

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



Zur Erlangung des akademischen Grades eines
DOKTORS DER NATURWISSENSCHAFTEN

(Dr. rer. nat.)

Fakultät für Chemie und Biowissenschaften

Karlsruher Institut für Technologie (KIT) - Universitätsbereich
vorgelege

Dissertation

von

Emanuel Heitlinger

geboren in

Schwäbisch Gmünd

Dekan:

Referent: Prof. Dr. Horst Taraschewski

Korreferent: Prof. Mark Blaxter

Tag der mündlichen Präsentation:

Abstract

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

Gene regulatory networks, as a bridge between genotype and phenotype, are thought to play a central role both in the response to stress (e.g. from sofar unexperienced environmental stressors) and in the divergence and eventually establishment of reproductive barriers between populations. An additional feature of parasite gene-expression is the theoretically deduced need to express only a single allele of a polymorphic parasite infection locus.

In the present project the differences in gene-expression in *A. crassus* populations should be illuminated. The parasite was introduced to Europe 30 years ago, spread successfully in a new host and established stable populations.

Zusammenfassung

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

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

To my grandmother Ruth my brother Roman and my wife Silvia

Acknowledgements

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

Contents

List of Figures	vii
List of Tables	ix
Glossary	xi
1 Introduction	1
1.1 The study organism: <i>Anguillicola crassus</i>	1
1.1.1 Ecological significance	1
1.1.2 Evolutionary significance	6
1.1.2.1 The eel-host	6
1.1.2.2 Interest in <i>A. crassus</i> based on its phylogeny	8
1.1.2.3 A taxonomy of common garden experiments and the divergence of <i>A. crassus</i> populations	11
1.2 DNA sequencing	16
1.2.1 Two out of three: DNA sequencing and the central dogma of molecular biology	16
1.2.2 The history and methods of high-throughput DNA-sequencing	19
1.2.3 DNA-sequencing in Nematodes	19
1.2.4 Advances in sequencing technology	21
1.2.4.1 Pyro-sequencing	23
1.2.4.2 Illumina-Solexa sequencing	25
1.2.5 Computational methods in DNA-sequence analysis	27
1.2.6 Applications of NGS in ecology and evolution	29
1.3 Gene-expression and evolutionary divergence	30

CONTENTS

2 Aims of the project	31
2.1 Preliminary aims	31
2.2 Final aim	31
3 Pilot sequencing (Sanger method)	33
4 Evaluation of an assembly strategy for pyrosequencing reads	39
4.1 Overview	39
4.2 The <code>newbler</code> first-order assembly	39
4.3 The <code>mira</code> -assembly and the second-order assembly	41
4.4 Data-categories in the second-order assembly	42
4.5 Contribution of first-order assemblies to second-order contigs	44
4.6 Evaluation of the assemblies	44
4.7 Measurements on second-order assembly	46
4.7.1 Contig coverage	50
4.7.2 Example use of the contig-measurements	50
4.8 Finalising the fullest assembly set	52
5 Pyrosequencing of the <i>A. crassus</i> transcriptome	53
6 Transcriptomic divergence in a common garden experiment	55
6.1 Infection experiments	55
6.2 Sample preparation and sequencing	57
6.3 Examination of data-quality	58
6.4 Expression differences in general linear models	58
6.4.1 Foo	61
6.5 Expression differences between worms in European and Japanese Eels .	61
6.6 Expression differences between worms in the European and Taiwanese worm-population	61
7 Discussion	65
7.1 Sanger-method pilot-sequencing	65
7.2 454-pyrosequencing	65
7.3 Experimental infections	67

CONTENTS

8 Materials & methods	69
8.1 Sampling of worms from wild eels (Sanger- and pyro-sequencing)	69
8.2 RNA-extraction and cDNA synthesis for Sanger- and 454-sequencing	69
8.3 Cloning for Sanger-sequencing	70
8.4 Pilot Sanger-sequencing	71
8.5 454-pyro-sequencing	72
8.6 Transcriptomic divergence in a common garden experiment	75
8.6.1 Experimental infection of eels	75
8.7 RNA extraction and preparation of sequencing libraries	76
8.8 Mapping and read-counts	77
8.9 Statistical analysis	77
8.10 Collapsing of orthologous for two model-species	77
References	79

CONTENTS

List of Figures

1.1	Transcontinental dispersal of <i>A. crassus</i> :	2
1.2	Life-cycle of <i>A. crassus</i>	4
1.3	Difference between worms in the swimbladder of the European eel and the Japanese eel	6
1.4	Phylogeny of the genus <i>Anguillicola</i> based nLSU	9
1.5	Phylogeny of the genus <i>Anguillicola</i> based on COXI	10
1.6	Phylogeny of nematode clade III based on nuclear small ribosomal subunit	12
1.7	Differences in developmental speed	14
1.8	Major macromolecules bearing biological sequence information:	16
1.9	The structure of a protein coding gene and it's mRNA	17
1.10	Falling sequencing costs	22
1.11	Schematic representation of pyrosequencing	24
1.12	Schematic representation of illumina sequencing	26
3.1	Proportion of rRNA in different libraries for <i>A. crassus</i> and <i>A. japonica</i>	34
3.2	GC-content of sequences from <i>A. japonica</i> and <i>A. crassus</i>	36
4.1	Number of contigs/isotigs splitted	40
4.2	Origing of reads	43
4.3	Contribution to second-order assembly	45
4.4	Base-content and reference-transcriptome coverage in percent of bases .	47
4.5	Base-content and reference-transcriptome coverage	48
4.6	Base-content and reference-transcriptome coverage in percent of proteins coverd to at least 80%	49
6.1	Recovery of worms in cross-infection experiment	56

LIST OF FIGURES

6.2	Distances between RNA-seq read-count for different samples	59
6.3	Principle coordinate plot for expression inRNA-seq libraries	60
6.4	Variance/mean stabilized expression values for contigs different between populations	62

List of Tables

3.1	Screening statistics for pilot sequencing	35
3.2	Annotaion of putative host-derived sequences in the <i>A. crassus</i> -dataset .	38
4.1	Statistics for the first-order assemblies	41
4.2	number of reads in assemblies	42
4.3	Example for assembly-measurements	51
4.4	finalizing the assembly	52
6.1	Linear model for recovery	57
6.2	Summary of RNA preparation	63
6.3	Mapping Summary	64
8.1	PCR protocol for insert amplification	70

GLOSSARY

Glossary

	days after an individual has been infected
ORF	Open Reading Frame; a region in a DNA-sequence begining with a start-codon and not containing a stop-codon. For example a region within a processed mRNA transcript being transcribed into a protein
SNP	Single Nucleotide Polymorphism; variation occurring in a single nucleotide between two closely related homlogous sequences. Leading to for example to allelic differences within a population or even the homologous chromosomes in an individual
DNA	Desoxy Ribonucleic Acid; a chemical molecule bearing the heritable genetic information in all life on earth
dpi	Days post infection; In infection experiments, a point in time given in

GLOSSARY

1

Introduction

1.1 The study organism: *Anguillicola crassus*

1.1.1 Ecological significance

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

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

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

1. INTRODUCTION

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

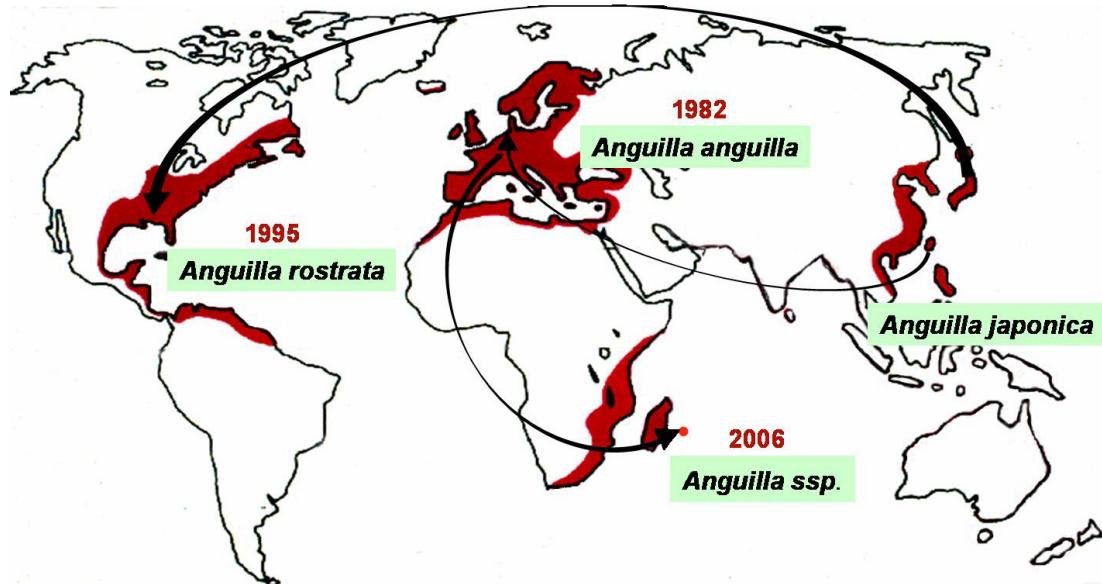


Figure 1.1: Transcontinental dispersal of *A. crassus*: - Invasions of different continents by different source-populations are illustrated using arrows. Red color indicates the range of the eel species targeted by the invasion. Modified from (11), based on data reviewed in (4) and newer findings in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

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

1.1 The study organism: *Anguillicola crassus*

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

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

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

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

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

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

1. INTRODUCTION

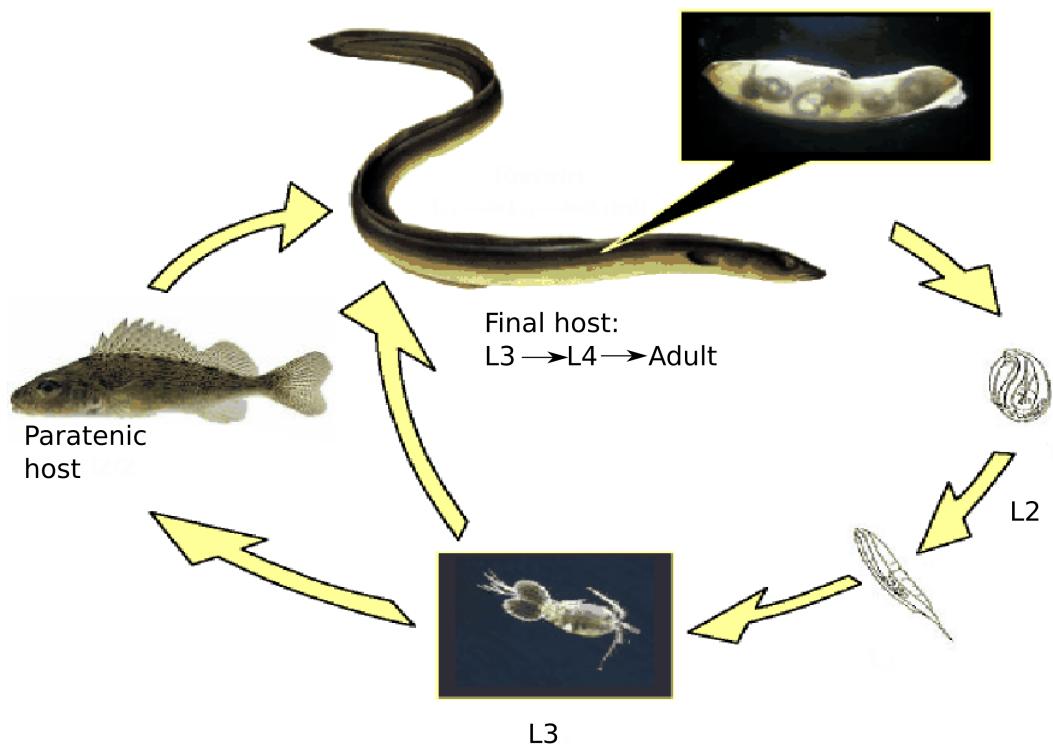


Figure 1.2: Life-cycle of *A. crassus* - Adult females deposit already hatched L2 in the lumen of the swimbladder. Larvae migrate through the *ductus pneumaticus* and the intestine into the open water. Copepods serve as intermediate host where infective L3-larvae develop. These can be transported and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

1.1 The study organism: *Anguillicola crassus*

in 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* (28, 29, 30), including even amphibians and larvae of aquatic insects (31).

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

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

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

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

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

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

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

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

1. INTRODUCTION

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

Interestingly the differences in the two host also affect the size and life-history of the worm: In European eels the nematodes are bigger and develop and reproduce faster (35).

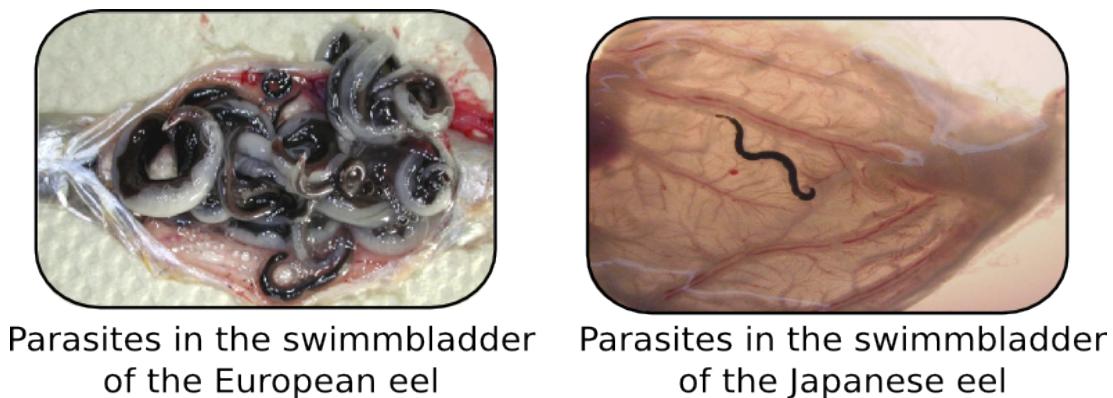


Figure 1.3: Difference between worms in the swimbladder of the European eel and the Japanese eel - Note the bigger size and higher number of worm in a typically infected European eel. In comparison in the Japanese eel worms are smaller and intensities of infection are much lower. The dark brown matter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

1.1.2 Evolutionary significance

1.1.2.1 The eel-host

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

1.1 The study organism: *Anguillicola crassus*

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

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

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

A decline of epidemiological parameters for European populations of *A. crassus* has been hypothesized based on data published over two decades. This decline however, has not been confirmed in an explicit meta-analysis. If it would be present, possible

1. INTRODUCTION

explanations would include lower population density of the eel (likely (32)), an evolution of the eel host towards better resistance (rather unlikely; see above), and an evolution of *A. crassus* towards lower or at least altered virulence (part of the present investigation).

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

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

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

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

The recent anthropogenic host-switches of *A. crassus* from *An. japonica* to *An. anguilla* and *An. rostrata* constitute additional acquisitions of phylogenetically well separated hosts. This affinity for host-switching may be an evolutionary relict found

1.1 The study organism: *Anguillicola crassus*

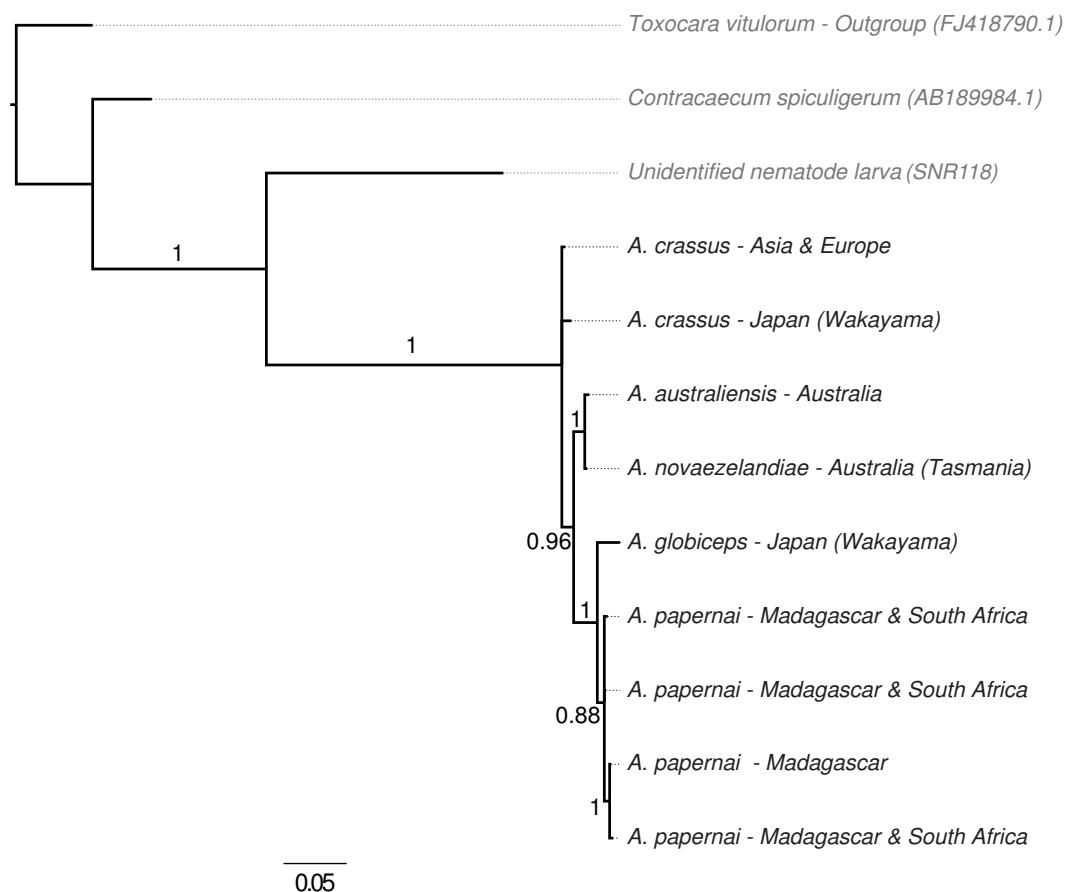
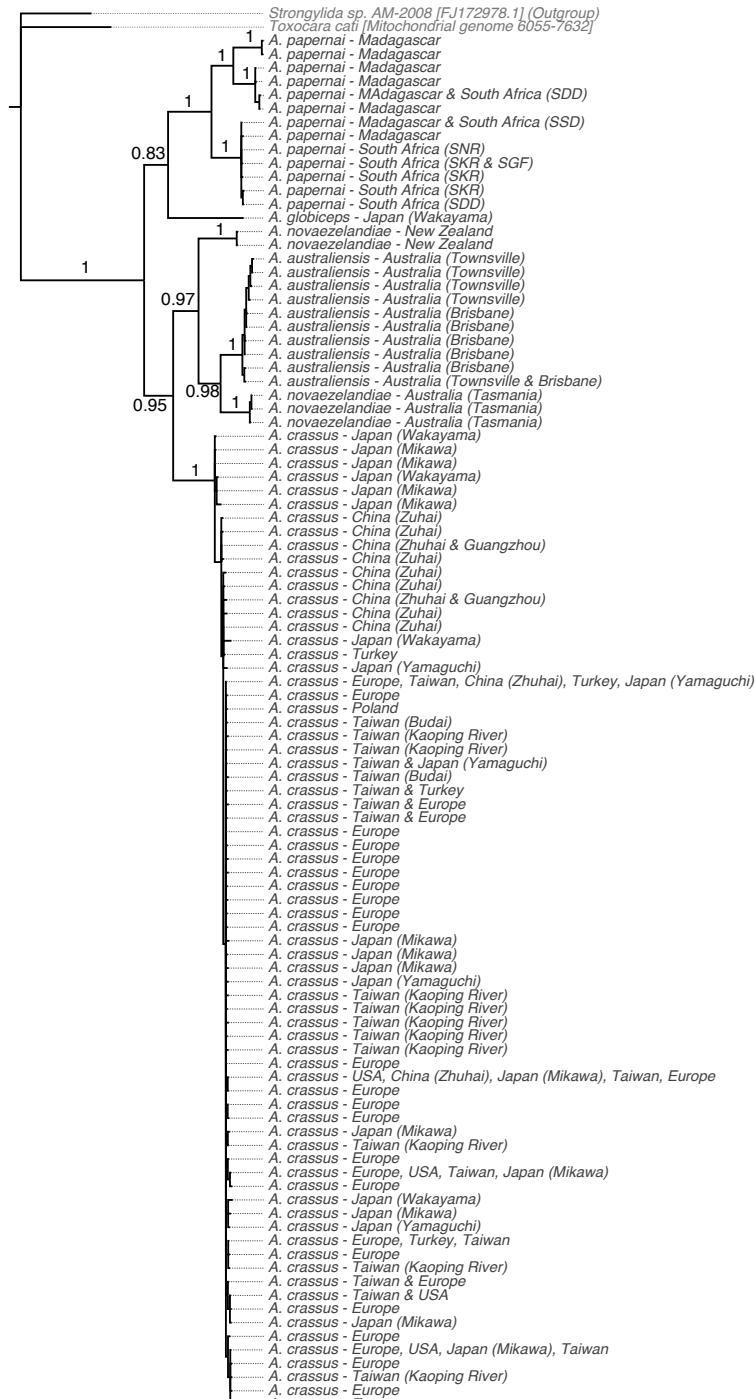


Figure 1.4: Phylogeny of the genus *Anguillicola* based nLSU - Phylogram inferred from nuclear large ribosomal subunit (nLSU) of *Anguillicola* and outgroups using Bayesian Inference. Labels on internal branches indicate Bayesian posterior probabilities. From Laetsch et al. CITE!!

1. INTRODUCTION



03

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

1.1 The study organism: *Anguillicola crassus*

only in one of the two clades separating *A. crassus* on a morphological sub-species level !!CITE Laetsch.

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

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

This phylogenetic position makes the Anguillicolidae an interesting system as out-group taxa to understand the evolution of parasitic phenotypes in the Spirurina. The recent anthropogenic expansion of *A. crassus* to new host species provides the opportunity to observe phenotypic modifications and possible also changes in an early genetic divergence.

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

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

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

1. INTRODUCTION

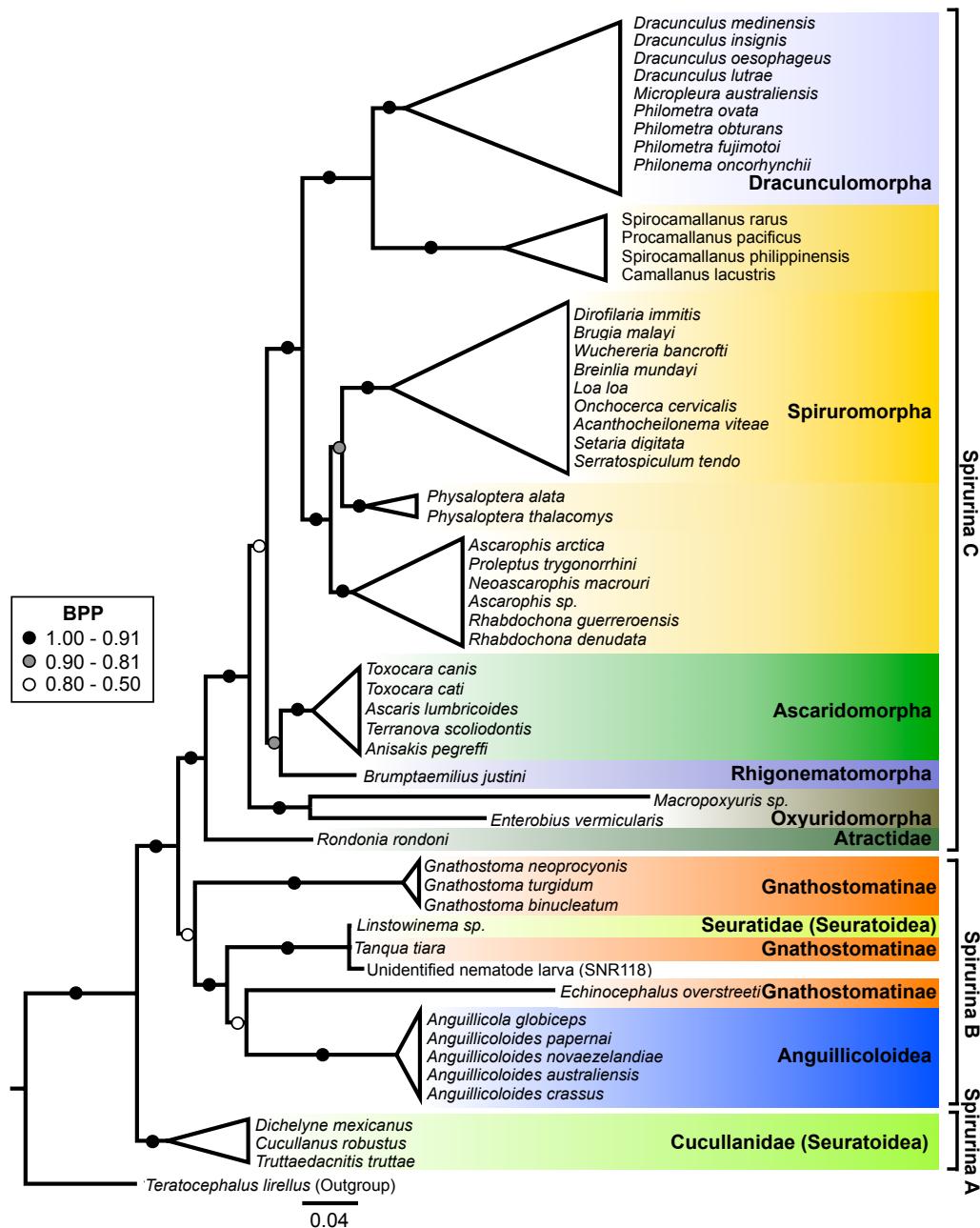


Figure 1.6: Phylogeny of nematode clade III based on nuclear small ribosomal subunit - Phylogram inferred from nuclear small ribosomal subunit for Spirurina using Bayesian Inference. Branches are collapsed to highlight major groups. Labels on internal branches indicate Bayesian posterior probabilities.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

1.1 The study organism: *Anguillicola crassus*

by environment (GxE) interactions. In other words: there are situations in which the differences in genotypes could be visible only under special environmental conditions.

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

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

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

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

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

1. INTRODUCTION

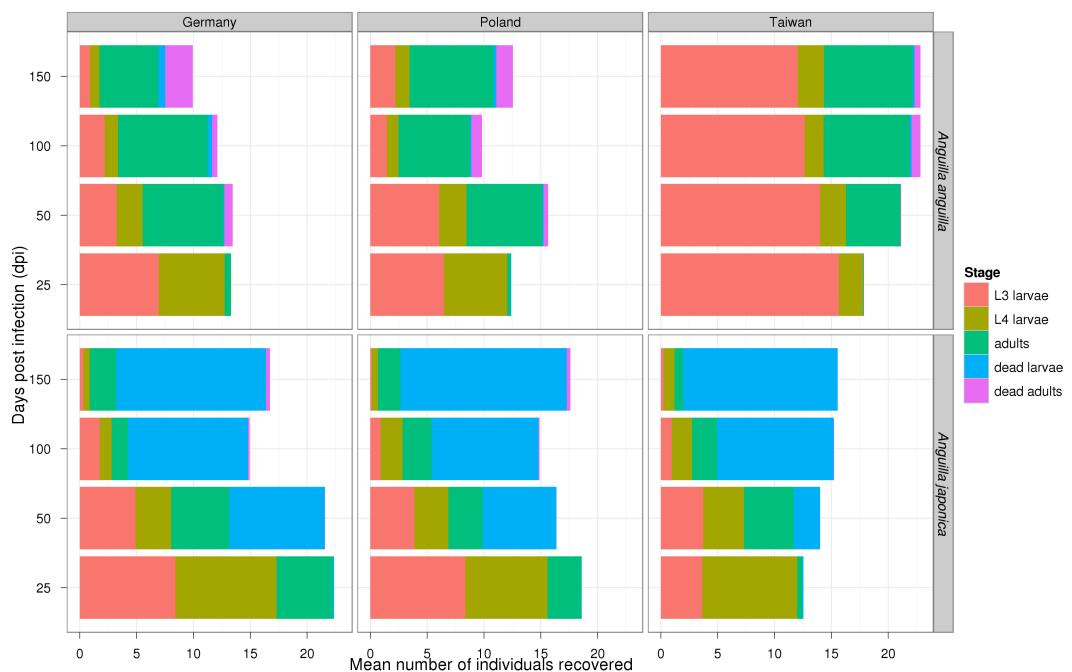


Figure 1.7: Differences in developmental speed - Three populations of *A. crassus* (panels in rows) were raised in two different hosts (panels in columns). Eels were dissected at 4 different time points post infection (dpi). Bars represent means of recovered individuals from three different life-cycle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

1.1 The study organism: *Anguillicola crassus*

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

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

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

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

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

1. INTRODUCTION

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

1.2 DNA sequencing

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

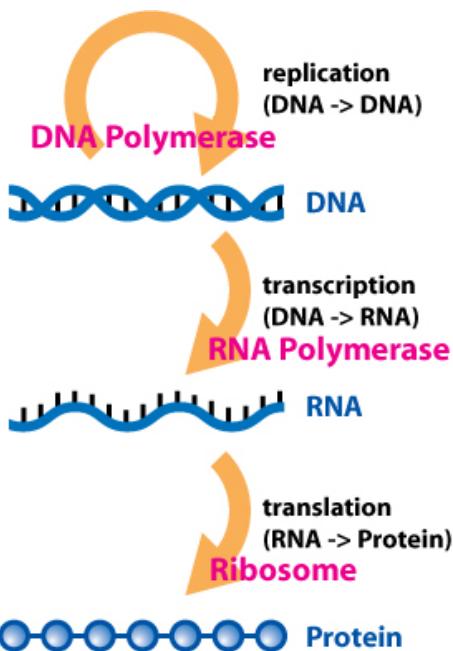


Figure 1.8: Major macromolecules bearing biological sequence information: - A schematic view on the flow of genetic information in a cellular life: Enzymes (red font) process macromolecules carrying genetic information from DNA to RNA, from RNA to protein. Picture from wikipedia.

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

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

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

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

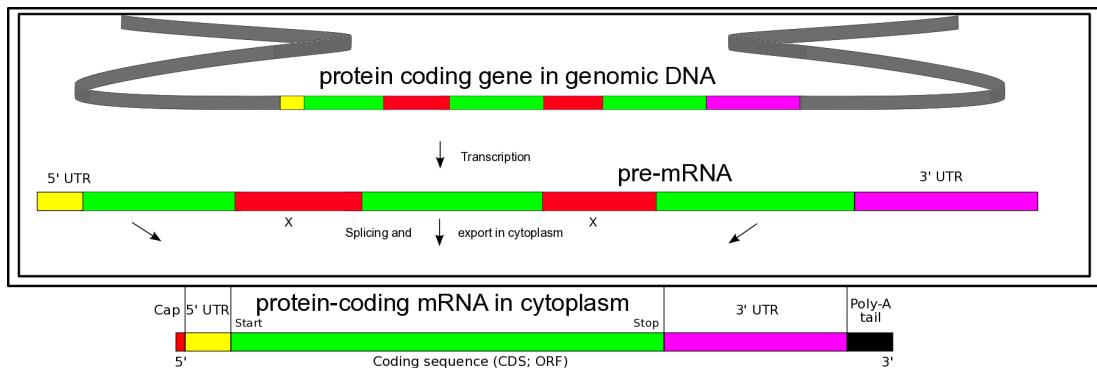


Figure 1.9: The structure of a protein coding gene and its mRNA - A schematic view of posttranscriptional modifications in an eukaryotic gene. Introns are spliced, 5' and 3' structures are added and the mRNA molecule is exported into the cytoplasm. Note that the double stranded nature of the genomic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

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

1. INTRODUCTION

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

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

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

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

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

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

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

In parallel to the mentioned genome-project transcriptome projects were needed and used to predict genes and to extrapolate their number (102).

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

1.2.3 DNA-sequencing in Nematodes

In 1998 *Caenorhabditis elegans* had become the first multicellular organism with a sequenced genome (98). Soon it was noted, that in addition to 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 in the phylum nematoda (103). The breadth and

1. INTRODUCTION

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

The genome sequence of *Caenorhabditis elegans* was soon complemented by the genome of *Caenorhabditis briggsae* (105), a second nematode from the genus *Caenorhabditis* sequenced a satellite system for comparative genomics inside this genus. As a second satellite model in clade V the necromenic *Pristionchus pacificus* (living in close association with beetles) has a published draft genome (106).

The first published genome of a parasitic nematode in the Spirurina was the draft genome of *Brugia malayi* (107) and as a second genome in the Spirurina recently the genome of *Ascaris suum* has been published (108).

Also in the remaining clades of the nematoda genome sequencing flourished: For the animal-parasite *Trichinella spiralis* from clade I (109), the plant parasites *Meloidogyne incognita* (110) and *Meloidogyne hapla* (111) as well as the pinewood nematode *Bursaphelenchus xylophilus* (112) (a plant parasite using a beetle as an vector) from clade IV recently genome sequences have been analyzed and published.

The current revolution in sequencing methodology (see 1.2.4) brings into sight many more sequenced nematode genomes (including that of *A. crassus*). The 959 nematode genomes initiative promotes such sequencing of nematode genomes and makes working-drafts of genome-assemblies available for analytical purposes on a **blast**-server (113).

Before the advent of NGS the lack of genomic information in many species of nematodes promoted the use of ESTs as a tool for gene-discovery. Partial genomes *sensu* (114) were successfully interrogated for a large array of genes interesting for various scientific communities. In nematode parasites of vertebrates, pathogenic factors were described as potential vaccine candidates (115). 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 (115). 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 (116) (an anti-inflammatory, cellular response, thought to be non-effective against helminths). Further molecules involved in host-parasite interaction, which have been identified in transcriptome-projects include abundant larval transcripts of *B. malayi* (Bm-ALT) (117) and venom like allergens (Bm-VLA) (118).

In some of these studies secreted proteins were in the center of interest. They could potentially be excreted by the nematode to allow movement and food-uptake but also to interact with the host's immune system. The detection of signal-peptides for secretion using *in silico* analysis of ESTs has been used to highlight candidate genes for example in *Nippostrongylus brasiliensis* (119), and across all nematode ESTs (120).

Over the years sequence information derived Sanger-sequencing derived EST-data and whole genome sequencing has been collected and updated into the nembase transcriptome databases (121, 122). The recent compendium nembase4 describes clustering of 679,480 raw ESTs in 233,295 clusters from 62 species (123).

Obviously NGS also leaves its marks currently in nematode transcriptomics: While NGS analysis on the transcriptomes of *Ancylostoma caninum* (124), *Pristionchus pacificus* (125), *Litomosoides sigmodontis* (126) and *Ascaris suum* (127) have been published a recent review (128) lists 8 datasets for other species already available in public repositories.

1.2.4 Advances in sequencing technology

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

The technologies portrayed here and used in the work underlying this thesis can't work on single molecules and thus target molecules have to be amplified like in Sanger-sequencing. This amplification has to produce spatially separated templates. Immobilization on a solid surface to archive this clonal amplification is used in preparation of both pyrosequencing and for the illumina-platform (129). The detailed implementation

1. INTRODUCTION

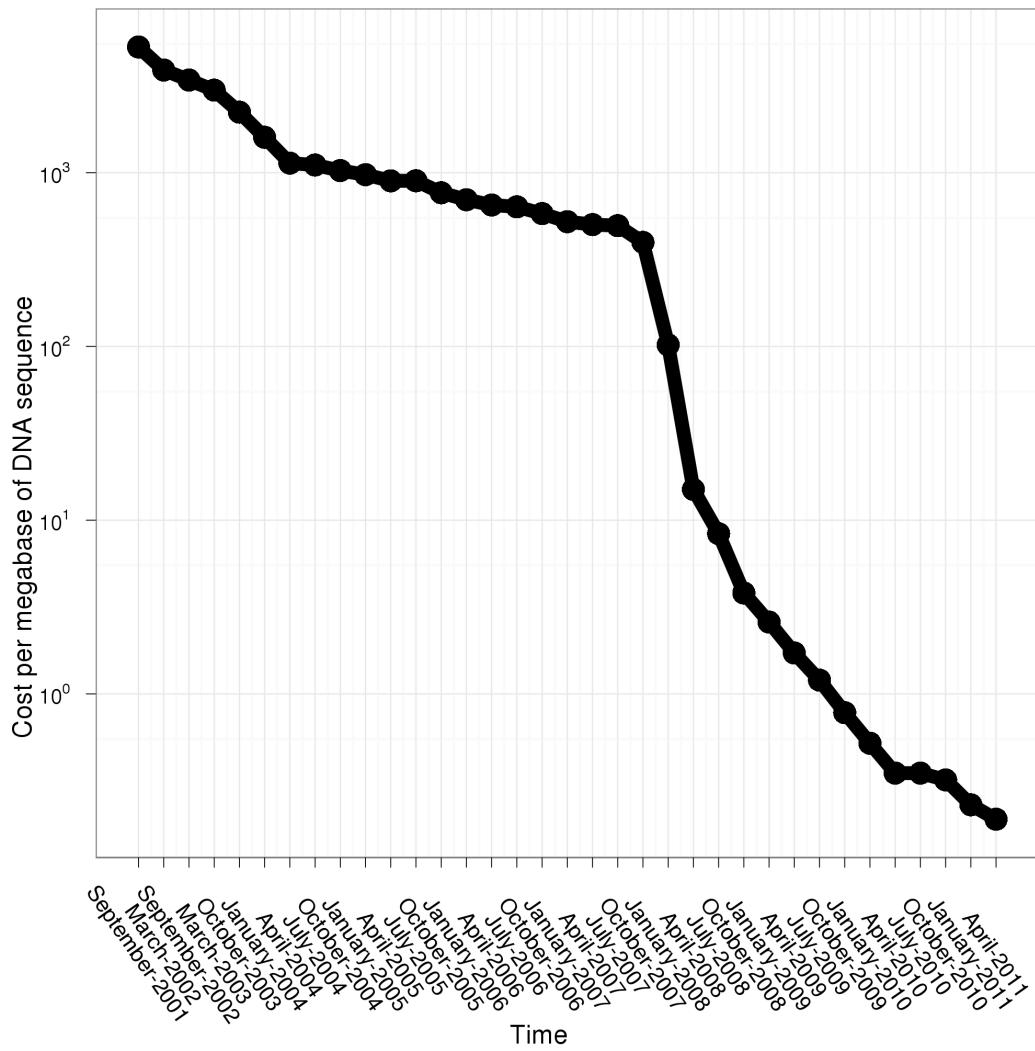


Figure 1.10: Falling sequencing costs - Sequencing costs falling due to advances in Solexa-sequencing: Due to improved read-length and data-volume on this platform per base sequencing-prices for many applications tumble into free fall. Data provided by National Human Genome Research Institute, NHGRI.

of this solid-state amplification in each technology differs and will be explained in the corresponding sub-chapter.

One cumbersome aspect of the need for amplification is the high amount of DNA starting-material ($3 - 20\mu\text{g}$) required (129). Other disadvantages include, that mutations during clonal amplification in templates can disguise error as sequence variants. Nucleotide composition of the target may also introduce amplification bias and thus biased product yield (130). This in turn leads to underrepresentation of certain molecules. The last point can be detrimental in quantitative applications, such as RNA-seq (90). However, while alternative single molecule approaches exist (eg. (131, 132)) and can be applied to address the above stated problems (133, 134), to date these technologies are in throughput and reliability not competitive for most real life applications.

1.2.4.1 Pyro-sequencing

Prior to pyrosequencing (or 454-sequencing; named by the company making it commercially available) an emulsion PCR is used to clonally amplify DNA molecules attached to beads (figure 1.11): After fragmentation by mechanical shearing or ultrasound (136) (see figure 1.11), DNA is ligated to adapters, denatured and single stranded molecules are attached to complementary sequence on a bead. Emulsion of beads in oil together with enzymes under conditions that favor one bead per water/enzyme droplet allows PCR in micro-scale reactions. This covers each bead with multiple copies of one target molecule. The beads are then distributed over the wells of a fiber-optic slide, the so called picolitre plate. A single bead per well is covered with enzymes on the surface of smaller beads. These enzymes are used in the actual pyrosequencing reaction originally developed by Pål Nyrén in the 1990s (137). The release of inorganic PPi as a result of nucleotide incorporation by polymerase starts a cascade of enzymatic reactions. The released PPi is converted to ATP by ATP sulfurylase, providing energy for luciferase to oxidize luciferin and to generate light. The added nucleotide is known as nucleotides are flushed over the plate one at a time. A high resolution camera records the emission of light. The intensity of emitted light is proportional to the number of nucleotides incorporated.

The ability to distinguish length of homopolymeric runs of the same nucleotide decreases with the length of such homopolymer runs (138). Current “Titanium chemistry”

1. INTRODUCTION

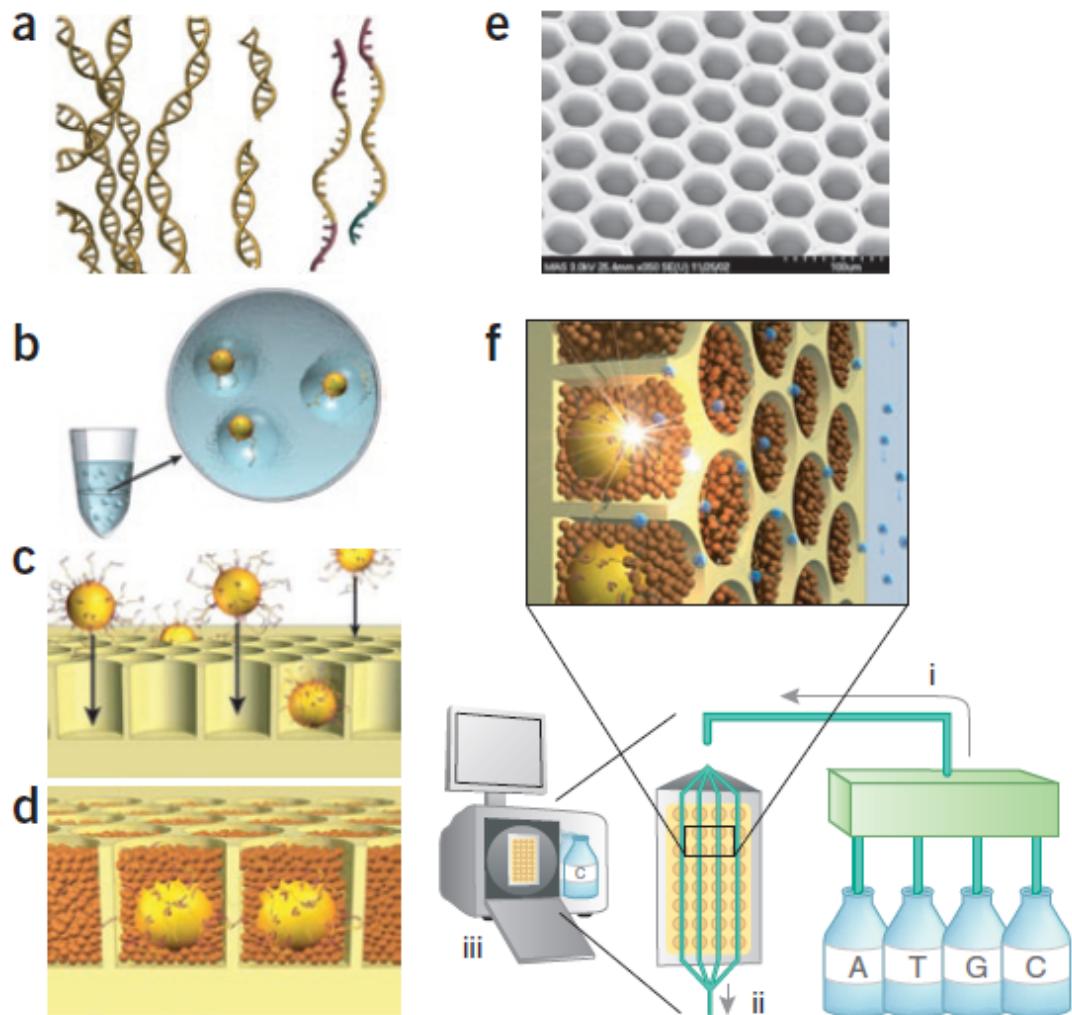


Figure 1.11: Schematic representation of pyrosequencing - (a) DNA (genomic or transcriptomic) is isolated, fragmented, ligated to adapters and denatured into single strands (b) Under conditions that favor one fragment per bead fragments are bound to beads. These beads are isolated and compartmentalized in the droplets of an emulsion and PCR (a mixture of reagents in oil). Within each droplet DNA is amplified, and beads are obtained which carrying millions of copies of a unique DNA template. (c) After denaturation of DNA, beads are deposited into wells of a fiber-optic slide (called picolitre plate). (d) Immobilized enzymes carried on smaller beads are added to each well and a solid phase pyrophosphate sequencing reaction is initiated. (e) A portion of a fiber-optic slide, in a scanning electron micrograph (prior to bead deposition) (f) Major subsystems of the 454 sequencing instrument: a fluidic assembly holding nucleotides separately (object i), the well-containing picolitre-plate in a flow cell (object ii), a CCD camera assembly and the user interface for instrument control (object iii) (135)

is producing read of > 350 bases length, “FLX chemistry” (used up to 2009) was able to produce reads of roughly 250 bases length (139).

This longer read length of 454-sequencing (140) compared to other NGS technologies (see 1.2.4.2), allows *de novo* assembly of Expressed Sequence Tags (ESTs) in organisms lacking previous genomic or transcriptomic data (for a comprehensive list of studies using this approach before Oct 2010 see (126)).

1.2.4.2 Illumina-Solexa sequencing

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

The Illumina-Solexa platform uses bridge-amplification to produce clonal copies of DNA molecules in clusters on a glass slide (figure 1.12): Fragmented, double-stranded DNA is therefore ligated to a pair of oligonucleotide-adapters in a forked configuration (the adapter-ends have non-complementary sequence). Two primers are used in an initial amplification and a double-stranded molecule with a different adapter on either end is produced. Denatured single-strands are then annealed to complementary adapters on the surface of a glass slide. Using the 3' end of the surface-bound oligonucleotide as a primer, a new strand is synthesized. Subsequently the adapter sequence at the 3' end of newly synthesized copied strand is bound to another surface-bound complementary oligonucleotide. This results in a bridge-structure and generation of a new priming-site for synthesis after denaturation. Multiple cycles of this kind of solid-state PCR result in growth of clusters on the surface of the glass-slide (141).

In the actual sequencing reaction these clusters are sequenced using a sequencing by synthesis technique: polymerase and all four nucleotides simultaneously are flushed over the class slide in successive cycles. To avoid incorporation of multiple nucleotides, “removable terminator”-nucleotides are used, which allow only incorporation of one nucleotide per strand pre cycle. These nucleotides are labeled each with a different removable fluorophore. Transient incorporation of a nucleotide is detected using a high resolution camera after laser-induced excitation. The fluorophore is removed and next cycle initiated (141).

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.

1. INTRODUCTION

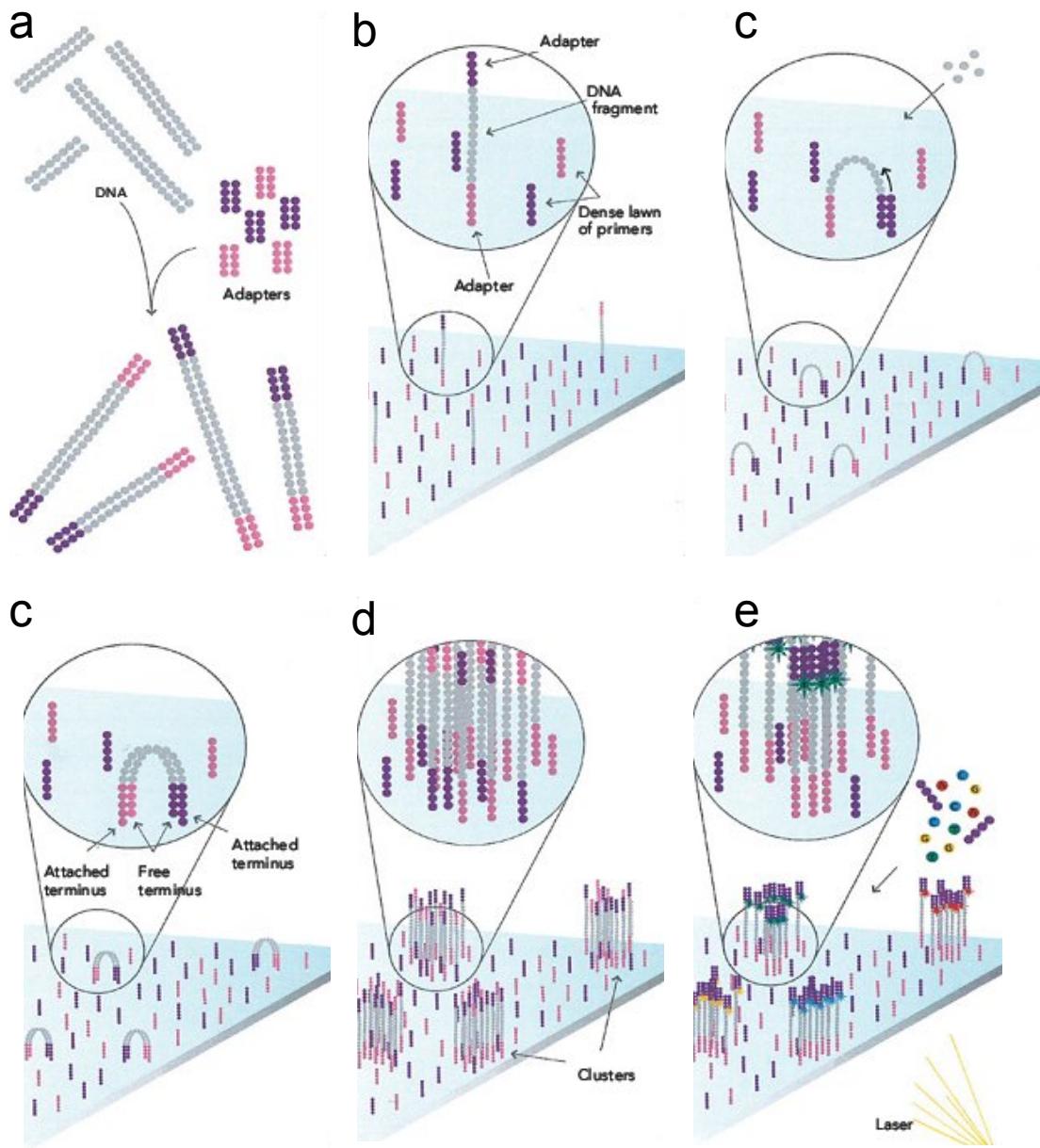


Figure 1.12: Schematic representation of illumina sequencing - (a) DNA (genomic or transcriptomic) is isolated, fragmented and ligated to adapters. (b) Single stranded fragments are bound to a glass-slide. (c-d) Solid-phase bridge amplification using unlabeled nucleotides, primers (binding the adapters)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

An slight alteration of the above method, which is extremely useful to inform assembly, is paired-end sequencing: After the first sequencing (as above), the original template strand is used to regenerate the complementary strand. This complementary strand then acts as a template for the second sequencing producing complementary sequence from the other end of the molecule. Using template molecules of a certain size range, sequence information can be obtained spanning 200-500 bases (the possible span of a nucleotide bridge in bridge-amplification) (141).

Additionally recent increases in read length (from 35 bases in 2008 to over 100 bases in 2011) are beginning to allow *de novo* sequencing and assembly of large eukaryotic genomes (e.g. that of the giant panda (142)) and transcriptomes (143) (but see also 1.2.5 for methodical challenges). In the same period throughput also increased from 6,000,000 reads in 2008 to 20,000,000 reads in 2011 per lane of the instrument.

The high throughput of the Illumina-Solexa platform makes it also first choice for gene expression analysis (144): RNA-seq has revolutionized transcriptomics both in model and non-model organisms (90). SuperSAGE (145) using expression-tags provides the benefit of classical SAGE-analysis (146) with those of the ultra heigh throughput of Illumina-Solexa sequencing.

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

This fostered use and development of new methods to assemble large-scale shotgun sequences, as higher coverage but shorter read-length (and also lower accuracy) are increasing the computational complexity of the assembly-problem (reviewed in (147)).

1.2.5 Computational methods in DNA-sequence analysis

In the context of computational tools another common characteristic of all DNA-sequencing methods has to be emphasized: Read-length is usually shorter than the length of the target molecule to be sequenced. This potential problem is solved by oversampling the target molecule, producing overlapping sequence. The amount of redundancy of the overlap is termed coverage (e.g. 10-fold coverage means a base is sequenced 10 times redundantly) the method as such is referred to as shotgun-sequencing and has - shortly after sequencing chemistry - been described by Sanger (148). Soon

1. INTRODUCTION

computer programs were necessary to align sequences, to compute overlaps and consensus sequences (149) and the process of computationally reconstructing the target molecule was termed sequence-assembly (150).

What is a contig...

The first step in this 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 (151) the second “Needleman-Wunsch” global alignment (152). Of course these alignment methods have usages outside of sequence assembly in general sequence comparison, including protein sequence.

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

With the advent of next generation sequencing (see 1.2.4) even the heuristic approach of **Blast** or it’s mapping equivalent **Blat** (153) 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.

Ssaha2 (154) is able to speed up sequence searches by orders of magnitude building a hash table indexing k-tuples (k contiguous bases, implicitly also done in the seeding step of **Blast/Blat**). Then sorting of matching indices gives regions of high similarity without an alignment. These are then aligned using a banded Smith-Waterman algorithm.

Burrows-Wheeler Aligner (BWA) (155) 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.

Obviously the above mentioned programs for sequence comparison or mapping only solve a subset of the problem, which *de novo* assembly constitutes. New algorithmic

approaches involve construction of a de Bruijn-graph. In most formulations of the 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 (156).

On top of this complexity de novo assembly of transcriptomes 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 reads may be covered uneven 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 a chimeric transcript. (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 (157).

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. However the power for the resolution of alternate splicing isoforms (d) is reduced (especially for lower-coverage transcripts). Recent versions of **gsAssembler** (also called **Newbler**; Roche/454) provide an opportunity to assess alternative splicing (158).

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.

1.2.6 Applications of NGS in ecology and evolution

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

A study on trout in Lake Superior (159) used an approach similar to the approach in the work presented here: Fish, which show two different phenotypes were raised in a common environment, demonstrating the genetic fixation of the phenotypic trait. 454 sequencing was then used to measure the gene expression levels and successfully identified 40 genes from two biochemical pathways being differently expressed. However, in addition to showing divergent evolution of gene-expression, this study highlighted the limitations of 454 sequencing for gene-expression analysis.

1. INTRODUCTION

Positive or diversifying selection on parasite proteins from the host-parasite interface can lead to a overabundance of non-synonymous changes (altering the protein sequence) over synonymous polymorphisms e.g. in *Plasmodium* (160).

1.3 Gene-expression and evolutionary divergence

Today, both theoretical arguments as well as field and laboratory data suggest that evolution, including divergence of populations, can occur very rapidly given the right selective pressure. Such situations provide us with the opportunity of examining how divergence and even speciation work at the molecular genetic level (161).

In *Drosophila* variation of gene-expression within a single species can be attributed more to trans-regulatory elements, while expression divergent between species is dominated by cis-regulatory differences (162). Furthermore sterility of hybrids between species of this genus has been shown to result from incompatibilities in gene-regulatory networks (163).

In parasites gene-expression is thought to evolve towards avoidance of co-expression: For polymorphism to be positively selected it requires the evolution of a regulator locus or the evolution of polymorphism is followed by the evolution of a regulator locus is (164).

Two virulence factor LbGAP in venom-producing tissues that the major virulence factor in the wasp *Leptopilina boulardi* differs only quantitatively. The regulation of gene expression might thus be major mechanism at the origin of intraspecific variation of virulence (165).

2

Aims of the project

2.1 Preliminary aims

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

2.2 Final aim

In a common garden environment

2. AIMS OF THE PROJECT

3

Pilot sequencing (Sanger method)

In preparation of high-throughput transcriptome sequencing of the swimbladder nematode *A. crassus* expressed sequence tags (ESTs) were generated using traditional Sanger-technology. 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 *Anguilla japonica* (3 libraries from 3 cDNA preparations) were sequenced.

Initial quality screening

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

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

rRNA screening

The further screening of sequences revealed a high abundance of rRNA (see Figure 3.1) ranging from 71.67% to 91.67% of obtained sequences. High abundances of rRNA were also found in the libraries from host liver tissue (see table 3.1), ranging from

3. PILOT SEQUENCING (SANGER METHOD)

71.67% to 77.42%. This contamination in libraries from both species was mainly responsible for a low amount of sequences being of sufficient quality for submission to NCBI-dbEST. At this point for the *A. japonica*-dataset 36 sequences were submitted to NCBI-dbEST under the Library Name “*Anguilla japonica* liver” and were assigned the accession LIBEST_027503.

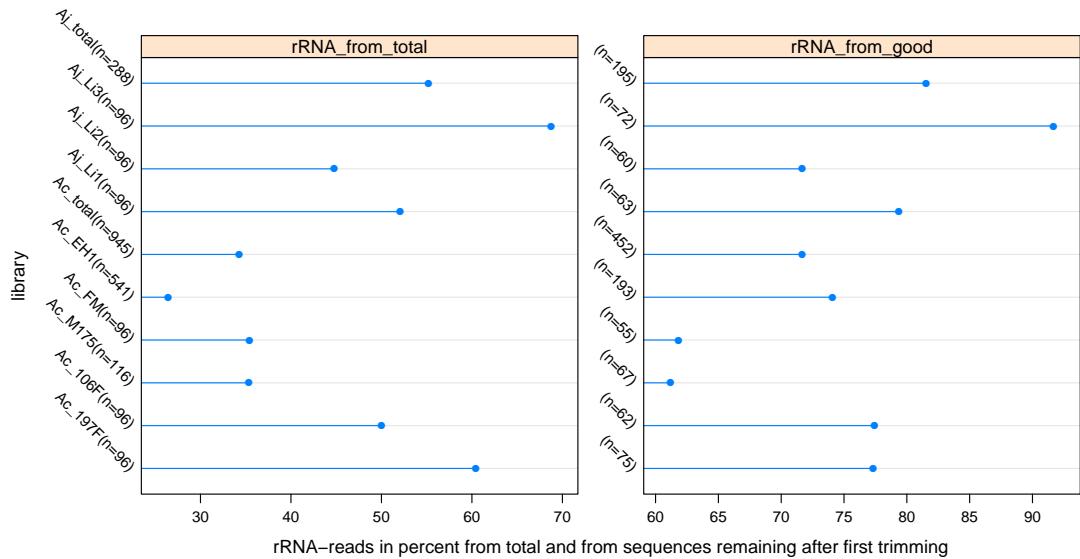


Figure 3.1: Proportion of rRNA in different libraries for *A. crassus* and *A. japonica* - rRNA abundance as proportion of the raw sequencing-reads (rRNA from total) and as proportion of the reads after quality screening (rRNA from good). Libraries starting with “Ac_” are from *A. crassus*, libraries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

Screening for host-contamination

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

	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.

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 it has to be noted that these 13 ESTs are still accessible through the same library name “Adult *Anguillilcola crassus*” and library-identifier LIBEST_027505 and are taxonomically attributed to *A. crassus* on NCBI-dbEST.

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

Blast-annotations obtained (by similarity searches against NCBI-nr, bit-score 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).

3. PILOT SEQUENCING (SANGER METHOD)

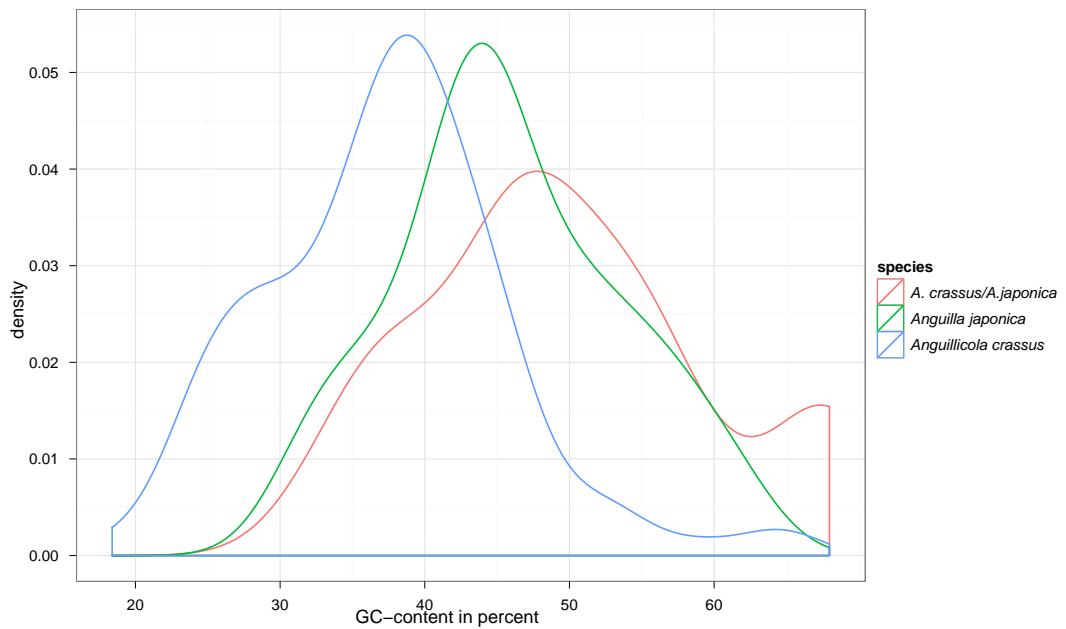


Figure 3.2: GC-content of sequences from *A. japonica* and *A. crassus* - The Japanese eel has a slightly higher GC-content than the parasite: This sequence characteristic is useful for separation of sequences from the host-parasite interface, note the higher GC-content of the sequences from *A. crassus*,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

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 “valid” i.e. not clearly host origin.

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

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

3. PILOT SEQUENCING (SANGER METHOD)

sequence	hit identifier	hit description	species	bit-score	e-value
Ac_EHif_005B07	gb AAQ97992.1	cyclin G1	<i>Danio rerio</i>	67.0	9e-10
Ac_EHif_01A02	gb ACO10003.1	Nicotinamide ribo- side kinase 2	<i>Osmerus mordax</i>	333	1e-89
Ac_EHif_01C10	gb ADF80517.1	ferritin M subunit	<i>Sciaenops ocellatus</i>	328	5e-88
Ac_EHir_004A04	ref XP_003340320.1	cytoplasmic actin	<i>Monodelphis domestica</i>	102	3e-20
Ac_EHir_005B07	gb ABN80454.1	cyclin G1	<i>Poecilia reticulata</i>	90.5	8e-17
Ac_EHir_009C03	ref NP_001122208.1	THAP domain containing protein 4	<i>Danio rerio</i>	176	1e-42
Ac_EHir_01A07	sp P80946.1	Hemoglobin subunit beta	<i>Anguilla anguilla</i>	283	1e-74
Ac_FMf_08F03	ref XP_003226802.1	cohesin subunit SA-2-like isoform 2	<i>Anolis carolinensis</i>	219	8e-56
Ac_MI75_01H02	emb CAQ87569.1	NKEF-B protein	<i>Plecoglossus altivelis</i>	365	3e-99
Ac_197FF_01E04	ref XP_002121150.1	CUB and sushi domain-containing protein 3	<i>Ciona intestinalis</i>	80.5	2e-13
Ac_EHif_01D07	ref XP_002606965.1	hypothetical protein	<i>Branchiostoma floridae</i>	82.8	3e-14
Ac_MI75_01B06	ref XP_422710.2	hypothetical protein	<i>Gallus gallus</i>	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

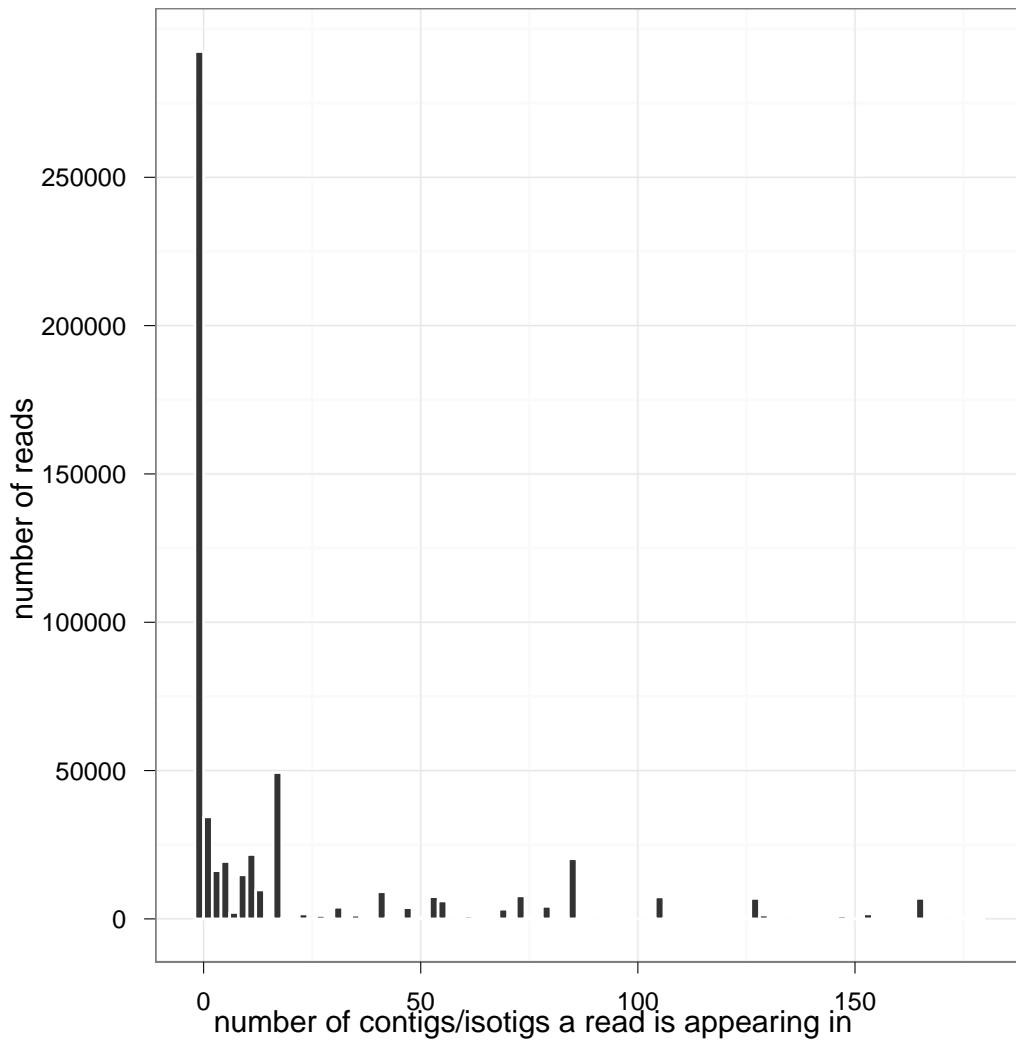
The pre-processed *Anguilllicola crassus* data-set consisting of 100491819 bases in 353055 reads (58617 generated using “FLX-chemistry”, 294438 using “Titanium-chemistry”) was assembled following an approach proposed by (126): Two assemblies were generated, one using `newbler v2.6` (140), the other using `mira v3.2.1` (166). The resulting assemblies (referred to as first-order assemblies) were merged with `Cap3` (167) into a combined assembly (referred to as second-order assembly).

4.2 The newbler first-order assembly

During transcriptome-assembly (with options `-cdna -urt`) `newbler` can split individual reads spanning the breakpoints of alternate isoforms, to assemble e.g. the first portion of the reads in one contig, the second portion in two different contigs. Later multiple so called isotigs would be constructed and reported, one for each putative transcript-variant. While this approach could be helpful for the detection of alternate isoforms, it also produces short contigs (especially at error-prone edges of high-coverage transcripts) when the building of isotigs fails. The read-status report and the assembly output in `ace`-format the program provides are including short contigs only used during the assembly-process, but not reported in the contigs-file used in transcriptome-assembly projects (`454Isotigs.fna`). Therefore to get all reads not included in contigs (i.e. a consistent definition of “singleton”) it was necessary to add all reads appearing only

4. EVALUATION OF AN ASSEMBLY STRATEGY FOR PYROSEQUENCING READS

in contigs not reported in the fasta-file to the reported singletons. The number of singletons increased in this step from the 26211 reported to 109052. We later also address the usefulness of `newbler`'s report vs. the expanded singleton-category, but for the meantime we define singletons as all reads not present in a given assembly.



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

4.1). If information would correspond to real isoforms it should be about an order of magnitude lower. This fact emphasises the need for further processing of the contigs. The maximal number of read-splits in a given contig and it's usefulness will be discussed later in greater detail.

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

The `mira` assembly (with options `-job=denovo,est,accurate,454`) 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 remainder of this text deals with the exploratory analysis of how information from both estimates of the transcriptome are integrated into the final second-order assembly.

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

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

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

4. EVALUATION OF AN ASSEMBLY STRATEGY FOR PYROSEQUENCING READS

4.4 Data-categories in the second-order assembly

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

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

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

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

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

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

4.4 Data-categories in the second-order assembly

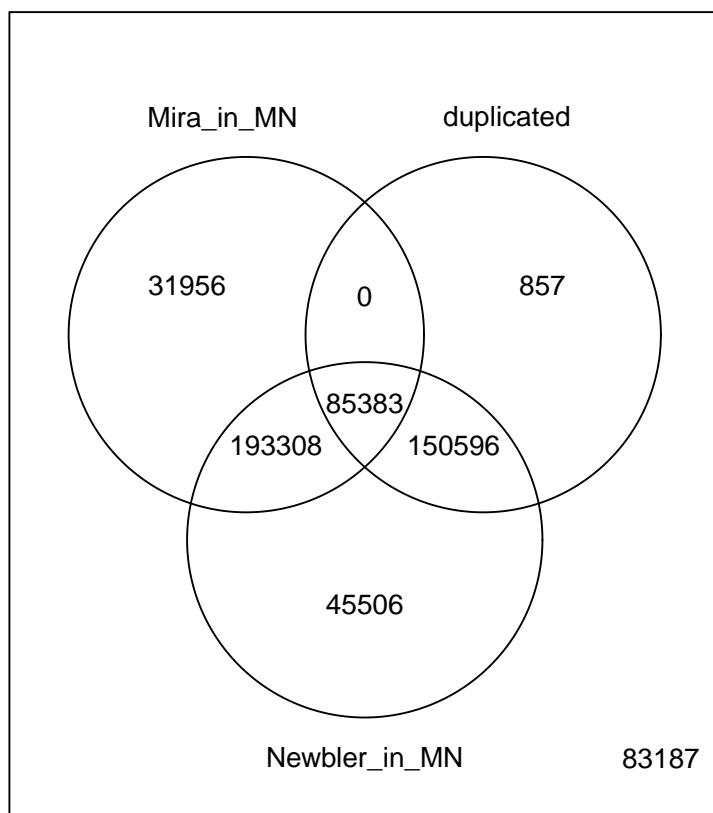


Figure 4.2: Origing of reads - Reads in the most reliable (MN) assembly-category are categorized by the way they entered the assembly: Although they are in a highly credible contig, reads can still have entered from only one first order assembly (Mira_in_MN or Newbler_in_MN). The intersection gives the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

4. EVALUATION OF AN ASSEMBLY STRATEGY FOR PYROSEQUENCING READS

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

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

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

4.6 Evaluation of the assemblies

To further compare assemblies (**mira**, **newber** first-order assemblies including or excluding their singletons) and the second-order assembly (including different contigs-categories and singletons) we evaluated the number of bases or proteins their contigs and singletons (partially) cover in the related model-nematodes, *Caenorhabditis elegans* and *Brugia malayi*. To this purpose we used **blast** (**blastx** e-value cut-off 1e-5) and a custom perl-script provided by S. Kumar.

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

The database-coverage for the two reference species can then be 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 on a second place and **mira** producing the lowest reference-coverage. When adding the contigs considered lower quality supported by only one assembler to the second-order assembly the reference-coverage increased moderately.

Including singletons the **mira** and **newber** assemblies were of increased size. A comparison of the **newbler**'s reported singletons with all singletons added to the **newbler**-assembly shows, that the reported singletons increased reference-coverage to the same

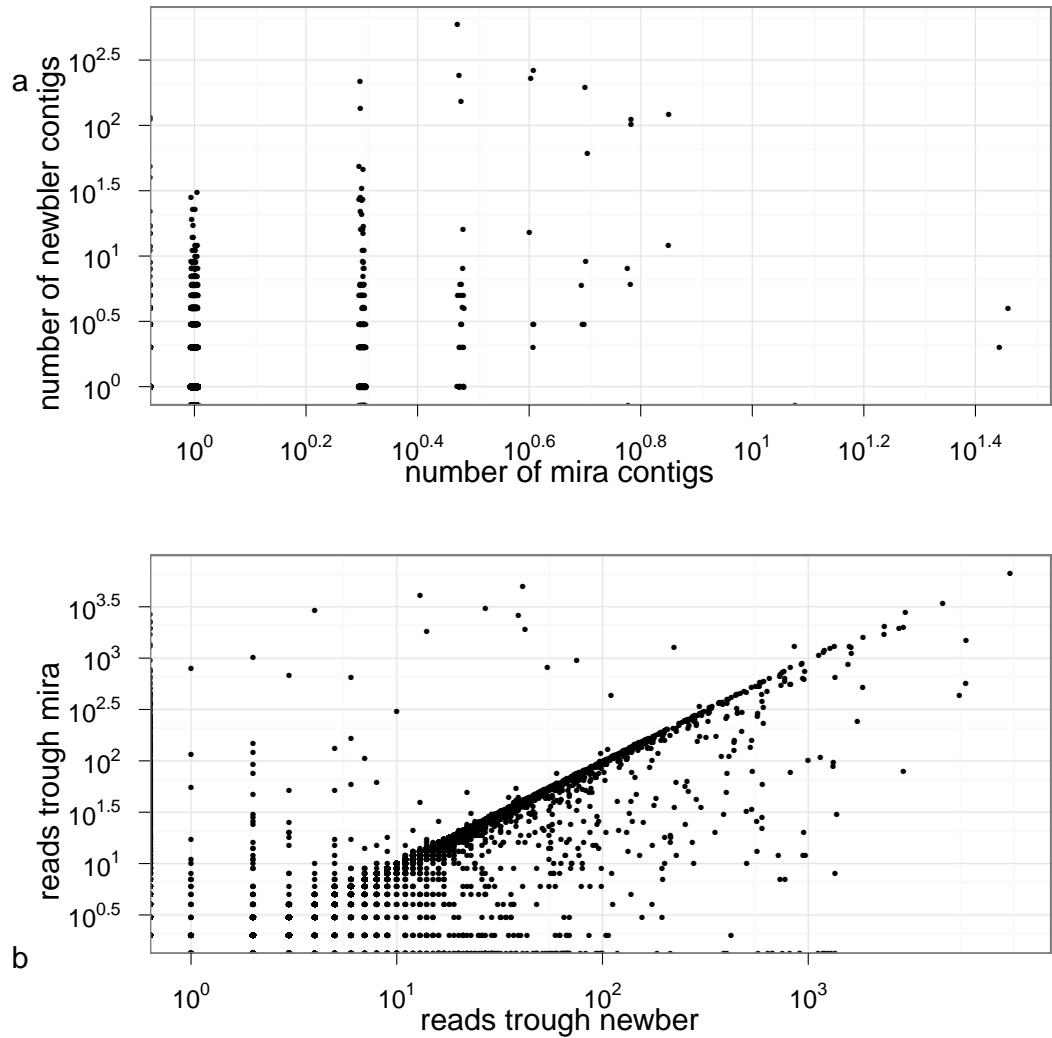


Figure 4.3: Contribution to second-order assembly - Number of first-order contigs from both first-order assemblies for each second order contig (a) number of reads through newbler and mira for each second-order contig (b)

4. EVALUATION OF AN ASSEMBLY STRATEGY FOR PYROSEQUENCING READS

amount than all singletons, while the non-reported singletons only increased the size of the assembly. It can be concluded, that the latter contain hardly any additional information but only error-prone or variant reads.

The second-order assembly including the intersection of first-order singletons performed similar to the **newbler** assembly for the number of bases 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 not the singleton of the second-order assembly (as defined in 4.2) but the intersection of **newbler**'s “reported singletons” and **mira**'s singletons were considered a very parsimonious assembly with high reference-coverage (termed fullest assembly; and labeled FU in the plots above) was obtained.

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

When database-proteins covered for at least to 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. We give most weight in our analysis to these results as in average longer correct contigs will allow finding the highest number of putative full-length genes.

Given this evaluation we defined a “minimal adequate” assembly as the subset of contigs of the second-order assembly supported by both assemblers (labeled NM above).

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

4.7 Measurements on second-order assembly

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

- number of **mira** and **newbler** first-order contigs

4.7 Measurements on second-order assembly

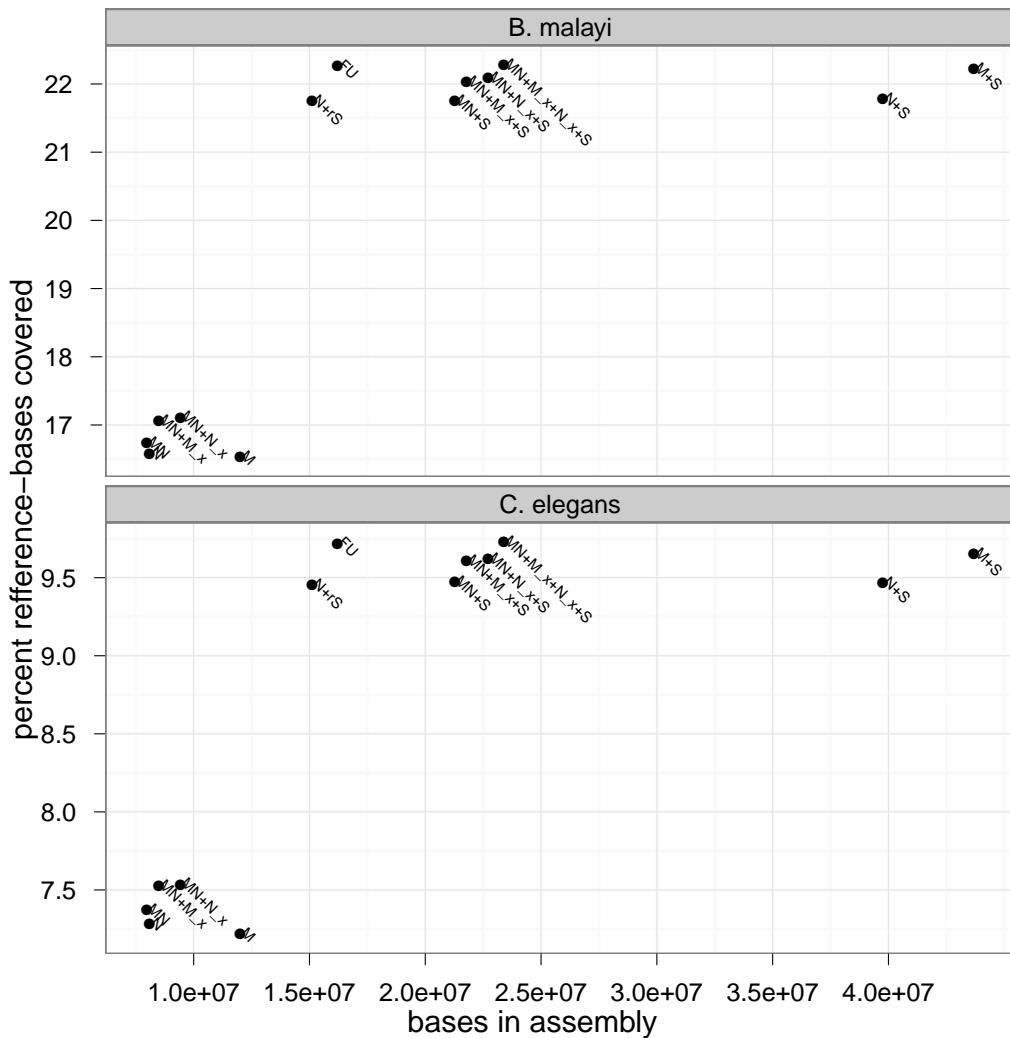


Figure 4.4: Base-content and reference-transcriptome coverage in percent of bases - for different assemblies and assembly-combinations; M = mira; N = newbler; M + S = mira + singletons; N + S = newbler plus singletons; N + Sr = newbler plus singletons reported in readstatus.txt; MN = second-order contigs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

4. EVALUATION OF AN ASSEMBLY STRATEGY FOR PYROSEQUENCING READS

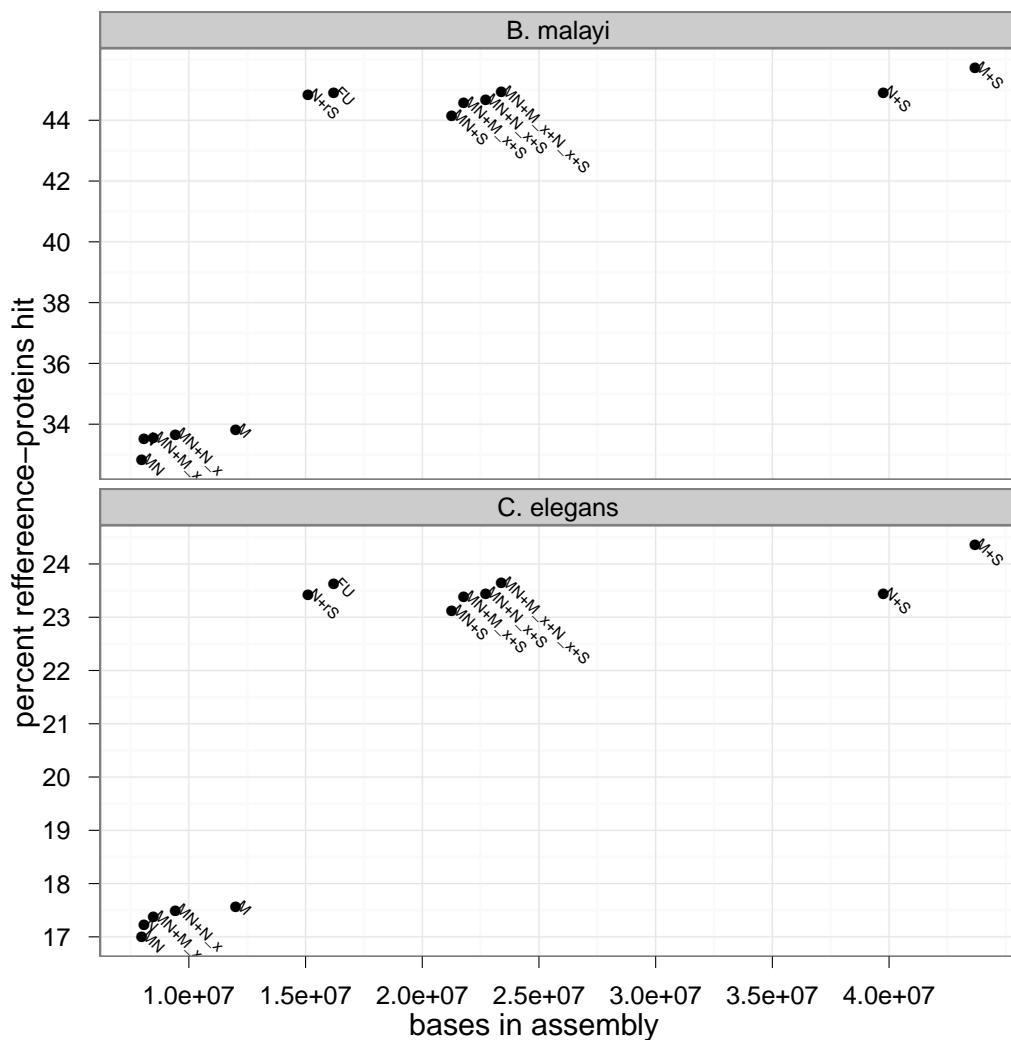


Figure 4.5: Base-content and reference-transcriptome coverage - in percent of proteins hit for different assemblies and assembly-combinations (for category-abbreviations see figure 4.4)

4.7 Measurements on second-order assembly

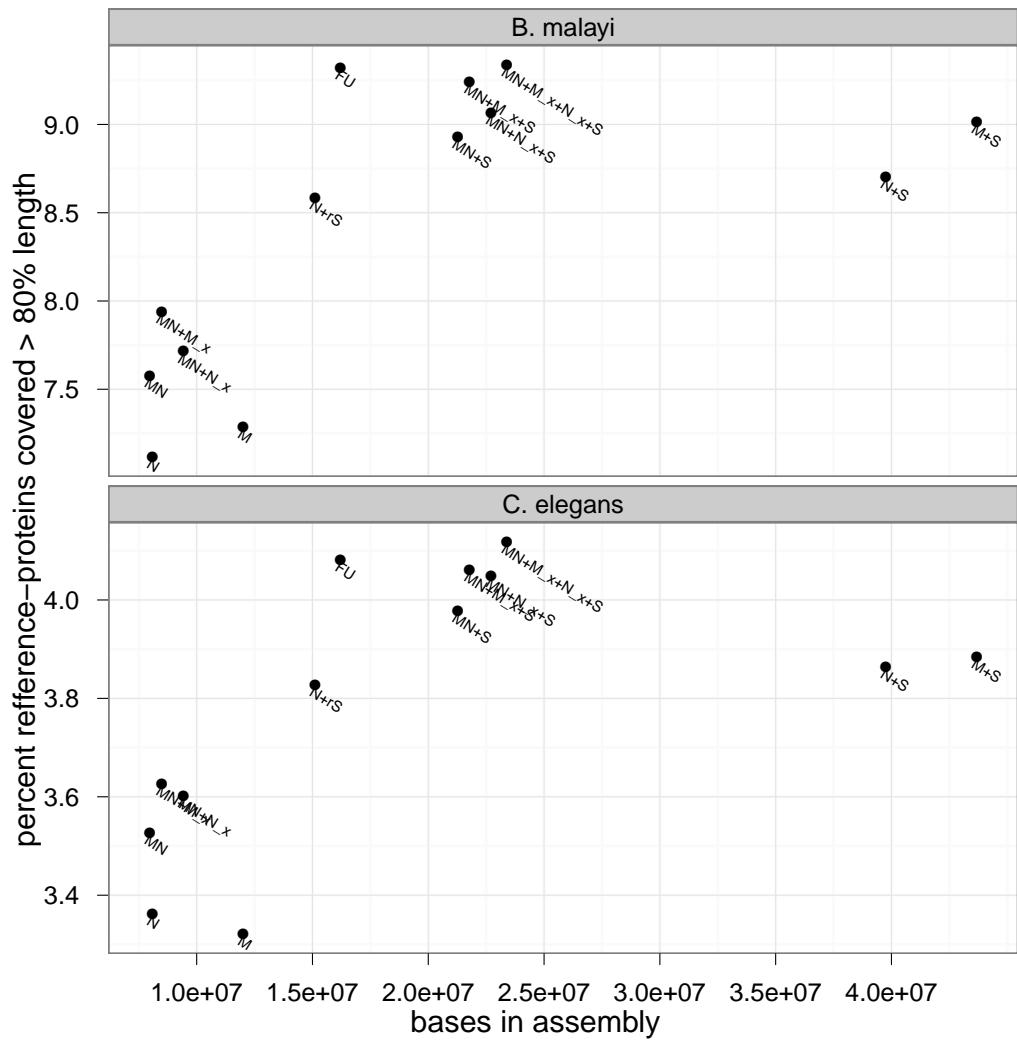


Figure 4.6: Base-content and reference-transcriptome coverage in percent of proteins covered to at least 80% - of their length for different assemblies and assembly-combinations (for category-abbreviations see figure 4.4)

4. EVALUATION OF AN ASSEMBLY STRATEGY FOR PYROSEQUENCING READS

- number of reads through `mira` and reads through `newbler`
- number of reads being split by `newbler` in first-order assembly
- number of read-split events in the first-order assembly (equals the sum of reads multiplied by number of contigs a read has been split into)
- maximal number of first-order contigs a read in the contig has been split into during `newbler`-assembly
- the number of reads same-read-paires from the `newbler` and `mira` first order-assembly merged in a second order contig
- cluster-id of the contig: All contigs “connected” by sharing reads (similar to the graph clustering reported in (158)).
- number of other second order contigs containing the same read (size of the cluster)

4.7.1 Contig coverage

As well defined coverage-information is not readyly avaiable from the output of this combined assembly aproach (although we followed individual reads through the process) we inferred coverage by mapping the reads used for assembly against the fullest assembly using ssaha2 (154) with parameters (-kmer 13 -skip 3 -seeds 6 -score 100 -cmatch 10 -ckmer 6 -output sam -best 1). We converted the `sam`-ouput via a sorted `bam`-file to `pileup`-format using `samtools` (168).

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

- mean per base coverage
- mean unique per base coverage

4.7.2 Example use of the contig-measurements

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

4.7 Measurements on second-order assembly

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

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

Contig1047 is in the well trusted MN category of contigs. It consists of only one contig from each first-order assembly (newbler_contigs and mira_contigs), each containing a set of reads of moderate size: 16 from `newbler` (reads_through_newbler) 26 from `mira` (reads_through_mira). 8 of the 16 reads `newbler` used in its one assembled contig were also assembled to a 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`-first order contigs.

A comprehensive interpretation of the other example-contigs depicted is left to the

4. EVALUATION OF AN ASSEMBLY STRATEGY FOR PYROSEQUENCING READS

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

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

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

4.8 Finalising the fullest assembly set

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

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

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

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

5

Pyrosequencing of the *A. crassus* transcriptome

454

5. PYROSEQUENCING OF THE *A. CRASSUS* TRANSCRIPTOME

6

Transcriptomic divergence in a common garden experiment

6.1 Infection experiments

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

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

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

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

6. TRANSCRIPTOMIC DIVERGENCE IN A COMMON GARDEN EXPERIMENT

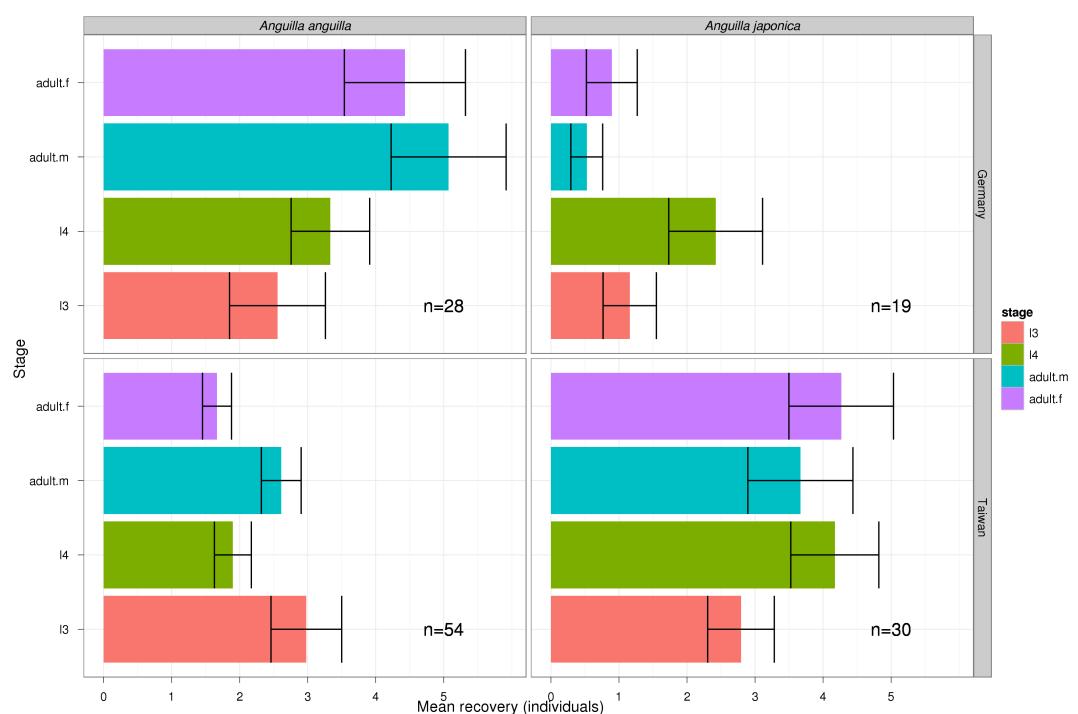


Figure 6.1: Recovery of worms in cross-infection experiment - Mean numbers of worms recovered after 55-57 dpi for sample sizes given as n=x. Error-bars indicate the standard error (s.e.) of the mean. Stages are L3-larvae (I3), L4-larvae (I4), adult females (adult.f) and adult males (adult.m).

6.2 Sample preparation and sequencing

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

Table 6.1: Linear model for recovery of adult worms. The estimate gives the mean of the distribution of adult worms for the factor values in the rows. The intercept is set to Aa R (*An. anguilla* and the European population) further rows give variations for each factor. Std. Error is the standard error of this value. 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 for each of the two sexes of worms. This resulted in a total 24 RNA-extractions prepared for sequencing: 3 individual female worms from each experimental group were 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 small exception from this form of statistical independence: From eel AJ/R3 a male worm as well as a female worm had to be prepared. It was impossible to extract enough RNA from all but the biggest male worms especially of the Japanese eel/European worm combination, leaving no other choice. Because of the small size of male worms it was generally not possible to randomly choose individuals. Preparation of sufficient amounts of RNA was only achieved in pools of the biggest individuals. All male worms were thus 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 over lanes in a blocking design to further guarantee statistical independence from sequencing-lane effects. Each pool of eight was sequenced on two lanes, giving in total six lanes of data and two technical replicates for each library. Sequencing resulted in a total of 263.668.952 raw sequencing read-pairs.

6. TRANSCRIPTOMIC DIVERGENCE IN A COMMON GARDEN EXPERIMENT

6.3 Examination of data-quality

Reads were mapped against the fullest assembly (see 4.8 and ??) using `BWA` (155). Of the 263.668.952 raw read-pairs 173.602.387 mapped uniquely to the assembly and were counted on a per-library base.

Read counts for one library were originally obtained from two different technical replicate lanes of a flow-cell. The technical replicate of read-counts demonstrated very low differences as inferred from a clustering analysis using variance stabilized data and euclidian distances as implemented in `DESeq` (196) (see ??).

After this analysis we summed up read-counts for technical replicates of each library. 158.232.523 read-pairs were left after removal of hits to contigs for which non-*A. crassus* origin had been inferred in the curation of the 454-transcriptome assembly.

After another screening for spurious read-counts to low covered transcripts and to transcripts of low reliability (lowCA in the 454-assembly; see ??) 137.477.156 read-pairs were left for further analysis. Distribution of these read-pairs over libraries showed roughly three 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).

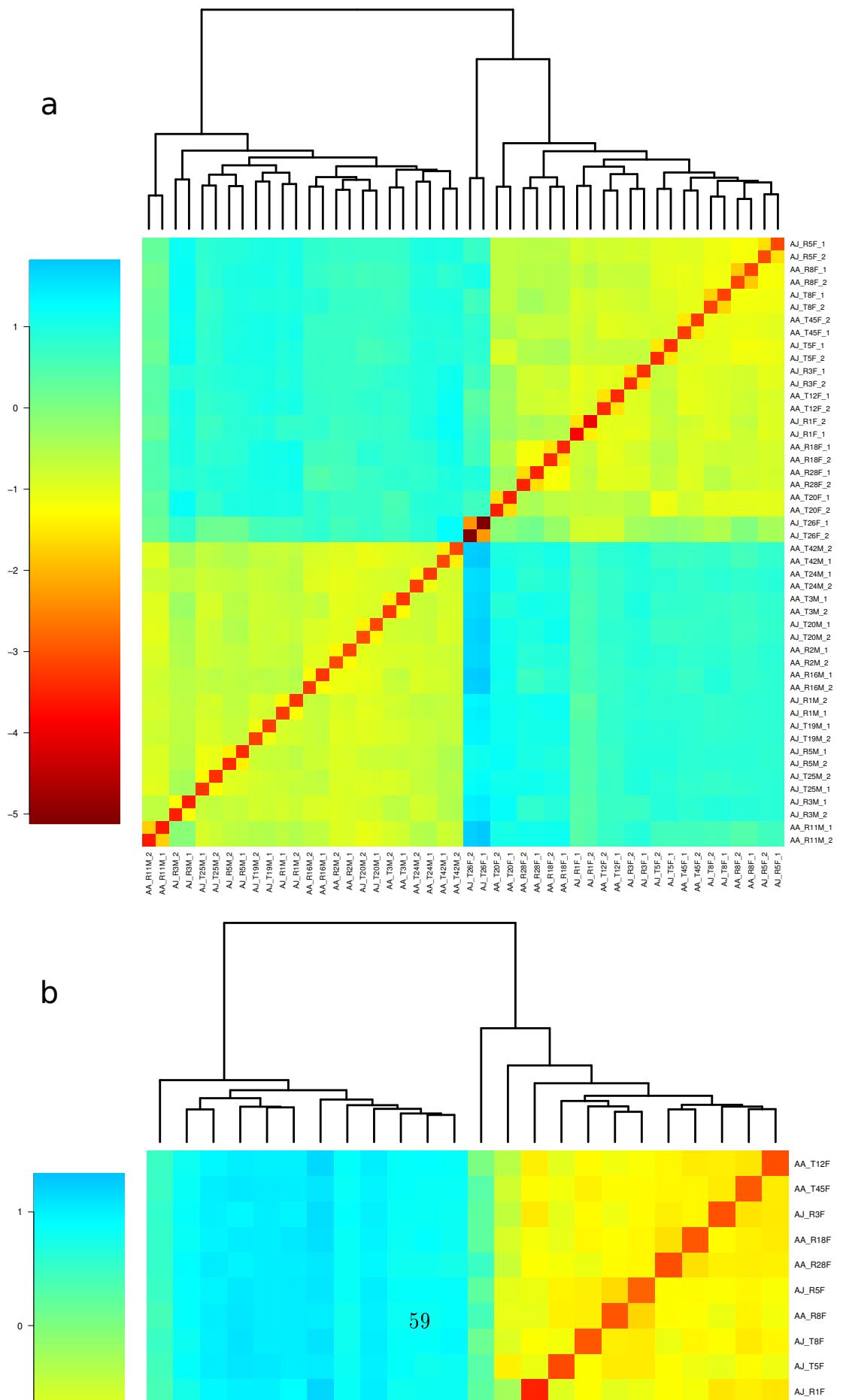
These reads mapped to xxx contigs from our 454 assembly, making them the basis for all further investigations.

6.4 Expression differences in general linear models

Generalized linear models (GLMs) were used as implemented in the R-package `edgeR`. Using these models we highlighted xxx contigs as differentially expressed between male and female worms at a false discovery rate of 5%. At the same threshold XX contigs showed significant differential response to the host-species, XX contigs according to the population of the worm.

In these models interaction-terms were prevalent. XX contigs showed differential expression according to

6.4 Expression differences in general linear models



6. TRANSCRIPTOMIC DIVERGENCE IN A COMMON GARDEN EXPERIMENT

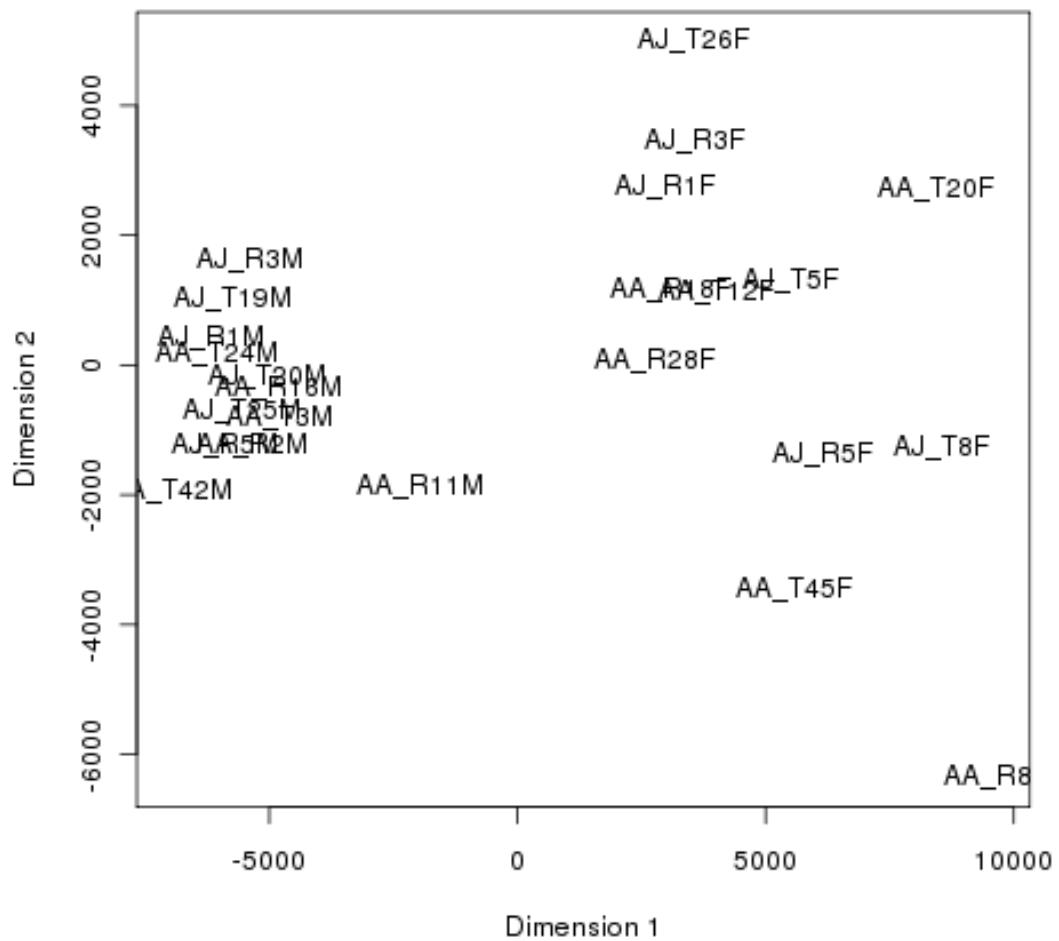


Figure 6.3: Principle coordinate plot for expression inRNA-seq libraries - Distance between sample-pairs is the root-mean-square deviation (Euclidean distance) for the most differentially expressed genes. Distances can be interpreted as the log₂-fold-change of the genes with the biggest changes, i.e.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

6.5 Expression differences between worms in European and Japanese Eels

6.4.1 Foo

6.5 Expression differences between worms in European and Japanese Eels

6.6 Expression differences between worms in the European and Taiwanese worm-population

6. TRANSCRIPTOMIC DIVERGENCE IN A COMMON GARDEN EXPERIMENT

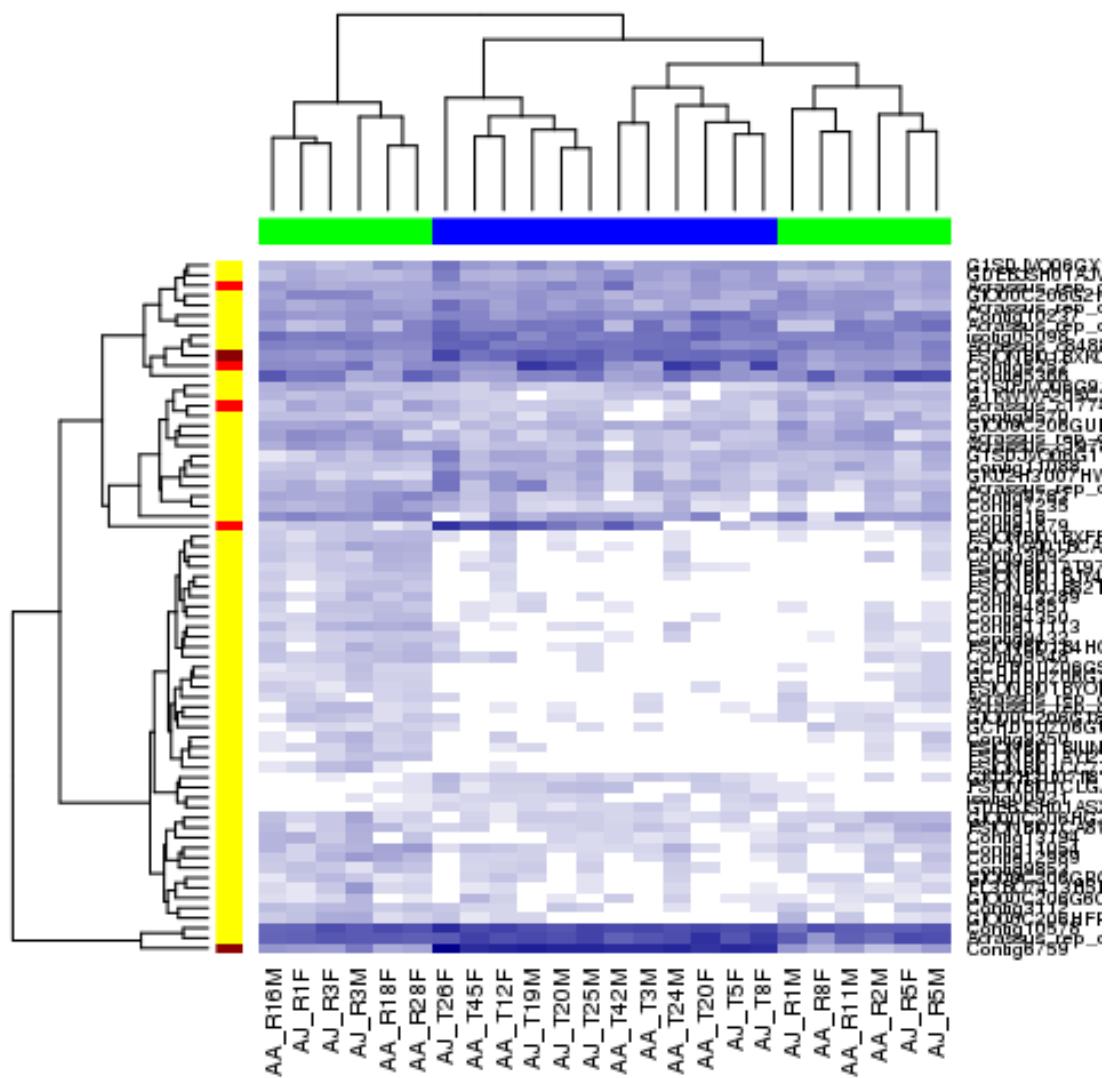


Figure 6.4: Variance/mean stabilized expression values for contigs different between populations - foobar

6.6 Expression differences between worms in the European and Taiwanese worm-population

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

Table 6.2: A summary of all 24 samples prepared for RNA-seq - The label of the RNA preparation follows a convention based on the eel species (host; first two letter of label, AA for *An. anguilla* AJ for *An. japonica*), worm population (population - R for European, T for Taiwanese; letter after the slash) and sex of worm(s) in preparation (F for female, M for male; last letter in label). Additionally the intensity of infection (number of adult worms found in the infected eel; intensity) and the number of worms pooled in the preparation (only male worms are pooled for RNA extraction, individual female worms were used). Finally RNA-concentration (conc in prep) in the preparation is given in μg per ml

6. TRANSCRIPTOMIC DIVERGENCE IN A COMMON GARDEN EXPERIMENT

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

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

7

Discussion

7.1 Sanger-method pilot-sequencing

In was not achieved to reproducibly alleviate the rRNA-levels in libraries prepared for sequencing.

This has probably been due to the fact that extraction of total-RNA from worms filled with host blood resulted in low amounts of starting material, and amplification using standard kits did not allow to reproducibly alleviate rRNA abundance. As the same problems existed in preparation of liver tissue of the host species it seems likely that the blood of eels contains substances limiting the success of specific amplification protocols. In fact it is known that compounds like hemoglobin can inhibit PCR reactions (169) and reverse transcription (170).

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

7.2 454-pyrosequencing

We are providing transcriptome-data for the parasite *A. crassus*, enabling a broad spectrum of molecular research on this ecologically and economically important species.

We emphasize the importance of screening for xenobiotics. We consider this aspect important in any deep transcriptome project. First the depth of sequencing is leading to the generation of large amounts of off-target data from a “metatranscriptomic community” associated with a target organism. Second due to the abundance of laboratory contamination and the possibility of cross-contamination if libraries are sequenced only

7. DISCUSSION

on a subset of a picotiter-plate (i.e. without the use of barcodes distinguishing between samples (171)) non-biological contamination can be introduced. However, in the context of a parasite (or an infected host) the screening for off-target data and contamination becomes even more important: Correct inference of biological origin for a given contig constitutes a prerequisite for most downstream analysis or the interpretation of results.

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

For the remaining off-target and contamination problem we archived separation of sequences in two steps, one before assembly, one afterward. Both screening-steps had to rely solely on sequence comparison. The screening-step before assembly has to employ lower stringency as sequence comparisons on sequence as short as reads are less informative than on longer contig-sequence. In our case of *A. crassus*, neither of the two host species has genomic data available for use in similarity searches. A publicly available transcriptome-data-set for European eel (172) in addition to a unpublished data-set for the same species was augmented with a data-set generated from the Japanese eel sequenced for the purpose of screening *A. crassus*-sequences in the present project. The pre-assembly screening had the rationale of facilitating the assembly process reducing the amount of divergent sequence from two host-species and the amount of extensively covered rRNA sequence. In our sequencing we were not able to reproducibly alleviate the rRNA coverage. This has probably been due to the fact that extraction of total-RNA from worms filled with host blood resulted in low amounts of starting material, and amplification using standard kits did not allow to reproducibly alleviate rRNA abundance. As the same problems existed in preparation of liver tissue of the host species it seems likely that the blood of eels contains substances limiting the success of specific amplification protocols. In fact it is known that compounds like hemoglobin can inhibit PCR reactions (169) and reverse transcription (170).

Although raw reads with rRNA hits were screened out prior to assembly, it was still possible to gain insights from these off-target data, as we assembled and annotated screening databases. Some of the rRNA data especially from the L2 library showed high similarity to flagellate eukaryotes. It could be possibly derived from an unknown protist living in the swimbladder of eels (possibly as a commensal of *A. crassus*), from where the L2 larvae for RNA-preparation were washed out. This seems worth further investigation, especially as it has been controversial whether encapsulated objects in the swimbladder of eels could be attributed solely to *A. crassus* (44) or to opportunist coinfections.

7.3 Experimental infections

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

First the ratio of transitions to transversions (ti/tv) increased. Such an increase is explainable by the removal of “noise” associated with common homopolymer-errors (138). Assuming that errors would be independent of transversion-transition bias erroneous SNPs would have a ti/tv of 0.5 and thereby lower the overall value. Other explanations for these observations are hard to find so it can be concluded that removing noise from homopolymer sequencing-error ti/tv increases. The value of XXX XXXX outside, XXX inside ORFs) is in good agreement with the overall ti/tv of humans (2.16) or *Drosophila* (2.07 (173)).

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 XXX to XXXX after full screening. Similar to ti/tv it the most plausible explanation is the removal of error, as unbiased error would lead to a dn/ds of 1. While dn/ds is not unproblematic to interpret within populations (174), assuming negative (purifying) selection on most protein-coding genes lower values seem more plausible, also in comparison with other studies (see further text).

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

The benefit of this screening was mainly a reduction of non-synonymous SNPs in high coverage contigs. When it was applied dn/ds did not scale with coverage. Working with an estimate of dn/ds independent of coverage, efforts to control for sampling a biased by sampling depth (i.e. coverage) like developed (175) and used (176) could be avoided.

7.3 Experimental infections

Such experiments have their problems because environmental factors, such as the general quality of the environment (i.e. water temperature) can interact with the host-environment (75).

7. DISCUSSION

8

Materials & methods

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

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

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

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

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

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

8. MATERIALS & METHODS

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 10 μ M. The protocol for PCR cycling is shown

Initial denaturation	94 °C	5min
Denaturation	94 °C	30s
Annealing	54 °C	45s
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 3.5 μ M 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

8.4 Pilot Sanger-sequencing

Ac_106F: Female from Taiwanese aquaculture

Ac_M175: Male from Taiwanese aquaculture

Ac_FM: Female from Taiwanese aquaculture

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

For *Anguilla japonica* the following three libraries:

Aj_Li1: liver of an eel from aquaculture

Aj_Li2: liver of an eel from aquaculture

Aj_Li3: liver of an eel from aquaculture

8.4 Pilot Sanger-sequencing

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

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

The last field indicates the plate number (two digits), the row (one letter) and the column (two digits) of the corresponding clone. For first quality trimming trace2seq, a tool derived from trace2dbEST (both part of PartiGene (114)) was used, briefly it performs quality trimming using phred(177) and trimming of vector sequences using cross-match(178). 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

8. MATERIALS & METHODS

expression used for this was:

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

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

Sequences were screened for host contamination by a comparison of BLAST searches against the version of nempep4 and a fish protein database. Sequences producing better bit scores 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.

8.5 454-pyro-sequencing

cDNA preparation and sequencing

RNA was extracted from individual adult male and female nematodes and from a population of L2 larvae (Table 1). RNA was reverse transcribed and amplified into cDNA using the MINT-cDNA synthesis kit (Evrogen, Moscow, Russia). For host contamination screening a liver-sample from an uninfected *An. japonica* was also processed. Emulsion PCR was performed for each cDNA library according to the 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, we confirmed that these data were not laboratory contaminants by screening Roche 454 data produced on the same run in independent sequencing lanes.

Trimming, quality control and assembly

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

kit using `cross-match` (178) (with parameters `-minscore 20 -minmatch 10`). `Seqclean` (179) was used to identify and remove poly-A-tails, low quality, repetitive and short (<100 base) sequences. All reads were compared to a set of screening databases using `BLAST` (expect value cutoff $E < 1e-5$, low complexity filtering turned off: `-F F`). The databases used were (a) a host sequence database comprising an assembly of the *An. japonica* Roche 454 data, a unpublished assembly of *An. anguilla* Sanger dideoxy sequenced expressed sequence tags (made available to us by Gordon Cramb, University of St Andrews) and transcripts from EelBase (172) a publicly available transcriptome database for the European eel; (b) a database of ribosomal RNA (rRNA) sequences from eel species derived from our Roche 454 data and EMBL-Bank; and (c) a database of rRNA sequences identified in our *A. crassus* data by comparing the reads to known nematode rRNAs from EMBL-Bank. This last database notably also contained xenobiont rRNA sequences. Reads with matches to one of these databases over more than 80% of their length and with greater than 95% identity were removed from the dataset. Screening and trimming information was written back into sff-format using `sfffile` (Roche 454). The filtered and trimmed data were assembled using the combined assembly approach (126): Two assemblies were generated, one using `Newbler v2.6` (140) (with parameters `-cdna -urt`), the other using `Mira v3.2.1` (166) (with parameters `-job=denovo,est,accurate,454`). The resulting two assemblies were combined into one using `Cap3` (167) at default settings and contigs were labeled by whether they derived from both assemblies or one assembly only (for a detailed analysis of the assembly categories see the supporting Methods file).

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 (121, 123). For each contig, the highest-scoring match was recorded as long as it spanned more than 50% of the contig. We also compared the contigs to the NCBI non-redundant nucleotide (NCBI-nt) and protein (NCBI-nr) databases, recording the taxonomy of all best matches with expect values better than $1e-05$. TUGs with a best hit to non-Metazoans and to Chordata within Metazoa were additionally excluded from further analysis.

8. MATERIALS & METHODS

Protein prediction and annotation

Protein translations were predicted from the contigs using **prot4EST** (version 3.0b) (180). Proteins were predicted either by joining single high scoring segment pairs (HSPs) from a **BLAST** search of uniref100 (181), or by **ESTscan** (182), using as training data the *Brugia malayi* complete proteome back-translated using a codon usage table derived from the **BLAST** HSPs, or, if the first two methods failed, simply the longest ORF in the contig. For contigs where the protein prediction required insertion or deletion of bases in the original sequence, we also imputed an edited sequence for each affected contig. Annotations with Gene Ontology (GO), Enzyme Commission (EC) and Kyoto Encyclopedia of Genes and Genomes (KEGG) terms were inferred for these proteins using **Annot8r** (version 1.1.1) (183), using the annotated sequences available in uniref100 (181). 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** (184) 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** (185) (command line utility **iprscan** (version 4.6) with options **-cli -format raw -iprlookup -seqtype p -goterms**) was used to obtain domain based annotations for the high credibility assembly (highCA) derived contigs.

We recorded the presence of a lethal rnai-phenotype in the *C. elegans* ortholog of each TUG using the biomart-interface (186) to wormbase v. 220 through the R-package **biomaRt** (187).

Single nucleotide polymorphism analysis

We mapped the raw reads against the the complete set of contigs, replacing imputed sequences for originals where relevant, using **ssaha2** (154) (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** (168), discarding reads mapping to multiple regions. **VarScan** (188) (**pileup2snp**) was used with default parameters on pileup-files to output lists of single nucleotide polymorphisms (SNPs) and their locations. For enrichment analysis of GO-terms we used the R-package **GOstats** (189).

Using **Samtools** (168) (**mpileup -u**) and **Vcftools** (190) (**view -gcv**) we genotyped individual libraries for the list of previously found overall SNPs. Genotype-calls were

8.6 Transcriptomic divergence in a common garden experiment

accepted at a phred-scaled genotype quality threshold of 10. In addition to the relative heterozygosity (number of homozygous sites/number of heterozygous sites) we used the R package `Rhh` (191) to calculate internal relatedness (192), homozygosity by loci (193) and standardized heterozygosity (194) from these data.

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

Gene-expression analysis

Read-counts were obtained from the `bam`-files generated also for genotyping using the R-package `Rsamtools` (195). TUGs with less than 48 reads over all libraries were excluded from analysis, as diagnostic plot (not shown) indicated a lack of statistical power for lower overall expression. We used the R-package `DESeq` (196) to assess statistical significance of differences in counts according to groups of libraries.

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

8.6 Transcriptomic divergence in a common garden experiment

8.6.1 Experimental infection of eels

An. anguilla were obtained from the Albe-Fishfarm in Haren-Raijtenbroek, 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 (197). During the infection period water temperature was held constant at 20°C. Eels were kept in 160-liter tanks in groups of 5-10 individuals and continuously provided

8. MATERIALS & METHODS

with fresh, oxygenated water and commercial fish pellets (Dan-Ex 2848, Dana Feed A/S Ltd, Horsens, Denmark).

L2 larvae used for the infection were collected from the swimbladders of wild yellow and silver eels from the River Rhine near Karlsruhe and from Lake Müggelsee near Berlin in Germany. Taiwanese larvae were obtained from aquaculture in Kao Ping and from 150km further north in Tainan County in Taiwan. They were stored at 4°C for no longer than 2 weeks before copepods were infected. Mixed species samples of uninfected copepods were collected from a small pond near Karlsruhe, known to be free of eels. They were infected individually in wells of micro-titer plates at an intensity of 10 L2-larvae per copepod. One week after infection they were placed in bigger tanks. Twice a week yeast was provided as food and at 21 dpi the L3 were harvested with a tissue potter using a modified procedure developed by (198). 50 L3 were suspended in 100 µl RPMI-1640 medium (Qiagen, Hilden, Germany) and eels were infected.

55-57 days post infection (dpi) eels were euthanized and dissected. After determination of the sex of adult worms under a binocular microscope (Semi 2000, Zeiss, Germany), they were immediately immersed in RNAlater (Qiagen, Hilden, Germany).

8.7 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 (Qiagen, Hilden, Germany) (see table XXX).

The Paired-End TruSeqTM RNA sample preparation kit (illumina) was followed to build cDNA libraries with insert sizes of roughly 270 bp for paired-end sequencing: Poly-A⁺ oligo-dT 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 2nd strand reaction, overhangs were repaired to form blunt ends, a single “A” nucleotide was added at the 3' end and paired end sequencing adapters were ligated with a complementary “T”-overhang. At this step multiple different indexed paired-end adapters were used to enable multiplexing of the 24 different sequencing libraries, in 3 pools of 8 samples each, on one lane of the instrument. Molecules having adapter sequences were enriched in the mix using PCR and the libraries were controlled 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 the first primer Rd1 SP. The original template strand

was then used to regenerate the complementary strand, the original strand was removed and complementary strand acted as a template for the second read sequenced primed by the second sequencing primer Rd1 SP.

8.8 Mapping and read-counts

8.9 Statistical analysis

8.10 Collapsing of orthologous for two model-species

8. MATERIALS & METHODS

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 crassus* sp. n. (Philometridae, Anguillicolidae).** *Japanese Journal for Parasitology*, 23(5):275–279, 1974. 1
- [2] B SURES, K KNOPF, AND H TARASCHEWSKI. **Development of *Anguillicola crassus* (Dracunculoidea, Anguillicolidae) in experimentally infected Balearic congers *Ariosoma balearicum* (Anguilloidea, Congridae).** *Diseases of Aquatic Organisms*, 39(1):75–8, December 1999. 1
- [3] H. TARASCHEWSKI. **Hosts and Parasites as Aliens.** *Journal of Helminthology*, 80(02):99–128, 2007. 1
- [4] R. S. KIRK. **The impact of *Anguillicola crassus* on European eels.** *Fisheries Management & Ecology*, 10(6):385–394, 2003. 1, 2
- [5] LAMIA GARGOURI BEN ABDALLAH AND FADHILA MAAMOURI. **Spatio-temporal dynamics of the nematode *Anguillicola crassus* in Northeast Tunisian lagoons.** *Comptes Rendus Biologies*, 329(10):785–789, October 2006. 1
- [6] ABDECHAHID LOUKILI AND DRISS BELGHYTI. **The dynamics of the nematode *Anguillicola crassus*, Kuvalaha 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 eels *Anguilla anguilla* in freshwater and marine habitats in Iceland in comparison with other parasite communities of eels in Europe.** *Folia Parasitologica*, 54(2):141, 2007. 1
- [8] K. KNOPF, J. WUERTZ, B. SURES, AND H. TARASCHEWSKI. **Impact of low water temperature on the development of *Anguillicola crassus* in the final host *Anguilla anguilla*.** *Diseases of Aquatic Organisms*, 33:143–149, 1998. 1
- [9] R. S. KIRK, C. R. KENNEDY, AND J. W. LEWIS. **Effect of salinity on hatching, survival and infectivity of *Anguillicola crassus* (Nematoda: Dracunculoidea) larvae.** *Diseases of Aquatic Organisms*, 40(3):211–8, April 2000. 1
- [10] SÉBASTIEN WIELGOSS, HORST TARASCHEWSKI, AXEL MEYER, AND THIERRY WIRTH. **Population structure of the parasitic nematode *Anguillicola crassus*, an invader of declining North Atlantic eel stocks.** *Molecular Ecology*, 17(15):3478–95, August 2008. 1, 2, 15
- [11] MÜNDERLE. **Ökologische, morphometrische und genetische Untersuchungen an Populationen des invasiven Schwimmblasen-Nematoden *Anguillicola crassus* aus Europa und Taiwan.** PhD thesis, University of Karlsruhe, 2005. 2, 4
- [12] PIERRE SASAL, HORST TARASCHEWSKI, PIERRE VALADE, HENRI GRONDIN, SÉBASTIEN WIELGOSS, AND FRANTIŠEK MORAVEC. **Parasite communities in eels of the Island of Reunion (Indian Ocean): a lesson in parasite introduction.** *Parasitology Research*, 102(6):1343–1350, May 2008. 2, 3
- [13] W NEUMANN. **Schwimmblasenparasit *Anguillicola* bei Aalen.** *Fischer und Teichwirt*, page 322, 1985. 2
- [14] H. KOOPS AND F. HARTMANN. **Anguillicola-infestations in Germany and in German eel imports.** *Journal of Applied Ichthyology*, 5(1):41–45, 1989. 2
- [15] S. WIELGOSS, F. HOLLANDT, T. WIRTH, AND A. MEYER. **Genetic signatures in an invasive parasite of *Anguilla anguilla* correlate with differential stock management.** *J. Fish Biol.*, 77:191–210, Jul 2010. 2
- [16] LT FRIES, DJ WILLIAMS, AND SKEN JOHNSON. **Occurrence of *Anguillicola crassus*, an exotic parasitic swim bladder nematode of eels, in the Southeastern United States.** *Transactions of the American Fisheries Society*, 125(5):794–797, 1996. 3
- [17] A. M. BARSE AND D. H. SECOR. **An exotic nematode parasite of the American eel.** *Fisheries*, 24(2):6–10, 1999. 3
- [18] ANN M. BARSE, SCOTT A. MCGUIRE, MELISSA A. VINOORES, LAURA E. EIERNAN, AND JULIE A. WEEDER. **The swimbladder nematode *Anguillicola crassus* in American eels (*Anguilla rostrata*) from middle and upper regions of Chesapeake bay.** *Journal of Parasitology*, 87(6):1366–1370, December 2001. 3
- [19] FRANTISEK MORAVEC, KAZUYA NAGASAWA, AND MUNENORI MIYAKAWA. **First record of ostracods as natural intermediate hosts of *Anguillicola crassus*, a pathogenic swimbladder parasite of eels *Anguilla* spp.** *Diseases of Aquatic Organisms*, 66(2):171–3, September 2005. 3
- [20] O. L. M. HAENEN, T. A. M. VAN WIJNGAARDEN, M. H. T. VAN DER HEIJDEN, J. HÖGLUND, J. B. J. W. CORNELISSEN, L. A. M. G. VAN LEENGOD, F. H. M. BORGSTEED, AND W. B. VAN MUISWINKEL. **Effects of experimental infections with different doses of *Anguillicola crassus* (Nematoda, Dracunculoidea) on European eel (*Anguilla anguilla*).** *Aquaculture*, 141(1–2):101–8, July 2006. PMID: 16956057. 3
- [21] M. POLZER AND H. TARASCHEWSKI. **Identification and characterization of the proteolytic enzymes in the developmental stages of the eel-pathogenic nematode *Anguillicola crassus*.** *Parasitology Research*, 79(1):24–7, 1993. 3
- [22] D. DE CHARLEROY, L. GRIZEZ, K. THOMAS, C. BELPAIRE, AND F. OLLEVIER. **The life cycle of *Anguillicola crassus*.** *Diseases of Aquatic Organisms*, 8(2):77–84, 1990. 3

REFERENCES

- [23] J WÜRTZ, K KNOPF, AND H TARASCHEWSKI. Distribution and prevalence of *Anguillicola crassus* (Nematoda) in eels *Anguilla anguilla* of the rivers Rhine and Naab, Germany. *Diseases of Aquatic Organisms*, **32**(2):137–43, March 1998. 3
- [24] K. THOMAS, FP OLLEVIER, ET AL. Population biology of *Anguillicola crassus* in the final host *Anguilla anguilla*. *Diseases of aquatic organisms*, 1992. 3
- [25] F S LEFEBVRE AND A J CRIVELLI. Anguillicolosis: dynamics of the infection over two decades. *Diseases of Aquatic Organisms*, **62**(3):227–32, December 2004. 3
- [26] M MÜNDEL, H TARASCHEWSKI, B KLAR, C W CHANG, J C SHIAO, K N SHEN, J T HE, S H LIN, AND W N TZENG. Occurrence of *Anguillicola crassus* (Nematoda: Dracunculoidea) in Japanese eels *Anguilla japonica* from a river and an aquaculture unit in SW Taiwan. *Diseases of Aquatic Organisms*, **71**(2):101–8, July 2006. 3, 5
- [27] K. THOMAS AND F. OLLEVIER. Paratenic hosts of the swimbladder nematode *Anguillicola crassus*. *Diseases of Aquatic Organisms*, **13**:165–174, 1992. 3
- [28] M. PIETROCK AND T. MEINELT. Dynamics of *Anguillicola Crassus* Larval Infections in a Paratenic Host, the Ruffe (*Gymnocephalus Cernuus*) from the Oder River on the Border of Germany and Poland. *Journal of Helminthology*, **76**(03):235–240, 2002. 3, 5
- [29] LESZEK ROLBIECKI. Can the DAB (*Limanda limanda*) be a paratenic host of *Anguillicola crassus* (Nematoda: Dracunculoidea)? The Gulf of Gdańsk and Vistula Lagoon (Poland) example. *Wiadomości Parazytologiczne*, **50**(2):317–22, 2004. 5
- [30] C SZÉKELY. Dynamics of *Anguillicola crassus* (Nematoda: Dracunculoidea) larval infection in paratenic host fishes of Lake Balaton, Hungary. *Acta Veterinaria Hungarica*, **43**(4):401–22, 1995. 5
- [31] F. MORAVEC AND B. SKORIKOVA. Amphibians and larvae of aquatic insects as new paratenic hosts of *Anguillicola crassus* (Nematoda: Dracunculoidea), a swimbladder parasite of eels. *DISEASES OF AQUATIC ORGANISMS*, **34**:217–222, 1998. 5
- [32] M. SCHABUSS, C.R. KENNEDY, R. KONECNY, B. GRILITSCH, W. RECKENDORFER, F. SCHIEMER, AND A. HERZIG. Dynamics and Predicted Decline of *Anguillicola Crassus* Infection in European Eels, *Anguilla Anguilla*, in Neusiedler See, Austria. *Journal of Helminthology*, **79**(02):159–167, 2005. 5, 8
- [33] F.W. TESCH. *Der Aal: Biologie und Fischerei*. Paul Parey, 1983. 5
- [34] T. WIRTH AND L. BERNATCHEZ. Decline of North Atlantic eels: a fatal synergy? *Proc. Biol. Sci.*, **270**:681–688, Apr 2003. 5
- [35] K KNOPF AND M MAHNKE. Differences in susceptibility of the European eel (*Anguilla anguilla*) and the Japanese eel (*Anguilla japonica*) to the swimbladder nematode *Anguillicola crassus*. *Parasitology*, **129**(Pt 4):491–6, October 2004. 5, 6
- [36] K KNOPF. The swimbladder nematode *Anguillicola crassus* in the European eel *Anguilla anguilla* and the Japanese eel *Anguilla japonica*: differences in susceptibility and immunity between a recently colonized host and the original host. *Journal of Helminthology*, **80**(2):129–36, June 2006. 5
- [37] MATTHEW J GOLLOCK, CLIVE R KENNEDY, AND J ANNE BROWN. Physiological responses to acute temperature increase in European eels *Anguilla anguilla* infected with *Anguillicola crassus*. *Diseases of Aquatic Organisms*, **64**(3):223–8, May 2005. 5
- [38] A.P. PALSTRA, D.F.M. HEPPENER, V.J.T. VAN GINEKEN, 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
- [39] B. SURES AND K. KNOPF. Parasites as a threat to freshwater eels? *Science*, **304**(5668):209–11, Apr 2004. 5
- [40] J WÄIJRTZ AND H TARASCHEWSKI. Histopathological changes in the swimbladder wall of the European eel *Anguilla anguilla* due to infections with *Anguillicola crassus*. *Diseases of Aquatic Organisms*, **14**(39):121–134, 2000. 5
- [41] A BEREGI, K MOLNÁR, L BÉKÉSI, AND C SZÉKELY. Radiodiagnostic method for studying swimbladder inflammation caused by *Anguillicola crassus* (Nematoda: Dracunculoidea). *Diseases of Aquatic Organisms*, **34**(2):155–60, October 1998. 5
- [42] G. FAZIO, P. SASAL, C. DA SILVA, B. FUMET, J. BOISSIER, R. LECOMTE-FINIGER, AND H. MONÉ. Regulation of *Anguillicola crassus* (Nematoda) infrapopulations in their definitive host, the European eel, *Anguilla anguilla*. *Parasitology*, **135**(1):1–10, 2008. 5
- [43] K KNOPF AND R LUCIUS. Vaccination of eels (*Anguilla japonica* and *Anguilla anguilla*) against *Anguillicola crassus* with irradiated L3. *Parasitology*, **135**(5):633–40, April 2008. 5
- [44] EMANUEL HEITLINGER, DOMINIK LAETSCH, URSZULA WECLAWSKI, YU-SAN HAN, AND HORST TARASCHEWSKI. Massive encapsulation of larval *Anguillicoloides crassus* in the intestinal wall of Japanese eels. *Parasites and Vectors*, **2**(1):48, 2009. 5, 6, 66
- [45] K. AARESTRUP, F. OKLAND, M. M. HANSEN, D. RIGHTON, P. GARGAN, M. CASTONGUAY, L. BERNATCHEZ, P. HOWEY, H. SPARHOLT, M. I. PEDERSEN, AND R. S. MCKINLEY. Oceanic spawning migration of the European eel (*Anguilla anguilla*). *Science*, **325**:1660, Sep 2009. 6
- [46] M. KUROKI, J. AOYAMA, M. J. MILLER, T. YOSHINAGA, A. SHINODA, S. HAGIHARA, AND K. TSUKAMOTO. Sympatric spawning of *Anguilla marmorata* and *Anguilla japonica* in the western North Pacific Ocean. *J. Fish Biol.*, **74**:1853–1865, Jun 2009. 7

REFERENCES

- [47] T. D. AILS, M. M. HANSEN, G. E. MAES, M. CASTONGUAY, L. RIEMANN, K. AARESTRUP, P. MUNK, H. SPARHOLT, R. HANEL, AND L. BERNATCHEZ. All roads lead to home: panmixia of European eel in the Sargasso Sea. *Mol. Ecol.*, **20**:1333–1346, Apr 2011. 7
- [48] J. M. PUJOLAR, G. A. DE LEO, E. CICCOTTI, AND L. ZANE. Genetic composition of Atlantic and Mediterranean recruits of European eel *Anguilla anguilla* based on EST-linked microsatellite loci. *J. Fish Biol.*, **74**:2034–2046, Jun 2009. 7
- [49] T. WIRTH AND L. BERNATCHEZ. Genetic evidence against panmixia in the European eel. *Nature*, **409**:1037–1040, Feb 2001. 7
- [50] S. PALM, J. DANNEWITZ, T. PRESTEGAARD, AND H. WICKSTROM. Panmixia in European eel revisited: no genetic difference between maturing adults from southern and northern Europe. *Heredity*, **103**:82–89, Jul 2009. 7
- [51] J. DANNEWITZ, G. E. MAES, L. JOHANSSON, H. WICKSTROM, F. A. VOLCKAERT, AND T. JARVI. Panmixia in the European eel: a matter of time.. *Proc. Biol. Sci.*, **272**:1129–1137, Jun 2005. 7
- [52] J. M. PUJOLAR, D. BEVACQUA, F. CAPOCCHI, E. CICCOTTI, G. A. DE LEO, AND L. ZANE. Genetic variability is unrelated to growth and parasite infestation in natural populations of the European eel (*Anguilla anguilla*). *Mol. Ecol.*, **18**:4604–4616, Nov 2009. 7
- [53] S. D. COTE, A. STIEN, R. J. IRVINE, J. F. DALLAS, F. MARSHALL, O. HALVORSEN, R. LANGVATN, AND S. D. ALBON. Resistance to abomasal nematodes and individual genetic variability in reindeer. *Mol. Ecol.*, **14**:4159–4168, Nov 2005. 7
- [54] J. M. RIJKS, J. I. HOFFMAN, T. KUIKEN, A. D. OSTERHAUS, AND W. AMOS. Heterozygosity and lungworm burden in harbour seals (*Phoca vitulina*). *Heredity*, **100**:587–593, Jun 2008. 7
- [55] M. DIONNE. Pathogens as potential selective agents in the wild. *Mol. Ecol.*, **18**:4523–4525, Nov 2009. 7
- [56] M. K. OLIVER, S. TELFER, AND S. B. PIERTNEY. Major histocompatibility complex (MHC) heterozygote superiority to natural multi-parasite infections in the water vole (*Arvicola terrestris*). *Proc. Biol. Sci.*, **276**:1119–1128, Mar 2009. 7
- [57] P. ILMONEN, D. J. PENN, K. DAMJANOVICH, L. MORRISON, L. GHOTBI, AND W. K. POTTS. Major histocompatibility complex heterozygosity reduces fitness in experimentally infected mice. *Genetics*, **176**:2501–2508, Aug 2007. 7
- [58] K. MATHIAS WEGNER, MARTIN KALBE, JOACHIM KURTZ, THORSTEN B. H. REUSCH, AND MANFRED MILINSKI. Parasite Selection for Immunogenetic Optimality. *Science*, **301**(5638):1343, September 2003. 7
- [59] DJ CONWAY AND SD POLLEY. Measuring immune selection. *Parasitology (London. Print)*, **125**:3–16, 2002. 7
- [60] C. M. L. PRESS AND Ø. EVENSEN. The morphology of the immune system in teleost fishes. *Fish & Shellfish Immunology*, **9**(4):309–318, 1999. 7
- [61] M E NIELSEN AND M D ESTEVE-GASSENT. The eel immune system: present knowledge and the need for research. *Journal of Fish Diseases*, **29**(2):65–78, 2006. 7
- [62] B. STAR, A. J. NEDERBRAGT, S. JENTOFT, U. GRIMHOLT, M. 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. VOUGEL, 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. 7
- [63] J. HIKIMA, T. S. JUNG, AND T. AOKI. Immunoglobulin genes and their transcriptional control in teleosts. *Dev. Comp. Immunol.*, **35**:924–936, Sep 2011. 7
- [64] S. KALUJNAIA, I. S. MCWILLIAM, V. A. ZAGUINAICO, A. L. FEILEN, J. NICHOLSON, N. HAZON, C. P. CUTLER, AND G. CRAMB. Transcriptomic approach to the study of osmoregulation in the European eel *Anguilla anguilla*. *Physiol. Genomics*, **31**:385–401, Nov 2007. 7
- [65] H. TARASCHEWSKI AND F. MORAVEC. Revision of the genus *Anguillicolæ* Yamaguti, 1935 (Nematoda: *Anguillicolidae*) of the swimbladder of eels, including descriptions of two new species, *A. novaezealandiae* sp. n. and *A. papernai* sp. n. *Folia Parasitol (Praha)*, **35**(2):125–146, 1988. 8
- [66] S. YAMAGUTI. Studies on the helminth fauna of Japan, part 9. Nematodes of fishes. *Japanese Journal of Zoology*, **6**, 1933. 8
- [67] T. H. JOHNSTON AND P. M. MAWSON. Some nematodes parasitic in Australian freshwater fish. *Transactions of the Royal Society of South Australia*, **64**(2):340–352, 1940. 8
- [68] FRANTISEK MORAVEC. *Dracunculoid and anguillicoloid nematodes parasitic in vertebrates*. Academia, 2006. 8
- [69] YUKI MINEGISHI, JUN AOYAMA, JUN G. INOUE, MASAKI MIYA, MUTSUMI NISHIDA, AND KATSUMI TSUKAMOTO. Molecular phylogeny and evolution of the freshwater eels genus *Anguilla* based on the whole mitochondrial genome sequences. *Molecular Phylogenetics and Evolution*, **34**(1):134–146, 2005. 8
- [70] MARK L. BLAXTER, PAUL DE LEY, JAMES R. GAREY, LEO X. LIU, PATSY SCHELDEMAN, ANDY VIERSTRAETE, JACQUES R. VANFLETEREN, LAURA Y. MACKEY, MARK DORRIS, LINDA M. FRISSE, J. T. VIDA, AND W. KELLEY THOMAS. A molecular evolutionary framework for the phylum Nematoda. *Nature*, **392**(6671):71–75, March 1998. 11

REFERENCES

- [71] S. A. NADLER, R. A. CARRENO, H. MEJIA-MADRID, J. ULLBERG, C. PAGAN, R. HOUSTON, AND J.-P. HUGOT. Molecular Phylogeny of Clade III Nematodes Reveals Multiple Origins of Tissue Parasitism. *Parasitology*, **134**(10):1421–1442, 2007. 11
- [72] MARTINA WIJOVÁ, FRANTISEK MORAVEC, ALES HORÁK, AND JULIUS LUKES. Evolutionary relationships of Spirurina (Nematoda: Chromadorea: Rhabditida) with special emphasis on dracunculoid nematodes inferred from SSU rRNA gene sequences. *International Journal for Parasitology*, **36**(9):1067–75, August 2006. 11
- [73] A. KERNER. *The natural history of plants, their forms, growth, reproduction, and distribution. Translated by F. W. Oliver.*, 1895. 11
- [74] G. BONNIER. Recherches expérimentales sur la adaptation des plantes au climat alpin. *Ann. Scie. Nat. (Bot.)*, **20**:217–358, 1895. 11
- [75] O. KALTZ AND J. A. SHYKOFF. Local adaptation in host-parasite systems. *Heredity*, pages 361–370, May 1998. 13, 67
- [76] T. A. MOUSSEAU AND D. A. ROFF. Natural selection and the heritability of fitness components. *Heredity*, **59** (Pt 2):181–197, Oct 1987. 15
- [77] J. N. THOMPSON, S. L. NUISMER, AND R. GOMULKIEWICZ. Coevolution and maladaptation. *Integr. Comp. Biol.*, **42**:381–387, Apr 2002. 15
- [78] J. N. THOMPSON. *The geographic mosaic of coevolution*. University of Chicago Press, 2005. 15
- [79] S. L. NUISMER AND S. GANDON. Moving beyond common-garden and transplant designs: insight into the causes of local adaptation in species interactions. *Am. Nat.*, **171**:658–668, May 2008. 15
- [80] F. H. C. CRICK. The biological replication of macromolecules. In *Symp. Soc. Exp. Biol.*, **12**, pages 138–163, 1958. 16
- [81] CRICK F. Central dogma of molecular biology. *Nature*, **226**:1198–1199, Jun 1970. [PubMed:5422595]. 16
- [82] M. LYNCH. The lower bound to the evolution of mutation rates. *Genome Biol Evol*, **3**:1107–1118, 2011. 16
- [83] Y. WAN, M. KERTESZ, R. C. SPITALE, E. SEGAL, AND H. Y. CHANG. Understanding the transcriptome through RNA structure. *Nat. Rev. Genet.*, **12**:641–655, Sep 2011. 17
- [84] H. GUO, N. T. INGOLIA, J. S. WEISSMAN, AND D. P. BARTEL. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, **466**:835–840, Aug 2010. 17
- [85] G. RUVKUN. Molecular biology. Glimpses of a tiny RNA world. *Science*, **294**:797–799, Oct 2001. 17
- [86] G. DIECI, M. PRETI, AND B. MONTANINI. Eukaryotic snoRNAs: a paradigm for gene expression flexibility. *Genomics*, **94**:83–88, Aug 2009. 17
- [87] W. DENG, X. ZHU, G. SKOGRB?, Y. ZHAO, Z. FU, Y. WANG, H. HE, L. CAI, H. SUN, C. LIU, B. LI, B. BAI, J. WANG, D. JIA, S. SUN, H. HE, Y. CUI, Y. WANG, D. BU, AND R. CHEN. Organization of the *Caenorhabditis elegans* small non-coding transcriptome: genomic features, biogenesis, and expression. *Genome Res.*, **16**:20–29, Jan 2006. 17
- [88] F. H. CRICK. The origin of the genetic code. *J. Mol. Biol.*, **38**:367–379, Dec 1968. 17
- [89] E. KIM, A. MAGEN, AND G. AST. Different levels of alternative splicing among eukaryotes. *Nucleic Acids Res.*, **35**:125–131, 2007. 18
- [90] Z. WANG, M. GERSTEIN, AND M. SNYDER. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**:57–63, Jan 2009. 18, 23, 27
- [91] J. ARMENGaud. Proteogenomics and systems biology: quest for the ultimate missing parts. *Expert Rev Proteomics*, **7**:65–77, Feb 2010. 18
- [92] B. SCHWANHAUSER, D. BUSSE, N. LI, G. DITTMAR, J. SCHUCHHARDT, J. WOLF, W. CHEN, AND M. SELBACH. Global quantification of mammalian gene expression control. *Nature*, **473**:337–342, May 2011. 18
- [93] F. SANGER, S. NICKLEN, AND A. R. COULSON. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.*, **74**:5463–5467, Dec 1977. 19
- [94] H. SWERDLOW AND R. GESTELAND. Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic Acids Res.*, **18**:1415–1419, Mar 1990. 19
- [95] W. FIERS, R. CONTRERAS, F. DUERINCK, G. HAEGERMAN, D. ISERENTANT, J. MERREGAERT, W. MIN JOU, F. MOLEMAN, A. RAEYMAEKERS, A. VAN DEN BERGHE, G. VOLCKAERT, AND M. YSEBAERT. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature*, **260**:500–507, Apr 1976. 19
- [96] F. R. BLATTNER, G. PLUNKETT, C. A. BLOCH, N. T. PERNA, V. BURLAND, M. RILEY, J. COLLADO-VIDES, J. D. GLASNER, C. K. RODE, G. F. MAYHEW, J. GREGOR, N. W. DAVIS, H. A. KIRKPATRICK, M. A. GOEDEN, D. J. ROSE, B. MAU, AND Y. SHAO. The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**:1453–1462, Sep 1997. 19
- [97] A. GOFFEAU, B. G. BARRELL, H. BUSSEY, R. W. DAVIS, B. DUJON, H. FELDMANN, F. GALIBERT, J. D. HOHEISEL, C. JACQ, M. JOHNSTON, E. J. LOUIS, H. W. MEWES, Y. MURAKAMI, P. PHILIPPSEN, H. TETTELIN, AND S. G. OLIVER. Life with 6000 genes. *Science*, **274**:563–567, Oct 1996. 19
- [98] THE C. ELEGANS SEQUENCING CONSORTIUM. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, **282**:2012–2018, Dec 1998. 19

REFERENCES

- [99] M. D. ADAMS, S. E. CELNIKER, R. A. HOLT, C. A. EVANS, J. D. GOCAYNE, P. G. AMANATIDES, S. E. SCHERER, P. W. LI, R. A. HOSKINS, R. F. GALLO, ET AL. **The genome sequence of *Drosophila melanogaster***. *Science*, **287**(5461):2185, 2000. 19
- [100] R. H. WATERSTON, K. LINDBLAD-TOH, E. BIRNEY, J. ROGERS, J. F. ABRIL, P. AGARWAL, R. AGARWALA, R. AINSCOUGH, M. ALEXANDERSSON, P. AN, S. E. ANTONARAKIS, J. ATTWOOD, R. BAERTSCH, J. BAILEY, K. BARLOW, S. BECK, E. BERRY, B. BIRREN, T. BLOOM, P. BORK, M. BOTCHERBY, N. BRAY, M. R. BRENT, D. G. BROWN, S. D. BROWN, C. BULT, J. BURTON, J. BUTLER, R. D. CAMPBELL, P. CARNINCI, S. CAWLEY, F. CHIAROMONTE, A. T. CHINWALLA, D. M. CHURCH, M. CLAMP, C. CLEE, F. S. COLLINS, L. L. COOK, R. R. COPLEY, A. COULSON, O. COURRONNE, 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. FELSENFIELD, 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. GRAPHAM, T. A. GRAVES, E. D. GREEN, S. GREGORY, R. GUIGO, M. GUYER, R. C. HARDISON, D. HAUSLER, Y. HAYASHIZAKI, L. W. HILLIER, A. HINRICH, 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. MCCLAY, 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. SHOWKEEN, S. SIMS, J. B. SINGER, G. SLATER, A. SMIT, D. R. SMITH, B. SPENCER, A. STABENAU, N. STANGE-TOMMANN, C. SUGNET, M. SUYAMA, G. TESLER, J. THOMPSON, D. TORRENTS, E. TREVASKIS, J. TROMP, C. UCLA, A. URETA-VIDAL, J. P. VINSON, A. C. VON NIEDERHAUSERN, C. M. WADE, M. WALL, R. J. WEBER, R. B. WEISS, M. C. WENDL, A. P. WEST, K. WETTERSTRAND, R. WHEELER, S. WHELAN, J. WIERZBOWSKI, D. WILLEY, S. WILLIAMS, R. K. WILSON, E. WINTER, K. C. WORLEY, D. WYMAN, S. YANG, S. P. YANG, E. M. ZDOBNOV, M. C. ZODY, AND E. S. LANDER. **Initial sequencing and comparative analysis of the mouse genome**. *Nature*, **420**:520–562, Dec 2002. 19
- [101] J. C. VENTER, M. D. ADAMS, E. W. MYERS, P. W. LI, R. J. MURAL, G. G. SUTTON, H. O. SMITH, M. YANDELL, C. A. EVANS, R. A. HOLT, J. D. GO-
- CAYNE, P. AMANATIDES, R. M. BALLEW, D. H. HUNSON, 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 MIKLLOS, 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-TREIDEH, E. BEASLEY, K. BIDDICK, V. BONAZZI, R. BRANDON, M. CARGILL, I. CHANDRAMOLISWARAN, R. CHARLAB, K. CHATURVEDI, Z. DENG, V. DI FRANCESCO, P. DUNN, K. ELBECK, 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. YOSEPH, D. ALLEN, A. BASU, J. BAXENDALE, L. BLICK, M. CAMINHA, J. CARNES-STINE, P. CAULK, Y. H. CHIANG, M. COYNE, C. DAHLKE, A. MAYS, M. DOMBROSKI, M. DONNELLY, D. ELY, S. ESPARHAM, C. FOSLER, H. GIRE, S. GLANOWSKI, K. GLASSER, A. GLODEK, M. GOROKHOV, K. GRAHAM, B. GROPMAN, M. HARRIS, J. HEIL, S. HENDERSON, J. HOOVER, D. JENNINGS, C. JORDAN, J. JORDAN, J. KASHA, L. KAGAN, C. KRAFT, A. LEVITSKY, M. LEWIS, X. LIU, J. LOPEZ, D. MA, W. MAJOROS, J. McDANIEL, S. MURPHY, M. NEWMAN, T. NGUYEN, N. NGUYEN, M. NODELL, S. PAN, J. PECK, M. PETERSON, W. ROWE, R. SANDERS, J. SCOTT, M. SIMPSON, T. SMITH, A. SPRAGUE, T. STOCKWELL, R. TURNER, E. VENTER, M. WANG, M. WEN, D. WU, M. WU, A. XIA, A. ZANDIEH, AND X. ZHU. **The sequence of the human genome**. *Science*, **291**:1304–1351, Feb 2001. 19
- [102] C. FIELDS, M. D. ADAMS, O. WHITE, AND J. C. VENTER. **How many genes in the human genome?** *Nat. Genet.*, **7**:345–346, Jul 1994. 19

REFERENCES

- [103] MARK BLAXTER. ***Caenorhabditis elegans* Is a Nematode.** *Science*, **282**(5396):2041–2046, December 1998. 19
- [104] M. B. GERSTEIN, Z. J. LU, E. L. VAN NOSTRAND, C. CHENG, B. I. ARSHINOFF, T. LIU, K. Y. YIP, R. ROBILLOTTO, A. RECHTSTEINER, K. IKEGAMI, P. ALVES, A. CHATEIGNER, M. PERRY, M. MORRIS, R. K. AUERBACH, X. FENG, J. LENG, A. VIELLE, W. NIU, K. KRISHNARAKRAI, 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. EGELOHOFER, S. ERGAN, G. EUSSKIRCHEN, 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. HENKOFF, S. R. HENZ, A. HINRICH, 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. LATTORRE, A. LEAHY, S. LEWIS, P. LLOYD, L. LOCHOVSKY, R. F. LOWDON, Y. LUBLING, R. LYNE, M. MACCoss, S. D. MACKOWIAK, M. MANGONE, S. MCKAY, D. MECHANAS, 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. RYZANOV, M. SAROV, R. SASIDHARAN, A. SBONER, P. SCHEID, E. SEGAL, H. SHIN, C. SHOU, F. J. SLACK, C. SLIGHSTAM, R. SMITH, W. C. SPENCER, E. O. STINSON, S. TAING, T. TAKASAKI, D. VAFFEADS, 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. GUNSLAS, 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 mod-ENCODE project.** *Science*, **330**:1775–1787, Dec 2010. 20
- [105] LINCOLN D. STEIN, ZHIRONG BAO, DARIN BLASIER, THOMAS BLUMENTHAL, MICHAEL R. BRENT, NANSHENG CHEN, ASIF CHINWALLA, LAURA CLARKE, CHRIS CLEE, AVRIL COGHLAN, ALAN COULSON, PETER D'EUSTACHIO, DAVID H. A. FITCH, LUCINDA A. FULTON, ROBERT E. FULTON, SAM GRIFFITHS-JONES, TODD W. HARRIS, LADEANA W. HILLIER, RAVI KAMATH, PATRICIA E. KUWABARA, ELAINE R. MARDIS, MARCO A. MARRA, TRACIE L. MINER, PATRICK MINX, JAMES C. MULLIKIN, ROBERT W. PLUMB, JANE ROGERS, JACQUELINE E. SCHEIN, MARC SOHRMANN, JOHN SPIETH, JASON E. STAJICH, CHAOCHUN WEI, DAVID WILLEY, RICHARD K. WILSON, RICHARD DURBIN, AND ROBERT H. WATERSTON. **The Genome Sequence of *Caenorhabditis briggsae*: A Platform for Comparative Genomics.** *PLoS Biology*, **1**(2):e45 EP –, November 2003. 20
- [106] C. DIETERICH, S. W. CLIFTON, L. N. SCHUSTER, A. CHINWALLA, K. DELEHAUNTY, I. DINKELACKER, L. FULTON, R. FULTON, J. GODFREY, P. MINX, M. MITREVA, W. ROESELER, H. TIAN, H. WITTE, S. P. YANG, R. K. WILSON, AND R. J. SOMMER. **The *Pristionchus pacificus* genome provides a unique perspective on nematode lifestyle and parasitism.** *Nat. Genet.*, **40**:1193–1198, Oct 2008. 20
- [107] ELODIE GHEDIN, SHILIANG WANG, DAVID SPIRO, ELISABET CALER, QI ZHAO, JONATHAN CRABTREE, JONATHAN E. ALLEN, ARTHUR L. DELCHER, DAVID B. GUILIANO, DIEGO MIRANDA-SAAVEDRA, SAMUEL V. ANGIOLI, TODD CREASY, PAOLO AMEDEO, BRIAN HAAS, NAJIB M. EL-SAYED, JENNIFER R. WORTMAN, TAMARA FELDBLYUM, LUKE TALLON, MICHAEL SCHATZ, MARTIN SHUMWAY, HEAN KOO, STEVEN L. SALZBERG, SETH SCHOBEL, MIRHAELA PERTEA, MIHAI POP, OWEN WHITE, GEOFFREY J. BARTON, CLOTILDE K. S. CARLOW, MICHAEL J. CRAWFORD, JENNIFER DAUB, MATTHEW W. DIMMIC, CHRIS F. ESTES, JEREMY M. FOSTER, MEHUL GANATRA, WILLIAM F. GREGORY, NICHOLAS M. JOHNSON, JINMING JIN, RICHARD KOMUNIECKI, IAN KORF, SANJAY KUMAR, SANDRA LANEY, BEN-WEN LI, WEN LI, TIM H. LINDBLOM, SARA LUSTIGMAN, DONG MA, CLAUDE V. MAINA, DAVID M. A. MARTIN, JAMES P. MCCARTER, LARRY McREYNOLDS, MAKEDONKA MITREVA, THOMAS B. NUTMAN, JOHN PARKINSON, JOSE M. PEREGRIN-ALVAREZ, CATHERINE POOLE, QINGHU REN, LORI SAUNDERS, ANN E. SLUDER, KATHERINE SMITH, MARIO STANKE, THOMAS R. UNNASCH, JENNA WARE, AGUAN D. WEI, GARY WEIL, DERYCK J. WILLIAMS, YINHUA ZHANG, STEVEN A. WILLIAMS, CLAIRE FRASER-LIGGETT, BARTON SLATKO, MARK L. BLAXTER, AND ALAN L. SCOTT. **Draft Genome of the Filarial Nematode Parasite *Brugia malayi*.** *Science*, **317**(5845):1756–1760, September 2007. 20
- [108] 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. NEISUM, P. W. STERNBERG, H. YANG, J. WANG, J. WANG, AND R. B. GASER. ***Ascaris suum* draft genome.** *Nature*, Oct 2011. 20
- [109] M. MITREVA, D. P. JASMER, D. S. ZARLENGA, Z. WANG, S. ABUDUCKER, J. MARTIN, C. M. TAYLOR, Y. YIN, L. FULTON, P. MINX, S. P. YANG, W. C. WARREN, R. S. FULTON, V. BHONAGIRI, X. ZHANG, K. HALLSWORTH-PEPIN, S. W. CLIFTON, J. P. MCCARTER, J. APPLETON, E. R. MARDIS, AND R. K. WILSON. **The draft genome of the parasitic nematode *Trichinella spiralis*.** *Nat. Genet.*, **43**:228–235, Mar 2011. 20
- [110] P. ABAD, J. GOUZY, J. M. AURY, P. CASTAGNONE-SERENO, E. G. DANCHIN, E. DELEURY, L. PERFUS-BARBECH, 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. EWANKIN, B. FAVERY, E. GRENIER, B. HENRISSAT, J. T. JONES, V. LAUDET, A. G. MAULE, H. QUESNEVILLE, M. N. ROSSO, T. SCHIEX, G. SMANT, J. WEISSENBACH, AND P. WINCKER. **Genome sequence of the metazoan plant-parasitic nematode *Meloidogyne incognita*.** *Nat. Biotechnol.*, **26**:909–915, Aug 2008. 20
- [111] C. H. OPPERMANN, 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. SCHAFER, E. SCHOLL, B. R. SOSINSKI, V. P. THOMAS, AND

REFERENCES

- E. WINDHAM. Sequence and genetic map of *Meloidogyne hapla*: A compact nematode genome for plant parasitism. *Proc. Natl. Acad. Sci. U.S.A.*, **105**:14802–14807, Sep 2008. 20
- [112] T. KIKUCHI, J. A. COTTON, J. J. DALZELL, K. HASEGAWA, N. KANZAKI, P. McVEIGH, T. TAKANASHI, I. J. TSAI, S. A. ASSEFA, P. J. COCK, T. D. OTTO, M. HUNT, A. J. REID, A. SANCHEZ-FLORES, K. TSUCHIHARA, T. YOKOI, M. C. LARSSON, J. MIWA, A. G. MAULE, N. SAHASHI, J. T. JONES, AND M. BERRIMAN. Genomic Insights into the Origin of Parasitism in the Emerging Plant Pathogen *Bursaphelenchus xylophilus*. *PLoS Pathog.*, **7**:e1002219, Sep 2011. 20
- [113] S. KUMAR, P. H. SCHIFFER, AND M. BLAXTER. 959 Nematode Genomes: a semantic wiki for coordinating sequencing projects. *Nucleic Acids Res.*, Nov 2011. [DOI:10.1093/nar/gkr826] [PubMed:22058131]. 20
- [114] JOHN PARKINSON, ALASDAIR ANTHONY, JAMES WASMUTH, RALF SCHMID, ANN HEDLEY, AND MARK BLAXTER. PartiGene—constructing partial genomes. *Bioinformatics*, **20**(9):1398–1404, June 2004. 20, 71
- [115] R. M. MAIZELS, N. GOMEZ-ESCOBAR, W. F. GREGORY, J. MURRAY, AND X. ZANG. Immune evasion genes from filarial nematodes. *Int. J. Parasitol.*, **31**:889–898, Jul 2001. 20, 21
- [116] RICK M. MAIZELS, ADAM BALIC, NATALIA GOMEZ-ESCOBAR, MEERA NAIR, MATT D. TAYLOR, AND JUDITH E. ALLEN. Helminth parasites; masters of regulation. *Immunological Reviews*, **201**(1):89–116, 2004. 21
- [117] NATALIA GOMEZ-ESCOBAR, WILLIAM F. GREGORY, COLLETTE BRITTON, LINDA MURRAY, CRAIG CORTON, NEIL HALL, JEN DAUB, MARK L. BLAXTER, AND RICK M. MAIZELS. Abundant larval transcript-1 and -2 genes from *Brugia malayi*: diversity of genomic environments but conservation of 5' promoter sequences functional in *Caenorhabditis elegans*. *Molecular and Biochemical Parasitology*, **125**(1-2):59–71, 2002. 21
- [118] J. MURRAY, W. F. GREGORY, N. GOMEZ-ESCOBAR, A. K. ATMADJA, AND R. M. MAIZELS. Expression and immune recognition of *Brugia malayi* VAL-1, a homologue of vespid venom allergens and *Ancylostoma* secreted proteins. *Mol. Biochem. Parasitol.*, **118**:89–96, Nov 2001. [PubMed:11704277]. 21
- [119] YVONNE HARCUS, JOHN PARKINSON, CECILIA FERNANDEZ, JENNIFER DAUB, MURRAY SELKIRK, MARK BLAXTER, AND RICK MAIZELS. Signal sequence analysis of expressed sequence tags from the nematode *Nippostrongylus brasiliensis* and the evolution of secreted proteins in parasites. *Genome Biology*, **5**(6):R39, 2004. 21
- [120] SHIVASHANKAR H. NAGARAJ, ROBIN B. GASSER, AND SHOBA RANGANATHAN. Needles in the EST Haystack: Large-Scale Identification and Analysis of Excretory-Secretory (ES) Proteins in Parasitic Nematodes Using Expressed Sequence Tags (ESTs). *PLoS Neglected Tropical Diseases*, **2**(9):e301, 2008. 21
- [121] JOHN PARKINSON, CLAIRE WHITTON, RALF SCHMID, MARIAN THOMSON, AND MARK BLAXTER. % bf NEMBASE: a resource for parasitic nematode ESTs. *Nucl. Acids Res.*, **32**(suppl_1):D427–430, 2004. 21, 73
- [122] JAMES Wasmuth, RALF SCHMID, ANN HEDLEY, AND MARK BLAXTER. On the Extent and Origins of Genic Novelty in the Phylum Nematoda. *PLoS Neglected Tropical Diseases*, **2**(7):e258, July 2008. 21
- [123] B. ELSWORTH, J. Wasmuth, AND M. BLAXTER. NEMBASE4: The nematode transcriptome resource. *Int. J. Parasitol.*, **41**:881–894, Jul 2011. 21, 73
- [124] Z. WANG, S. ABUBUCKER, J. MARTIN, R. K. WILSON, J. HAWDON, AND M. MITREVA. Characterizing *Ancylostoma caninum* transcriptome and exploring nematode parasitic adaptation. *BMC Genomics*, **11**:307, 2010. 21
- [125] N. BORCHERT, C. DIETERICH, K. KRUG, W. SCHUTZ, S. JUNG, A. NORDHEIM, R. J. SOMMER, AND B. MACEK. Proteogenomics of *Pristionchus pacificus* reveals distinct proteome structure of nematode models. *Genome Res.*, **20**:837–846, Jun 2010. 21
- [126] S. KUMAR AND M. L. BLAXTER. Comparing de novo assemblers for 454 transcriptome data. *BMC Genomics*, **11**:571, Oct 2010. 21, 25, 39, 73
- [127] J. WANG, B. CZECH, A. CRUNK, A. WALLACE, M. MITREVA, G. J. HANNON, AND R. E. DAVIS. Deep small RNA sequencing from the nematode *Ascaris* reveals conservation, functional diversification, and novel developmental profiles. *Genome Res.*, **21**:1462–1477, Sep 2011. 21
- [128] M. BLAXTER, S. KUMAR, G. KAUR, G. KOUTSOUVOULOS, AND B. ELSWORTH. Genomics and transcriptomics across the diversity of the Nematoda. *Parasite Immunol.*, Nov 2011. 21
- [129] M. L. METZKER. Sequencing technologies - the next generation. *Nat. Rev. Genet.*, **11**:31–46, Jan 2010. 21, 23
- [130] O. HARISMENDY, P. C. NG, R. L. STRAUSBERG, X. WANG, T. B. STOCKWELL, K. Y. BEESON, N. J. SCHORK, S. S. MURRAY, E. J. TOPOL, S. LEVY, AND K. A. FRAZER. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol.*, **10**:R32, 2009. 23
- [131] K. E. STEINMANN, C. E. HART, J. F. THOMPSON, AND P. M. MILOS. Helicos single-molecule sequencing of bacterial genomes. *Methods Mol. Biol.*, **733**:3–24, 2011. 23
- [132] W. TIMP, U. M. MIRSAIDOV, D. WANG, J. COMER, A. AKSIMENTIEV, AND G. TIMP. Nanopore Sequencing: Electrical Measurements of the Code of Life. *IEEE Trans Nanotechnol*, **9**:281–294, May 2010. 23
- [133] T. RAZ, M. CAUSEY, D. R. JONES, A. KIEU, S. LETOVSKY, D. LIPSON, E. THAYER, J. F. THOMPSON, AND P. M. MILOS. RNA sequencing and quantitation using the Helicos Genetic Analysis System. *Methods Mol. Biol.*, **733**:37–49, 2011. 23

REFERENCES

- [134] F. OZSOLAK AND P. M. MILOS. **Single-molecule direct RNA sequencing without cDNA synthesis.** *Wiley Interdiscip Rev RNA*, **2**:565–570, 2011. 23
- [135] J. M. ROTHBERG AND J. H. LEAMON. **The development and impact of 454 sequencing.** *Nat. Biotechnol.*, **26**:1117–1124, Oct 2008. 24
- [136] M. LARGUINHO, H. M. SANTOS, G. DORIA, H. SCHOLZ, P. V. BAPTISTA, AND J. L. CAPELO. **Development of a fast and efficient ultrasonic-based strategy for DNA fragmentation.** *Talanta*, **81**:881–886, May 2010. 23
- [137] P. NYREN. **The history of pyrosequencing.** *Methods Mol. Biol.*, **373**:1–14, 2007. 23
- [138] S. BALZER, K. MALDE, AND I. JONASSEN. **Systematic exploration of error sources in pyrosequencing flowgram data.** *Bioinformatics*, **27**:i304–309, Jul 2011. 23, 67
- [139] R. C. NOVAIS AND Y. R. THORSTENSON. **The evolution of Pyrosequencing® for microbiology: From genes to genomes.** *J. Microbiol. Methods*, **86**:1–7, Jul 2011. 25
- [140] 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. HELGESON, 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. MAKHJANI, K. E. McDADE, M. P. MCKENNA, E. W. MYERS, E. NICKERSON, J. R. NOBILE, R. PLANT, B. P. PUG, M. T. RONAN, G. T. ROTH, G. J. SARKIS, J. F. SIMONS, J. W. SIMPSON, M. SRINIVASAN, K. R. TARTARO, A. TOMASZ, K. A. VOGT, G. A. VOLKMER, S. H. WANG, Y. WANG, M. P. WEINER, P. YU, R. F. BEGLEY, AND J. M. ROTHBERG. **Genome sequencing in microfabricated high-density picolitre reactors.** *Nature*, **437**:376–380, Sep 2005. 25, 39, 73
- [141] 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, O. 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. 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. EGEBUJOR, 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. KREBTUKOVA, 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. McCUALEY, 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. ROBERT 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. SEGAANE, 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. SZAKOWSKI, C. L. TREGIDGO, G. TURCATTI, S. VANDEVONDELE, Y. VERHOVSKY, S. M. VIRK, S. WAKELIN, G. C. WALCOTT, J. WANG, G. J. WORSLEY, J. YAN, L. YAU, M. ZUERLEIN, J. ROGERS, J. C. MULLIKIN, M. E. HURLES, N. J. MCCOOKE, J. S. WEST, F. L. OAKS, P. L. LUNDBERG, D. KLENERMAN, R. DURBIN, AND A. J. SMITH. **Accurate whole human genome sequencing using reversible terminator chemistry.** *Nature*, **456**:53–59, Nov 2008. 25, 27
- [142] R. LI, W. FAN, G. TIAN, H. ZHU, L. HE, J. CAI, Q. HUANG, Q. CAI, B. LI, Y. BAI, Z. ZHANG, Y. ZHANG, W. WANG, J. LI, F. WEI, H. LI, M. JIAN, J. LI, Z. ZHANG, R. NIELSEN, D. LI, W. GU, Z. YANG, Z. XUAN, O. A. RYDER, F. C. LEUNG, Y. ZHOU, J. CAO, X. SUN, Y. FU, X. FANG, X. GUO, B. WANG, R. HOU, F. SHEN, B. MU, P. NI, R. LIN, W. QIAN, G. WANG, C. YU, W. NIE, J. WANG, Z. WU, H. LIANG, J. MIN, Q. WU, S. CHENG, J. RUAN, M. WANG, Z. SHI, M. WEN, B. LIU, X. REN, H. ZHENG, D. DONG, K. COOK, G. SHAN, H. ZHANG, C. KOSIOL, X. XIE, Z. LU, H. ZHENG, Y. LI, C. C. STEINER, T. T. LAM, S. LIN, Q. ZHANG, G. LI, J. TIAN, T. GONG, H. LIU, D. ZHANG, L. FANG, C. YE, J. ZHANG, W. HU, A. XU, Y. REN, G. ZHANG, M. W. BRUFORD, Q. LI, L. MA, Y. GUO, N. AN, Y. HU, Y. ZHENG, Y. SHI, Z. LI, Q. LIU, Y. CHEN, J. ZHAO, N. QU, S. ZHAO, F. TIAN, X. WANG, H. WANG, L. XU, X. LIU, T. VINAR, Y. WANG, T. W. LAM, S. M. YIU, S. LIU, H. ZHANG, D. LI, Y. HUANG, X. WANG, G. YANG, Z. JIANG, J. WANG, N. QIN, L. LI, J. LI, L. BOLUND, K. KRISTIANSEN, G. K. WONG, M. OLSON, X. ZHANG, S. LI, H. YANG, J. WANG, AND J. WANG. **The sequence and de novo assembly of the giant panda genome.** *Nature*, **463**:311–317, Jan 2010. 27
- [143] B. FELDMAYER, C. W. WHEAT, N. KREZDORN, B. ROTTER, AND M. PFENNINGER. **Short read Illumina data for the de novo assembly of a non-model snail species transcriptome (*Radix balthica*, Basommatophora, Pulmonata), and a comparison of assembler performance.** *BMC Genomics*, **12**:317, 2011. 27

REFERENCES

- [144] J. H. MALONE AND B. OLIVER. **Microarrays, deep sequencing and the true measure of the transcriptome.** *BMC Biol.*, **9**:34, 2011. 27
- [145] H. MATSUMURA, K. YOSHIDA, S. LUO, D. H. KRUGER, G. KAHL, G. P. SCHROTH, AND R. TERAUCHI. **High-throughput SuperSAGE.** *Methods Mol. Biol.*, **687**:135–146, 2011. 27
- [146] V. E. VELCULESCU, L. ZHANG, B. VOGELSTEIN, AND K. W. KINZLER. **Serial analysis of gene expression.** *Science*, **270**:484–487, Oct 1995. 27
- [147] J. R. MILLER, S. KOREN, AND G. SUTTON. **Assembly algorithms for next-generation sequencing data.** *Genomics*, **95**:315–327, Jun 2010. 27
- [148] 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]. 27
- [149] R. STADEN. **A strategy of DNA sequencing employing computer programs.** *Nucleic Acids Res.*, **6**:2601–2610, Jun 1979. 28
- [150] T. R. GINGERAS AND R. J. ROBERTS. **Steps toward computer analysis of nucleotide sequences.** *Science*, **209**:1322–1328, Sep 1980. 28
- [151] T. F. SMITH AND M. S. WATERMAN. **Identification of common molecular subsequences.** *J. Mol. Biol.*, **147**:195–197, Mar 1981. 28
- [152] T. F. SMITH, M. S. WATERMAN, AND W. M. FITCH. **Comparative biosequence metrics.** *J. Mol. Evol.*, **18**:38–46, 1981. 28
- [153] W. J. KENT. **BLAT—the BLAST-like alignment tool.** *Genome Res.*, **12**:656–664, Apr 2002. 28
- [154] Z. NING, A. J. COX, AND J. C. MULLIKIN. **SSAHA: a fast search method for large DNA databases.** *Genome Res.*, **11**:1725–1729, Oct 2001. 28, 50, 74
- [155] H. LI AND R. DURBIN. **Fast and accurate long-read alignment with Burrows-Wheeler transform.** *Bioinformatics*, **26**:589–595, Mar 2010. 28, 58
- [156] D. R. ZERBINO AND E. BIRNEY. **Velvet: algorithms for de novo short read assembly using de Bruijn graphs.** *Genome Res.*, **18**:821–829, May 2008. 29
- [157] M. G. GRABHERR, B. J. HAAS, M. YASSOUR, J. Z. LEVIN, D. A. THOMPSON, I. AMIT, X. ADICONIS, L. FAN, R. RAYCHOWDHURY, Q. ZENG, Z. CHEN, E. MAUCELI, N. HACOHEN, A. GNIRKE, N. RHIND, F. DI PALMA, B. W. BIRREN, C. NUSBAUM, K. LINDBLAD-TOH, N. FRIEDMAN, AND A. REGEV. **Full-length transcriptome assembly from RNA-Seq data without a reference genome.** *Nat. Biotechnol.*, **29**:644–652, Jul 2011. 29
- [158] T. S. SCHWARTZ, H. TAE, Y. YANG, K. MOCKAITIS, J. L. VAN HEMERT, S. R. PROULX, J. H. CHOI, AND A. M. BRONIKOWSKI. **A garter snake transcriptome: pyrosequencing, de novo assembly, and sex-specific differences.** *BMC Genomics*, **11**:694, 2010. 29, 50
- [159] F. GOETZ, D. ROSAUER, S. SITAR, G. GOETZ, C. SIM-CHICK, 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. 29
- [160] M. K. HUGHES AND A. L. HUGHES. **Natural selection on Plasmodium surface proteins.** *Mol. Biochem. Parasitol.*, **71**:99–113, Apr 1995. 30
- [161] S. VIA. **The Ecological Genetics of Speciation.** *The American Naturalist*, **159**(S3):1–7, 2002. 30
- [162] 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. 30
- [163] 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. 30
- [164] S. L. NUISMER AND S. P. OTTO. **Host-parasite interactions and the evolution of gene expression.** *PLoS Biol.*, **3**:e203, Jul 2005. 30
- [165] D. COLINET, A. SCHMITZ, D. CAZES, J. L. GATTI, AND M. POIRIE. **The origin of intraspecific variation of virulence in an eukaryotic immune suppressive parasite.** *PLoS Pathog.*, **6**:e1001206, 2010. 30
- [166] B. CHEVREUX, T. PFISTERER, B. DRESCHER, A. J. DRIESSEL, W. E. MULLER, T. WETTER, AND S. SUHAL. **Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs.** *Genome Res.*, **14**:1147–1159, Jun 2004. 39, 73
- [167] X. HUANG AND A. MADAN. **CAP3: A DNA sequence assembly program.** *Genome Res.*, **9**:868–877, Sep 1999. 39, 73
- [168] HENG LI, BOB HANDSAKER, ALEC WYSOKER, TIM FENNEL, JUE RUAN, NILS HOMER, GABOR MARTH, GONĀGALO R. ABECASIS, AND RICHARD DURBIN. **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics*, **25**(16):2078–2079, 2009. 50, 74
- [169] I. G. WILSON. **Inhibition and facilitation of nucleic acid amplification.** *Appl. Environ. Microbiol.*, **63**:3741–3751, Oct 1997. 65, 66
- [170] M. A. VALASEK AND J. J. REPA. **The power of real-time PCR.** *Adv Physiol Educ*, **29**:151–159, Sep 2005. 65, 66
- [171] N. J. LENNON, R. E. LINTNER, S. ANDERSON, P. ALVAREZ, A. BARRY, W. BROCKMAN, R. DAZA, R. L. ERlich, G. GIANNOUKOS, L. GREEN, A. HOLLINGER, C. A. HOOVER, D. B. JAFFE, F. JUHN, D. McCARTHY, D. PERRIN, K. PONCHNER, T. L. POWERS, K. RIZZOLO, D. ROBBINS, E. RYAN, C. RUSS, T. SPARROW, J. STALKER, S. STEELMAN, M. WEIAND, A. ZIMMER, M. R. HENN, C. NUSBAUM, AND R. NICOL. **A scalable, fully automated process for construction of sequence-ready barcoded libraries for 454.** *Genome Biol.*, **11**:R15, 2010. 66

REFERENCES

- [172] 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. 66, 73
- [173] ANDREW ADEY, HILARY MORRISON, XU XUN, JACOB KITZMAN, EMILY TURNER, BETHANY STACKHOUSE, ALEXANDRA MACKENZIE, NICHOLAS CARUCCIO, XIUQING ZHANG, JACOB SHENDURE, EMILY TURNER, BETHANY STACKHOUSE, ALEXANDRA MACKENZIE, NICHOLAS CARUCCIO, XIUQING ZHANG, AND JAY SHENDURE. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol.*, 11(12):R119, 2010. 67
- [174] S. KRYAZHIMSKIY AND J. B. PLOTKIN. The population genetics of dN/dS. *PLoS Genet.*, 4:e1000304, Dec 2008. 67
- [175] 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. 67
- [176] 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. 67
- [177] BRENT EWING, LADEANA HILLIER, MICHAEL C. WENDL, AND PHIL GREEN. Base-Calling of automated sequencer traces using Phred. I. Accuracy Assessment. *Genome Res.*, 8(3):175–185, March 1998. 71
- [178] PHIL GREEN. PHRAP documentation., 1994. 71, 73
- [179] 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. 73
- [180] JAMES WASMUTH AND MARK BLAXTER. prot4EST: Translating Expressed Sequence Tags from neglected genomes. *BMC Bioinformatics*, 5(1):187, 2004. 74
- [181] A. BAIROCH, L. BOUGUELERET, S. ALTAIRAC, V. AMENDOLIA, A. AUCHINCLOSS, G. ARGOUUD-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. ROGLI, S. DUVAUD, A. ESTREICHER, L. FAMIGLIETTI, M. FEUERMANN, S. GEHANT, N. FARRIOL-MATHIS, S. FERRO, E. GASTEIGER, A. GATEAU, V. GERRITSSEN, 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. SONESON, S. STAELHI, 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. EBERRHARDT, A. FEDOTOV, R. FOULGER, J. GARAVELLI, R. GOLIN, A. HORNE, R. HUNTELEY, 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. 74
- [182] C ISELI, CV JONGENEEL, AND P BUCHER. ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc Int Conf Intell Syst Mol Biol*, pages 138–148, 1999. 74
- [183] RALF SCHMID AND MARK L BLAXTER. annot8r: GO, EC and KEGG annotation of EST datasets. *BMC Bioinformatics*, 9:180, 2008. 74
- [184] 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. 74
- [185] E. M. ZDOBNOV AND R. APWEILER. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, 17:847–848, Sep 2001. 74
- [186] A. KASPRZYK. BioMart: driving a paradigm change in biological data management. *Database (Oxford)*, 2011:bar049, 2011. 74
- [187] 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. 74
- [188] 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. 74
- [189] S. FALCON AND R. GENTLEMAN. Using GOstats to test gene lists for GO term association. *Bioinformatics*, 23:257–258, Jan 2007. 74
- [190] P. DANECEK, ÅÄÄ. AUTON, G. ABECASIS, ALBERS CA., E. BANKS, M.A. DEPRISTO, HANDSAKER RE., LUNTER G., MARSH GT., SHERRY ST., McVEAN GT., DURBIN T., AND 1000 GENOMES PROJECT. The variant call format and VCFtools. *Bioinformatics*, 27:2156–2158, Aug 2011. 74

REFERENCES

- [191] 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. 75
- [192] 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. 75
- [193] 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. 75
- [194] W. COLTMAN, PILKINGTON J. G., SMITH J. A., AND PEMBERTON J.M. **Parasite-mediated selection against inbred Soay sheep in a free-living, island population.** *Evolution*, **81**:1259–1267, 1999. 75
- [195] MARTIN MORGAN AND HERVÉ PAGÈS. **Rsamtools: Import aligned BAM file format sequences into R / Bioconductor.** R package version 1.4.3. 75
- [196] S. ANDERS AND W. HUBER. **Differential expression analysis for sequence count data.** *Genome Biol.*, **11**:R106, 2010. 58, 75
- [197] JH BOON, VMH CANNERTS, H. AUGUSTIJN, MAM MACHIELS, D. DE CHARLEROY, AND F. OLLEVIER. **The effect of different infection levels with infective larvae of *Anguillicola crassus* on haematological parameters of European eel (*Anguilla anguilla*).** *Aquaculture*, **87**(3-4):243–253, 1990. 75
- [198] O.L.M. HAENEN, T.A.M. VAN WIJNGAARDEN, AND F.H.M. BORGSTEDE. **An improved method for the production of infective third-stage juveniles of *Anguillicola crassus*.** *Aquaculture(Amsterdam)*, **123**(1-2):163–165, 1994. 76

Declaration

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

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

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

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

KARLSRUHE,

Emanuel G. Heitlinger

Redtenbacherstr. 9
76133 Karlsruhe
Germany

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

Born September, 12th in Schwäbisch Gmünd, Germany

Education

2008-2012 Doctoral studies, Karlsruhe Institute of Technology

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

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

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

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

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

Main subject: Zoology

Subsidiary subjects: Genetics, Botany

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

June 2000 Abitur (general qualification for university entrance).

1987-1991 Primary 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-parasite interactions, Transcriptomics, Genomics

Research

Peer Reviewed Publications

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

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

Conference Presentations

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

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

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

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

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

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

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

Honors, Awards, & Fellowships

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

Last updated: December 3, 2011