plastic tubes.

Worms from the European eel were sampled in Sniardwy Lake, Poland (53.751959N ,21.730957E) by Urszula Weclawski and from the Linkenheimer Altrhein, Germany (49.0262N; 8.310556E), following a procedure similar to the one described above for worms from Taiwan.

8.2 RNA-extraction and cDNA synthesis for Sanger- and 454-sequencing

Total RNA was extracted from single, whole worms using the RNeasy kit (Quiagen, Hilden, Germany), following the manufacturers protocol. Alternatively parts of the liver of the host species *Anguilla japonica*, which also had been preserved in RNAlater

were used for RNA extraction, following the same protocol.

The Evrogen MINT cDNA synthesis kit (Evrogen, Moscow, Russia) was then used to amplify mRNA transcripts according to the manufacturers protocol. It uses an adapter sequence at 3' the end of a a poly dT-primer for first strand synthesis and adds a second adapter complementary to the bases at the 5' end of the transcripts by terminal transferase activity and template switching. Using these adapters it is possible to specifically amplify mRNA enriched for full-length transcripts.

8.3 Cloning for Sanger-sequencing

The obtained cDNA preparations were undirectionally cloned into TOPO2PCR-vectors (Invitrogen, Carlsbad, USA) and TOP10 chemically competent cells (Invitrogen, Carlsbad, USA) were transformed with this construct. The cells were plated on LB-medium-agarose containing Kanamycin (5mg/ml), xGal (5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside) and IPTG (Isopropyl-β-D-1-thiogalactopyranosid). After 24h of incubation at 36 °C cells were picked into 96-well micro-liter-plates containing liquid LB-medium and Kanamycin (5mg/ml) and incubated for another 24h. Subsequently 2ml of the cells were used as template for amplification of the insert by PCR using the primers

Forward M13F(GTAAAACGACGGCCAGT) and

Reverse M13R(GGCAGGAAACAGCTATGACC)

in a concentration of 10μM. The protocol for PCR cycling is shown

Inital denaturation	94 °C	5min	
Denaturation	94 °C	30s	
Annealing	54 °C	45s	35 cycles
Elongation	72 °C	2min	
Filnal Elongation	72 °C	10min	

Table 8.1: PCR protocol for insert amplification

Amplification products were controlled on gel and cleaned using SAP (Shrimp Alkaline Phosphatase) and ExoI (Exonuclease I). Sequencing reactions were performed using the BigDye-Terminator kit and PCR-primers (forward or reverse) in a concentration of 3.5μM and sequenced on an ABI 3730 DNA Analyzer (Applied Biosystems, Foster City, California, USA). For *A. crassus* the following libraries were prepared:

Ac_197F: Female from Taiwanese aquaculture

Ac_106F: Female from Taiwanese aquaculture

Ac_M175: Male from Taiwanese aquaculture

Ac_FM: Female from Taiwanese aquaculture

Ac_EH1: Same cDNA preparation as Ac_FM, but sequenced by students in a practical

For *Anguilla japonica* the following three libraries:

Aj_Li1: liver of an eel from aquaculture

Aj_Li2: liver of an eel from aquaculture

Aj_Li3: liver of an eel from aquaculture

8.4 Pilot Sanger-sequencing

The original sequencing-chromatographs ("trace-files") were renamed according to the NERC environmental genomics scheme. "Ac" was used as project-identifier for *Anguillicoloides crassus*, "Aj" for *Anguilla japonica*. In *Anguillicoloides* sequences information on the sequencing primer (forward or reverse PCR primer *Anguilla japonica* sequences were all sequenced using the forward PCR primer) was stored in the middle "library"-field, resulting in names of the following form:

- `Ac_[\d|\w]{2,4}(f|r)_\d\d\w\d\d`
- `Aj_[\d|\w]{2,4}_\d\d\w\d\d`

The last field indicates the plate number (two digits), the row (one letter) and the column (two digits) of the corresponding clone. For first quality trimming trace2seq, a tool derived from trace2dbEST (both part of PartiGene (115)) was used, briefly it performs quality trimming using phred(220) and trimming of vector sequences using cross-match(221). The adapters used by the MINT kit were trimmed by supplying them in the vector-file used for trimming along with the TOPO2PCR-vector. After processing with trace2seq additional quality trimming was performed on the produced sequence-files using a custom script. This trimming was intended to remove artificial sequences produced when the sequencing reaction starts at the 3' end of the transcript at the poly-A tail. These sequences typically consist of numerous homo-polymer-runs throughout their length caused by "slippage" of the reaction. The basic perl regular

expression used for this was:

```
/(.*A{5,}|T{5,}|G{5,}|C{5,}.*){$lengthfac,}/g
```

Where `$lengthfac` was set to the length of the sequence devided by 70 and rounded to the next integer. So only one homo-polymer-run of more then 5 bases was allowed per 75 bases.

Sequences were screened for host contamination by a comparison of `BLAST` searches against the version of nempep4 and a fish protein database. Sequences producing better bit scores againt fish proteins than nematode proteins were labeled as host-contamination.

Only the trace-files corresponding to the sequences still regarded as good after this step were processed with trace2dbEST. Additionally to the processing of traces already included in trace2seq sequences were preliminary annotated using `BLAST` versus the NCBI-NR non-redundant protein database and EST-submission-files were produced.

8.5 454-pyro-sequencing

cDNA preparation and sequencing

RNA was extracted from individual adult male and female nematodes and from a population of L2 larvae (Table 1). RNA was reverse transcribed and amplified into cDNA using the MINT-cDNA synthesis kit (Evrogen, Moscow, Russia). For host contamination screening a liver-sample from an uninfected *An. japonica* was also processed. Emulsion PCR was performed for each cDNA library according to the manufacturerâĂŹs protocols (Roche/454 Life Sciences), and sequenced on a Roche 454 Genome Sequencer FLX. All samples were sequenced using the FLX Titanium chemistry, except for the taiwanese female sample T2, which was sequenced using FLX standard chemistry, to generate between 99,000 and 209,000 raw reads. For the L2 larval library, which had a larger number of non-*A. crassus*, non-*Anguilla* reads, we confirmed that these data were not laboratory contaminants by screening Roche 454 data produced on the same run in independent sequencing lanes.

Trimming, quality control and assembly

Raw sequences were extracted in `fasta`-format (with the corresponding qualities files) using `sffinfo` (Roche/454) and screened for adapter sequences of the MINT-amplification-

kit using `cross-match` (221) (with parameters `-minscore 20 -minmatch 10`). `Seqclean` (184) was used to identify and remove poly-A-tails, low quality, repetitive and short (<100 base) sequences. All reads were compared to a set of screening databases using `BLAST` (expect value cutoff E<1e-5, low complexity filtering turned off: -F F). The databases used were (a) a host sequence database comprising an assembly of the *An. japonica* Roche 454 data, a unpublished assembly of *An. anguilla* Sanger dideoxy sequenced expressed sequence tags (made available to us by Gordon Cramb, University of St Andrews) and transcripts from EeelBase (222) a publicly available transcriptome database for the European eel; (b) a database of ribosomal RNA (rRNA) sequences from eel species derived from our Roche 454 data and EMBL-Bank; and (c) a database of rRNA sequences identified in our *A. crassus* data by comparing the reads to known nematode rRNAs from EMBL-Bank. This last database notably also contained xenobiont rRNA sequences. Reads with matches to one of these databases over more than 80% of their length and with greater than 95% identity were removed from the dataset. Screening and trimming information was written back into sff-format using `sfffile` (Roche 454). The filtered and trimmed data were assembled using the combined assembly approach (127): Two assemblies were generated, one using `Newbler v2.6` (142) (with parameters `-cdna -urt`), the other using `Mira v3.2.1` (181) (with parameters `-job=denovo,est,accurate,454`). The resulting two assemblies were combined into one using `Cap3` (182) at default settings and contigs were labeled by whether they derived from both assemblies or one assembly only (for a detailed analysis of the assembly categories see the supporting Methods file).

Evalutaion of the assemblies

The ace-files for all three (two first-order, one second-order) assemblies were interrogated for the fate of single reads. This was used to tabulate the full read-first-order-second-order-associations.

Based on reads shared between clusters we collapsed reads linked by such read-paths, assigned a cluster-id and recorded the size of the cluster.

`Blast` (`blastx -e 1e-5`) was used to search the complete proteomes of *C. elegans* (as present in wormbase v.220) and the complete proteome of *B. malayi* (as present in uniref 100) for the contigs and singtletons of all investigated assemblies. A custom perl-script (provided by S. Kumar) was then used to mask all bases in the database covered. For each sequence in the database the size of the masked region was then determined and statistics were created summarizing the number of database-sequences with any coverage, the number with coverage over 80% of teir sequence-length and the

overall proportion of bases covered.

Post-assembly classification and taxonomic assignment of contigs

After assembly contigs were assessed a second time for host and other contamination by comparing them (using `Blast`) to the three databases defined above, and also to nembase4, a nematode transcriptome database derived from whole genome sequencing and EST assemblies (122, 124). For each contig, the highest-scoring match was recorded as long as it spanned more than 50% of the contig. We also compared the contigs to the NCBI non-redundant nucleotide (NCBI-nt) and protein (NCBI-nr) databases, recording the taxonomy of all best matches with expect values better than 1e-05. TUGs with a best hit to non-Metazoans and to Chordata within Metazoa were additionally excluded from further analysis.

Protein prediction and annotation

Protein translations were predicted from the contigs using `prot4EST` (version 3.0b) (185). Proteins were predicted either by joining single high scoring segment pairs (HSPs) from a `BLAST` search of uniref100 (223), or by `ESTscan` (224), using as training data the *Brugia malayi* complete proteome back-translated using a codon usage table derived from the `BLAST` HSPs, or, if the first two methods failed, simply the longest ORF in the contig. For contigs where the protein prediction required insertion or deletion of bases in the original sequence, we also imputed an edited sequence for each affected contig. Annotations with Gene Ontology (GO), Enzyme Commission (EC) and Kyoto Encyclopedia of Genes and Genomes (KEGG) terms were inferred for these proteins using `Annot8r` (version 1.1.1) (186), using the annotated sequences available in uniref100 (223). Up to 10 annotations based on a `BLAST` similarity bitscore cut-off of 55 were obtained for each annotation set. The complete *B. malayi* proteome (as present in uniref100) and the complete *C. elegans* proteome (as present in wormbase v.220) were also annotated in the same way. `SignalP V4.0` (188) was used to predict signal peptide cleavage sites and signal anchor signatures for the *A. crassus*-transcriptome and similarly again for the proteomes of the tow model-worms. Additionally `InterProScan` (187) (command line utility `iprscan` (version 4.6) with options `-cli -format raw -iprlookup -seqtype p -goterms`) was used to obtain domain based annotations for the high credibility assembly (highCA) derived contigs.

We recorded the presence of a lethal rnai-phenotype in the *C. elegans* ortholog of each TUG using the biomart-interface (225) to wormbase v. 220 through the R-package

`biomaRt` (226).

Single nucleotide polymorphism analysis

We mapped the raw reads against the the complete set of contigs, replacing imputed sequences for originals where relevant, using `ssaha2` (157) (with parameters `-kmer 13 -skip 3 -seeds 6 -score 100 -cmatch 10 -ckmer 6 -output sam -best 1`). From the `ssaha2` output, pileup-files were produced using `samtools` (183), discarding reads mapping to multiple regions. `VarScan` (162) (`pileup2snp`) was used with default parameters on pileup-files to output lists of single nucleotide polymorphisms (SNPs) and their locations. For enrichment analysis of GO-terms we used the R-package `GOstats` (227).

Using `Samtools` (183) (`mpileup -u`) and `Vcftools` (163) (`view -gcv`) we genotyped individual libraries for the list of previously found overall SNPs. Genotype-calls were accepted at a phred-scaled genotype quality threshold of 10. In addition to the relative heterozygosity (number of homozygous sites/number of heterozygous sites) we used the R package `Rhh` (192) to calculate internal relatedness (189), homozygosity by loci (190) and standardized heterozygosity (191) from these data.

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

Gene-expression analysis

Read-counts were obtained from the `bam`-files generated also for genotyping using the R-package `Rsamtoools` (228). TUGs with less than 48 reads over all libraries were excluded from analysis, as diagnostic plot (not shown) indicated a lack of statistical power for lower overall expression. We used the R-package `DESeq` (168) (version 1.6.1) to assess statistical significance of differences in counts according to groups of libraries.

Additionally we collapsed TUGs by their orthologous assignment in *C. elegans* and *B. malayi*. We used the sums of counts for these orthologous-groups to asses the influence of mapping to our potentially fragmented reference. For both model-nematodes fold-change and p-values were obtained the same way than for the contigs and merged with these.

8.6 Transcriptomic divergence in a common garden experiment

8.6.1 Experimental infeciton of eels

An. anguilla were obtained from the Albe-Fishfarm in Haren-RÃijtenbrock, Germany. *An. japonica* were caught at the glass-eel stage in the estuary of Kao-ping 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 infecions with *A. crassus* in both eel-species was confirmed by dissection of 10 individuals of each species.

After an acclimatisation periode 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 (229). During the infection periode water temperature was held constant at 20°C. Eels were kept in 160-liter tanks in groups of 5-10 individuals and continuously provided with fresh, oxygenated water and commercial fish pellets (Dan-Ex 2848, Dana Feed A/S Ltd, Horsens, Denmark).

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 aquacultre in Kao Ping and from 150km further north in Tainan County in 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. They were infected individually in wells of micro-titer plates at an intensity of 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 the L3 were harvested with a tissue potter using a modified procedure developed by (230). 50 L3 were suspended in 100 μl RPMI-1640 medium (Quiagen, Hilden, Germany) and eels were infected.

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), they were immediately immersed in RNAlater (Quiagen, Hilden, Germany).

8.7 RNA exrtaction 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 6.2).

The Paired-End TruSeqTM RNA sample preparation kit (illumina) was followed to build cDNA libraries with insert sizes of roughly 270 bp for paired-end sequencing: Poly-T oligo-attached magnetic beads were used for purification of mRNA and to simultaniously 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 2nd 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. At this step multiple different indexed paired-end adapters were used to enable multiplexing of the 24 different sequening libraries, in 3 pools of 8 samples each. 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 Rd1 SP.

8.8 Mapping and normalistion of read-counts

All sequencing reads were mapped to the fullest 454 assembly (as defined in 4.8; we were including TUGs inferred as host or xenobiont origin as filter) using `BWA` (158) (version 0.5.9-r16; `BWA aln` and `BWA sampe` with default options) and processed with `samtools` (183) (version 0.1.18; `samtools view -uS -q 1`) to only allwo uniquely mapping reads. All reads mapping to host- and xenobiont off-target data were removed during downstream evaluation.

Counts were summed for technical replicates and counts to lowCA-derived contigs were disregarded for statistics on a contigs-base as well as spuriouse read counts to contigs with less than 32 mapping reads in total (see however 8.10 for how these counts were used in further tests of reference fragmentation).

The remaining counts were normalized using `DESeq` (version 1.6.1) and all tables summarizing read-counts are based on these normalized counts. Additionally we obtained "variance stabilized data" in an expression matrix scaled around a mean of 0 and with a standrard deviation of 1 for each gene and library. These data were used in all heatmap and multivariate visualisations.

8.9 Statistical analysis with GLMs

Generalized linear models were used as implemented in `edgeR` (version 2.4.1) (169). In the maximal fitted model expression was regresses on worm-sex, host-species and parasite population, including all their interactions and modeled as negative binomial distributed. The full model thus contained terms $S_i + H_j + P_k + (SH)_{ij} + (SxP)_{ik} + (HxP)_{jk} + (SxHxP)_{ijk}$ where S_i is the effect of the ith sex (male or female), H_j is the effect of the ith host specis (*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 generalized linear models was respected considering (removing) all interaction effects of a main-term (e.g. $(SxP)_{ik}$, $(SH)_{ij}$ and $(SxHxP)_{ijk}$) 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 (231). Differential expression was inferred at a false discovery rate (FDR) of 5% (adjusted p-value of 0.05).

8.10 Count-collapsing for orthologous from two model-species

In order to test the influence of deficiencies (i.e. fragmentation) of the assembly on mapping and read-counts we summed read counts over orthologous sequence in *C. elegans* and *B. malayi*. For this purpose we used all reference contigs (also lowCA-derived contigs to allow inference of fragmental mapping to those, but not contigs of non *A. crassus* origin) Differential expression for these orthologous-counts was analysed the same way as for contigs. Contigs were filtered based on inferrence 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 the two orthologous evaluations.

8.11 Multivarite confirmation of linear models

We 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 of worm-population) the appropriate constrained 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`.

References

[1] A Kuwahara, H Niimi, and H Itagaki. **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*, **23**(5):275–279, 1974. 1

[2] B Sures, K Knopf, and H Taraschewski. **Development of Anguillicola crassus (Dracunculoidea, Anguillicolidae) in experimentally infected Balearic congers Ariosoma balearicum (Anguilloidea, Congridae)**. *Diseases of Aquatic Organisms*, **39**(1):75–8, December 1999. 1

[3] H. Taraschewski. **Hosts and Parasites as Aliens**. *Journal of Helminthology*, **80**(02):99–128, 2007. 1

[4] R. S. Kirk. **The impact of Anguillicola crassus on European eels**. *Fisheries Management & Ecology*, **10**(6):385–394, 2003. 1, 2

[5] Lamia Gargouri Ben Abdallah and Fadhila Maamouri. **Spatio-temporal dynamics of the nematode Anguillicola crassus in Northeast Tunisian lagoons**. *Comptes Rendus Biologies*, **329**(10):785–789, October 2006. 1

[6] Abdechahid Loukili and Driss Belghyti. **The dynamics of the nematode Anguillicola crassus, Kuvahara 1974 in eel Anguilla anguilla (L. 1758) in the Sebou estuary (Morocco)**. *Parasitology Research*, **100**(4):683–686, March 2007. 1

[7] A. Kristmundsson and S. Helgason. **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*, **54**(2):141, 2007. 1

[8] K. Knopf, J. Wuertz, B. Sures, and H. Taraschewski. **Impact of low water temperature on the development of Anguillicola crassus in the final host Anguilla anguilla**. *Diseases of Aquatic Organisms*, **33**:143–149, 1998. 1

[9] R S Kirk, C R Kennedy, and J W Lewis. **Effect of salinity on hatching, survival and infectivity of Anguillicola crassus (Nematoda: Dracunculoidea) larvae**. *Diseases of Aquatic Organisms*, **40**(3):211–8, April 2000. 1

[10] Sébastien Wielgoss, Horst Taraschewski, Axel Meyer, and Thierry Wirth. **Population structure of the parasitic nematode Anguillicola crassus, an invader of declining North Atlantic eel stocks**. *Molecular Ecology*, **17**(15):3478–95, August 2008. 1, 2, 15, 105

[11] Münderle. *Ökologische, morphometrische und genetische Untersuchungen an Populationen des invasiven Schwimmblasen-Nematoden Anguillicola crassus aus Europa und Taiwan*. PhD thesis, University of Karlsruhe, 2005. 2, 4

[12] Pierre Sasal, Horst Taraschewski, Pierre Valade, Henri Grondin, Sébastien Wielgoss, and František Moravec. **Parasite communities in eels of the Island of Reunion (Indian Ocean): a lesson in parasite introduction**. *Parasitology Research*, **102**(6):1343–1350, May 2008. 2, 3

[13] W Neumann. **Schwimblasenparasit Anguillicola bei Aalen**. *Fischer und Teichwirt*, page 322, 1985. 2

[14] H. Koops and F. Hartmann. **Anguillicola-infestations in Germany and in German eel imports**. *Journal of Applied Ichthyology*, **5**(1):41–45, 1989. 2

[15] S. Wielgoss, F. Hollandt, T. Wirth, and A. Meyer. **Genetic signatures in an invasive parasite of Anguilla anguilla correlate with differential stock management**. *J. Fish Biol.*, **77**:191–210, Jul 2010. 2

[16] LT Fries, DJ Williams, and Sken Johnson. **Occurrence of Anguillicola crassus, an exotic parasitic swim bladder nematode of eels, in the Southeastern United States**. *Transactions of the American Fisheries Society*, **125**(5):794–797, 1996. 3

[17] A. M. Barse and D. H. Secor. **An exotic nematode parasite of the American eel**. *Fisheries*, **24**(2):6–10, 1999. 3

[18] Ann M. Barse, Scott A. McGuire, Melissa A. Vinores, Laura E. Eierman, and Julie A. Weeder. **The swimbladder nematode Anguillicola crassus in American eels (Anguilla rostrata) from middle and upper regions of Chesapeake bay**. *Journal of Parasitology*, **87**(6):1366–1370, December 2001. 3

[19] Frantisek Moravec, Kazuya Nagasawa, and Munenori Miyakawa. **First record of ostracods as natural intermediate hosts of Anguillicola crassus, a pathogenic swimbladder parasite of eels Anguilla spp**. *Diseases of Aquatic Organisms*, **66**(2):171–3, September 2005. 3

[20] O.L.M. Haenen, T.A.M. van Wijngaarden, M.H.T. van der Heijden, J. Höglund, J.B.J.W. Cornelissen, L.A.M.G. van Leengoed, F.H.M. Borgsteede, and W.B. van Muiswinkel. **Effects of experimental infections with different doses of Anguillicola crassus (Nematoda, Dracunculoidea) on European eel (Anguilla anguilla)**. *Aquaculture*, **141**(1-2):101–8, July 2006. PMID: 16956057. 3

[21] M Polzer and H Taraschewski. **Identification and characterization of the proteolytic enzymes in the developmental stages of the eel-pathogenic nematode Anguillicola crassus**. *Parasitology Research*, **79**(1):24–7, 1993. 3, 104

[22] D. De Charleroy, L. Grisez, K. Thomas, C. Belpaire, and F. Ollevier. **The life cycle of Anguillicola crassus**. *Diseases of Aquatic Organisms*, **8**(2):77–84, 1990. 3

REFERENCES

[23] J Würtz, K Knopf, and H Taraschewski. **Distribution and prevalence of Anguillicola crassus (Nematoda) in eels Anguilla anguilla of the rivers Rhine and Naab, Germany**. *Diseases of Aquatic Organisms*, **32**(2):137–43, March 1998. 3

[24] K. Thomas, FP Ollevier, et al. **Population biology of Anguillicola crassus in the final host Anguilla anguilla**. *Diseases of aquatic organisms*, 1992. 3

[25] F S Lefebvre and A J Crivelli. **Anguillicolosis: dynamics of the infection over two decades**. *Diseases of Aquatic Organisms*, **62**(3):227–32, December 2004. 3

[26] M Münderle, H Taraschewski, B Klar, C W Chang, J C Shiao, K N Shen, J T He, S H Lin, and W N Tzeng. **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*, **71**(2):101–8, July 2006. 3, 5

[27] K. Thomas and F. Ollevier. **Paratenic hosts of the swimbladder nematode Anguillicola crassus**. *Diseases of Aquatic Organisms*, **13**:165–174, 1992. 3

[28] M. Pietrock and T. Meinelt. **Dynamics of Anguillicola Crassus Larval Infections in a Paratenic Host, the Ruffe (Gymnocephalus Cernuus) from the Oder River on the Border of Germany and Poland**. *Journal of Helminthology*, **76**(03):235–240, 2002. 3, 5

[29] Leszek Rolbiecki. **Can the DAB (Limanda limanda) be a paratenic host of Anguillicola crassus (Nematoda; Dracunculoidea)? The Gulf of Gdańsk and Vistula Lagoon (Poland) example**. *Wiadomości Parazytologiczne*, **50**(2):317–22, 2004. 5

[30] C Székely. **Dynamics of Anguillicola crassus (Nematoda: Dracunculoidea) larval infection in paratenic host fishes of Lake Balaton, Hungary**. *Acta Veterinaria Hungarica*, **43**(4):401–22, 1995. 5

[31] F. Moravec and B. Skorikova. **Amphibians and larvae of aquatic insects as new paratenic hosts of Anguillicola crassus (Nematoda: Dracunculoidea), a swimbladder parasite of eels**. *DISEASES OF AQUATIC ORGANISMS*, **34**:217–222, 1998. 5

[32] M. Schabuss, C.R. Kennedy, R. Konecny, B. Grillitsch, W. Reckendorfer, F. Schiemer, and A. Herzig. **Dynamics and Predicted Decline of Anguillicola Crassus Infection in European Eels, Anguilla Anguilla, in Neusiedler See, Austria**. *Journal of Helminthology*, **79**(02):159–167, 2005. 5, 8

[33] F.W. Tesch. *Der Aal: Biologie und Fischerei*. Paul Parey, 1983. 5

[34] T. Wirth and L. Bernatchez. **Decline of North Atlantic eels: a fatal synergy?** *Proc. Biol. Sci.*, **270**:681–688, Apr 2003. 5

[35] K Knopf and M Mahnke. **Differences in susceptibility of the European eel (Anguilla anguilla) and the Japanese eel (Anguilla japonica) to the swimbladder nematode Anguillicola crassus**. *Parasitology*, **129**(Pt 4):491–6, October 2004. 5, 6

[36] K Knopf. **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*, **80**(2):129–36, June 2006. 5

[37] Matthew J Gollock, Clive R Kennedy, and J Anne Brown. **Physiological responses to acute temperature increase in European eels Anguilla anguilla infected with Anguillicola crassus**. *Diseases of Aquatic Organisms*, **64**(3):223–8, May 2005. 5

[38] A.P. Palstra, D.F.M. Heppener, V.J.T. van Ginneken, C. Sz?kely, and G.E.E.J.M. van den Thillart. **Swimming performance of silver eels is severely impaired by the swim-bladder parasite Anguillicola crassus**. *Journal of Experimental Marine Biology and Ecology*, **352**(1):244–256, November 2007. 5

[39] B. Sures and K. Knopf. **Parasites as a threat to freshwater eels?** *Science*, **304**(5668):209–11, Apr 2004. 5

[40] J WÃijrtz and H Taraschewski. **Histopathological changes in the swimbladder wall of the European eel Anguilla anguilla due to infections with Anguillicola crassus**. *Diseases of Aquatic Organisms*, **14**(39):121–134, 2000. 5

[41] A Beregi, K Molnár, L Békési, and C Székely. **Radiodiagnostic method for studying swimbladder inflammation caused by Anguillicola crassus (Nematoda: Dracunculoidea)**. *Diseases of Aquatic Organisms*, **34**(2):155–60, October 1998. 5

[42] G. Fazio, P. Sasal, C. Da Silva, B. Fumet, J. Boissier, R. Lecomte-Finiger, and H. Monè. **Regulation of Anguillicola crassus (Nematoda) infrapopulations in their definitive host, the European eel, Anguilla anguilla**. *Parasitology*, **135**(-1):1–10, 2008. 5

[43] K Knopf and R Lucius. **Vaccination of eels (Anguilla japonica and Anguilla anguilla) against Anguillicola crassus with irradiated L3**. *Parasitology*, **135**(5):633–40, April 2008. 5, 104

[44] Emanuel Heitlinger, Dominik Laetsch, Urszula Weclawski, Yu-San Han, and Horst Taraschewski. **Massive encapsulation of larval Anguillicoloides crassus in the intestinal wall of Japanese eels**. *Parasites and Vectors*, **2**(1):48, 2009. 5, 6, 102

[45] K. Aarestrup, F. Okland, M. M. Hansen, D. Righton, P. Gargan, M. Castonguay, L. Bernatchez, P. Howey, H. Sparholt, M. I. Pedersen, and R. S. McKinley. **Oceanic spawning migration of the European eel (Anguilla anguilla)**. *Science*, **325**:1660, Sep 2009. 6

[46] M. Kuroki, J. Aoyama, M. J. Miller, T. Yoshinaga, A. Shinoda, S. Hagihara, and K. Tsukamoto. **Sympatric spawning of Anguilla marmorata and Anguilla japonica in the western North Pacific Ocean**. *J. Fish Biol.*, **74**:1853–1865, Jun 2009. 7

[47] T. D. Als, M. M. Hansen, G. E. Maes, M. Castonguay, L. Riemann, K. Aarestrup, P. Munk, H. Sparholt, R. Hanel, and L. Bernatchez. **All roads lead to home: panmixia of European eel in the Sargasso Sea**. *Mol. Ecol.*, **20**:1333–1346, Apr 2011. 7

[48] J. M. Pujolar, G. A. De Leo, E. Ciccotti, and L. Zane. **Genetic composition of Atlantic and Mediterranean recruits of European eel Anguilla anguilla based on EST-linked microsatellite loci**. *J. Fish Biol.*, **74**:2034–2046, Jun 2009. 7

[49] T. Wirth and L. Bernatchez. **Genetic evidence against panmixia in the European eel**. *Nature*, **409**:1037–1040, Feb 2001. 7

[50] S. Palm, J. Dannewitz, T. Prestegaard, and H. Wickstrom. **Panmixia in European eel revisited: no genetic difference between maturing adults from southern and northern Europe**. *Heredity*, **103**:82–89, Jul 2009. 7

[51] J. Dannewitz, G. E. Maes, L. Johansson, H. Wickstrom, F. A. Volckaert, and T. Jarvi. **Panmixia in the European eel: a matter of time.**. *Proc. Biol. Sci.*, **272**:1129–1137, Jun 2005. 7

[52] J. M. Pujolar, D. Bevacqua, F. Capoccioni, E. Ciccotti, G. A. De Leo, and L. Zane. **Genetic variability is unrelated to growth and parasite infestation in natural populations of the European eel (Anguilla anguilla)**. *Mol. Ecol.*, **18**:4604–4616, Nov 2009. 7

[53] S. D. Cote, A. Stien, R. J. Irvine, J. F. Dallas, F. Marshall, O. Halvorsen, R. Langvatn, and S. D. Albon. **Resistance to abomasal nematodes and individual genetic variability in reindeer**. *Mol. Ecol.*, **14**:4159–4168, Nov 2005. 7

[54] J. M. Rijks, J. I. Hoffman, T. Kuiken, A. D. Osterhaus, and W. Amos. **Heterozygosity and lungworm burden in harbour seals (Phoca vitulina)**. *Heredity*, **100**:587–593, Jun 2008. 7

[55] M. Dionne. **Pathogens as potential selective agents in the wild**. *Mol. Ecol.*, **18**:4523–4525, Nov 2009. 7

[56] M. K. Oliver, S. Telfer, and S. B. Piertney. **Major histocompatibility complex (MHC) heterozygote superiority to natural multi-parasite infections in the water vole (Arvicola terrestris)**. *Proc. Biol. Sci.*, **276**:1119–1128, Mar 2009. 7

[57] P. Ilmonen, D. J. Penn, K. Damjanovich, L. Morrison, L. Ghotbi, and W. K. Potts. **Major histocompatibility complex heterozygosity reduces fitness in experimentally infected mice**. *Genetics*, **176**:2501–2508, Aug 2007. 7

[58] K. Mathias Wegner, Martin Kalbe, Joachim Kurtz, Thorsten B. H. Reusch, and Manfred Milinski. **Parasite Selection for Immunogenetic Optimality**. *Science*, **301**(5638):1343, September 2003. 7

[59] DJ CONWAY and SD POLLEY. **Measuring immune selection**. *Parasitology(London. Print)*, **125**:3–16, 2002. 7

[60] C.M.L. Press and Ø. Evensen. **The morphology of the immune system in teleost fishes**. *Fish & Shellfish Immunology*, **9**(4):309–318, 1999. 7

[61] M E Nielsen and M D Esteve-Gassent. **The eel immune system: present knowledge and the need for research**. *Journal of Fish Diseases*, **29**(2):65–78, 2006. 7

[62] B. Star, A. J. Nederbragt, S. Jentoft, U. Grimholt, M. Malmstr?m, T. F. Gregers, T. B. Rounge, J. Paulsen, M. H. Solbakken, A. Sharma, O. F. Wetten, A. Lanzen, R. Winer, J. Knight, J. H. Vogel, B. Aken, O. Andersen, K. Lagesen, A. Tooming-Klunderud, R. B. Edvardsen, K. G. Tina, M. Espelund, C. Nepal, C. Previti, B. O. Karlsen, T. Moum, M. Skage, P. R. Berg, T. Gj?en, H. Kuhl, J. Thorsen, K. Malde, R. Reinhardt, L. Du, S. D. Johansen, S. Searle, S. Lien, F. Nilsen, I. Jonassen, S. W. Omholt, N. C. Stenseth, and K. S. Jakobsen. **The genome sequence of Atlantic cod reveals a unique immune system**. *Nature*, **477**:207–210, Sep 2011. 7

[63] J. Hikima, T. S. Jung, and T. Aoki. **Immunoglobulin genes and their transcriptional control in teleosts**. *Dev. Comp. Immunol.*, **35**:924–936, Sep 2011. 7

[64] S. Kalujnaia, I. S. McWilliam, V. A. Zaguinaiko, A. L. Feilen, J. Nicholson, N. Hazon, C. P. Cutler, and G. Cramb. **Transcriptomic approach to the study of osmoregulation in the European eel Anguilla anguilla**. *Physiol. Genomics*, **31**:385–401, Nov 2007. 7

[65] H. Taraschewski and F. Moravec. **Revision of the genus Anguillicola Yamaguti, 1935 (Nematoda: Anguillicolidae) of the swimbladder of eels, including descriptions of two new species, A. novaezelandiae sp. n. and A. papernai sp. n.** *Folia Parasitol (Praha)*, **35**(2):125–146, 1988. 8

[66] S. Yamaguti. **Studies on the helmith fauna of Japan, part 9. Nematodes of fishes.** *Japanese Journal of Zoology*, **6**, 1933. 8

[67] T.H. JOHNSTON and P.M. Mawson. **Some nematodes parasitic in Australian freshwater fish.** *Transactions of the Royal Society of South Australia*, **64**(2):340–352, 1940. 8

[68] Frantisek Moravec. *Dracunculoid and anguillicoloid nematodes parasitic in vertebrates.* Academia, 2006. 8

[69] Yuki Minegishi, Jun Aoyama, Jun G. Inoue, Masaki Miya, Mutsumi Nishida, and Katsumi Tsukamoto. **Molecular phylogeny and evolution of the freshwater eels genus Anguilla based on the whole mitochondrial genome sequences**. *Molecular Phylogenetics and Evolution*, **34**(1):134–146, 2005. 8

[70] Mark L. Blaxter, Paul De Ley, James R. Garey, Leo X. Liu, Patsy Scheldeman, Andy Vierstraete, Jacques R. Vanfleteren, Laura Y. Mackey, Mark Dorris, Linda M. Frisse, J. T. Vida, and W. Kelley Thomas. **A molecular evolutionary framework for the phylum Nematoda**. *Nature*, **392**(6671):71–75, March 1998. 11

REFERENCES

[71] S.?A. NADLER, R.?A. CARRENO, H. MEJ?A-MADRID, J. ULLBERG, C. PAGAN, R. HOUSTON, and J.-P. HUGOT. **Molecular Phylogeny of Clade III Nematodes Reveals Multiple Origins of Tissue Parasitism**. *Parasitology*, **134**(10):1421–1442, 2007. 11

[72] Martina Wijová, Frantisek Moravec, Ales Horák, and Julius Lukes. **Evolutionary relationships of Spirurina (Nematoda: Chromadorea: Rhabditida) with special emphasis on dracunculoid nematodes inferred from SSU rRNA gene sequences**. *International Journal for Parasitology*, **36**(9):1067–75, August 2006. 11

[73] A Kerner. **The natural history of plants, their forms, growth, reproduction, and distribution**. *Translated by F. W. Oliver.*, 1895. 11

[74] G Bonnier. **Recherches expÃŕrimentales sur lâĂŹadaptation des plants au climat alpin**. *Ann. Scie. Nat. (Bot.)*, **20**:217–358, 1895. 11

[75] O Kaltz and J. A. Shykoff. **Local adaptation in hostâĂŞparasite systems**. *Heredity*, pages 361–âĂŞ370, May 1998. 13

[76] T. A. Mousseau and D. A. Roff. **Natural selection and the heritability of fitness components**. *Heredity*, **59 (Pt 2)**:181–197, Oct 1987. 15

[77] J. N. Thompson, S. L. Nuismer, and R. Gomulkiewicz. **Coevolution and maladaptation**. *Integr. Comp. Biol.*, **42**:381–387, Apr 2002. 15

[78] J.N. Thompson. *The geographic mosaic of coevolution*. University of Chicago Press, 2005. 15

[79] S. L. Nuismer and S. Gandon. **Moving beyond common-garden and transplant designs: insight into the causes of local adaptation in species interactions**. *Am. Nat.*, **171**:658–668, May 2008. 15

[80] F.H.C. Crick. **The biological replication of macromolecules**. In *Symp. Soc. Exp. Biol*, **12**, pages 138–163, 1958. 16

[81] Crick F. **Central dogma of molecular biology**. *Nature*, **226**:1198–1199, Jun 1970. [PubMed:5422595]. 16

[82] M. Lynch. **The lower bound to the evolution of mutation rates**. *Genome Biol Evol*, **3**:1107–1118, 2011. 16

[83] Y. Wan, M. Kertesz, R. C. Spitale, E. Segal, and H. Y. Chang. **Understanding the transcriptome through RNA structure**. *Nat. Rev. Genet.*, **12**:641–655, Sep 2011. 17

[84] H. Guo, N. T. Ingolia, J. S. Weissman, and D. P. Bartel. **Mammalian microRNAs predominantly act to decrease target mRNA levels**. *Nature*, **466**:835–840, Aug 2010. 17

[85] G. Ruvkun. **Molecular biology. Glimpses of a tiny RNA world**. *Science*, **294**:797–799, Oct 2001. 17

[86] G. Dieci, M. Preti, and B. Montanini. **Eukaryotic snoRNAs: a paradigm for gene expression flexibility**. *Genomics*, **94**:83–88, Aug 2009. 17

[87] W. Deng, X. Zhu, G. Skogerb?, Y. Zhao, Z. Fu, Y. Wang, H. He, L. Cai, H. Sun, C. Liu, B. Li, B. Bai, J. Wang, D. Jia, S. Sun, H. He, Y. Cui, Y. Wang, D. Bu, and R. Chen. **Organization of the Caenorhabditis elegans small non-coding transcriptome: genomic features, biogenesis, and expression**. *Genome Res.*, **16**:20–29, Jan 2006. 17

[88] F. H. Crick. **The origin of the genetic code**. *J. Mol. Biol.*, **38**:367–379, Dec 1968. 17

[89] E. Kim, A. Magen, and G. Ast. **Different levels of alternative splicing among eukaryotes**. *Nucleic Acids Res.*, **35**:125–131, 2007. 18

[90] Z. Wang, M. Gerstein, and M. Snyder. **RNA-Seq: a revolutionary tool for transcriptomics**. *Nat. Rev. Genet.*, **10**:57–63, Jan 2009. 18, 23, 27

[91] J. Armengaud. **Proteogenomics and systems biology: quest for the ultimate missing parts**. *Expert Rev Proteomics*, **7**:65–77, Feb 2010. 18

[92] B. Schwanhausser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach. **Global quantification of mammalian gene expression control**. *Nature*, **473**:337–342, May 2011. 18

[93] F. Sanger, S. Nicklen, and A. R. Coulson. **DNA sequencing with chain-terminating inhibitors**. *Proc. Natl. Acad. Sci. U.S.A.*, **74**:5463–5467, Dec 1977. 19

[94] H. Swerdlow and R. Gesteland. **Capillary gel electrophoresis for rapid, high resolution DNA sequencing**. *Nucleic Acids Res.*, **18**:1415–1419, Mar 1990. 19

[95] W. Fiers, R. Contreras, F. Duerinck, G. Haegeman, D. Iserentant, J. Merregaert, W. Min Jou, F. Molemans, A. Raeymaekers, A. Van den Berghe, G. Volckaert, and M. Ysebaert. **Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene**. *Nature*, **260**:500–507, Apr 1976. 19

[96] F. R. Blattner, G. Plunkett, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, and Y. Shao. **The complete genome sequence of Escherichia coli K-12**. *Science*, **277**:1453–1462, Sep 1997. 19

[97] A. Goffeau, B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, E. J. Louis, H. W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, and S. G. Oliver. **Life with 6000 genes**. *Science*, **274**:563–567, Oct 1996. 19

[98] The C. elegans Sequencing Consortium. **Genome sequence of the nematode C. elegans: a platform for investigating biology**. *Science*, **282**:2012–2018, Dec 1998. 19

[99] M.D. Adams, S.E. Celniker, R.A. Holt, C.A. Evans, J.D. Gocayne, P.G. Amanatides, S.E. Scherer, P.W. Li, R.A. Hoskins, R.F. Galle, et al. **The genome sequence of Drosophila melanogaster**. *Science*, **287**(5461):2185, 2000. 19

[100] R. H. Waterston, K. Lindblad-Toh, E. Birney, J. Rogers, J. F. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, P. An, S. E. Antonarakis, J. Attwood, R. Baertsch, J. Bailey, K. Barlow, S. Beck, E. Berry, B. Birren, T. Bloom, P. Bork, M. Botcherby, N. Bray, M. R. Brent, D. G. Brown, S. D. Brown, C. Bult, J. Burton, J. Butler, R. D. Campbell, P. Carninci, S. Cawley, F. Chiaromonte, A. T. Chinwalla, D. M. Church, M. Clamp, C. Clee, F. S. Collins, L. L. Cook, R. R. Copley, A. Coulson, O. Couronne, J. Cuff, V. Curwen, T. Cutts, M. Daly, R. David, J. Davies, K. D. Delehaunty, J. Deri, E. T. Dermitzakis, C. Dewey, N. J. Dickens, M. Diekhans, S. Dodge, I. Dubchak, D. M. Dunn, S. R. Eddy, L. Elnitski, R. D. Emes, P. Eswara, E. Eyras, A. Felsenfeld, G. A. Fewell, P. Flicek, K. Foley, W. N. Frankel, L. A. Fulton, R. S. Fulton, T. S. Furey, D. Gage, R. A. Gibbs, G. Glusman, S. Gnerre, N. Goldman, L. Goodstadt, D. Grafham, T. A. Graves, E. D. Green, S. Gregory, R. Guigo, M. Guyer, R. C. Hardison, D. Haussler, Y. Hayashizaki, L. W. Hillier, A. Hinrichs, W. Hlavina, T. Holzer, F. Hsu, A. Hua, T. Hubbard, A. Hunt, I. Jackson, D. B. Jaffe, L. S. Johnson, M. Jones, T. A. Jones, A. Joy, M. Kamal, E. K. Karlsson, D. Karolchik, A. Kasprzyk, J. Kawai, E. Keibler, C. Kells, W. J. Kent, A. Kirby, D. L. Kolbe, I. Korf, R. S. Kucherlapati, E. J. Kulbokas, D. Kulp, T. Landers, J. P. Leger, S. Leonard, I. Letunic, R. Levine, J. Li, M. Li, C. Lloyd, S. Lucas, B. Ma, D. R. Maglott, E. R. Mardis, L. Matthews, E. Mauceli, J. H. Mayer, M. McCarthy, W. R. McCombie, S. McLaren, K. McLay, J. D. McPherson, J. Meldrim, B. Meredith, J. P. Mesirov, W. Miller, T. L. Miner, E. Mongin, K. T. Montgomery, M. Morgan, R. Mott, J. C. Mullikin, D. M. Muzny, W. E. Nash, J. O. Nelson, M. N. Nhan, R. Nicol, Z. Ning, C. Nusbaum, M. J. O'Connor, Y. Okazaki, K. Oliver, E. Overton-Larty, L. Pachter, G. Parra, K. H. Pepin, J. Peterson, P. Pevzner, R. Plumb, C. S. Pohl, A. Poliakov, T. C. Ponce, C. P. Ponting, S. Potter, M. Quail, A. Reymond, B. A. Roe, K. M. Roskin, E. M. Rubin, A. G. Rust, R. Santos, V. Sapojnikov, B. Schultz, J. Schultz, M. S. Schwartz, S. Schwartz, C. Scott, S. Seaman, S. Searle, T. Sharpe, A. Sheridan, R. Shownkeen, S. Sims, J. B. Singer, G. Slater, A. Smit, D. R. Smith, B. Spencer, A. Stabenau, N. Stange-Thomann, C. Sugnet, M. Suyama, G. Tesler, J. Thompson, D. Torrents, E. Trevaskis, J. Tromp, C. Ucla, A. Ureta-Vidal, J. P. Vinson, A. C. Von Niederhausern, C. M. Wade, M. Wall, R. J. Weber, R. B. Weiss, M. C. Wendl, A. P. West, K. Wetterstrand, R. Wheeler, S. Whelan, J. Wierzbowski, D. Willey, S. Williams, R. K. Wilson, E. Winter, K. C. Worley, D. Wyman, S. Yang, S. P. Yang, E. M. Zdobnov, M. C. Zody, and E. S. Lander. **Initial sequencing and comparative analysis of the mouse genome**. *Nature*, **420**:520–562, Dec 2002. 19

[101] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigo, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu. **The sequence of the human genome**. *Science*, **291**:1304–1351, Feb 2001. 19

[102] M. D. Adams, J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, C. R. Merril, A. Wu, B. Olde, and R. F. Moreno. **Complemen-**

tary DNA sequencing: expressed sequence tags and human genome project. *Science*, **252**:1651–1656, Jun 1991. 19

[103] C. Fields, M. D. Adams, O. White, and J. C. Venter. **How many genes in the human genome?** *Nat. Genet.*, **7**:345–346, Jul 1994. 19

[104] Mark Blaxter. **Caenorhabditis elegans Is a Nematode**. *Science*, **282**(5396):2041–2046, December 1998. 20

[105] M. B. Gerstein, Z. J. Lu, E. L. Van Nostrand, C. Cheng, B. I. Arshinoff, T. Liu, K. Y. Yip, R. Robilotto, A. Rechtsteiner, K. Ikegami, P. Alves, A. Chateigner, M. Perry, M. Morris, R. K. Auerbach, X. Feng, J. Leng, A. Vielle, W. Niu, K. Rhrissorrakrai, A. Agarwal, R. P. Alexander, G. Barber, C. M. Brdlik, J. Brennan, J. J. Brouillet, A. Carr, M. S. Cheung, H. Clawson, S. Contrino, L. O. Dannenberg, A. F. Dernburg, A. Desai, L. Dick, A. C. Dose, J. Du, T. Egelhofer, S. Ercan, G. Euskirchen, B. Ewing, E. A. Feingold, R. Gassmann, P. J. Good, P. Green, F. Gullier, M. Gutwein, M. S. Guyer, L. Habegger, T. Han, J. G. Henikoff, S. R. Henz, A. Hinrichs, H. Holster, T. Hyman, A. L. Iniguez, J. Janette, M. Jensen, M. Kato, W. J. Kent, E. Kephart, V. Khivansara, E. Khurana, J. K. Kim, P. Kolasinska-Zwierz, E. C. Lai, I. Latorre, A. Leahey, S. Lewis, P. Lloyd, L. Lochovsky, R. F. Lowdon, Y. Lubling, R. Lyne, M. MacCoss, S. D. Mackowiak, M. Mangone, S. McKay, D. Mecenas, G. Merrihew, D. M. Miller, A. Muroyama, J. I. Murray, S. L. Ooi, H. Pham, T. Phippen, E. A. Preston, N. Rajewsky, G. Ratsch, H. Rosenbaum, J. Rozowsky, K. Rutherford, P. Ruzanov, M. Sarov, R. Sasidharan, A. Sboner, P. Scheid, E. Segal, H. Shin, C. Shou, F. J. Slack, C. Slightam, R. Smith, W. C. Spencer, E. O. Stinson, S. Taing, T. Takasaki, D. Vafeados, K. Voronina, G. Wang, N. L. Washington, C. M. Whittle, B. Wu, K. K. Yan, G. Zeller, Z. Zha, M. Zhong, X. Zhou, J. Ahringer, S. Strome, K. C. Gunsalus, G. Micklem, X. S. Liu, V. Reinke, S. K. Kim, L. W. Hillier, S. Henikoff, F. Piano, M. Snyder, L. Stein, J. D. Lieb, and R. H. Waterston. **Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project**. *Science*, **330**:1775–1787, Dec 2010. 20

[106] Lincoln D. Stein, Zhirong Bao, Darin Blasiar, Thomas Blumenthal, Michael R. Brent, Nansheng Chen, Asif Chinwalla, Laura Clarke, Chris Clee, Avril Coghlan, Alan Coulson, Peter D'Eustachio, David H. A. Fitch, Lucinda A. Fulton, Robert E. Fulton, Sam Griffiths-Jones, Todd W. Harris, LaDeana W. Hillier, Ravi Kamath, Patricia E. Kuwabara, Elaine R. Mardis, Marco A. Marra, Tracie L. Miner, Patrick Minx, James C. Mullikin, Robert W. Plumb, Jane Rogers, Jacqueline E. Schein, Marc Sohrmann, John Spieth, Jason E. Stajich, Chaochun Wei, David Willey, Richard K. Wilson, Richard Durbin, and Robert H. Waterston. **The Genome Sequence of Caenorhabditis briggsae: A Platform for Comparative Genomics**. *PLoS Biology*, **1**(2):e45 EP –, November 2003. 20

[107] C. Dieterich, S. W. Clifton, L. N. Schuster, A. Chinwalla, K. Delehaunty, I. Dinkelacker, L. Fulton, R. Fulton, J. Godfrey, P. Minx, M. Mitreva, W. Roeseler, H. Tian, H. Witte, S. P. Yang, R. K. Wilson, and R. J. Sommer. **The Pristionchus pacificus genome provides a unique perspective on nematode lifestyle and parasitism**. *Nat. Genet.*, **40**:1193–1198, Oct 2008. 20

[108] Elodie Ghedin, Shiliang Wang, David Spiro, Elisabet Caler, Qi Zhao, Jonathan Crabtree, Jonathan E. Allen, Arthur L. Delcher, David B. Guiliano, Diego Miranda-Saavedra, Samuel V. Angiuoli, Todd Creasy, Paolo Amedeo, Brian Haas, Najib M. El-Sayed, Jennifer R. Wortman, Tamara Feldblyum, Luke Tallon, Michael Schatz, Martin Shumway, Hean Koo, Steven L. Salzberg, Seth Schobel, Mihaela Pertea, Mihai Pop, Owen White, Geoffrey J. Barton, Clotilde K. S. Carlow, Michael J. Crawford, Jennifer Daub, Matthew W. Dimmic, Chris F. Estes, Jeremy M. Foster, Mehul Ganatra, William F. Gregory, Nicholas M. Johnson, Jinming Jin, Richard Komuniecki, Ian Korf, Sanjay Kumar, Sandra Laney, Ben-Wen Li, Wen Li, Tim H. Lindblom, Sara Lustigman, Dong Ma, Claude V. Maina, David M. A. Martin, James P. McCarter, Larry McReynolds, Makedonka Mitreva, Thomas B. Nutman, John Parkinson, Jose M. Peregrin-Alvarez, Catherine Poole, Qinghu Ren, Lori Saunders, Ann E. Sluder, Katherine Smith, Mario Stanke, Thomas R. Unnasch, Jenna Ware, Aguan D. Wei, Gary Weil, Deryck J. Williams, Yinhua Zhang, Steven A. Williams, Claire Fraser-Liggett, Barton Slatko, Mark L. Blaxter, and Alan L. Scott. **Draft Genome of the Filarial Nematode Parasite Brugia malayi**. *Science*, **317**(5845):1756–1760, September 2007. 20

[109] A. R. Jex, S. Liu, B. Li, N. D. Young, R. S. Hall, Y. Li, L. Yang, N. Zeng, X. Xu, Z. Xiong, F. Chen, X. Wu, G. Zhang, X. Fang, Y. Kang, G. A. Anderson, T. W. Harris, B. E. Campbell, J. Vlaminck, T. Wang, C. Cantacessi, E. M. Schwarz, S. Ranganathan, P. Geldhof, P. Nejsum, P. W. Sternberg, H. Yang, J. Wang, J. Wang, and R. B. Gasser. **Ascaris suum draft genome**. *Nature*, Oct 2011. 20

[110] M. Mitreva, D. P. Jasmer, D. S. Zarlenga, Z. Wang, S. Abubucker, J. Martin, C. M. Taylor, Y. Yin, L. Fulton, P. Minx, S. P. Yang, W. C. Warren, R. S. Fulton, V. Bhonagiri, X. Zhang, K. Hallsworth-Pepin, S. W. Clifton, J. P. McCarter, J. Appleton, E. R. Mardis, and R. K. Wilson. **The draft genome of the parasitic nematode Trichinella spiralis**. *Nat. Genet.*, **43**:228–235, Mar 2011. 20

[111] P. Abad, J. Gouzy, J. M. Aury, P. Castagnone-Sereno, E. G. Danchin, E. Deleury, L. Perfus-Barbeoch, V. Anthouard, F. Artiguenave, V. C. Blok, M. C. Caillaud, P. M. Coutinho, C. Dasilva, F. De Luca, F. Deau, M. Esquibet, T. Flutre, J. V. Goldstone, N. Hamamouch, T. Hewezi, O. Jaillon, C. Jubin, P. Leonetti, M. Magliano, T. R. Maier, G. V. Markov, P. McVeigh, G. Pesole, J. Poulain, M. Robinson-Rechavi, E. Sallet, B. Segurens, D. Steinbach, T. Tytgat, E. Ugarte, C. van Ghelder, P. Veronico, T. J. Baum, M. Blaxter, T. Bleve-Zacheo, E. L. Davis, J. J. Ewbank, B. Favery, E. Grenier, B. Henrissat, J. T. Jones, V. Laudet, A. G. Maule, H. Quesneville, M. N. Rosso, T. Schiex, G. Smant, J. Weissenbach, and P. Wincker. **Genome sequence of the metazoan**

plant-parasitic nematode Meloidogyne incognita. *Nat. Biotechnol.*, **26**:909–915, Aug 2008. 20

[112] C. H. Opperman, D. M. Bird, V. M. Williamson, D. S. Rokhsar, M. Burke, J. Cohn, J. Cromer, S. Diener, J. Gajan, S. Graham, T. D. Houfek, Q. Liu, T. Mitros, J. Schaff, R. Schaffer, E. Scholl, B. R. Sosinski, V. P. Thomas, and E. Windham. **Sequence and genetic map of Meloidogyne hapla: A compact nematode genome for plant parasitism**. *Proc. Natl. Acad. Sci. U.S.A.*, **105**:14802–14807, Sep 2008. 20

[113] T. Kikuchi, J. A. Cotton, J. J. Dalzell, K. Hasegawa, N. Kanzaki, P. McVeigh, T. Takanashi, I. J. Tsai, S. A. Assefa, P. J. Cock, T. D. Otto, M. Hunt, A. J. Reid, A. Sanchez-Flores, K. Tsuchihara, T. Yokoi, M. C. Larsson, J. Miwa, A. G. Maule, N. Sahashi, J. T. Jones, and M. Berriman. **Genomic Insights into the Origin of Parasitism in the Emerging Plant Pathogen Bursaphelenchus xylophilus**. *PLoS Pathog.*, **7**:e1002219, Sep 2011. 20

[114] S. Kumar, P. H. Schiffer, and M. Blaxter. **959 Nematode Genomes: a semantic wiki for coordinating sequencing projects**. *Nucleic Acids Res*, Nov 2011. [DOI:10.1093/nar/gkr826] [PubMed:22058131]. 20

[115] John Parkinson, Alasdair Anthony, James Wasmuth, Ralf Schmid, Ann Hedley, and Mark Blaxter. **PartiGene–constructing partial genomes**. *Bioinformatics*, **20**(9):1398–1404, June 2004. 20, 103, 111

[116] R. M. Maizels, N. Gomez-Escobar, W. F. Gregory, J. Murray, and X. Zang. **Immune evasion genes from filarial nematodes**. *Int. J. Parasitol.*, **31**:889–898, Jul 2001. 20, 21

[117] Rick M. Maizels, Adam Balic, Natalia Gomez-Escobar, Meera Nair, Matt D. Taylor, and Judith E. Allen. **Helminth parasites; masters of regulation**. *Immunological Reviews*, **201**(1):89–116, 2004. 21

[118] Natalia Gomez-Escobar, William F. Gregory, Collette Britton, Linda Murray, Craig Corton, Neil Hall, Jen Daub, Mark L. Blaxter, and Rick M Maizels. **Abundant larval transcript-1 and -2 genes from Brugia malayi: diversity of genomic environments but conservation of 5' promoter sequences functional in Caenorhabditis elegans**. *Molecular and Biochemical Parasitology*, **125**(1-2):59–71, 2002. 21

[119] J. Murray, W. F. Gregory, N. Gomez-Escobar, A. K. Atmadja, and R. M. Maizels. **Expression and immune recognition of Brugia malayi VAL-1, a homologue of vespid venom allergens and Ancylostoma secreted proteins**. *Mol. Biochem. Parasitol.*, **118**:89–96, Nov 2001. [PubMed:11704277]. 21

[120] Yvonne Harcus, John Parkinson, Cecilia Fernandez, Jennifer Daub, Murray Selkirk, Mark Blaxter, and Rick Maizels. **Signal sequence analysis of expressed sequence tags from the nematode Nippostrongylus brasiliensis and the evolution of secreted proteins in parasites**. *Genome Biology*, **5**(6):R39, 2004. 21, 103

[121] Shivashankar H. Nagaraj, Robin B. Gasser, and Shoba Ranganathan. **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*, **2**(9):e301, 2008. 21

[122] John Parkinson, Claire Whitton, Ralf Schmid, Marian Thomson, and Mark Blaxter. % bf NEMBASE: a resource for parasitic nematode ESTs. *Nucl. Acids Res.*, **32**(suppl_1):D427–430, 2004. 21, 61, 114

[123] James Wasmuth, Ralf Schmid, Ann Hedley, and Mark Blaxter. **On the Extent and Origins of Genic Novelty in the Phylum Nematoda**. *PLoS Neglected Tropical Diseases*, **2**(7):e258, July 2008. 21, 103

[124] B. Elsworth, J. Wasmuth, and M. Blaxter. **NEMBASE4: The nematode transcriptome resource**. *Int. J. Parasitol.*, **41**:881–894, Jul 2011. 21, 61, 114

[125] Z. Wang, S. Abubucker, J. Martin, R. K. Wilson, J. Hawdon, and M. Mitreva. **Characterizing Ancylostoma caninum transcriptome and exploring nematode parasitic adaptation**. *BMC Genomics*, **11**:307, 2010. 21, 104

[126] N. Borchert, C. Dieterich, K. Krug, W. Schutz, S. Jung, A. Nordheim, R. J. Sommer, and B. Macek. **Proteogenomics of Pristionchus pacificus reveals distinct proteome structure of nematode models**. *Genome Res.*, **20**:837–846, Jun 2010. 21

[127] S. Kumar and M. L. Blaxter. **Comparing de novo assemblers for 454 transcriptome data**. *BMC Genomics*, **11**:571, Oct 2010. 21, 25, 41, 58, 113

[128] J. Wang, B. Czech, A. Crunk, A. Wallace, M. Mitreva, G. J. Hannon, and R. E. Davis. **Deep small RNA sequencing from the nematode Ascaris reveals conservation, functional diversification, and novel developmental profiles**. *Genome Res.*, **21**:1462–1477, Sep 2011. 21

[129] M. Blaxter, S. Kumar, G. Kaur, G. Koutsouvoulos, and B. Elsworth. **Genomics and transcriptomics across the diversity of the Nematoda**. *Parasite Immunol*, Nov 2011. 21

[130] C. Cantacessi, B. E. Campbell, N. D. Young, A. R. Jex, R. S. Hall, P. J. Presidente, J. L. Zawadzki, W. Zhong, B. Aleman-Meza, A. Loukas, P. W. Sternberg, and R. B. Gasser. **Differences in transcription between free-living and CO2-activated third-stage larvae of Haemonchus contortus**. *BMC Genomics*, **11**:266, 2010. 21

[131] M. L. Metzker. **Sequencing technologies - the next generation**. *Nat. Rev. Genet.*, **11**:31–46, Jan 2010. 23

[132] O. Harismendy, P. C. Ng, R. L. Strausberg, X. Wang, T. B. Stockwell, K. Y. Beeson, N. J. Schork, S. S. Murray, E. J. Topol, S. Levy, and K. A. Frazer. **Evaluation of next generation sequencing platforms for population targeted sequencing studies**. *Genome Biol.*, **10**:R32, 2009. 23

REFERENCES

[133] K. E. Steinmann, C. E. Hart, J. F. Thompson, and P. M. Milos. **Helicos single-molecule sequencing of bacterial genomes**. *Methods Mol. Biol.*, **733**:3–24, 2011. 23

[134] W. Timp, U. M. Mirsaidov, D. Wang, J. Comer, A. Aksimentiev, and G. Timp. **Nanopore Sequencing: Electrical Measurements of the Code of Life**. *IEEE Trans Nanotechnol*, **9**:281–294, May 2010. 23

[135] T. Raz, M. Causey, D. R. Jones, A. Kieu, S. Letovsky, D. Lipson, E. Thayer, J. F. Thompson, and P. M. Milos. **RNA sequencing and quantitation using the Helicos Genetic Analysis System**. *Methods Mol. Biol.*, **733**:37–49, 2011. 23

[136] F. Ozsolak and P. M. Milos. **Single-molecule direct RNA sequencing without cDNA synthesis**. *Wiley Interdiscip Rev RNA*, **2**:565–570, 2011. 23

[137] J. M. Rothberg and J. H. Leamon. **The development and impact of 454 sequencing**. *Nat. Biotechnol.*, **26**:1117–1124, Oct 2008. 24

[138] M. Larguinho, H. M. Santos, G. Doria, H. Scholz, P. V. Baptista, and J. L. Capelo. **Development of a fast and efficient ultrasonic-based strategy for DNA fragmentation**. *Talanta*, **81**:881–886, May 2010. 23

[139] P. Nyren. **The history of pyrosequencing**. *Methods Mol. Biol.*, **373**:1–14, 2007. 23

[140] S. Balzer, K. Malde, and I. Jonassen. **Systematic exploration of error sources in pyrosequencing flowgram data**. *Bioinformatics*, **27**:i304–309, Jul 2011. 25, 66, 103

[141] R. C. Novais and Y. R. Thorstenson. **The evolution of PyrosequencingÂő for microbiology: From genes to genomes**. *J. Microbiol. Methods*, **86**:1–7, Jul 2011. 25

[142] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. Alenquer, T. P. Jarvie, K. B. Jirage, J. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg. **Genome sequencing in microfabricated high-density picolitre reactors**. *Nature*, **437**:376–380, Sep 2005. 25, 41, 113

[143] D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, J. M. Boutell, J. Bryant, R. J. Carter, R. Keira Cheetham, A. J. Cox, D. J. Ellis, M. R. Flatbush, N. A. Gormley, S. J. Humphray, L. J. Irving, M. S. Karbelashvili, S. M. Kirk, H. Li, X. Liu, K. S. Maisinger, L. J. Murray, B. Obradovic, T. Ost, M. L. Parkinson, M. R. Pratt, I. M. Rasolonjatovo, M. T. Reed, R. Rigatti, C. Rodighiero, M. T. Ross, A. Sabot, S. V. Sankar, A. Scally, G. P. Schroth, M. E. Smith, V. P. Smith, A. Spiridou, P. E. Torrance, S. S. Tzonev, E. H. Vermaas, K. Walter, X. Wu, L. Zhang, M. D. Alam, C. Anastasi, I. C. Aniebo, D. M. Bailey, I. R. Bancarz, S. Banerjee, S. G. Barbour, P. A. Baybayan, V. A. Benoit, K. F. Benson, C. Bevis, P. J. Black, A. Boodhun, J. S. Brennan, J. A. Bridgham, R. C. Brown, A. A. Brown, D. H. Buermann, A. A. Bundu, J. C. Burrows, N. P. Carter, N. Castillo, M. Chiara E Catenazzi, S. Chang, R. Neil Cooley, N. R. Crake, O. O. Dada, K. D. Diakoumakos, B. Dominguez-Fernandez, D. J. Earnshaw, U. C. Egbujor, D. W. Elmore, S. S. Etchin, M. R. Ewan, M. Fedurco, L. J. Fraser, K. V. Fuentes Fajardo, W. Scott Furey, D. George, K. J. Gietzen, C. P. Goddard, G. S. Golda, P. A. Granieri, D. E. Green, D. L. Gustafson, N. F. Hansen, K. Harnish, C. D. Haudenschild, N. I. Heyer, M. M. Hims, J. T. Ho, A. M. Horgan, K. Hoschler, S. Hurwitz, D. V. Ivanov, M. Q. Johnson, T. James, T. A. Huw Jones, G. D. Kang, T. H. Kerelska, A. D. Kersey, I. Khrebtukova, A. P. Kindwall, Z. Kingsbury, P. I. Kokko-Gonzales, A. Kumar, M. A. Laurent, C. T. Lawley, S. E. Lee, X. Lee, A. K. Liao, J. A. Loch, M. Lok, S. Luo, R. M. Mammen, J. W. Martin, P. G. McCauley, P. McNitt, P. Mehta, K. W. Moon, J. W. Mullens, T. Newington, Z. Ning, B. Ling Ng, S. M. Novo, M. J. O'Neill, M. A. Osborne, A. Osnowski, O. Ostadan, L. L. Paraschos, L. Pickering, A. C. Pike, A. C. Pike, D. Chris Pinkard, D. P. Pliskin, J. Podhasky, V. J. Quijano, C. Raczy, V. H. Rae, S. R. Rawlings, A. Chiva Rodriguez, P. M. Roe, J. Rogers, M. C. Rogert Bacigalupo, N. Romanov, A. Romieu, R. K. Roth, N. J. Rourke, S. T. Ruediger, E. Rusman, R. M. Sanches-Kuiper, M. R. Schenker, J. M. Seoane, R. J. Shaw, M. K. Shiver, S. W. Short, N. L. Sizto, J. P. Sluis, M. A. Smith, J. Ernest Sohna Sohna, E. J. Spence, K. Stevens, N. Sutton, L. Szajkowski, C. L. Tregidgo, G. Turcatti, S. Vandevondele, Y. Verhovsky, S. M. Virk, S. Wakelin, G. C. Walcott, J. Wang, G. J. Worsley, J. Yan, L. Yau, M. Zuerlein, J. Rogers, J. C. Mullikin, M. E. Hurles, N. J. McCooke, J. S. West, F. L. Oaks, P. L. Lundberg, D. Klenerman, R. Durbin, and A. J. Smith. **Accurate whole human genome sequencing using reversible terminator chemistry**. *Nature*, **456**:53–59, Nov 2008. 25, 27

[144] R. Li, W. Fan, G. Tian, H. Zhu, L. He, J. Cai, Q. Huang, Q. Cai, B. Li, Y. Bai, Z. Zhang, Y. Zhang, W. Wang, J. Li, F. Wei, H. Li, M. Jian, J. Li, Z. Zhang, R. Nielsen, D. Li, W. Gu, Z. Yang, Z. Xuan, O. A. Ryder, F. C. Leung, Y. Zhou, J. Cao, X. Sun, Y. Fu, X. Fang, X. Guo, B. Wang, R. Hou, F. Shen, B. Mu, P. Ni, R. Lin, W. Qian, G. Wang, C. Yu, W. Nie, J. Wang, Z. Wu, H. Liang, J. Min, Q. Wu, S. Cheng, J. Ruan, M. Wang, Z. Shi, M. Wen, B. Liu, X. Ren, H. Zheng, D. Dong, K. Cook, G. Shan, H. Zhang, C. Kosiol, X. Xie, Z. Lu, H. Zheng, Y. Li, C. C. Steiner, T. T. Lam, S. Lin, Q. Zhang, G. Li, J. Tian, T. Gong, H. Liu, D. Zhang, L. Fang, C. Ye, J. Zhang, W. Hu, A. Xu, Y. Ren, G. Zhang, M. W. Bruford, Q. Li, L. Ma, Y. Guo, N. An, Y. Hu, Y. Zheng, Y. Shi, Z. Li, Q. Liu, Y. Chen, J. Zhao, N. Qu, S. Zhao, F. Tian, X. Wang, H. Wang, L. Xu, X. Liu, T. Vinar,

Y. Wang, T. W. Lam, S. M. Yiu, S. Liu, H. Zhang, D. Li, Y. Huang, X. Wang, G. Yang, Z. Jiang, J. Wang, N. Qin, L. Li, J. Li, L. Bolund, K. Kristiansen, G. K. Wong, M. Olson, X. Zhang, S. Li, H. Yang, J. Wang, and J. Wang. **The sequence and de novo assembly of the giant panda genome**. *Nature*, **463**:311–317, Jan 2010. 27

[145] B. Feldmeyer, C. W. Wheat, N. Krezdorn, B. Rotter, and M. Pfenninger. **Short read Illumina data for the de novo assembly of a non-model snail species transcriptome (Radix balthica, Basommatophora, Pulmonata), and a comparison of assembler performance**. *BMC Genomics*, **12**:317, 2011. 27

[146] J. H. Malone and B. Oliver. **Microarrays, deep sequencing and the true measure of the transcriptome**. *BMC Biol.*, **9**:34, 2011. 27

[147] H. Matsumura, K. Yoshida, S. Luo, D. H. Kruger, G. Kahl, G. P. Schroth, and R. Terauchi. **High-throughput SuperSAGE**. *Methods Mol. Biol.*, **687**:135–146, 2011. 27

[148] V. E. Velculescu, L. Zhang, B. Vogelstein, and K. W. Kinzler. **Serial analysis of gene expression**. *Science*, **270**:484–487, Oct 1995. 27

[149] J. R. Miller, S. Koren, and G. Sutton. **Assembly algorithms for next-generation sequencing data**. *Genomics*, **95**:315–327, Jun 2010. 27

[150] F. Sanger, A. R. Coulson, B. G. Barrell, A. J. Smith, and B. A. Roe. **Cloning in single-stranded bacteriophage as an aid to rapid DNA sequencing**. *J. Mol. Biol.*, **143**:161–178, Oct 1980. [PubMed:6260957]. 28

[151] R. Staden. **A strategy of DNA sequencing employing computer programs**. *Nucleic Acids Res.*, **6**:2601–2610, Jun 1979. 28

[152] T. R. Gingeras and R. J. Roberts. **Steps toward computer analysis of nucleotide sequences**. *Science*, **209**:1322–1328, Sep 1980. 28

[153] T. F. Smith and M. S. Waterman. **Identification of common molecular subsequences**. *J. Mol. Biol.*, **147**:195–197, Mar 1981. 28

[154] T. F. Smith, M. S. Waterman, and W. M. Fitch. **Comparative biosequence metrics**. *J. Mol. Evol.*, **18**:38–46, 1981. 28

[155] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. **Basic local alignment search tool**. *J. Mol. Biol.*, **215**:403–410, Oct 1990. 28

[156] W. J. Kent. **BLAT–the BLAST-like alignment tool**. *Genome Res.*, **12**:656–664, Apr 2002. 28

[157] Z. Ning, A. J. Cox, and J. C. Mullikin. **SSAHA: a fast search method for large DNA databases**. *Genome Res.*, **11**:1725–1729, Oct 2001. 28, 52, 115

[158] H. Li and R. Durbin. **Fast and accurate long-read alignment with Burrows-Wheeler transform**. *Bioinformatics*, **26**:589–595, Mar 2010. 29, 84, 117

[159] D. R. Zerbino and E. Birney. **Velvet: algorithms for de novo short read assembly using de Bruijn graphs**. *Genome Res.*, **18**:821–829, May 2008. 29

[160] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev. **Full-length transcriptome assembly from RNA-Seq data without a reference genome**. *Nat. Biotechnol.*, **29**:644–652, Jul 2011. 29

[161] T. S. Schwartz, H. Tae, Y. Yang, K. Mockaitis, J. L. Van Hemert, S. R. Proulx, J. H. Choi, and A. M. Bronikowski. **A garter snake transcriptome: pyrosequencing, de novo assembly, and sex-specific differences**. *BMC Genomics*, **11**:694, 2010. 29, 52, 102

[162] D. C. Koboldt, K. Chen, T. Wylie, D. E. Larson, M. D. McLellan, E. R. Mardis, G. M. Weinstock, R. K. Wilson, and L. Ding. **VarScan: variant detection in massively parallel sequencing of individual and pooled samples**. *Bioinformatics*, **25**:2283–2285, Sep 2009. 30, 66, 115

[163] P. Danecek, âĂĂA. Auton, G. Abecasis, Albers CA., E. Banks, MA. DePristo, Handsaker RE., Lunter G., Marth GT., Sherry ST., McVean GT., Durbin T., and 1000 Genomes Project. **The variant call format and VCFtools**. *Bioinformatics*, **27**:2156–2158, Aug 2011. 30, 77, 115

[164] L. W. Hillier, G. T. Marth, A. R. Quinlan, D. Dooling, G. Fewell, D. Barnett, P. Fox, J. I. Glasscock, M. Hickenbotham, W. Huang, V. J. Magrini, R. J. Richt, S. N. Sander, D. A. Stewart, M. Stromberg, E. F. Tsung, T. Wylie, T. Schedl, R. K. Wilson, and E. R. Mardis. **Whole-genome sequencing and variant discovery in C. elegans**. *Nat. Methods*, **5**:183–188, Feb 2008. 30

[165] S. U. Franssen, J. Gu, N. Bergmann, G. Winters, U. C. Klostermeier, P. Rosenstiel, E. Bornberg-Bauer, and T. B. Reusch. **Transcriptomic resilience to global warming in the seagrass Zostera marina, a marine foundation species**. *Proc. Natl. Acad. Sci. U.S.A.*, **108**:19276–19281, Nov 2011. 30

[166] G. Smyth. **Limma: linear models for microarray data**. *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420, 2005. 30

[167] M. D. Robinson and G. K. Smyth. **Small-sample estimation of negative binomial dispersion, with applications to SAGE data**. *Biostatistics*, **9**:321–332, Apr 2008. 30

[168] S. Anders and W. Huber. **Differential expression analysis for sequence count data**. *Genome Biol.*, **11**:R106, 2010. 30, 78, 84, 105, 115

[169] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data**. *Bioinformatics*, **26**:139–140, Jan 2010. 30, 118

REFERENCES

[170] T. J. Hardcastle and K. A. Kelly. **baySeq: empirical Bayesian methods for identifying differential expression in sequence count data**. *BMC Bioinformatics*, **11**:422, 2010. 30

[171] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat. Genet.*, **25**:25–29, May 2000. 30

[172] E. C. Dimmer, R. P. Huntley, Y. Alam-Faruque, T. Sawford, C. O'Donovan, M. J. Martin, B. Bely, P. Browne, W. Mun Chan, R. Eberhardt, M. Gardner, K. Laiho, D. Legge, M. Magrane, K. Pichler, D. Poggioli, H. Sehra, A. Auchincloss, K. Axelsen, M. C. Blatter, E. Boutet, S. Braconi-Quintaje, L. Breuza, A. Bridge, E. Coudert, A. Estreicher, L. Famiglietti, S. Ferro-Rojas, M. Feuermann, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, J. James, S. Jimenez, F. Jungo, G. Keller, P. Lemercier, D. Lieberherr, P. Masson, M. Moinat, I. Pedruzzi, S. Poux, C. Rivoire, B. Roechert, M. Schneider, A. Stutz, S. Sundaram, M. Tognolli, L. Bougueleret, G. Argoud-Puy, I. Cusin, P. Duek-Roggli, I. Xenarios, and R. Apweiler. **The UniProt-GO Annotation database in 2011**. *Nucleic Acids Res*, Nov 2011. 30

[173] S. Via. **The Ecological Genetics of Speciation**. *The American Naturalist*, **159**(S3):1–7, 2002. 31

[174] C. J. McManus, J. D. Coolon, M. O. Duff, J. Eipper-Mains, B. R. Graveley, and P. J. Wittkopp. **Regulatory divergence in Drosophila revealed by mRNA-seq**. *Genome Res.*, **20**:816–825, Jun 2010. 31

[175] W. Haerty and R. S. Singh. **Gene regulation divergence is a major contributor to the evolution of Dobzhansky-Muller incompatibilities between species of Drosophila**. *Mol. Biol. Evol.*, **23**:1707–1714, Sep 2006. 31

[176] F. Goetz, D. Rosauer, S. Sitar, G. Goetz, C. Simchick, S. Roberts, R. Johnson, C. Murphy, C. R. Bronte, and S. Mackenzie. **A genetic basis for the phenotypic differentiation between siscowet and lean lake trout (Salvelinus namaycush)**. *Mol. Ecol.*, **19 Suppl 1**:176–196, Mar 2010. 31

[177] M. Dassanayake, JS Haas, HJ Bohnert, and JM Cheeseman. **Shedding light on an extremophile lifestyle through transcriptomics**. *New Phytologist*, **183**(3):764–775, 2009. 31

[178] M. K. Hughes and A. L. Hughes. **Natural selection on Plasmodium surface proteins**. *Mol. Biochem. Parasitol.*, **71**:99–113, Apr 1995. 32

[179] S. L. Nuismer and S. P. Otto. **Host-parasite interactions and the evolution of gene expression**. *PLoS Biol.*, **3**:e203, Jul 2005. 32

[180] D. Colinet, A. Schmitz, D. Cazes, J. L. Gatti, and M. Poirie. **The origin of intraspecific variation of virulence in an eukaryotic immune suppressive parasite**. *PLoS Pathog.*, **6**:e1001206, 2010. 32

[181] B. Chevreux, T. Pfisterer, B. Drescher, A. J. Driesel, W. E. Muller, T. Wetter, and S. Suhai. **Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs**. *Genome Res.*, **14**:1147–1159, Jun 2004. 41, 58, 113

[182] X. Huang and A. Madan. **CAP3: A DNA sequence assembly program**. *Genome Res.*, **9**:868–877, Sep 1999. 41, 113

[183] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, GonÃ¢alo R. Abecasis, and Richard Durbin. **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics*, **25**(16):2078–2079, 2009. 52, 77, 115, 117

[184] G. Pertea, X. Huang, F. Liang, V. Antonescu, R. Sultana, S. Karamycheva, Y. Lee, J. White, F. Cheung, B. Parvizi, J. Tsai, and J. Quackenbush. **TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets**. *Bioinformatics*, **19**:651–652, Mar 2003. 59, 113

[185] James Wasmuth and Mark Blaxter. **prot4EST: Translating Expressed Sequence Tags from neglected genomes**. *BMC Bioinformatics*, **5**(1):187, 2004. 61, 114

[186] Ralf Schmid and Mark L Blaxter. **annot8r: GO, EC and KEGG annotation of EST datasets**. *BMC Bioinformatics*, **9**:180, 2008. 61, 63, 64, 114

[187] E. M. Zdobnov and R. Apweiler. **InterProScan–an integration platform for the signature-recognition methods in InterPro**. *Bioinformatics*, **17**:847–848, Sep 2001. 62, 63, 102, 114

[188] T. N. Petersen, S. Brunak, G. von Heijne, and H. Nielsen. **SignalP 4.0: discriminating signal peptides from transmembrane regions**. *Nat. Methods*, **8**:785–786, 2011. 62, 114

[189] W. Amos, J. W. Wilmer, K. Fullard, T. M. Burg, J. P. Croxall, D. Bloch, and T. Coulson. **The influence of parental relatedness on reproductive success**. *Proc. Biol. Sci.*, **268**:2021–2027, Oct 2001. 77, 115

[190] J. M. Aparicio, J. Ortego, and P. J. Cordero. **What should we weigh to estimate heterozygosity, alleles or loci?** *Mol. Ecol.*, **15**:4659–4665, Dec 2006. 77, 115

[191] W. ColtMan, Pilkington J. G., Smith J. A., and Pember ton J.M. **Parasite-mediated selection against inbred Soay sheep in a free-living, island population**. *Evolution*, **81**:1259–1267, 1999. 77, 115

[192] J. S. Alho, K. Valimaki, and J. Merila. **Rhh: an R extension for estimating multilocus heterozygosity and heterozygosity-heterozygosity correlation**. *Mol Ecol Resour*, **10**:720–722, Jul 2010. 77, 115

[193] I. G. Wilson. **Inhibition and facilitation of nucleic acid amplification**. *Appl. Environ. Microbiol.*, **63**:3741–3751, Oct 1997. 101

[194] M. A. Valasek and J. J. Repa. **The power of real-time PCR**. *Adv Physiol Educ*, **29**:151–159, Sep 2005. 101

[195] K. Ohashi, F. Takizawa, N. Tokumaru, C. Nakayasu, H. Toda, U. Fischer, T. Moritomo, K. Hashimoto, T. Nakanishi, and J. M. Dijkstra. **A molecule in teleost fish, related with human MHC-encoded G6F, has a cytoplasmic tail with ITAM and marks the surface of thrombocytes and in some fishes also of erythrocytes**. *Immunogenetics*, **62**:543–559, Aug 2010. 101

[196] K. al Sabti. **Micronuclei induced by selenium, mercury, methylmercury and their mixtures in binucleated blocked fish erythrocyte cells**. *Mutat. Res.*, **320**:157–163, Jan 1994. 101

[197] M. C. Hale, J. R. Jackson, and J. A. Dewoody. **Discovery and evaluation of candidate sex-determining genes and xenobiotics in the gonads of lake sturgeon (Acipenser fulvescens)**. *Genetica*, **138**:745–756, Jul 2010. 102

[198] A. Papanicolaou, R. Stierli, R. H. Ffrench-Constant, and D. G. Heckel. **Next generation transcriptomes for next generation genomes using est2assembly**. *BMC Bioinformatics*, **10**:447, 2009. 102

[199] J. Emmersen, S. Rudd, H. W. Mewes, and I. V. Tetko. **Separation of sequences from host-pathogen interface using triplet nucleotide frequencies**. *Fungal Genet. Biol.*, **44**:231–241, Apr 2007. 102

[200] S. T. O'Neil, J. D. Dzurisin, R. D. Carmichael, N. F. Lobo, S. J. Emrich, and J. J. Hellmann. **Population-level transcriptome sequencing of nonmodel organisms Erynnis propertius and Papilio zelicaon**. *BMC Genomics*, **11**:310, 2010. 102, 104

[201] R. Gregory, A. C. Darby, H. Irving, M. B. Coulibaly, M. Hughes, L. L. Koekemoer, M. Coetzee, H. Ranson, J. Hemingway, N. Hall, and C. S. Wondji. **A De Novo Expression Profiling of Anopheles funestus, Malaria Vector in Africa, Using 454 Pyrosequencing**. *PLoS ONE*, **6**:e17418, 2011. 102

[202] A. Kunstner, J. B. Wolf, N. Backstrom, O. Whitney, C. N. Balakrishnan, L. Day, S. V. Edwards, D. E. Janes, B. A. Schlinger, R. K. Wilson, E. D. Jarvis, W. C. Warren, and H. Ellegren. **Comparative genomics based on massive parallel transcriptome sequencing reveals patterns of substitution and selection across 10 bird species**. *Mol. Ecol.*, **19 Suppl 1**:266–276, Mar 2010. 102

[203] H. Yang, X. Chen, and W. H. Wong. **Completely phased genome sequencing through chromosome sorting**. *Proc. Natl. Acad. Sci. U.S.A.*, **108**:12–17, Jan 2011. 103

[204] Andrew Adey, Hilary Morrison, X Asan, Xu Xun, Jacob Kitzman, Emily Turner, Bethany Stackhouse, Alexandra MacKenzie, Nicholas Caruccio, Xiuqing Zhang, Jacob Shendure, Emily Turner, Bethany Stackhouse, Alexandra MacKenzie, Nicholas Caruccio, Xiuqing Zhang, and Jay Shendure. **Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition**. *Genome Biol.*, **11**(12):R119, 2010. 103

[205] S. Kryazhimskiy and J. B. Plotkin. **The population genetics of dN/dS**. *PLoS Genet.*, **4**:e1000304, Dec 2008. 104

[206] E. Novaes, D. R. Drost, W. G. Farmerie, G. J. Pappas, D. Grattapaglia, R. R. Sederoff, and M. Kirst. **High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome**. *BMC Genomics*, **9**:312, 2008. 104

[207] W. J. Swanson, A. Wong, M. F. Wolfner, and C. F. Aquadro. **Evolutionary expressed sequence tag analysis of Drosophila female reproductive tracts identifies genes subjected to positive selection**. *Genetics*, **168**:1457–1465, Nov 2004. 104

[208] W. J. Swanson, A. G. Clark, H. M. Waldrip-Dail, M. F. Wolfner, and C. F. Aquadro. **Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in Drosophila**. *Proc. Natl. Acad. Sci. U.S.A.*, **98**:7375–7379, Jun 2001. 104, 107

[209] T. Miyata and T. Yasunaga. **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.*, **16**:23–36, Sep 1980. 104

[210] K Knopf, A Madriles Helm, R Lucius, W Bleiss, and H Taraschewski. **Migratory response of European eel (Anguilla anguilla) phagocytes to the eel swimbladder nematode Anguillicola crassus**. *Parasitology Research*, **102**(6):1311–6, May 2008. 104

[211] K Molnár. **Formation of parasitic nodules in the swimbladder and intestinal walls of the eel Anguilla anguilla due to infections with larval stages of Anguillicola crassus**. *Diseases of Aquatic Organisms*, **20**(3):163–170, 1994. 104

[212] A. L. Veuthey and G. Bittar. **Phylogenetic relationships of fungi, plantae, and animalia inferred from homologous comparison of ribosomal proteins**. *J. Mol. Evol.*, **47**:81–92, Jul 1998. 104

[213] A. L. Scott. **Nematode sperm**. *Parasitol. Today (Regul. Ed.)*, **12**:425–430, Nov 1996. 105

[214] S. J. Gould and R. C. Lewontin. **The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme**. *Proceedings of the Royal Society of London. Series B, Biological Sciences (1934-1990)*, **205**(1161):581–598, 1979. 107

[215] R. Nielsen. **Adaptionism-30 years after Gould and Lewontin**. *Evolution*, **63**:2487–2490, Oct 2009. 107

[216] X. Yang, E. E. Schadt, S. Wang, H. Wang, A. P. Arnold, L. Ingram-Drake, T. A. Drake, and A. J. Lusis. **Tissue-specific expression and regulation of sexually dimorphic genes in mice**. *Genome Res.*, **16**:995–1004, Aug 2006. 107

[217] K. Thoemke, W. Yi, J. M. Ross, S. Kim, V. Reinke, and D. Zarkower. **Genome-wide analysis of sex-enriched gene expression during C. elegans larval development**. *Dev. Biol.*, **284**:500–508, Aug 2005. 107

[218] A. D. Cutter and S. Ward. **Sexual and temporal dynamics of molecular evolution in C. elegans development**. *Mol. Biol. Evol.*, **22**:178–188, Jan 2005. 107

[219] Z. F. Jiang and C. A. Machado. **Evolution of sex-dependent gene expression in three recently diverged species of Drosophila**. *Genetics*, **183**:1175–1185, Nov 2009. 107

[220] Brent Ewing, LaDeana Hillier, Michael C. Wendl, and Phil Green. **Base-Calling of automated sequencer traces using Phred. I. Accuracy Assessment**. *Genome Res.*, **8**(3):175–185, March 1998. 111

[221] Phil Green. *PHRAP documentation.*, 1994. 111, 113

[222] A. Coppe, J. M. Pujolar, G. E. Maes, P. F. Larsen, M. M. Hansen, L. Bernatchez, L. Zane, and S. Bortoluzzi. **Sequencing, de novo annotation and analysis of the first Anguilla anguilla transcriptome: EelBase opens new perspectives for the study of the critically endangered European eel**. *BMC Genomics*, **11**:635, 2010. 113

[223] A. Bairoch, L. Bougueleret, S. Altairac, V. Amendolia, A. Auchincloss, G. Argoud-Puy, K. Axelsen, D. Baratin, M. C. Blatter, B. Boeckmann, J. Bolleman, L. Bollondi, E. Boutet, S. B. Quintaje, L. Breuza, A. Bridge, E. deCastro, L. Ciapina, D. Coral, E. Coudert, I. Cusin, G. Delbard, D. Dornevil, P. D. Roggli, S. Duvaud, A. Estreicher, L. Famiglietti, M. Feuermann, S. Gehant, N. Farriol-Mathis, S. Ferro, E. Gasteiger, A. Gateau, V. Gerritsen, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, N. Hulo, J. James, S. Jimenez, F. Jungo, V. Junker, T. Kappler, G. Keller, C. Lachaize, L. Lane-Guermonprez, P. Langendijk-Genevaux, V. Lara, P. Lemercier, V. Le Saux, D. Lieberherr, T. d. e. O. Lima, V. Mangold, X. Martin, P. Masson, K. Michoud, M. Moinat, A. Morgat, A. Mottaz, S. Paesano, I. Pedruzzi, I. Phan, S. Pilbout, V. Pillet, S. Poux, M. Pozzato, N. Redaschi, S. Reynaud, C. Rivoire, B. Roechert, M. Schneider, C. Sigrist, K. Sonesson, S. Staehli, A. Stutz, S. Sundaram, M. Tognolli, L. Verbregue, A. L. Veuthey, L. Yip, L. Zuletta, R. Apweiler, Y. Alam-Faruque, R. Antunes, D. Barrell, D. Binns, L. Bower, P. Browne, W. M. Chan, E. Dimmer, R. Eberhardt, A. Fedotov, R. Foulger, J. Garavelli, R. Golin, A. Horne, R. Huntley, J. Jacobsen, M. Kleen, P. Kersey, K. Laiho, R. Leinonen, D. Legge, Q. Lin, M. Magrane, M. J. Martin, C. O'Donovan, S. Orchard, J. O'Rourke, S. Patient, M. Pruess, A. Sitnov, E. Stanley, M. Corbett, G. di Martino, M. Donnelly, J. Luo, P. van Rensburg, C. Wu, C. Arighi, L. Arminski, W. Barker, Y. Chen, Z. Z. Hu, H. K. Hua, H. Huang, R. Mazumder, P. McGarvey, D. A. Natale, A. Nikolskaya, N. Petrova, B. E. Suzek, S. Vasudevan, C. R. Vinayaka, L. S. Yeh, and J. Zhang. **The Universal Protein Resource (UniProt) 2009**. *Nucleic Acids Res.*, **37**:D169–174, Jan 2009. 114

[224] C Iseli, CV Jongeneel, and P Bucher. **ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences.** *Proc Int Conf Intell Syst Mol Biol*, pages 138–148, 1999. 114

[225] A. Kasprzyk. **BioMart: driving a paradigm change in biological data management**. *Database (Oxford)*, **2011**:bar049, 2011. 114

[226] S. Durinck, P. T. Spellman, E. Birney, and W. Huber. **Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt**. *Nat Protoc*, **4**:1184–1191, 2009. 115

[227] S. Falcon and R. Gentleman. **Using GOstats to test gene lists for GO term association**. *Bioinformatics*, **23**:257–258, Jan 2007. 115

[228] Martin Morgan and Hervé Pagès. *Rsamtools: Import aligned BAM file format sequences into R / Bioconductor*. R package version 1.4.3. 115

[229] JH Boon, VMH Cannaerts, H. Augustijn, MAM Machiels, D. De Charleroy, and F. Ollevier. **The effect of different infection levels with infective larvae of Anguillicola crassus on haematological parameters of European eel (Anguilla anguilla)**. *Aquaculture*, **87**(3-4):243–253, 1990. 116

[230] O.L.M. Haenen, T.A.M. van Wijngaarden, and F.H.M. Borgsteede. **An improved method for the production of infective third-stage juveniles of Anguillicola crassus**. *Aquaculture(Amsterdam)*, **123**(1-2):163–165, 1994. 116

[231] Y. Benjamini and Y. Hochberg. **Controlling the false discovery rate: a practical and powerful approach to multiple testing**. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995. 118

9

Additional tables and figures

9.1 Additional tables

9.2 Additional figures

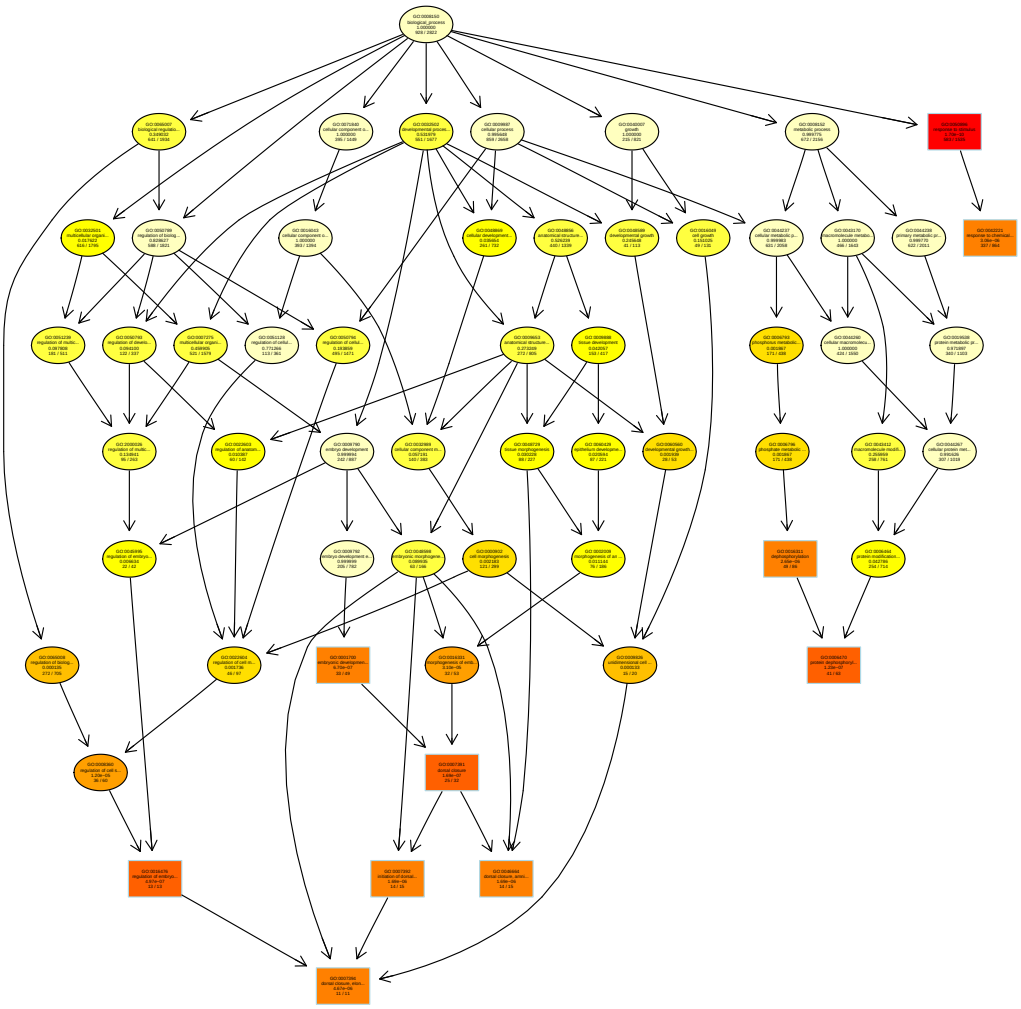

Figure 9.1: GO biological process graph for enriched terms in DE according to sex - Subgraph of the GO-ontology biological process category induced by the top 10 terms identified as enriched in DE genes between male and female worms. Boxes indicate the 10 most significant terms. Box color represents the relative significance, ranging from dark red (most significant) to light yellow (least significant). In each node the categoy-identifier, a (eventualy 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.

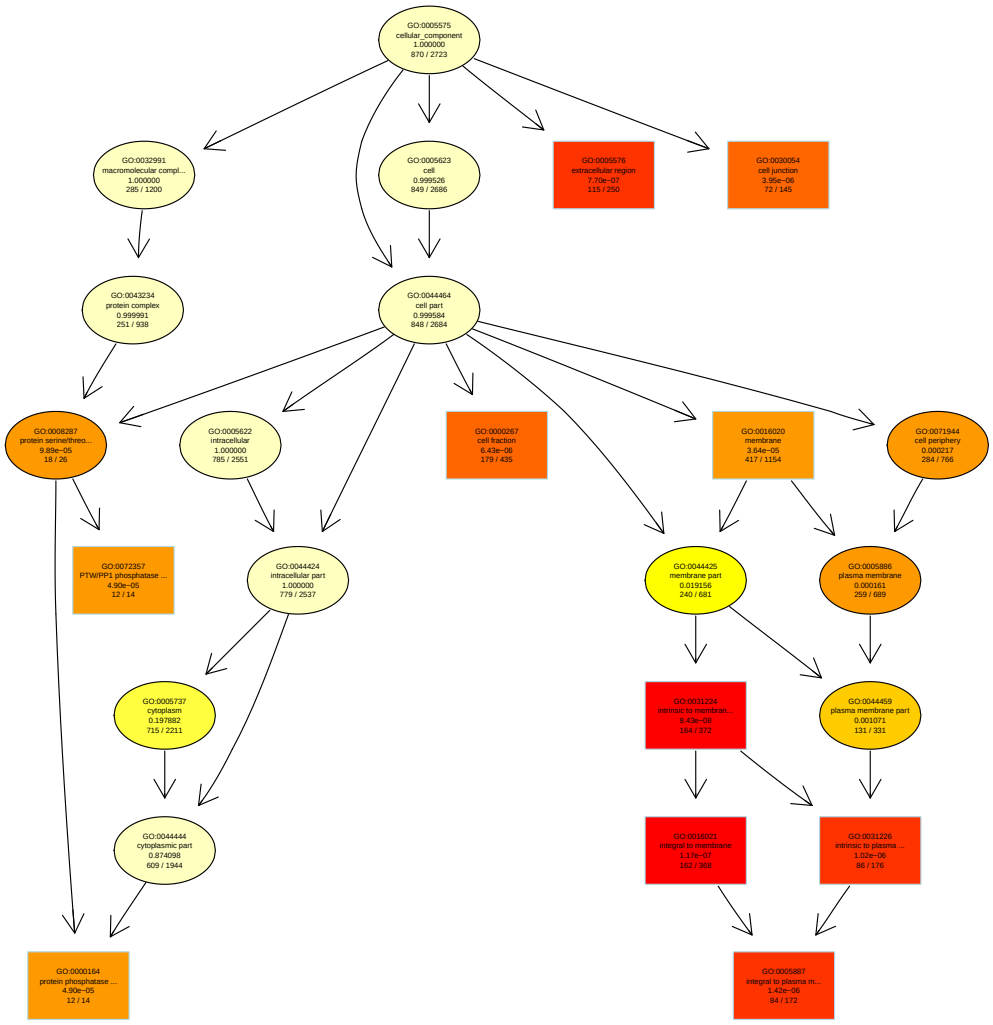

Figure 9.2: GO cellular compartment graph for enriched terms in DE according to sex - Subgraph of the GO-ontology cellular compartment category induced by the top 10 terms identified as enriched in DE genes between male and female worms. Boxes indicate the 10 most significant terms. Box color represents the relative significance, ranging from dark red (most significant) to light yellow (least significant). In each node the categoy-identifier, a (eventualy 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.

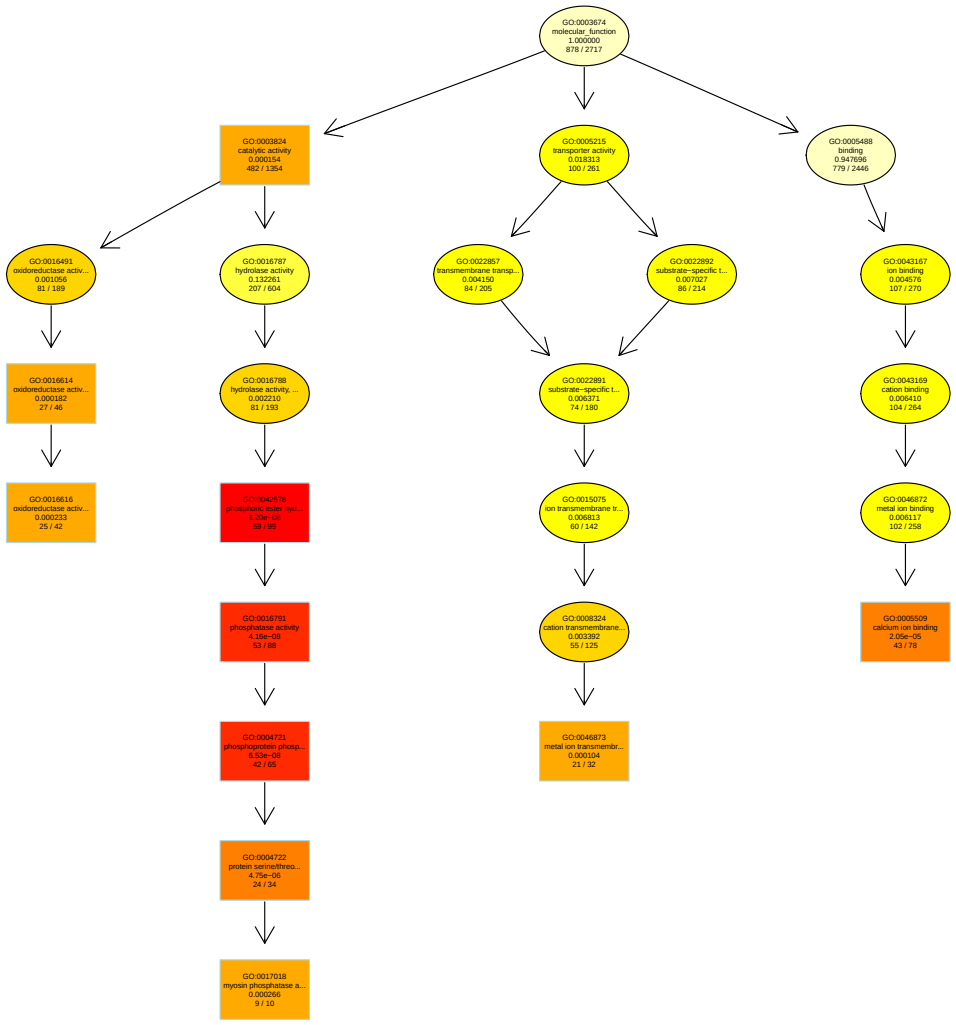

Figure 9.3: GO molecular function graph for enriched terms in DE according to sex - Subgraph of the GO-ontology molecular function category induced by the top 10 terms identified as enriched in DE genes between male and female worms. Boxes indicate the 10 most significant terms. Box color represents the relative significance, ranging from dark red (most significant) to light yellow (least significant). In each node the categoy-identifier, a (eventualy 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.

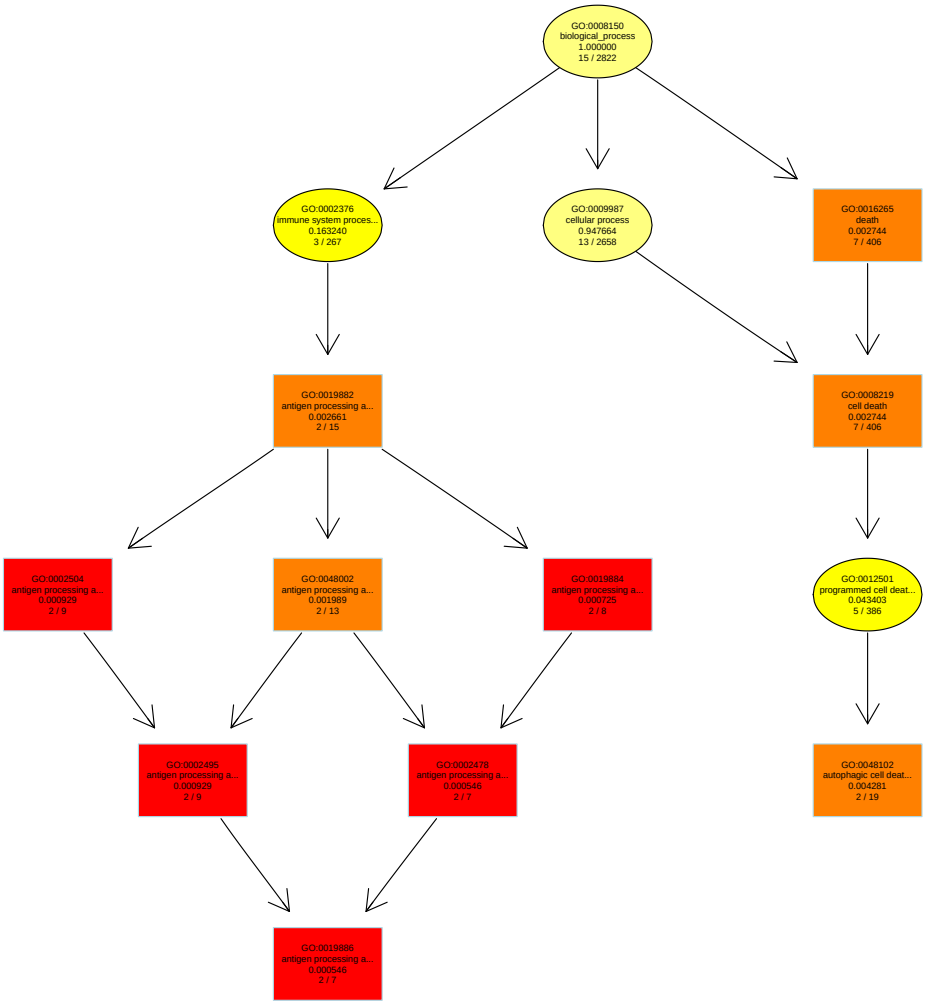

Figure 9.4: GO biological process graph for enriched terms in DE according to eel-host - Subgraph of the GO-ontology biological process category induced by the top 10 terms identified as enriched in DE genes between different host species. Boxes indicate the 10 most significant terms. Box color represents the relative significance, ranging from dark red (most significant) to light yellow (least significant). In each node the categoy-identifier, a (eventualy 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.

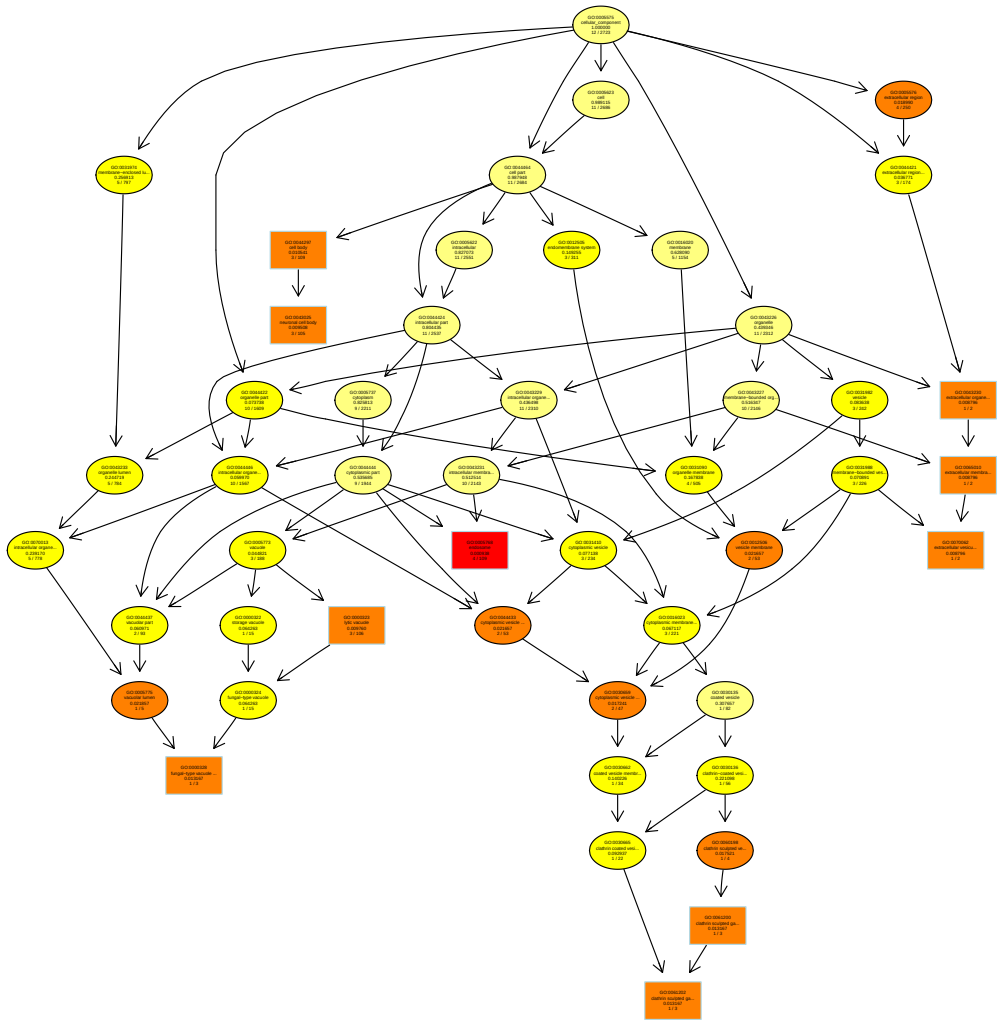

Figure 9.5: GO cellular compartment graph for enriched terms in DE according to eel-host - Subgraph of the GO-ontology cellular comparment category induced by the top 10 terms identified as enriched in DE genes between different host species. Boxes indicate the 10 most significant terms. Box color represents the relative significance, ranging from dark red (most significant) to light yellow (least significant). In each node the categoy-identifier, a (eventualy 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.

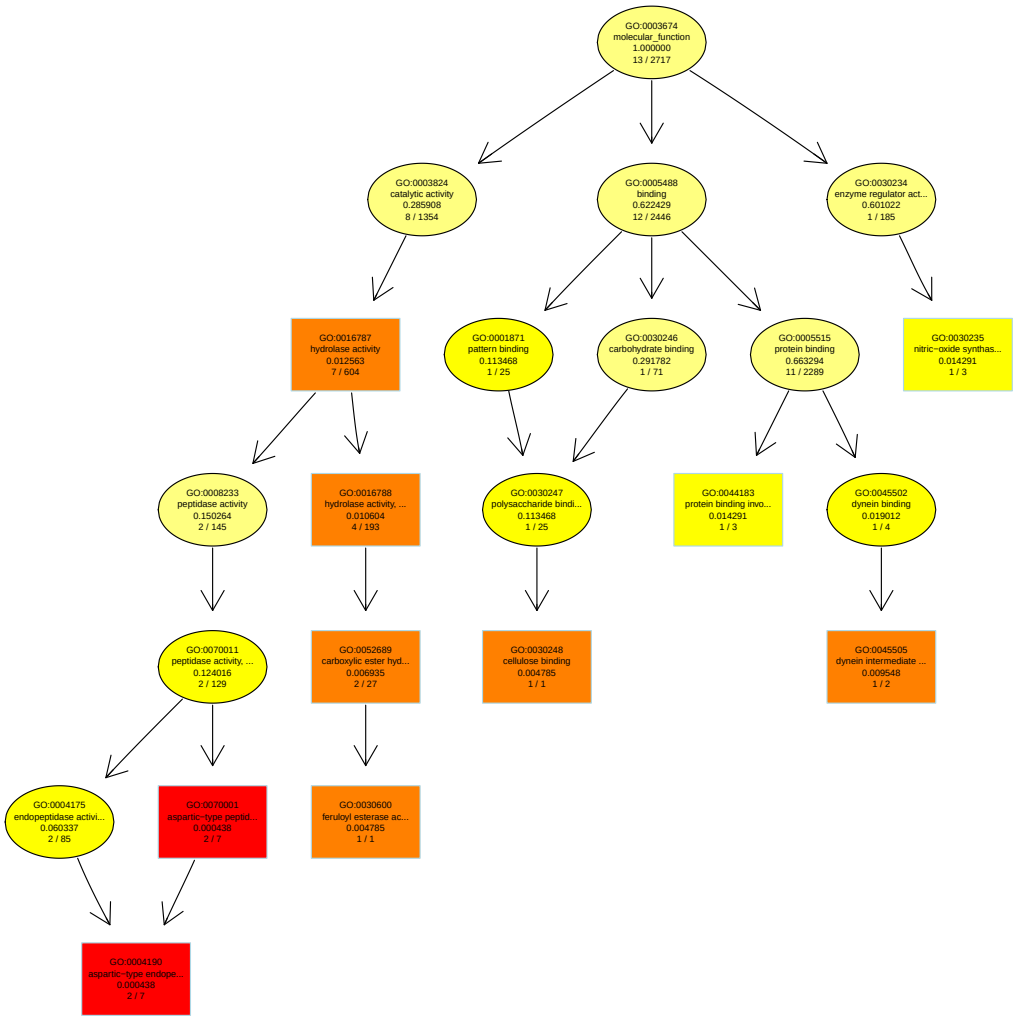

Figure 9.6: GO molcular function graph for enriched terms in DE according to eel-host - Subgraph of the GO-ontology molecular function category induced by the top 10 terms identified as enriched in DE genes between different host species. Boxes indicate the 10 most significant terms. Box color represents the relative significance, ranging from dark red (most significant) to light yellow (least significant). In each node the categoy-identifier, a (eventualy 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.

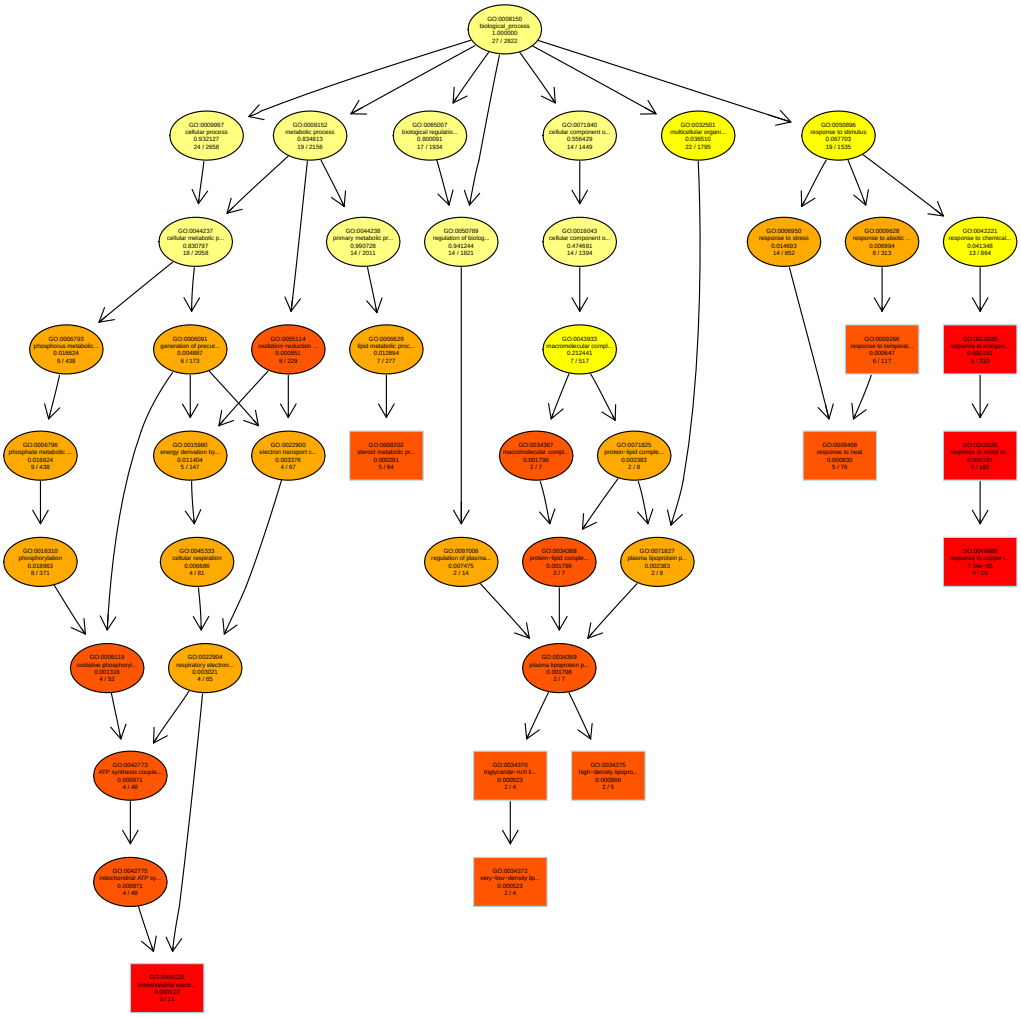

Figure 9.7: GO biological process graph for enriched terms in DE according to worm-population - Subgraph of the GO-ontology cellular comparment category induced by the top 10 terms identified as enriched in DE genes between different parasite populations. Boxes indicate the 10 most significant terms. Box color represents the relative significance, ranging from dark red (most significant) to light yellow (least significant). In each node the categoy-identifier, a (eventualy 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.

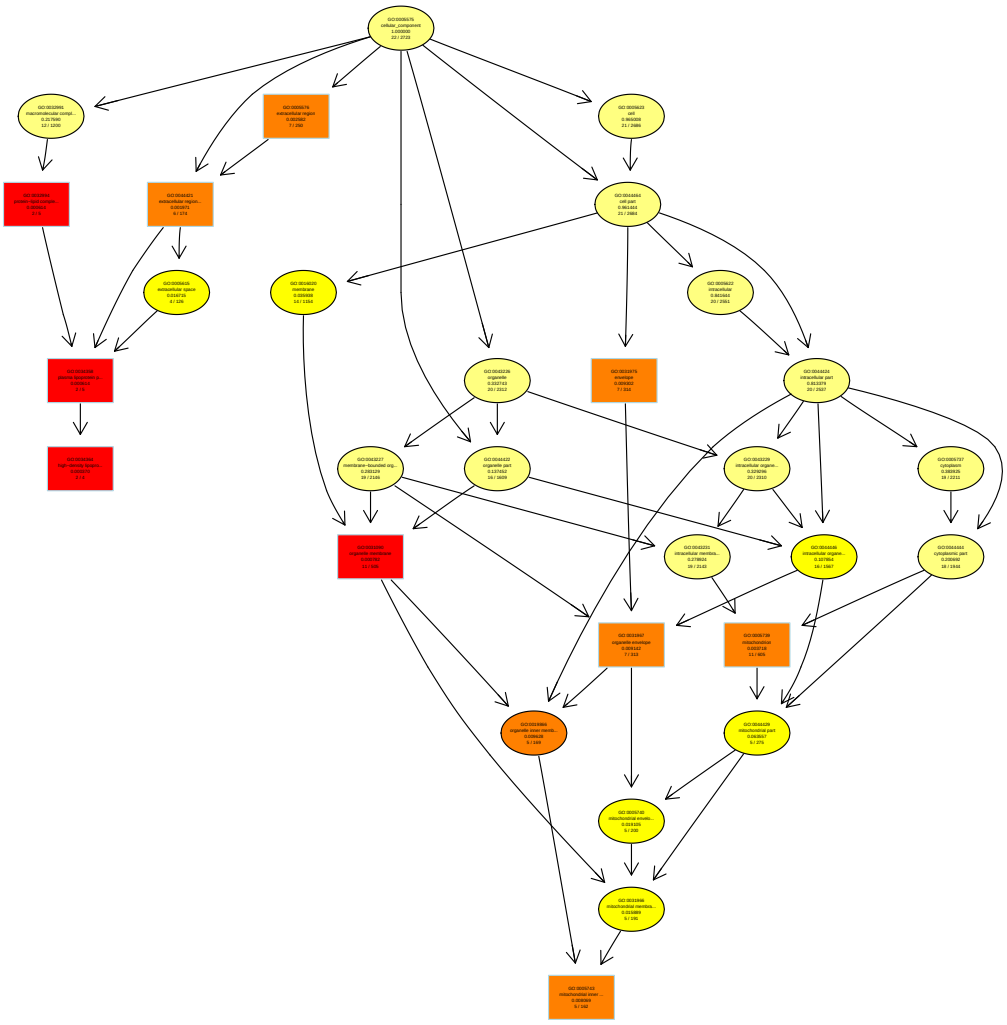

Figure 9.8: GO cellular compartment graph for enriched terms in DE according to worm-population - Subgraph of the GO-ontology biological process category induced by the top 10 terms identified as enriched in DE genes between different parasite populations. Boxes indicate the 10 most significant terms. Box color represents the relative significance, ranging from dark red (most significant) to light yellow (least significant). In each node the categoy-identifier, a (eventualy 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.

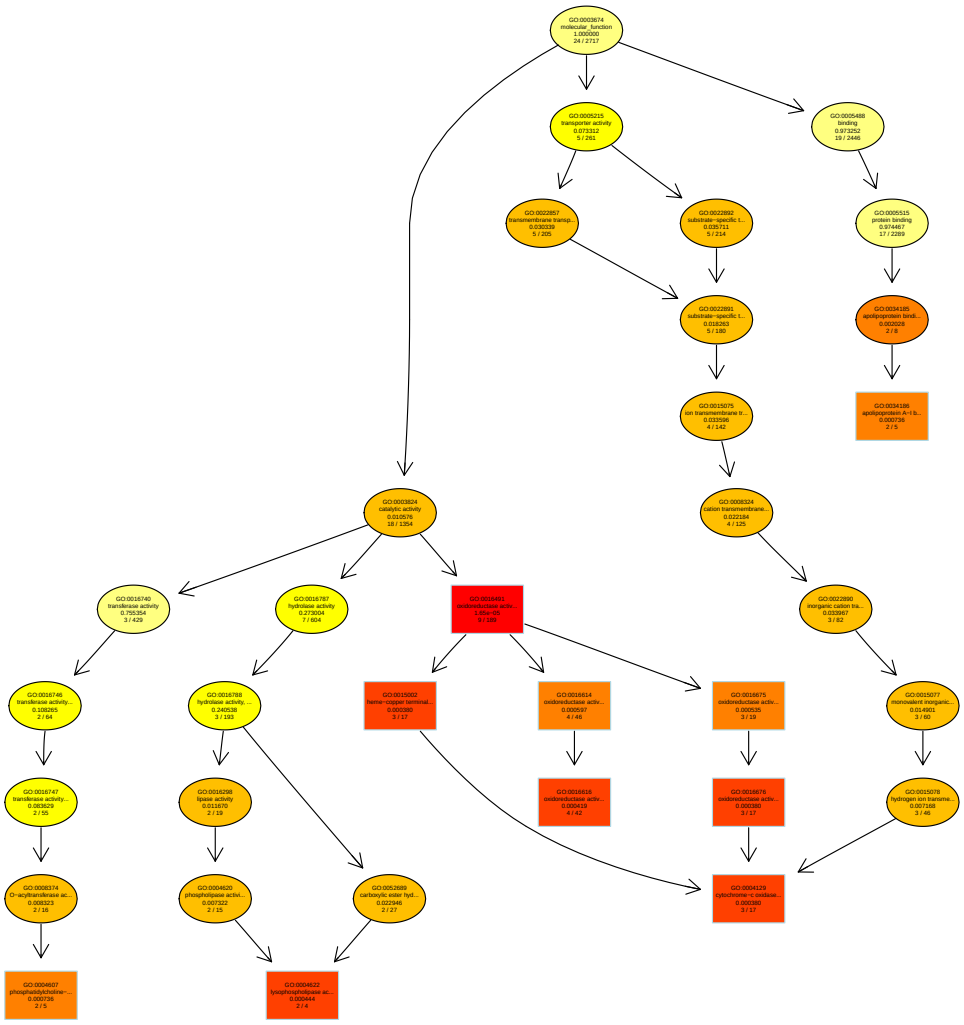

Figure 9.9: GO molecular function graph for enriched terms in DE according to worm-population - Subgraph of the GO-ontology molecular function category induced by the top 10 terms identified as enriched in DE genes between different parasite populations. Boxes indicate the 10 most significant terms. Box color represents the relative significance, ranging from dark red (most significant) to light yellow (least significant). In each node the categoy-identifier, a (eventualy 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.

Declaration

I herewith declare that I have produced this paper without the prohibited assistance of third parties and without making use of aids other than those specified; notions taken over directly or indirectly from other sources have been identified as such. This paper has not previously been presented in identical or similar form to any other German or foreign examination board.

Chapter 5 was in similar form submitted for publication to BMC Genomics, in the course of manuscript preparation Mark Blaxter edited the text.

Chapter 6 is in similar intended for publication in Plos biology, Mark Blaxter edited parts of the text.

The thesis work was conducted from May 2008 to December 2011 under the supervision of Prof. Dr. Horst Taraschewski at the Karlsruhe Institute of Technology and Prof. Mark Blaxter at the University of Edinburgh.

KARLSRUHE,

Emanuel G. Heitlinger

Redtenbacherstr. 9
76133 Karlsruhe
Germany

Phone: (+49) 721 292-5588
Email: emanuelheitligner@gmail.com

Born September, 12th in Schwäbisch Gmünd, Germany

Education

2008-2012 Doctoral studies, Karlsruhe Institute of Technology

Dissertation: Divergence of an intoduced parasite: a transcriptomic perspective on *Anguillicola crassus*

Supervisors: Prof. Dr. Horst Taraschewski and Prof. Mark Blaxter.

2007-2008 Work on diploma thesis, University of Karlsruhe, Zoological Institute, Department for Parasitology and Ecology

Thesis title: Vergleichende licht- und elektronenmikroskopische Untersuchungen am Intestinaltrakt des invasiven Schwimmblasennematoden *Anguillicola crassus* aus verschiedenen Aalarten.

2001-2007 Undergraduate studies in Biology, University of Karlsruhe.

Main subject: Zoology

Subsidiary subjects: Genetics, Botany

1991-2000 Secondary school, Privat-Gymnasium St.Paulusheim Bruchsal.

June 2000 Abitur (general qualification for university entrance).

1987-1991 Pirmary school, Kraichtal Oberöwisheim.

Employment

2008-2011 Research assistant, Karlsruhe Institute of Technology, Zoological Institute, Department for Parasitology and Ecology.

2000-2001 Alternative military service, youth centre Bruchsal.

Fields of Research Interest

Ecology and evolution of host-parastite interactions, Transcriptomics, Genomics

Research

Peer Reviewed Publications

Emanuel G Heitlinger, Dominik R Laetsch, Urszula Weclawski, Yu-San and Horst Taraschewski (2009) Massive encapsulation of larval *Anguillicoloides crassus* in the intestinal wall of Japanese eels. *Parasites & Vectors*, 2:48.

Dominik R Laetsch, Emanuel G Heitlinger, Horst Taraschewski, Steven A Nadler and Mark L Blaxter (2012) The phylogenetics of Anguillicolidae (Nematoda: Anguillicolidea), swimbladder parasites of eels. *under review BMC Evolutionary Biology.*

Conference Presentations

3rd Status Symposium, Volkswagen Foundation Funding Initiative Evolutionary Biology, November 7-11 2011, Sylt, Germany, Oral presentation: "Divergence of an intoduced parasite: a transcriptomic perspective on *Anguillicola crassus*".

2nd Status Symposium, Volkswagen Foundation Funding Initiative Evolutionary Biology, May 9-12 2010, Frauenchiemsee, Germany, Oral presentation: "The transcriptome of *Anguillicoloides crassus* sampled by pyrosequencing".

24th Biannual conference of the German society of parasitology (DGP), March 16-19 2010, Münster, Germany. 2 oral presentations: "The transcriptome of *Anguillicoloides crassus* sampled by pyrosequencing" and "Massive encapsulation of larval *Anguillicoloides crassus* in the intestinal wall of the Japanese eel".

Mind the gap: joining empirical and theoretical population genetics, October 2-3 2009, Freiburg, Germany. Oral Presentation: "Divergence between European and Asian populations of the swimbladder nematode *Anguillicoloides crassus*".

1st Status Symposium, Volkswagen Foundation Funding Initiative Evolutionary Biology, February 25-27 2009, Münster, Germany. Poster: "Divergence between East Asian and European populations of the swimbladder-nematode *Anguillicola crassus*".

Xth European Multicolloquium of Parasitology - EMOP 10, August 24-28, 2008, Paris, France. Oral Presentation: "Divergence between East Asian and European

populations of the swimbladder-nematode *Anguillicola crassus*".

Honors, Awards, & Fellowships

2008 Volkswagen Stiftung PhD Fellowship, Funding Initiative Evolutionary Biology, full funding of research position and research material

Last updated: December 11, 2011