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.)
von der Fakultät für Chemie und Biowissenschaften
des Karlsruher Institut für Technologie (KIT) - Universitätsbereich
genehmigte
Dissertation
von

Emanuel Heitlinger
geboren in
Schwäbisch Gmünd

Dekan: Prof. Dr. Martin Bastmeyer

Referent: Prof. Dr. Horst Taraschewski

Korreferent: Prof. Mark Blaxter

Tag der mündlichen Prüfung: 07. Feb. 2012

Abstract

The ability to expand into new environments and niches, despite being highly adapted to a habitual environment, is a fascinating feat of organisms. 30 years ago *Anguillicola crassus* was introduced from Asia, where it parasitises *Angilla japonica*, to Europe and spread here in the new host species *Anguilla anguilla*. Whether and how much phenotypic plasticity or rapid adaptation to differential selection are contributing to its success in invading new host-populations is a question of substantial evolutionary interest.

Gene regulatory networks, as an important link between genotype and phenotype, are thought to play a central role both in the response to stress (e.g. from as yet unexperienced environmental stressors) and in local adaptation.

In the present project, differential gene-expression in *A. crassus* populations was assessed using next generation sequencing on the 454 and Illumina platforms and genetic components of differences were isolated in cross-inoculation experiments with both Asian and European host-species and parasite populations.

Several proteases were shown to be under positive selection on the sequence level, highlighting this group of enzymes as possible targets of an immune-attack on *A. crassus*. On the gene-expression level, the extent of heritable change was large in comparison to the effect of modification in different host-environments. Mitochondrially encoded subunits of the respiratory chain and other genes connected to aerobic respiration showed divergent expression patterns in European vs. Asian parasite populations; cuticle collagen genes showed "adapted" patterns of expression in present day sympatric host-parasites pairs.

These results identified gene-expression phenotypes, confirming the divergence of European *A. crassus* populations. Such phenotypes will be more accessible to popluation-genetic analysis invetigating selection than complex life history traits.

Zusammenfassung

Die Fähigkeit sich in neuen Umgebungen und Nieschen auszubreiten, obwohl sie höchst angepasst an ihren angestammten Lebensraum sind, stellt eine faszinierende Leistung von Lebenwesen dar. Vor 30 Jahren wurde der Schwimmblasen-Nematode *Anguillicola crassus* aus Asien, wo er *Anguilla japonica* parastiert, nach Europa eingeschleppt und breitete sich hier in der neuen Wirtsart *Anguilla anguilla* aus. Ob und in wie weit phänotypische Plastizität oder die schnelle Anpassung an unterschiedliche Selektionsdrücke zum Erfolg der Invasion beitragen stellt eine Frage von großer evolutionsbiologischer Bedeutung dar.

Gen-regulatorische Netzwerke, als eine Verbindung zwischen Genotyp and Phänotyp, haben eine zentrale Rolle sowohl in der Antwort auf Stress (etwa durch eine veränderte Umwelt) als auch in der lokalen Anpassung.

Im hier vorgestellen Projekt wurden Unterschiede in der Gen-Expression zwischen Populationen von *A. crassus* mit Hilfe von neuer Sequenziertechology (454 und Illumina) untersucht und erbliche Komponenten dieser Unterschiede in einem Kreuzinfektions-Experiment mit asiatischen und europäischen Wirten und Parasiten isoliert.

Mehrere Peptidasen zeigten Spuren positiver Selektion auf der Sequenz-Ebene und heben diese Gruppe von Enzymen als ein mögliches Ziel des Immunangriffs auf *A. crassus* hervor. Auf der Expressions-Ebene überwiegen erbliche Veränderungen gegenüber Modifikationen in unterschiedlicher Wirts-Umgebung. Mitochondrial codierte Enzyme der Atmungskette und andere Enzyme in Verbindung mit aerober Atmung zeigten unterschiedliche Expression in uropäischen und asiatischen Populationen des Parasiten, Collagen-Gene der Cuticula zeigten “angepasste” Expressionsmuster in Wirt-Parsit Paaren gemeinsamer Herkunft.

Diese Resultate identifizieren Gen-Expressions Phänotypen und bestätigen die Divergenz der europäischen *A. crassus* Populationen. Solche Phänotypen werden einer populationsgenetischen Analyse, die einen Zusammenhang mit Selektion untersucht, besser zugänglich sein als komplizierte Merkmale der Entwicklung.

NATUR! [...] Es ist ein ewiges Leben, Werden und Bewegen in ihr, und doch rückt sie nicht weiter. Sie verwandelt sich ewig, und ist kein Moment Stillestehen in ihr. Fürs Bleiben hat sie keinen Begriff, und ihren Fluch hat sie ans Stillestehen gehängt. Sie ist fest. Ihr Tritt ist gemessen, ihre Ausnahmen selten, ihre Gesetze unwandelbar.

J. W. GOETHE

NATURE! [...] Incessant life, development, and movement are in her, but she advances not. She changes for ever and ever, and rests not a moment. Quietude is inconceivable to her, and she has laid her curse upon rest. She is firm. Her steps are measured, her exceptions rare, her laws unchangeable.

Translation by T. H. HUXLEY for the first issue of nature magazine.

For my grandmother Ruth and my wife Silvia

Acknowledgements

I would like to acknowledge the following persons for their tremendous help realising this project (in alphabetical order):

Mark Blaxter, Stephen Bridgett, Timothee Cezard, Yun-San Han, Karim Gharbi, Sujai Kumar, Dominik Laetsch, Jenna Mann, Anna Montazam, Trevor Petney, Horst Taraschewski, Marian Thomson, Urmi Trivedi, Urszula Weclawski, Nicola Wrobel and the thousands of unnamed individuals who have coded for free software and open source projects used in my research.

Contents

List of Figures

LIST OF FIGURES

LIST OF FIGURES

List of Tables

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 has attracted the 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 its life, and this has been proposed as one of the reasons for its success invading new continents (3).

A. crassus colonised Europe in the early 1980s 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) and currently *A. crassus* is found in all but the northernmost populations 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 its colonisation process and history. Their data are in good agreement with previous knowledge about the history of introduction

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

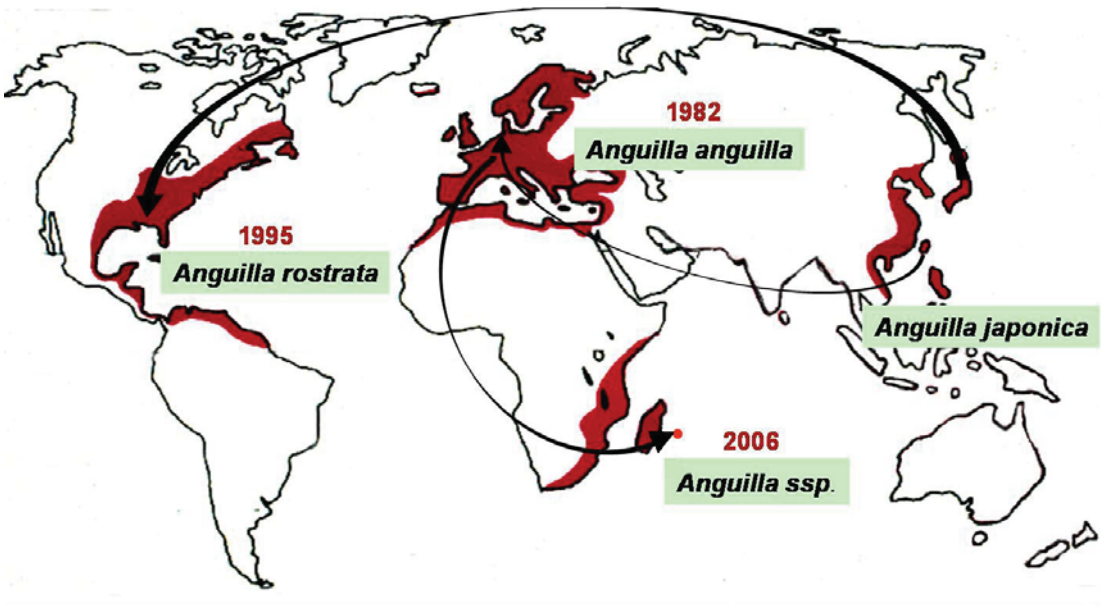

Figure 1.1: Transcontinental dispersal of *A. crassus* - Invasions of different continents by different source-populations are illustrated using arrows. Red colour indicates the range of the eel species targeted by the invasion. Modified form (11), based on data reviewed in (4) and newer findings in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

A. crassus was first recorded in 1982 in North-West Germany, and this record was published in a German fisheries magazine in 1985 (13). The import of Japanese eels from Taiwan to the harbour 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 the population genetic structure using the above mentioned microsatellites. Furthermore, from the fact that genetic diversity is highest in northern regions of Germany and gradually declines to the south, Wielgoss *et al.* (10) 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 sporadic 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 (16). These Turkish

haplotypes cluster with Taiwanese haplotypes and the introduction source would be similar to the main introduction (see also figure 1.5).

A second colonisation of *A. crassus*, succeeded in North-America: since the 1990s populations of the American eel (*Anguilla rostrata*) have been invaded as novel hosts (17, 18, 19). Wielgoss *et al.* (10) identified Japan as the most likely source of this American population of *A. crassus* using microsatellite data. Laetsch *et al.* (16) showed that all source-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 (20). 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 (21), i.a. using a trypsin-like proteinase(22). 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 its intestine and finally into the water (23). The time needed for the completion of a typical life-cycle from egg to reproducing female is important 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 leads to 2-3 generations completed per year in Europe and a total of circa 60-90 generations since introduction.

High prevalences of the parasite of above 70% (24, 25), as well as high intensities of infections have been reported, throughout the newly colonised area (26). In the natural host in Asia prevalence and intensities are lower than in Europe (27).

One of the possible differences between Asian and European population of *A. crassus* could be the widespread use of paratenic hosts in European waters (28, 29). Such a use of paratenic hosts has not been reported from the Asian range of the parasite and there is speculation that the use and availability of paratenic hosts could be a factor explaining the success of invasion or even the higher epidemiological parameters in

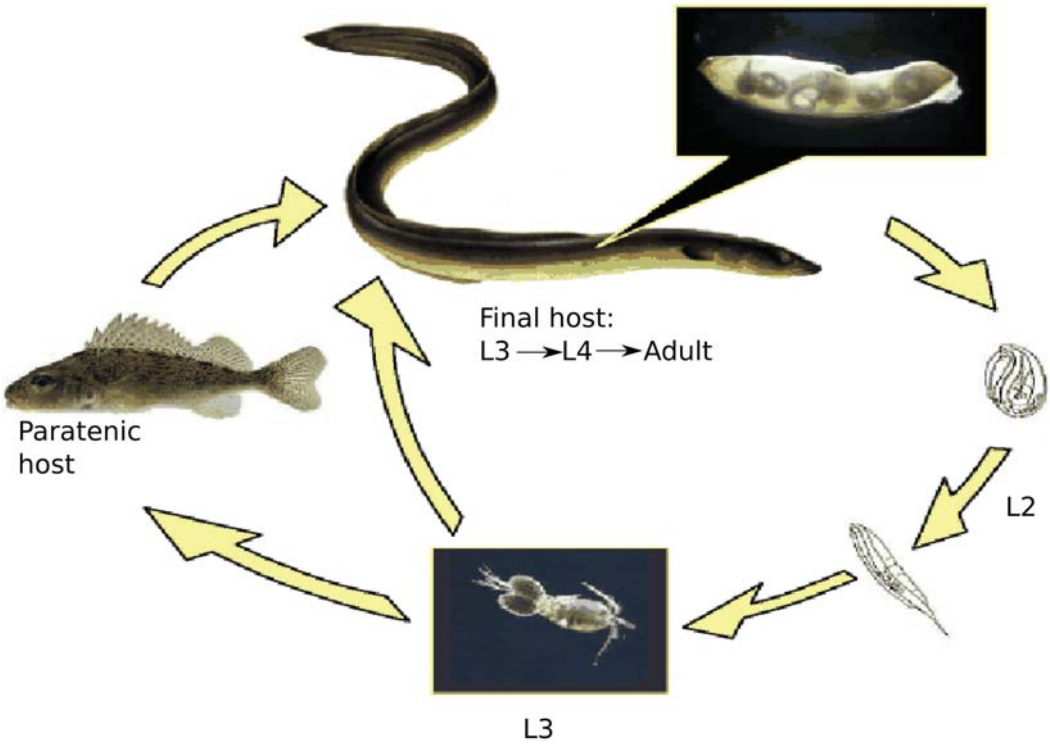

Figure 1.2: Life-cycle of *A. crassus* - Adult females deposit already hatched L2 in the lumen of the swimbladder. Larvae migrate through the *ductus pneumaticus* and the intestine into the open water. Copepods serve as intermediate host where infective L3-larvae develop. These can be transported and 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 moult to adults worms arrive in the lumen of the swimbladder, feed on blood and reproduce. Modified from (11).

Europe compared to Asia. However, the lack of evidence for the use of paratenic host in Asia is rather likely to be a result of the lack of appropriate studies in Asian water systems, given the broad spectrum of paratenic hosts used by *A. crassus* in Europe (28, 30, 31), including even amphibians and larvae of aquatic insects (32).

Also, the abundance of the final hosts *An. anguilla* and *An. japonica* itself could have an effect on epidemiological parameters (33). This host-density, however, is thought to be similar for each of two host-species in its endemic area (34), and more explicitly it is in in parallel rapidly declining for the last decades both in Asia and Europe (35).

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 definitive eel-host permitting a differential survival of the larval and the adult parasites (36, 37).

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 (38). Responses in *An. anguilla* show hallmarks of pathology, including thickening (39) and inflammation (40) of the swimbladder wall, infiltration with white blood cells and dilated blood vessels.

Especially these changes in the tissue of the swimbladder wall have been shown to influence swimming behaviour and it has been speculated that infected eels may fail to complete their spawning migration (41). While nobody would claim Anguillicolosis (the condition caused by *Anguillicola* infection) to be the main reason for the decline of eel stocks, it could very well be a cofactor (42) adding to the main factor of overfishing of glass-eels (35).

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 (although high) range (43).

In contrast to the European eel, the Japanese eel is capable of killing larvae of the parasite after vaccination with irradiated larvae (44) or under high infection pressure. Such mortality of *A. crassus* larvae has been reported in the swimbladder wall of *An. japonica* in the wild (27) and under high infection pressure even more pronounced in the intestinal wall (45).

Furthermore, it has been shown that the establishment of encapsulated larvae inside the intestinal wall is related to the death 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 were found in single, non-repeated experimental infections of Japanese eels, while larvae are killed in the swimbladder wall. These observations show 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 (45) and thus a prominent role of host-immunity in the natural host *An. japonica*.

Interestingly, the differences in the two host-species also affect the size and life-history of the worm: in European eels the nematodes are bigger and develop and reproduce faster (37).

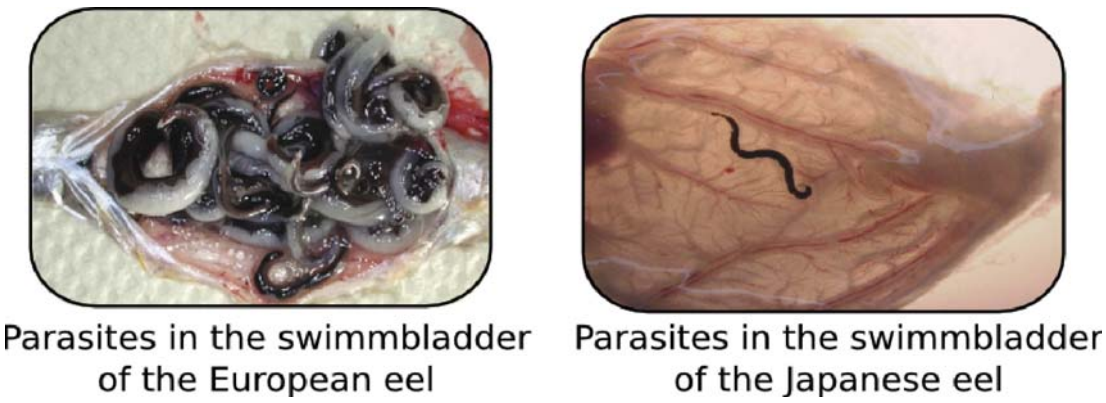

Figure 1.3: Difference between worms in the swimbladder of the European eel and the Japanese eel. - Note the bigger size and higher number of worm in a typically infected European eel. In comparison in the Japanese eel worms are smaller and intensities of infection are much lower. The dark brown matter 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 *Anguilla* spp. 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 (46). The Japanese eel in its endemic area migrates to the west of the southern West Mariana Ridge (47). 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 (48), European eels as a species are considered panmictic (49): signals for population structure, initially interpreted as evidence against panmixia (50), have been shown to be an artefact of temporal variation between cohorts of juvenile eels (48, 51, 52). Such panmixia reduces the effectiveness of selection. Uninfected populations participating in reproduction make rapid local adaptation to a parasite less likely.

Interestingly it has been shown, that individual genetic heterozygosity in *An. anguilla* is no predictor for *A. crassus* infestation (53). 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 (54, 55). Variation at highly polymorphic loci is one of the cornerstones of host-adaptation (56). Once variation is present in a population, overdominance (or heterozygote superiority) can favour heterozygous individuals (57, 58). Matching parasite antigens and allowing them to be presented as an epitopes on professional antigen presenting cells, the MHC class II molecule, for example, has been demonstrated to be under diversifying selection in many vertebrate species. Sticklebacks 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 (59). Conversely the vertebrate immune system, and especially its memory component, are thought to be driving positive selection on antigens of microorganisms (60).

Morphological and functional differences between the immune systems of teleost fishes and other vertebrates (especially mammals) are prevalent (61). 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 (62). 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 (63), this gene shows conservation in the adaptive immune system of jawed vertebrates (64) and its presence has been confirmed in transcriptome data for *An. anguilla* (65).

A decline of prevalence and mean intensities for European populations of *A. crassus* has been hypothesised 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 (33)), evolution of the eel host towards better resistance (rather unlikely; see above), and 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* is the only genus in the family Anguillicolidae. It comprises five morphospecies (66): in East Asia, in addition to *A. crassus*, *Anguillicola globiceps* Yamaguti, 1935 (67) is found in *An. japonica*. *Anguillicola novaezelandiae* is endemic to New Zealand and South-Eastern Australia in *Anguilla australis* and *Anguillicola australiensis* Johnston et Mawson, 1940 (68) parasitises the long-fin eel *Anguilla reinhardtii* in North-eastern Australia. Finally *Anguillicola 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 of swimbladder-nematodes but *A. globiceps*, to the rank of a genus (69). In the meantime 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* (16). In this study, *A. crassus* was identified as the basal species in the genus, analysing 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 places *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

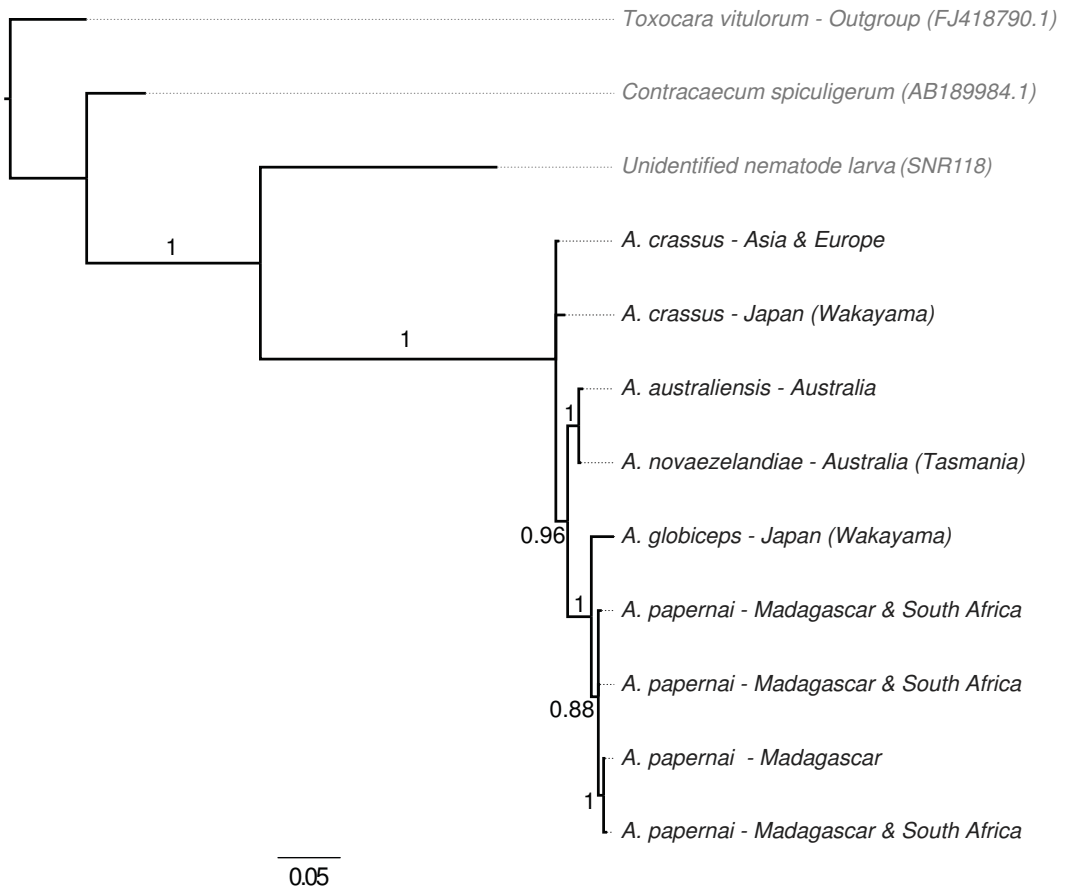

Figure 1.4: Phylogeny of the genus *Anguillicola* based nLSU - Phylogram inferred from nuclear large ribosomal subunit (nLSU) of *Anguillicola* and outgroups using Bayesian inference. Labels on internal branches indicate Bayesian posterior probabilities. From (16).

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 (70): 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 a 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 separated hosts. This affinity for host-switching may be an evolutionary relic found only in one of the two clades (putative cryptic species) into which *A. crassus* can be divided (16).

The to date most likely phylogenetic hypothesis places the genus *Anguillicola* at a basal position in the Spirurina (clade III *sensu* (71)), one of 5 major clades of nematodes (72, 73). 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 (16) 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 favours a complex life history as the ancestral state for the Spirurina.

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

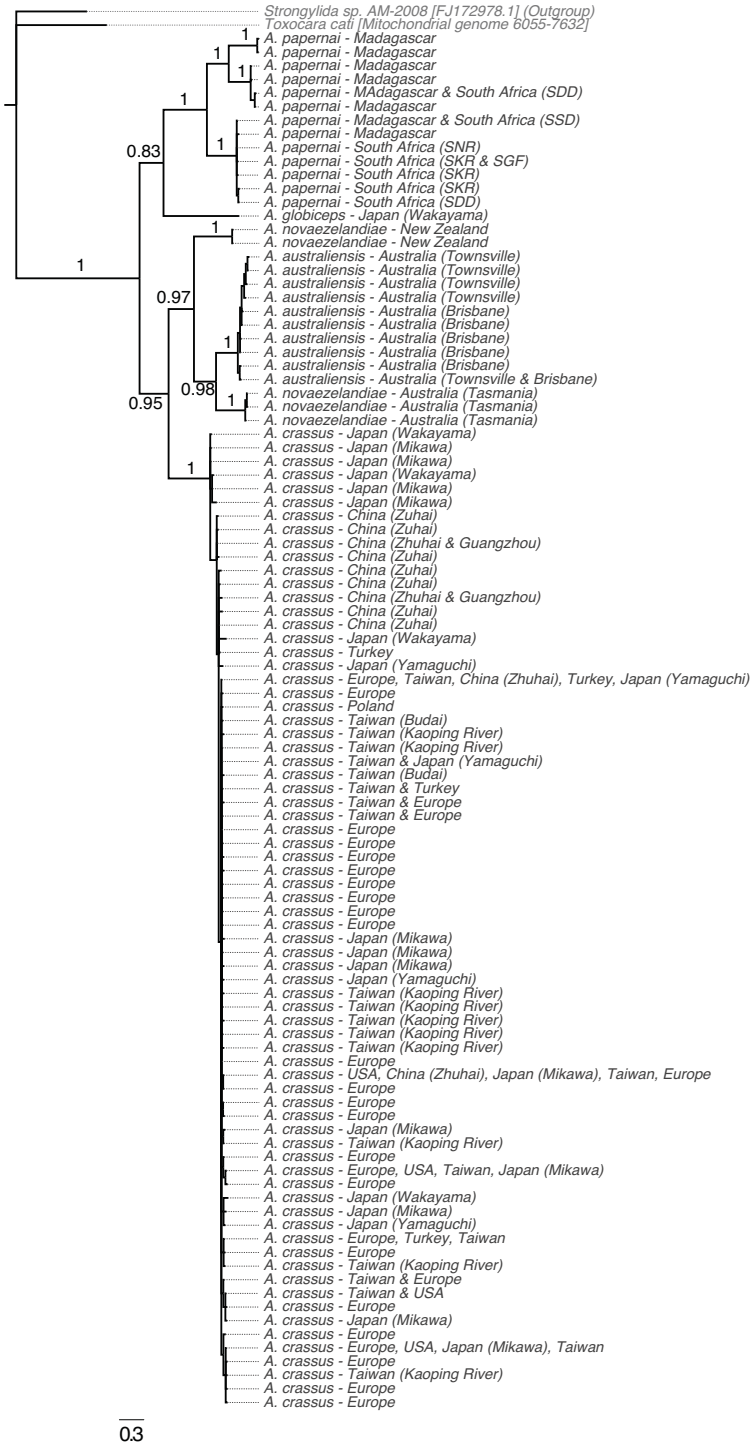

Figure 1.5: Phylogeny of the genus *Anguillicola* based on COXI - Phylogram inferred for *Anguillicola* and outgroups based on mitochondrial cytochrome C oxidase subunit I (COXI) using Bayesian inference. Labels on internal branches indicate Bayesian posterior probabilities. From (16).

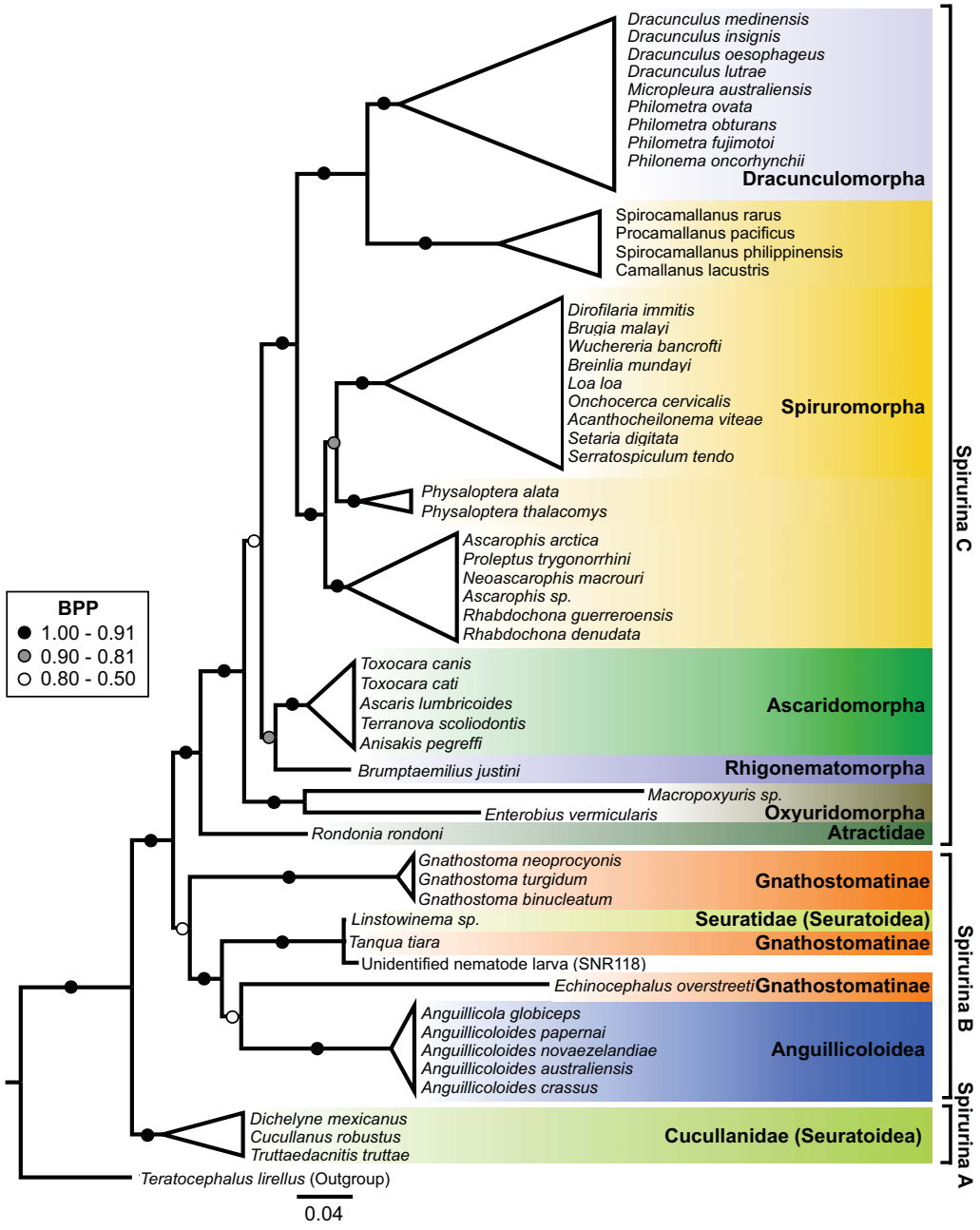

Figure 1.6: Phylogeny of nematode clade III based on nSSU - Phylogram inferred from nuclear small ribosomal subunit for Spirurina using Bayesian inference. Branches are collapsed to highlight major groups. Labels on internal branches indicate Bayesian posterior probabilities. From (16).

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

The phenotype of an organism can respond to changes in the environment through either evolutionary or nongenetic processes. Common-garden and transplant experiments are a method to separate genetic components (G) of phenotypic differences from environmental (E) influences. They have been used for almost as long as scientists have investigated evolution (74, 75).

The goal of a classical common garden experiment is the exclusion of environmental factors: by carefully choosing a 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 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.

These limitations to 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 two-factorial design in which interactions between genotype and environment (GxE) can be incorporated into an analytic 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 one hosts-species (or vice versa) such an experiment can be best described as an "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* spp. project it is the parasite, as definitive genetic differences between the host-species are not part of 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" (76). The infection of both host-species with both parasite populations allows the incorporation of genotype by genotype (GxG) effects into an analytic model. This approach is chosen in the experiments presented in this thesis.

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

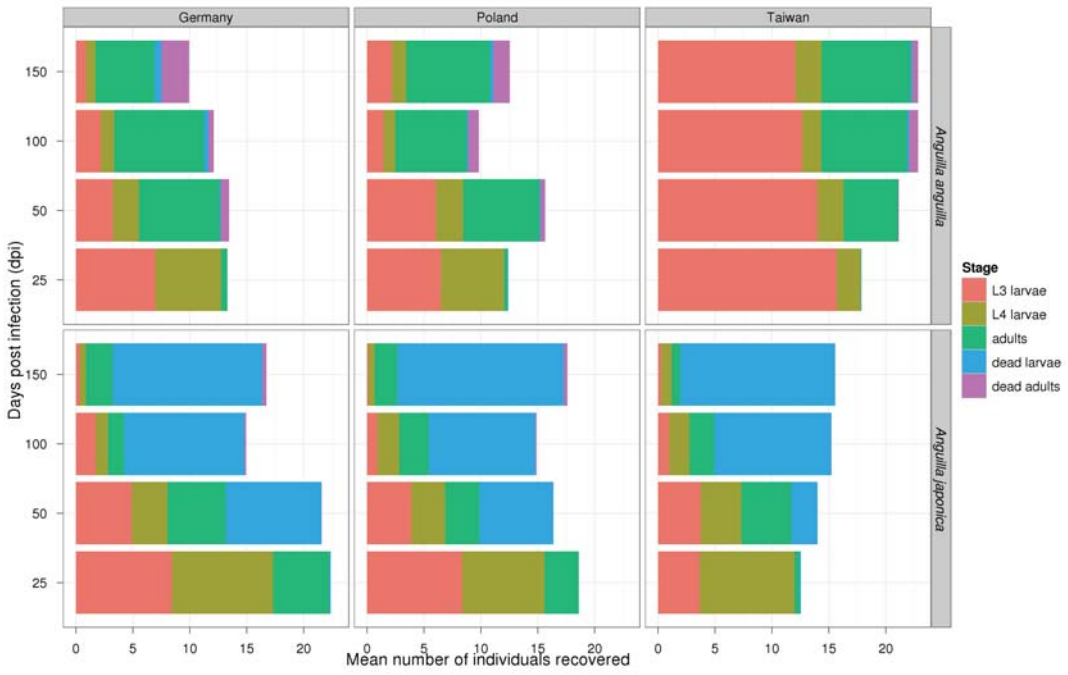

Figure 1.7: Differences in developmental speed - Three populations of *A. crassus* (panels in columns) were raised in two different hosts (panels in rows). Eels were dissected at 4 different time points post infection (dpi). Bars represent means of recovered individuals from three different life-cycle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

In these experiments differences between the two European populations and the Taiwanese population of worms were examined. 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 at up to 50 dpi there were less living adults observed for worms from

the Taiwanese population and fewer dead adult worms were recorded for the Taiwanese population beginning from 50 dpi.

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 show 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 an 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 (77). The inferred faster development of the European population of *A. crassus* can thus be regarded as a highly interesting candidate-phenotype for adaptation. However, the slightly delayed development of the Taiwanese population even in the natural host *An. japonica* would constitute an maladaptation (78) 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 interactions as predicted by the theory of geographic mosaic of coevolution (79). Such an experimental design, a "reciprocal cross-inoculation under reciprocal transplant conditions" (80), 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 not to 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 thus is largely the object of 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

Two kinds of macromolecules carry all the information evolution has 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 a transient manifestation of this information (81). In all cellular life, genetic information flows from replicating DNA to RNA in a process called transcription and from RNA to protein in a process called translation (82) (see figure 1.8).

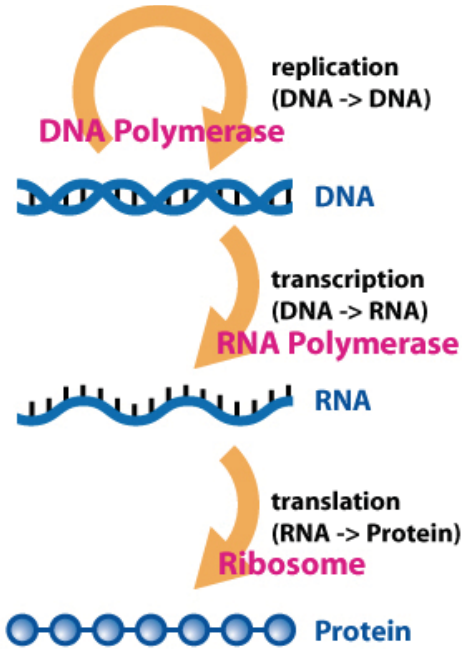

Figure 1.8: Major macromolecules bearing biological sequence information - A schematic view of the flow of genetic information in a cellular life: enzymes (red font) process macromolecules carrying genetic information from DNA to RNA, from RNA to protein. Picture from wikipedia.

The relatively inert DNA is adapted to carry information over generations and to limit the number of mutations (also by evolving low error in polymerase) (83). The single stranded, more reactive RNA, on the other hand, can create secondary structures by base-pairing with itself or other (macro-) molecules. It is involved in numerous cellular processes making use of this reactivity (84): microRNAs (miRNAs) regulate translation by binding mRNA, initiate degradation and thus decrease its levels (85, 86), 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 (87). In addition, a variety of poorly understood other non-protein coding RNA (ncRNA) families exist (88). 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 a codon recognised in the coding sequence (CDS) of a messenger RNA (mRNA) molecule and a protein is synthesised (89).

These mRNAs (like the untranslated RNAs above) have been transcribed from genomic DNA (see figure 1.9). Eukaryotic mRNAs have a special structure to preventing and regulating 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.

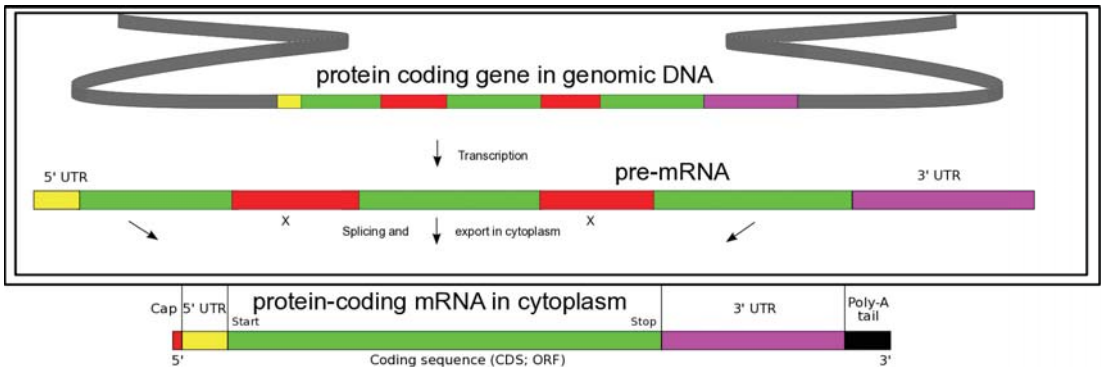

Figure 1.9: The structure of a protein coding gene and its mRNA - A schematic view of posttranscriptional modifications in an eukaryotic gene. Introns are spliced, 5' and 3' structures are added and the mRNA molecule is exported into the cytoplasm. Note that the double stranded nature of the genomic DNA (grey) is not indicated in this diagram 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 interspersed in coding regions. This excision 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 (90). Only after the processing of pre-mRNA to mature mRNA is the molecule released into the cytoplasm where it eventually can be translated (see above).

The complete set of transcripts in a cell is called the transcriptome. One of the major goals 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 categorisation 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 (mRNA): finding start sites for both transcription from the genome and for translation into protein, 5' and 3' ends, splicing patterns and other post-transcriptional modifications (91).

Transcriptome-projects and transcriptomic data have been invaluable in determining the structure of the genome (information gained from the transcriptome provides information about genomic features), but they are also at the centre of one of the major challenges in biology linking genotypes to phenotypes. The "expression" of the gene in a literal sense would be the phenotype visible for natural selection. It is known that posttranslational modification and 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 (92, 93). However, overall levels of mRNA abundance correlate well with protein abundance (94). Measurement of mRNA levels is methodically less demanding than measurement of protein 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 mention of protein sequences in the results of this document are derived from computational prediction based on the 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 (more or less specifically) mRNA is reverse transcribed into complementary DNA (cDNA). The RNA-dependent DNA-polymerase (reverse transcriptase) used for this process is originally found in retroviruses. Amplification and reverse-transcription protocols often achieve (more or less) specific amplification of mRNA from the other RNA species using its poly-A tail as primer or adapter binding site.

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

For almost three decades the method developed by F. Sanger (95) 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 labelling of ddNTPs allowed all four reactions to be performed together. Additionally modern machines use the chain-termination method combined with capillary gel electrophoresis (96) in a highly parallelized way.

Due to these advancements it was possible to tackle the sequencing of bigger genomes, than those of the phages in the first years of DNA sequencing (97): the bacterium *Haemophilus influenzae* in 1995 (**?**), the baker's yeast *Saccharomyces cerevisiae* in 1996 (98), the nematode *Caenorhabditis elegans* in 1998 (99), the fruit fly *Drosophila melanogaster* in 2000 (100) and the mouse *Mus musculus* in 2002 (101) were the first cellular organisms with sequenced genomes. For these laboratory model-organisms, multi-national consortia were needed financing and coordinating sequencing and analysis in multi-million dollar projects. This "first generation of genomics" culminated in the publication of the human genome in 2001 (102).

In parallel to the mentioned genome-projects, transcriptome projects were conducted. Single pass Sanger-sequencing reads called expressed sequence tags (ESTs) were mapped to genomic sequence, identifying coding regions (103). First estimates of the number of genes in the human genome, for example, were based on the extrapolation of the number of genes found with this method in early sequenced regions of the genome (104).

Costs and labour 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 the costs and labour associated with this method prohibitive for a large scale application in non-model organisms.

1.2.3 DNA-sequencing in nematodes

As mentioned above in 1998 *Caenorhabditis elegans* had become the first multicellular organism with a sequenced genome (99). Soon it was noted that in addition to its use as a general model system for the metazoa and beyond, knowledge gained in this species

has the potential to be even more valuable within the phylum nematoda (105). The breadth and detail of genomic information available for *C. elegans* to date is illustrated by a recent publication, using transcriptomics to provide detailed annotation of the diverse functional genomic elements and their interactions at single base resolution (106). With this amount of data digested into usable information *C. elegans* continues be an invaluable resource in nematode genomics: 21,000 protein coding genes, over 5,000 RNA genes and 100.2 megabases (Mb) of overall sequence still provide the most thoroughly investigated comparative basis for new genome or transcriptome projects started in the Nematoda.

The genome sequence of *Caenorhabditis elegans* was soon complemented by the genome of *Caenorhabditis briggsae* (107), a second nematode from the genus sequenced as a satellite-system for comparative genomics. As a second more distant satellite-model in clade V the necromenic *Pristionchus pacificus* (living in close association with beetles) was sequenced (108).

The first published genome of a parasitic nematode in the Spirurina was the genome of *Brugia malayi* (109), and only very recently, as second parasite from this clade, *Ascaris suum* had its genome published (110).

Also in the remaining clades of the nematoda genome sequencing flourished: for the animal-parasite *Trichinella spiralis* from clade I (111), the plant parasites *Meloidogyne incognita* (112) and *Meloidogyne hapla* (113), as well as the the pinewood nematode *Bursaphelenchus xylophilus* (114) (a plant parasite using a beetle as an vector) from clade IV genome sequences have been recently analysed 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 analytic purposes on a `Blast`-server (115).

Before the advent of next generation sequencing (NGS; see 1.2.4), the lack of genomic information on many species of nematodes promoted the use of ESTs as a tool for gene-discovery. Partial genomes *sensu* (116) were successfully searched for a large array of genes interesting to various scientific communities. In nematode parasites of vertebrates, pathogenic factors were described as potential vaccine candidates (117). 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 (117). 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 (118) (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) (119) and venom like allergens (Bm-VLA) (120).

In some of these studies, secreted proteins were in the centre 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* (121), and across all nematode ESTs (122).

Over the years sequence information derived from EST-data and whole genome sequencing has been collected and updated into the nembase transcriptome databases (123, 124). The recent compendium nembase4 describes clustering of 679,480 raw ESTs in 233,295 clusters from 62 species (125). This database provides an invaluable collection of confirmed information for comparison, validation and hypothesis generation when new transcriptomes are analysed as in the present project.

Obviously, NGS currently also leaves its mark in nematode transcriptomics: NGS analysis on the transcriptomes of *Ancylostoma caninum* (126), *Pristionchus pacificus* (93), *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*, a pyrosequencing-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 a rapid and cost-effective generation of genome-scale DNA-sequence data. Labour and costs associated with DNA-sequences have been drastically reduced during the last 5 years (see figure 1.10).

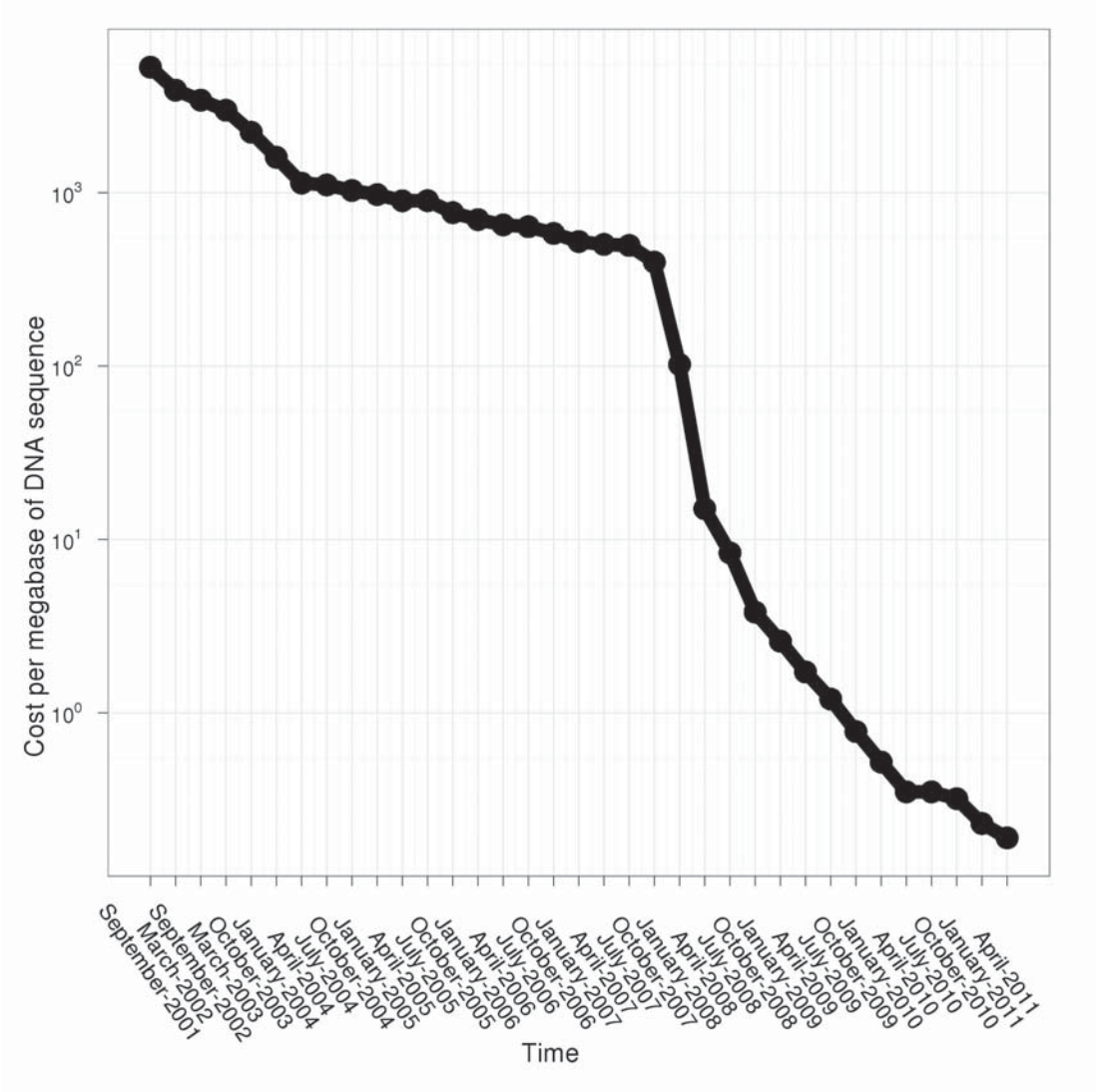

Figure 1.10: Falling sequencing costs - First (till 2005) sequencing costs were falling due to the improvements in Sanger-sequencing and the invention of pyrosequencing. Later (since 2008) advances in Solexa-sequencing are beating down the price. Due to improved read-length and throughput on this platform 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 not work on single molecules and thus target molecules have to be amplified as in Sanger-sequencing. This amplification has to produce spatially separated templates. In particular, new methods to address this methodological need are at the heart of the new technologies. Immobilisation on a solid surface to archive clonal amplification is used in the 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.

1.2.4.1 Pyrosequencing

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 (133) (see figure 1.11), the DNA is ligated to adapters, denatured and single stranded molecules are attached to a complementary sequence on a bead. An emulsion of beads in oil together with enzymes under conditions that favour 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 fibre-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 (134). 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 oxidise luciferin and to generate light. The added nucleotide is known as the nucleotides are flushed over the plate one at a time. A high resolution camera records the emission of light. The intensity of emitted light is proportional to the number of nucleotides incorporated.

The ability to distinguish the length of homopolymeric runs of the same nucleotide decreases with the length of such homopolymer runs (135). Current “Titanium chemistry” is producing reads of > 350 bases length, “FLX chemistry” (used up to 2009) was able to produce reads of roughly 250 bases length (136).

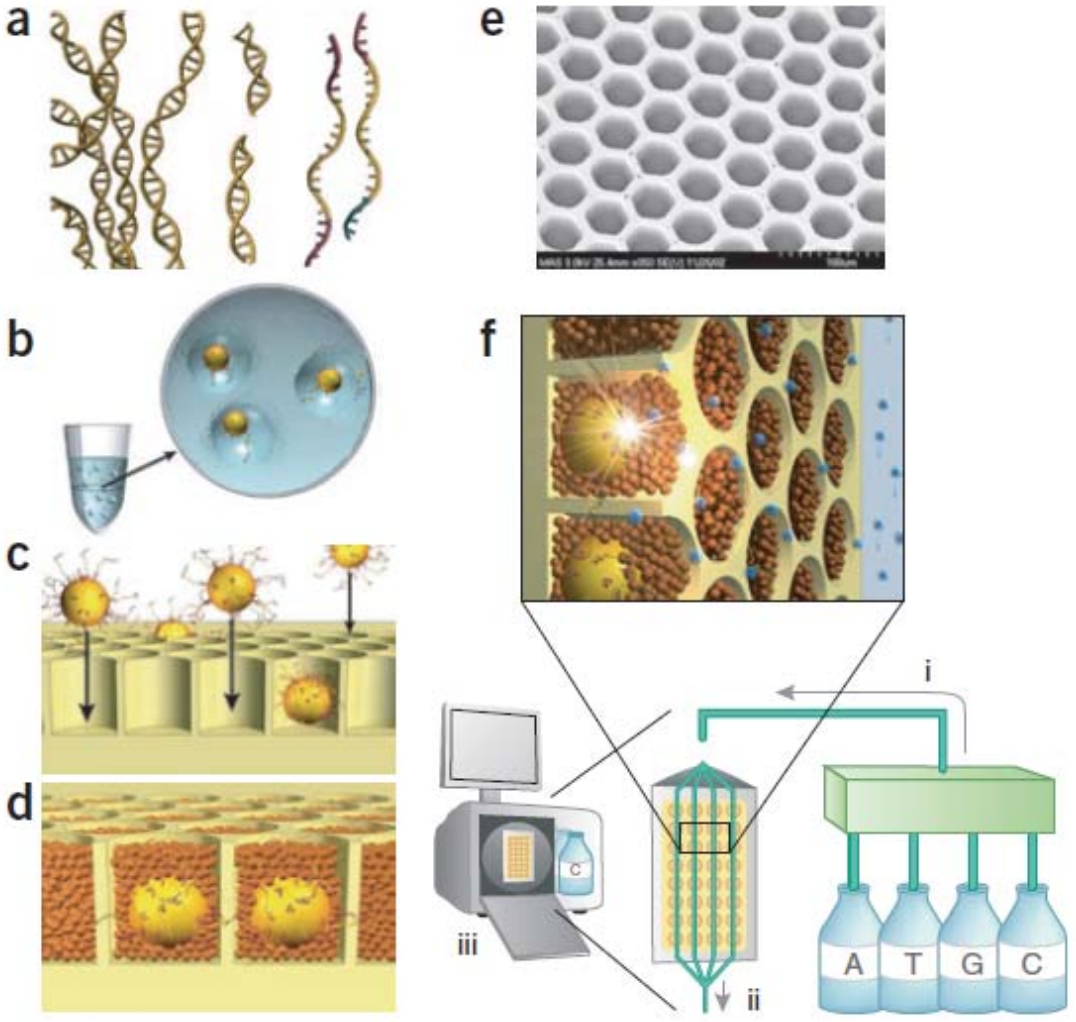

Figure 1.11: Schematic representation of pyrosequencing - (a) DNA (genomic or transcriptomic) is isolated, fragmented, ligated to adapters and denatured into single strands (b) Under conditions that favour one fragment per bead fragments are bound to beads. These beads are isolated and compartmentalised in the droplets of an emulsion and PCR (a mixture of reagents in oil). Within each droplet DNA is amplified, and beads are obtained carrying millions of copies of a unique DNA template. (c) After denaturation of DNA, beads are deposited into wells of a fibre-optic slide (called picolitre plate). (d) Immobilised enzymes carried on smaller beads are added to each well and a solid phase pyrophosphate sequencing reaction is initiated. (e) A portion of a fibre-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) (132).

This longer read length of 454-sequencing (137) compared to other NGS technologies (see 1.2.4.2), allows *de novo* assembly of transcripts in organisms lacking previous genomic or transcriptomic data (127).

1.2.4.2 Illumina-Solexa sequencing

Illumina-Solexa technology is to date (Dec. 2011) the most competitive commercial sequencing platform, 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 synthesised. Subsequently the adapter sequence at the 3' end of newly synthesised 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 (138).

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 glass 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 fluorophore along with a nucleotide is detected using a high resolution camera after laser-induced excitation. The fluorophore is removed and next cycle is initiated (138).

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

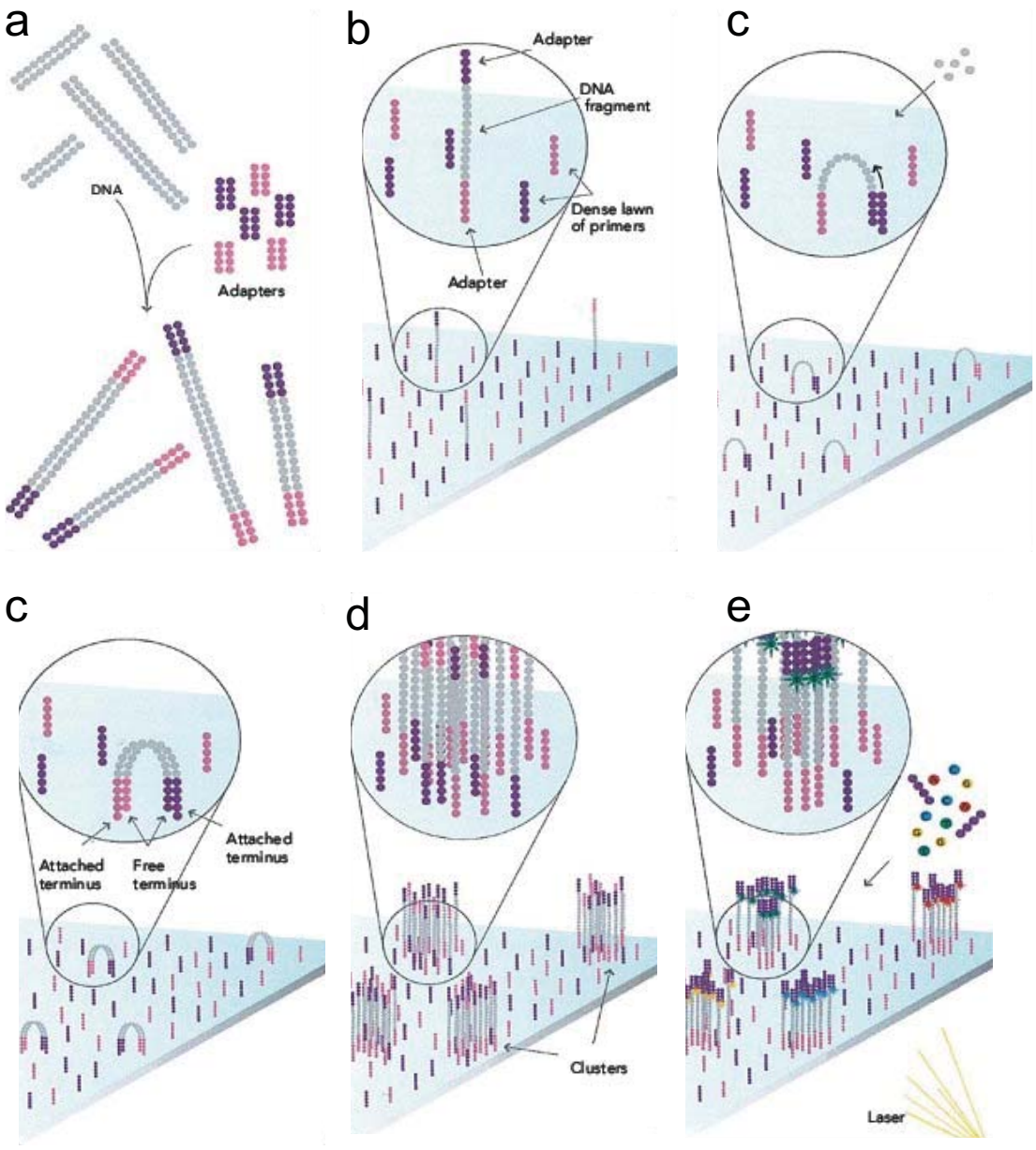

Figure 1.12: Schematic representation of Illumina-sequencing - (a) DNA (genomic or transcriptomic) is isolated, fragmented and ligated to adapters. (b) Single stranded fragments are bound to a glass-slide. (c-d) Solid-phase bridge amplification using unlabeled nucleotides, primers (binding the adapters) 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. An image of the emitted fluorescence under laser excitation is taken . Step (e) is repeated multiple times for the length of the DNA-sequence. Modified from Seqanswers-forum.

strand then acts as a template for the second sequencing reaction producing a 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) (138).

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 the assembly of large eukaryotic genomes (e.g. that of the giant panda (139)) and transcriptomes (140) (but see also 1.2.5 for methodical challenges). In the same period throughput also increased from roughly 6,000,000 reads in 2008 to roughly 20,000,000 reads in 2011 on one of six lanes (compartments of the glass-slide) of the instrument.

The high throughput of the Illumina-Solexa platform also makes it first choice for gene expression analysis (141): RNA-seq has revolutionised transcriptomics both in model and non-model organisms (91), replacing microarray technology as the method of choice for gene-expression measurements (142). SuperSAGE (143) using expression-tags provides the benefit of classical SAGE-analysis (144) with those of the ultra high throughput of Illumina-Solexa sequencing.

1.2.5 Computational methods in DNA-sequence analysis

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

This has fostered the 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 (145)).

In the context of computational tools this common characteristic of all DNA-sequencing methods has to be emphasised: 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 sequences. 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 was - shortly after sequencing chemistry - described by F. Sanger (146). Soon computer programs were necessary to align sequences and to compute overlaps and consensus sequences (147), this process of computationally reconstructing the target molecule was termed sequence-assembly (148). The reconstructed target molecules are termed contigs, derived from contiguous sequence. In an (hardly achieved) optimal genome-assembly a contig would thus represent a chromosome, in an optimal transcriptome assembly there would be a contig for every transcript of the organism.

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 (149), the second "Needleman-Wunsch" global alignment (150). Of course these alignment methods have usages outside of sequence assembly in general sequence comparison, including protein sequence.

The program `Blast` (151), for example, enables the 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 normalised 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 its mapping equivalent `Blat` (152) 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 themselves to this sub-problem. In mapping only the positions (and the qualities) of a match relative to an already sequenced longer contig are investigated. `Ssaha2` (153) is able to speed up such sequence searches by orders of magnitude. It builds a hash table indexing k-tuples (k contiguous bases, an approach implicitly also used in the seeding step of `Blast/Blat`). Then sorting of matching indices shows regions of high similarity without an alignment, but these regions can then be aligned using a banded Smith-Waterman algorithm.

`Burrows-Wheeler Aligner` (`BWA`) (154) 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 construction 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 (155).

On top of the complexity found in the *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 unevenly across their length. (b) In highly expressed transcripts more erroneous bases are found in total. (c) Transcripts from adjacent loci can overlap and can be erroneously 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 (156).

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 (157).

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.

The downstream analysis of assembled sequence is also highly complex and processing of potentially biased, multidimensional data into biological relevant knowledge provides additional computational and statistical challenges.

Inference of single nucleotide polymorphism (SNPs) requires statistical categorisation in true polymorphisms and sequencing errors. Tools like `VarScan` (158) or `VCFtools` (159) 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` (160) additionally uses 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 (161).

Assessment of the statistical significance of differences in read counts (from transcriptomic data; also called "digital transcriptomics"), needs some special treatment in comparison to the well established methods for microarray-data (162). 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 (163) instead of a normal distribution for microarray data. To allow testing for low numbers of replicates, the software 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` (164), `edgeR` (165) and `baySeq` (166).

The functional interpretation of results (from SNP-calling or digital transcriptomics) linking them to biological meaningful annotation needs a standardised vocabulary in a datastructure across species and databases. Gene ontology (GO) provides such a vocabulary of controlled terms. The terms are organised in a directed, acyclic graph. This means, that a hierarchical structure links lower level "child"-terms (more specific) to higher level "parent"-terms (less specific) through a standardised set of directional relations. Back-links forming circles are not allowed (167, 168). For exp ample, "endopeptidase activity" "is a" "peptidase activity", not the other way round. The "is a" in the previous sentence is such a directional relation and other possible links would be e.g. "part of" or "regulates".

1.2.6 Applications in ecology and evolution and gene-expression divergence

Pyrosequencing in particular has been used to study the transcriptomes of organisms with ecological and evolutionary significance. Numerous studies have characterised transcriptomes to enable further research in such species (reviewed in (169)). Many of them are comparable to chapter 5 of this thesis. In addition to general annotation

often expression levels are compared between libraries, SNPs and genetic variation is identified and correlations of these "measurements" are investigated. A dedicated experimental approach using a transcriptomic readout, like presented in chapter 6 of this thesis, is not as common yet. Nevertheless, without the aim to be comprehensive, some examples should be mentioned.

A study on two phylogenetically distant mangrove species chose to sequence the transcriptomes from their natural habitats. Comparing expression levels of the two species convergent evolution of gene expression was found and connected to the ecological niche. From the fact that closer relatives of both studied species, living in different ecological niches, do not show the same similarities, the study concluded an adaptation of gene expression to the similar environment (170).

A study on trout in Lake Superior (171) used an approach similar to that used in the work presented here: Fish showing 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. Expression levels estimated from 454-sequencing did not correlate well with expression-levels estimated from reverse transcription quantitative polymerase chain reaction (RTqPCR).

In the seagrass *Zostera marina* northern and southern populations were subjected to heat stress for a short time-period in a common garden setup. The transcriptome was analysed using pyrosequencing both during and after the heat wave. From different patterns of not the direct response to heat but the resilience of expression patterns after a heat wave the authors concluded an adaptation of the southern population to heat. The ability to return to normal expression levels after a perturbation event was furthermore hypothesised as the adapted trait (161).

Other aspects of the central questions regarding the evolution of gene-expression levels are better addressed in laboratory model-organisms. In *Drosophila*, for example, variation of gene-expression (measured using RNA-seq) within a single species has been shown to be more attributed trans-regulatory elements, while expression divergent between species is dominated by cis-regulatory differences (172). In general the perceived gap between laboratory and ecological model-organisms is closed from both sides. One

side is the establishment of genomic and transcriptomic data for thus far (by molecular biologists) neglected organisms interesting because their evolutionary ecology. On the other side laboratory model organisms are more and more put in their ecological context, as exemplified by the (above mentioned) investigation of natural variability in free living strains and species of *Drosophila* or the analysis of polymorphism in natural populations of *C. elegans* (173).

Before the advent of NGS investigations on the evolution of gene expression in both laboratory and ecological/evolutionary model organisms used microarray technology. For example, fitting with the above mentioned research, sterility of hybrids between species of *Drosophila* has been shown to result from incompatibilities in gene-regulatory networks (174). A more detailed discussion of results from such studies, as related to my work, is provided in the discussion (chapter 7.3) of this thesis.

2

Aims of the project

2.1 Preliminary aims

In order to investigate the response of the transcriptome to environmental stimuli or alternatively, a genetic fixation of such a response, the responding units (transcripts) had to be established first. Ensuring the quality of these computationally constructed transcript-models (contigs) and screening for host- and other xenobiont-derived sequences were central aims of this preparatory part of the project. These goals were pursued using bioinformatic analysis of Sanger- and pyrosequencing data, with the aim of guaranteeing reliable inference based on this reference data.

2.2 Final aim

Not only gene-expression studies were enabled based on the sequence of this reference transcriptome, but also questions could be addressed regarding general aspects of the evolutionary biology of *A. crassus*. Aims addressable at the sequence-level were the characterisation of the transcriptome in relations to related parasitic nematodes and the inference of positive selection using data on polymorphism.

The genetic component of expression differences was then elucidated in reciprocal transplant experiments. As the final aim of these experiments, the relative contributions of physiological plasticity of gene-expression versus rapid, heritable, evolutionary change will be illuminated. I hypothesise that divergent expression phenotypes between European and Asian populations will be found.

3

Pilot sequencing (Sanger method)

3.1 Overview

This chapter reports a small pilot-project investigating the RNA-extraction and cDNA preparation in preparation for high-throughput transcriptome sequencing of the swimbladder nematode *A. crassus*. I generated expressed sequence tags (ESTs) using traditional Sanger-technology and conducted a first assessment of the sequence diversity expected in deeper sequencing. Especially the expected coverage of 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.

3.2 Initial quality screening

The initial quality screening revealed a high number of sequences that had to be discarded due to failed sequencing reactions (sequences being too short after quality trimming by `trace2seq`) in the library prepared by students. For sequences of *An. 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 *An. japonica* 64 (22.22%) of the sequences were excluded because of homopolymer-runs considered to be

artificial. This resulted in 452 of the nematode and 195 of the host reads being regarded of sufficient quality for further processing after base-calling and quality screening.

3.3 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 the 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 number of sequences being of sufficient quality for submission to NCBI-dbEST. At this point for the *An. japonica*-dataset, 36 sequences were submitted to NCBI-dbEST under the Library Name "*Anguilla japonica* liver" and were assigned the accession LIBEST_027503.

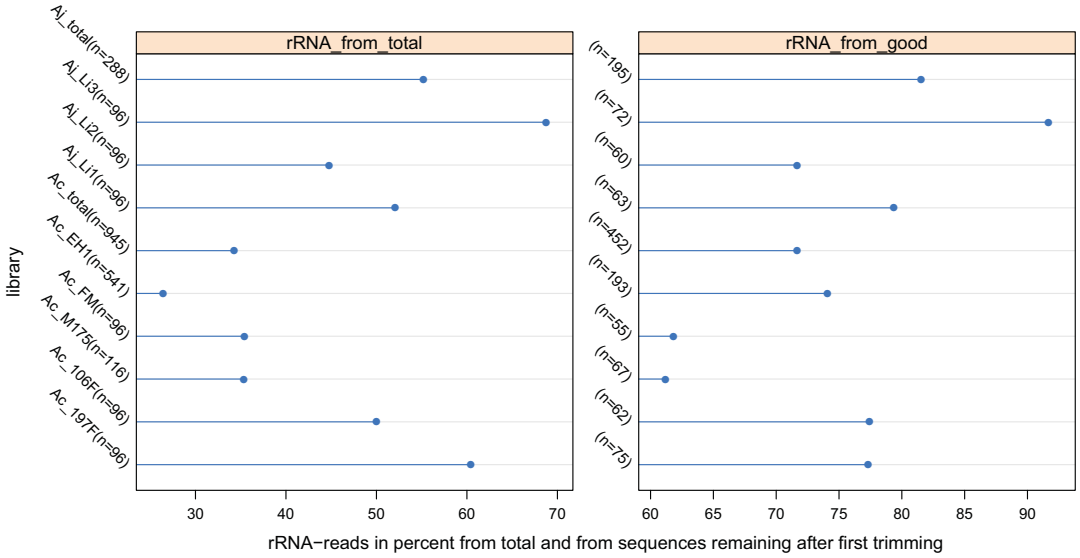

Figure 3.1: Proportion of rRNA in different libraries for *A. crassus* and *An. japonica* - rRNA abundance as proportion of the raw sequencing-reads (rRNA from total) and as proportion of the reads after quality screening (rRNA from good). Libraries starting with "Ac_" are from *A. crassus*, libraries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

	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 submitted to NCBI-dbEST.

3.4 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 large 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 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 essentially to 115 ESTs. However, 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

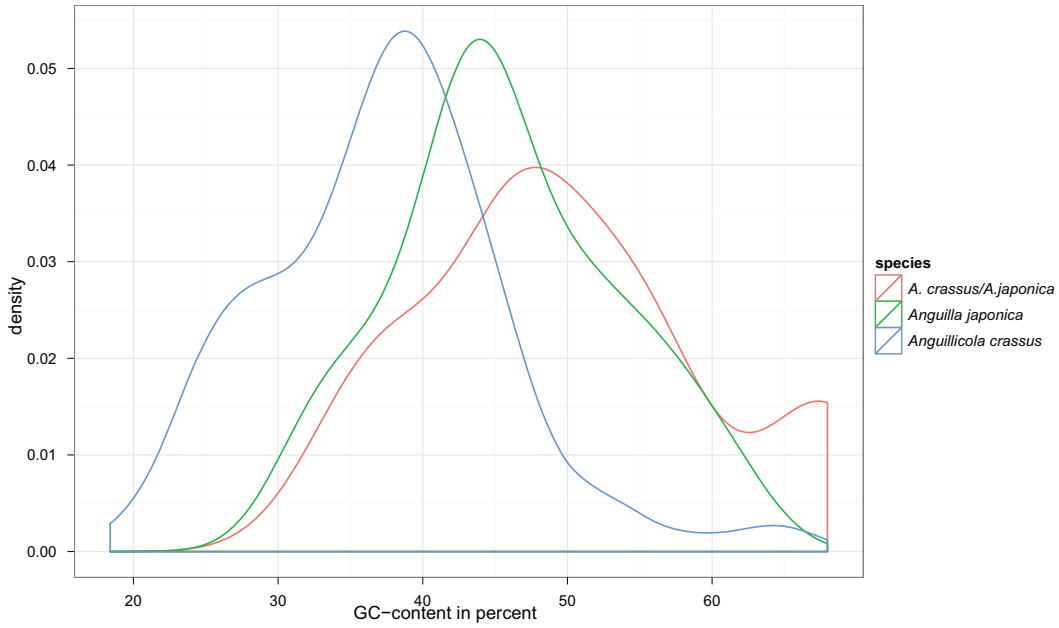

Figure 3.2: GC-content of sequences from *An. japonica* and *A. crassus* - The Japanese eel has a slightly higher GC-content than the parasite: This sequence characteristic is useful for separation of sequences from the host-parasite interface, note the higher GC-content of the sequences from *A. crassus*,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

distribution (see Figure 3.2). *A. crassus* had a lower mean GC-content (37.32 ± 8.36 mean $\pm$ sd) than *An. 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 threshold 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 highly similar to "hemoglobin anodic subunit" from the European eel. Others were annotated with best hits to highly expressed housekeeping 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 threshold of 55 bits.

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

However, 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 freshwater and sequences are probably derived from a commensal, symbiont or pathogen of eels or swimbladder-nematodes. These off-target data were left in the submission file.

For 66 (58.4%) of the remaining 113 ESTs annotations were obtained from orthologous sequences. All of 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: Annotation 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 chapter 5: the sequence-assembly. The quality of this sequence assembly constitutes a fundamental foundation of the later chapters.

The pre-processed *A. crassus* data-set consisting of 100,491,819 bases in 353,055 reads (58,617 generated using "FLX-chemistry", 294,438 using "Titanium-chemistry") was assembled following an approach proposed by (127): two assemblies were generated, one using `Newbler v2.6` (137), the other using `Mira v3.2.1` (175). The resulting assemblies (referred to as first-order assemblies) were merged with `Cap3` (176) 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 for the evaluation of downstream results. As a perfect assembly with each contig representing a single full transcript is illusive and every contig constitutes a hypothesis, it becomes important to validate and question analyses based on as much information as possible. Thus a comprehensive set of assembly derived statistics is presented.

4.2 The Newbler first-order assembly

During transcriptome-assembly Newbler can split individual reads spanning the breakpoints of alternate isoforms, to assemble, for example, 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 include 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 from the 26,211 reported to 109,052. I later also address the usefulness of Newbler's report vs. the expanded singleton-category, but in the meantime I 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 number of contigs Newbler split one read into (in some cases more 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 maximum number of read-splits in a given contig and its 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 chapter 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 of slightly smaller size than the Newbler assembly. The

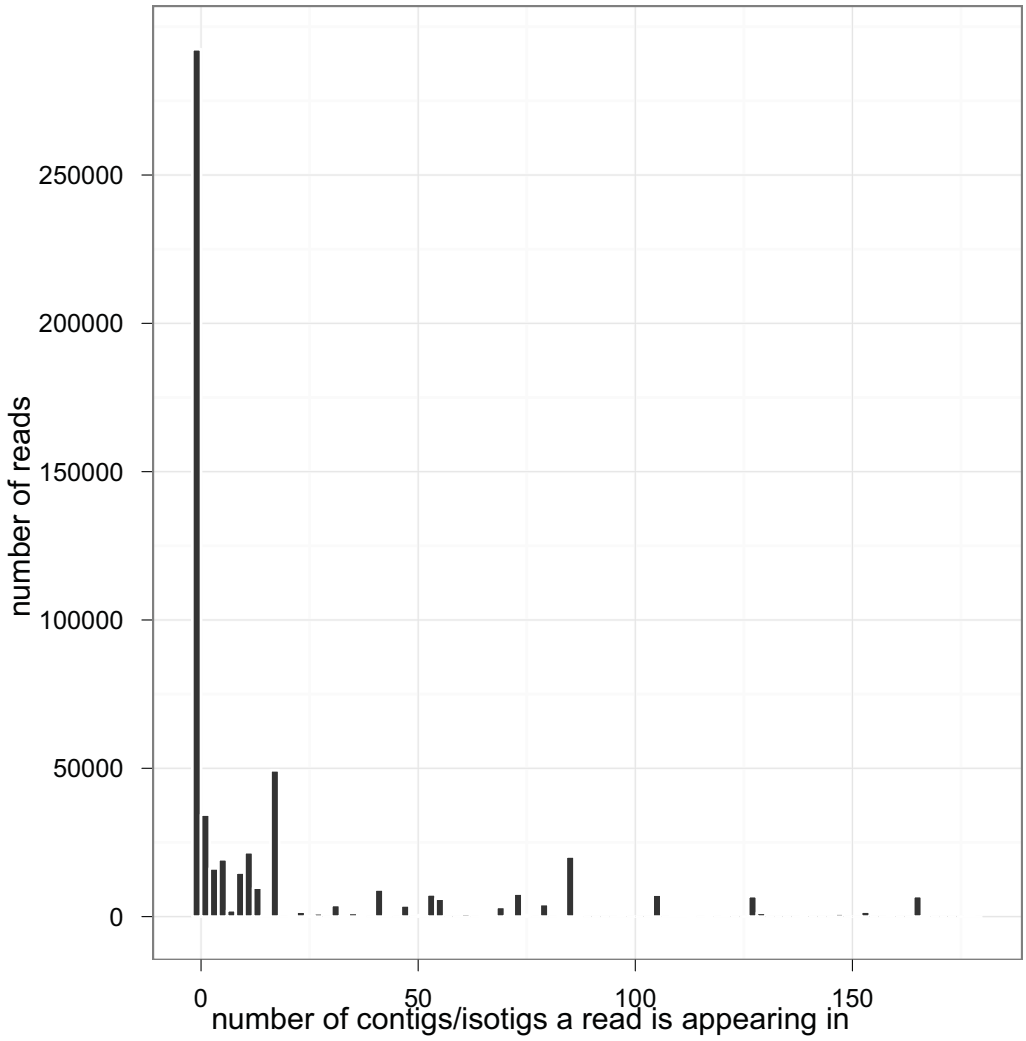

Figure 4.1: Number of contigs/isotigs split - A histogram of the number of contigs or isotigs `Newbler` split a single read into.

	Newbler	Mira	Second-order(MN)
Max length	6,300	6,352	6,377
Number of contigs	15,934	22,596	14,064
Number of Bases	8,085,922	12,010,349	8,139,143
N50	579	579	662
Number of contigs in N50	4,301	6,749	3,899
non ATGC bases	375	29,962	5,245
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 assembly (for which only the most reliable category of contigs (MN) is shown; see 4.4).

second-order assembly had on average longer contigs than both first-order assemblies and a higher weighted median contig size (N50).

4.4 Data-categories in the second-order assembly

Three main categories of assembled sequence data can be distinguished in the second-order assembly, 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 intersection of the `Newbler`-singletons (as defined in 4.2) and the `Mira`-singletons. 47,669 reads fell into this category. A second category of sequence contains the first-order contigs which 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 contributions from both first-order assemblies (MN in table 4.2). For this last, most reliable (MN) category, reads contained in the assembly can be categorised 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` split during

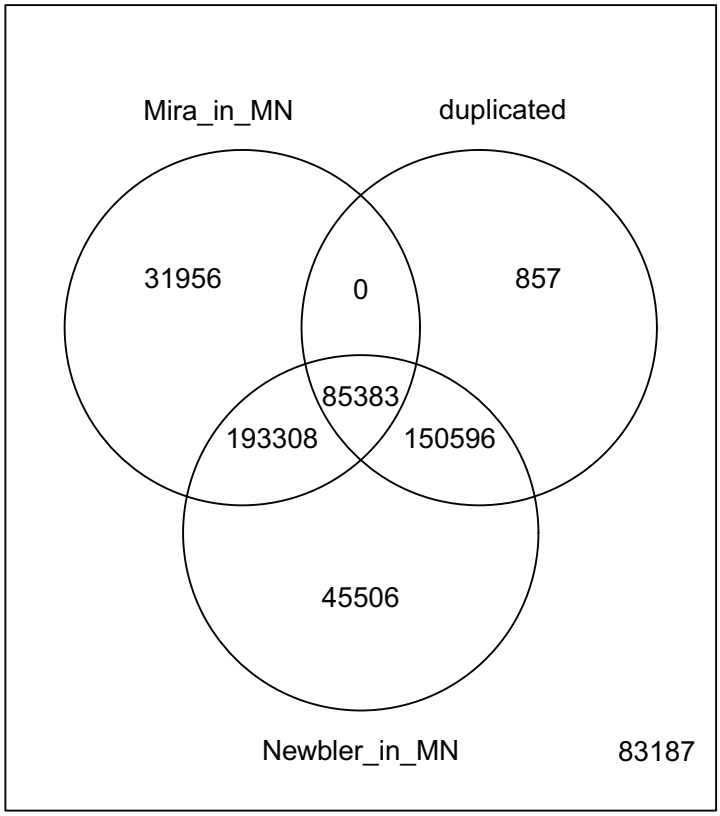

Figure 4.2: Origin of reads - Reads in the most reliable (MN) assembly-category are categorised by the way they entered the assembly: Although they are in a highly credible contig, reads can still have entered from only one first order assembly (Mira_in_MN or Newbler_in_MN). The intersection gives the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

first-order assembly. Interestingly, most reads `Newbler` split ended in the high-quality category of the second order assembly only.

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 4.3a 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 4.3b). 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 the splitting of reads.

4.6 Evaluation of the assemblies

To further compare assemblies (`Mira, Newbler` first-order assemblies including or excluding their singletons) and the second-order assembly (including different contigs-categories and singletons) I 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

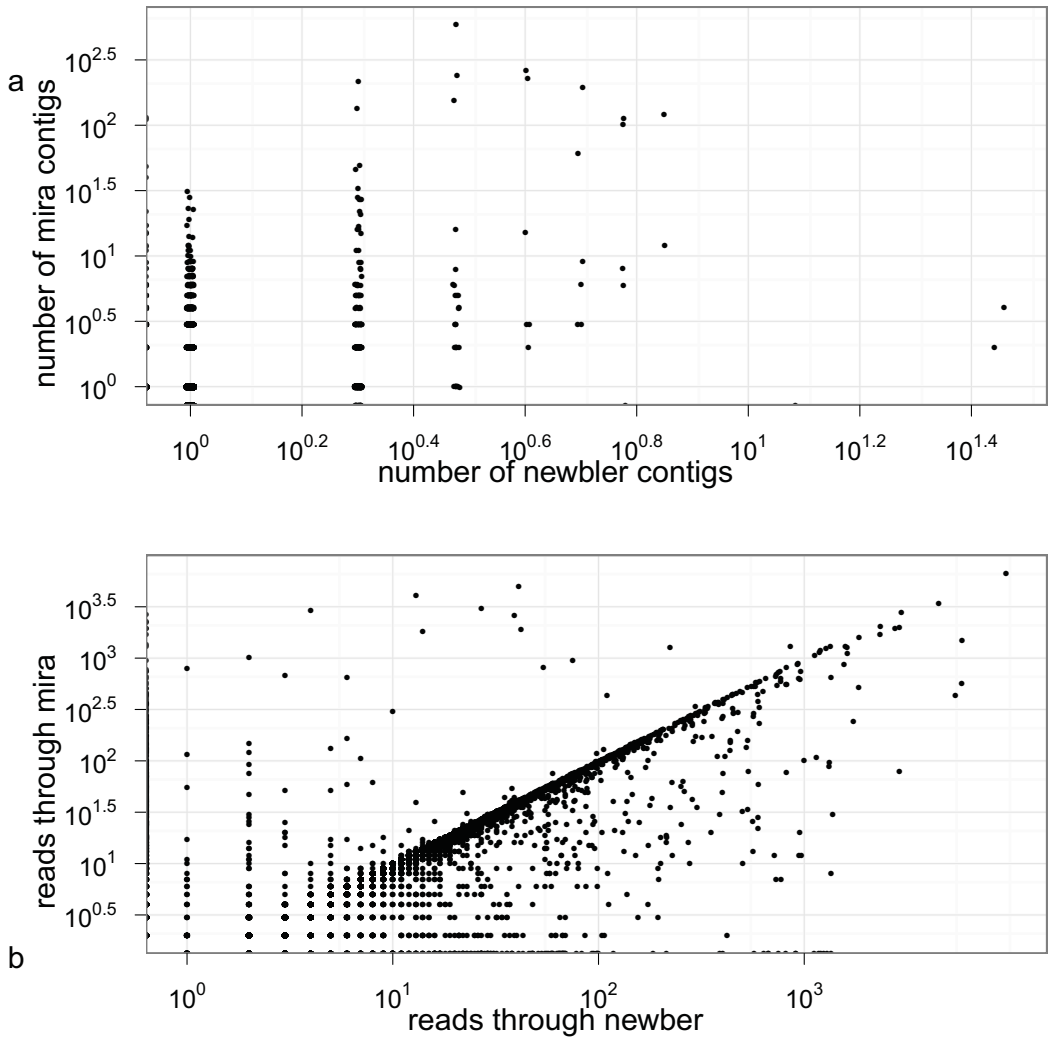

Figure 4.3: Contribution to second-order assembly - Number of first-order contigs from both first-order assemblies for each second order contig (a) number of reads through `Newbler` and `Mira` for each second-order contig (b).

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 plotted against the 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 in 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 `Newbler` assemblies were of increased size. A comparison of the `Newbler's` reported singletons with all singletons added to the `Newbler`-assembly shows that the reported singletons increased reference-coverage to the same amount as 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 similarly to the `Newbler` assembly for the number of bases covered, but was larger in size. Adding the less reliable set of one-assembler supported second-order-contigs the assembly covered the most bases in both references. When the singleton of the second-order assembly (as defined in 4.2) were not included but only the intersection of `Newbler's` "reported singletons" and `Mira's` singletons, 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 well. Excluding singletons it covered similar numbers of database-proteins to the `Newbler`-assembly and and was outperformed by the `Mira`-assembly, although the latter was again shown to be least parsimonious. The same general picture emerged from this analysis when singletons were considered additionally. `Newbler` and second-order assemblies covered similar amounts of reference-data.

When database-proteins covered for at least 80% of their length are considered, the second-order assembly showed its 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. I give most weight in my analysis to these results

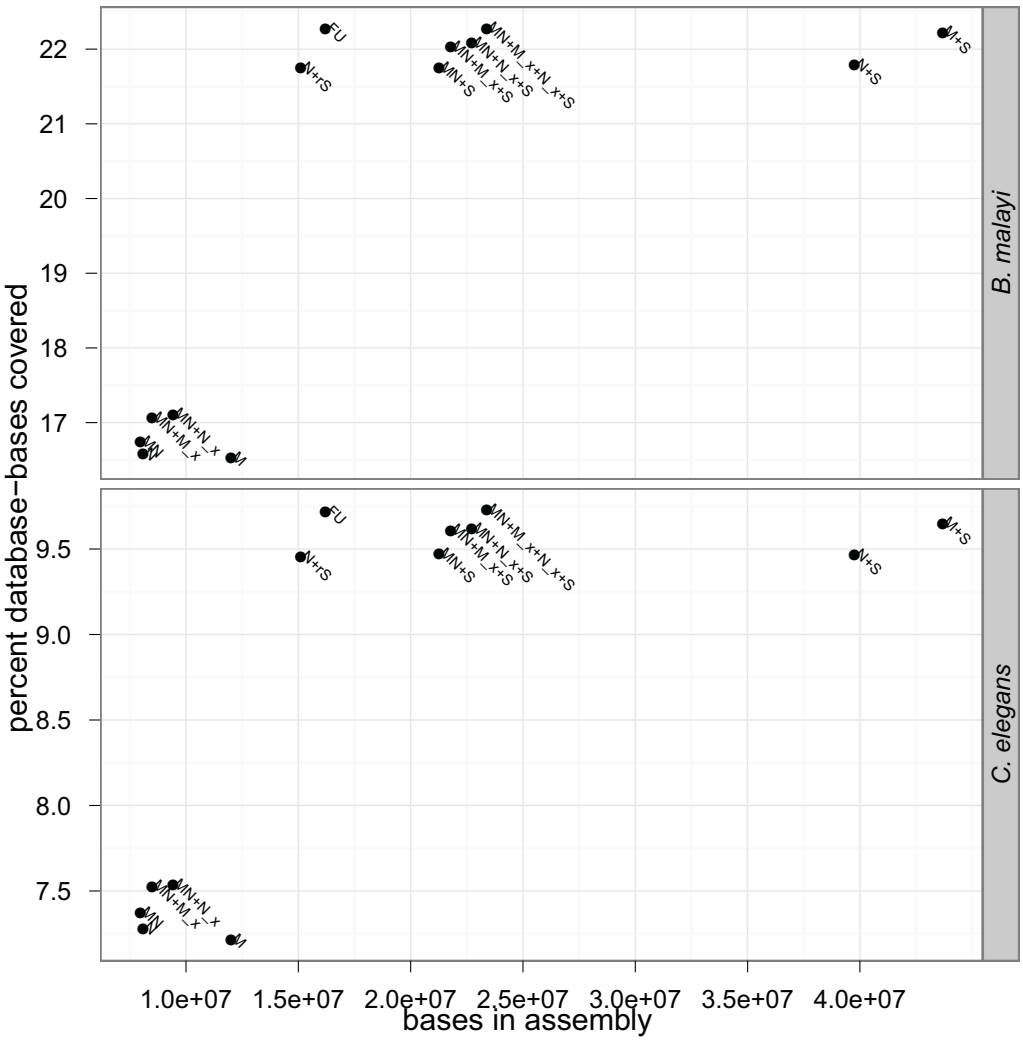

Figure 4.4: Base-content and reference-transcriptome coverage in percent of bases - for different assemblies and assembly-combinations; M = `Mira`; N = `Newbler`; $M + S$ = `Mira` + singletons; $N + S$ = `Newbler` plus singletons; $N + rS$ = `Newbler` plus singletons reported in readstatus.txt; MN = second-order contigs supported by both first-order; $MN + N\_x$ = second-order MN plus contigs only supported by `Newbler` (N_x = N_n and N_1); $MN + M\_x$ = 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

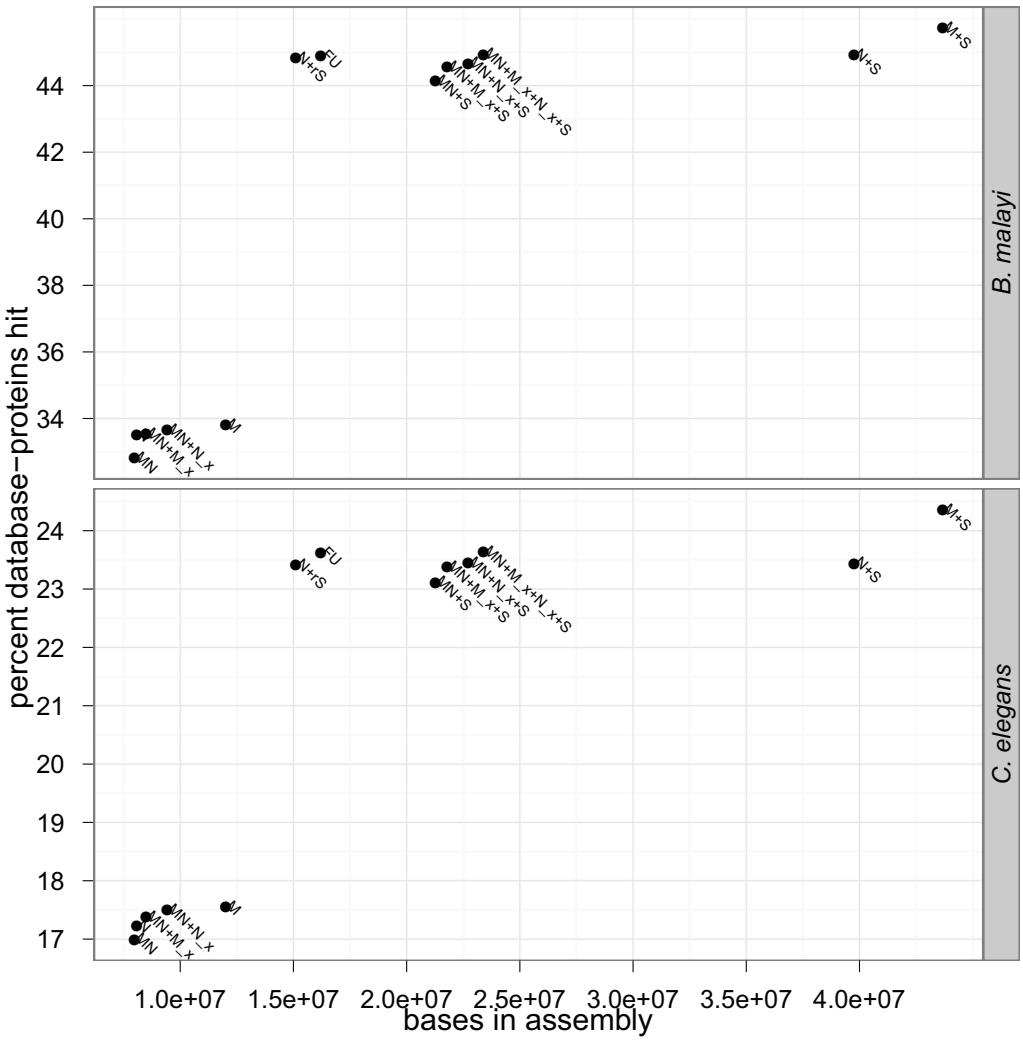

Figure 4.5: Base-content and reference-transcriptome coverage in percent of proteins hit - in percent of proteins hit for different assemblies and assembly-combinations (for category-abbreviations see figure 4.4)

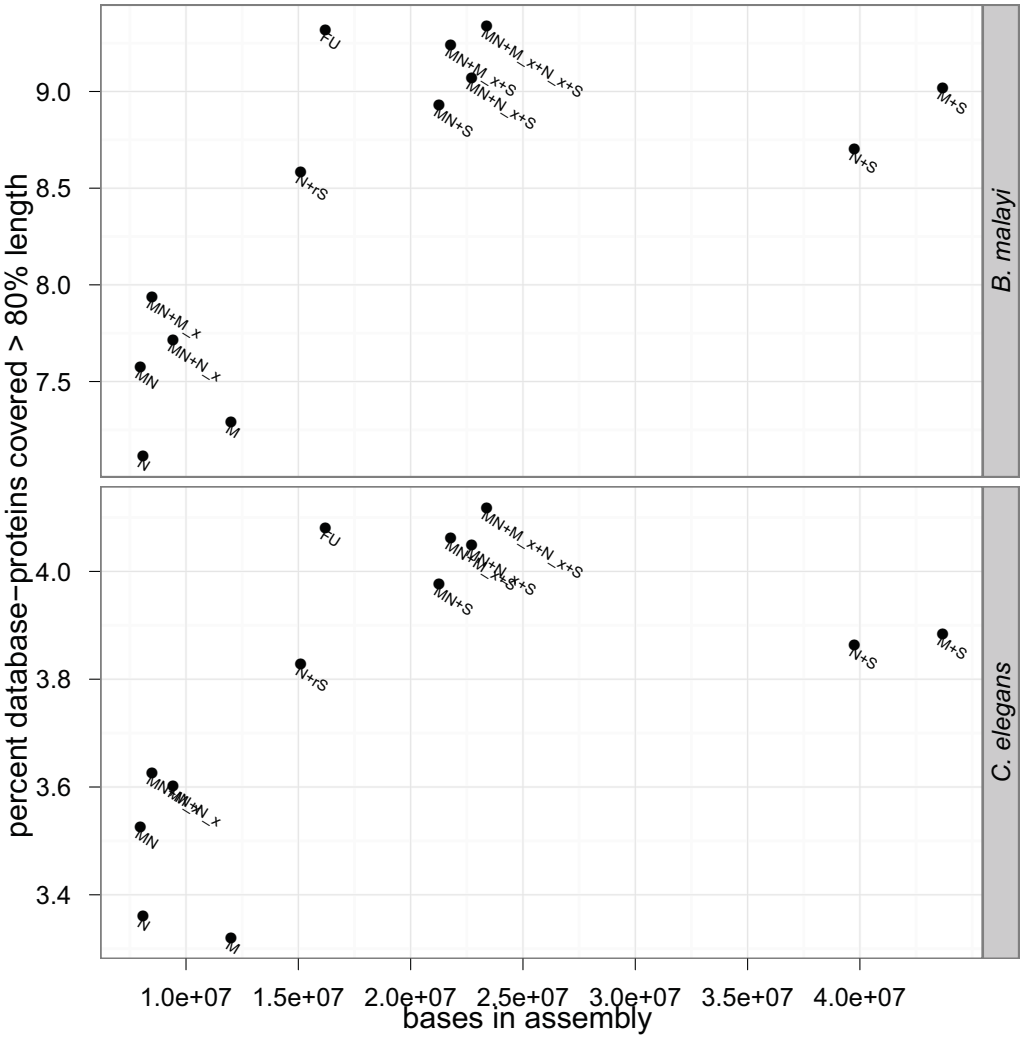

Figure 4.6: Base-content and reference-transcriptome coverage in percent of proteins covered to at least 80% - of their length for different assemblies and assembly-combinations (for category-abbreviations see figure 4.4)

as in average longer correct contigs will allow finding the highest number of putative full-length genes.

Given this evaluation I defined a "minimal adequate" assembly as the subset of contigs of the second-order assembly supported by both assemblers (labeled MN above). Given the performance of the singletons `Newbler` reported. I 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 Measurements on second-order assembly

Based on the tracking of reads through the complicated assembly process, I calculated the following statistics for each contig in the second-order assembly.

- 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 same-read-pairs 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 were assigned the same id (similar to the graph clustering reported in (157)).
- 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 readily available from the output of this combined assembly approach (although I followed individual reads through the process) I inferred coverage by mapping the reads used for assembly against the fullest assembly using `ssaha2` (153) :

- mean per base coverage

- mean unique per base coverage

The ratio of mean per base coverage and unique per base coverage (the standard for assessing coverage) can be used as to asses the redundancy of a contig.

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. 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 for assembly-measurements - Measurements on contigs, 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 different `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 two 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 admitted 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-pairs from both first-order assemblies and 8 split-reads from `Newbler`-fist 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. Cluster size correlates as expected with the ratio of unique/non-unique coverage, as contigs in clusters contain redundant sequences also found in other contigs.

These measurements were used in all later analyses to evaluate likelihood of mis-assembly artefacts as an influence on a given set of biological relevant contigs. All gene-sets mentioned later (in chapter 5) were thus, as a matter of routine, controlled for unusual patterns in the contig meta-data.

4.8 Finalising the fullest assembly set

As additional measure in order to minimise the amount of sequence with artificially inferred isoform-breakpoints, I 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 I removed all well trusted MN-category contigs having no coverage at all and the contigs (and singletons) from other categories having no unique coverage.

Thereby I reduced my dataset to 40187 tentative unique genes (TUGs), redefining the "fullest assembly" dataset. Based on the above evaluation I 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).

	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(-36)**	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. Only the contigs in bold listed here were not screened from the assembly (7 MN-contigs).

4. EVALUATION OF AN ASSEMBLY STRATEGY FOR PYROSEQUENCING READS

5

Pyrosequencing of the *A. crassus* transcriptome

5.1 Overview

In this chapter the transcriptome assembly of *A. crassus* is analysed in its biological context. It constitutes 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, I assembled 353,055 into 11,371 contigs spanning 6,575,121 bases and additionally obtained 21,147 singleton and lower quality contigs spanning 6,157,974 bases. I obtained annotations for ca. 60% of the contigs and 40% of the tentatively unique genes (TUGs) confirming the high quality of especially the contigs. I identified 5,112 high quality single nucleotide polymorphisms (SNPs) and suggest 199 of them as most suitable markers for population-genetic studies. Correlation between different analyses provided further insights and confirmed biologically relevant expectations: I found an overabundance of predicted signal peptide cleavage sites in sequence conserved in Nematoda and novel in *A. crassus*, correlations between coding polymorphism and differential expression, between coding polymorphism and peptide cleavage sites and between conservation and presence of orthologs with lethal RNAi-phenotypes in *C. elegans*. GO-term analysis identified an enrichment of peptidases and subunits of the respiratory chain for transcripts under positive selection. Enzymes for energy metabolism were also found enriched in genes differentially expressed between European and Asian *A. crassus*.

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 I 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 756,363 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 585,949 high-quality reads (spanning 169,863,104 bases). In the eel dataset from 159,370 raw reads 135,072 were assembled after basic quality screening.

I then screened the *A. crassus* reads for contamination by host (30,071 matched previously sequenced eel genes or my own *An. japonica* 454 transcriptome, which had been assembled into 10,639 mRNA contigs. (181,783 reads matched large or small subunit nuclear or mitochondrial ribosomal RNA sequences of *A. crassus*) . In addition to fish mRNAs, I identified (and removed) 5,286 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 (see also chapter 4)

I assembled the remaining 353,055 reads (spanning 100,491,819 bases) using the combined assembler strategy (127) and Roche 454 GSassembler (`Newbler` version 2.6) and `Mira` (version 3.21) (175). From this I derived 13,851 contigs that were supported by both assembly algorithms, 3,745 contigs only supported by one of the assembly algorithms and 22,591 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

library	E1	E2	L2	M	T1	T2
life.st	adult f	adult f	L2 larvae	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
over.32	3528	34051	10444	21219	22435	1602

Table 5.1: Pyrosequencing library statistics - 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` (177). 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). over.32 = number of reads mapping to contigs with overall coverage of more than 32 reads (considered in gene-expression analysis).

transcriptomes). In the highCA parsimony and low redundancy is prioritised, while in the complete assembly (highCA plus lowCA) completeness is prioritised. The 40,187 sequences (contig consensuses and singletons) in the complete assembly are referred to below as tentatively unique genes (TUGs).

I screened the complete assembly for residual host contamination, and identified 3,441 TUGs that had higher, significant similarity to eel (and chordate) sequences (my 454 ESTs and EMBLBank Chordata proteins) than to nematode sequences (125).

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

1,153 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 I 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.

I excluded all TUGs with best hits outside Metazoa and my assembly thus has 32,518 TUGs, spanning 12,733,095 bases (of which 11,371 are highCA-derived, and span 6,575,121 bases) that are likely to derive from *A. crassus*.

5.5 Protein prediction

For 32,411 TUGs a protein was predicted using `prot4EST` (178) (see table 5.2). The full open reading frame was obtained in 353 TUGs, while while for 2,683 the 5' end and for 8,283 the 3' end was complete. In 13,379 TUGs the corrected sequence with the imputed ORF was slightly changed compared to the raw sequence.

	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	6157974	6575121	12733095
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 8.5.5), 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.

5.6 Annotation

I obtained basic annotations with orthologous sequences from *C. elegans* for 9,554 TUGs, from *B. malayi* for 9,662 TUGs, from nempep (123, 125) for 11,617 TUGs and with uniprot proteins for 11,113 TUGs.

I used `annot8r` (179) to assign gene ontology (GO) terms for 6,509 TUGs, Enzyme Commission (EC) numbers for 2,458 TUGs and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway annotations for 3,844 TUGs (see table 5.2). Additionally 5,125 highCA derived contigs were annotated with GO terms through `InterProScan` (180). Nearly one third (6,987) of the *A. crassus* TUGs were annotated with at least one identifier, and 1,829 had GO, EC and KEGG annotations (see figure 5.1).

I compared my *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.

I inferred presence of signal peptide cleavage sites in the predicted protein sequence using `SignalP` (181). I 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).

I 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 highCA derived contigs were categorised 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*.

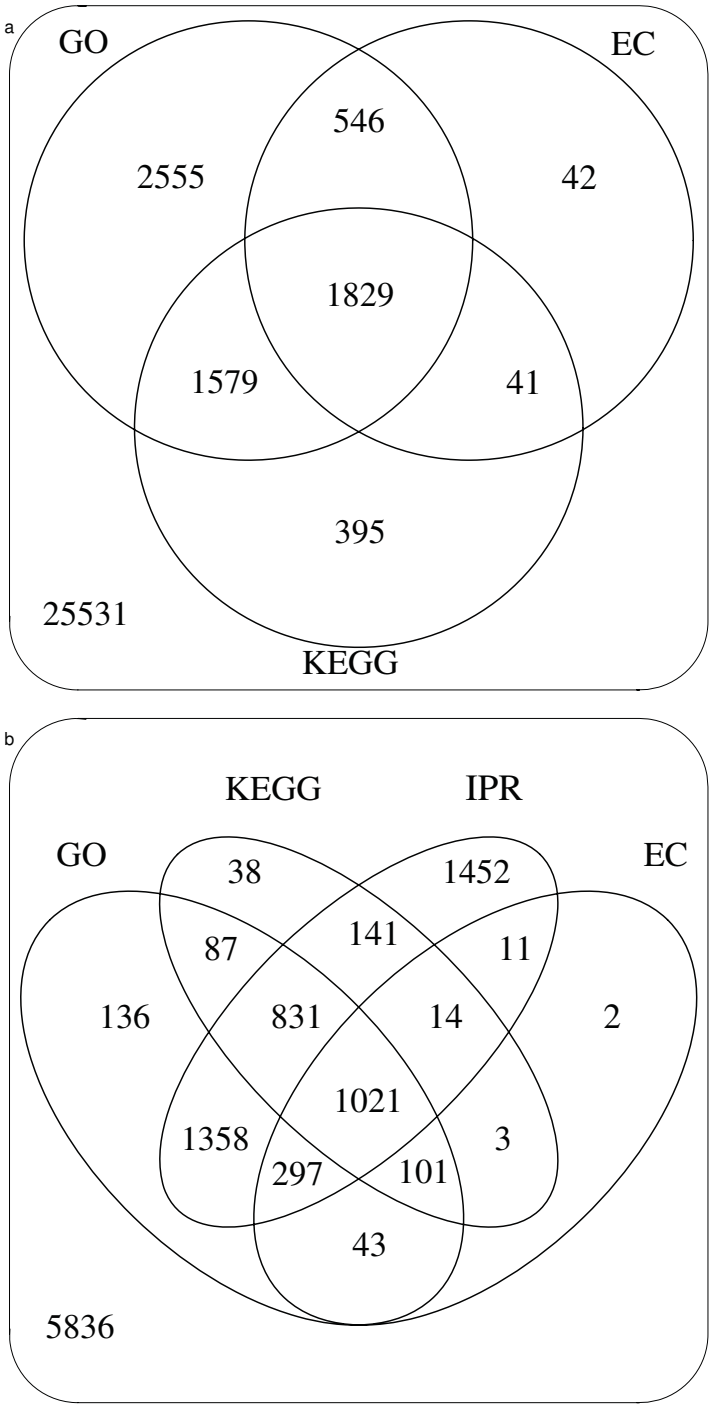

Figure 5.1: Annotation using different identifiers - Number of annotations obtained for Gene Ontology (GO), Enzyme Commission (EC) and Kyoto Encyclopedia of Genes and Genomes (KEGG) terms through `Annot8r` (179) for all TUGs (a) and for highCA derived contigs (b). The latter includes additional domain-based annotations obtained with `InterProScan` (180).

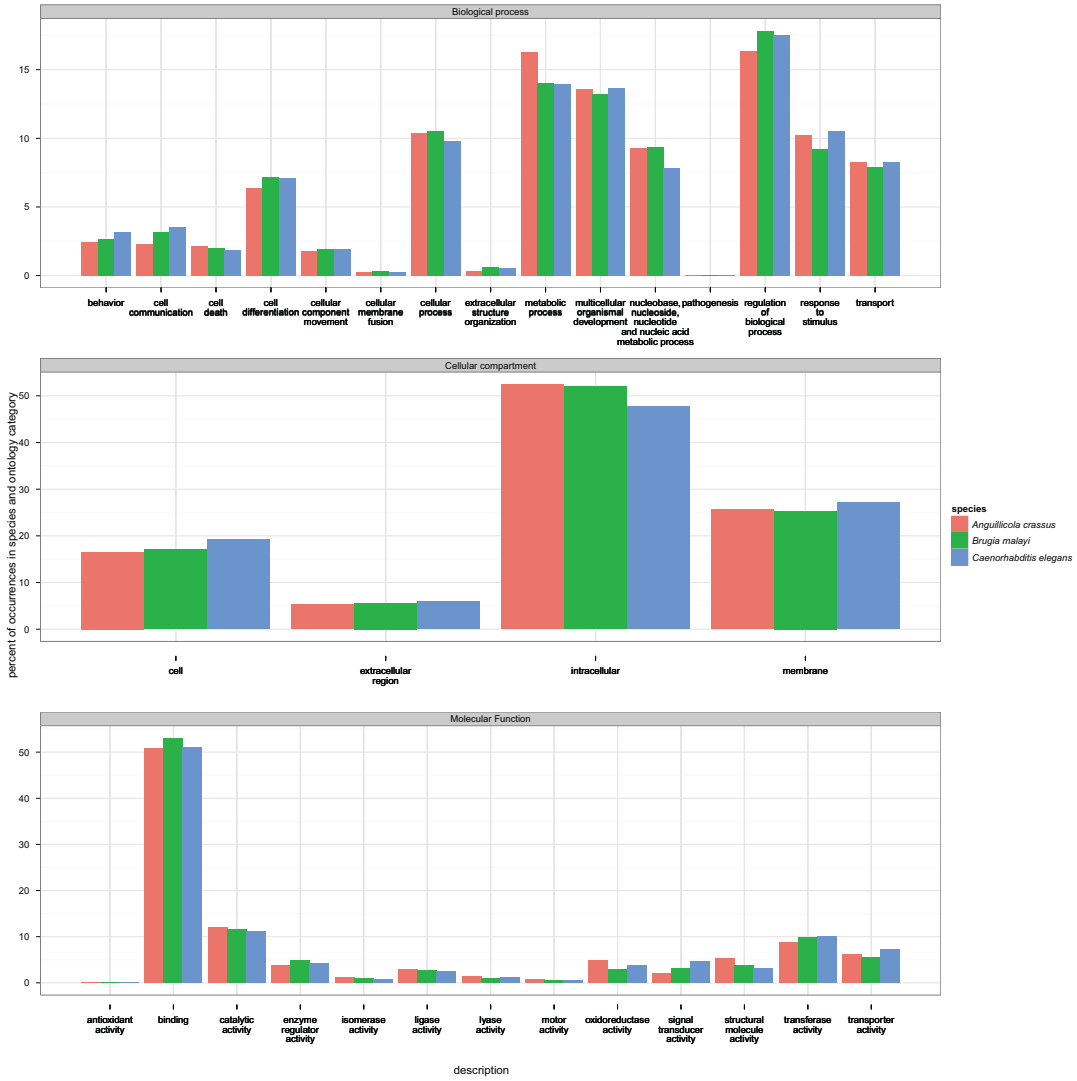

Figure 5.2: Cross-taxa comparison of annotation - For Gene Ontology (GO) categories molecular function, cellular compartment and biological process the proportion (for each ontology-category and species) of terms in high level GO-slim categories is given as obtained through `Annot8r` (179).

	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 categorised as novel in *A. crassus* (novel.in.Ac). Both novelty and conservation can be derived from this (numbers for conservation would be the cumulative sum of lower-level novelty).

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

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

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

I called single nucleotide polymorphisms (SNPs) on the 1,099,419 bases of the TUGs that had coverage of more then 8-fold available using `VARScan` (158). I excluded SNPs predicted to have more than 2 alleles or that mapped to an undetermined (N) base in the reference, and retained 10,458 SNPs. The ratio of transitions (ti; 6,890) to transversion (tv; 3568) in this set was 1.93. Using the prot4EST predictions and the corrected sequences, 7,153 of the SNPs were predicted to be inside an ORF, with 2,310 at codon first positions, 1,819 at second positions and 3,024 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. I filtered these SNPs to exclude those that might be associated with analytic bias. As Roche 454 sequences have well-known systematic errors associated with homopolymeric nucleotide sequences (135), I analysed the effect of exclusion of SNPs in, or close to, homopolymer regions. I 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 I 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. I also observed a relationship between TUG dn/ds and TUG coverage,

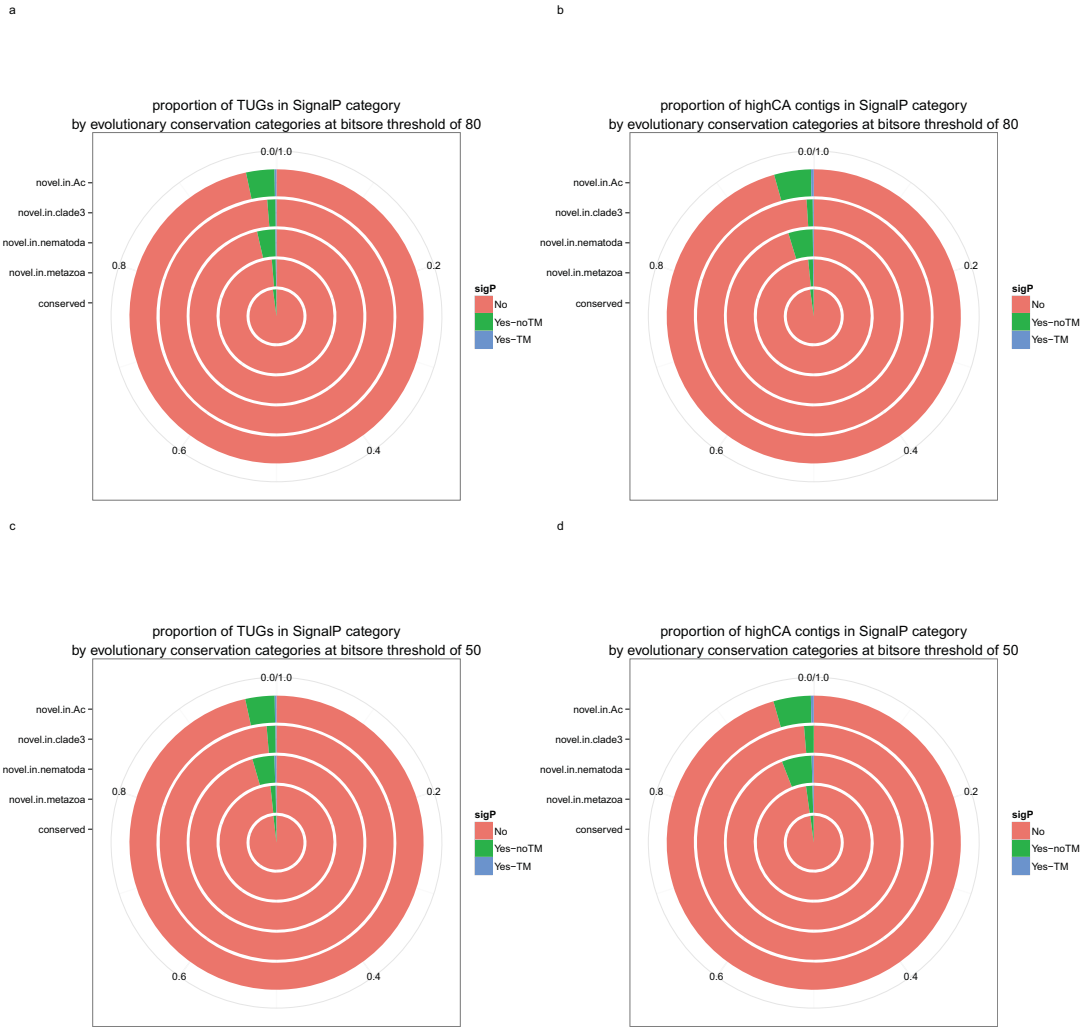

Figure 5.3: Enrichment of signal-positives for categories of evolutionary conservation - Proportions of `SignalP`-predictions for each category of evolutionary conservation. Generally - across bit-score thresholds - TUGS novel in nematodes and in *A. crassus* have the highest proportion of signal-positives. sigP = signalIP-prediction; Yes-noTM, cleavage site predicted; Yes-TM, transmembrane-anchor predicted.

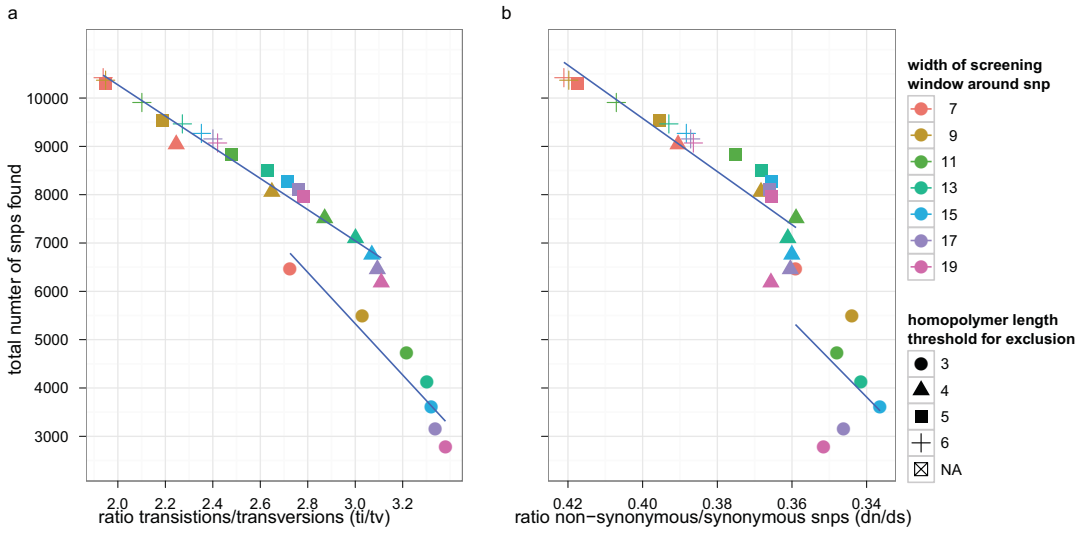

Figure 5.4: Homopolymer screening for SNP-calling - When SNPs in or adjacent to homopolymeric regions are removed changes in ti/tv (a) and dn/ds (b) are observed: As the overall number of SNPs is reduced both ratios change to more plausible values. Note the reversed axis for dn/ds to plot these lower 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).

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). My filtered SNP dataset includes 5,112 SNPs. I retained 4.65 SNPs per kb of contig sequence, with 8.37 synonymous SNPs per 1,000 synonymous bases and 2.4 non-synonymous SNPs per 1,000 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.

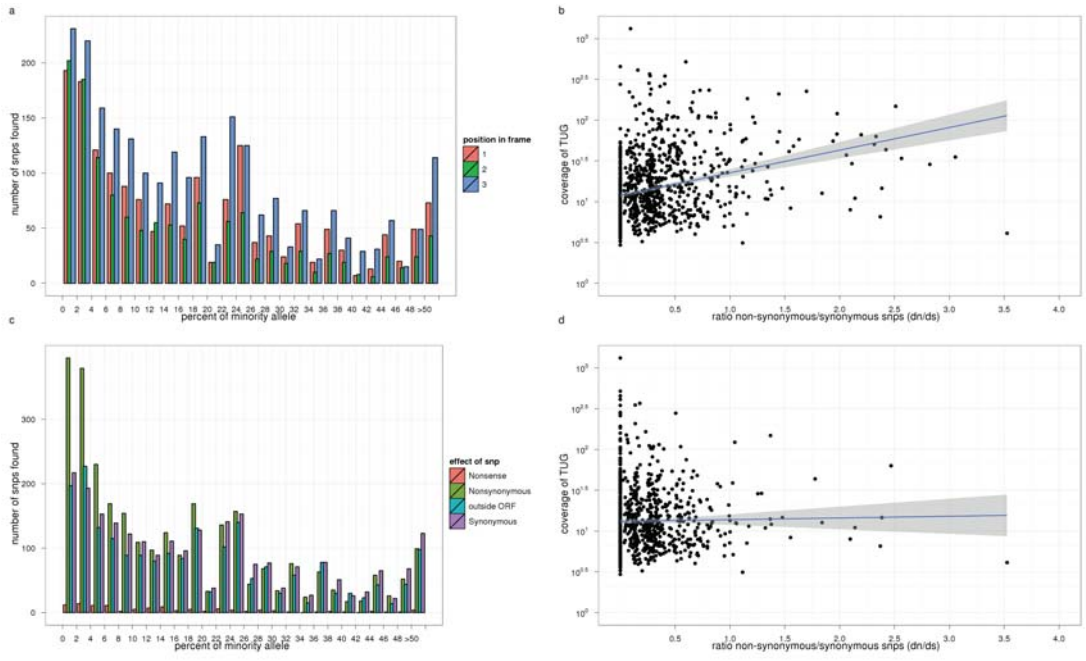

Figure 5.5: SNP-calling and SNP categories - Overabundance of SNPs at (a) codon-position two and of (c) non-synonymous SNPs for low percentages of the minority allele. (b) Significant positive correlation of coverage and dn/ds before removing these SNPs at a threshold of 7% ($p < 0.001$, $R^2 = 0.015$) and (d) no significant correlation afterwards (R^2 <0.001, p =0.211).

5.9 Polymorphisms associated with biological processes

I consolidated my 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 I defined TUGs with a dn/ds higher than 0.5 as positively selected. I identified over-represented GO ontology terms associated with these putatively positively selected genes (see table 5.4 and additional figures **??**, **??** and **??**).

GO.ID	Term	Annotated	Significant	Expected	p-value
Molecular function					
GO:0008233	peptidase activity	43	12	5.26	0.0028
GO:0015179	L-amino acid transmembrane transporter activity	2	2	0.24	0.0147
GO:0016787	hydrolase activity	110	20	13.45	0.0262
GO:0043021	ribonucleoprotein binding	6	3	0.73	0.0266
GO:0005102	receptor binding	26	7	3.18	0.0288
GO:0046982	protein heterodimerization activity	16	5	1.96	0.0348
GO:0004129	cytochrome-c oxidase activity	3	2	0.37	0.0407
GO:0004540	ribonuclease activity	3	2	0.37	0.0407
GO:0005275	amine transmembrane transporter activity	3	2	0.37	0.0407
GO:0005342	organic acid transmembrane transporter activity	3	2	0.37	0.0407
GO:0005275	amine transmembrane transporter activity	3	2	0.37	0.0407
GO:0005342	organic acid transmembrane transporter activity	3	2	0.37	0.0407
GO:0015002	heme-copper terminal oxidase activity	3	2	0.37	0.0407
GO:0015171	amino acid transmembrane transporter activity	3	2	0.37	0.0407
GO:0016675	oxidoreductase activity, acting on a heme group of donors	3	2	0.37	0.0407

Continued on next page

Table 5.4 – continued from previous page

GO.ID	Term	Annotated	Significant	Expected	p-value
GO:0016676	oxidoreductase activity, acting on a heme group of donors, oxygen as acceptor	3	2	0.37	0.0407
GO:0046943	carboxylic acid transmembrane transporter activity	3	2	0.37	0.0407
GO:0047035	testosterone dehydrogenase (NAD+) activity	3	2	0.37	0.0407
GO:0015077	monovalent inorganic cation transmembrane transporter activity	12	4	1.47	0.0471
Biological process					
GO:0009081	branched chain family amino acid metabolic process	3	3	0.36	0.0017
GO:0042594	response to starvation	15	6	1.82	0.0052
GO:0006914	autophagy	12	5	1.45	0.0090
GO:0006520	cellular amino acid metabolic process	44	11	5.33	0.0102
GO:0007281	germ cell development	17	6	2.06	0.0105
GO:0090068	positive regulation of cell cycle process	17	6	2.06	0.0105
GO:0009308	amine metabolic process	57	13	6.90	0.0118
GO:0051325	interphase	23	7	2.79	0.0139
GO:0051329	interphase of mitotic cell cycle	23	7	2.79	0.0139
GO:0010564	regulation of cell cycle process	34	9	4.12	0.0140
GO:0051726	regulation of cell cycle	52	12	6.30	0.0143
GO:0005997	xylulose metabolic process	2	2	0.24	0.0145
GO:0006739	NADP metabolic process	2	2	0.24	0.0145
GO:0009744	response to sucrose stimulus	2	2	0.24	0.0145

Continued on next page

Table 5.4 – continued from previous page

GO.ID	Term	Annotated	Significant	Expected	p-value
GO:0010172	embryonic body morphogenesis	2	2	0.24	0.0145
GO:0015807	L-amino acid transport	2	2	0.24	0.0145
GO:0019321	pentose metabolic process	2	2	0.24	0.0145
GO:0034285	response to disaccharide stimulus	2	2	0.24	0.0145
GO:0050885	neuromuscular process controlling balance	2	2	0.24	0.0145
GO:0006915	apoptosis	78	16	9.45	0.0147
GO:0009056	catabolic process	149	26	18.04	0.0148
GO:0031571	mitotic cell cycle G1/S transition DNA damage checkpoint	14	5	1.70	0.0187
GO:0044106	cellular amine metabolic process	55	12	6.66	0.0224
GO:0009063	cellular amino acid catabolic process	10	4	1.21	0.0234
GO:0000082	G1/S transition of mitotic cell cycle	15	5	1.82	0.0255
GO:0030330	DNA damage response, signal transduction by p53 class mediator	15	5	1.82	0.0255
GO:0033238	regulation of cellular amine metabolic process	15	5	1.82	0.0255
GO:0042770	signal transduction in response to DNA damage	15	5	1.82	0.0255
GO:0072331	signal transduction by p53 class mediator	15	5	1.82	0.0255
GO:0006401	RNA catabolic process	6	3	0.73	0.0259
GO:0010638	positive regulation of organelle organization	6	3	0.73	0.0259
GO:0042981	regulation of apoptosis	64	13	7.75	0.0312

Continued on next page

Table 5.4 – continued from previous page

GO.ID	Term	Annotated	Significant	Expected	p-value
GO:0043067	regulation of programmed cell death	64	13	7.75	0.0312
GO:0009310	amine catabolic process	11	4	1.33	0.0335
GO:0051084	'de novo' posttranslational protein folding	11	4	1.33	0.0335
GO:0008219	cell death	93	17	11.26	0.0370
GO:0016265	death	93	17	11.26	0.0370
GO:0012501	programmed cell death	86	16	10.41	0.0371
GO:0010941	regulation of cell death	66	13	7.99	0.0396
GO:0000393	spliceosomal conformational changes to generate catalytic conformation	3	2	0.36	0.0400
GO:0006123	mitochondrial electron transport, cytochrome c to oxygen	3	2	0.36	0.0400
GO:0006865	amino acid transport	3	2	0.36	0.0400
GO:0009313	oligosaccharide catabolic process	3	2	0.36	0.0400
GO:0031023	microtubule organizing center organization	3	2	0.36	0.0400
GO:0045292	nuclear mRNA cis splicing, via spliceosome	3	2	0.36	0.0400
GO:0045840	positive regulation of mitosis	3	2	0.36	0.0400
GO:0051262	protein tetramerization	3	2	0.36	0.0400
GO:0051289	protein homotetramerization	3	2	0.36	0.0400
GO:0051297	centrosome organization	3	2	0.36	0.0400
GO:0051785	positive regulation of nuclear division	3	2	0.36	0.0400
GO:2000242	negative regulation of reproductive process	3	2	0.36	0.0400
GO:0007286	spermatid development	7	3	0.85	0.0415

Continued on next page

Table 5.4 – continued from previous page

GO.ID	Term	Annotated	Significant	Expected	p-value
GO:0009267	cellular response to starvation	7	3	0.85	0.0415
GO:0048515	spermatid differentiation	7	3	0.85	0.0415
GO:0016071	mRNA metabolic process	47	10	5.69	0.0437
GO:0006458	'de novo' protein folding	12	4	1.45	0.0457
GO:0022607	cellular component assembly	103	18	12.47	0.0484
Cellular compartment					
GO:0030532	small nuclear ribonucleoprotein complex	7	4	0.84	0.005
GO:0005682	U5 snRNP	2	2	0.24	0.014
GO:0015030	Cajal body	2	2	0.24	0.014
GO:0046540	U4/U6 x U5 tri-snRNP complex	2	2	0.24	0.014
GO:0016607	nuclear speck	6	3	0.72	0.025
GO:0005739	mitochondrion	136	23	16.35	0.031
GO:0005604	basement membrane	3	2	0.36	0.039
GO:0060198	clathrin sculpted vesicle	3	2	0.36	0.039

Table 5.4: Over-representation of GO-terms in positively selected - GO-terms over-represented in contigs putatively under positive selection. Horizontal lines separate categories of the GO-ontology. First category is molecular function, second biological process, last cellular compartment. P values (pval) for over- representation (Fishters exact test) 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) see also additional figures **??**, **??** and **??**.

Within the molecular function category, "peptidase activity" was the most significantly overrepresented term and had twelve TUGs supporting the overrepresentation. The highlighted twelve peptidases annotated with eleven unique orthologs in *C. elegans* and *B. malayi*. 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 and to amino and fatty acid catabolic processes.

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).

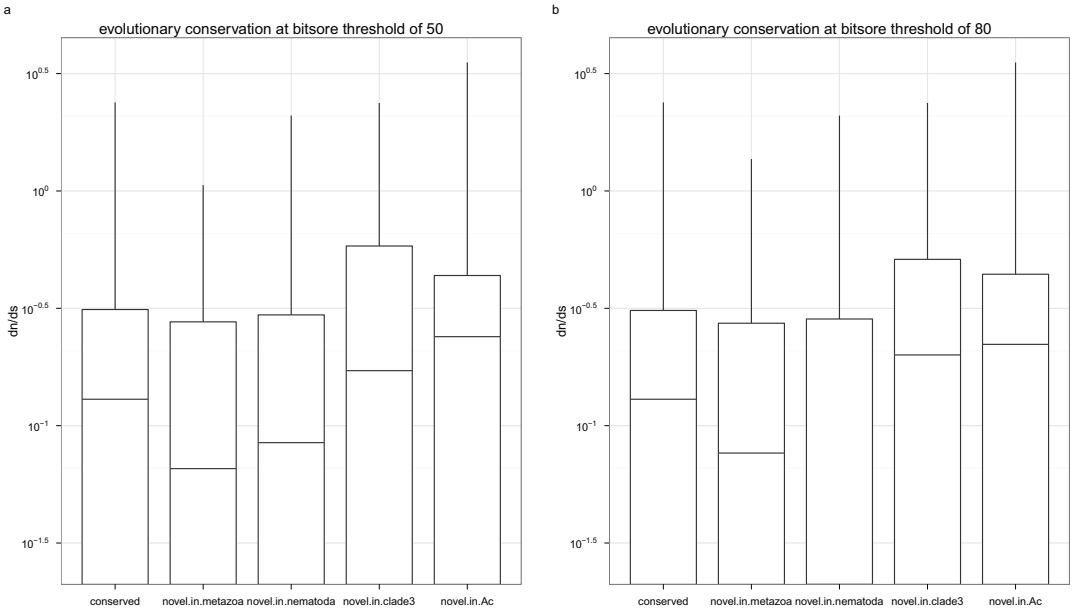

Figure 5.6: Positive selection and evolutionary conservation - Box-plots for dn/ds in TUGs according to different categories of evolutionary conservation. Significant comparisons are novel.in.metazoa - novel.in.Ac (0.005 and 0.015), novel.in.nematoda - novel.in.Ac (0.005 and 0.002), novel.in.nematoda - 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

I used `Samtools`(182) and `Vcftools`(159) 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; (183)), homozygosity by loci (ho.loci; (184)) and standardised heterozygosity (std.het; (185)) 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 (186) confirmed the genome-wide significance of these markers.

Internal relatedness (183), homozygosity by loci (184) and standardised heterozygosity (185) 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).

I confirmed the genome-wide significance of these estimates using heterozygosity-heterozygosity correlation (186). 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

I also analysed gene-expression inferred from mapping. Of the 353,055 reads 252,388 (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).

The number of reads mapping is given for each library in table 5.1, to get unbiased estimates of expression I removed also all contigs with a coverage lower than 32 reads overall and thus analysed 658 contigs.

Using the statistics of of Audic and Claverie (187) and filtering for relevant contrasts, 54 contigs showed an expression predominantly in the male library, 56 contigs in the female library. 56 contigs were primarily expressed in the libraries from Taiwan, 22 contigs in the European library.

Overrepresentation of of GO-terms differentially expressed between the male and female libraries highlighted especially ribosomal proteins, oxidoreductases and collagen processing enzymes as enriched (table 5.6 and additional figures **??**, **??** and **??**). These ribosomal proteins were all overexpressed in the male library, oxidoreductases and collagen processing enzymes were all overexpressed female libraries.

GO.ID	Term	Annotated	Significant	Expected	p-value
Molecular function					
GO:0005198	structural molecule activity	51	18	8.28	0.00019
GO:0016706	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxyge...	3	3	0.49	0.00407
GO:0004656	procollagen-proline 4-dioxygenase activity	2	2	0.32	0.02595
GO:0031543	peptidyl-proline dioxygenase activity	2	2	0.32	0.02595
GO:0034641	cellular nitrogen compound metabolic process	159	37	25.03	0.00020
Biological process					
GO:0048731	system development	146	35	22.98	0.00020
GO:0034621	cellular macromolecular complex subunit organization	73	22	11.49	0.00026
GO:0006807	nitrogen compound metabolic process	162	37	25.50	0.00034
GO:0032774	RNA biosynthetic process	70	21	11.02	0.00043

Continued on next page

Table 5.6 – continued from previous page

GO.ID	Term	Annotated	Significant	Expected	p-value
GO:0071822	protein complex subunit organization	71	21	11.18	0.00055
GO:0043933	macromolecular complex subunit organization	82	23	12.91	0.00063
GO:0000022	mitotic spindle elongation	19	9	2.99	0.00080
GO:0051231	spindle elongation	19	9	2.99	0.00080
GO:0044281	small molecule metabolic process	188	40	29.59	0.00082
GO:0006139	nucleobase-containing compound metabolic process	139	32	21.88	0.00157
GO:0048856	anatomical structure development	188	39	29.59	0.00241
GO:0071841	cellular component organization or biogenesis at cellular level	139	31	21.88	0.00408
GO:0090304	nucleic acid metabolic process	105	25	16.53	0.00546
GO:0071842	cellular component organization at cellular level	135	30	21.25	0.00559
GO:0016070	RNA metabolic process	96	23	15.11	0.00797
GO:0040007	growth	138	30	21.72	0.00847
GO:0050789	regulation of biological process	198	39	31.17	0.00952
GO:0042274	ribosomal small subunit biogenesis	10	5	1.57	0.01084
GO:0009791	post-embryonic development	116	26	18.26	0.01151
GO:0007275	multicellular organismal development	221	42	34.79	0.01156
GO:0022414	reproductive process	105	24	16.53	0.01280
GO:0042157	lipoprotein metabolic process	7	4	1.10	0.01335

Continued on next page

Table 5.6 – continued from previous page

GO.ID	Term	Annotated	Significant	Expected	p-value
GO:0007051	spindle organization	27	9	4.25	0.01435
GO:0007052	mitotic spindle organization	27	9	4.25	0.01435
GO:0040009	regulation of growth rate	62	16	9.76	0.01599
GO:0040010	positive regulation of growth rate	62	16	9.76	0.01599
GO:0018988	molting cycle, protein-based cuticle	23	8	3.62	0.01616
GO:0010467	gene expression	114	25	17.94	0.01935
GO:0042303	molting cycle	24	8	3.78	0.02127
GO:0071840	cellular component organization or biogenesis	171	34	26.92	0.02143
GO:0032501	multicellular organismal process	241	44	37.94	0.02183
GO:0009416	response to light stimulus	8	4	1.26	0.02360
GO:0032502	developmental process	227	42	35.73	0.02409
GO:0008543	fibroblast growth factor receptor signaling pathway	2	2	0.31	0.02437
GO:0018401	peptidyl-proline hydroxylation to 4-hydroxy-L-proline	2	2	0.31	0.02437
GO:0019471	4-hydroxyproline metabolic process	2	2	0.31	0.02437
GO:0019511	peptidyl-proline hydroxylation	2	2	0.31	0.02437
GO:0046887	positive regulation of hormone secretion	2	2	0.31	0.02437
GO:0071570	cement gland development	2	2	0.31	0.02437
GO:0000279	M phase	44	12	6.93	0.02555
GO:0009792	embryo development ending in birth or egg hatching	123	26	19.36	0.02787
GO:0016043	cellular component organization	167	33	26.29	0.02838

Continued on next page

Table 5.6 – continued from previous page

GO.ID	Term	Annotated	Significant	Expected	p-value
GO:0009152	purine ribonucleotide biosynthetic process	5	3	0.79	0.02925
GO:0009260	ribonucleotide biosynthetic process	5	3	0.79	0.02925
GO:0002164	larval development	106	23	16.69	0.03108
GO:0042254	ribosome biogenesis	21	7	3.31	0.03144
GO:0000003	reproduction	137	28	21.56	0.03399
GO:0022613	ribonucleoprotein complex biogenesis	26	8	4.09	0.03482
GO:0065007	biological regulation	217	40	34.16	0.03874
GO:0007010	cytoskeleton organization	57	14	8.97	0.03908
GO:0045927	positive regulation of growth	68	16	10.70	0.03978
GO:0071843	cellular component biogenesis at cellular level	27	8	4.25	0.04344
GO:0048518	positive regulation of biological process	127	26	19.99	0.04357
GO:0034645	cellular macromolecule biosynthetic process	103	22	16.21	0.04358
GO:0000226	microtubule cytoskeleton organization	32	9	5.04	0.04471
GO:0007017	microtubule-based process	32	9	5.04	0.04471
GO:0006364	rRNA processing	18	6	2.83	0.04643
GO:0044267	cellular protein metabolic process	134	27	21.09	0.04769
GO:0002119	nematode larval development	104	22	16.37	0.04876
GO:0009059	macromolecule biosynthetic process	104	22	16.37	0.04876
GO:0030529	ribonucleoprotein complex	62	20	9.84	0.00022
GO:0043228	non-membrane-bounded organelle	115	28	18.25	0.00178

Continued on next page

Table 5.6 – continued from previous page

GO.ID	Term	Annotated	Significant	Expected	p-value
GO:0043232	intracellular non-membrane-bounded organelle	115	28	18.25	0.00178
GO:0044444	cytoplasmic part	258	48	40.95	0.00181
GO:0043227	membrane-bounded organelle	251	47	39.84	0.00274
GO:0043231	intracellular membrane-bounded organelle	251	47	39.84	0.00274
GO:0005829	cytosol	149	33	23.65	0.00306
GO:0031981	nuclear lumen	66	18	10.48	0.00538
GO:0005618	cell wall	17	7	2.70	0.00922
GO:0070013	intracellular organelle lumen	92	22	14.60	0.01115
GO:0043226	organelle	270	48	42.86	0.01309
GO:0043229	intracellular organelle	270	48	42.86	0.01309
GO:0030312	external encapsulating structure	18	7	2.86	0.01324
GO:0044446	intracellular organelle part	193	38	30.63	0.01332
GO:0009536	plastid	27	9	4.29	0.01507
GO:0044422	organelle part	195	38	30.95	0.01703
GO:0043233	organelle lumen	95	22	15.08	0.01721
GO:0022627	cytosolic small ribosomal subunit	15	6	2.38	0.01909
GO:0031974	membrane-enclosed lumen	97	22	15.40	0.02257
Cellular compartment					
GO:0045169	fusome	2	2	0.32	0.02477
GO:0070732	spindle envelope	2	2	0.32	0.02477
GO:0015935	small ribosomal subunit	16	6	2.54	0.02684
GO:0005737	cytoplasm	275	48	43.65	0.02798
GO:0009507	chloroplast	25	8	3.97	0.02868
GO:0005791	rough endoplasmic reticulum	5	3	0.79	0.02991
GO:0005811	lipid particle	30	9	4.76	0.03102

Continued on next page

Table 5.6 – continued from previous page

GO.ID	Term	Annotated	Significant	Expected	p-value
GO:0005773	vacuole	46	12	7.30	0.03833

Table 5.6: Over-representation of GO-terms in differentially expressed between male and female worms - Significance level (p.value) for over-representation are given along with the number of differentially expressed contigs (Significant) and the number of contigs with this annotation analysed (Annotated) and the description of the GO-term (Term). For a graph of induced GO-terms see also additional figures **??**, **??** and **??**.

Overrepresentation of of GO-terms differentially expressed between libraries from worms of European and Asian origin highlighted catalytic activity especially related to energy metabolism (table 5.7 and additional figures **??**, **??** and **??**). Acyltransferase contigs were all upregulated in the European libraries. However, the expression patterns for other contigs connected to metabolism did not show concerted up or down-regulation (e.g. for "steroid biosynthetic process" 2 contigs were downregulated in the European library, 3 contigs upregulated).

GO.ID	Term	Annotated	Significant	Expected	p-value
Molecular function					
GO:0016408	C-acyltransferase activity	3	3	0.37	0.0018
GO:0016747	transferase activity, transferring acyl groups other than amino-acyl groups	4	3	0.50	0.0065
GO:0003824	catalytic activity	158	27	19.62	0.0088
GO:0016746	transferase activity, transferring acyl groups	8	4	0.99	0.0099
GO:0001871	pattern binding	2	2	0.25	0.0151
GO:0003682	chromatin binding	2	2	0.25	0.0151
GO:0003985	acetyl-CoA C-acetyltransferase activity	2	2	0.25	0.0151
GO:0008061	chitin binding	2	2	0.25	0.0151
GO:0030247	polysaccharide binding	2	2	0.25	0.0151
GO:0003713	transcription coactivator activity	6	3	0.75	0.0273

Continued on next page

Table 5.7 – continued from previous page

GO.ID	Term	Annotated	Significant	Expected	p-value
GO:0005543	phospholipid binding	6	3	0.75	0.0273
GO:0004090	carbonyl reductase (NADPH) activity	3	2	0.37	0.0417
GO:0008289	lipid binding	12	4	1.49	0.0483
GO:0016853	isomerase activity	12	4	1.49	0.0483
Biological process					
GO:0016126	sterol biosynthetic process	5	4	0.60	0.00083
GO:0048732	gland development	9	5	1.08	0.00173
GO:0016125	sterol metabolic process	6	4	0.72	0.00228
GO:0006694	steroid biosynthetic process	10	5	1.20	0.00316
GO:0006338	chromatin remodeling	4	3	0.48	0.00596
GO:0006695	cholesterol biosynthetic process	4	3	0.48	0.00596
GO:0044281	small molecule metabolic process	188	30	22.63	0.00748
GO:0008202	steroid metabolic process	12	5	1.44	0.00825
GO:0042180	cellular ketone metabolic process	57	13	6.86	0.00845
GO:0023051	regulation of signaling	28	8	3.37	0.01087
GO:0019219	regulation of nucleobase-containing compound metabolic process	41	10	4.94	0.01412
GO:0001655	urogenital system development	2	2	0.24	0.01416
GO:0001822	kidney development	2	2	0.24	0.01416
GO:0006611	protein export from nucleus	2	2	0.24	0.01416
GO:0007528	neuromuscular junction development	2	2	0.24	0.01416
GO:0009953	dorsal/ventral pattern formation	2	2	0.24	0.01416

Continued on next page

Table 5.7 – continued from previous page

GO.ID	Term	Annotated	Significant	Expected	p-value
GO:0048581	negative regulation of post-embryonic development	2	2	0.24	0.01416
GO:0048741	skeletal muscle fiber development	2	2	0.24	0.01416
GO:0051124	synaptic growth at neuromuscular junction	2	2	0.24	0.01416
GO:0070050	neuron homeostasis	2	2	0.24	0.01416
GO:0072001	renal system development	2	2	0.24	0.01416
GO:0006082	organic acid metabolic process	54	12	6.50	0.01489
GO:0019752	carboxylic acid metabolic process	54	12	6.50	0.01489
GO:0043436	oxoacid metabolic process	54	12	6.50	0.01489
GO:0008152	metabolic process	266	37	32.02	0.01526
GO:0006355	regulation of transcription, DNA-dependent	30	8	3.61	0.01697
GO:0019953	sexual reproduction	44	10	5.30	0.02361
GO:0048747	muscle fiber development	6	3	0.72	0.02503
GO:0051171	regulation of nitrogen compound metabolic process	51	11	6.14	0.02556
GO:0009966	regulation of signal transduction	21	6	2.53	0.02842
GO:0032787	monocarboxylic acid metabolic process	21	6	2.53	0.02842
GO:0051252	regulation of RNA metabolic process	33	8	3.97	0.03036
GO:0048545	response to steroid hormone stimulus	16	5	1.93	0.03141
GO:0065008	regulation of biological quality	81	15	9.75	0.03399
GO:0050794	regulation of cellular process	151	24	18.18	0.03420

Continued on next page

Table 5.7 – continued from previous page

GO.ID	Term	Annotated	Significant	Expected	p-value
GO:0010033	response to organic substance	60	12	7.22	0.03487
GO:0048609	multicellular organismal reproductive process	60	12	7.22	0.03487
GO:0002026	regulation of the force of heart contraction	3	2	0.36	0.03923
GO:0007416	synapse assembly	3	2	0.36	0.03923
GO:0007431	salivary gland development	3	2	0.36	0.03923
GO:0007435	salivary gland morphogenesis	3	2	0.36	0.03923
GO:0007559	histolysis	3	2	0.36	0.03923
GO:0007595	lactation	3	2	0.36	0.03923
GO:0016271	tissue death	3	2	0.36	0.03923
GO:0022612	gland morphogenesis	3	2	0.36	0.03923
GO:0030518	steroid hormone receptor signaling pathway	3	2	0.36	0.03923
GO:0030522	intracellular receptor mediated signaling pathway	3	2	0.36	0.03923
GO:0030879	mammary gland development	3	2	0.36	0.03923
GO:0034612	response to tumor necrosis factor	3	2	0.36	0.03923
GO:0035070	salivary gland histolysis	3	2	0.36	0.03923
GO:0035071	salivary gland cell autophagic cell death	3	2	0.36	0.03923
GO:0035220	wing disc development	3	2	0.36	0.03923
GO:0035272	exocrine system development	3	2	0.36	0.03923
GO:0043628	ncRNA 3'-end processing	3	2	0.36	0.03923
GO:0045540	regulation of cholesterol biosynthetic process	3	2	0.36	0.03923
GO:0050808	synapse organization	3	2	0.36	0.03923

Continued on next page

Table 5.7 – continued from previous page

GO.ID	Term	Annotated	Significant	Expected	p-value
GO:0051091	positive regulation of sequence-specific DNA binding transcription factor activity	3	2	0.36	0.03923
GO:0051262	protein tetramerization	3	2	0.36	0.03923
GO:0051289	protein homotetramerization	3	2	0.36	0.03923
GO:0090181	regulation of cholesterol metabolic process	3	2	0.36	0.03923
GO:0032504	multicellular organism reproduction	61	12	7.34	0.03954
GO:0002165	instar larval or pupal development	7	3	0.84	0.04016
GO:0003015	heart process	7	3	0.84	0.04016
GO:0007589	body fluid secretion	7	3	0.84	0.04016
GO:0048872	homeostasis of number of cells	7	3	0.84	0.04016
GO:0060047	heart contraction	7	3	0.84	0.04016
GO:0006351	transcription, DNA-dependent	41	9	4.94	0.04017
GO:0009308	amine metabolic process	41	9	4.94	0.04017
GO:0006066	alcohol metabolic process	35	8	4.21	0.04262
GO:0006357	regulation of transcription from RNA polymerase II promoter	12	4	1.44	0.04362
GO:0009968	negative regulation of signal transduction	12	4	1.44	0.04362
GO:0010648	negative regulation of cell communication	12	4	1.44	0.04362
GO:0023057	negative regulation of signaling	12	4	1.44	0.04362
GO:0007165	signal transduction	69	13	8.31	0.04443
GO:0007276	gamete generation	42	9	5.06	0.04652

Continued on next page

Table 5.7 – continued from previous page

GO.ID	Term	Annotated	Significant	Expected	p-value
GO:0009888	tissue development	42	9	5.06	0.04652
GO:0044237	cellular metabolic process	255	35	30.69	0.04950
Cellular compartment					
GO:0031967	organelle envelope	47	12	5.52	0.0033
GO:0031975	envelope	48	12	5.64	0.0040
GO:0005740	mitochondrial envelope	29	8	3.41	0.0116
GO:0005643	nuclear pore	2	2	0.23	0.0135
GO:0046930	pore complex	2	2	0.23	0.0135
GO:0005739	mitochondrion	93	17	10.92	0.0184
GO:0031966	mitochondrial membrane	28	7	3.29	0.0322
GO:0005902	microvillus	3	2	0.35	0.0374
GO:0044429	mitochondrial part	36	8	4.23	0.0432

Table 5.7: Over-representation of GO-terms in differentially expressed between worms from Asia and Europe - Significance level (p.value) for over-representation are given along with the number of differentially expressed contigs (Significant) and the number of contigs with this annotation analysed (Annotated) and the description of the GO-term (Term). For a graph of incuced GO-terms see also additional figures **??**, **??** and **??**.

Enrichment of signal-positives was not found in any category of overexpressed genes. Differntially expressed genes also showed no pattern of enrichement in conservation categories and no enrichment of *C. elegans* orthologs with lethal/non-lethal RNAi-phenotypes.

Significantly elevated dn/ds was found for contigs differentially expressed according to worm-origin (Fisher's exact test p=0.005; also both up- or downregulated were significant). Contigs overexpressed in the female libraries showed elevated levels of dn/ds (Fisher's exact test p=0.035). In contrast male overexpressed genes showed decreased levels of dn/ds (Fisher's exact test p=0.015). Within these groups there was no correlation between dn/ds and log-fold-change values for gene-expression.

6

Transcriptomic divergence in a common garden experiment

6.1 Infection experiments

Dissection of eels 55-57 after infection (dpi) showed higher recovery of European worms in *An. anguilla* and higher recovery of Taiwanese worms in *An. japonica*, compared to the other parasite populations. In other words, in host-parasite combinations of matching origin, more parasites were recovered.

In the host-species/parasite-population pairs found in nature roughly eight or nine adult worms could be recovered 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 individuals more were recorded from both larval stages in the host/parasite combination found in nature.

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

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 as a fitness component.

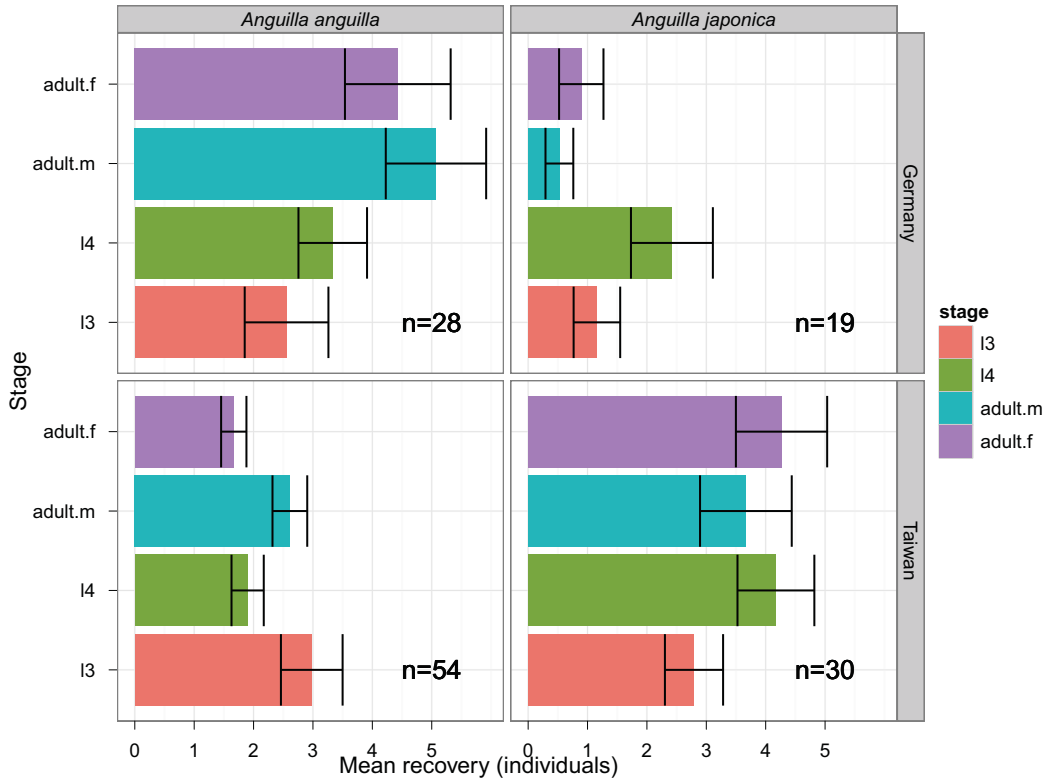

Figure 6.1: Recovery of worms in coinoculation experiment - Mean numbers of worms recovered after 55-57 dpi for sample sizes given as n=x. Error-bars indicate the standard error (s.e.) of the mean. Recovered lifecycle stages of the parasite are listed separately as L3-larvae (l3), L4-larvae (l4), adult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

	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 populations) further rows give variations for each factor. Std. Error is the standard error of this value. Additionally the probability of a t-value as small or smaller than the observed t-value are given. The signature of local adaptation is visible in the highly significant interaction term.

6.2 Sample preparation and sequencing

Three biological replicates were obtained from each of the two worm populations in each of the two eel-host species for each of the two sexes of worms. This resulted in a total of 24 RNA-extractions prepared for sequencing: 3 individual female worms from each experimental group were chosen randomly to give in total twelve females. Additionally, from three individual male worms, and from 9 pools of male worms RNA was extracted (see table 6.2). Pools consisted of worms from one infected eel individual each. All worms or worm-pools were derived from infections of different eel individuals, with one exception from this form of statistical independence: from *An. japonica* European male worms as well as a female worm had to be prepared from the same eel individuals. It was impossible to extract enough RNA from all but the biggest male worms especially of the Japanese eel/European worm combination, leaving no other choice. Because of the small size of male worms it was generally not possible to randomly choose individuals. Preparation of sufficient amounts of RNA was only achieved in pools of the biggest individuals. All male worms were thus chosen for preparation based on large size, even when pools of worms were used.

Sequencing was performed in three multiplexed pools of eight libraries each. The samples were partitioned into these pools spreading replicates for each condition over all three pools 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, each read having a length of 51 bases and 270 bases mean insert size between the read pairs.

6.3 Examination of data-quality

Reads were mapped against the fullest pyrosequencing-assembly (see 4.8) using `BWA` (154). Of the 263,668,952 raw read-pairs 173,602,387 mapped uniquely to the assembly and were counted on a per-library base.

The technical replicates demonstrated very low differences as inferred from a clustering analysis using variance stabilised data and transposed euclidean distances between samples (see figure 6.2 a).

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

After another screening for spurious read-counts to low coverage transcripts and to transcripts of low reliability (lowCA in the 454-assembly; see 4.8) 137,477,156 read-

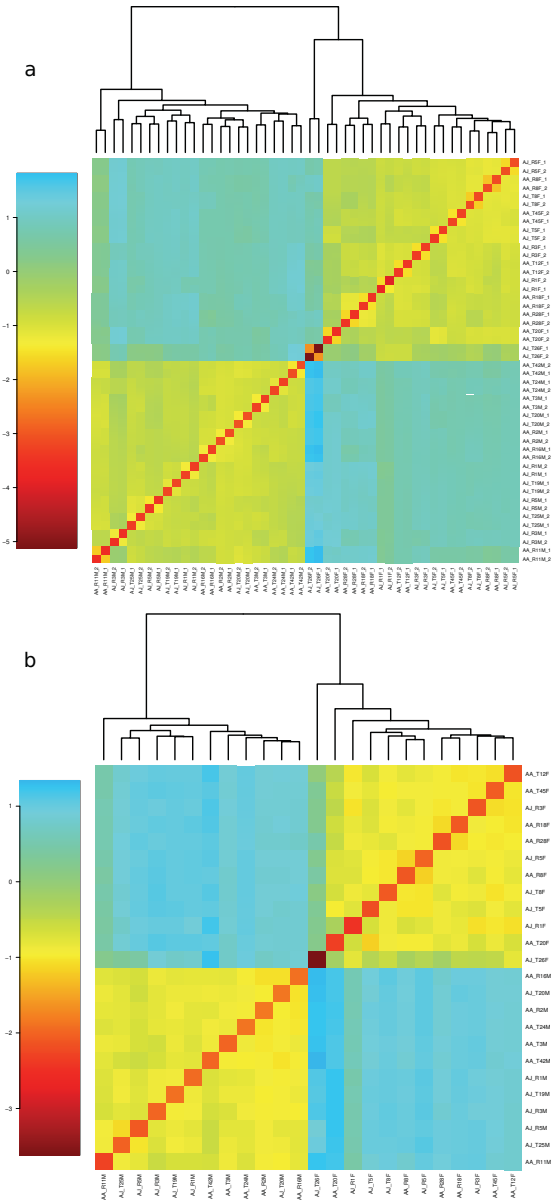

Figure 6.2: Distances between RNA-seq read-count for different samples - Euclidean distance (square distance between the two count vectors) for variance stabilised read-counts for all libraries including technical replicates; Red indicates low distance (high similarity), blue high distance (low similarity). a) Data before screening and summation of technical replicates. All technical replicates are clustered very closely, the distance between an outlier female sample (AJ_T26F) is high. b) Same illustration after summation of technical replicates and screening. Distance between outlier-sample and other female samples is reduced.

pairs were left for further analysis. Distribution of these read-pairs over libraries showed roughly 2.7-fold differences, with a mean of 5,728,215 reads and a range from 3,422,526 read-pairs for library AJ_R3M to 9,453,468 read-pairs for library AA_R8F (see 6.3).

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

In addition to hierarchical cluster analysis, also principal component analysis grouped libraries according to the sex of worms (the largest effect), but was unable to identify libraries with expression correlated in more subtle ways (see figure 6.2 b). Between-sample distance confirmed the hierarchical library clustering. Sex of the worms defined the overall distances between libraries, host- or population-differences were not visible in an overall effect in the top differentially expressed (DE) genes (see figure 6.3). Male samples showed a smaller distance in congruence due to the fact that they were made from pooled individuals balancing expression differences for individual worms.

6.4 Orthologous screening for expression differences

For the 7,520 contigs with expression values 4,382 *C. elegans*-orthologs and 4,292 *B. malayi*-orthologs were determined based on the annotation of our pyrosequencing-assembly (see 5.6). This resulted in 3,596 contigs with an expression measurement, having a measurement also for both corresponding orthologs (or group of orthologs) in both model-species and thus being available for analysis.

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

6.5 Expression differences in generalised linear models

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

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

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

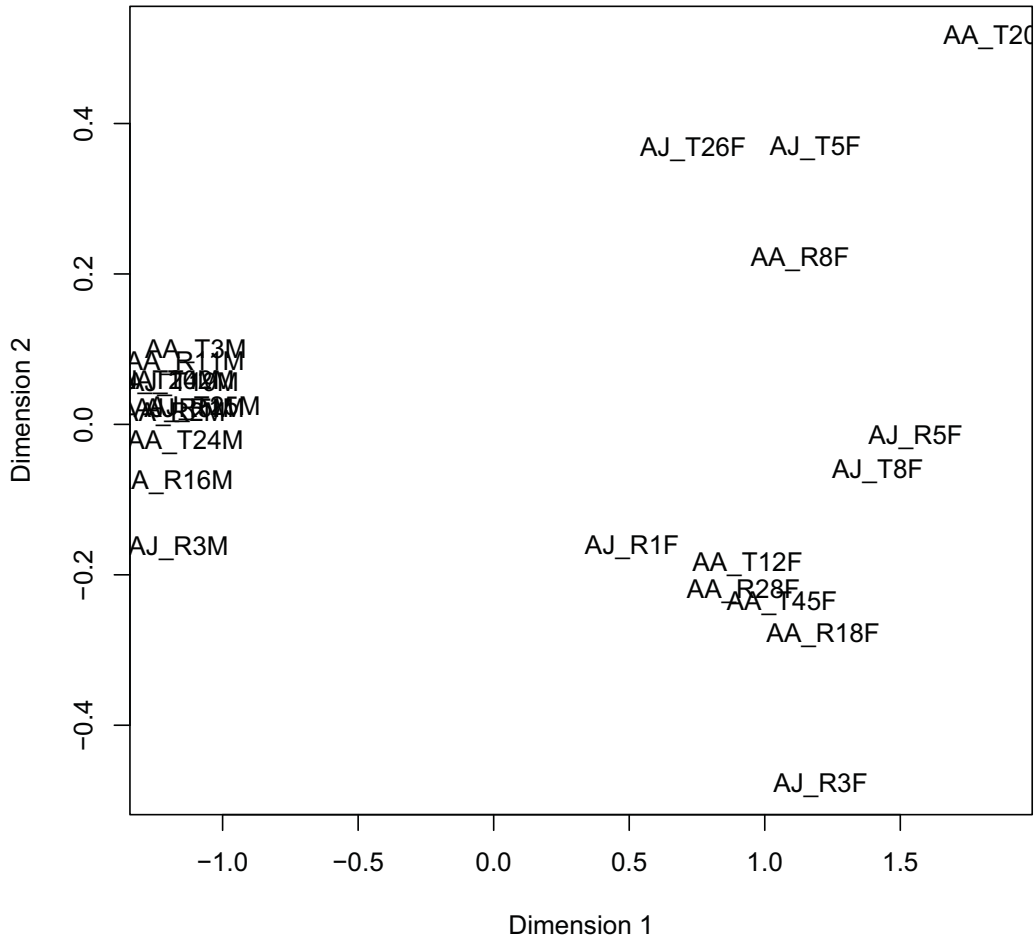

Figure 6.3: Principle coordinate plot for expression in RNA-seq libraries - Distance between sample-pairs is the root-mean-square deviation (Euclidean distance) for the most differentially expressed (DE) genes. Distances can be interpreted as the log2-fold-change of the genes with the biggest changes,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

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

Table 6.2: A summary of 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) and sex of worm(s) in preparation (F for female, M for male; last letter in label). The European samples were from two locations: river Rhine (R,) and Müggelsee near Berlin (B), the Taiwanese samples were from from Kao Ping River (K) and Yunlin county (Y). Additionally the intensity of infection (number of adult worms found in the infected eel; intensity) and the number of worms pooled in the preparation (only male worms are pooled for RNA extraction, individual female worms were used). Finally RNA-concentration in the preparation (conc in prep) is given in μg per ml.

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: Mapping Summary - Mapping is summarised for all 24 libraries. Rows indicate different libraries (worms or worm-pools as indicated in 6.2) raw.reads gives the number of read-pairs sequenced, raw.mapped the number of reads mapping uniquely with their best hit, tax.mapped the number of reads after subtraction of reads to putative eel-host derived contigs and screened after subtraction of all reads mapping not to the highCA-derived assembly or to contigs with overall counts less than 32.

29 (4 OC) in the European populations.

An important observation in these models is the prevalence of co-occurring significance of simple main effects. Expression changes overlapping for two main effects mean a significant difference in expression according to both factors. These differences are in the same direction for a combination of the factors. Most contigs DE according to the main effects of host-species or worm-population were also DE according to the sex of the worm. There was also a number of contigs differing for all three predictors in the same way. No contigs were observed DE in both the host-species and worm-population in the same direction but not according to worm-sex. From the 68 contigs DE in different *A. crassus*-populations, 38 were also DE according to worm sex and 16 according to all three main effects (see figure 6.4).

In addition, interaction-effects were also observed. 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 sympatric host-species/parasite-population combinations as found in nature: 7 contigs (0 OC) showed differential expression according to the worm-sex/eel-species interaction, 12 (3 OC) to worm-sex/parasite-population, 13 (2 OC) to host-species/parasite-population, 1 (0 OC) contig showed significance for the 3-way interaction (see figure 6.4). It should be noted, that conclusions drawn from of simple main effects do not necessarily hold for contigs with significant interaction effects (e.g. significantly higher expression in European population can then mean higher values only in one of the host-species).

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

6.6 Confirmation of contig categories through principal component analysis

I performed constrained redundancy analysis for the effects of eel-host and worm-population. This technique, similarly to principal components analysis, can partition the variance into orthogonal components, and additionally constrain one of the components to the factor of interest. I found that 7% of the variance in contigs DE between eel-hosts and 11% of the variance in contigs DE between worm-population explained by

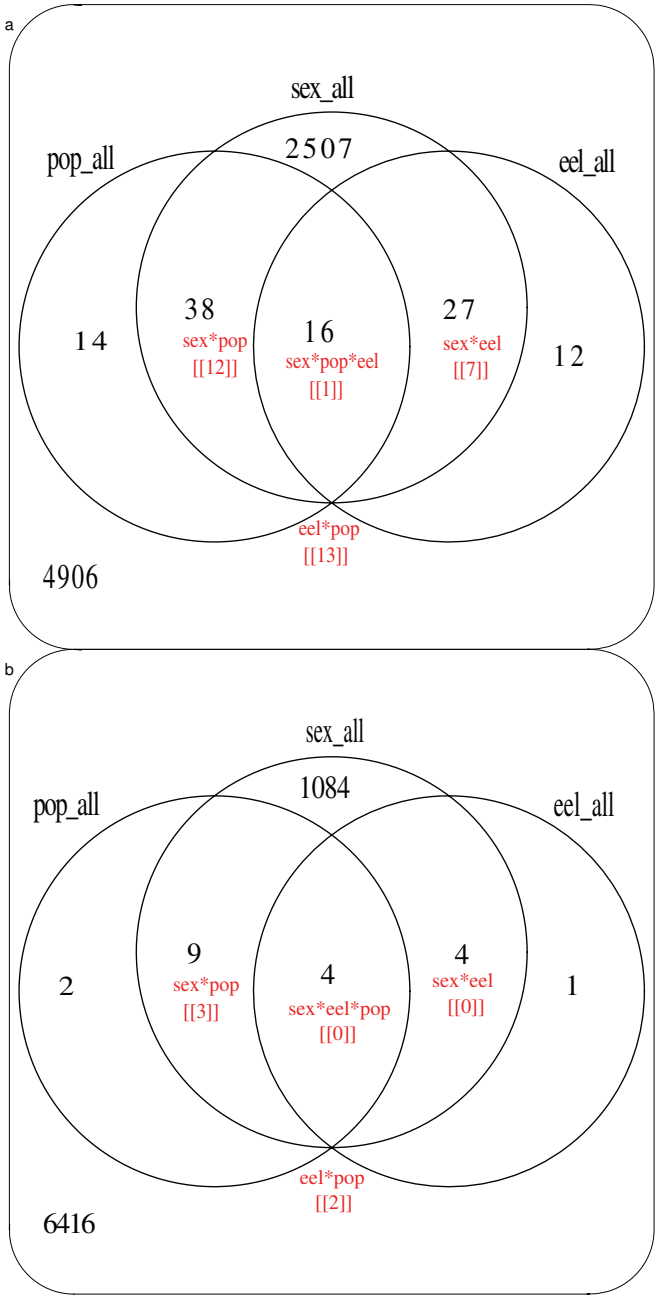

Figure 6.4: Venn diagram of contigs significant for different terms in `edgeR`-GLMs - Overlap between differences in simple main effects are given as black numbers in the Venn-Diagram. Numbers outside the circles in the lower left corner indicated non-significant contigs. The number of significant contigs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

the corresponding factor. In both evaluations more than 50% of the remaining variance could be explained by a single principal component, to which sex contributed over 99% (loading) (see figure 6.5 a and 6.6 a). When only OC-DE contigs were considered the explained variance for difference between eel-host dropped to 3.3% and the explained variance for differences between worm-population was raised to 23%, while the sex-effect explained 70% and 50% of the variance (see figure 6.5 b and 6.6 b). Significance of the constrained component evaluated by a permutation-test could be established at a $p < 0.05$ threshold for all but the OC eel-host DE subset.

6.7 Biological processes associated with DE contigs

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

For the differences between male and female worms enriched annotations can be summarised into three broad categories: Terms over-represented due to spermatogenesis (e.g. PP1-phosphatase and ester hydrolase are important for spermatogenesis in *C. elegans* (188, 189)) embryo development (many obvious terms) and terms for other processes more related to metabolic differences between males and females (such as oxidoreductase activity; see table 6.4 but also additional figures 9.10, 9.11 and 9.12).

GO.ID	Term	Annotated	Significant	Expected	p-value
Molecular function					
GO:0042578	phosphoric ester hydrolase activity	99	59	31.99	1.2e-08
GO:0016791	phosphatase activity	88	53	28.44	4.2e-08
GO:0004721	phosphoprotein phosphatase activity	65	42	21.00	6.5e-08
GO:0004722	protein serine/threonine phosphatase act...	34	24	10.99	4.8e-06
GO:0005509	calcium ion binding	78	43	25.21	2.1e-05
GO:0046873	metal ion transmembrane transporter acti...	32	21	10.34	0.00010
GO:0003824	catalytic activity	1354	482	437.55	0.00015

Continued on next page

Table 6.4 – continued from previous page

GO.ID	Term	Annotated	Significant	Expected	p-value
GO:0016614	oxidoreductase activity, acting on CH-OH...	46	27	14.86	0.00018
GO:0016616	oxidoreductase activity, acting on the C...	42	25	13.57	0.00023
GO:0017018	myosin phosphatase activity	10	9	3.23	0.00027
Biological process					
GO:0050896	response to stimulus	1535	583	504.78	1.7e-10
GO:0006470	protein dephosphorylation	63	41	20.72	1.2e-07
GO:0007391	dorsal closure	32	25	10.52	1.7e-07
GO:0016476	regulation of embryonic cell shape	13	13	4.27	5.0e-07
GO:0001700	embryonic development via the syncytial ...	49	33	16.11	6.7e-07
GO:0007392	initiation of dorsal closure	15	14	4.93	1.7e-06
GO:0046664	dorsal closure, amnioserosa morphology c...	15	14	4.93	1.7e-06
GO:0016311	dephosphorylation	86	49	28.28	2.6e-06
GO:0042221	response to chemical stimulus	864	337	284.12	3.1e-06
GO:0007394	dorsal closure, elongation of leading ed...	11	11	3.62	4.7e-06
Cellular compartment					
GO:0031224	intrinsic to membrane	372	164	118.85	8.4e-08
GO:0016021	integral to membrane	368	162	117.58	1.2e-07
GO:0005576	extracellular region	250	115	79.88	7.7e-07
GO:0031226	intrinsic to plasma membrane	176	86	56.23	1.0e-06
GO:0005887	integral to plasma membrane	172	84	54.95	1.4e-06

Continued on next page

Table 6.4 – continued from previous page

GO.ID	Term	Annotated	Significant	Expected	p-value
GO:0030054	cell junction	145	72	46.33	3.9e-06
GO:0000267	cell fraction	435	179	138.98	6.4e-06
GO:0016020	membrane	1154	417	368.70	3.6e-05
GO:0000164	protein phosphatase type 1 complex	14	12	4.47	4.9e-05
GO:0072357	PTW/PP1 phosphatase complex	14	12	4.47	4.9e-05

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.

For the lower number of contigs DE between host-species inference of higher order terms was obviously only possible to a limited extent and in part also unnecessary, because annotations can be interpreted at face value. However, annotations for contigs DE between eel-hosts highlighted redundant terms associated with "antigen processing and presentation" proteins which are in mammals usually involved in antigen processing and cleavage of the invariant chain of the MHCII complex. These terms led to Contig566 and Contig26 and their *B. malayi*-orthologs "aspartic protease BmAsp-1, identical" and "eukaryotic aspartyl protease family protein". In blood feeding helminths these enzymes are in contrast usually involved in early cleavage events during the digestion of host haemoglobin (190).

For contigs DE between worm populations despite the limited number of DE contigs, enrichment analysis identified "oxidoreductase activity" as an informative significantly enriched higher level term (see figure 7.1). The biological processes "response to metal ion" and "mitochondiral electron transport" (see figure6.7) confirmed an evaluation linking these mainly to enzymes used in respiratory processes and highlighted additionally enzymes from lipid metabolism (especially β-oxidation of fatty acids) related to respriration and the availability of oxygen.

6.8 Clustering analysis

For the remainder of the text I will concentrate on these differences of the European and Taiwanese populations and mention the other differences only as far as they are

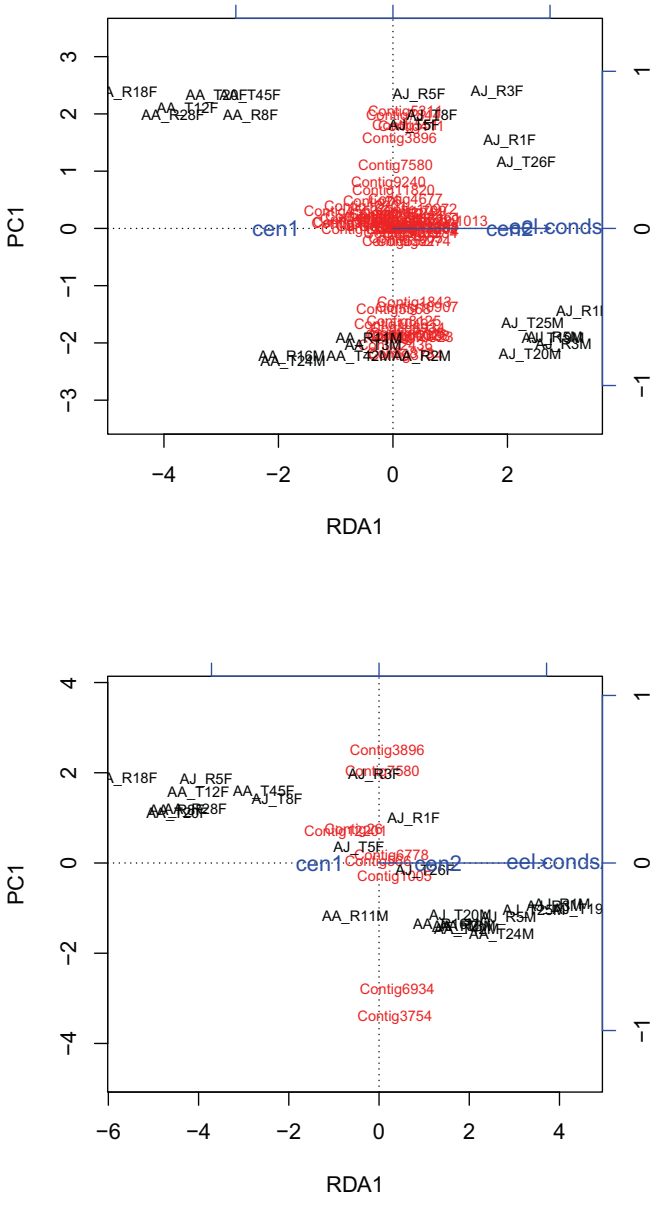

Figure 6.5: Constrained redundancy analysis for host-DE contigs - Eel-host differences are displayed as constrained component on the x-axis, the sex contributed >99% (loading) to the principal component on the y-axis. (a) Host differences partition the variance in samples in like expected for all contigs, the constrained component showed significance. (b) For OC contigs the constrained component fails to to partition the variance as expected, the component showed no significance for this subset of the data.

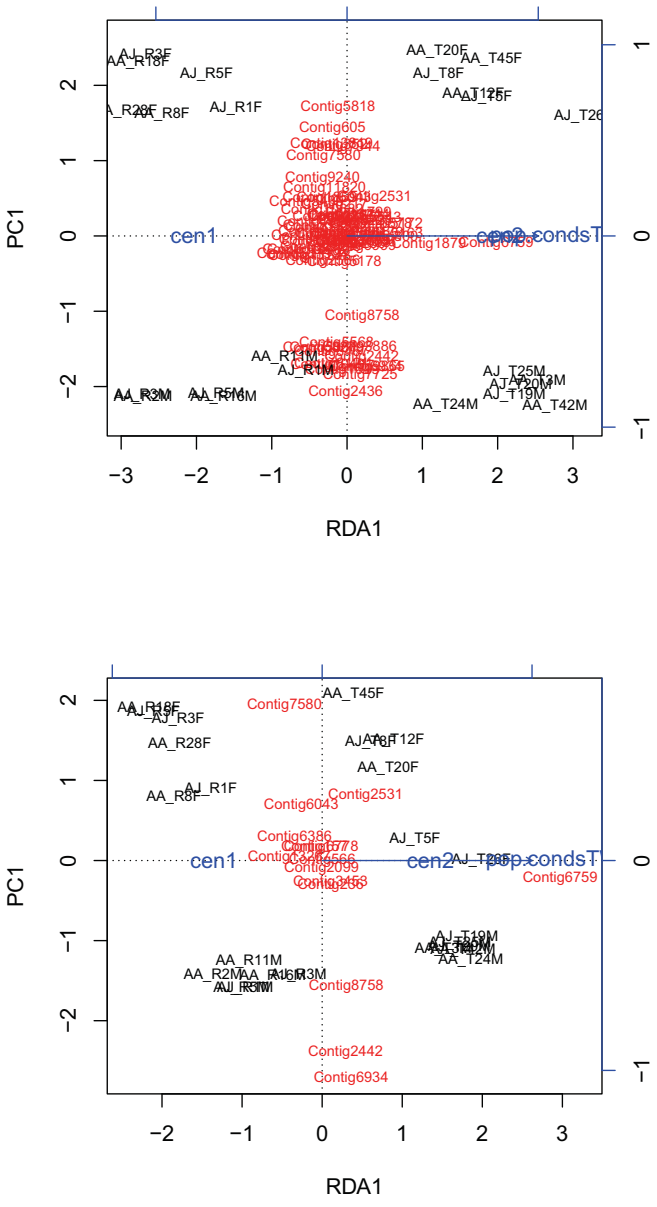

Figure 6.6: Constrained redundancy analysis for population-DE contigs - Population differences are displayed as constrained component on the x-axis, the principal component on the y-axis corresponds to the sex of the worm. Host differences partition the variance in samples like expected for all contigs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

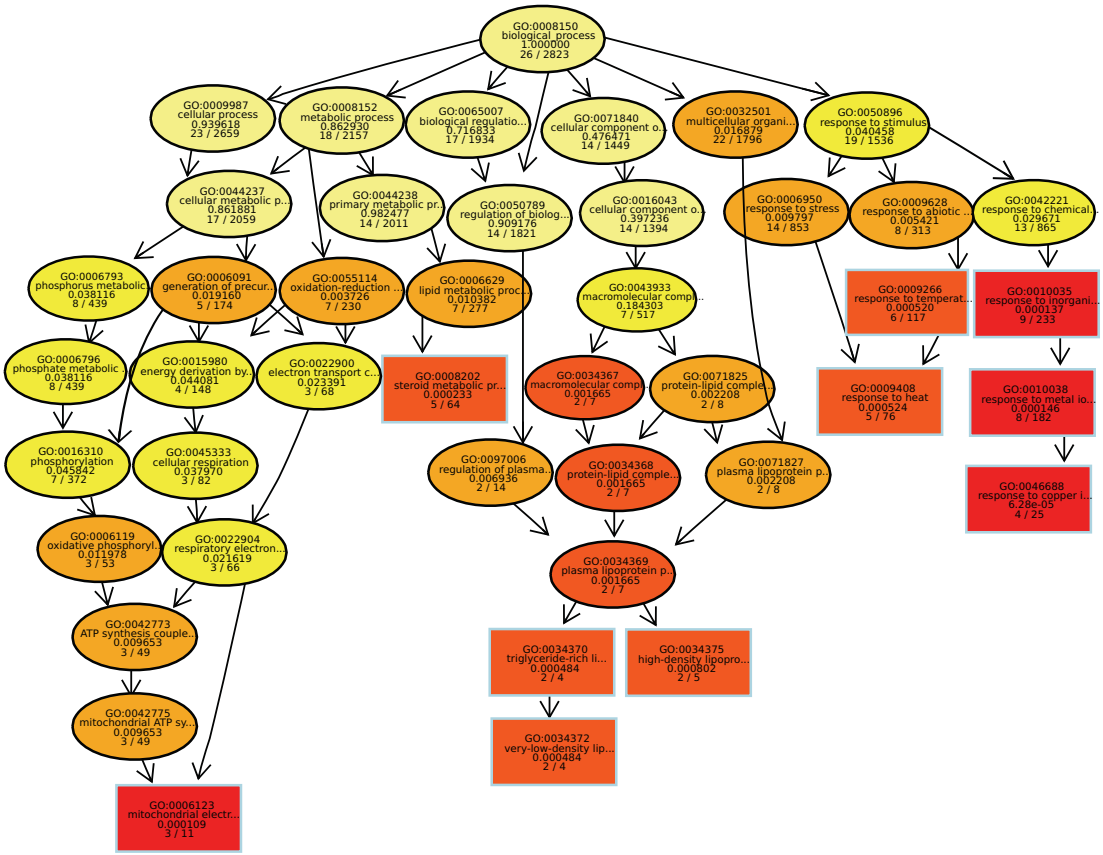

Figure 6.7: GO biological process graph for enriched terms in DE according to worm-population - Subgraph of the GO-ontology biological process category induced by the top 10 terms identified as enriched in DE genes between different parasite populations. Boxes indicate the 10 most significant terms. Box colour represents the relative significance, ranging from dark red (most significant) to light yellow (least significant). In each node the category-identifier, a (eventually truncated) description of the term, the significance for enrichment and the number of DE / total number of annotated gene is given. Black arrows indicate a is "is-a" relationship.

related to this focal factor. In 9.2 however, graphical analyses of the same type are presented for other factors.

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

Hierarchical clustering analyses of genes DE between populations confirmed the results of principal component based multivariate analysis. The main factor grouping libraries was the sex of the worm. A sub-grouping of samples fully according to European and Taiwanese populations was only observed for male worms. In female worms other unmeasured co-factors were preventing a clustering fully according to this factor. In male worm however, library clustering even followed a pattern of similar expression in according to the second factor of eel-host. These statements are true for both the full set of contigs (see figure 6.8) and OC contigs (see 7.2).

Clustering of genes revealed three co-regulated groups in the full set of contigs and the OC set. The first of gene-clusters (top in 6.8 and 7.2) was in sex-subgroups mainly following an expression pattern differing between populations. The second gene-group was much larger in the full set than in the OC set of contigs (middle in 6.8). It was only very weakly reacting to any other factor but sex and was very sparsely annotated (therefore this group was much smaller in the OC set 7.2). The third gene-group found again in both the full and OC contigs (bottom in 6.8 and 7.2) was reacting on both the host and population factor in a converse way. Contigs in this cluster were mainly found to be significant for interaction effects.

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

6.9 Single gene differences

Tables on single transcript values of OC contigs DE between eel-hosts and populations can be found in additional tables 9.3 and 9.4. Obviously for some contigs differences significant in the model are rendered inaccessibly by comparing simple mean values because of superposed interaction effects or overwhelming general effects of worm sex.

Cytochrome C oxidase subunit 2 (COXII) shows the clearest of all expression pat-

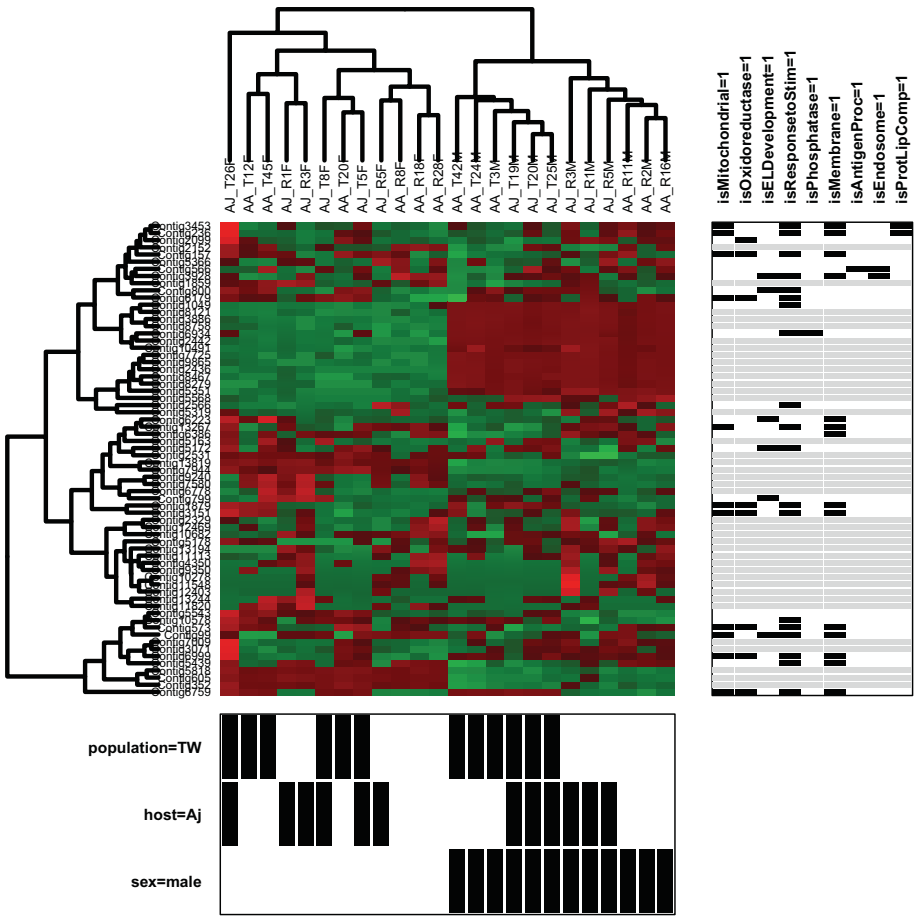

Figure 6.8: Clustering of expression values for contigs DE between populations - A heatmap of variance/mean stabilised expression values. Deprograms are based on hierarchical clustering. Green indicates expression below the mean, red above the mean. Experimental conditions are indicated by black bars for groups of samples (columns) below the plot. Presence GO-term annotation for contigs (rows) are given as black bars right to the plot: isOxidoreductase = GO:0016491, oxidoreductase activity; isMitochondrial = GO:0005739, mitochondrion; isELDevelopment = GO:0002164, larval development or GO:0009791, post-embryonic development; isResponsetoStim = GO:0050896, response to stimulus; isPhosphatase = GO:0016791, phosphatase; isMembrane = GO:0016020, membrane; isAntigenProc = GO:0002478, antigen processing and presentation of exogenous peptide antigen; isEndosome = GO:0005768, endosome; isProtLipComp = GO:0032994, protein-lipid complex. Grey bars indicate no annotation available.

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

7

Discussion

7.1 Pilot-sequencing

In was not achieved to 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 haemoglobin can inhibit PCR reactions (191) and reverse transcription (192).

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

In fact all sequenced ESTs, for which host-origin was inferred were later found also in pyrosequencing: The observation of haemoglobin and ferritin subunits from *An. anguilla* are expected, as fish erythrocytes contain a nucleus and still transcribe genes actively (193). These are typical proteins for the functioning of red blood cells. The observation of fish cyclin G1 and cohesin, genes expressed in mitosis, is remarkable, as fish erythrocytes are thought to exhibit low rates of mitosis (194). Other observations 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 (from all sequencing technologies) very promising, it is however beyond the scope of the present thesis.

7.2 Pyrosequencing

I 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, I have used exhaustive filtering to attempt to remove all host-derived, and host-associated organism-derived contamination from the data. To do this I 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 (45) 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 (157), and an analysis of lake sturgeon tested and rejected hypotheses of horizontal gene-transfer when xenobiont sequences was identified (195). A custom pipeline for transcriptome assembly from pyrosequencing reads (196) proposed the use of EST3 (197) to infer sequence origin based simply on nucleotide frequency. I was not able to use this approach successfully, probably due to the fact that xenobiont sequences in my data set derive from multiple sources with different GC content and codon usage.

Compared to other NGS transcriptome sequencing projects (198), 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. (199)), comparable to the mira sub-assembly in my 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 my protein prediction. This bias is near-ubiquitous in deep transcriptome sequencing projects (e.g. (200)).

I was able to obtain high-quality annotations for a large set of TUGs: For 40% of the complete assembly and 60% of my 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` (180).

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 my dataset. This implies that my transcriptome is truly a representative partial genome (116) 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 (124). 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* (121), where signal positives were reported as less conserved. Interestingly enrichment of signal sequence bearing TUGs in my dataset was constrained to sequences novel in nematodes and *A. crassus* (i.e. not to the level of clade III). This may be explained, with 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 my 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.

I 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 (135). 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 (201)) or *Drosophila* (2.07 (202)). 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 (203), the assumption of negative (purifying) selection on most protein-coding genes makes lower mean values seem more plausible. I 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 (204) and (198)) could be avoided.

Also in comparison with published intra-species values of dn/ds my final estimate seems plausible: in transcripts from the female reproductive tract of *Drosophila* dn/ds was 0.15 (205) and 0.21 in the male reproductive tract (206) (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* (126) 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 positive selection (205) instead of the classical threshold of 1 (207). The use of this threshold for positive selection led to the identification of over-represented of GO-term highlighting very interesting transcripts:

Twelve peptidases under positive selection (from 43 with a dn/ds obtained) meant an enrichment in the category. All twelve 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 (22). The twelve proteinases under positive selection could be the targets of the adaptive immunity developed against *A. crassus* (44, 208), which is often only elicited against subtypes of larvae (209).

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 (210) and should be 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 my study are in agreement with microsatellite data (10) showing reduced heterozygosity in European populations of *A. crassus*.

I employed methods to developed for the comparison of cDNA-libraries to make inference about possible differential gene-expression according to experimental groups (origin of sequencing-libraries) (187). Such approaches are widely used with pyrosequencing-data (e.g. (126)). For the statistically valid comparison of conditions however, the unit of replication would be the individual library and approaches respecting this fact would be desirable. However, I was not able to use the R-packages `DESeq` (164) or `edgeR` (165) developed for count data from deep sequencing (but more targeted towards RNA-seq on the solexa-platform) as both repetition and throughput of my pyrosequencing experiment were too low. As a result the differentially expressed genes are by no means significant for the investigated conditions, but just for the specific cDNA-libraries. With these reservations we identified genes differentially expressed between libraries prepared from worms of different sex and worms from different origin.

Genes over-expressed in male *A. crassus* comprise major sperm proteins well known for their high expression in nematode sperm (211). A surprise was the overexpression of ribosomal proteins in the male library.

That collagen processing enzymes are overexpressed in female worms, filled with developing embryos and larvae, is in line with a complicated regulation and modulation of collagen in nematode larval development (212).

The overexpression acetyl-CoA acetyltransferase in European worms are interesting especially because of the role of these enzymes in fatty-acid β-oxidation in peroxisomes and mitochondria (213). Together with a change in steroid metabolism and the enrichment of mitochondrially localised enzymes these are suggestive of changes in energy metabolism of *A. crassus* from different origins. Possible explanations would include a change to more or less aerobic processes in worms in Europe due to their bigger size and/or increased availability of nutrients.

Contigs overexpressed in the female libraries showed elevated levels of dn/ds but genes overexpressed in males decreased levels of dn/ds. The first finding is unexpected, as overexpressed in female libraries will also contain contigs related to larval development (such as the collagen modifying enzymes discussed above), these larval transcripts in turn are expected to be under purifying selection because of pleiotropic effects of genes in early development (214). Also the second finding is in slight contrast to published results for male specific traits and transcripts, often showing hallmarks of positive selection (206, 215). In *Ancylostoma caninum* however, female-specific transcripts

showed an enrichment of "parasitism genes" (126) and a possible expansion would be a similar enrichment of positively selected parasitism related genes in my dataset. For males the decreased dn/ds can be explained by the high number of ribosomal proteins, which are all showing very low levels of dn/ds (that these proteins are found differentially expressed remains puzzling though), while single transcripts e.g. major sperm protein (expressed in the male library only) showed elevated dn/ds but did not level the overall effect. But this also has a positive aspect: it is unlikely that correlation of differential expression with positive selection results from mapping artefacts, as all the ribosomal proteins identified overexpressed in males have very low dn/ds.

Genes differential expressed according to worm-origin (in either direction) showed significantly elevated levels of dn/ds. This is interpretable as a correlation between sequence evolution and phenotypic modification in different host-environments or even correlation between sequence evolution and evolution of gene-expression. Thus, whether expression of these genes is modified in different hosts or evolved rapidly in a contemporary divergence between European and Asian populations of *A. crassus*, is in the centre of a future research program building on the reference transcriptome presented here. For such an analysis it is important to disentangle the influence of the host and the nematode population in a coinoculation experiment. Such a project will also use the individual worm as the level of replication for "conditions" (that is, worm-population and host-species) to allow rigid hypothesis testing. Based on the pilot evaluation presented here differences in these factors are expected overlap with differences in male vs. female worms and the careful cross-examination of the above factors with worm-sex is advised.

The *A. crassus* transcriptome provides a basis of molecular research on this important species. It further provides insight in the evolution of parasitism complementing the catalogue of available transcriptomic data with a member of the Spirurina phylogenetically distant to so far sequenced parasites in this clade. Differences in energy metabolism between European and Asian *A. crassus* constitute a candidate phenotype relevant for phenotypic modification or contemporary divergent evolution as well as for the long term evolution of parasitism.

7.3 Transcriptomic divergence in a common garden experiment

7.3.1 Recovery and adaptation

With some reservations discussed below my observation of higher recovery of adult worms from sympatric *A. crassus-Anguilla* spp. host-parasite combinations imply local adaptation of different worm populations to host-species.

The percentage of recovered European worms is in agreement with data from Knopf & Mahnke (37): roughly one-third for the host-parasite combination sympatric in Europe and only little over 10% for European worms applied back to *An. japonica*. This pattern of recovery was precisely inverted for the Taiwanese population of *A. crassus*, for which recovery was thus roughly 30% in the sympatric *An. japonica* and only 10% in *An. anguilla*. These data are not in complete agreement with findings by Weclawski *et al.* (unpublished; see 1.1.2.3), who recorded recovery at only slightly different time-points after infection (25, 50, 100 and 150 dpi). Similar to my study 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 of this diverged population in the Japanese eel. A possible explanation for these different results are interactions of host-parasite genotypes conditional on the environment (GxGxE interactions, see also 1.1.2.3). It is imaginable that the environment provided in the common-garden setting slightly differed between the two experiments (despite the fact that these experiments were performed in the same experimental setup).

It has to be emphasised 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 obvious direct fitness-consequences, a so called fitness-component. Fitness is defined as the differential contribution to the next generation, therefore such a fitness-component would ideally be a measurement on a single individual, and individual life-time reproductive success would be an ideal measurement. However, techniques to measure individual life-time reproductive 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 only a proxy, interpretable as a fitness-component. It is a composite measurement of the speed of development from previous lifecycle stages (or speed of migration towards the swimbladder) and of survival. While survival is surely an important component of fitness, it is not completely clear whether fast development and/or migration to the swimbladder are. It is possible that under certain conditions slower development could lead to higher fitness, if it

would, for example allow development without attracting the attention of the immune system.

Another slight problem with recovery in these experiments is that it is a mean measurement over many individuals. If one would want to find genotype associations with the most suboptimal phenotype it would not be possible to isolate individuals bearing this trait, because these would be dead or still on their way migrating to the swimbladder. Apart from the problem of clear definition and measurement, lifecycle traits are also notoriously complex in the underlying genetic architecture (216).

When I later venture into adaptive interpretations of the observed gene-expression differences it has to be remembered that these constitute nothing more than a molecular phenotype. This phenotype is not necessarily a fitness-component. It is one of the dangers of genomic data to forget the fundamental lesson from the debate initiated by Gould and Lewontin in 1979 (217). Briefly, while functional changes are often caused by selection, differences in function do not necessarily demonstrate the past or present action of selection. There is no way to infer the action of selection based on functional considerations, and even if selection can be inferred otherwise, it is not necessarily a particular observed variable trait that selection acted on (218).

7.3.2 Variance, stringency of analysis and general pattern

I decided on a study design using pools of individuals for one sex (males) and single individuals for the other. A study on *Fundulus heteroclitus* revealed that approximately 18% of the transcripts are differentially expressed between individual fish from the same population, grown under controlled environmental conditions (219). And it thus not surprising that between individual variation in female samples was leading to higher variance of these female samples compared to pooled male samples in my study.

This interindividual variation in gene-expression under a particular environmental condition is generally agreed to be closely linked to a genetic basis (220). For example in a cross between two parental strains of yeast the genetic component of variation was estimated from haploid segregants to be 84% (221). The genetic component was found to be the main factor determining expression level variability between two strains, sexes and ages of *Drosophila melanogaster* for 267 (7%) from 3,931 genes and at least 25% of the transcriptome were estimated to be affected mainly by genotypic factors in any of the groups (222). Variation in the regulation of gene-expression is thought to constitute a major source of evolutionary novelty (223).

A second study from the line of research on *Fundulus heteroclitus* (224) used genetic relatedness as inferred from phylogenetics to separate variation in gene-expression in

a common experimental environment into a neutral component and a selected component, this way removing variation most likely accounted for by the shared neutral evolutionary history. My case of *A. crassus* is potentially simpler: the investigated European populations are direct descendants and thus a subset of a Taiwanese source population. In fact I studied two European and two Taiwanese populations as a few hundred kilometers between the geographical origins of the two different locations in Germany and Taiwan probably constitute a barrier to gene-flow in a parasite with an aquatic intermediate host. However, I treated worms from both European and Taiwanese populations as replicates (and use the terminology of one European and one Asian population throughout the text) with the rationale of increasing variance for random genetic differences and raising the bar for potentially adaptive differences to be detected.

Given the sampling of only twelve Taiwanese worms the question could be raised, whether these constitute a representative sample of the true source population, of which a sub-population was funding European populations. A microsatellite study indicated gene-flow even between populations of *A. crassus* separated by thousands of kilometers in Asia (Japan and Taiwan) (10). Given the high interconnectivity of Taiwanese water systems used for aquaculture both by man-build structural links and anthropogenic exchange of fish, a sampling from two Taiwanese populations similarly neutrally diverged from the true European funding population seems very unlikely. The worms sampled from Taiwan can thus be regarded a sample of the (meta-)population appropriate for finding differences in relation to the source of the introduction.

Of no surprise was the abundance of differential expression between male and female worms in roughly one third of the genes. A large number of genes are known to be sex-specific, regulating ovulation and spermatogenesis throughout the metazoa and especially in nematodes (214). On top of these sex-specific genes there are large numbers of genes differently expressed due to differences in metabolism between males and females. Estimates for *Drosophila* based on similar sample sizes to those used in my study range between one- and two-thirds of the transcriptome showing sex-biased expression (222). In the liver transcriptome of *Mus musculus*, even 70% of transcripts have been shown to differ between sexes (225) (note however that this study used 169 female and 165 male mice to guarantee the finding of even the most subtle differences). Given the scale of these differences in other species my estimate of roughly one third of the transcripts in *A. crassus* showing differential expression according to the sex of the worms implies conservative thresholds used in the statistical analysis and moderate power for detection of differences.

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

In addition biological implications could produce false negatives in such an evaluation: All genes duplicated in *A. crassus* (a) and following antithetic expression patterns will be evaluated negatively, as will duplicated genes in any of the model-species (b) following such a pattern. However, there is no other choice then applying these stringent conditions to screen for artefacts producing the same patterns based on mapping to fragmented (a) or chimeric (b) reference contigs. I think that an evaluation based on this scrutinised confidence in an assembly previously computed from 454-data is even more appropriate then an analysis solely based on counts collapsed for orthologs excluding only possible fragmentation artefacts (as used e.g. in (161)).

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

Nevertheless it is surprising that less than 1% of transcripts were detected differentially expressed between worms in different host-species and less then 0.3% were confirmed with the orthologous-summation method. This was an unexpected finding, as the differences in the immune response of the host species have a big influence on other phenotypes of worms (36). In addition to the low number of genes, multifactorial analysis revealed that below 10% of the variance could be explained by host-species effect, even in significantly differential regulated genes for this factor.

Although these differences between worms in different host species were the most marginal of any of the factors, it is possible to connect some (at least two) of the genes to a prominent physiological difference: the digestion of haemoglobin. Two different aspartic proteases (both confirmed through orthologs, one of them differing for all three main effects, the other for an interaction of worm-sex and host species) known to be involved in the first steps of digestion of haemoglobin from other nematodes (190) were overexpressed in worms in *An. anguilla*. This expression phenotype could potentially be linked to the often observed phenotype of bigger size of *A. crassus* in this host (36), as the main contribution to this increase in size is the larger volume of host-blood taken

up by the parasite. Accordingly the parasite probably digests haemoglobin at a higher rate.

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

Another important finding was the large overlap in contigs expressed differentially depending on worm-sex and worm-population. Such an overlap is expected if genes expressed differentially according to sex are evolving faster towards a differential expression according to other factors. Faster evolution of reproductive (and especially male specific) traits has been shown in many species at a phenotypic and at a sequence level (215). In *Drosophila*, male reproductive proteins have been shown to evolve at elevated levels and under positive selection (206). Moreover, gene expression should evolve at a higher rate in sex-specific genes. Indeed the transcriptomes of *Drosophila* species show that interspecific expression divergence is sex dependent and the action of sex-dependent natural selection during species divergence has been inferred from this (226, 227).

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

7.3.3 Functions of genes with genetically fixed expression differences

From a functional perspective, genes identified to differ between populations can be categorised as important in general metabolic processes instead of specific host-parasite interactions. This constitutes a negative evaluation of one of my *a priori* hypotheses based on finding parasite-specific genes, identified as vaccine candidates in a number of nematodes, within the genes modified or diverged in my study (1.2.3). However, more direct host-parasite interactions are expected in tissue-dwelling larval stages (L3 and L4) and in fact most immunomodulators are expressed predominantly in these stages (118). Adults of *A. crassus* could thus be the wrong lifecycle stages to detect such expression differences, if they existed.

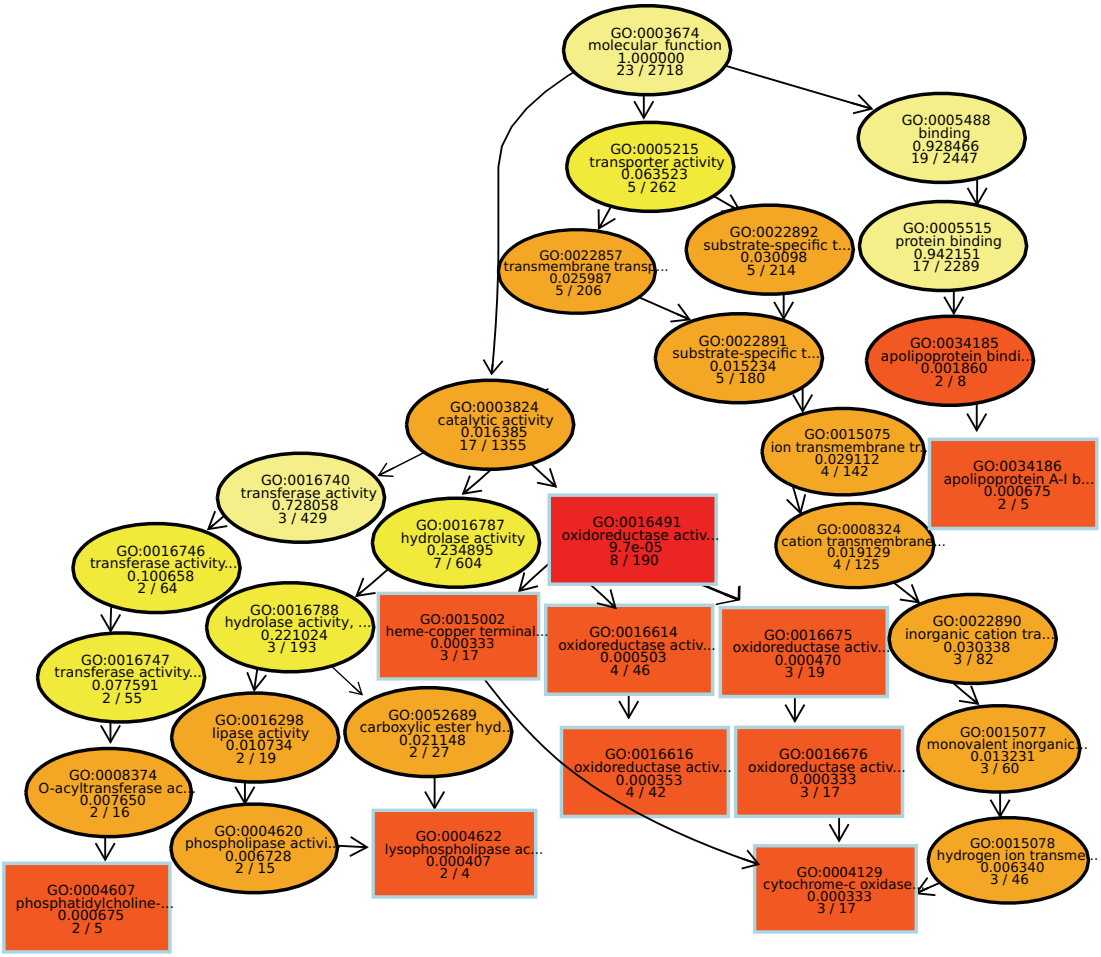

Figure 7.1: GO molecular function graph for enriched terms in DE according to worm-population - Subgraph of the GO-ontology molecular function category induced by the top 10 terms identified as enriched in DE genes between different parasite populations. Boxes indicate the 10 most significant terms. Box colour represents the relative significance, ranging from dark red (most significant) to light yellow (least significant). In each node the category-identifier, a (eventually truncated) description of the term, the significance for enrichment and the number of DE / total number of annotated genes is given. Black arrows indicate an "is-a" relationship.

7.3.3.1 Metabolism

Instead enzymes and enzyme subunits important for aerobic respiration are especially expressed at lower levels in European *A. crassus*. In fact, most transcripts significantly differing between populations were annotated as "oxidoreductase" in gene-ontology (GO). Downregulation of cytochrome C oxidase subunit 2 (COXII) in the European population of *A. crassus* was the most persistent finding. This downregulation was confirmed by the low expression of the same contig in the European libraries compared to higher expression in all three libraries from Taiwanese worms in pyrosequencing. Cytochrome C oxidase subunits 1-3 are are essential components of respiratory chain complex IV, the cytochrome c oxidase. They are encoded in the mitochondrial genome and coordinate catalytic heme and copper cofactors (228).

In fact, not only enrichment analysis highlighted oxidoreductases, but expression values of COXII clustered with other enzymes related to the state of energy metabolism: two lecitin:cholesterol acyltransferase transcripts are putative recently duplicated genes. They showed slightly divergent protein sequences but hit the same orthologs in *C. elegans* and *B. malayi*. They also shared very similar expression profiles. Expression of different cholesterol acyltransferases has been shown to vary in response to the presence of heme and anaerobiosis in yeast (229). 3-hydroxyacyl-CoA dehydrogenase (involved fatty-acid β-oxidation (230)), malate/L-lactate dehydrogenase (from the anaerobic glycolytic pathway or the Krebs-cycle (231)) and aspartyl proteases (involved in the digestion of host haemoglobin in helminths (190)) completed this particular cluster.

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

metabolism would be perturbing the pleiotropic effects of the persistent core-change.

Such a scenario also provokes speculation about the adaptive value of such a change in a core metabolic process: aerobic respiration is a potential source for oxidative stress providing a steady source of reactive oxygen species (ROS) as electrons are leaking from the respiratory chain as superoxide anions. It is well established that such ROS production is especially harmful to blood-feeding parasites, as free inorganic iron, as well as heme, have the potential to generate additional ROS (232). Anaerobic metabolism is thus thought to occur in many haematophagous parasites as a counter-measure against oxidative stress from haemoglobin catabolism (233). It could thus be hypothesised that the bigger size and the larger amount of eel-blood ingested leading to a higher rate of haemoglobin digestion provided the selective pressure to reduce aerobic respiration. Additionally helminths can simply get too large to maintain oxygen diffusion to mitochondriae in the absence of a cardiovascular system. As yet proton-pumping electron transport constitutes the most profitable energy-providing process, the mitochondriae of facultatively anaerobic helminths produce a proton gradient for the use of ATPase with the help of terminal electron acceptors other than O_2 (234). Such an alternate electron sink is fumarate used in many helminths in a process called malat dismutation (235).

An interesting implication is that such metabolic differences could potentially be visible ultrastructurally. Indeed in my own diploma thesis (236) I identified two different kinds of mitochondriae, one with standard christae-like morphology, the other with unusual sacculus-like morphology in *A. crassus*. Additionally I observed less electron-dense inclusions (probably lipid reserves) in bigger worms and more glycogen granulae. The fact that such lipids are less usable under anaerobic conditions led me to the hypothesis that bigger worms are using less aerobic processes. Reanalysing this data and probably obtaining new data with additional histochemical staining methods could be a way to put gene-expression into a physiological perspective. Furthermore, a biochemical examination of isolated mitochondriae could highlight changes in the mitochondrial respiratory chain under *in vitro* conditions (237). Such direct measurements of COX enzyme activity (using well established assays (238)) would be desirable to establish even the validity of the first logical step in these adaptive speculations that underexpression of COXII is leading to decreased enzyme activity. It would be counterintuitive to expect higher enzyme activity when COXII mRNA levels are low, but, for example, in *Schistosoma mansoni* COXI over-expression in praziquantel-resistant strains is leading rather to decreased enzyme activity (239).

The sensitivity to perturbation of mitochondrial genes for respiratory chain com-

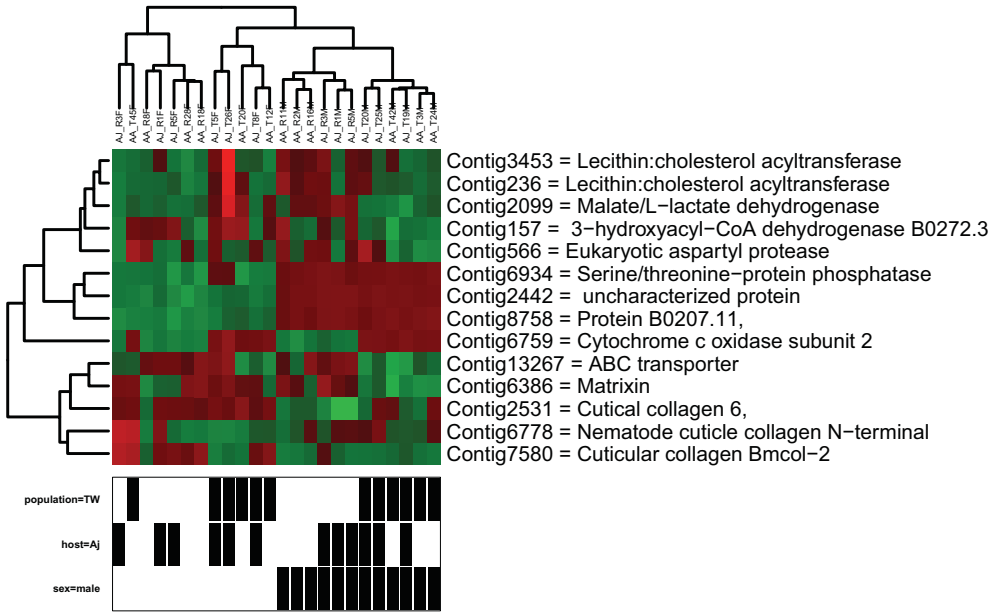

Figure 7.2: Clustering of expression values for OC contigs DE between populations - A heatmap of variance/mean stabilised expression values. Deprograms are based on hierarchical clustering. Green indicates expression below the mean, red above the mean. Experimental conditions are indicated by black bars for groups of samples (columns) below the plot. Expression levels for libraries are clustering mainly according to the sex of worms. However, in both male and female worms subordinate clusters are following a worm-population and to a lesser extend (and mainly in males) a host-species pattern. Below contig-names uniprot names are given for ortholog genes in *B. malayi*. Genes are clustering according to annotation-profiles: the top cluster represents genes important in energy metabolism. They cluster with COXII, which shows clear overexpression in - without any exception - all libraries from Taiwanese worms.

plexes in nematode parasites is underlined by their up-regulation after depletion of Wolbachia from filarial nematodes (240, 241). Wolbachia are obligate endosymbiont bacteria of some clade III nematodes, they are supplying heme to non-haematophagous parasites in the absence of an intrinsic pathway for heme synthesis (109) (which is absent also in free living *C. elegans* (242)). While my sequence analysis suggests the absence of wolbachial symbionts in *A. crassus*, such studies support a central role of host or endosymbiont derived heme for respiratory processes and suggest a propensity for evolutionary change in related processes (in Filaria even acquisition of an endosymbiont).

Assuming a genetically fixed lower expression of COXII in European *A. crassus* as a driver for other metabolic differences does not imply a simple regulation of the expression itself, or a genetically simple change underlying the changed expression phenotype. Regulation of the mitochondrially encoded genes has been extensively integrated into the regulatory network of eukaryotic cells and is controlled by and interacting with nuclear transcription factors (243).

Intriguingly overexpression of respiratory chain enzymes was limited to cytochrome c oxidase transcripts (COXII and to lesser extent also COXI and COXIII). Mitochondrial transcription produces multiple polycistronic unmatured transcripts, which are cleaved and modified in their expression post-transcriptionally. Cleavage occurs at t-RNA sequences interspersed between protein coding genes and can be imperfect to leave some transcripts polycistronic in a matured state. Nevertheless, due to posttranscriptional modification individual transcripts can be expressed uncoordinated, even when expressed on the same unmatured polycistronic transcript (244). The addition of poly-A tails, for example, is vital for stability of mature transcripts in metazoans. The mitochondrial genome contains only very little untranscribed sequence, is polyploid (once homoplasmic, essentially maternally inherited like haploid) and transmitted completely linked, with very scarce recombination events (228).

Cis-regulatory change in a control region would thus be very easily detectable in my transcriptome data. Even if the sequence variation leading to the observed expression phenotypes would locate to the untranscribed hypervariable mitochondrial control region (in D-Loop associated promoters), selection on such a variant would render the whole mitochondrial genome inadequate for phylogenetic analysis, as a variant sweeping to fixation would have removed polymorphism from the complete mitochondrial genome due to the prefect linkage (245). If a sweep would be presently ongoing, high levels of heteroplasmy would be found in single individuals (**?**). Such a pattern has not been found in populations of *A. crassus* in Europe when COXI was used as a

marker (10, 16) (see also figure 1.5) and is also not visible from preliminary analysis of polymorphism in mitochondrial genes in my RNA-seq data.

Functional constraints are also expected regarding the mechanism by which the expression of COXII could evolve. Most infective L3 larvae of parasitic nematodes rely on aerobic respiration (246). Dixenous parasites like *A. crassus* migrate through tissues of definitive hosts, where oxygen is readily available, after leaving the haeomocoel of the intermediate host. Enzyme subunits building a functioning aerobic respiratory chain are thus likely to be expressed at earlier lifecycle stages of *A. crassus* and elevated anaerobiosis is expected to be restricted to the adult stages.

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

RNAi screens in *C. elegans* for increased lifespan focus on genes leading to lower oxygen consumption and altered mitochondrial morphology and function (247). Such candidate genes will provide an additional link back to functional considerations once screening for genomic regions with signature of selection will highlight candidate loci.

7.3.3.2 Collagens

A second group of genes differentially expressed in populations of *A. crassus* emerged from both cluster and enrichment analyses. Two transcripts in this cluster were significant for interaction effects between host-species and parasite-population, they were annotated as collagens. For both genes this meant an "adjusted" (to avoid the suggestive "adapted") expression difference leading to a lower expression in sympatric host-species/parasite-population pairs. Cuticle collagens are a large multigene family (Interpro lists 164 entries for "Nematode cuticle collagen, N-terminal" for *C. elegans* and 51 for *B. malayi*), containing extensive repeat regions: roughly 50% Gly-X-Y residues, often Gly-Pro-Hpy. In the genome of *B. malayi* 82 genes encoding collagen repeats have been found (109). It was thus very important to have orthologous confirmation for these two contigs, as misassembly could have easily lead false positives here.

The two collagens were clustered with a third contig sharing a collagen-annotation (failing to be significant for the interaction term probably because of low overall expression) and a contig annotated as "Matrixin" (a metallo-proteinase assumed to be involved in remodelling of the extracellular matrix (248)) and a ABC-transporter family

protein.

Functional speculations are more difficult for collagen than for the respiratory chain enzymes. The cuticle constitutes an exoskeleton and a barrier between the worm and its host-environment. Synthesis of most collagens is believed to occur at negligible levels in adult male worms and is rather constrained to discrete temporal periods in larval development, the moults (212). The differential expression could thus be due to changes in larval development or due to alternations in the low-level, steady renewal of the adult cuticle and remodelling of the extracellular matrix of hypodermis cells. Some considerations would favour of the second explanation: in *C. elegans* genes expressed after reproductive maturity evolve faster than genes expressed earlier in development (214). This suggests a model of elevated pleiotropic effects in genes expressed at earlier stages of development and hence more conserved expression patterns in larval stages. Independent of these considerations, both the primary assembly and the constant remodelling of the cuticle involve complex post-translational processes hardly accessible at the transcriptomic level: a zipper-like nucleation/growth mechanism leads to the folding of a triple helix of and heterotrimers and homotrimers (246). If and how differential expression of two particular collagens interferes with this process requests further research. As for the metabolic differences, differential expression patterns could be reflected in morphology. One approach would be to measure thickness and density of the cuticle of worms from coinoculation experiments.

7.4 Outlook

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

least in low frequency. Thus, hundreds of individual worms from Taiwanese populations will be tested as new funding becomes available. Appropriate *A. crassus* samples stored in RNA-later are readily available from broad sampling for the present transcriptome projects from populations of worms in both wild and cultured *An. japonica*.

An assembly of the mitochondrial genome of *A. crassus* from preliminary genome-sequencing data (discussed below) and the identification of the poly-cistronic unmatured and, if present, matured transcripts (similar to (244)), will further inform and validate the analysis of the expression of mitochondrial genes. Additionally, disentangling assembly artefacts complicating mapping from real nuclear or even mitochondrial (**?**) pseudogenes of mitochondrial genes will help increasing the power of expression analysis and furthermore permit the analysis of interaction of such pseudogenes with the expression of functional genes.

Multiple starting points also exist for further functional examination of metabolic change, as mentioned throughout the text. However, the search for ultimate causes for evolutionary change *sensu* (249) will potentially be even more rewarding.

I will expand the RNA-seq analysis presented here to study allele-specific expression and the association between gene expression and sequence variants. This kind of quantitative expression trait locus (eQTL) analysis is possible as both sequence and expression information are available from the present RNA-seq data. Both simple cis-acting variation in promoter or enhancer regions, as well as trans-acting variation can theoretically be detected (250). To detect trans-acting variants, however, might be impossible with the (for population studies) relative low number of sequenced individuals, as it relies on statistical associations requiring broad sampling. Yet, cis-acting variation, more readily detectable as allele-specific variation, is unlikely to explain variations in mitochondrial gene expressions for the reasons discussed above.

Therefore, large scale meta-population wide sampling must not be limited to an evaluation of the divergent gene-expression phenotypes, but has to further elucidate the population genetic relationships between Taiwanese and European worms. A future research program will thus need to employ population-scale sampling of genotype data, densely spread across the genome. Genotyping of many European *A. crassus* from different populations and comparison with many individual genomes from different Asian populations will enable tests for selection: based on the fact that around selected variants nucleotide diversity is reduced by hitchhiking of neutral variation in so called selective sweeps (251), a punctual increase of population differentiation measured by the fixation index F_{st} (252) in regions linked to selected variants can be measured. Other well established population genetic measurements include Tajima's D, a measure

based on the allele frequency spectrum (253). When these methods are applied on a genome wide scale the neutral null-expectation to separate a loss in variability based on selection from neutral loss due to demography is given by the diversity across all regions of the genome. A microsatellite study (10) as well as my own evaluations (based on pyrosequencing see 5.10) and RNA-seq (data not shown) indicate only a moderate genetic bottleneck caused by the introduction of *A. crassus* to Europe and thus the necessary neutral diversity as a background for these tests will be present.

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

One methods enabling such population wide genotyping emerging from NGS technology is the sequencing of restriction-site associated DNA (RAD) markers. Preparation of RAD libraries involves digestion of genomic DNA with a restriction enzyme. Individually tagged adaptors can then be ligated to the fragments and individual samples can be pooled. The choice of restriction enzyme is important to optimise the number of restriction sites (depth of sampling the genome) relative to the number of individual samples being investigated (254). In the case of *A. crassus* this optimisation also concerns the minimisation of restriction sites in host-genome, as present in unavoidable contamination.

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

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

It has been possible to isolate roughly 1μg of genomic DNA from a big individual worm. Only ca. 20% of the DNA were derived from the genome of the eel-host (see figure 7.3). As only 300 ng of DNA material (with low amounts of contamination with

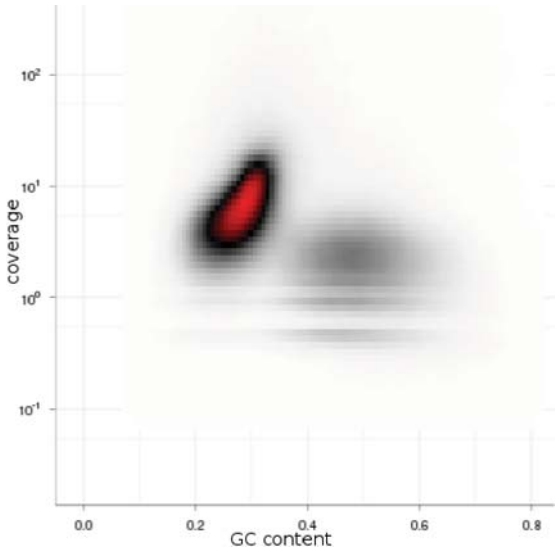

Figure 7.3: GC-content and coverage for a preliminary genome assembly - A preliminary assembly of roughly 10 Gb sequence data in over 110 million reads. The analysis of GC-content and coverage identifies host-contamination at higher GC, but lower coverage. Coverage and GC-content separate two distinct data-sources: a lower GC/higher coverage nematode subset and a higher GC/lower coverage eel subset (confirmed by `Blast`). For this sequencing library only 10-20% of the reads are lost to eel-host derived off-target data. The preliminary assembly was provided by Sujai Kumar from Mark Blaxter's lab.

host-blood) are needed for RAD-sequencing, this can be achieved in most big specimen of *A. crassus*.

For both reference genome assembly and annotation and for the future genome-scans I will continue to collaborate with Mark Blaxter's laboratory at the University of Edinburgh. This group is actively developing methods especially for RAD-sequencing and applying them to questions in evolutionary model-species (255).

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

A prime example for a research program on the evolution of ecologically important traits is provided by the Stickleback *Gasterosteus aculeatus*: QTL-mapping has been performed to fine-map the loss of lateral plates in freshwater populations (254) and parallel adaptation has been investigated using population genomics (256). Both approaches used RAD-sequencing. The sophistication and depth of insight available in such an evolutionary model species is underlined by research on adaptive reduction of pelvic structures, an evolutionary trajectory shown to be favoured by the localisation of the underlying change in an instable region of the genome (257).

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

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

8

Materials & methods

8.1 Sampling of worms from wild eels for Sanger- and pyrosequencing

Cultured eels were acquired from an aquaculture directly adjacent to Kaoping river (22.6418N; 120.4440E) 15km stream upwards from its 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 pyrosequencing

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 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

Initial denaturation	94 °C	5min	
Denaturation	94 °C	30s	
Annealing	54 °C	45s	35 cycles
Elongation	72 °C	2min	
Final 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 *Anguillicola crassus*, "Aj" for *An. japonica*. In *Anguillicola* sequences information on the sequencing primer (forward or reverse PCR primer; *An. 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` (116)) was used, briefly it performs quality trimming using `phred` (258) and trimming of vector sequences using `cross-match` (259). 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 divided 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 nempep (125) (version 4) and a fish protein database. Sequences producing better bit scores against 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 Pyrosequencing

8.5.1 cDNA preparation and sequencing

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

8.5.2 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` (259) (with parameters `-minscore 20 -minmatch 10`). `Seqclean` (177) 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, an 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 (260) a publicly available transcriptome database for the European eel; (b) a database of ribosomal RNA (rRNA) sequences from eel species derived from my Roche 454 data and EMBL-Bank; and (c) a database of rRNA sequences identified in my *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` (137) (with parameters `-cdna -urt`), the other using `Mira v3.2.1` (175) (with parameters `-job=denovo,est,accurate,454`). The resulting two assemblies were combined into one using `Cap3` (176) at default settings and contigs were labeled by whether they derived from both assemblies or one assembly only.

8.5.3 Evaluation 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.

`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 singletons 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 summarising the number of database-sequences with any coverage, the number with coverage over 80% of their sequence-length and the overall proportion of bases covered.

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

To estimate contig-coverage I converted `sam`-output generated with `ssaha2`(153) via a sorted `bam`-file to `pileup`-format using `samtools` (182). For a second evaluation I excluded best-hits mapping to multiple contigs before converting the `sam`-file to obtain unique coverage.

8.5.4 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 (123, 125). For each contig, the highest-scoring match was recorded as long as it spanned more than 50% of the contig. I 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.

8.5.5 Protein prediction and annotation

Protein translations were predicted from the contigs using `prot4EST` (version 3.0b) (178). Proteins were predicted either by joining single high scoring segment pairs (HSPs) from a `Blast` search of uniref100 (261), or by `ESTscan` (262), 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, I 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) (179), using the annotated sequences available in uniref100 (261). 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` (181) 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` (180) (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.

I recorded the presence of a lethal RNAi-phenotype in the *C. elegans* ortholog of each TUG using the biomart-interface (263) to wormbase v. 220 through the R-package `biomaRt` (264).

8.5.6 Single nucleotide polymorphism analysis

I mapped the raw reads against the the complete set of contigs, replacing imputed sequences for originals where relevant, using `ssaha2` (153) (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` (182), discarding reads mapping to multiple regions. `VarScan` (158) (`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 I used the R-package `GOstats` (265).

Using `Samtools` (182) (`mpileup -u`) and `Vcftools` (159) (`view -gcv`) I 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) I used the R package `Rhh` (186) to calculate internal relatedness (183), homozygosity by loci (184) and standardised heterozygosity (185) from these data.

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

8.5.7 Gene-expression analysis

Read-counts were obtained from the `bam`-files generated also for genotyping using the R-package `Rsamtoools` (266). 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. I used the R-package `DESeq` (164) (version 1.6.1) to assess statistical significance of differences in counts according to groups of libraries.

Additionally I collapsed TUGs by their orthologous assignment in *C. elegans* and *B. malayi*. I used the sums of counts for these orthologous-groups to asses the influence of mapping to my 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.

8.5.8 Enrichment analysis

We used Mann-Whitney u-tests to test the influence of factors on dn/ds values, when multiple contrasts between groups (factors) were investigated we used Nemenyi-Damico-Wolfe-Dunn tests. For overrepresentation of one group (factor) in another group (factor) we used Fisher's exact test.

Prior to analysis of GO-term over-representation (based on dn/ds or expression values) we used the R-package `annotationDbi` (267) to obtain a full list of associations (also with higher-level terms) from `annot8r`-annotations. We then used the R-package `topGO` (268) to traverse the annotation-graph and analyse each node in the annotation for over-representation of the associated term in the focal gene-set compared to an appropriate universal gene-set (all contigs with dn/ds values or all contigs analysed for gene-expression) with the "classic" method and Fisher's exact test.

8.6 Transcriptomic divergence in a common garden experiment

8.6.1 Experimental infection of eels

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

After an acclimatisation period of 4 weeks (*An. anguilla*) or when they reached a size of $>$ 35cm (*An. japonica*) eels were infected using a stomach tube as described in (269). During the infection period water temperature was held constant at 20°C. Eels were kept in 160-litre tanks in groups of 5-10 individuals and continuously provided with fresh, oxygenated water and once every two days with commercial fish pellets (Dan-Ex 2848, Dana Feed A/S Ltd, Horsens, Denmark) *at libitum*.

L2 larvae used for the infection were collected from the swimbladders of wild yellow and silver eels from the River Rhine near Karlsruhe and from Lake Müggelsee near Berlin in Germany. Taiwanese larvae were obtained from eels from an aquaculture adjacent to Kao Ping River in south Taiwan and from a second aquaculture in Yunlin

county, approximately 150 km further north on the west coast of Taiwan. They were stored at 4°C for no longer than 2 weeks before copepods were infected. Mixed species samples of uninfected copepods were collected from a small pond near Karlsruhe, known to be free of eels (and *Anguillicola*). They were infected individually in wells of microtiter plates at an intensity of roughly 10 L2-larvae per copepod. One week after infection they were placed in bigger tanks. Twice a week yeast was provided as food and at 21 dpi infective L3 were harvested with using a tissue potter as described by (270). 50 L3 for infection of individual eel were suspended in 100 μl RPMI-1640 medium (Quiagen, Hilden, Germany) and eels were infected as described above.

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

8.6.2 RNA extraction and preparation of sequencing libraries

RNA was extracted from 12 individual female worms and for 12 pools of male worms using the RNeasy-kit (Quiagen, Hilden, Germany) (see table 6.2).

The paired-end TruSeqTM RNA sample preparation kit (Illumina) was followed to build sequencing libraries with insert sizes of roughly 270 bp for paired-end sequencing from cDNA libraries: briefly, poly-T oligo-attached magnetic beads were used for purification of mRNA and to simultaneously fragment the RNA. The RNA was then primed with random hexamer primers for cDNA synthesis and reverse transcribed into first strand cDNA using reverse transcriptase. The cDNA was cleaned from the second strand reaction, overhangs were repaired to form blunt ends, a single "A"-nucleotide was added at the 3' end and paired end sequencing adapters ware ligated with a complementary "T"-overhang. In this step multiple differently indexed paired-end adapters were used to enable multiplexing of the 24 different sequencing libraries in 3 pools of 8 samples each. These three pools all contained one random replicate each for each treatment combination ensuring complete statistical independence of replicates. Molecules having adapter sequences were enriched in the mix using PCR and the libraries were controled for quality and quantity on the BioAnalyzer (Agilent). Clusters were generated by bridge amplification. The resulting clusters were sequenced on the Genome Analyzer IIX in combination with the paired-end module. The first read was sequenced using using the first primer Rd1 SP. The original template strand was then used to regenerate the complementary strand, the original strand was removed and complementary strand acted as a template for the second read, sequenced primed by the second sequencing

primer Rd2 SP.

8.6.3 Mapping and normalisation of read-counts

All sequencing reads were mapped to the fullest 454 assembly (as defined in 4.8; I were including TUGs inferred as host or xenobiont origin as filter) using `BWA` (154) (version 0.5.9-r16; `BWA aln` and `BWA sampe` with default options) and processed with `samtools` (182) (version 0.1.18; `samtools view -uS -q 1`) to only allow 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 spurious read counts to contigs with less than 32 mapping reads in total (see however 8.6.5 for how these counts were used in further tests of reference fragmentation).

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

8.6.4 Statistical analysis with generalised linear models (GLMs)

The R-package `edgeR` (version 2.4.1) (165) was used to build negative binomial generalised linear models, as these specialised GLMs outperformed GLMs in `DESeq` in speed and reliability of convergence. Modeled were based on a negative binomial distribution and the dispersion parameter for each transcript was approximated with a trend depending on the overall level of expression. In the maximal fitted model expression was regressed on worm-sex, host-species and parasite population, including all their interactions. The full model thus contained terms $S_i + H_j + P_k + (SH)_{ij} + (SP)_{ik} + (HP)_{jk} + (SHP)_{ijk} + \varepsilon$, where ε is the residual variance, S_i is the effect of the ith sex (male or female), H_j is the effect of the ith host species (*An. anguilla* or *An. japonica*), P_k is the effect of the kth population (European or Asian), $(SH)_{ij}$ is the sex-by-species interaction and similarly for the other interactions.

The hierarchical nature of generalised linear models was respected considering (re-

moving) all interaction effects of a main-term (e.g. $(SP)_{ik}$, $(SH)_{ij}$ and $(SHP)_{ijk}$) when analysing models for the significance of that term (e.g. S_i). Resulting p-values were corrected for multiple testing using the method of Benjamini and Hochberg (271) and differential expression was inferred at a false discovery rate (FDR) of 5% (adjusted p-value of 0.05).

8.6.5 Count-collapsing for orthologs from two model-species

In order to test the influence of deficiencies (i.e. fragmentation) of the assembly on mapping and read-counts I summed read counts over orthologous sequence in *C. elegans* and *B. malayi*. For this purpose I used all reference contigs (also lowCA-derived contigs to allow inference of fragmentary 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 inference from orthologous counts merging the two orthologous evaluations and the contig evaluation. Differential expression was accepted at a FDR of 5% for the contig evaluation and 10% for both of the two orthologous evaluations.

8.6.6 Multivariate confirmation of linear models

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

8.6.7 GO-term enrichment analysis

Enrichment analysis was performed as described above for pyrosequencing data (see 8.5.8).

8.6.8 Clustering analysis

The R-package `HeatmapPlus` was used on variance stabilised expression values to visualise hierarchical clusters similar to the method of (272). The results were displayed along with annotations stored in a Bioconductor eSet-class object.

8.7 General coding methods

The bulk of analysis (unless otherwise cited) presented in this paper was carried out in `R` (273) using custom scripts. I used a method provided in the R-packages `Sweave` (274) and `Weaver` (275) for "reproducible research" combining `R` and LaTeXcode in a single file. The complete reproducible compilations were only carried out for sub-chapters of this document, the thesis-document was then compiled from plain LaTeXsub-documents. Nevertheless all intermediate data files needed to compile sub-document of the thesis from data-sources are provided upon request. For general visualisation I used the R-packages `ggplot2` (276) and `VennDiagram` (277).

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

[6] A. Loukili and D. 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] S. Wielgoss, H. Taraschewski, A. Meyer, and T. 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, 3, 16, 115, 119, 127, 130

[11] M. 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] P. Sasal, H. Taraschewski, P. Valade, H. Grondin, S. Wielgoss, and F. 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] D. R. Laetsch, E. G. Heitlinger, H. Taraschewski, S. A. Nadler, and M. Blaxter. **The phylogenetics of Anguillicolidae (Nematoda: Anguillicolidea), swimbladder parasites of eels**. *BMC Evolutionary Biology*, under review. 2, 3, 8, 9, 10, 11, 12, 127

[17] A. M. Barse, S. A. McGuire, M. A. Vinores, L. E. Eierman, and J. 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

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

[19] L. T. Fries, D. J. Williams, and S. 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

[20] F. Moravec, K. Nagasawa, and M. 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

[21] 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

[22] 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, 114

[23] 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

[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] 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

[26] 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

[27] M. Münderle, G. 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

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

[30] L. 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

[31] 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

[32] 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

[33] 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

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

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

[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, 120

[37] 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, 117

[38] 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

[39] J. Würtz 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*, **39**(2):121–34, 2000. 5

[40] 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

[41] A.P. Palstra, D.F.M. Heppener, V.J.T. van Ginneken, C. Székely, 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

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

[43] 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

[44] 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, 114

[45] E. Heitlinger, D. Laetsch, U. Weclawski, Y. S. Han, and H. Taraschewski. **Massive encapsulation of larval Anguillicoloides crassus in the intestinal wall of Japanese eels**. *Parasites and Vectors*, **2**(1):48, 2009. 5, 6, 112

[46] 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. 7

[47] 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

[48] 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

[49] 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

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

[53] 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

[54] 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

[55] 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

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

[59] 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

[60] D. J. Conway and S. D. Polley. **Measuring immune selection**. *Parasitology(London. Print)*, **125**:3–16, 2002. 7

[61] 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

[62] 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. 8

[63] B. Star, A. J. Nederbragt, S. Jentoft, U. Grimholt, M. Malmstrøm, 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. Gjøen, 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. 8

[64] J. Hikima, T. S. Jung, and T. Aoki. **Immunoglobulin genes and their transcriptional control in teleosts**. *Dev. Comp. Immunol.*, **35**:924–936, Sep 2011. 8

[65] 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. 8

[66] 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

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

[68] 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

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

[70] Y. Minegishi, J. Aoyama, J. G. Inoue, M. Miya, M. Nishida, and K. 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. 10

[71] M. Blaxter, P. De Ley, J.R. Garey, L. X. Liu, P. Scheldeman, A. Vierstraete, J.R. Vanfleteren, L.Y. Mackey, M Dorris, L.M. Frisse, J.T. Vida, and W.K. Thomas. **A molecular evolutionary framework for the phylum Nematoda**. *Nature*, **392**(6671):71–75, March 1998. 10

[72] S. A. Nadler, R. A. Carreno, H. Meja-Madrid, J. Ullberg, C. 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. 10

[73] M. Wijová, F. Moravec, A. Horák, and J. 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. 10

REFERENCES

[74] G. Bonniér. **Recherches expérimentales sur l´adaptation des plants au climat alpin**. *Ann. Scie. Nat. (Bot.)*, **20**:217–358, 1895. 13

[75] A. Kerner. **The natural history of plants, their forms, growth, reproduction, and distribution**. *Translated by F. W. Oliver.*, 1895. 13

[76] O Kaltz and J. A. Shykoff. **Local adaptation in host-parasite systems**. *Heredity*, pages 361–370, May 1998. 13

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

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

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

[80] 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

[81] F. Crick. **The biological replication of macromolecules**. In *Symp. Soc. Exp. Biol*, **12**, pages 138–163, 1958. 17

[82] F. Crick. **Central dogma of molecular biology**. *Nature*, **226**:1198–1199, Jun 1970. 17

[83] M. Lynch. **The lower bound to the evolution of mutation rates**. *Genome Biol Evol*, **3**:1107–1118, 2011. 17

[84] 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

[85] 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

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

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

[88] 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

[89] F. Crick. **The origin of the genetic code**. *J. Mol. Biol.*, **38**:367–379, Dec 1968. 18

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

[91] Z. Wang, M. Gerstein, and M. Snyder. **RNA-Seq: a revolutionary tool for transcriptomics**. *Nat. Rev. Genet.*, **10**:57–63, Jan 2009. 19, 28

[92] J. Armengaud. **Proteogenomics and systems biology: quest for the ultimate missing parts**. *Expert Rev Proteomics*, **7**:65–77, Feb 2010. 19

[93] 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. 19, 22

[94] 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. 19

[95] 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

[96] H. Swerdlow and R. Gesteland. **Capillary gel electrophoresis for rapid, high resolution DNA sequencing**. *Nucleic Acids Res.*, **18**:1415–1419, Mar 1990. 20

[97] 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. 20

[98] 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. 20

[99] The C. elegans Sequencing Consortium. **Genome sequence of the nematode C. elegans: a platform for investigating biology**. *Science*, **282**:2012–2018, Dec 1998. 20

[100] 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. 20

[101] 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. 20

[102] 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. 20

[103] 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. **Complementary DNA sequencing: expressed sequence tags and human genome project**. *Science*, **252**:1651–1656, Jun 1991. 20

[104] 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. 20

[105] M. Blaxter. **Caenorhabditis elegans Is a Nematode**. *Science*, **282**(5396):2041–2046, dec 1998. 21

[106] 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. 21

[107] L. D. Stein, Z. Bao, D. Blasiar, T. Blumenthal, M. R. Brent, N. Chen, A. Chinwalla, L. Clarke, C. Clee, A. Coghlan, A. Coulson, P. D'Eustachio, D. H. Fitch, L. A. Fulton, R. E. Fulton, S. Griffiths-Jones, T. W. Harris, L. W. Hillier, R. Kamath, P. E. Kuwabara, E. R. Mardis, M. A. Marra, T. L. Miner, P. Minx, J. C. Mullikin, R. W. Plumb, J. Rogers, J. E. Schein, M. Sohrmann, J. Spieth, J. E. Stajich, C. Wei, D. Willey, R. K. Wilson, R. Durbin, and R. H. Waterston. **The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics**. *PLoS Biol.*, **1**:E45, Nov 2003. http://www.ncbi.nlm.nih.gov/pubmed/14624247. 21

[108] 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. 21

[109] E. Ghedin, S. Wang, D. Spiro, E. Caler, Q. Zhao, J. Crabtree, J. E. Allen, A. L. Delcher, D. B. Guiliano, D. Miranda-Saavedra, S. V. Angiuoli, T. Creasy, P. Amedeo, B. Haas, N. M. El-Sayed, J. R. Wortman, T. Feldblyum, L. Tallon, M. Schatz, M. Shumway, H. Koo, S. L. Salzberg, S. Schobel, M. Pertea, M. Pop, O. White, G. J. Barton, C. K. Carlow, M. J. Crawford, J. Daub, M. W. Dimmic, C. F. Estes, J. M. Foster, M. Ganatra, W. F. Gregory, N. M. Johnson, J. Jin, R. Komuniecki, I. Korf, S. Kumar, S. Laney, B. W. Li, W. Li, T. H. Lindblom, S. Lustigman, D. Ma, C. V. Maina, D. M. Martin, J. P. McCarter, L. McReynolds, M. Mitreva, T. B. Nutman, J. Parkinson, J. M. Peregrin-Alvarez, C. Poole, Q. Ren, L. Saunders, A. E. Sluder, K. Smith, M. Stanke, T. R. Unnasch, J. Ware, A. D. Wei, G. Weil, D. J. Williams, Y. Zhang, S. A. Williams, C. Fraser-Liggett, B. Slatko, M. L. Blaxter, and A. L. Scott. **Draft genome of the filarial nematode parasite Brugia malayi**. *Science*, **317**:1756–1760, Sep 2007. http://www.ncbi.nlm.nih.gov/pubmed/17885136. 21, 126, 127

[110] 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. 21

[111] 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. 21

[112] 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. 21

[113] 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. 21

[114] 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. 21

[115] S. Kumar, P. H. Schiffer, and M. Blaxter. **959 Nematode Genomes: a semantic wiki for coordinating sequencing projects**. *Nucleic Acids Res*, Nov 2011. 21

[116] J. Parkinson, A. Anthony, J. Wasmuth, R. Schmid, A. Hedley, and M. Blaxter. **PartiGene–constructing partial genomes**. *Bioinformatics*, **20**(9):1398–1404, June 2004. 21, 113, 135

[117] 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. 21, 22

[118] R. M. Maizels, A. Balic, N. Gomez-Escobar, M. Nair, M. D. Taylor, and J. E. Allen. **Helminth parasites; masters of regulation**. *Immunological Reviews*, **201**(1):89–116, 2004. 22, 121

[119] N. Gomez-Escobar, W. F. Gregory, C. Britton, L. Murray, C. Corton, N. Hall, J. Daub, M. Blaxter, and R. 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. 22

[120] 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. 22

[121] Y. Harcus, J. Parkinson, C. Fernandez, J. Daub, M. Selkirk, M. Blaxter, and R. 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. 22, 113

[122] S. H. Nagaraj, R. B. Gasser, and S. 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. 22

[123] J. Parkinson, C. Whitton, R. Schmid, M. Thomson, and M. Blaxter. % bf NEMBASE: a resource for parasitic nematode ESTs. *Nucl. Acids Res.*, **32**(suppl_1):D427–430, 2004. 22, 64, 138

[124] J. Wasmuth, R. Schmid, A. Hedley, and M. Blaxter. **On the Extent and Origins of Genic Novelty in the Phylum Nematoda**. *PLoS Neglected Tropical Diseases*, **2**(7):e258, July 2008. 22, 113

[125] B. Elsworth, J. Wasmuth, and M. Blaxter. **NEMBASE4: The nematode transcriptome resource**. *Int. J. Parasitol.*, **41**:881–894, Jul 2011. 22, 62, 64, 136, 138

[126] 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. 22, 114, 115, 116

[127] S. Kumar and M. L. Blaxter. **Comparing de novo assemblers for 454 transcriptome data**. *BMC Genomics*, **11**:571, Oct 2010. 22, 26, 43, 60, 137

[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. 22

[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. 22

[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. 22

[131] M. L. Metzker. **Sequencing technologies - the next generation**. *Nat. Rev. Genet.*, **11**:31–46, Jan 2010. 24

[132] J. M. Rothberg and J. H. Leamon. **The development and impact of 454 sequencing**. *Nat. Biotechnol.*, **26**:1117–1124, Oct 2008. 25

[133] 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. 24

[134] P. Nyren. **The history of pyrosequencing**. *Methods Mol. Biol.*, **373**:1–14, 2007. 24

[135] S. Balzer, K. Malde, and I. Jonassen. **Systematic exploration of error sources in pyrosequencing flowgram data**. *Bioinformatics*, **27**:i304–309, Jul 2011. 24, 68, 113

[136] 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. 24

[137] 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. 26, 43, 137

[138] 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. 26, 28

[139] 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. 28

[140] 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. 28

[141] J. H. Malone and B. Oliver. **Microarrays, deep sequencing and the true measure of the transcriptome**. *BMC Biol.*, **9**:34, 2011. 28

[142] P. A. 't Hoen, Y. Ariyurek, H. H. Thygesen, E. Vreugdenhil, R. H. Vossen, R. X. de Menezes, J. M. Boer, G. J. van Ommen, and J. T. den Dunnen. **Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms**. *Nucleic Acids Res.*, **36**:e141, Dec 2008. 28

[143] 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. 28

[144] V. E. Velculescu, L. Zhang, B. Vogelstein, and K. W. Kinzler. **Serial analysis of gene expression**. *Science*, **270**:484–487, Oct 1995. 28

[145] J. R. Miller, S. Koren, and G. Sutton. **Assembly algorithms for next-generation sequencing data**. *Genomics*, **95**:315–327, Jun 2010. 28

[146] 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]. 29

[147] R. Staden. **A strategy of DNA sequencing employing computer programs**. *Nucleic Acids Res.*, **6**:2601–2610, Jun 1979. 29

[148] T. R. Gingeras and R. J. Roberts. **Steps toward computer analysis of nucleotide sequences**. *Science*, **209**:1322–1328, Sep 1980. 29

[149] T. F. Smith and M. S. Waterman. **Identification of common molecular subsequences**. *J. Mol. Biol.*, **147**:195–197, Mar 1981. 29

[150] T. F. Smith, M. S. Waterman, and W. M. Fitch. **Comparative biosequence metrics**. *J. Mol. Evol.*, **18**:38–46, 1981. 29

[151] 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. 29

[152] W. J. Kent. **BLAT–the BLAST-like alignment tool**. *Genome Res.*, **12**:656–664, Apr 2002. 29

[153] 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. 29, 54, 138, 139

[154] H. Li and R. Durbin. **Fast and accurate long-read alignment with Burrows-Wheeler transform**. *Bioinformatics*, **26**:589–595, Mar 2010. 30, 94, 142

[155] 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. 30

[156] 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. 30

[157] 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. 30, 54, 112

[158] 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, 68, 139

[159] Danecek, P. and Auton,âĂŹA. and Abecasis, G. and Albers CA. and Banks, E. and DePristo, MA. and Handsaker RE. and Lunter G. and Marth GT. and Sherry ST. and McVean GT. and Durbin T. and the 1000 Genomes Project. **The variant call format and VCFtools**. *Bioinformatics*, **27**:2156–2158, Aug 2011. 30, 77, 139

[160] 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. 31

[161] 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. 31, 32, 120

[162] G. Smyth. **Limma: linear models for microarray data**. *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420, 2005. 31

[163] 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. 31

[164] S. Anders and W. Huber. **Differential expression analysis for sequence count data**. *Genome Biol.*, **11**:R106, 2010. 31, 115, 139

[165] 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. 31, 115, 142

[166] T. J. Hardcastle and K. A. Kelly. **baySeq: empirical Bayesian methods for identifying differential expression in sequence count data**. *BMC Bioinformatics*, **11**:422, 2010. 31

[167] 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. 31

[168] 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. 31

[169] R. Ekblom and J. Galindo. **Applications of next generation sequencing in molecular ecology of non-model organisms**. *Heredity (Edinb)*, **107**:1–15, Jul 2011. 31

[170] M. Dassanayake, J. S. Haas, H. J. Bohnert, and J. M. Cheeseman. **Shedding light on an extremophile lifestyle through transcriptomics**. *New Phytologist*, **183**(3):764–775, 2009. 32

[171] 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. 32

[172] 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. 32

[173] H. S. Rane, J. M. Smith, U. Bergthorsson, and V. Katju. **Gene conversion and DNA sequence polymorphism in the sex-determination gene fog-2 and its paralog ftr-1 in Caenorhabditis elegans**. *Mol. Biol. Evol.*, **27**:1561–1569, Jul 2010. 33

[174] 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. 33

[175] 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. 43, 60, 137

[176] X. Huang and A. Madan. **CAP3: A DNA sequence assembly program**. *Genome Res.*, **9**:868–877, Sep 1999. 43, 137

[177] 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. 61, 137

[178] J. Wasmuth and M. Blaxter. **prot4EST: Translating Expressed Sequence Tags from neglected genomes**. *BMC Bioinformatics*, **5**(1):187, 2004. 62, 138

[179] R. Schmid and Blaxter M. **annot8r: GO, EC and KEGG annotation of EST datasets**. *BMC Bioinformatics*, **9**:180, 2008. 64, 65, 66, 138

[180] E. M. Zdobnov and R. Apweiler. **InterProScan–an integration platform for the signature-recognition methods in InterPro**. *Bioinformatics*, **17**:847–848, Sep 2001. 64, 65, 112, 138

[181] 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. 64, 138

REFERENCES

[182] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. R. Abecasis, and R. Durbin. **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics*, **25**(16):2078–2079, 2009. 77, 138, 139, 142

[183] 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. 78, 139

[184] 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. 78, 139

[185] 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. 78, 139

[186] 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. 78, 139

[187] S. Audic and J. M. Claverie. **The significance of digital gene expression profiles**. *Genome Res.*, **7**:986–995, Oct 1997. 79, 115

[188] M. Fardilha, S.L.C. Esteves, L. Korrodi-Gregório, S. Pelech, O.A.B. da Cruz e Silva, and E. da Cruz e Silva. **Protein phosphatase 1 complexes modulate sperm motility and present novel targets for male infertility**. *Molecular human reproduction*, **17**(8):466–477, 2011. 102

[189] H. Smith. **Sperm motility and MSP**. *WormBook: The Online Review of C. Elegans Biology*, 2006. 102

[190] J. P. Dalton, P. J. Brindley, D. P. Knox, C. P. Brady, P. J. Hotez, S. Donnelly, S. M. O'Neill, G. Mulcahy, and A. Loukas. **Helminth vaccines: from mining genomic information for vaccine targets to systems used for protein expression**. *Int. J. Parasitol.*, **33**:621–640, May 2003. 104, 120, 123

[191] I. G. Wilson. **Inhibition and facilitation of nucleic acid amplification**. *Appl. Environ. Microbiol.*, **63**:3741–3751, Oct 1997. 111

[192] M. A. Valasek and J. J. Repa. **The power of real-time PCR**. *Adv Physiol Educ*, **29**:151–159, Sep 2005. 111

[193] 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. 111

[194] 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. 111

[195] 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. 112

[196] 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. 112

[197] 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. 112

[198] 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. 112, 114

[199] 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. 112

[200] 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. 112

[201] 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. 113

[202] A. Adey, H. Morrison, X. Asan, X. Xun, J. Kitzman, E. Turner, B. Stackhouse, A. MacKenzie, N. Caruccio, X. Zhang, and J. Shendure. **Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition**. *Genome Biol.*, **11**(12):R119, 2010. 113

[203] S. Kryazhimskiy and J. B. Plotkin. **The population genetics of dN/dS**. *PLoS Genet.*, **4**:e1000304, Dec 2008. 114

[204] 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. 114

[205] 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. 114

[206] 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. 114, 115, 121

[207] 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. 114

[208] 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. 114

[209] 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. 114

[210] 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. 114

[211] A. L. Scott. **Nematode sperm**. *Parasitol. Today (Regul. Ed.)*, **12**:425–430, Nov 1996. 115

[212] I. L. Johnstone. **Cuticle collagen genes. Expression in Caenorhabditis elegans**. *Trends Genet.*, **16**:21–27, Jan 2000. 115, 128

[213] B. Middleton. **The oxoacyl-coenzyme A thiolases of animal tissues**. *Biochem. J.*, **132**:717–730, Apr 1973. 115

[214] 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. 115, 119, 128

[215] W. G. Eberhard. **Evolutionary conflicts of interest: are female sexual decisions different?** *Am. Nat.*, **165 Suppl 5**:19–25, May 2005. 115, 121

[216] W. G. Hill, M. E. Goddard, and P. M. Visscher. **Data and theory point to mainly additive genetic variance for complex traits**. *PLoS Genet.*, **4**:e1000008, Feb 2008. 118

[217] 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. 118

[218] R. Nielsen. **Adaptionism-30 years after Gould and Lewontin**. *Evolution*, **63**:2487–2490, Oct 2009. 118

[219] M. F. Oleksiak, G. A. Churchill, and D. L. Crawford. **Variation in gene expression within and among natural populations**. *Nat. Genet.*, **32**:261–266, Oct 2002. 118

[220] J. A. Stamatoyannopoulos. **The genomics of gene expression**. *Genomics*, **84**:449–457, Sep 2004. 118

[221] R. B. Brem, G. Yvert, R. Clinton, and L. Kruglyak. **Genetic dissection of transcriptional regulation in budding yeast**. *Science*, **296**:752–755, Apr 2002. 118

[222] W. Jin, R. M. Riley, R. D. Wolfinger, K. P. White, G. Passador-Gurgel, and G. Gibson. **The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster**. *Nat. Genet.*, **29**:389–395, Dec 2001. 118, 119

[223] J. R. True and E. S. Haag. **Developmental system drift and flexibility in evolutionary trajectories**. *Evol. Dev.*, **3**:109–119, 2001. 118

[224] A. Whitehead and D. L. Crawford. **Neutral and adaptive variation in gene expression**. *Proc. Natl. Acad. Sci. U.S.A.*, **103**:5425–5430, Apr 2006. 118

[225] 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. 119

[226] 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. 121

[227] S. V. Nuzhdin, M. L. Wayne, K. L. Harmon, and L. M. McIntyre. **Common pattern of evolution of gene expression level and protein sequence in Drosophila**. *Mol. Biol. Evol.*, **21**:1308–1317, Jul 2004. 121

[228] B. Lemire. **Mitochondrial genetics**. *WormBook*, pages 1–10, 2005. 123, 126

[229] M. Valachovic, V. Klobucnikova, P. Griac, and I. Hapala. **Heme-regulated expression of two yeast acyl-CoA:sterol acyltransferases is involved in the specific response of sterol esterification to anaerobiosis**. *FEMS Microbiol. Lett.*, **206**:121–125, Jan 2002. 123

[230] S. Y. Yang and M. Elzinga. **Association of both enoyl coenzyme A hydratase and 3-hydroxyacyl coenzyme A epimerase with an active site in the amino-terminal domain of the multifunctional fatty acid oxidation protein from Escherichia coli**. *J. Biol. Chem.*, **268**:6588–6592, Mar 1993. 123

[231] G. Sturm, C. Hirschhäuser, and F. Zilliken. **Vergleichende Bestimmung von Enzymaktivitäten in Fasciola hepatica und Rinderleber**. *Hoppe-Seyler´s Zeitschrift für physiologische Chemie*, **350**(1):696–700, 1969. 123

[232] S. Q. Toh, A. Glanfield, G. N. Gobert, and M. K. Jones. **Heme and blood-feeding parasites: friends or foes?** *Parasit Vectors*, **3**:108, 2010. 124

[233] P. L. Oliveira and M. F. Oliveira. **Vampires, Pasteur and reactive oxygen species. Is the switch from aerobic to anaerobic metabolism a preventive antioxidant defence in blood-feeding parasites?** *FEBS Lett.*, **525**:3–6, Aug 2002. 124

[234] A. G. Tielens, C. Rotte, J. J. van Hellemond, and W. Martin. **Mitochondria as we don't know them**. *Trends Biochem. Sci.*, **27**:564–572, Nov 2002. 124

[235] A. G. Tielens. **Energy generation in parasitic helminths**. *Parasitol. Today (Regul. Ed.)*, **10**:346–352, Sep 1994. 124

[236] Emanuel Heitlinger. **Vergleichende licht- und elektronenmikroskopische Untersuchungen am Intestinaltrakt des invasiven Schwimmblasennematoden Anguillicola crassus aus verschiedenen Aalarten**, 2008. 124

[237] L. I. Grad, L. C. Sayles, and B. D. Lemire. **Isolation and functional analysis of mitochondria from the nematode Caenorhabditis elegans**. *Methods Mol. Biol.*, **372**:51–66, 2007. 124

[238] R. A. Capaldi, M. F. Marusich, and J. W. Taanman. **Mammalian cytochrome-c oxidase: characterization of enzyme and immunological detection of subunits in tissue extracts and whole cells**. *Meth. Enzymol.*, **260**:117–132, 1995. 124

[239] C. Pereira, P. G. Fallon, J. Cornette, A. Capron, M. J. Doenhoff, and R. J. Pierce. **Alterations in cytochrome-c oxidase expression between praziquantel-resistant and susceptible strains of Schistosoma mansoni**. *Parasitology*, **117 (Pt 1)**:63–73, Jul 1998. 124

[240] E. Ghedin, T. Hailemariam, J. V. DePasse, X. Zhang, Y. Oksov, T. R. Unnasch, and S. Lustigman. **Brugia malayi gene expression in response to the targeting of the Wolbachia endosymbiont by tetracycline treatment**. *PLoS Negl Trop Dis*, **3**:e525, 2009. 126

[241] U. Strubing, R. Lucius, A. Hoerauf, and K. M. Pfarr. **Mitochondrial genes for heme-dependent respiratory chain complexes are up-regulated after depletion of Wolbachia from filarial nematodes**. *Int. J. Parasitol.*, **40**:1193–1202, Aug 2010. 126

[242] A. U. Rao, L. K. Carta, E. Lesuisse, and I. Hamza. **Lack of heme synthesis in a free-living eukaryote**. *Proc. Natl. Acad. Sci. U.S.A.*, **102**:4270–4275, Mar 2005. 126

[243] T. L. Ulery, S. H. Jang, and J. A. Jaehning. **Glucose repression of yeast mitochondrial transcription: kinetics of derepression and role of nuclear genes**. *Mol. Cell. Biol.*, **14**:1160–1170, Feb 1994. 126

[244] T. T. Torres, M. Dolezal, C. Schlotterer, and B. Ottenwalder. **Expression profiling of Drosophila mitochondrial genes via deep mRNA sequencing**. *Nucleic Acids Res.*, **37**:7509–7518, December 2009. 126, 129

[245] N. Galtier, B. Nabholz, S. Glemin, and G. D. Hurst. **Mitochondrial DNA as a marker of molecular diversity: a reappraisal**. *Mol. Ecol.*, **18**:4541–4550, Nov 2009. 126

[246] M.W. Kennedy and W. Harnett. *Parasitic nematodes: molecular biology, biochemistry, and immunology*. CABI, 2001. 127, 128

[247] S. S. Lee, R. Y. Lee, A. G. Fraser, R. S. Kamath, J. Ahringer, and G. Ruvkun. **A systematic RNAi screen identifies a critical role for mitochondria in C. elegans longevity**. *Nat. Genet.*, **33**:40–48, Jan 2003. 127

[248] R. P. Mecham and Parks W. C., editors. *Matrix Metalloproteinases*. Academic Press, 1989. 127

[249] E. Mayr. **Cause and effect in biology**. *Science*, **134**(3489):1501–1506, 1961. 129

[250] W. Sun. **A Statistical Framework for eQTL Mapping Using RNA-seq Data**. *Biometrics*, August 2011. 129

[251] R. Nielsen, S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark, and C. Bustamante. **Genomic scans for selective sweeps using SNP data**. *Genome Res.*, **15**:1566–1575, November 2005. 129

[252] S. Wright. **The genetical structure of populations**. *Annals of Human Genetics*, **15**(1):323–354, 1949. 129

[253] F. Tajima. **Statistical method for testing the neutral mutation hypothesis by DNA polymorphism**. *Genetics*, **123**:585–595, November 1989. 130

[254] N. A. Baird, P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver, Z. A. Lewis, E. U. Selker, W. A. Cresko, and E. A. Johnson. **Rapid SNP discovery and genetic mapping using sequenced RAD markers**. *PLoS ONE*, **3**:e3376, 2008. 130, 132

[255] J. W. Davey, P. A. Hohenlohe, P. D. Etter, J. Q. Boone, J. M. Catchen, and M. L. Blaxter. **Genome-wide genetic marker discovery and genotyping using next-generation sequencing**. *Nat. Rev. Genet.*, **12**:499–510, Jul 2011. 132

[256] P. A. Hohenlohe, S. Bassham, P. D. Etter, N. Stiffler, E. A. Johnson, and W. A. Cresko. **Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags**. *PLoS Genet.*, **6**:e1000862, Feb 2010. 132

[257] Y. F. Chan, M. E. Marks, F. C. Jones, G. Villarreal, M. D. Shapiro, S. D. Brady, A. M. Southwick, D. M. Absher, J. Grimwood, J. Schmutz, R. M. Myers, D. Petrov, B. Jonsson, D. Schluter, M. A. Bell, and D. M. Kingsley. **Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer**. *Science*, **327**:302–305, Jan 2010. 132

[258] B. Ewing, L. Hillier, M. C. Wendl, and P. Green. **Base-Calling of automated sequencer traces using Phred. I. Accuracy Assessment**. *Genome Res.*, **8**(3):175–185, March 1998. 135

[259] P. Green. *PHRAP documentation.*, 1994. 135, 137

[260] 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. 137

[261] 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. 138

[262] C. Iseli, C. V. 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. 138

[263] A. Kasprzyk. **BioMart: driving a paradigm change in biological data management**. *Database (Oxford)*, **2011**:bar049, 2011. 139

[264] 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. 139

[265] S. Falcon and R. Gentleman. **Using GOstats to test gene lists for GO term association**. *Bioinformatics*, **23**:257–258, Jan 2007. 139

[266] M. Morgan and H. Pagès. *Rsamtools: Import aligned BAM file format sequences into R / Bioconductor*. R package version 1.4.3. 139

[267] Herve Pages, Marc Carlson, Seth Falcon, and Nianhua Li. *AnnotationDbi: Annotation Database Interface*. R package version 1.16.10. 140

[268] Adrian Alexa and Jorg Rahnenfuhrer. *topGO: topGO: Enrichment analysis for Gene Ontology*, 2010. R package version 2.6.0. 140

[269] J. H. Boon, V. M. H. Cannaerts, H. Augustijn, M. A. M. 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. 140

[270] 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. 141

[271] 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. 143

[272] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. **Cluster analysis and display of genome-wide expression patterns**. *Proc. Natl. Acad. Sci. U.S.A.*, **95**:14863–14868, Dec 1998. 143

[273] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. 144

[274] Friedrich Leisch. **Sweave: Dynamic Generation of Statistical Reports Using Literate Data Analysis**. In Wolfgang Härdle and Bernd Rönz, editors, *Compstat 2002 — Proceedings in Computational Statistics*, pages 575–580. Physica Verlag, Heidelberg, 2002. ISBN 3-7908-1517-9. 144

[275] Seth Falcon. **Caching code chunks in dynamic documents**. *Computational Statistics*, **24**(2):255–261, 2009. 144

[276] Hadley Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009. 144

[277] H. Chen and P. C. Boutros. **VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R**. *BMC Bioinformatics*, **12**:35, 2011. 144

9

Additional tables and figures

9.1 Additional tables

9.1.1 Transcriptomic divergence in a common garden experiment

Table 9.1: 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	Expected	p-value
Molecular function					
GO:0004190	aspartic-type endopeptidase activity	7	2	0.03	0.00044
GO:0070001	aspartic-type peptidase activity	7	2	0.03	0.00044
GO:0030248	cellulose binding	1	1	0.00	0.00478
GO:0030600	feruloyl esterase activity	1	1	0.00	0.00478
GO:0052689	carboxylic ester hydrolase activity	27	2	0.13	0.00694
GO:0045505	dynein intermediate chain binding	2	1	0.01	0.00955
GO:0016788	hydrolase activity, acting on ester bond...	193	4	0.92	0.01060
GO:0016787	hydrolase activity	604	7	2.89	0.01256
GO:0030235	nitric-oxide synthase regulator activity	3	1	0.01	0.01429

Continued on next page

Table 9.1 – continued from previous page

GO.ID	Term	Annotated	Significant	Expected	p-value
GO:0044183	protein binding involved in protein fold...	3	1	0.01	0.01429
Biological process					
GO:0002478	antigen processing and presentation of e...	7	2	0.04	0.00055
GO:0019886	antigen processing and presentation of e...	7	2	0.04	0.00055
GO:0019884	antigen processing and presentation of e...	8	2	0.04	0.00073
GO:0002495	antigen processing and presentation of p...	9	2	0.05	0.00093
GO:0002504	antigen processing and presentation of p...	9	2	0.05	0.00093
GO:0048002	antigen processing and presentation of p...	13	2	0.07	0.00199
GO:0019882	antigen processing and presentation	15	2	0.08	0.00266
GO:0008219	cell death	406	7	2.16	0.00274
GO:0016265	death	406	7	2.16	0.00274
GO:0048102	autophagic cell death	19	2	0.10	0.00428
Cellular compartment					
GO:0005768	endosome	109	4	0.48	0.00094
GO:0043230	extracellular organelle	2	1	0.01	0.00880
GO:0065010	extracellular membrane-bounded organelle	2	1	0.01	0.00880
GO:0070062	extracellular vesicular exosome	2	1	0.01	0.00880
GO:0043025	neuronal cell body	105	3	0.46	0.00951
GO:0000323	lytic vacuole	106	3	0.47	0.00976
GO:0044297	cell body	109	3	0.48	0.01054
GO:0000328	fungal-type vacuole lumen	3	1	0.01	0.01317
GO:0061200	clathrin sculpted gamma-aminobutyric aci...	3	1	0.01	0.01317

Continued on next page

Table 9.1 – continued from previous page

GO.ID	Term	Annotated	Significant	Expected	p-value
GO:0061202	clathrin sculpted gamma-aminobutyric aci...	3	1	0.01	0.01317

Table 9.2: GO-terms enriched in DE between worm-populations - The top 10 enriched GO-categories are given for genes DE between the different worm populations.

GO.ID	Term	Annotated	Significant	Expected	p-value
Molecular function					
GO:0016491	oxidoreductase activity	189	9	1.67	1.7e-05
GO:0004129	cytochrome-c oxidase activity	17	3	0.15	0.00038
GO:0015002	heme-copper terminal oxidase activity	17	3	0.15	0.00038
GO:0016676	oxidoreductase activity, acting on a hem...	17	3	0.15	0.00038
GO:0016616	oxidoreductase activity, acting on the C...	42	4	0.37	0.00042
GO:0004622	lysophospholipase activity	4	2	0.04	0.00044
GO:0016675	oxidoreductase activity, acting on a hem...	19	3	0.17	0.00054
GO:0016614	oxidoreductase activity, acting on CH-OH...	46	4	0.41	0.00060
GO:0004607	phosphatidylcholine-sterol O-acyltransfe...	5	2	0.04	0.00074
Biological process					
GO:0034186	apolipoprotein A-I binding	5	2	0.04	0.00074
GO:0046688	response to copper ion	25	4	0.24	7.3e-05
GO:0006123	mitochondrial electron transport, cytoch...	11	3	0.11	0.00012
GO:0010035	response to inorganic substance	233	9	2.23	0.00019
GO:0010038	response to metal ion	182	8	1.74	0.00020
GO:0008202	steroid metabolic process	64	5	0.61	0.00028

Continued on next page

Table 9.2 – continued from previous page

GO.ID	Term	Annotated	Significant	Expected	p-value
GO:0034370	triglyceride-rich lipoprotein particle r...	4	2	0.04	0.00052
GO:0034372	very-low-density lipoprotein particle re...	4	2	0.04	0.00052
GO:0009408	response to heat	76	5	0.73	0.00063
GO:0009266	response to temperature stimulus	117	6	1.12	0.00065
Cellular compartment					
GO:0034375	high-density lipoprotein particle remode...	5	2	0.05	0.00087
GO:0034364	high-density lipoprotein particle	4	2	0.03	0.00037
GO:0032994	protein-lipid complex	5	2	0.04	0.00061
GO:0034358	plasma lipoprotein particle	5	2	0.04	0.00061
GO:0031090	organelle membrane	505	11	4.08	0.00078
GO:0044421	extracellular region part	174	6	1.41	0.00197
GO:0005576	extracellular region	250	7	2.02	0.00258
GO:0005739	mitochondrion	605	11	4.89	0.00372
GO:0005743	mitochondrial inner membrane	162	5	1.31	0.00807
GO:0031967	organelle envelope	313	7	2.53	0.00914
GO:0031975	envelope	314	7	2.54	0.00930

Table 9.3: Group-means for OC genes DE between eel species - Group means for expression counts are given for host combination *An. japonica* (Aj) and *An. anguilla* (Aa) with European (EU) and Taiwanese (TW) worm populations. Contig-names, annotation with protein names of *B. malayi* orthologs (second row for each contig) and wormbase transcripts identifiers (third row) are given along with the aggregated counts for these orthologs.

	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
T10B9.2.mean	557.65	662.20	1658.80	1004.08

Continued on next page

Table 9.3 – continued from previous page

	Aa:EU	Aa:TW	Aj:EU	Aj:TW
Contig12201.mean	514.90	549.58	116.02	99.56
Lipase family protein	502.48	553.48	119.47	101.09
F58B6.1.mean	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
Y39B6A.20.mean	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
F01D5.8.mean	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
K06A1.1.mean	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
F21F8.7.mean	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
F11G11.11.mean	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
F23B12.1.mean	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
C44C10.1.mean	231.55	1298.61	2272.71	38.23

Table 9.4: Group-means for OC genes DE between worm populations - Group means for expression counts are given for host combination *An. japonica* (Aj) and *An. anguilla* (Aa) with European (EU) and Taiwanese (TW) worm populations. Contig-names, annotation with protein names of *B. malayi* orthologs (second row for each contig) and wormbase transcripts identifiers (third row) are given along with the aggregated counts for these orthologs.

	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
F22E10.2.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
B0272.3.mean	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
F36A2.3.mean	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
M05B5.4.mean	2716.28	2886.46	2225.58	5278.32
Contig3453.mean	269.89	209.33	277.53	1032.13
Lecithin:cholesterol acyltransferase family protein1	2797.98	2969.10	2306.91	6119.67
M05B5.4.mean	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
Y76A2A.1.mean	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
ZK1290.3a.mean	20.86	51.95	26.08	36.53
Contig566.mean	642.74	484.47	337.05	691.06
Eukaryotic aspartyl protease family protein	651.38	496.17	377.95	733.26

Continued on next page

Table 9.4 – continued from previous page

	Aa:EU	Aa:TW	Aj:EU	Aj:TW
F21F8.7.mean	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
T01B6.1.mean	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
H36L18.1.mean	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
MTCE.31.mean	5865.67	19455.08	9437.50	41673.94
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
F11G11.11.mean	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
F23B12.1.mean	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
C44C10.1.mean	231.55	1298.61	2272.71	38.23
Contig8758.mean	390.97	715.11	602.46	494.53
Protein B0207.11, putative	383.10	687.32	626.45	510.14
T08G11.2.mean	389.74	701.10	633.78	511.74

9.2 Additional figures

9.2.1 Pyrosequencing of the *A. crassus* transcriptome

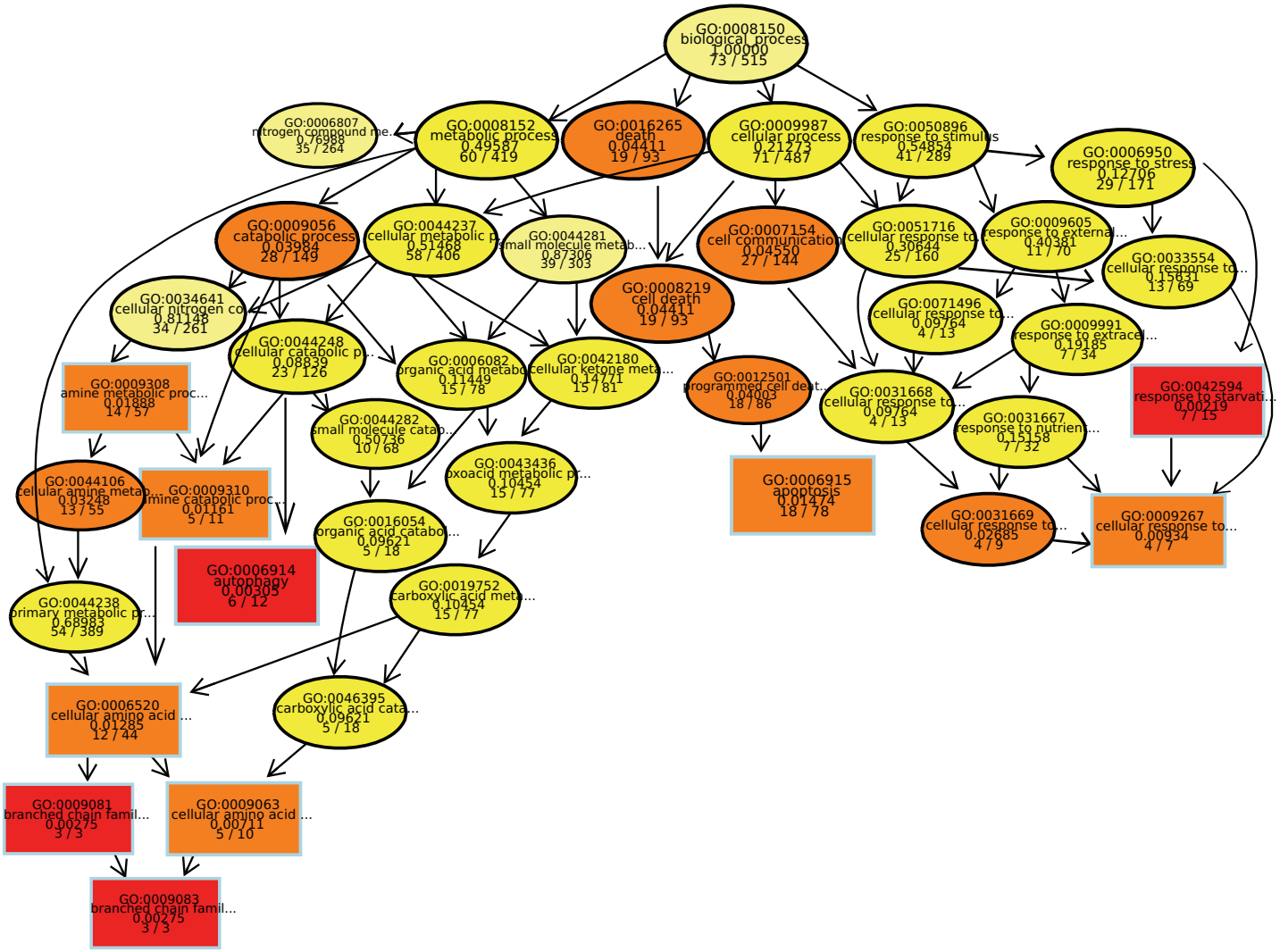

Figure 9.1: GO biological process graph for enriched terms in contigs under positive selection - Subgraph of the GO-ontology biological process category induced by the top 10 terms identified as enriched contigs under positive selection. Boxes indicate the 10 most significant terms. Box colour represents the relative significance, ranging from dark red (most significant) to light yellow (least significant). In each node the category-identifier, a (eventually truncated) description of the term, the significance for enrichment and the number of DE / total number of annotated genes is given. Black arrows indicate an "is-a" relationship.

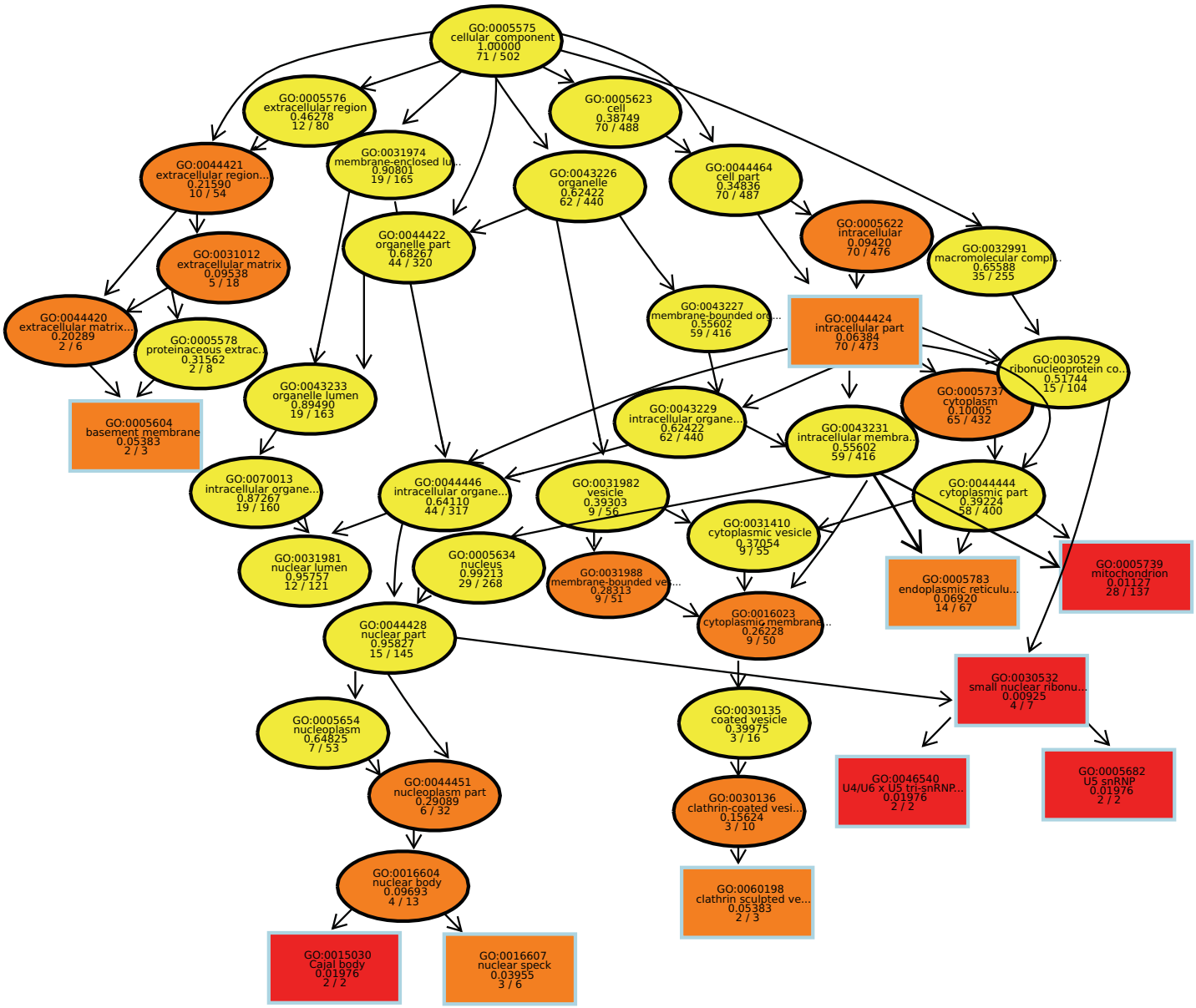

Figure 9.2: GO cellular compartment graph for enriched terms in contigs under positive selection - Subgraph of the GO-ontology cellular compartment category induced by the top 10 terms identified as enriched contigs under positive selection. Boxes indicate the 10 most significant terms. Box colour represents the relative significance, ranging from dark red (most significant) to light yellow (least significant). In each node the category-identifier, a (eventually truncated) description of the term, the significance for enrichment and the number of DE / total number of annotated genes is given. Black arrows indicate an "is-a" relationship.

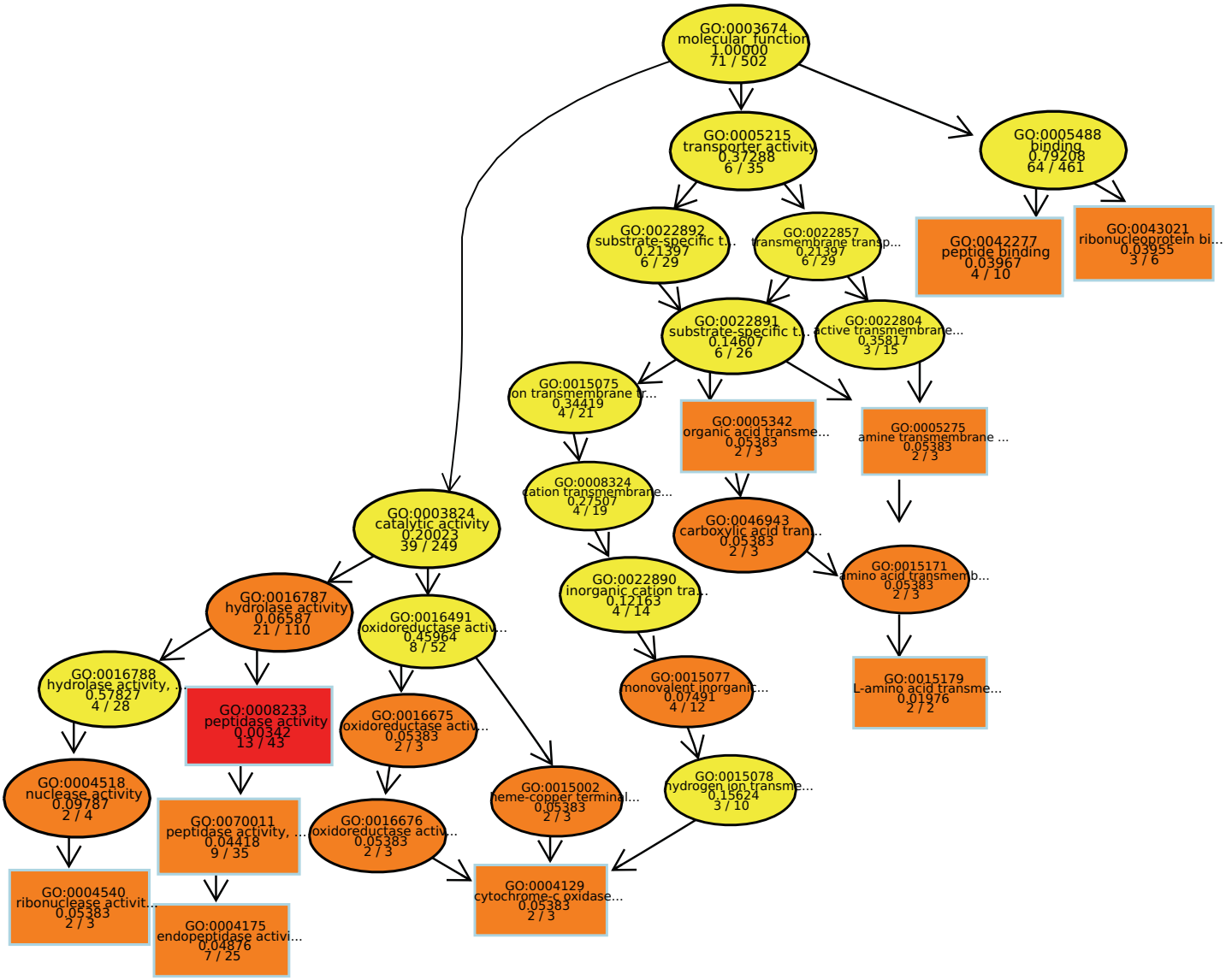

Figure 9.3: GO molecular function graph for enriched terms in contigs under positive selection - Subgraph of the GO-ontology biological process category induced by the top 10 terms identified as enriched contigs under positive selection. Boxes indicate the 10 most significant terms. Box colour represents the relative significance, ranging from dark red (most significant) to light yellow (least significant). In each node the category-identifier, a (eventually truncated) description of the term, the significance for enrichment and the number of DE / total number of annotated gene is given. Black arrows indicate an "is-a" relationship.

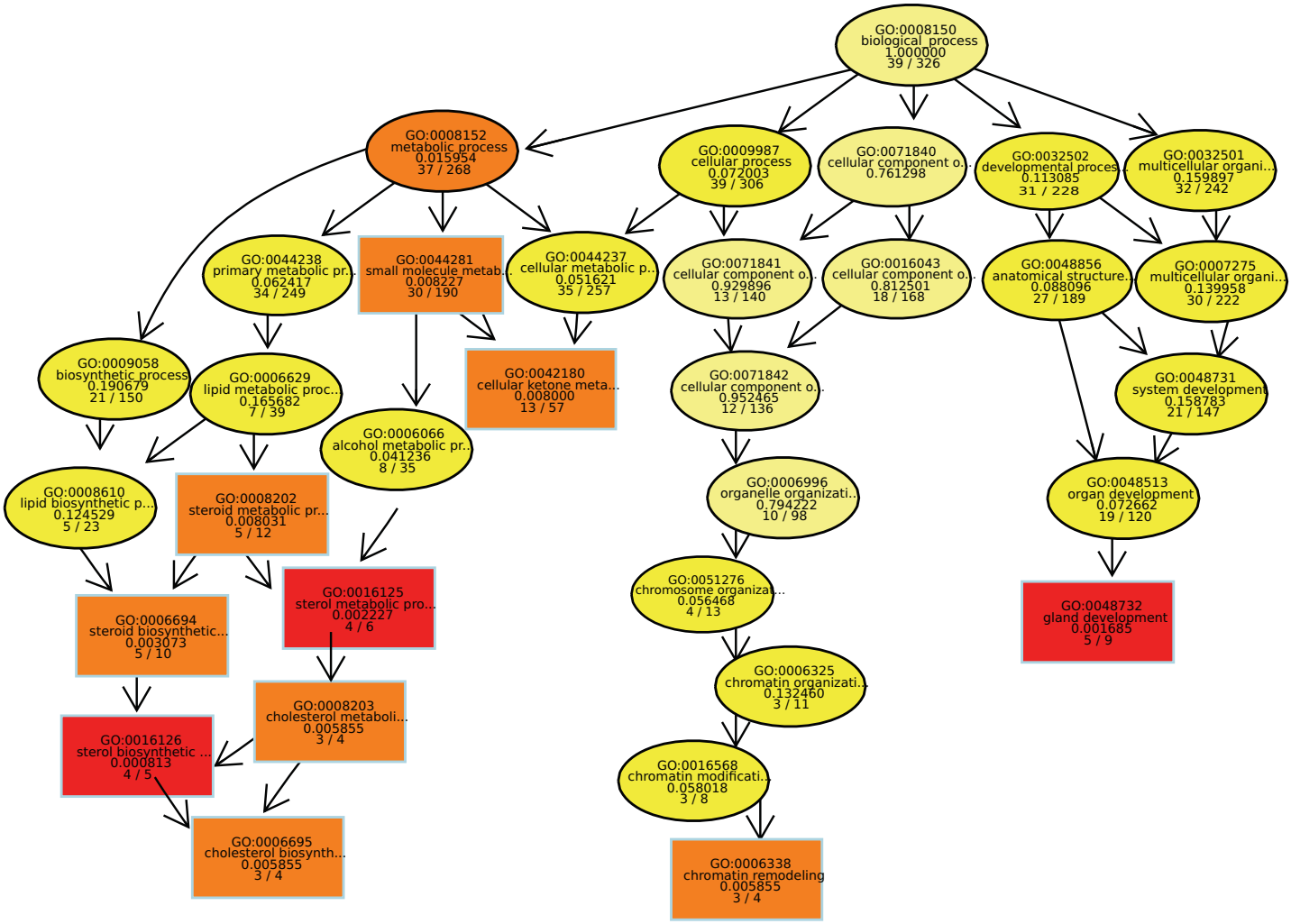

Figure 9.4: GO biological process graph for enriched terms in pyrosequencing-DE genes between worm-origin - Subgraph of the GO-ontology biological process category induced by the top 10 terms identified as enriched in DE genes between worms from Asia and Europe. Boxes indicate the 10 most significant terms. Box colour represents the relative significance, ranging from dark red (most significant) to light yellow (least significant). In each node the category-identifier, a (eventually truncated) description of the term, the significance for enrichment and the number of DE / total number of annotated genes is given. Black arrows indicate an "is-a" relationship.

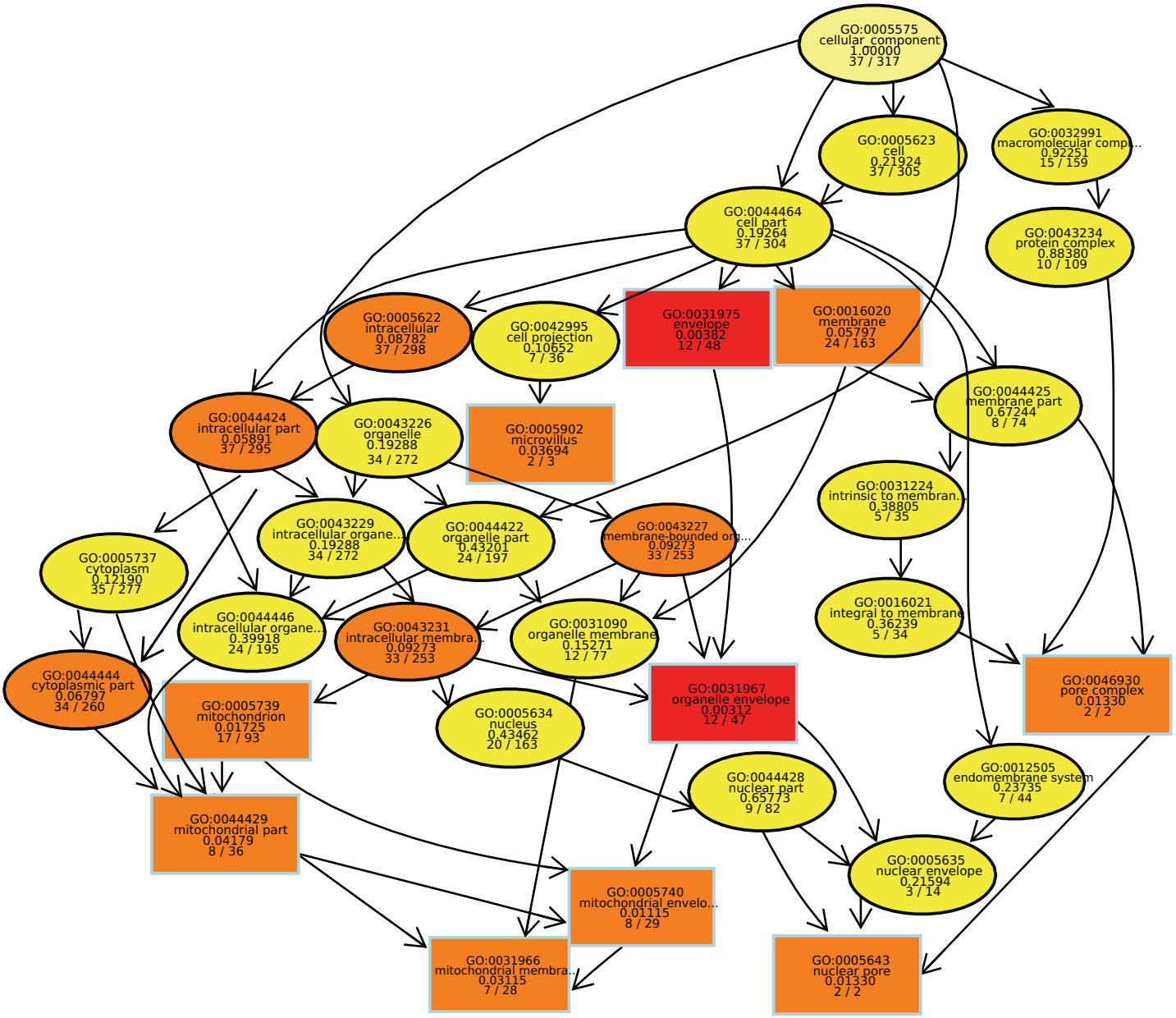

Figure 9.5: GO cellular compartment graph for enriched terms in pyrosequencing-DE genes between worm-origin - Subgraph of the GO-ontology cellular compartment category induced by the top 10 terms identified as enriched in DE genes between worms from Asia and Europe. Boxes indicate the 10 most significant terms. Box colour represents the relative significance, ranging from dark red (most significant) to light yellow (least significant). In each node the category-identifier, a (eventually truncated) description of the term, the significance for enrichment and the number of DE / total number of annotated genes is given. Black arrows indicate an "is-a" relationship.

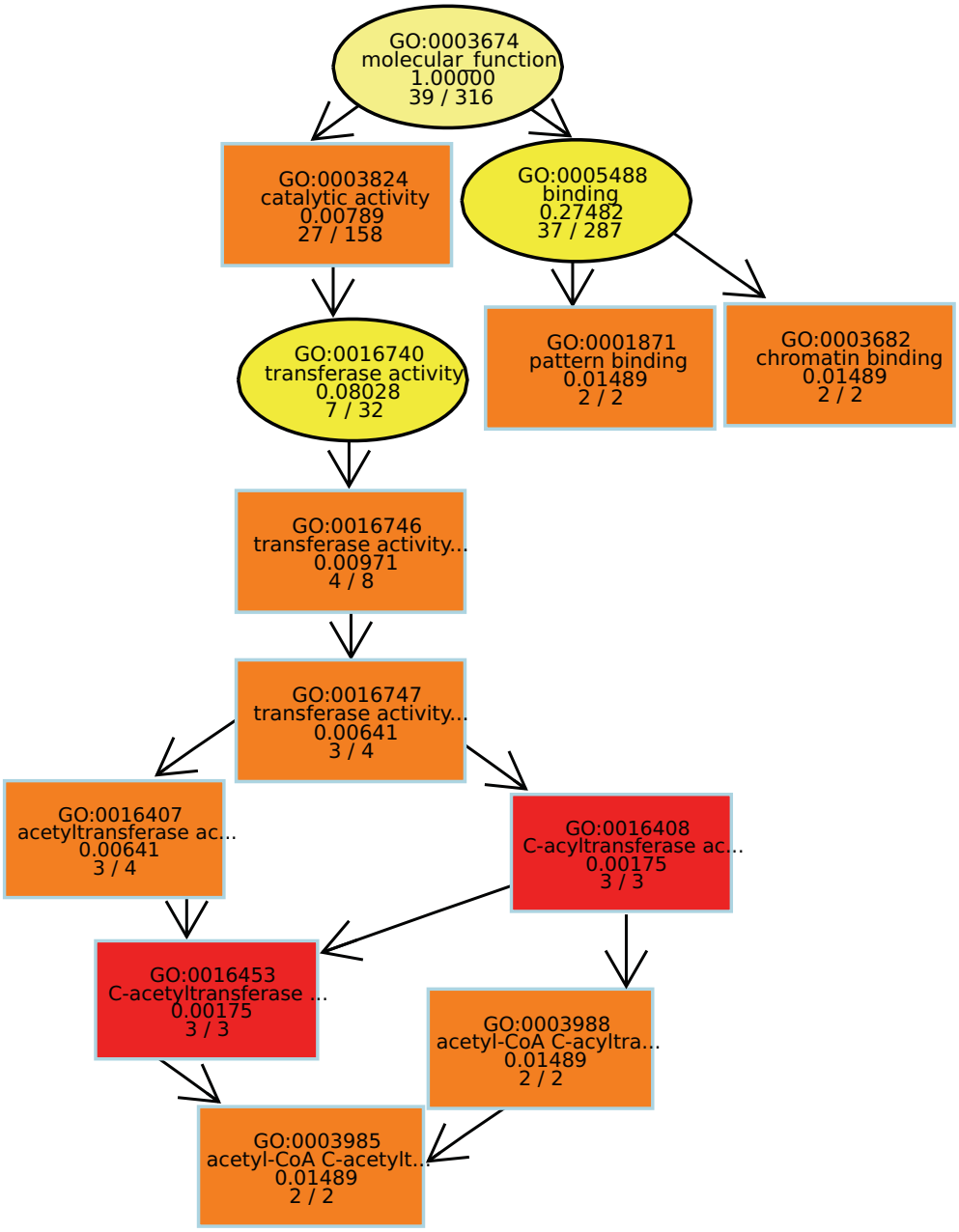

Figure 9.6: GO molecular function graph for enriched terms in pyrosequencing-DE genes between worm-origin - Subgraph of the GO-ontology molecular function category induced by the top 10 terms identified as enriched in DE genes between worms from Asia and Europe. Boxes indicate the 10 most significant terms. Box colour represents the relative significance, ranging from dark red (most significant) to light yellow (least significant). In each node the category-identifier, a (eventually truncated) description of the term, the significance for enrichment and the number of DE / total number of annotated gene is given. Black arrows indicate an "is-a" relationship.

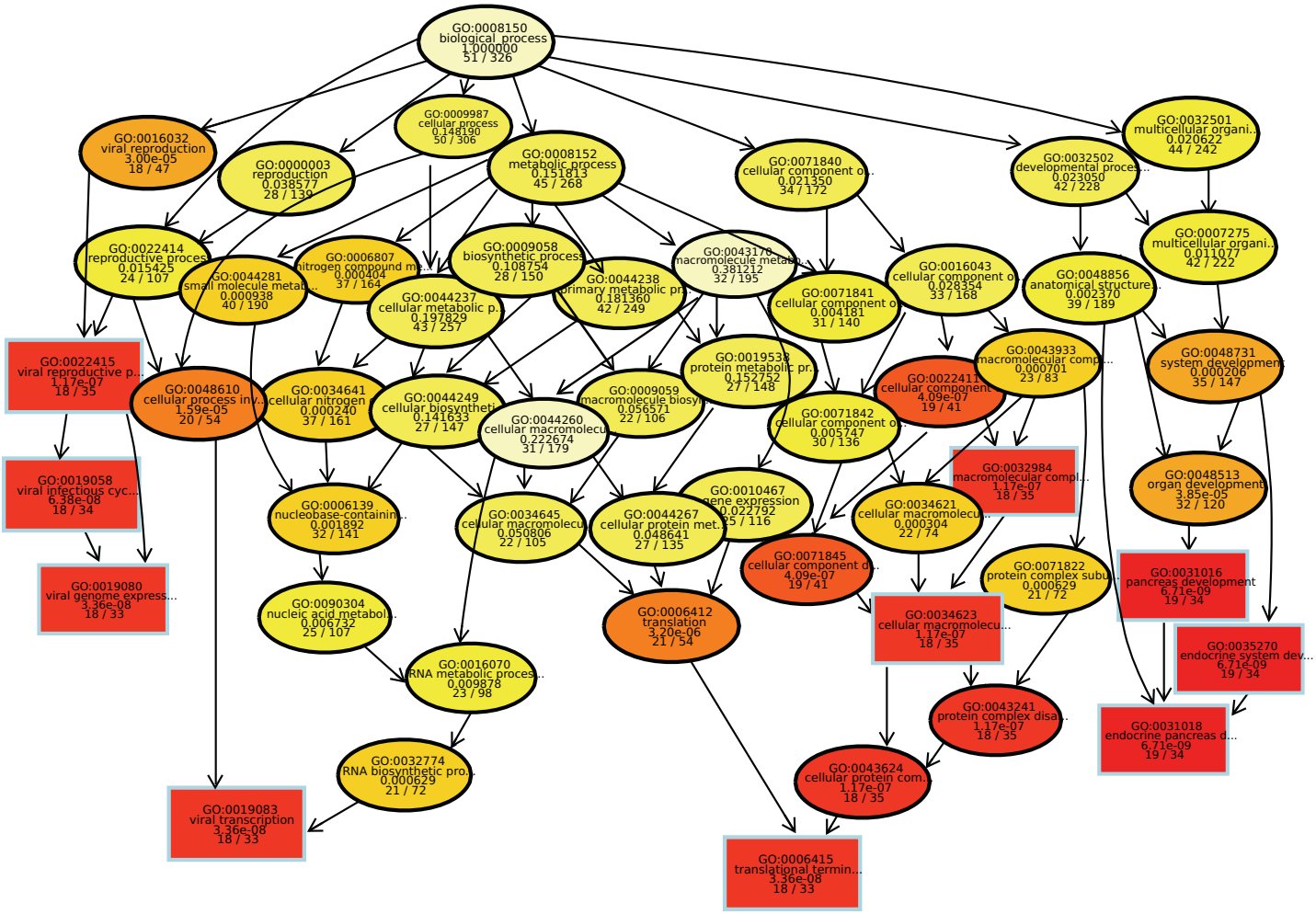

Figure 9.7: GO biological process graph for enriched terms in pyrosequencing-DE genes between worm-sex - Subgraph of the GO-ontology cellular compartment category induced by the top 10 terms identified as enriched in DE genes between female and male worms. Boxes indicate the 10 most significant terms. Box colour represents the relative significance, ranging from dark red (most significant) to light yellow (least significant). In each node the category-identifier, a (eventually truncated) description of the term, the significance for enrichment and the number of DE / total number of annotated genes is given. Black arrows indicate an "is-a" relationship.

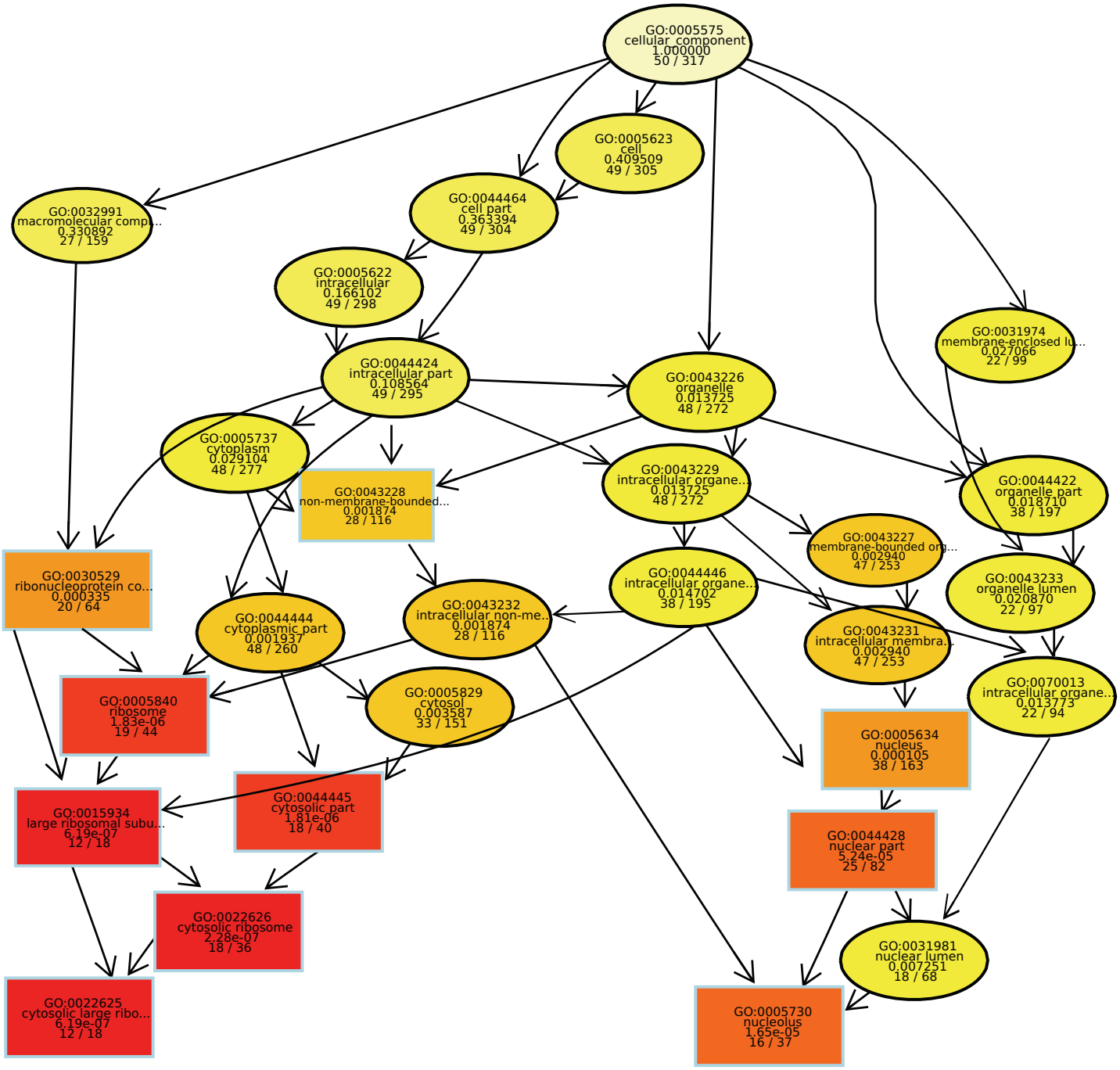

Figure 9.8: GO cellular compartment graph for enriched terms in pyrosequencing-DE genes between worm-sex - Subgraph of the GO-ontology cellular compartment category induced by the top 10 terms identified as enriched in DE genes between female and male worms. Boxes indicate the 10 most significant terms. Box colour represents the relative significance, ranging from dark red (most significant) to light yellow (least significant). In each node the category-identifier, a (eventually truncated) description of the term, the significance for enrichment and the number of DE / total number of annotated genes is given. Black arrows indicate an "is-a" relationship.

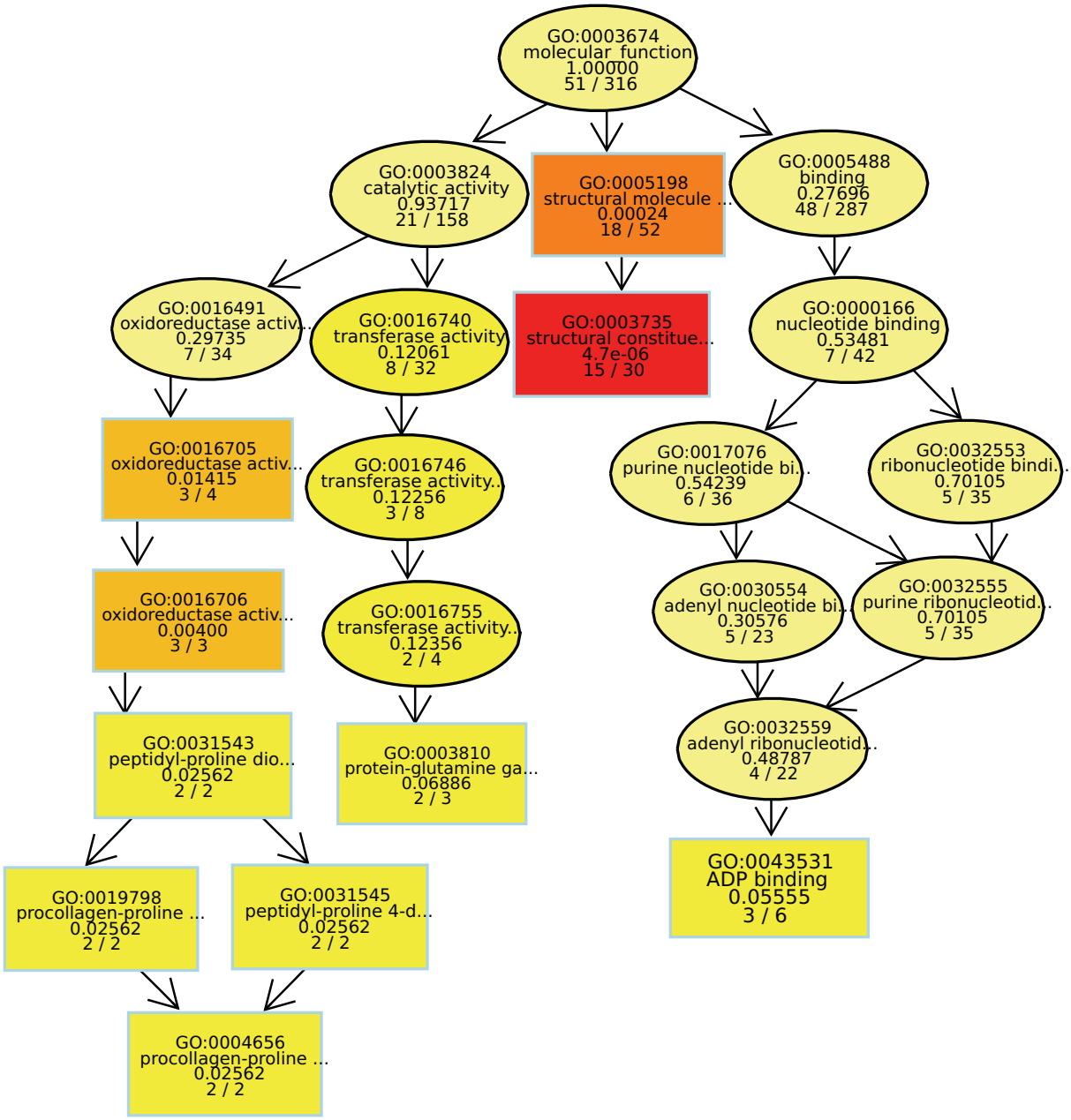

Figure 9.9: GO molecular function graph for enriched terms in pyrosequencing-DE genes between worm-sex - Subgraph of the GO-ontology cellular compartment category induced by the top 10 terms identified as enriched in DE genes between female and male worms. Boxes indicate the 10 most significant terms. Box colour represents the relative significance, ranging from dark red (most significant) to light yellow (least significant). In each node the category-identifier, a (eventually truncated) description of the term, the significance for enrichment and the number of DE / total number of annotated genes is given. Black arrows indicate an "is-a" relationship.

9.2.2 Transcriptomic divergence in a common garden experiment

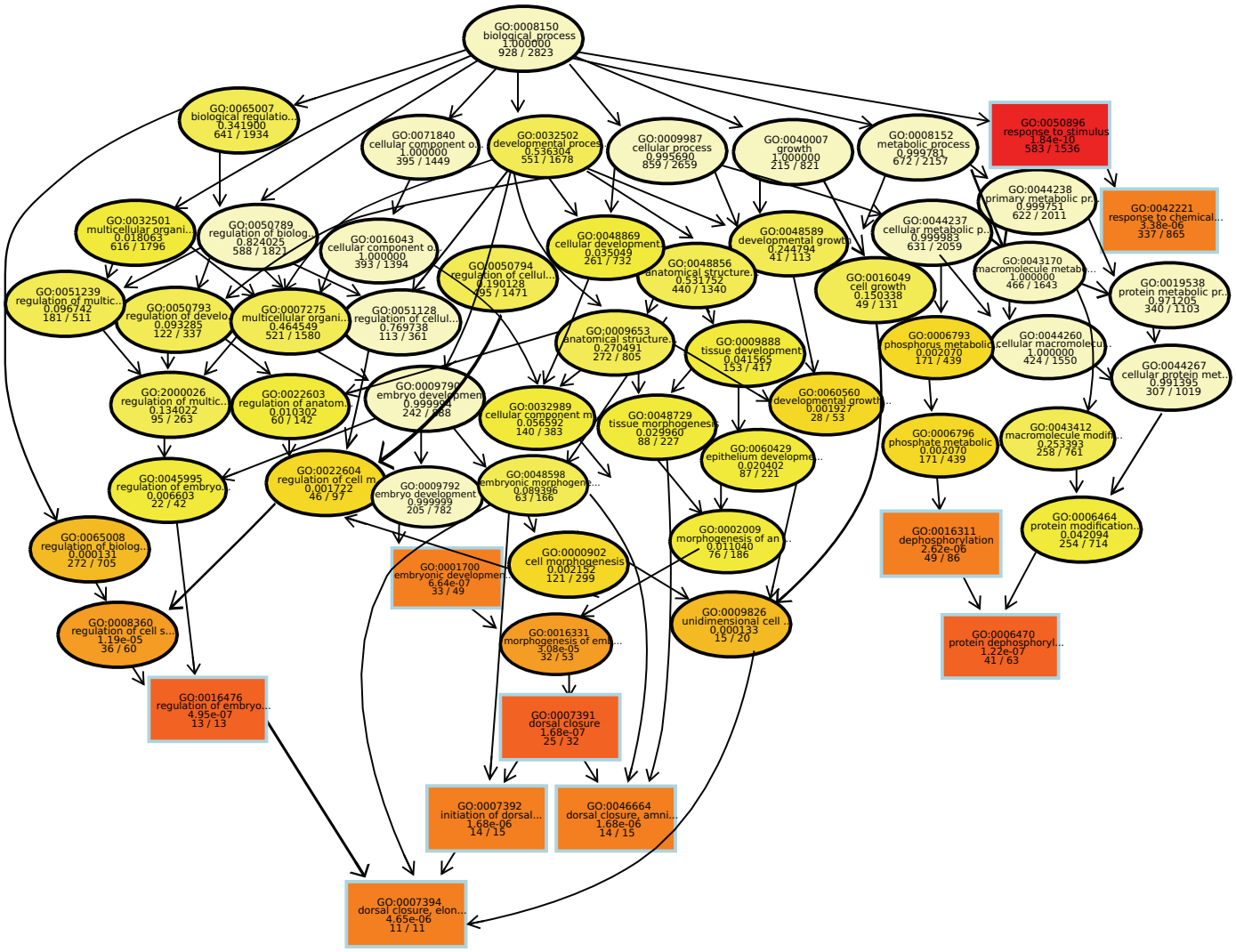

Figure 9.10: GO biological process graph for enriched terms in DE according to sex - Subgraph of the GO-ontology biological process category induced by the top 10 terms identified as enriched in DE genes between male and female worms. Boxes indicate the 10 most significant terms. Box colour represents the relative significance, ranging from dark red (most significant) to light yellow (least significant). In each node the category-identifier, a (eventually truncated) description of the term, the significance for enrichment and the number of DE / total number of annotated genes is given. Black arrows indicate an "is-a" relationship.

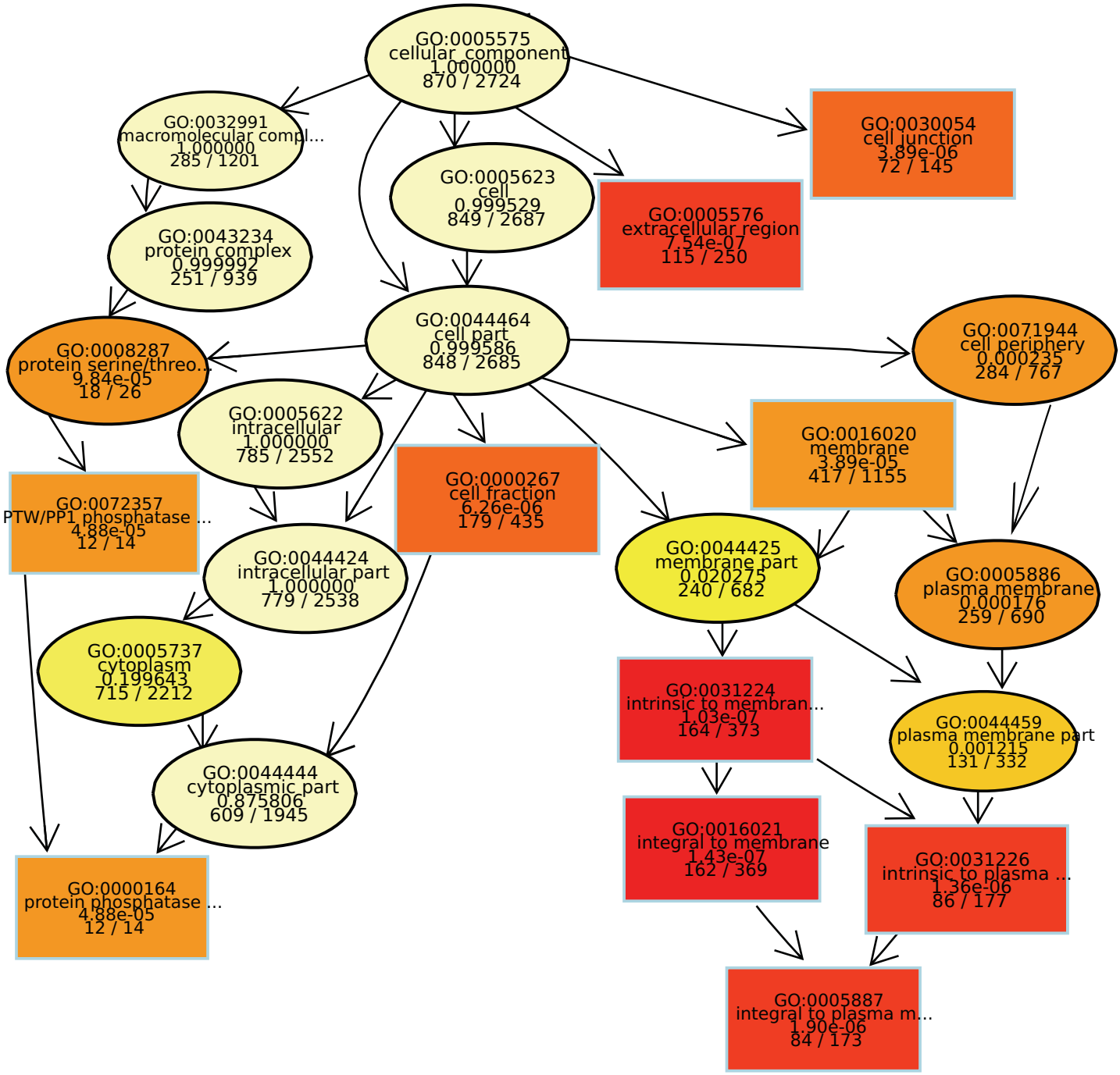

Figure 9.11: GO cellular compartment graph for enriched terms in DE according to sex - Subgraph of the GO-ontology cellular compartment category induced by the top 10 terms identified as enriched in DE genes between male and female worms. Boxes indicate the 10 most significant terms. Box colour represents the relative significance, ranging from dark red (most significant) to light yellow (least significant). In each node the category-identifier, a (eventually truncated) description of the term, the significance for enrichment and the number of DE / total number of annotated gene is given. Black arrows indicate an "is-a" relationship.

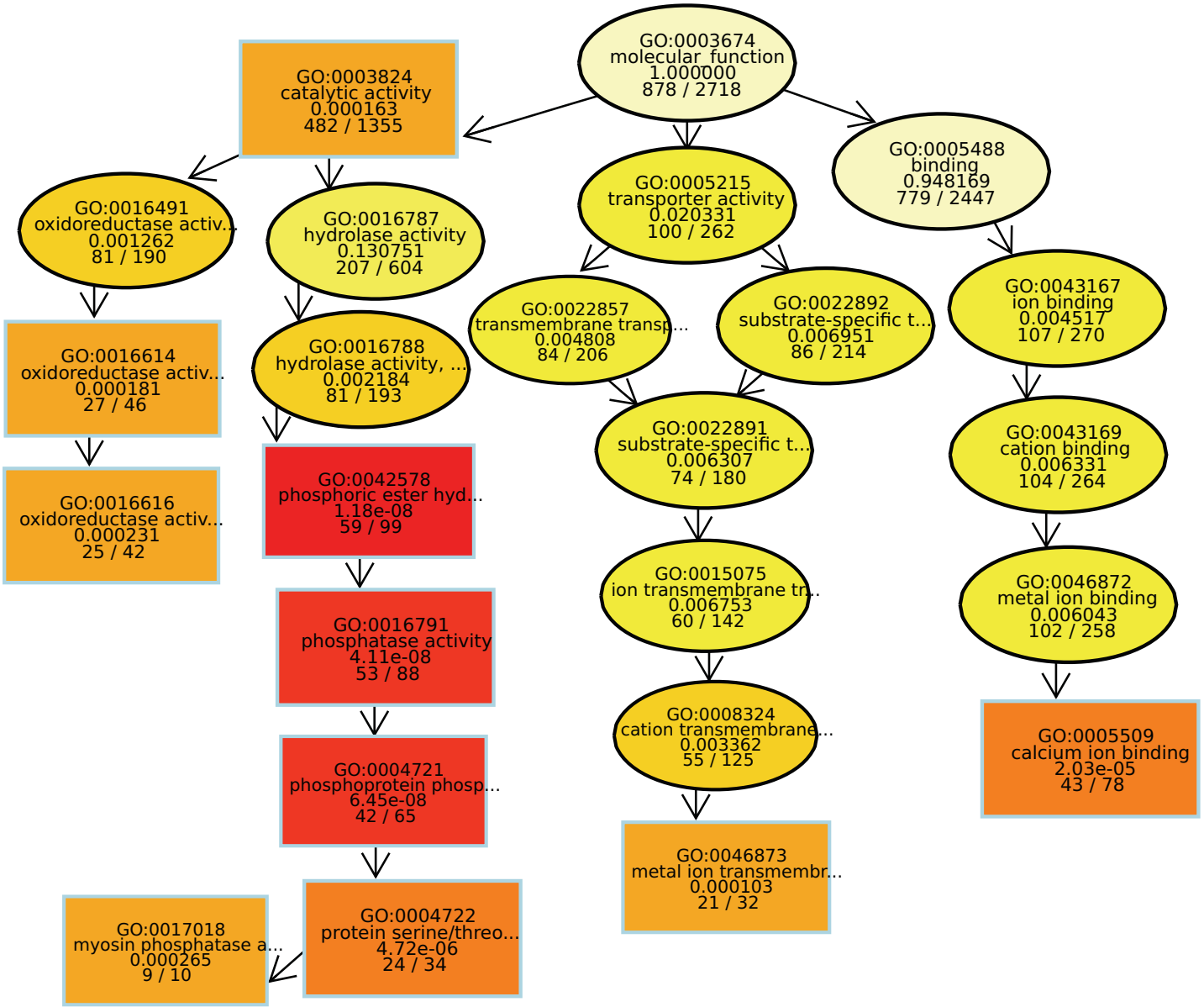

Figure 9.12: GO molecular function graph for enriched terms in DE according to sex - Subgraph of the GO-ontology molecular function category induced by the top 10 terms identified as enriched in DE genes between male and female worms. Boxes indicate the 10 most significant terms. Box colour represents the relative significance, ranging from dark red (most significant) to light yellow (least significant). In each node the category-identifier, a (eventually truncated) description of the term, the significance for enrichment and the number of DE / total number of annotated genes is given. Black arrows indicate an "is-a" relationship.

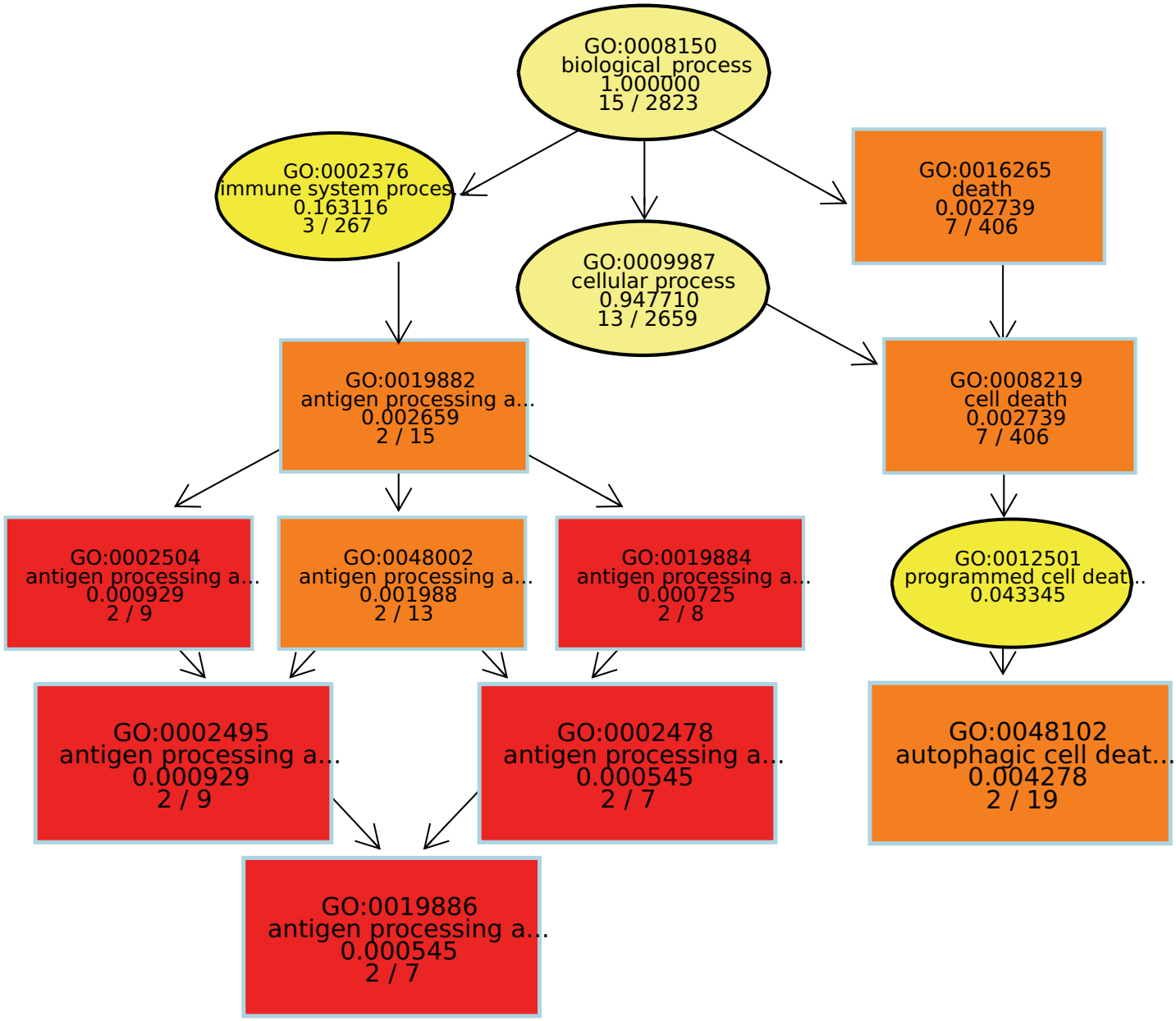

Figure 9.13: GO biological process graph for enriched terms in DE according to eel-host - Subgraph of the GO-ontology biological process category induced by the top 10 terms identified as enriched in DE genes between different host species. Boxes indicate the 10 most significant terms. Box colour represents the relative significance, ranging from dark red (most significant) to light yellow (least significant). In each node the category-identifier, a (eventually truncated) description of the term, the significance for enrichment and the number of DE / total number of annotated gene is given. Black arrows indicate an "is-a" relationship.

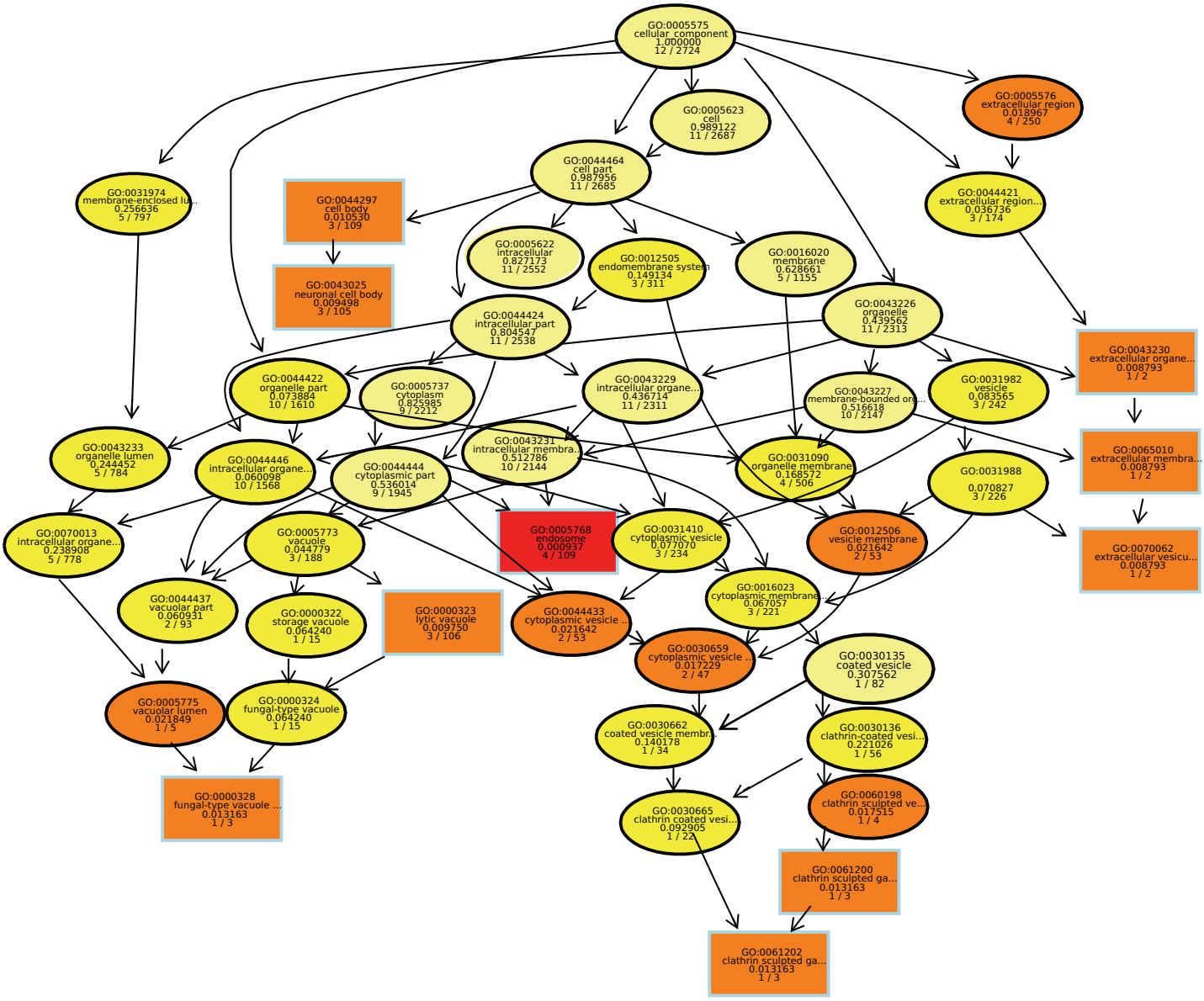

Figure 9.14: GO cellular compartment graph for enriched terms in DE according to eel-host - Subgraph of the GO-ontology cellular compartment category induced by the top 10 terms identified as enriched in DE genes between different host species. Boxes indicate the 10 most significant terms. Box colour represents the relative significance, ranging from dark red (most significant) to light yellow (least significant). In each node the category-identifier, a (eventually truncated) description of the term, the significance for enrichment and the number of DE / total number of annotated gene is given. Black arrows indicate an "is-a" relationship.

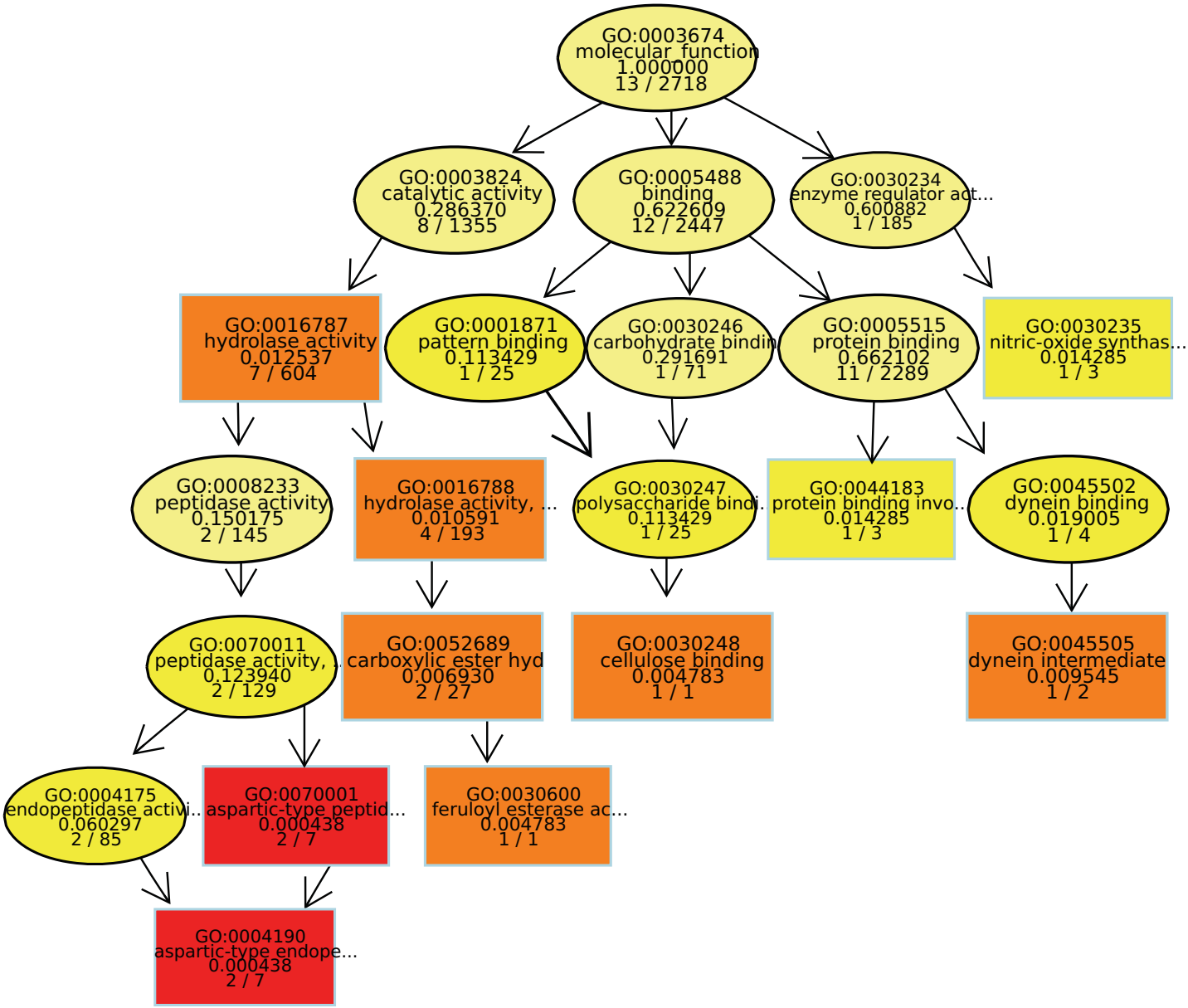

Figure 9.15: GO molecular function graph for enriched terms in DE according to eel-host - Subgraph of the GO-ontology molecular function category induced by the top 10 terms identified as enriched in DE genes between different host species. Boxes indicate the 10 most significant terms. Box colour represents the relative significance, ranging from dark red (most significant) to light yellow (least significant). In each node the category-identifier, a (eventually truncated) description of the term, the significance for enrichment and the number of DE / total number of annotated genes is given. Black arrows indicate an "is-a" relationship.

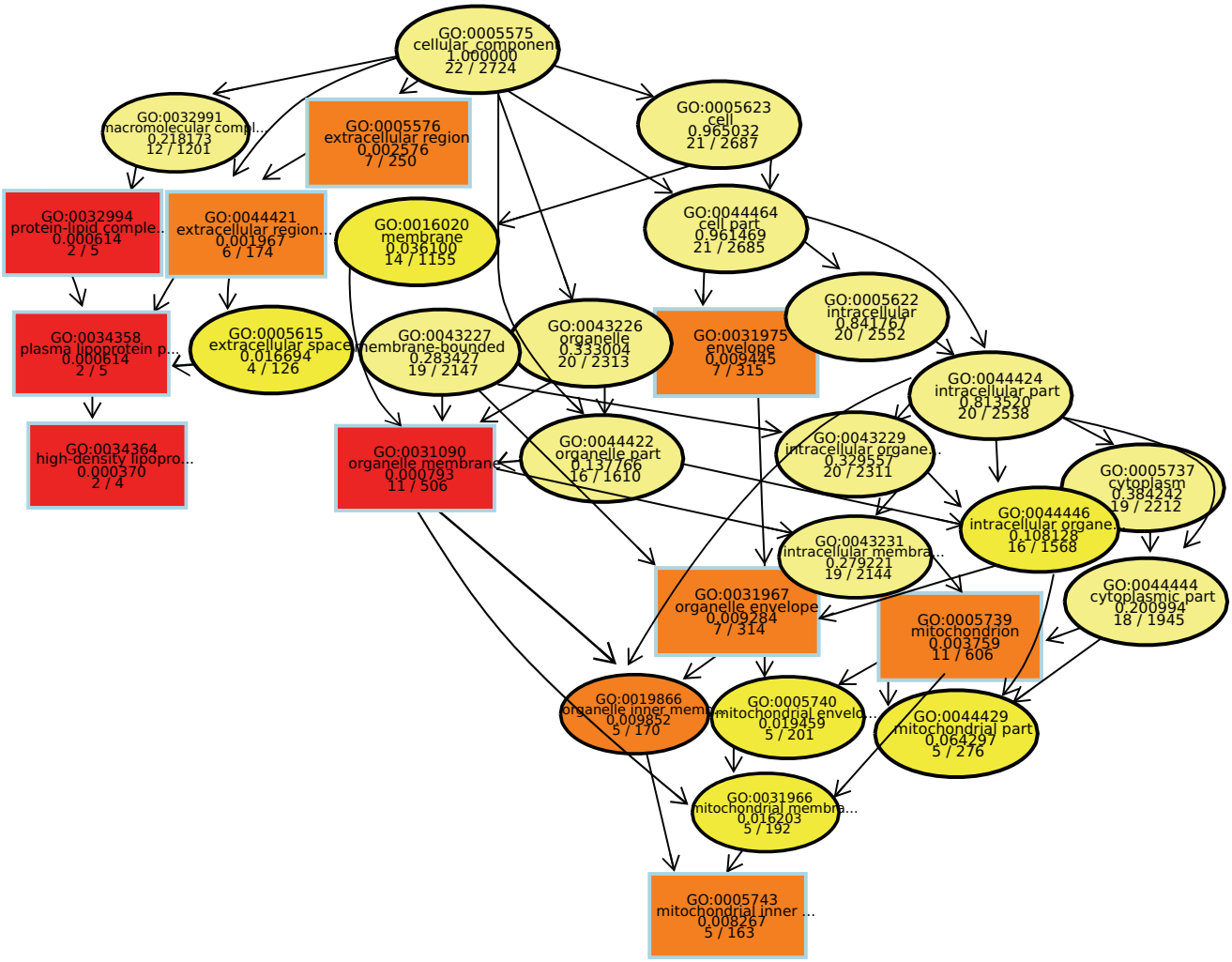

Figure 9.16: GO cellular compartment graph for enriched terms in DE according to worm-population - Subgraph of the GO-ontology biological process category induced by the top 10 terms identified as enriched in DE genes between different parasite populations. Boxes indicate the 10 most significant terms. Box colour represents the relative significance, ranging from dark red (most significant) to light yellow (least significant). In each node the category-identifier, a (eventually truncated) description of the term, the significance for enrichment and the number of DE / total number of annotated genes is given. Black arrows indicate an “is-a” relationship.

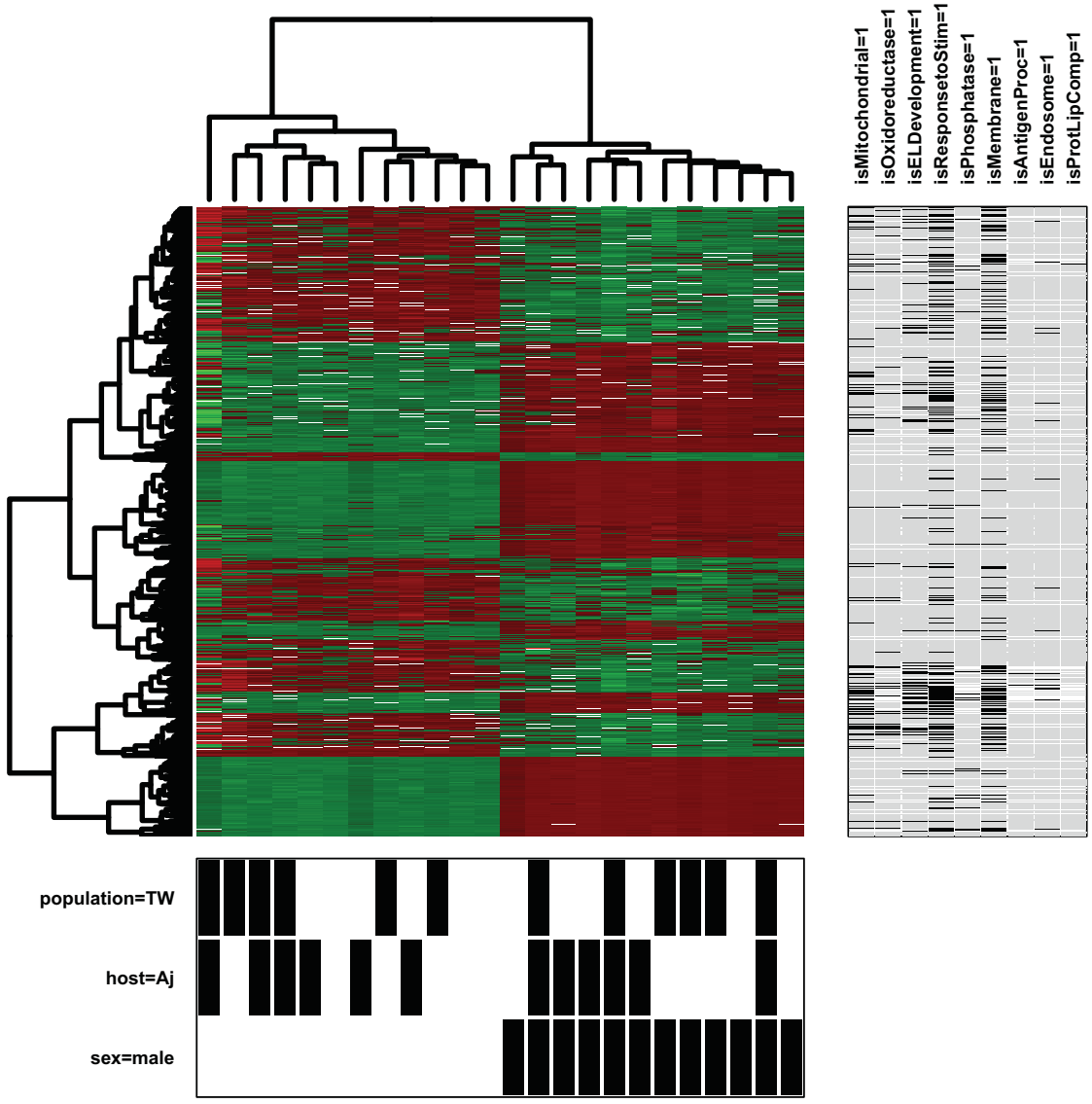

Figure 9.17: Clustering of expression values for contigs DE between female and male worms - A heatmap of variance/mean stabilised expression values. Deprograms are based on hierarchical clustering. Green indicates expression below the mean, red above the mean. Experimental conditions are indicated by black bars for groups of samples (columns) below the plot. Presence GO-term annotation for contigs (rows) are given as black bars right to the plot: isOxidoreductase = GO:0016491, oxidoreductase activity; isMitochondrial = GO:0005739, mitochondrion; isELDevelopment = GO:0002164, larval development or GO:0009791, post-embryonic development; isResponsetoStim = GO:0050896, response to stimulus; isPhosphatase = GO:0016791, phosphatase; isMembrane = GO:0016020, membrane; isAntigenProc = GO:0002478, antigen processing and presentation of exogenous peptide antigen; isEndosome = GO:0005768, endosome; isProtLipComp = GO:0032994, protein-lipid complex. Grey bars indicate no annotation available.

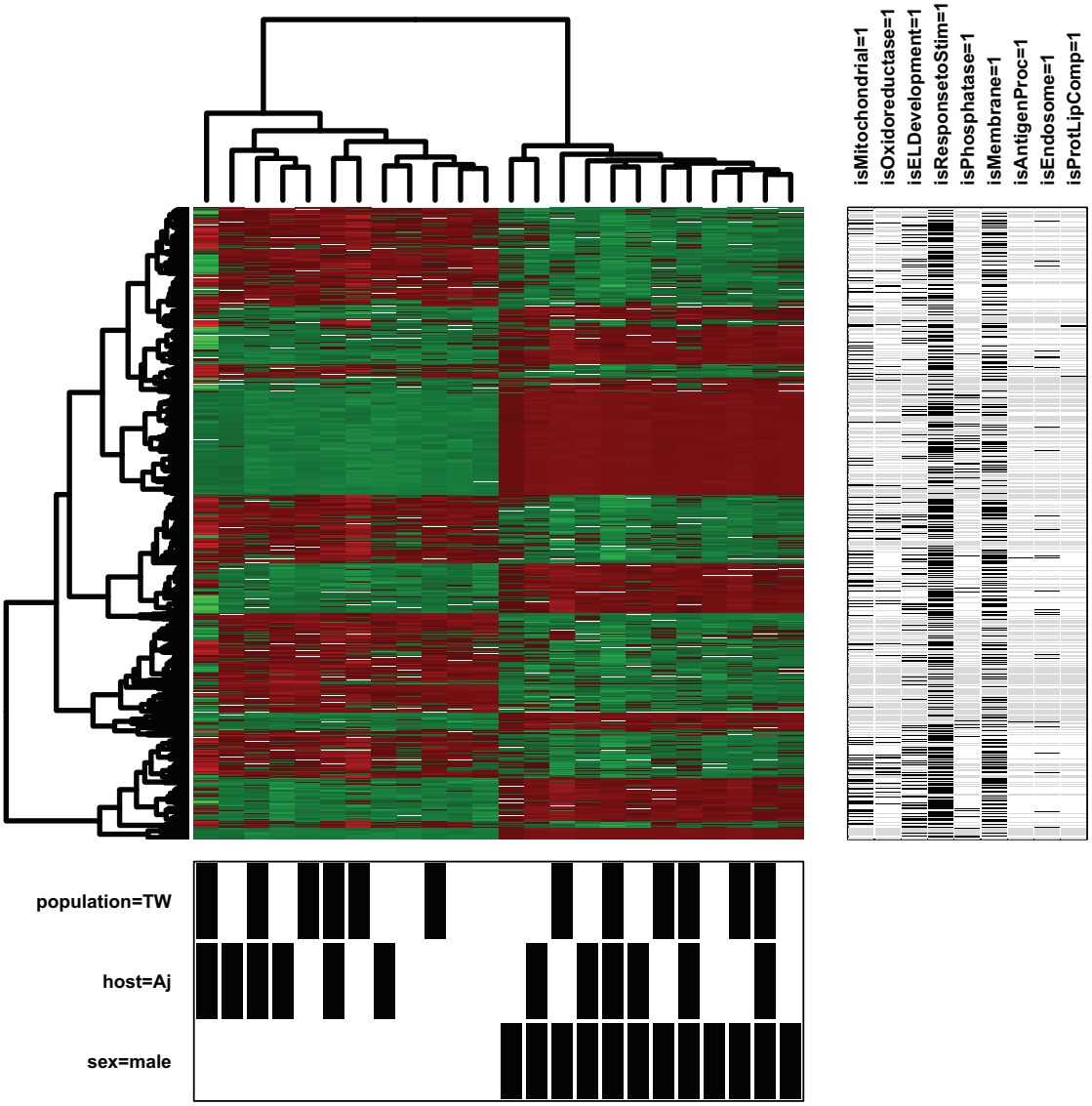

Figure 9.18: Clustering of expression values for OC contigs DE between female and male worms - A heatmap of variance/mean stabilised expression values. Deprograms are based on hierarchical clustering. Green indicates expression below the mean, red above the mean. Experimental conditions are indicated by black bars for groups of samples (columns) below the plot. Presence GO-term annotation for contigs (rows) are given as black bars right to the plot: isOxidoreductase = GO:0016491, oxidoreductase activity; isMitochondrial = GO:0005739, mitochondrion; isELDevelopment = GO:0002164, larval development or GO:0009791, post-embryonic development; isResponsetoStim = GO:0050896, response to stimulus; isPhosphatase = GO:0016791, phosphatase; isMembrane = GO:0016020, membrane; isAntigenProc = GO:0002478, antigen processing and presentation of exogenous peptide antigen; isEndosome = GO:0005768, endosome; isProtLipComp = GO:0032994, protein-lipid complex. Grey bars indicate no annotation available.

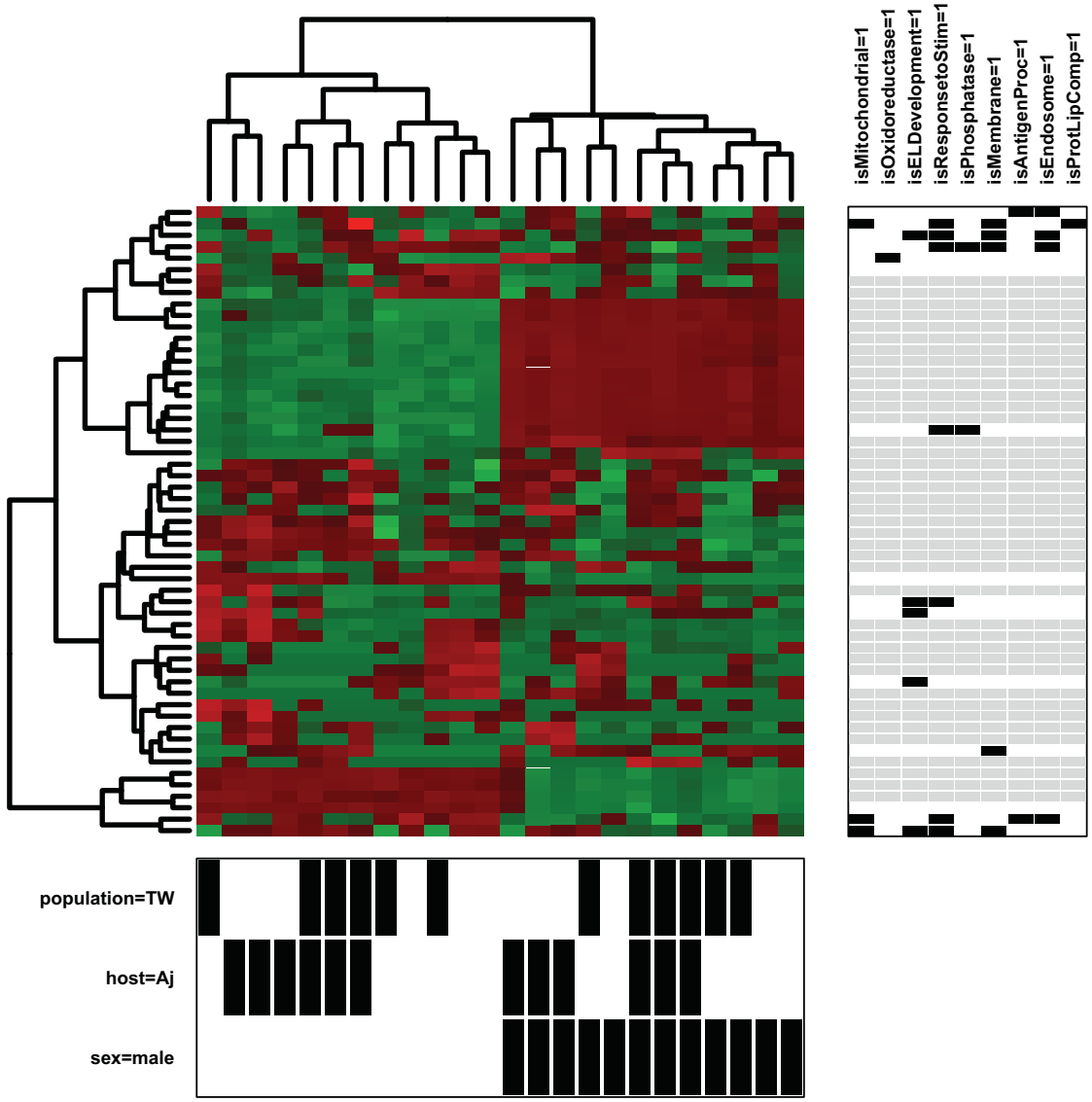

Figure 9.19: Clustering of expression values for contigs DE between worms in *An. japonica* and *An. anguilla* - A heatmap of variance/mean stabilised expression values. Deprograms are based on hierarchical clustering. Green indicates expression below the mean, red above the mean. Experimental conditions are indicated by black bars for groups of samples (columns) below the plot. Presence GO-term annotation for contigs (rows) are given as black bars right to the plot: isOxidoreductase = GO:0016491, oxidoreductase activity; isMitochondrial = GO:0005739, mitochondrion; isELDevelopment = GO:0002164, larval development or GO:0009791, post-embryonic development; isResponsetoStim = GO:0050896, response to stimulus; isPhosphatase = GO:0016791, phosphatase; isMembrane = GO:0016020, membrane; isAntigenProc = GO:0002478, antigen processing and presentation of exogenous peptide antigen; isEndosome = GO:0005768, endosome; isProtLipComp = GO:0032994, protein-lipid complex. Grey bars indicate no annotation available.

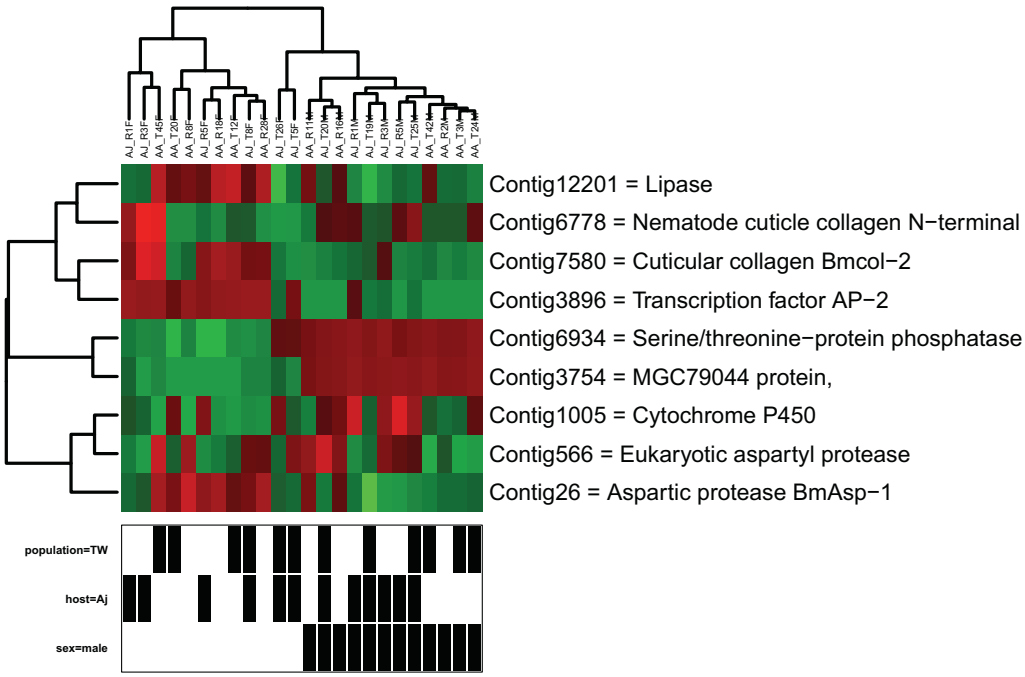

Figure 9.20: Clustering of expression values for OC contigs DE between worms in *An. japonica* and *An. anguilla* - A heatmap of variance/mean stabilised expression values. Deprograms are based on hierarchical clustering. Green indicates expression below the mean, red above the mean. Experimental conditions are indicated by black bars for groups of samples (columns) below the plot. Below contig-names uniprot names are given for ortholog genes in *B. malayi*.

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. The law and statutes of the Karlsruhe Institute of Technology ensuring good scientific practice were followed as in force at the relevant time.

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, March 19, 2012,

Emanuel G. Heitlinger

Redtenbacherstr. 9
76133 Karlsruhe
Germany

Phone: (+49) 721 9822588
Email: emanuelheitlinger@gmail.com
Born September, 12th 1980 in Schwäbisch Gmünd, Germany

Education

2008-2012 Doctoral studies, Karlsruhe Institute of Technology.

Dissertation: Divergence of an introduced population of the swimbladder-nematode *Anguillicola crassus* - a transcriptomic perspective.

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.

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

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.*

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

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. Two 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: March 19, 2012