Bioinformatic Analysis of microRNA Genes in Free-Living and Parasitic Nematodes

Rina Ahmed

November 2014

DISSERTATION

zur Erlangung des akademischen Grades der Doktorin
der Naturwissenschaften (Dr. rer. nat.)

eingereicht im Fachbereich Mathematik und Informatik
der Freien Universität Berlin

Begutachtet von:
Prof. Dr. Martin Vingron
Prof. Dr. Kris Gunsalus

1. Gutachter: Prof. Dr. Martin Vingron
2. Gutachterin: Prof. Dr. Kris Gunsalus

Disputation: 26. Februar 2015

Preface

The work that led to this thesis is part of two collaborative projects in which I participated. This thesis presents results from both projects. A panel of different bioinformatics and statistical methods suitable to analyze small RNA deep sequencing data were identified and developed. Individual contributions for each project will be detailed here:

Flexbar Project The work presented in Chapter 3 was published in the special issue "Next-Generation Sequencing Approaches in Biology" in the journal *Biology*[1]. The Flexible Barcode and Adapter Remover (FLEXBAR) originated from the Flexible Adapter Remover (FAR) and has been developed by Matthias Dodt in the bioinformatics group of Dr. Christoph Dieterich. As part of this project, I developed the adapter removal feature for SOLiD color space reads and focused on the application of small RNA-seq in letter and color space. Additionally, I was involved in the design of FAR and in the development of specific features of the subsequently added barcode detection function for demultiplexing. The final version of FLEXBAR (paper version) has been extensively revised and enhanced by Johannes Röhr through the introduction of novel and extended features, a cleanup in the source code, redesigned command-line interface, and optimized parameter settings.

miRNA Project The bioinformatics workflow and the analysis and results presented in Chapter 2 and 4 were published in *Genome Biology and Evolution*[2]. As part of this collaborative project, I designed and performed all computational experiments. The experimental data sets were generated in the group of Dr. Christoph Dieterich at the Berlin Institute for Medical Systems Biology (BIMSB) which is part of the Max-Delbrück-Center for Molecular Medicine (MDC). The total RNA of the parasite samples (*Strongyloides ratti*) were kindly provided by our collaborator PD. Dr. Norbert W. Brattig from the Bernhard Nocht Institute for Tropical Medicine in Hamburg. All

next-generation sequencing was performed in the group of Dr. Wei Chen at BIMSB.

Acknowledgements The research presented in this thesis was funded by the MDC-NYU Exchange Program and was carried out at BIMSB in the group of Dr. Christoph Dieterich and the Center for Genomics and Systems Biology at the New York University (NYU) in the group of Prof. Dr. Kris Gunsalus. In the following, I would like to thank all people who have supported and helped me throughout my PhD studies:

First of all, I would like to thank my supervisor Christoph Dieterich for giving me the opportunity to pursue this research in his lab and introducing me to the fascinating world of next-generation sequencing. I am very grateful for his ideas, support, and fruitful discussions throughout the years and for giving me the opportunity to attend international conferences. I especially want to thank my co-supervisor Kris Gunsalus for giving me the chance to work in a very stimulating research environment and tightly connected with the wet lab group of Fabio Piano. Kris provided a creative and open minded working environment and I am very grateful for her dedication and support inside and outside the lab. I would also like to thank Martin Vingron for taking the time to supervise my PhD thesis as my University advisor. Furthermore, I am very grateful to all present and former members of the Dietrich, Gunsalus, and Piano groups and all people from BIMSB for creating this inspiring working atmosphere with plenty of joyful coffee breaks. Special thanks goes to Nikolaus Rajewsky and Jutta Steinkötter for their guidance and dedication and for providing this excellent research program. Moreover, I have to thank Jennifer Stewart and Sabrina Deter for helping me with all organizational issues and loads of paper work. Last but not least, I would like to thank my friends and family for their tremendous support and patience.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Objectives and Thesis Structure

Objectives The goal of this work is to identify and develop bioinformatics methods and computational strategies to analyze small RNA deep sequencing data from free-living and parasitic nematodes. By doing so, I want to address the question whether miRNA genes impact developmental arrest and long-term survival in dauer and dauer-like stages, i.e. the infective stage of parasites. In particular, I want to address the long-standing hypothesis that dauer and infective larvae share a common origin. This investigation is specifically focused on determining whether shared 'dauer-infective' miRNA expression signatures exist. This work will expand on previous studies and will present a comprehensive profiling of known and novel miRNA genes in free-living and parasitic nematodes with emphasis on developmentally arrested stages. Furthermore, it will be the first to identify miRNA genes in any *Strongyloides* parasite. The questions that I will address are significant because they will reveal important aspects of dauer and dauer-like biology with emphasis on potential conserved miRNA regulatory mechanisms in free-living and parasitic nematodes.

Thesis Outline In Chapter 1 I provide an introduction to the animal models investigated in this study, the biological mechanisms involved in post-transcriptional regulation, and the experimental techniques of next-generation sequencing. Chapter 2 describes the bioinformatics methods and strategies used and implemented to analyze the small RNA deep sequencing data investigated in this study. In particular, six computational analysis steps of the applied bioinformatics workflow are described in detail. Moreover, the concept and usage of the flexible barcode and adapter remover

FLEXBAR, which was developed in our group, is introduced in Chapter 3. In Chapter 4 the methods introduced in Chapter 2 and 3 are applied to small RNA sequencing data profiled from developmental arrested stages of free-living and parasitic nematodes and the results are presented. Finally, Chapter 5 summarizes the analyses presented, discusses the results, and puts them in a broader context. Additionally, future directions are outlined.

1.2 The Animal Models

Nematodes, or roundworms, are one of the most diverse group of animals and species-rich phyla, comprising an estimated ∼1-10 million species. Nevertheless, only ∼25,000 have been described to date[3]. Based on molecular studies nematodes are grouped into five major taxonomic groups (Figure 1.1)[4]. Nematodes inhabit a wide range of ecological niches. As free-living species, as well as parasites of plants and animals including humans, they occupy freshwater, terrestrial, and marine environments[5]. Blaxter *et al.* (1998) suggested that parasitism evolved at least seven times independently. Early reports of parasitic nematodes date back to mummies in Egypt[6].

Nematode adults can be found as either females, males or self-fertilizing hermaphrodites, depending on the species. Soil nematodes are normally very small (0.3 to 3 mm long), whereas parasites of insects and mammals can be many centimeters long. *Placentonema gigantissima*, the largest known nematode, lives in the placenta of sperm whales, has a body volume of ∼174 liters, and is up to 8-9 meters in length[6]. Despite inhabiting diverse ecological niches, the basic life style of nematodes is conserved and typically involves four larval molts[7].

1.2.1 *Caenorhabditis elegans*

Over the last decades, the free-living non-parasitic nematode *C. elegans* has been established as a key model system for research on neurobiology, embryogenesis, and gonadal development[13,14]. In the late 1970's and early 1980's, the cell lineage of every single somatic cell was mapped[15]. *Caenorhabditis elegans* is easy to maintain in the laboratory with a food supply of *Escherichia coli* (*E. coli*); self-fertilization is the typical mode of reproduction, and a complete life cycle spans three days (20°C)[16]. In 1998, *C. elegans* became the first metazoan to have its genome fully sequenced[17]; the completely assembled genome is 100 megabases (Mb) in size and contains an estimated

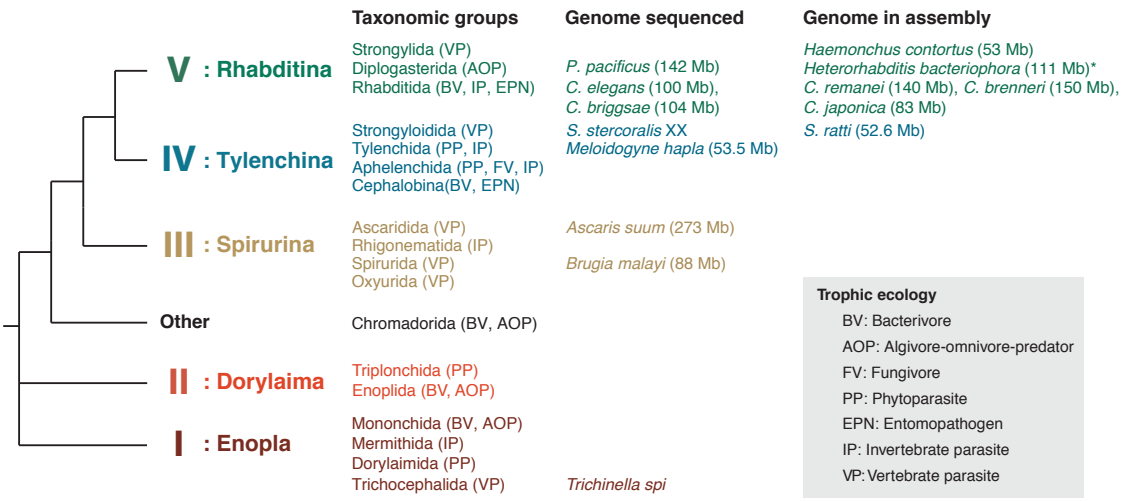

Figure 1.1: Major taxonomic groups of the phylum Nematoda
Overview of nematode phylogeny and the major taxonomic groups. Roman numerals indicate the clade according to Blaxter *et al.* (1998). The column 'Genome sequenced' lists all nematode species with published genome sequence whereas ongoing genome projects are listed in the column 'Genome in assembly'. The numbers in brackets indicate the genome sizes in megabases (Mb)[8–11]. Asterisk (*) denotes that *Heterorhabditis* form a genus of its own. Figure adapted and modified from Sommer and Streit (2011) and Blaxter (1998).

20,060 protein-coding genes[18].

Caenorhabditis elegans is an excellent model to study developmental responses to environmental changes. Under favorable conditions, development consists of embryogenesis followed by four larval stages (L1-L4) (Figure 1.2A)[16]. If embryos hatch in the absence of food, newly hatched L1 larvae enter L1 diapause and are able to survive in this state for several weeks[19]. If conditions sensed during L2 development are unfavorable, such as starvation and crowding, worms enter an alternative third larval stage, the dauer stage[20]. Dauer larvae are developmentally arrested, nonfeeding, stress-resistant, long-lived, and prevail in nature[21]. Dauer larvae are adapted morphologically and physiologically to remain in the environment without feeding for up to four to eight times the normal 2-week lifespan[22]. By transportation through insects or other invertebrates, dauer larvae can search for new food sources. The association of *C. elegans* dauer larvae with insects or other invertebrates for dispersal is not specific. Non-specific host association for transportation is called phoresy[23]. Recovery from dauer is initiated once environmental conditions become favorable, particularly a high food to pheromone ratio and low temperature[19]. The dauer stage is considered to be non-aging because

the duration of dauer stage does not affect post-dauer lifespan[22]. Strikingly, all life histories, whether continuous or interrupted, involve an identical pattern and sequence of cell division and cell fates[15].

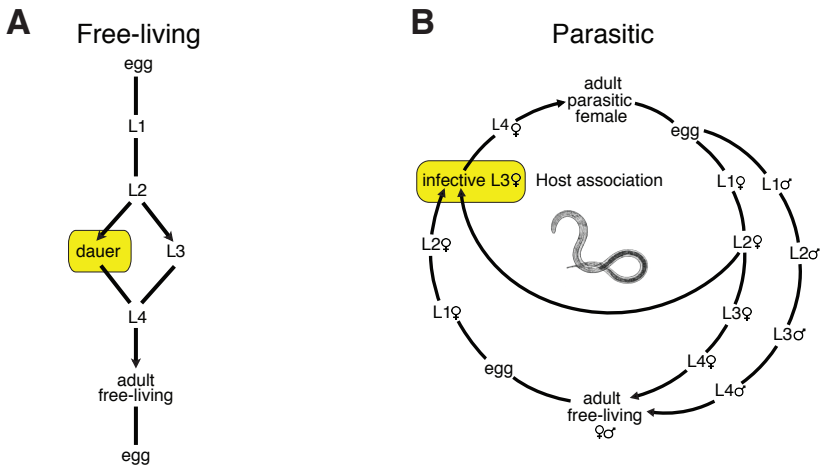

Figure 1.2: Life cycle of free-living and parasitic nematodes
(A) Under conditions that are favorable for reproduction, free-living larvae, such as *Caenorhabditis elegans*, develop through four larval stages. Under unfavorable environments L2 larvae enter dauer diapause. **(B)** Infective larvae of parasites, such as *Strongyloides ratti*, develop either directly or after a facultative sexual free-living adult generation.

The molecular mechanisms underlying dauer developmental transitions have been at least partially defined. As discussed above, entry and exit from dauer stage are developmental responses to specific chemosensory cues that inform the larva whether there is an abundant food supply to support reproduction. The nuclear receptor DAF-12 plays a critical role in the decision to enter dauer. Under optimal conditions DAF-12 is bound by the steroid hormone dafachronic acid (DA) resulting in nondauer development[24]. Under unfavorable conditions DA concentration is low and DAF-12 is bound by DIN-1 which specifies the dauer fate[25]. Other signaling pathways involved in dauer formation are the insulin/IGF and the TGF-β pathways[21]. Mutating components of these signaling pathways, i.e. dauer formation genes, results in dauer-constitutive (Daf-c) or dauer-defective (Daf-d) phenotypes. Daf-c mutants form dauer under normal conditions and Daf-d mutants fail to enter dauer under harsh conditions[26].

1.2.2 *Pristionchus pacificus*

Pristionchus pacificus (*P. pacificus*) is a nematode that has been established as a satellite model system to *C. elegans* for the study of evolutionary developmental biology[27]. *Pristionchus pacificus* has an assembled genome size of 142 Mb containing 24,231 protein-coding genes[11]. According to Blaxter and colleagues, who distinguish five major nematodes clades, both *P. pacificus* and *C. elegans* belong to the clade V nematodes (Figure 1.1)[4]. Based on nucleotide divergence, these two species were estimated to share a common ancestor 280-430 million years ago, and thus represent distantly related species of the same phylum[28]. *P. pacificus* lives in species-specific association with the oriental beetle *Exomala orientalis*. The self-fertilizing *P. pacificus* can be easily grown under laboratory conditions with *E. coli* as food source, where it can achieve a short generation time of 4 days (20°C). Similar to *C. elegans* under favorable conditions, *P. pacificus* develops through four larval stages (L1-L4) to the reproductive adult (Figure 1.2A). In conditions of starvation or crowding, *P. pacificus* enter the developmentally arrested, nonfeeding, and long-lived dauer stage. Dauer larvae actively invade the beetle and remain arrested until the death of the beetle. Upon the beetles death, *P. pacificus* dauer larvae resume development by feeding on bacteria, fungi, and other nematodes that grow on the insects carcass[29]. This species-specific association with a host, waiting for its death, and feeding on the microbes developing on the carcass, is called *necromeny*. It has been argued that phoretic (*C. elegans*) and necromenic (*P. pacificus*) associations serve as important pre-adaptations for the evolution of parasitism[23,30–33].

1.2.3 Relationship with Parasitic Nematodes

Nematode parasitism is a worldwide health problem with over 1 billion people being infected[34]. Due to their wide range of host targets ranging from plants to animals, including humans, parasitic nematodes are of importance to human and veterinary medicine, as well as agriculture[35]. However, the molecular mechanisms controlling the infection with parasites is poorly understood. Working with parasitic nematodes is usually complicated due to their more complex life cycles in comparison with *C. elegans* and *P. pacificus*, and only a few parasitic species can be cultured in the laboratory. One such example is the animal parasite *Strongyloides ratti* (*S. ratti*), whose life cycle has two phases: a parasitic phase and a free-living phase (Figure 1.2B). Parasitic adults live in the mucosa of the small intestine of rats. These are females that reproduce

by parthenogenesis, giving rise to both parasitic and free-living progeny that undergo two types of development: the so-called *homogonic* (direct) and *heterogonic* (indirect) modes of development. In homogonic development, after two larval stages (L1 and L2) the female offspring develop into infective L3 larvae (iL3), which infect a new host by skin penetration. The iL3 stage is developmentally arrested and will only develop further if it encounters a host. In heterogonic development, larvae develop through L1-L4 and become sexually reproducing free-living adults. All of the progeny of free-living adults develop into female iL3s. All males arise as free-living offspring of parasitic adults and pass through heterogonic development[36].

Dauer larvae of free-living nematodes like *C. elegans* and *P. pacificus* share morphological, behavioral and physiological traits with infective larvae of true parasitic species; dauer and infective larvae are the third larval stage, both have a slender appearance, a constricted esophagus, a closed mouth, and show host seeking-behavior like nictation[37–39]. Moreover, the dauer and infective larvae fate is determined by a conserved endocrine signaling mechanism. Notably, the DA/DAF-12 module which is required for dauer formation in *C. elegans*, is conserved in the necromenic nematode *P. pacificus* and the parasitic nematode *Strongyloides papillosus* (*S. papillosus*)[33]. Ogawa and colleagues showed that Δ7-DA strongly suppresses dauer formation in *P. pacificus*. Furthermore, the authors demonstrated that in the presence of Δ7-DA, the progeny of parasitic females of *S. papillosus* developed into free-living animals and that the formation of iL3s was completely inhibited. A different study observed similar results in the human parasite *Strongyloides stercoralis* and in the hookworm *Ancylostoma caninum*[40].

Several lines of evidence suggest that post-transcriptional regulatory mechanisms dominate the transition from dauer back into the reproductive life cycle. Intriguingly, RNA polymerase II transcription appears to be reduced in dauer larvae relative to other stages based on run-on transcription assays with isolated nuclei[41]. Moreover, the process of dauer exit is impaired by translational repression with cycloheximide but not by the inhibition of mRNA synthesis with either amanitin or actinomycin D[42,43], suggesting that mRNA synthesis is not necessary for dauer recovery[44]. Taken together, these results suggest that transcripts might be accumulated before dauer diapause or during dauer entry and that their activity is controlled during dauer and exit from dauer by post-transcriptional regulation.

1.3 Post-transcriptional Regulation of Gene Expression

The central dogma of molecular biology first formulated by Francis Crick in 1958 states that genetic information encoded in deoxyribonucleic acid (DNA) is transcribed to produce ribonucleic acid (RNA), which simply acts as a messenger molecule harboring the instructions for the translation of proteins which are ultimately responsible for cell phenotype[45,46]. Consequently, gene expression was thought to be regulated in a unidirectional way. It became clear that this view was too simplistic. To maintain proper cell functions, it is essential that all of the molecular steps necessary for the production of protein from DNA are highly regulated in a precise spatial and temporal manner.

In eukaryotes, gene expression is tightly regulated at the level of transcription by (i) transcription factors, proteins that bind to the DNA near transcription start sites and activate or inhibit the transcription of genes[47], and (ii) chromatin state, by epigenetic modifications that control the accessibility and thus the readability of genes through methylation and acetylation of histones[48]. Following transcription, messenger RNAs (mRNAs), are not directly translated into proteins because mRNA processing, localization, stability, and translation are regulated post-transcriptionally[49]. These regulatory processes are controlled and mediated by RNA-binding proteins (RBPs) and small (or short) non-coding RNAs (small ncRNAs) that form dynamic multicomponent ribonucleoprotein complexes (RNPs) that generally bind to functional sequences located in the untranslated regions (UTRs) of target mRNAs[50,51].

In 2001, the completion of the human genome project revealed that less than 2% of the human genome encodes information for protein-coding genes[52]. The rest of the genome was long termed as *junk* DNA without any functionality[53]. However, high-resolution transcriptomic studies revealed that the majority of the genome is transcribed into RNA without being processed into protein[54,55]. Instead these RNA transcripts constitute a large family generally termed non-coding RNAs (ncRNAs). Ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs) are examples of two well-known classes of ncRNAs. It has been suggested that the majority of ncRNAs are functional due to the fact that the percentage of ncRNAs transcribed in the genome is proportional to the complexity of the organism[56]. Indeed, a recent study of the ENCODE (Encyclopedia of DNA Elements) project assigned biochemical functions to 80% of the genome outside of protein-coding regions[57].

The large family of ncRNA transcripts consists of various different classes which can be distinguished by their function and molecular similarities, such as the length of a molecule, as in long ncRNAs (lncRNAs; defined as being longer than 200 nt) and small ncRNAs[58,59]. In addition, regulatory small ncRNAs can be divided into three major groups: (i) endogenous miRNAs, (ii) PIWI-interacting RNAs (piRNAs), and endogenous short interfering RNAs (endo-siRNAs). miRNAs generally silence host gene expression, pi-RNAs silence transposable elements in animal germ cells, and endo-siRNAs can be involved in host defense through viral RNA silencing[58]. In the following, I will emphasize miRNA genes, a key post-transcriptional regulator of almost all biological processes investigated[60].

1.3.1 microRNA Genes

miRNAs constitute a large family of ∼22 nt endogenous, small ncRNA molecules that downregulate the expression of protein-coding target genes at the post-transcriptional level[61]. The first miRNA, *lin-4*, was found in 1993. Initially, the authors identified a small RNA, *lin-4*, in the worm *C. elegans* that negatively regulates the production of the heterochronic gene LIN-14 by binding partially complementary to the 3' UTR of LIN-14[62,63]. LIN-14 encodes a protein that controls the devision timing of specific cells in *C. elegans* during postembryonic development. In 1981, Chalfie and colleagues showed that mutations in *lin-4* disrupt the temporal regulation of larval development, causing cell-devision patterns specific to the first larval stage to reiterate in later developmental stages. These studies demonstrated that the miRNA *lin-4*, is essential for a correct transition between developmental stages. Seven years later, in 2000, Ruvkun and colleagues discovered a second miRNA, *let-7*, again in *C. elegans*[65]. In contrast to *lin-4*, *let-7* was found to be highly conserved across the bilaterian phylogeny[66]. Since then, new miRNAs have been identified by small-RNA-cloning strategies in animal[67–69] and plant species[70,71]. More recently, thousands of miRNA genes have been discovered by experimental approaches, computational predictions, or combined strategies across the animal and plant kingdoms[72–74] and all published miRNA sequences and associated annotations have been catalogued in a public accessible repository called *miRBase*[75]. Increasing evidence suggest that miRNAs are not only key regulators of organismal development, but influence almost all biological processes investigated, including cellular differentiation, metabolism, and viral infection[60]. Consequently, dysregulation of miRNA function are observed in human pathologies such as cancer[76,77].

In animals, canonical miRNAs are typically transcribed by RNAPII, either through an independent promoter or as part of a host gene embedded within an intron of a protein-coding gene, as capped and polyadenylated primary miRNA (pri-miRNA) transcripts (Figure 1.3). Pri-miRNAs are often several kilobases (Kb) long and contain one or more stem-loop structures that house the functional ∼22 nt miRNA[78,79]. Subsequently, pri-miRNAs are cropped by the RNase III enzyme Drosha, which works in a complex with DGRC8 (in worms and flies also known as Pasha), generating a ∼65 nt precursor miRNA (pre-miRNA)[80,81]. In vertebrates and flies, pre-miRNAs (also called *hairpins*) are then exported from the nucleus into the cytoplasm by Exportin-5 (Exp5), a nuclear export protein[82–84]. Interestingly, in nematodes, a homolog of Exp5 is lacking and the cellular location of miRNA processing events is still unknown[85]. Once in the cytoplasm, pre-miRNAs are cleaved by Dicer (DCR-1 in worms and flies), another RNase III enzyme, into a ∼22 nt double-stranded RNA (dsRNA) with a characteristic 2-nt single-stranded 3' overhang on both ends.[86–89].

Apart from the canonical miRNA biogenesis pathway, alternative maturation strategies have been suggested recently (Figure 1.3). A distinct class of miRNAs located within short introns bypass Drosha cleavage and instead use the spliceosomal machinery to generate pre-miRNAs. These miRNAs, referred to as *mirtrons*, have been identified in flies, nematodes, and mammals[90–94]. Mirtrons are excised, debranched, and refolded by a lariat-debranching enzyme to form a hairpin structure that resembles a pre-miRNA which can be fed into the miRNA pathway. In addition, several mirtrons have been discovered in flies that are generated with extended 3' tails which need to be excised by the exosome before Dicer processing[95].

Following Dicer cleavage, one strand (or *arm*) of the RNA duplex, typically known as *guide* or *miRNA mature* (*miRNA*), is incorporated into the RNA-induced silencing complex (RISC) forming miRISC (miRNA-induced silencing complex), whereas the other strand, known as *passenger* or *miRNA star* (*miRNA**), is degraded. Studies indicate that the relative thermodynamic stability of the two ends of the RNA duplex determines which strand is loaded into RISC[96,97]. However, evidence suggests that miRNA sequences can be produced from both strands of a precursor at similar frequencies and that both strands might even be biologically functional[98–104]. Interestingly, the dominant strand produced from a precursor can vary in a cell-context and tissue-dependent fashion or between orthologous miRNAs[105–110].

To regulate target mRNAs, miRNAs bind primarily to 3' UTRs, inducing mRNA degradation and/or translational repression through multiple mechanisms[111,112]. miRNAs

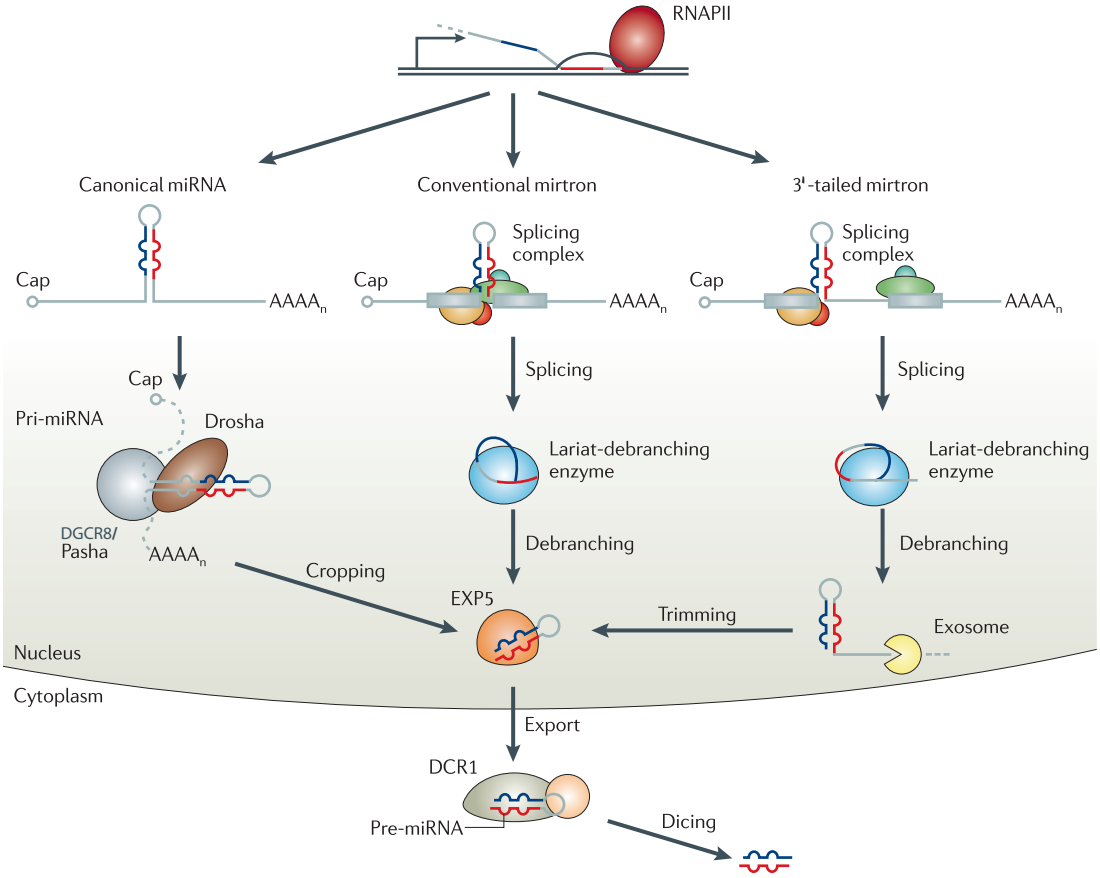

Figure 1.3: miRNA biogenesis pathway
Canonical animal miRNAs are transcribed by RNA polymerase II (RNAPII) as capped and polyadenylated primary miRNA (pri-miRNA) transcripts that are subsequently cleaved by the Drosha-complex releasing a ∼65 nt precursor miRNA (pre-miRNA). In an alternative pathway, conventional mirtrons and 3' tailed mirtrons are derived from small introns that are spliced, debranched and refolded to form a pre-miRNA. As the name suggests, 3' tailed mirtrons have tails at their 3' end that need to be trimmed by an exosome before further processing. Pre-miRNAs are then exported to the cytoplasm by Exportin-5 (Exp5), where they are processed by the Dicer-complex generating a ∼22 nt RNA duplex containing the guide and passenger strand of a miRNA. Figure adapted from Czech and Hannon (2011).

recognize their targets by partial complementarity and interactions involving the seed region, oftentimes nucleotide 2-8 of the miRNA, although other valid interactions modes for biological relevant miRNA/mRNA pairs exist[113]. Identifying miRNA target genes and their regulatory sequences is an outstanding problem in the miRNA research field, since apparent flexibility in miRNA-targeting rules suggest that other factors mediate

functional target interactions *in vivo* rather than just paring capacity. Despite the complex nature of this problem, multiple complementary approaches were implemented to tackle this problem[114]. For example, it is now possible to identify endogenous miRNA target sites by biochemical methods like CLIP-seq, HITS-CLIP, or PAR-CLIP, where sequences that are bound by specific RNA-binding proteins are isolated and then identified using high-throughput sequencing. Since Argonaute family proteins (Ago) form the core of miRISC, crosslinking and immunoprecipitation (CLIP) of Ago has been used to identify miRISC-binding sites on a genome-wide scale[115–118].

Because a single miRNA can target multiple mRNAs at the same time, miRNAs can potentially coordinate rapid changes in gene expression in response to environmental, developmental and physiological cues[113,119]. Several recent studies provide evidence that miRNA genes are involved in the regulation of lifespan as well as L1 and dauer diapause[120–126]. A recent study identified 17 microRNAs whose expression profiles are altered by dauer life history in comparison with continuous development[125]. More specifically, it has recently been shown that miRNAs play critical roles in the survival and recovery from starvation-induced L1 diapause[124], and that a feedback loop involving DAF-12 and *let-7* family miRNA members coordinate cell fate decisions with starvation-induced dauer arrest[121,122,126].

In summary, miRNA genes have emerged as key regulators in diverse biological pathways and pathologies, including organismal development, cellular differentiation, and metabolism[60]. Hence, determining the *miRNAome* and their target genes is of great importance and will help us to understand the functional content of genomes and regulatory networks involved. To date, major efforts have been made to profile and discover novel miRNA genes in eukaryotes by experimental approaches, computational predictions, or combined strategies; the latest miRBase[75] release (v21, June 2014) contains 28,645 precursor sequences from 223 different species*. The recently developed next-generation sequencing technologies are very promising methods for miRNA profiling and discovery.

1.4 Next-Generation Sequencing

Determining the order of nucleotides in DNA or RNA sequences is known as *sequencing*. The goal of the Human Genome Project (HGP), which was officially founded in 1990,

*http://www.mirbase.org; accessed June 2014

was to sequence the entire human DNA. After 10 years of hard work, a draft version of the human genome was announced in 2000 and published in February 2001[127,128]. The final completion of the human genome was announced in April 2003[129]. The sequencing approach used >20,000 large bacterial artificial chromosome (BAC) clones each containing a ∼100 Kb fragment of the human genome. Determined by physical mapping, these fragments provided an overlapping set (tiling path) through each human chromosome[127]. BAC-based sequencing is a complicated and complex approach involving many experimental steps that are time consuming and costly. Since the completion of the HGP, genome sequencing has moved away from BAC-based approaches toward whole-genome sequencing (WGS)[130]. Using WGS approaches, genomes can be sequenced more rapidly and are easier to read, but highly polymorphic or repetitive regions are difficult to assemble and remain fragmented after assembly. However, all BAC-based and WGS approaches rely on the same type of capillary sequencing instruments based on Sanger or di-deoxy terminator strategy (Figure 1.4A)[131].

In 2005, the first so-called next-generation sequencing (NGS) technologies (or 2nd generation sequencer) became commercially available[132]. NGS approaches are able to readout DNA or RNA sequences in a massively parallel manner at low cost per base. NGS has revolutionized diverse genomics applications, including *de novo* genome sequencing and re-sequencing, single nucleotide polymorphism (SNP) detection, transcriptome analysis including small non-coding RNAs, and chromatin immunoprecipitation[133]. Three NGS technologies are commonly used for massively parallel sequencing: Roche/454 pyrosequencing*, Illumina/Solexa sequencing by synthesis†, and Applied Biosystems (ABI) SOLiD™ sequencing by ligation‡. All 2nd generation sequencers are based on a template amplification phase before sequencing and can now produce hundreds of millions of short stretches of sequence, also called *reads*, of typically 35-400 base pairs (bp) in length (Sanger-based technologies produce reads up to 800 bp)[130,134]. Recently, so-called single-molecule sequencers (or 3rd generation sequencers), which avoid the amplification step, were introduced: the Helicos Heliscope™ § and Pacific Biosciences SMRT™ ¶ instruments[135]. However, highly repetitive and other genome-wide regions remain difficult to sequence. No current human genome is fully complete, fully accurate or certain to contain all rearrangements or haplotype information[136].

*http://www.454.com; accessed April 2014
†http://www.illumina.com; accessed April 2014
‡https://www.lifetechnologies.com; accessed April 2014
§http://www.helicosbio.com; website now defunct
¶http://www.pacificbiosciences.com; accessed April 2014

Despite the fact that all NGS platforms differ in their sequencing biochemistry, each workflow follows a similar cyclic-array sequencing strategy, in which a dense array of DNA features is sequenced by iterative cycles of enzymatic interrogation combined with imaging-based data detection (Figure 1.4B)[137,138]. Nevertheless, all technologies are quite divers in their features and overall performance (Table 1.1).

Table 1.1: Comparison of Sanger and next-sequencing platforms
All NGS platforms support single- and paired-end sequencing. Since NGS technologies are advancing rapidly, technical specifications and pricing are in flux. This table was adapted and modified from Liu *et al.* (2012) and Jessri and Farah (2014).

	Sanger 3730x	**Roche 454 GS FLX+**	**Illumina HiSeq™ 2500**	**SOLiD™ 5500XL**
Library preparation	*In vitro* cloning, picking, and growth	Emulsion PCR	Bridge amplification	Emulsion PCR
Sequencing principle	Dideoxy chain termination	Polymerase-based pyrosequencing	Polymerase-based sequencing by synthesis	Ligase-based sequencing and two-base encoding
Read length*	400 ∼ 900	700	36/50/100	50/75+35
Max output	1.9 ∼ 84 Kb	0.7 Gb	600 Gb	180 Gb
Run time	20 min ∼ 3 h	23 h	11 days; 27 h (rapid run)	7 days to 4 weeks
Machine cost	$95,000	$500,000	$690,000 (HiSeq 2000)	$495,000 (SOLiD 4)
Advantage	High quality; long read length	Long read length; short run time	High throughput; cost-effectiveness	High throughput; accuracy; inherent error correction through two-base encoding
Disadvantage	High cost; low throughput	High cost; high error rate in homopolymer repeats; low throughput; long hand-on time	Short read length	Short read length; long run time; complexity of library preparation

* measured in nucleotides

In the following, I will focus on the two NGS technologies (Illumina/Solexa and ABI SOLiD) that were applied in my research.

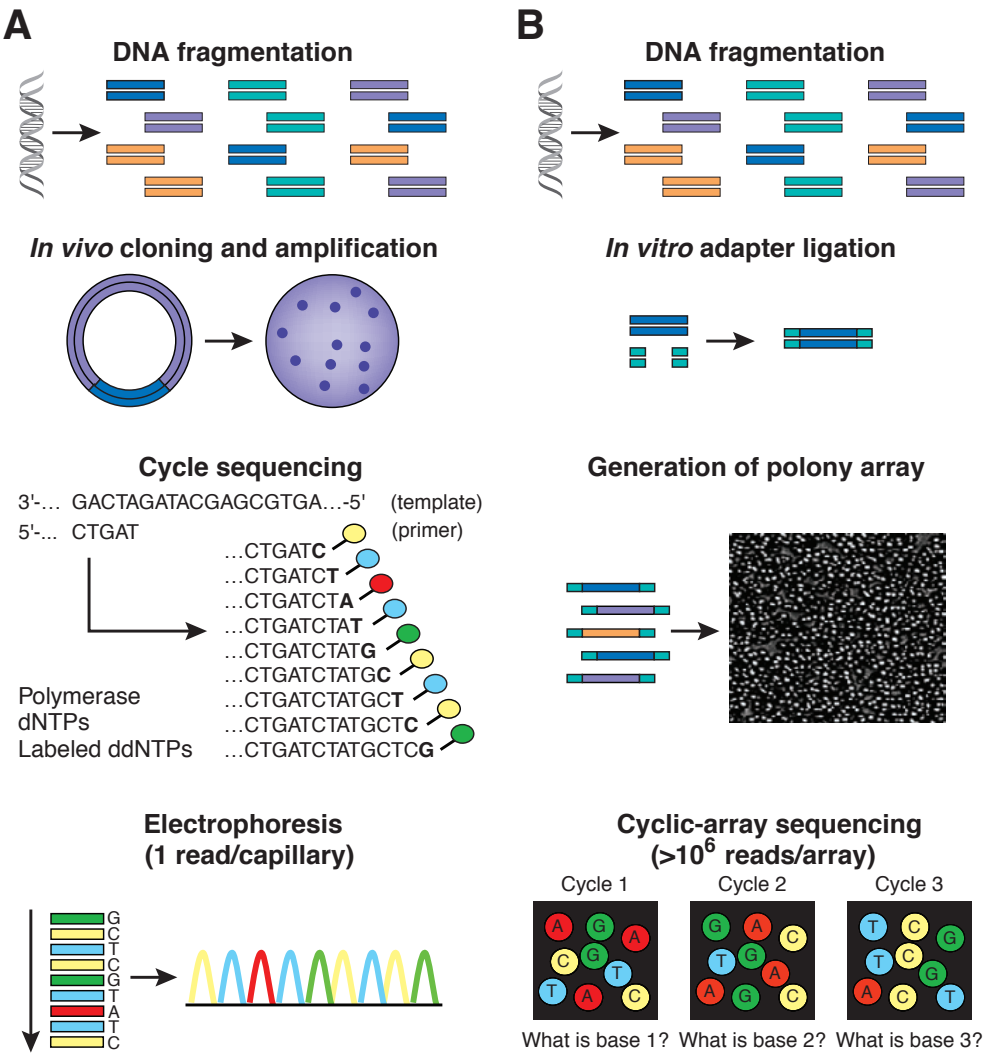

Figure 1.4: Workflow of Sanger versus next-generation sequencing
(A) In high-throughput shotgun *de novo* Sanger sequencing, randomly fragmented genomic DNA is cloned into a plasmid vector and then used to transform bacteria (e.g. *E. coli*). A single bacteria colony is picked and the plasmid DNA isolated, for each sequencing reaction. Each cycle sequencing reaction, in which cycles of template denaturation, primer annealing, and primer extension are performed, generates a ladder of fluorescently labeled dideoxynucleotides (ddNTPs). The sequence is determined by high-resolution electrophoretic separation in capillary-based polymer gel in one sequencing run. A detector generating a four-channel emission spectrum is passed by the fluorescently labeled fragments of discrete sizes to capture the sequencing trace. **(B)** In next-generation sequencing methods, an array of millions of immobilized colonies (or *polonies*) amplified by polymerase chain reaction (PCR) is generated as a result of treatment of fragmented genomic DNA after being ligated to common adapters. Each polony contains many copies of a single library fragment. A contiguous sequencing read is built for all array features in parallel in cyclic reactions by imaging-based detection of fluorescence labels. Figure adapted from Shendure and Ji (2008).

1.4.1 Illumina/Solexa System

Generally speaking, methods included in NGS technologies can be grouped into library preparation, sequencing, imaging, and data analysis[134]. Although the general cyclic-array sequencing strategy of different NGS technologies is conceptually similar (Figure 1.4B), important differences exist. Characteristics of the Illumina/Solexa system, which became commercially available in 2006 as the Genome Analyzer (GA), include the usage of bridge amplification for library preparation and a DNA polymerase-dependent sequencing by synthesis strategy (Figure 1.5A and B)[134].

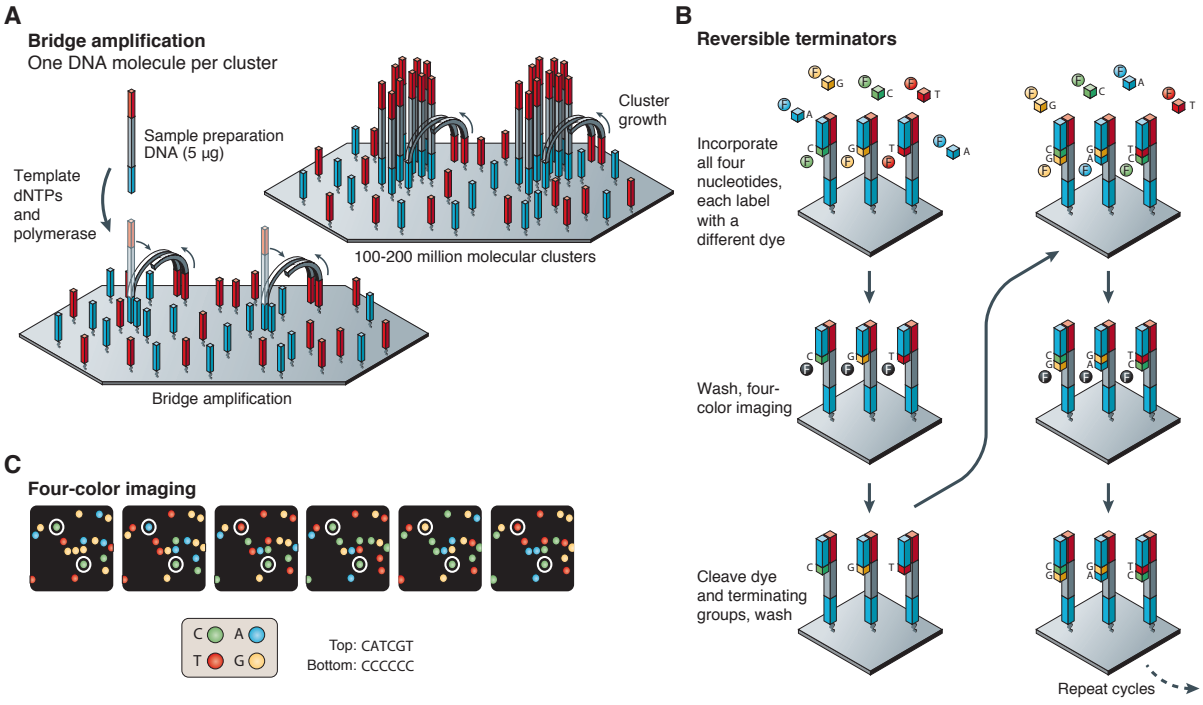

Figure 1.5: Illumina/Solexa sequencing system
(A) DNA template fragments are amplified through a process called *bridge amplification*. In this context, template fragments hybridize to complementary adapters on the slide with one end and bend to encounter a complementary second-end adapter. Immobilized template fragments are then amplified by a DNA polymerase to form clusters. **(B)** The four-color cyclic reversible termination method is used as sequencing strategy. Following bridge amplification, reversible terminators of all four nucleotides, each label carrying a different dye, are added simultaneously. During each cycle, one nucleotide at a time is incorporated by a DNA polymerase to all fragments on the slide. Following imaging, a cleavage step removes fluorescent dyes and regenerates the 3'-OH group. **(C)** Sequencing data from two amplified fragments are highlighted in the four-color images. Figure adapted and modified from Metzker (2010).

Following random fragmentation of input DNA to be sequenced, adapter sequences containing universal priming sites are ligated to the ends of the fragment templates. DNA templates flanked with adapters are then size fractionated to create the final template library. To be able to perform and detect billions of sequencing reactions simultaneously, the size fractionated library is clonally amplified and immobilized on a glass slide (flow cell), which is decorated by adapter sequences complementary to the library adapters. Clonal amplified clusters of identical library fragments are produced through a process called *bridge amplification*[142]. After the library is washed over the slide, DNA fragments hybridize to complementary adapters on the slide with one end and bend to encounter a complementary second-end adapter (Figure 1.5A). A DNA polymerase can now amplify the fragments to create a cluster of millions of copies of a single fragment at the same physical location. As a result, 100-200 million spatially separated template clusters are generated to which a universal sequencing primer can bind to initiate the sequencing process[134]. In addition to this single-end sequencing approach, all NGS technologies support paired-end sequencing, where sequence data from both ends of each template fragment is produced[143]. A single flow cell of the Illumina/Solexa machine consists of eight independent lanes each containing several million clusters. Therefore, eight independent samples can be sequenced in parallel during a single sequencing run[138]. Moreover, when using unique barcode sequences in a so called *multiplexed* sequencing run, it is possible to sequence up to 96 independent samples (12 samples per lane)*.

As sequencing strategy, the Illumina/Solexa system adopts a cyclic-array sequencing by synthesis approach, in which reversible terminators of all four nucleotides are added simultaneously for incorporation by a modified DNA polymerase (Figure 1.5B)[144]. During each cycle, just one fluorescent modified nucleotide carrying a unique base dye is added complementary to all immobilized fragments on the flow cell. The DNA synthesis is terminated due to a chemical block of the 3'-OH group of incorporated nucleotides. After each round of incorporation of reversible terminators, unincorporated nucleotides are washed away. Subsequently, an imaging system scans the flow cell stimulating individual dyes to emit light in specific wave lengths, which is detected by a sensitive camera (Figure 1.5C). Following imaging, a cleavage step regenerates the 3'-OH groups and removes the fluorescent dyes, such that the next cycle of reversible terminator incorporation can start. This process is repeated for a specific number of cycles as preset by user-defined instrument settings (up to 100x resulting in a read length of 100 nt;

*http://www.illumina.com/technology/multiplexing_sequencing_assay.ilmn; accessed April 15, 2014

Table 1.1)[134,138].

Each Illumina/Solexa system is distributed with software that generates primary readable data. This software includes a base calling algorithm that converts image-based signals into nucleotides and assigns quality values to each read. Poor-quality reads are directly removed by a quality checking step of the provided software[130]. However, depending on the experimental design and the genomic application, additional bioinformatics preprocessing steps are necessary for optimal results of downstream analysis (e.g. barcode detection after multiplexed sequencing runs or adapter removal in small RNA sequencing experiments). Detailed information about bioinformatics preprocessing steps involved in NGS analysis with emphasis on small RNA sequencing data is given in Materials & Methods 2.3.1.

Overall, the Illumina/Solexa platform is the most widely applied NGS approach and has the highest data output (Table 1.1)[134]. However, due to imperfect polymerase activity, the most common error type are substitutions and the accuracy decreases towards the 3' end of a read[134,145]. Illumina/Solexa machines produce at least 1 sequencing error every 200 bases (error percentage 0.5%)[143]. Moreover, Nakamura *et al.* (2011) suggested that together with possible biases during library construction and amplification, sequence-specific errors in sequencing are responsible for substantial coverage variations in read mapping when longer reads are produced. Despite these limitations, the Illumina/Solexa system is one of the most powerful technologies for DNA or RNA analyses including miRNA studies[93,106,146] and has been successfully applied within the modENCODE project[147].

1.4.2 ABI SOLiD™ System

The Applied Biosystems SOLiD (Support Oligonucleotide Ligation Detection) platform is based on the strategy described in Shendure *et al.* (2005) and on work by McKernan *et al.* (2006) at Agencourt Personal Genomics (Beverly, MA, USA) (acquired by Applied Biosystems (Foster City, CA, USA) in 2006). The SOLiD platform utilizes an entirely different approach for library amplification and sequencing compared to the Illumina/Solexa system. Instead of bridge amplification, ABI SOLiD applies a method called *emulsion PCR* (Figure 1.6A). Sequencing is accomplished by ligation using a DNA ligase[150] and two-base encoded probes (Figure 1.6B and C)[134].

As already mentioned, the preparation of a sequencing library is conceptually similar

in all NGS systems. Hence, to prepare the library for SOLiD sequencing, input DNA is fragmented, adapter ligated, size fractionated, and finally amplified by emulsion PCR. In this process, each DNA library fragment, after being single stranded, is captured on a magnetic bead. Each DNA-bead complex (millions of beads) is encapsulated in a single droplet, which is a mixture of an oil-aqueous emulsion containing primer, dNTPs, and polymerase. All fragments are then clonally amplified in parallel by PCR to thousands of copies each. As a result, all droplets contain one bead covered with copies of a single fragment. Following amplification, 100-200 million beads (each bearing amplification products) are chemically cross-linked to a glass slide to initiate the sequencing reaction (Figure 1.6A)[134]. SOLiD uses two slides per run, which can be partitioned into four or eight segments. Therefore, up to 16 individual samples can be sequenced in parallel during a single SOLiD run. Just as the Illumina/Solexa technology, SOLiD supports multiplexed sequencing with up to 96 unique barcodes for each of the segments, resulting in a multiplexing capacity of maximal 1,563 (96×16) individual samples per run*. This is much higher than the multiplexing capacity of Illumina/Solexa sequencers (up to 96 samples). In addition to single-end sequencing, SOLiD sequencers also support paired-end sequencing.

The sequencing methodology of SOLiD is based on sequential ligation with fluorescently labeled octamers (Figure 1.6B)[151]. In the first step, a universal sequencing primer complementary to the adapter sequence is hybridized to the template fragment. Each cycle of sequencing involves the ligation of one out of four different fluorescently labeled octamers to the sequencing primer. These oligonucleotide octamers are structured such that the identity of the first and second di-base is correlated with the identity of the fluorescent label attached at the end of the octamer. After ligation, non-ligated octamers are washed away and a fluorescence imaging system detects the identity of the ligated octamer and thus determines base 1 and 2 of the template sequence[152]. Next, the fluorescent label is cleaved from the octamer after the fifth base and the cycle of octamer ligation is repeated. Progressive rounds of octamer ligation enable the sequencing of every five bases, i.e. base 6 and 7, followed by 11 and 12, and so on. After completing a series of ligation cycles, the extended primer is denatured and the system is reset with another primer complementary to the $n-1$ position (one base shift) for the second round of ligation (i.e., 4 and 5, 9 and 10, 14 and 15, and so on). After three more rounds of primer reset the entire sequence of the template fragment is determined in color space[134].

*http://www.appliedbiosystems.com; accessed April 15, 2014

A unique and powerful characteristic of SOLiD is the use of two-base encoding, i.e. each base of the template is interrogated twice (two independent ligation reactions by two different primers) (Figure 1.6C). Two-base encoding enables an improved accuracy in miscall detection and thus the discrimination of measurement errors from SNPs [134,138]. However, it has been speculated that AT-rich and GC-rich regions are underrepresented in SOLiD data due to amplification biases during library preparation [153].

Like Illumina machines, SOLiD platforms are distributed with software that produces readable data in color space and assigns quality values to each read. The fact that SOLiD produces sequencing reads in color space rather than letter space (Illumina/Solexa and Roche/454) has some disadvantages: (i) direct translation of color space into letter space is not advisable due to the nature of the di-base dependent sequencing strategy (e.g. assuming that a single color was miscalled, all consecutive colors would be translated to incorrect nucleotide identities due to a 'frame shifted' translation), (ii) a reference template (e.g. genome or transcriptome) has to be available in order to determine sequencing reads in letter space, and (iii) limited availability of software solutions for downstream analysis (e.g. adapter removal or read mapping), since not all solutions support sequencing reads in color space. Nevertheless, the SOLiD system is a powerful technology that produces a large amount of sequencing data with high accuracy (Table 1.1). SOLiD has been successfully applied in various genomic applications including microRNA studies [109,146,154–157].

1.4.3 Small RNA Sequencing

All NGS technologies explained above need DNA as input material. Therefore, in order to sequence the transcriptome of organisms using NGS technologies, template RNA molecules need to be first reverse transcribed into cDNA. cDNA can then be used as input material for high-throughput RNA sequencing (RNA-seq), enabling researchers to discover, profile, and quantify RNA transcripts in a high-throughput manner [158]. Moreover, the high importance of miRNAs as gene regulators for a wide range of functionalities in a cell, lead to the development of small RNA high-throughput sequencing (small RNA-seq), a special protocol for deep sequencing of small non-coding RNAs [159]. Because different classes of small non-coding RNAs exist (e.g. miRNAs, piRNAs, or siRNAs), specific small RNA-seq protocols have to be applied.

In miRNA sequencing, input RNA is isolated and sized fractionated using a gel. Only gel bands containing short RNA molecules (18-30 nt) are used as input for library

construction. By doing this, longer RNA molecules like pre-miRNAs, tRNAs, rRNAs, etc. are discarded and mature miRNAs are filtered in the gel bands because of their short length (∼22 nt). To distinguish miRNAs from other small non-coding RNAs (e.g. piRNAs or siRNAs), most frequently used protocols require the small RNA molecules to have a 5' monophosphate and a 3' hydroxyl group, which are characteristic termini for Dicer products. To identify small RNAs with other terminal structures, such as piRNAs or siRNAs, alternative methods must be applied [58]. Following adapter ligation, library products, because of their short nature, do not require fragmentation compared to genome or mRNA sequencing and can be sequenced in a single read by any NGS technology. Note that mature miRNAs are shorter than the sequenced read due to a minimum read length of 36 nt (Table 1.1; older SOLiD platforms produced reads with 35 nt). Hence, traces of adapter sequence variable in length may be contained within the 3' end of a sequenced read and need to removed computationally in downstream bioinformatics analysis (see Materials & Methods 2.3.1.3 for details).

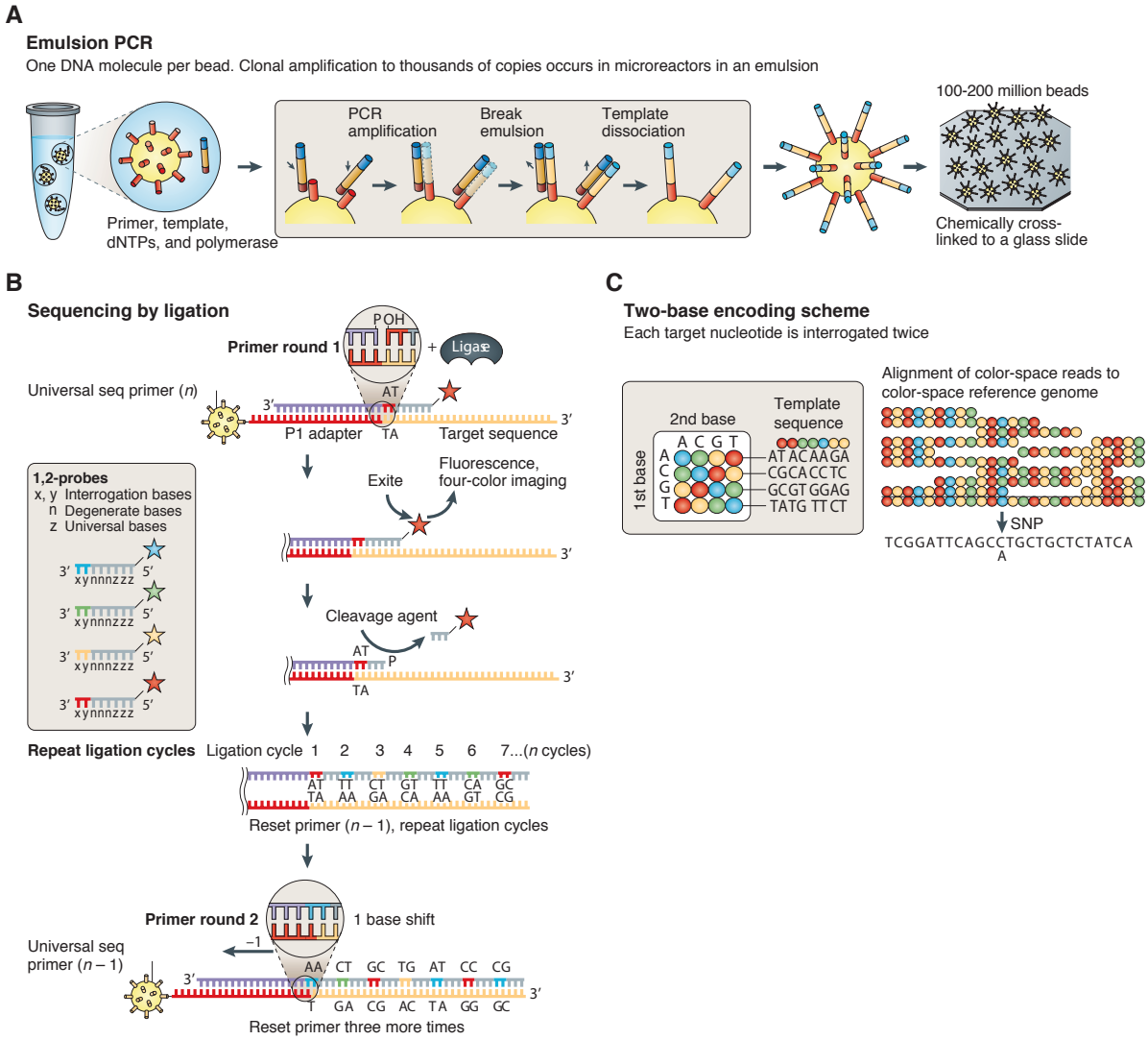

Figure 1.6: ABI SOLiD sequencing system
(A) DNA template fragments are amplified by emulsion PCR. In this process, a reaction mixture consisting of an oil-aqueous emulsion containing primer, dNTPs, and polymerase is created to encapsulate DNA-bead complexes in a single droplet. Within these droplets, PCR amplification is performed to clonally amplify all template fragments in parallel to thousands of copies each. Following amplification, beads are chemically cross-linked to a glass slide. **(B)** The four-color sequencing by ligation method is used as sequencing strategy. First, a universal sequencing primer is hybridized to the template fragment and a library of 1,2-probes (fluorescently labeled octamers) are added, which selectively hybridize and ligate to complementary positions. Following imaging, ligated probes are chemically cleaved after the fifth base and the cycle of probe ligation is repeated. After completing 10 ligation cycles, the extended primer is stripped and the system is reset for a second round of ligation with another primer complementary to the $n-1$ position (one base shift). To determine the entire sequence of the template fragment in a string of colors, three more ligation rounds are performed. **(C)** In the two-base encoding scheme, four di-bases are associated with one color and each base of the template is interrogated twice. Figure adapted and modified from Metzker (2010).

Chapter 2

Materials and Methods

The bioinformatics side has become the 'bottleneck' of all high-throughput based biological studies. A major problem is the handling and analysis of large-scale data sets produced by these experiments. In the following, I will first describe the experimental techniques and strategies applied to address the question whether miRNA genes impact developmental arrest and longterm survival in dauer and dauer-like stages, such as the infective stage of parasites (Section 2.1). I will then introduce the experimental data sets profiled (Section 2.2) and finally present the computational workflow and bioinformatics approaches developed and applied to analyze these data sets (Section 2.3).

2.1 Small RNA Sequencing

To address the role of miRNAs in the dauer/infective larvae fate, known and novel miRNA genes in *C. elegans*, *P. pacificus*, and *S. ratti* were profiled using a multiplatform sequencing approach (Figure 2.1). These experimental data sets were generated in the group of Dr. Christoph Dieterich (*Bioinformatics in Quantitative Biology*) at BIMSB which is part of the MDC. The total RNA of the parasite samples (*S. ratti*) were kindly provided by our collaborator PD. Norbert W. Brattig from the Bernhard Nocht Institute for Tropical Medicine in Hamburg. All next-generation sequencing was performed in the group of Dr. Wei Chen (*Scientific Genomics Platform*) at BIMSB.

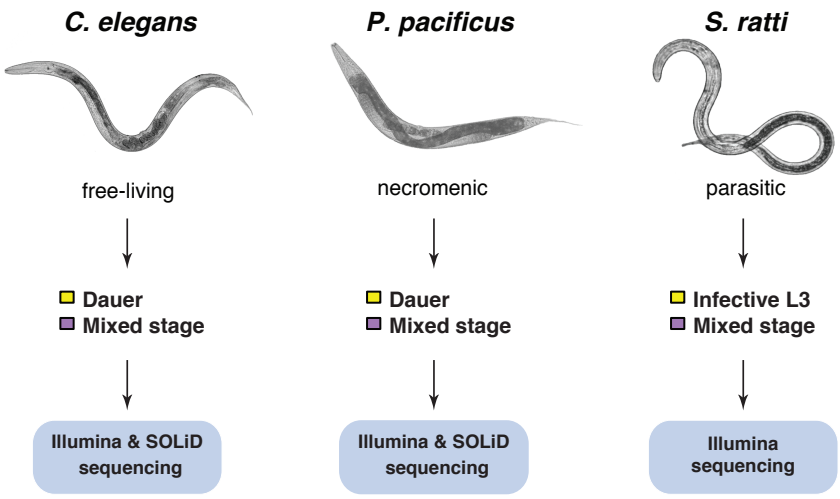

Figure 2.1: Experimental setup of microRNA gene profiling
Multiplatform small RNA deep sequencing was performed on mixed and dauer stage samples of *C. elegans* and *P. pacificus*. Illumina small RNA profiling was carried out on mixed and infective stages of *S. ratti*.

2.1.1 Nematode Strains and Culture

We used wild-type strains of three distinct species in all of our experiments. *Caenorhabditis elegans* and *Pristionchus pacificus* were grown on NGM plates with a lawn of *E. coli* strain OP50[160]. The *S. ratti* animals were maintained using Wistar rats by serial passage as previously described[161,162]. Approval was obtained from the Animal Protection Board of the City of Hamburg. The strains used in this study are as follows: N2, RS2333 (formerly known as PS312), and *S. ratti* Basel (Swiss Tropical Institute; provided by Dr. G. Pluschke).

For mixed-stage cultures of *C. elegans* and *P. pacificus* (Table 2.1; Data sets 1, 2, 5, and 6) 10 to 15 early adults were spotted on NGM plates, allowed to grow at 22°C for 5 days, and washed off with M9 for RNA extractions. Non-dauer (mixed-stage) and dauer samples for both species (Table 2.1; Data sets 3, 4, 7, and 8) were obtained from liquid cultures grown at 22°C starting with synchronized L1 larvae. Synchronized L1 larvae were sampled as follows: Gravid adult worms were treated with bleach to collect embryos[163]. Embryos were incubated in M9 buffer overnight at 22°C to hatch without food, causing the larvae to arrest at the L1 stage. To obtain non-dauer stages in liquid culture, we suspended 100 synchronized L1 larvae in 500 ml S-medium and

added 26nM (25S) Δ7-dafachronic acid on day four to prevent dauer formation. We added 0.5g OP50 per 100 ml worm culture on day 1, 4, and 7 as food source. Non-dauer mixed-stages were then purified on day 8 or 9 and used for RNA extraction. Dauer liquid culture was obtained from 10000 synchronized L1 larvae suspended in 500 ml S-medium. On day 1, 5, and 8 of culture, 0.5g OP50 per 100 ml worm culture was added. Dauer larvae were purified from liquid cultures on day 11 and 12. The culture was centrifuged to obtain a worm pellet. The dauer larvae were then collected and used for RNA extraction.

To isolate iL3s of *S. ratti*, we collected fecal pellets on days 6-16 after subcutaneous infection of male Wistar rats with 1800-2500 iL3s. Charcoal coprocultures were established, incubated at 26°C and assessed for vitality and sterility before further processing[164,165]. After 5-7 days incubation time, the Baermann method was used for the recovery of iL3s. Free-living stages were prepared in a similar way by reducing the incubation time to 24hrs.

2.1.2 Total RNA Isolation and Small RNA Library Generation

Total RNA was obtained from worm pellets using standard Trizol protocols. Quality and quantity of extracted product was assessed using Nanodrop and Bioanalyzer according to the manufactures protocol.

Libraries for deep sequencing were prepared from total RNA according to the manufacturer's protocol (SREK [small RNA Expression Kit]), Applied Biosystems, Forster City, CA, USA (sequencing), Illumina v1.5 protocol for small RNA sequencing, and the NEXTflex small RNA sequencing kit (Bioo Scientific; multiplexed libraries)].

2.2 Data Sets

2.2.1 Small RNA Sequencing Data

In total, we sequenced 10 small RNA samples using a multiplatform sequencing approach (Illumina GA II / HiSeq and ABI SOLiD). Detailed information on the experimental setup and the number of sequencing reads obtained is tabulated in Table 2.1. These data sets have been deposited in NCBI's Gene Expression Omnibus[166] and are

accessible through GEO Series accession number GSE41402*.

Table 2.1: Deep sequencing small RNA data sets profiled in nematodes
Ten deep sequencing data sets derived from *C. elegans*, *P. pacificus*, and *S. ratti.*

Data set	Species	Sample type	Platform	Read length[1]	#Raw reads
1	*C. elegans*	mixed-stage	Illumina GA II	36	20,557,719
2	*C. elegans*	mixed-stage	SOLiD	35	21,307,436
3	*C. elegans*	mixed-stage[2]	Illumina HiSeq	51	10,290,812
4	*C. elegans*	dauer[2]	Illumina HiSeq	51	10,349,552
5	*P. pacificus*	mixed-stage	Illumina GA II	36	27,208,332
6	*P. pacificus*	mixed-stage	SOLiD	35	25,717,306
7	*P. pacificus*	mixed-stage[2]	Illumina HiSeq	51	11,347,692
8	*P. pacificus*	dauer[2]	Illumina HiSeq	51	10,382,820
9	*S. ratti*	mixed-stage	Illumina GA II	36	27,540,069
10	*S. ratti*	infective L3	Illumina GA II	36	32,021,930

[1] measured in nucleotides
[2] libraries were multiplexed and sequenced on the same lane

2.2.2 Publicly Available Data

The genome of *C. elegans* and the 3' UTR, 5' UTR, exon, and intron coordinates were retrieved from WormBase† release WS204[18]. The *S. ratti* genome v1 was downloaded from the Wellcome Trust Sanger Institute‡. For *P. pacificus*, I used our in-house genome. The 3' UTR, 5' UTR, exon, and intron coordinates for *P. pacificus* were inferred from in-house deep sequencing mRNA transcriptome data.

Known miRNA genes from *C. elegans* (223 sequences) and *P. pacificus* (124 sequences) were downloaded from the miRBase[75] database§ (v18). 21U-RNA sequence annotations for *C. elegans* were retrieved from previous studies[167,168]. The 21U-RNA sequence information for *P. pacificus* was obtained from de Wit *et al.* (2009). I inferred the 21U-RNA coordinates by mapping the sequences to the respective genomes. Other non-coding RNA annotations, including rRNAs, tRNAs, snoRNAs, snRNAs, sbRNAs, as well as repetitive sequence and splice leader information, were obtained from an

*http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE41402
†http://www.wormbase.org
‡http://www.sanger.ac.uk
§http://www.mirbase.org

unpublished study (Peter F. Stadler - pers. comm.). *Strongyloides ratti* rRNA sequences were downloaded from the National Center for Biotechnology Information*. The Rfam[169] database (v11), including the Rfam sequences and covariance models, was downloaded from the Wellcome Trust Sanger Institute†.

Quantitative reverse transcription PCR (qRT-PCR) expression data of 107 miRNA genes from *C. elegans* were obtained from Karp *et al.* (2011). miRNA expression level changes calculated as $-\Delta\Delta C_T$ values measured from dauer versus L2m (late L2 - mid L3 stage) samples were taken from the same publication.

2.3 Bioinformatics Methods

As described in the previous section, my study is based on high-throughput deep sequencing data sets. A major problem arising in next-generation sequencing is the handling and analysis of generated large-scale data. To analyze these data sets, I developed a bioinformatics workflow, which consists of six distinct computational analysis steps (Figure 2.2). This workflow includes the analysis of small RNA deep sequencing data and reports known and novel miRNA genes, performs comparative investigations of miRNAs, and infers miRNA gene homologies. A number of bioinformatics methods and strategies were used and implemented to analyze these data sets and derive hypotheses about potential post-transcriptional regulators (miRNAs) conserved between free-living and parasitic nematodes. In the following, I will describe the different computational steps applied in more detail. If not mentioned otherwise, all computational analyses have been performed using available Bioconductor[170] packages or custom scripts implemented in Perl or R[171].

2.3.1 Preprocessing of Small RNA Sequencing Data

Next-generation sequencers produce millions of short sequences or reads (∼35-50 nt in length for small RNA profiling) in a short amount of time at low costs. The challenge is to analyze these high-throughput short reads in a systematic way. Typically, the bioinformatic problem starts with preprocessing of sequenced reads prior to mapping to a reference. Depending on the experimental design the preprocessing steps may include

*http://www.ncbi.nlm.nih.gov
†ftp://ftp.sanger.ac.uk/pub/databases/Rfam

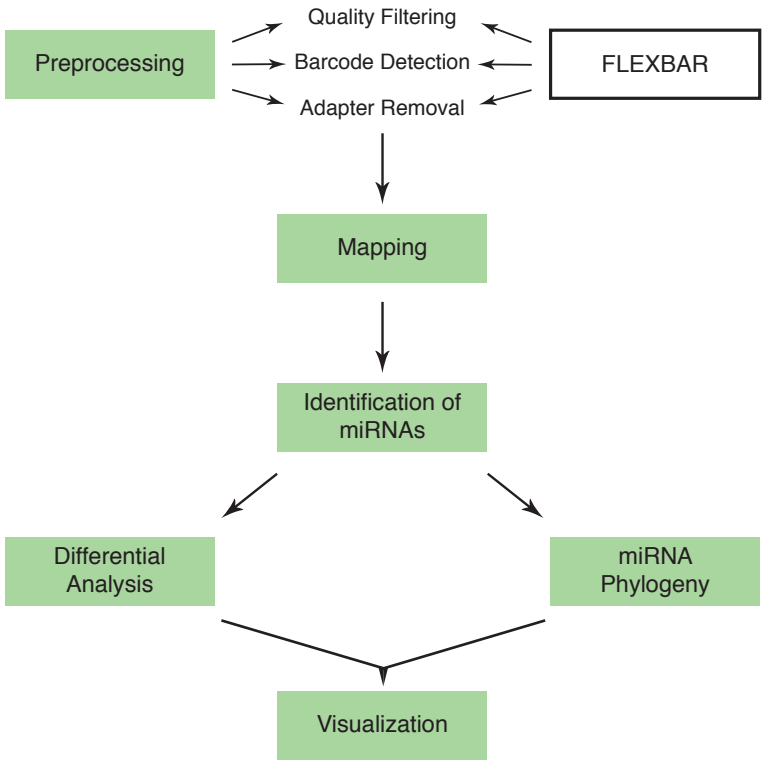

Figure 2.2: Bioinformatics analysis workflow for small RNA-seq data
The analysis workflow for small RNA-seq data developed in this study consists of six distinct computational steps: (i) preprocessing (quality filtering, barcode detection, and adapter removal), (ii) mapping to reference genome, (iii) identification of known and novel miRNA genes, (iv) measuring relative expression changes of miRNAs between samples, (v) inference of miRNA gene families and phylogeny, and (vi) visualization of differentially expressed miRNAs depending on inferred phylogenetic relationships.

quality filtering, barcode detection, and adapter sequence removal. In the following, I will discuss these different steps in detail.

2.3.1.1 Quality Filtering

Each NGS platform is distributed with software that converts image-based signals into nucleotides and assigns quality values to every read. This quality score Q is provided in *phred* format and can be used in three ways: (i) to asses the quality of sequences, (ii) for recognition and removal of low-quality sequences (end clipping), and (iii) determination of accurate consensus sequences. The phred format was developed during the Human

Genome Project[172,173] and is given by

$$Q = -10 \cdot log_{10}\ P$$

where P is the probability that a given nucleotide was called incorrectly. In Sanger sequencing, phred scores range from 0 to 93 using ASCII characters 33−126 in fastq files (Phred+33). Illumina uses its own Illumina-specific Phred+64 offset and depending on the platform employed, phred scores range from -5 to 40 using ASCII characters 59−104 or from 0 to 40 using ASCII 64−104. However, since the release of CASAVA v1.8 (software package distributed by Illumina), they moved away from their Illumina-specific offset and adopted the Sanger transformation, i.e. Phred+33. SOLiD quality values, which are assigned to each color, range from 0 to 45 using ASCII 64−109. It has to be mentioned that the exact relationship between phred scores and SOLiD quality values is unclear[174].

Several error sources can appear during next-generation sequencing and subsequent downstream analyses: (i) library preparation, (ii) sequencing process (insertions, deletions or mismatches), and (iii) bioinformatic processing (base calling and mapping to a reference). Therefore, bioinformatic postfiltering of low-quality reads (and if necessary read error correction) is an important step to reduce possible read based alignment errors in the mapping process.

2.3.1.2 Barcode Detection

With the increasing throughput of NGS machines, especially for organisms with a small genome size such as yeasts[175], worms[17] and flies[176], the number of sequenced and mapped reads is often higher than required for the experiment. Since the beginning of Sanger sequencing, multiplexed DNA sequencing (i.e. a method to analyze multiple biological samples at the same time) has been in use to reduce the sequencing costs per sample[177]. This strategy has been successfully adopted to Roche's 454 platforms, Illumina GA II or HiSeq, and ABI's SOLiD for different applications[178–186]. With the usage of barcodes, it is possible to sequence multiple samples in a single lane. In this context, a unique barcode identifier, typically 6 nt in length, is assigned to each sample. Barcodes can be introduced during PCR amplification of libraries, which are then sequenced separately, or by ligation of adapter sequences, which include barcodes. Studies comparing miRNA expression profiles obtained by these two methods show that

the quality of PCR-based barcoding is reproducible, whereas ligation-based barcoding introduces biases[187,188]. However, in both cases of barcode introduction before subsequent downstream analyses, a bioinformatic processing step is required that reliably detects barcode sequences and assigns reads to their corresponding biological sample.

Since barcode sequences are subject to false base calls resulting in mismatches or indels (depending on the technologies error model), it is crucial for a barcode detection tool to allow for errors in order to determine the actual barcode sequence for a correct assignment of reads. Thus, it is important that barcode detection tools are capable of dealing with vast amount of data in reasonable time and are flexible in their detection mode, i.e. PCR- or adapter-based barcoding, while being able to allow for false base calls. Moreover, when distinct sequencing platforms are in use that produce reads in letter and color space, it is preferable that tools cope with reads obtained by all platforms.

A number of programs are available that detect barcode sequences, e.g. Illumina's CASAVA software package (v1.8.2) includes a demultiplexing functionality of dual-indexed libraries (Illumina's PCR-based barcoding)* or Novobarcode which detects barcodes in Illumina indexed reads or within the 5' or 3' end of reads in letter space†. However, none of these programs includes all of the functionalities that were required for my tasks. Since we already implemented the Flexible Adapter Remover (FAR), a tool that detects strings in sequences by computing overlap alignments based on the Needleman-Wunsch algorithm[189] in letter and color space for the purpose of sequence adapter removal, we extended FAR's functionality and added a barcode detection feature. This tool is now known as the Flexible Barcode and Adapter Remover (FLEXBAR)[1]. I will explain FLEXBAR's barcode functionality and the underlying algorithm in Chapter 3.

2.3.1.3 Adapter Removal

Large amount of short sequence reads are produced in a massive parallelized strategy. All of the currently employed NGS technologies introduce sequence tags necessary for sample library preparation, which are ligated to the target sequence. Depending on the sequencing application and experimental setup, these artificially introduced sequence tags may overlap with the sequenced region and should be removed prior to subsequent

*http://support.illumina.com/sequencing/sequencing_software/casava.ilmn
†http://www.novocraft.com

downstream analyses. Importantly, in case of miRNA profiling, in which the biological signal is ∼22 nt in length, and the minimum read length of NGS platforms is 35-50 nt, all short reads contain adapter sequences or tags (partial sequence or entire). Hence, for optimal results in subsequent mapping procedures, these adapter sequences need to be detected and removed from short reads.

Several software solutions are available that detect and remove adapter sequences mostly in letter space. However, due to inherent error profiles of different sequencing technologies (sequencing errors generally increase towards the 3' end), perfect matching approaches between short reads and target sequences are insufficient. Thus, our goal was to implement a fast and flexible tool that detects and removes adapter sequences in reads from different next-generation sequencing technologies, i.e. in letter and color space. To accomplish this task, we originally implemented the Flexible Adapter Remover. FAR has now been extended and integrated into the Flexible Barcode and Adapter Remover. A detailed description of FLEXBAR's program features including adapter detection and removal functionality, and its evaluation compared to other sophisticated NGS data preprocessing tools is given in Chapter 3.

FLEXBAR detects and removes adapter sequences within reads by overlap sequence alignment based on the Needleman-Wunsch algorithm [189]. An overlap (or semi-global) alignment uses the same recurrence relations as a global alignment but does not penalize gaps at the end of the alignment. Adapter sequences can either be detected anywhere within a given read or specifically at the left- or right-end of a read (for details of different trim-end modes see Chapter 3). To detect these sequence tags in color space reads, the adapter sequence (in letter space) is first internally translated into color space (Figure 2.3A) and then aligned to each read (Figure 2.3B; RIGHT trim-end mode example). Due to the two-base encoding strategy of SOLiD, an additional character will be removed from the trimmed read.

FLEXBAR is applicable for a wide range of biological applications, such as detection of adapter sequences in small RNA-seq data or detection of splice-leaders in operons [1]. In general, it is possible to detect any kind of sequence anywhere in a given read sequence. An example of FLEXBAR's adapter removal application will be given in Chapter 3. With FLEXBAR at hand, successful detection of barcodes and/or removal of adapters in sequencing reads produce less mapping artefacts.

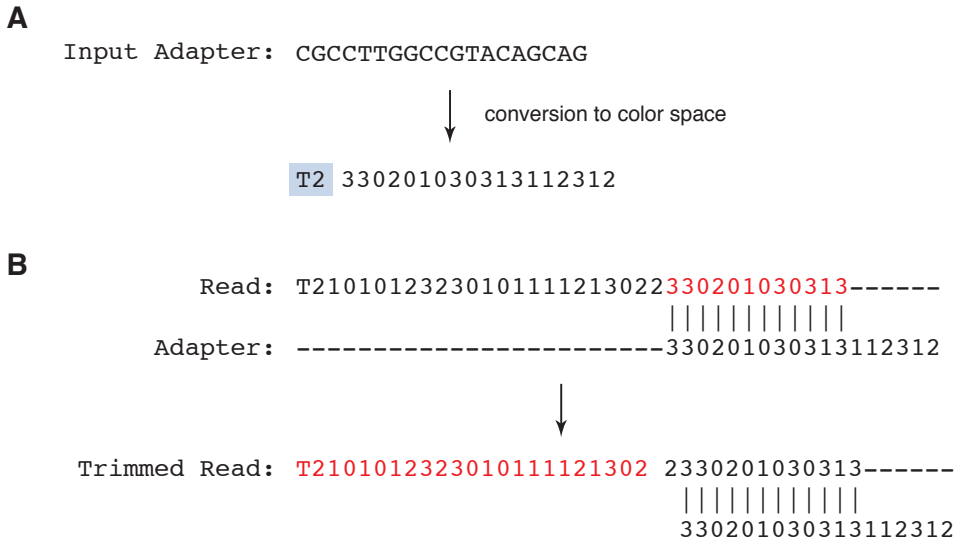

Figure 2.3: Processing strategy of adapter recognition and removal in color space **(A)** In general, adapter sequences are processed in letter space. If adapters have to be detected and removed in color space, FLEXBAR converts the letter space adapter automatically in color space prior to the subsequent alignment process. Moreover, the first two characters of the transformed color space adapter are removed before aligning due to the double encoding strategy of color space reads. **(B)** Overlap alignments are calculated using the Needleman-Wunsch algorithm[189] (here RIGHT trim-end mode). After adapter detection and removal in color space, an additional character - the right most character of the trimmed read - will be removed (RIGHT trim-end mode). Note: The trimmed read aligns perfectly with the reverse complement of cel-miR-34-5p in color space.

2.3.2 Mapping of Short Sequencing Reads to a Reference

After the initial preprocessing of short reads, most NGS analysis protocols continue with the crucial step of mapping reads the original sequences, i.e. the reference[174]. In most settings, the genome sequence of the species the reads have been generated from (if available), serves as the reference. However, any given sequence database can serve as a reference, e.g. transcriptome annotations can be used as a reference when mapping mRNA-seq for RNA splice identification[190].

The process of finding the most credible source within a reference for the observed sequencing read, given the knowledge of which species the sequence has come from, is the classical sequence alignment problem in bioinformatics. Determining the optimal alignment (or multiple) in the mapping step is complicated by various factors, including sequencing errors, short read length, genetic variation in the population, non-uniform

confidence in base calling, low reading accuracy in homopolymer stretches of identical nucleotides, and the huge amount of reads to be mapped, which requires for the algorithm to be optimized in speed and memory usage[133,174,191]. Traditional methods such as pure Smith-Waterman dynamic programming[192], BLAST[193] or BLAT[194] are not designed to map the huge amount of data as produced by NGS. Depending on the availability of computation power and random-access memory (RAM), these algorithms may map the reads in a few days[190]. However, large and expensive computer grid engines are not accessible to everyone and more efficient algorithms are needed. To date, a large ever growing number of programs have been implemented to overcome these challenges. Table 2.2 gives an overview of selected software solutions that are available to solve the problem of short-read alignment.

Table 2.2: Overview of selected short read mapping tools
This table was adapted and modified from Bao *et al.* (2011).

Program	Website	Open source	Quality score involved	Mapping strategy
Eland[195]	None	No	Yes	Hash the reads
Maq[196]	http://maq.sourceforge.net	Yes	Yes	Hash the reads
SHRiMP2[197]	http://compbio.cs.toronto.edu/shrimp/	Yes	Yes	Hash the reference
NovoAlign	http://www.novocraft.com	No	Yes	Hash the reference
Bowtie[198]	http://bowtie.cbcb.umd.edu	Yes	Yes	BWT-based, index the reference
Bowtie2[199]	http://bowtie-bio.sourceforge.net/bowtie2	Yes	Yes	BWT-based, index the reference
BWA[200]	http://bio-bwa.sourceforge.net/bwa.shtml	Yes	Yes	BWT-based, index the reference
SOAP2[201]	http://soap.genomics.org.cn/	Yes	Yes	BWT-based, index the reference

Most short-read alignment programs combine a two step procedure, i.e. a heuristic filtration technique followed by a verification step. In the filtration step a small set of most likely candidate regions that contain possible mapping locations are identified. In the verification step, once the smaller subset of most likely candidate regions has been determined, more accurate but slower alignment algorithms (e.g. Smith-Waterman) are applied on the limited subset. Although a large number of programs for short-read alignment exist, only a few fundamental concepts are implemented in the filtration step. These methods cover (i) hash table-based implementations, in which the hash

may be created using either the set of reads or the reference, and (ii) Burrows-Wheeler transformation[202] (BWT)-based methods, which facilitate rapid searching with low-memory usage by creating an efficient index of the reference (column 'Mapping strategy' in Table 2.2). BWT-based implementations typically use the full-text minute-space (FM) index data structure, which has been referred to as a compressed suffix array[203]. If designed into the alignment program, both of the above concepts can be applied to letter (Illumina, 454) and color space (SOLiD) reads. Regardless of the implemented approach in each program, there is a general tradeoff between sensitivity and speed[204].

In this study, the Bowtie (v0.12.5) and SHRiMP (v2.1.1) aligners were utilized. Bowtie, which indexes the reference using BWT, was applied for the identification of novel miRNA genes as part of the miRDeep2 pipeline (Section 2.3.3.2). For the quantification of miRNA expression levels and subsequent differential expression analysis, mapping results produced by the SHRiMP2 aligner were used (local alignment mode with parameters '−h 80% −−strata −o 20 −−max-alignments 20' and default otherwise). SHRiMP2, which uses a hash table-based strategy that indexes the reads, was specifically designed to map reads in color space, whereas Bowtie was originally implemented for letter space reads. Both short-read aligner provide a high accuracy with more then 79% correctly identified genuine matches, when evaluated on simulated Illumina single end data (15 million reads with 76bp length). While Bowtie utilized the least amount of RAM, SHRiMP2 performed best in terms of correctly mappable reads (around 96%), but at the expense of time and memory. The evaluation analysis using real data was generally consistent with the results from simulated data[133].

2.3.3 Identification of microRNA Genes from Small RNA-Seq Data

Next-Generation Sequencing has revolutionized diverse genomics applications including miRNA analysis. Advantages of NGS platforms compared to Sanger-based cloning methods or microarrays are a quantitative readout, which allows for digital gene expression (DGE) profiling, and at the same time the possibility to detect previously unknown miRNA genes or other genomic features at high speed and sensitivity and reduced costs[130]. Because of that, small RNA-seq (or microRNA-seq) has become the standard method for the analysis and discovery of miRNA genes.

2.3.3.1 Quantification of microRNA Expression Levels

After preprocessing and subsequent mapping of small RNA reads to the reference genome, the genomic mapping information of each read is used for the annotation of known and novel miRNAs and other small ncRNAs. miRNA and small ncRNA annotations can be downloaded from the miRBase*, Rfam† or UCSC‡ databases.

An important summary statistic for the quantification of miRNAs is the read count, i.e. the number of reads (or tags) assigned to a specific miRNA. Thus, if the genomic mapping coordinates of a read overlap with known genomic annotations from miRNAs (mature or precursor sequence) in the correct orientation (e.g. at least 80% of the read), the read is assumed to be a sequencing product of this particular miRNA; the corresponding read count is increased by one. Reads that mapped multiple times to the reference genome are assigned to each loci of a miRNA.

Read counts have been found to be a good linear approximation for the abundance of target transcripts[205]. However, due to sequence-specific biases arising from small RNA library preparation and sequencing technology, only relative quantification studies of miRNA expression levels are reliable[188,206,207]. Absolute quantification of miRNA expression levels is not possible[208].

2.3.3.2 Identification of Novel microRNA Genes

Originally, the experimental detection of novel miRNA genes involved the process of cloning followed by Sanger sequencing, which is an expensive and time-consuming procedure[209]. Additionally, miRNAs that are expressed in low copy numbers are difficult to detect. Small RNA-seq is a promising method for the detection of unknown miRNA genes at high speed and reduced costs. Small RNA-seq is particularly suited for the detection of low-abundance miRNAs, due to the high sequencing throughput.

To identify novel miRNA genes computationally, two types of miRNA prediction tools exist. One class of tools do not need experimental data. These tools rely on conservation information and specific miRNA characteristics. Since, conservation based tools depend on homology, only orthologous or paralogous miRNAs can be detected. Species-specific or previously not detected miRNA genes families will be missed. The other class of

*http://www.mirbase.org
†ftp://ftp.sanger.ac.uk/pub/databases/Rfam
‡http://genome.ucsc.edu

tools are based on small RNA-seq data. These tools use the small RNA-seq read pattern information for the prediction of novel miRNAs. The advantage of RNA-seq based tools is that they rely on experimental data and facilitate the identification of species-specific or unknown miRNA gene families. With the rapid development and improvement of NGS methods, computational strategies that employ small RNA-seq read patterns appear to be the most promising tools. In the following, I will focus on small RNA-seq based methods and their strategies for novel miRNA identification.

The first tools that were developed for the detection of miRNAs in small RNA-seq data, i.e. miRDeep[210] and mireap*, were published in 2008. Since then, many software tools have been developed, due to the successful application of small RNA-seq for miRNA identification even at low abundance in many areas. Recently, two studies were published that compared and evaluated software tools for miRNA identification in NGS data[211,212]. Eight software tools (miRDeep[210], miRanalyzer[213], miRExpress[214], miRTRAP[215], DSAP[216], mirTools[217], MIReNA[218], miRNAkey[219], and mireap*) were evaluated by Li *et al.* (2012) and three (miRDeep v1[210] and v2[220], miRanalyzer[213], and DSAP[216]) by Williamson *et al.* (2013). Both studies conclude that in general the best suitable software tool should be selected based on their specific input and output requirements, e.g. investigated species, availability and accuracy of reference genomes, available computational resources and time, or user friendliness. Nevertheless, Williamson and colleagues believe that miRDeep is the best solution for novel miRNA candidates[212] and Li and colleagues specifically recommend MiReNA, mireap, miRDeep, and miRanalyzer for novel miRNA prediction in *C. elegans*[211].

In this study, I applied miRDeep2 for the prediction of novel miRNA candidates, because it is easy to use and especially applicable for nematode species. In the following, I will present the general idea of miRDeep2.

The miRDeep2 program is a stand-alone application that predicts miRNAs from NGS data employing the characteristics of Dicer processing using Bayesian probabilities. The pipeline starts by mapping the reads to a reference genome. Then, consecutive reads that occur in close proximity are clustered and the region of potential precursor sequences is extended by 90 nt. Following the characteristics of Dicer processing, the algorithm assumes if a read originated from a miRNA, then it must either be a portion of the 5' arm, 3' arm, or loop sequence (Figure 2.4). Moreover, miRDeep2 assumes that the mature sequence (miRNA arm loaded into RISC) is more abundant in a cell

*http://sourceforge.net/projects/mireap; accessed November 5, 2013

than the other arm (in miRDeep2 termed as star sequence) or loop sequence and therefore is most abundant in the NGS data file. A probabilistic score for each potential miRNA is calculated based on the relative positions of reads within a predicted precursor and their frequencies (mature to star ratio), the thermodynamic stability of the secondary structure, the evidence of a 2-nt 3' overhang, and the conservation of the assumed miRNA in related species (if available). miRDeep2 incorporates Bowtie [198] for the mapping process, RNAfold [221] for the secondary structure prediction, and Randfold [222] for the estimation of the thermodynamic stability of the secondary structure. As input, miRDeep2 accommodates data produced by the Illumina and Roche 454 machine. miRDeep2 does not support the processing of color space data generated by the SOLiD machine.

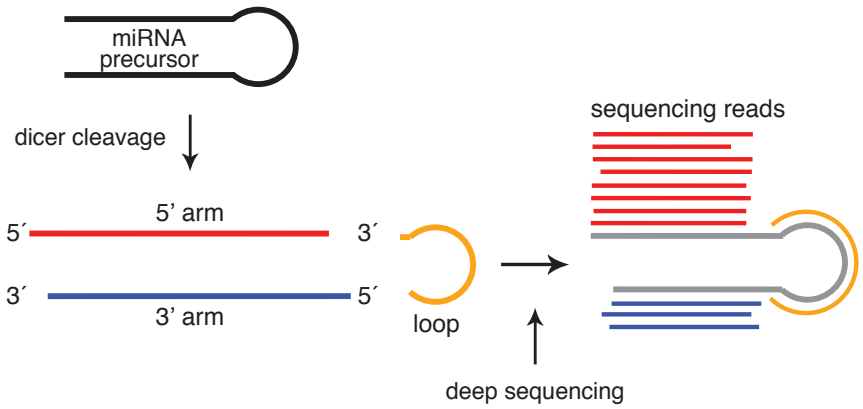

Figure 2.4: miRDeep2 prediction strategy
Following the characteristics of Dicer processing, each RNA product of a cleaved pre-miRNA (5' arm, 3' arm, or loop sequence) has a certain probability of being sequenced. Based on this assumption, sequenced reads from miRNAs will map to corresponding predicted precursor structures according to the three characteristic Dicer products. Since the functional miRNA (arm loaded into RISC) is assumed to be more abundant in a cell, it is sequenced more frequently than the other arm or loop sequence. Hence, statistics of read positions and read frequencies within a precursor signature are highly characteristic for miRNAs and are scored by miRDeep2. Figure adapted and modified from Friedländer *et al.* (2008).

In this study, a multiplatform NGS (Illumina and SOLiD) approach was applied to detect miRNA genes in different nematode species. Novel miRNA genes from Illumina data were identified using the miRDeep2 pipeline with default parameters and Illumina reads that matched with at least 18 nt to the genome were retained. In order to predict novel miRNA genes from color space data (SOLiD), I modified the miRDeep2 pipeline

and utilized the color space functionality of the Bowtie[198] read mapper (v0.12.5; option −C). SOLiD reads with at least 16 nt color space matches were retained. Due to the SOLiD color space two-base encoding strategy (for details see Introduction 1.4.2), the first and last nucleotide of mapped color space reads can not be correctly converted into letter space. Therefore, after the mapping process I added the first and last nucleotide to the letter space converted SOLiD read based on the reference genome sequence. Subsequently, read mappings from Illumina and SOLiD data were processed by the miRDeep2 core algorithm. To exclude false positive candidate miRNA loci, the initial list of all candidates (miRDeep2 score with signal-to-noise ratio ≥ 10 (except data set 6: signal-to-noise ratio = 9.4) were filtered against a database of other small non-coding RNAs, including rRNAs, tRNAs, snoRNAs, snRNAs, 21U-RNAs, sbRNAs, as well as repeats. Due to the absence of small ncRNA annotations in *S. ratti*, the initial miRNA candidate list was filtered against ribosomal RNAs and compared to the Rfam database[223] (v11). The remaining set of candidate miRNAs were manually inspected and curated to yield the final set of novel miRNA loci.

2.3.4 Differential Expression Analysis

Beyond the discovery of new species of miRNAs, a common goal in NGS analysis particularly in digital gene expression studies (DGE) is to quantitatively compare expression profiles between different biological samples[224,225]. In order to do so, miRNA count data need to be normalized prior to the identification of differentially expressed genes through statistical testing.

2.3.4.1 Normalizing microRNA Sequencing Data

Although NGS is the method of choice to profile miRNA genes these days, it is still error prone and systematic variations and biases are introduced during the experimental process. Sources of biases in microRNA-seq could be introduced by (i) the quality of the RNA sample, (ii) degenerated RNA and/or contamination with ribosomal RNA during the library preparation, (iii) sequence-specific ligation of adapter or barcode sequences to RNA, (iv) sequence-specific variations in enzyme efficiency, (v) reverse transcription, and (vi) PCR amplification[226]. Thus, when measuring relative expression changes of miRNAs between experiments, it is critical to consider these systematic variations. Furthermore, different total read counts are generated in different microRNA-seq li-

braries. Hence, it is essential to normalize microRNA-seq libraries before comparing the abundances of miRNAs between them[227].

The goal in normalization is to estimate systematic variations in different microRNA-seq experiments through the distinction between true biological signal and random noise. Oftentimes, simple total read count normalization is applied to remove differences in sequencing depth between libraries. More sophisticated normalization methods are desirable, since there is a great range of sources for systematic variations and biases. Moreover, previous studies have shown that the normalization method, rather than the differential expression (DE) model, largely determines the outcome of DE in RNA-seq studies[227–229]. Therefore, choosing the optimal normalization method is critical for DE. So far, a lot of effort has been invested for the development of normalizing methods for mRNA-seq data sets and one could expect that these normalizing methods could be adapted for microRNA-seq normalization. However, this is questionable because the total number of mRNA transcripts in a sample is magnitudes larger than the total number of miRNA molecules[227]. To answer this question Garmire and Subramaniam (2012) systematically evaluated seven commonly used normalization methods applied to high-throughput data for their applicability to microRNA-seq data, which can be grouped into two categories: (i) the ones that apply linear scaling or (ii) the ones that do not. The linear scaling methods that were investigated are scaling normalization, global normalization, Lowess normalization, and the Trimmed Mean of M-values (TMM), whereas the other category includes quantile normalization, variance stabilization (VSN), and the invariant method (INV)[227]. In oder to estimate systematic variations, each method makes different assumptions about true biological differences and random noise. Comparing these methods on several levels through multiple independent data sets revealed that Lowess and quantile normalization are best suited for the normalization of microRNA-seq data, while TMM, which is commonly applied for mRNA-seq normalization, performed worst[227].

The goal of quantile normalization (QN) is to make read count distributions across miRNA samples equal[230]. The same goal can be achieved with quantile-based scaling as applied by Schulte *et al.* (2010). The authors applied linear transformations using a scaling factor based on quantile-quantile (qq) plots (qq-scale normalization). The advantage is that qq-scale normalization is a linear transformation, whereas QN is nonscaling. Moreover, qq-scale normalization is an intuitive, data driven, and robust approach, since the scaling factor is calculated by the median of absolute differences of corresponding quantile values. An artifact in QN is that a gene whose expression value

is always high but not equal will have a low variance, due to the fact that QN shrinks differences for high values over proportionally. High abundance miRNAs (oftentimes high-count genes, i.e. a few genes whose read counts contribute to a large proportion of the total read count in a sample) are typically observed in miRNA DGE studies[231]. Therefore, applying QN to such data could be problematic and may remove expression variances of high-count miRNAs.

To remove potential biases in miRNA expression across developmentally arrested and non-arrested samples, I normalized raw read counts of each data set using reference-based qq-scale normalization[154]. All data sets were normalized by linear transformations using scaling factors. The scaling factors were computed based on the median of differences of corresponding quantile values of non-arrested (chosen as the reference) and arrested samples. The distribution of count values ≥ 5 in the paired data sets were compared in logarithmic space. In order to avoid problems associated with zero values, a pseudo count of one was added to read counts prior to normalization.

The *MA*-plot, a plot of log-intensity ratios (M-values) versus log-intensity averages (A-values) originally introduced for microarray gene expression data[232], is widely used to illustrate the dependency on intensities in high-throughput data. The *MA*-plot and specifically the median of the M-values gives an idea of how good a normalization procedure worked, i.e. a considerable deviation of the center of the distribution of M-values from zero indicates that additional normalization is needed[227]. Note that drawing definitive conclusions from such a qualitative comparison concerning the performance of the normalization methods is typically not possible. However, exploratory analyses generally help to shed light on the characteristics of the data and the impact of the normalization process on the data distribution.

In my study, the comparison of *MA*-plots from raw data (without normalization; left plot) and normalized data using reference-based qq-scale normalization (right plot) of detectable miRNAs in *P. pacificus* non-arrested vs. arrested samples (data set 7 and 8; Table 2.1) suggested that the normalization procedure worked satisfying (Figure 2.5; raw data with a median of -0.596; normalized data with a median of -0.009). Note that in this study no biological replicates were sequenced.

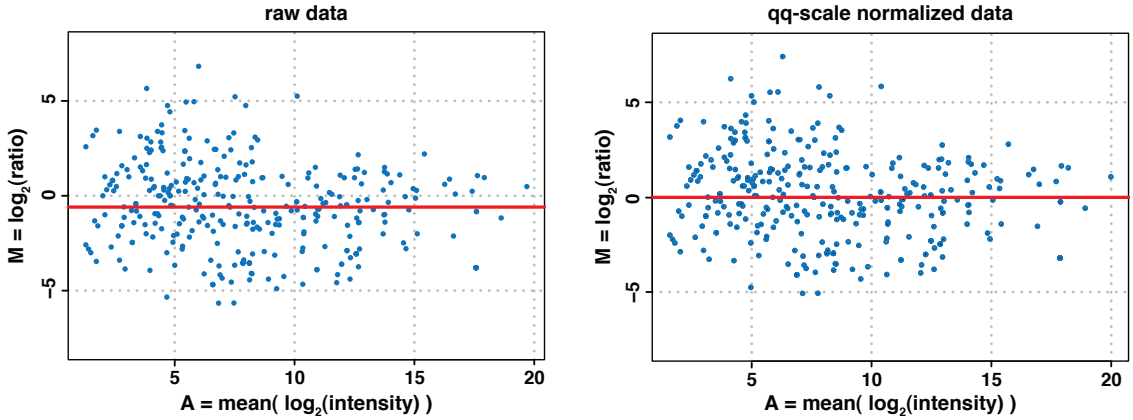

Figure 2.5: *MA*-plots of *P. pacificus* miRNA data before and after normalization
MA-plots show the distribution of miRNAs detectable in *P. pacificus* non-arrested vs. arrested samples in comparison before (raw data) and after normalization using reference-based qq-scale transformation (data set 7 and 8; Table 2.1). The horizontal red lines denote the mean of the M-values, which computed from the raw data deviates considerably from zero (median of -0.596), whereas the median of the normalized data is close to zero (median of -0.009).

2.3.4.2 Defining Differential Expression

After the normalization process of DGE data, the biological question whether a given miRNA gene is differentially expressed can be restated as a statistical problem of hypothesis testing: the simultaneous test for each miRNA whether the observed difference in read counts is significantly greater than what would be expected due to random variation under the null hypothesis of no association between the expression levels and experimental conditions (e.g. developmentally arrested and non-arrested worms)[232]. Hence, in order to determine miRNAs that differ significantly in their expression levels between different experimental conditions, a test statistic for each gene has to be computed. Moreover, the resulting p-value needs to be adjusted for multiple hypothesis testing, since several hundred miRNA genes may be tested simultaneously.

Numerous test statistics have been applied to model the problem of DE detection, e.g. χ^2/Fisher's exact test, binomial test, and Poisson tests[227–229,231,233,234]. In the χ^2/Fisher's exact test, each miRNA is associated with a 2×2 contingency table that include read counts of specific miRNAs from the reference library versus the condition of interest, as well as the sum of the read counts of all other miRNAs expressed. The χ^2 test is applied if all read counts are larger than 5. Otherwise the Fisher's exact test is used. A miRNA is called differentially expressed when the observed read counts are

greater or less than the expected read counts with a false discovery rate (FDR) of less than 5% in order to adjust for multiple hypothesis testing[227]. In the binomial test it is assumed that each miRNA is independently distributed from each other and appears either in the reference or the condition of interest and follows a binomial distribution with an expected probability $p = 0.5$ and $n = n_1 + n_2$, where n is the total number read counts in both libraries, n_1 the total number of read counts in the reference library, and n_2 the total number of read counts in the other library (condition of interest). As in the χ^2/Fisher's exact test, a miRNA is called differentially expressed when the observed read counts are greater or less than the expected read counts with a FDR <0.05. The Poisson test is done in a similar way than the binomial test[227].

In this study, I applied a simple two-sample comparison, since replicates were not profiled. For this, an exact two-sided binomial test was computed for each miRNA using the R function binomTest() from the Bioconductor package edgeR[235]. This test is closely related to Fisher's exact test for 2×2 contingency tables, but with the difference that for each miRNA it conditions on the total number of counts in the library, i.e. all miRNAs expressed in the library. By doing so, the library size variability between experiments is taken into account. The null hypothesis is that the expected read counts of a miRNA are in the same proportions as the library sizes with the probability of the reference library being

$$p_1 = \frac{n_1}{(n_1 + n_2)},$$

where n_1 is the total number of read counts in the reference library and n_2 the total number of read counts in the other library (condition of interest). In this approach the read counts in each library as a proportion of the whole follow a binomial distribution. The final set of differentially expressed miRNAs was defined by two criteria: i) absolute $\log_2$ fold change >1 and ii) exact two-sided binomial test with a p-value cutoff that corresponds to a FDR <0.05.

2.3.4.3 Correction for Multiple Hypothesis Testing

To define differentially expressed genes in high-throughput data as carried out in this study, simultaneous performance of statistical tests for many genes (multiple hypothesis testing) is involved. A correction of the p-value in multiple hypothesis testing is required, because the false positive rate of individual tests accumulates and therefore the chance of falsely calling differential expressed genes increases[236].

Classical methods control the family-wise error rate by adjusting each individual hypothesis significance level to ensure a least overall significance level. The most familiar method for multiple testing correction is the Bonferroni adjustment, which distributes the overall significance threshold α evenly on all performed tests n by requiring an overall significance level of at least α/n. However, this method is extremely stringent and not always appropriate especially when applied to large-scale biological data where the number of simultaneously performed tests can exceed many thousands[237]. In this regard, the control of the FDR, which is the expected proportion of incorrectly identified genes among the list of significant genes, achieves better power. The procedure for controlling the FDR, which was applied in this study, was introduced by Benjamini and Hochberg (1995) for independent p-values.

To ensure that the expected FDR is controlled at a given δ the Benjamini-Hochberg procedure works as follows[238]:

Consider testing m different hypothesis tests $H_1, H_2, ..., H_m$ based on the corresponding p-values $P_1, P_2, ..., P_m$. Let $P_{(1)} \leq P_{(2)} \leq ... \leq P_{(m)}$ be the ordered p-values, and denote by $H_{(i)}$ the null hypothesis corresponding to $P_{(i)}$. Find the largest index $k \in i$ for which

$$P_{(i)} \leq \frac{i}{m}\,\delta.$$

Subsequently, reject all $H_{(i)}$ $i = 1, 2, ..., k$ with p-values less than or equal to $P_{(k)}$.

2.3.5 Inference of microRNA Gene Families and Phylogeny

In this work, I investigated the evolution of miRNAs among free-living and parasitic nematodes. Hence, the inference of miRNA gene families and phylogeny was very important. Gene families of various ncRNA classes defined by means of sequence similarity are stored in the Rfam database[169]. A miRNA registry, which provides a resource for discovered miRNAs, was originally created as part of Rfam. With the increase of data on miRNAs, the miRNA registry was separated from Rfam and renamed to miRBase[75]. The miRBase database provides a grouping of miRNA families of miRBase deposited miRNAs based on conservation across precursors and manual curation using sequence alignments. Therefore, it is not possible to recreate miRBase family classifications from a set of rules (Sam Griffith-Jones - pers. comm.). However, novel miRNA sequences, as discovered in this study, required the regrouping of known families and/or the definition

of novel families. Therefore, I developed a strategy to infer miRNA gene families and phylogeny and grouped all miRNA loci including novel genes discovered in this study into families.

The inference of miRNA phylogeny can be divided into three steps: (i) grouping of miRNA gene families, (ii) computing multiple sequence-structure alignments of these families, and (iii) inference of phylogenetic trees using hierarchical clustering. In the following, I will describe these steps in detail. Moreover, I will provide an evaluation of the performance of this method in Section 4.5.

2.3.5.1 Grouping of microRNA Gene Families

It is common knowledge that miRNA genes usually belong to the same family if they share the same seed sequence. The assumption is that these miRNAs have similar targets and therefore similar functions, yet distinct spatial and temporal expression profiles. However, this definition of miRNA families does not distinguish between true homology and homoplasy. Thus, an in-depth phylogenetic analysis requires a careful distinction between homology and homoplasy. A homoplasy trait is shared by different taxa due to convergence, i.e. genes related by function show sequence similarity due to the same function and not because of common descent. For lack of widely accepted criteria to distinguish homology from homoplasy in miRNAs, I developed a novel strategy that captures sequence identity of miRNAs based on a 6mer miRNA target recognition site combined with supplemental pairing as presented by Bartel (2009).

To this end, I computed all-against-all pairwise global end-gap free alignments of miRNA 5'- and 3'-arms (first nucleotide was removed) using USEARCH[239] (v6.0.307; global alignment mode with parameters '−−allpairs_global −idprefix 6 −−gapopen 2.0I/1.0E −−gapext 1.0I/0.5 −query_cov 0.8 −target_cov 0.8 −−fulldp −−id 0.65'). Alignments were retained if they covered at least 80% of the target and query sequence with a minimal sequence identity of 65%. Sequence identity within an alignment was computed as the number of identical nucleotides divided by the number of columns in the alignment. These values were selected based on well-known miRNA families reported in the literature, e.g. *let-7* family. Additionally, the two largest alignment blocks had to cover 65% of the shorter sequence in each comparison. If that proportion happened to be 40-65%, the alignment was retained if it included the same arm for both, query and target sequence. The 40% threshold was chosen, because Bartel

reported that a 6mer (position 2-7) target recognition site combined with 3' supplementary pairing of at least 3 nt (position 13-15) is usually sufficient for binding[113]. Thus, if two alignment blocks exist that cover at least 9 out of 22 nucleotides (average miRNA length), an estimated 40% (9/22) sequence identity has to be present in the alignment. This value represents the lower bound necessary for target recognition. Because I wanted to avoid mistaking convergence for homology, I used a stringent miRNA family definition and added a second criteria in case sequence identity was comparably low (<65%): no occurrence of arm switching. In other words, assuming the query includes a 5' miRNA arm, the matching arm in the alignment has to originate from a 5' arm as well. Finally, to group miRNAs into families, I searched for connected components of an undirected graph, i.e. every pair of vertices is connected by a path, with nodes representing miRNA arms. miRNA arms were connected by edges if they form a valid alignment [see Figure 4.5 in Section 4.5 as graph example; R package igraph (v0.6.2)[240]].

2.3.5.2 Multiple Sequence-structure Alignments of RNA

Just as a pairwise alignment captures the relationship between two sequences (DNA, RNA, or proteins), a multiple sequence alignment (MSA) can show how sequences in a family relate to each other. In MSA construction, the goal is to produce columns of aligned residues (or nucleotides) that are structurally similar and related to each other, i.e. diverged form a common ancestor. Usually, an MSA has to be inferred from primary sequence alone considering structural and evolutionary conservation[241].

Automated generation of MSA is tedious and subject of extensive research in computational biology. Manually refined high quality alignments of proteins produced by biologist continue to be superior than purely automated methods[242]. The issue of automated methods is that the computation of an exact MSA is a nondeterministic polynomial time complete problem (NP-complete) given any sensible biological criterion and only feasible for unrealistically small data sets[243]. Thus, the computation of MSA depends on approximate algorithms or heuristics which are not guaranteed to give an optimal solution. The by far most widely implemented approach is the progressive alignment technique[244] which starts by pairwise alignment of the most similar sequences progressing to the most distantly related following a *guide tree*[245]. The most frequently used software solutions for protein and DNA sequences are[246] (i) ClustalW[247], (ii) MUSCEL[248], (iii) T-Coffee[249], (iv) MAFFT[250], (v) ProbCons[251],

and (vi) Kalign[252].

The computation of MSA for RNA molecules is even more complex. Sequence similarity among different RNA molecules is often remote within well-known families. In 1966, Madison *et al.* compared two yeast tRNA molecules (tyrosine and alanine) and concluded that in spite of limited sequence similarity, very similar based-paired structure models can be constructed[253]. Moreover, it is thought that functional secondary structures of RNA molecules are conserved in evolution[254]. Therefore, RNA alignment algorithms cannot rely on sequence alignment techniques alone, and should incorporate the information of a secondary structure model. Usually, this model has to be inferred from primary sequence data. Also, RNA alignments are complicated by long-range interactions due to base-pairing.

Multiple RNA sequence and structure-based alignments can be divided into two major classes, probabilistic and non-probabilistic approaches. Probabilistic approaches are based on stochastic context-free grammars (SCFG), the analogue to profile hidden Markov models (profile HMM). The quality of the computed multiple structural alignment strongly depends on an initial multiple alignment which is required as input (e.g. Cove[255], Infernal[256], and Pfold[257]). Non-probabilistic approaches, such as (i) RNAforester[258], (ii) MARNA[259], and (iii) PMcomp/PMmulti[260], require a known or predicted input structure. Simultaneous folding and aligning of two RNA sequences of length n was first introduced by Sankoff in 1985. However, this method is based on a dynamic programming algorithm and is not practical in terms of CPU time $O(n^6)$ or memory $O(n^4)$[261]. Since then, several Sankoff-style derivates have been developed. One efficient variant of these approaches is LocARNA (local alignment of RNA) and the multiple version mLocARNA[262]. LocARNA uses base pair probabilities computed by McCaskill's partition function folding algorithm[263] as structural input. mLocARNA computes multiple structural alignments following a progressive alignment strategy.

In this study, the goal was to investigate miRNA gene families and their relationship among distantly related nematodes. To visualize different miRNA families and derive phylogenetic trees, I computed respective pairwise or multiple sequence and structure-based alignments for each individual family with at least two precursors using mLocARNA[262] (v1.6.1). In miRNA target recognition, the seed sequence (here defined as positions 2-8 of the miRNA arm) is of great importance and crucial for miRNA regulation[113]. Therefore, mLocARNA was constrained to align each precursor at the seed sequence, which was determined by the most conserved arm for each miRNA. The resulting MSA were visualized using custom R scripts plus two R packages [R4RNA

(v0.1.4)[264] and Phangorn (v1.6.0)[265]] and the RNAalifold[266] webserver*.

2.3.5.3 Building Phylogenetic Trees

The Relationship of characters, i.e. any genetic, structural, or behavioral feature of a species including miRNA genes, can usually be represented by a phylogenetic tree. This tree can be inferred from observations upon existing organisms. In 1962, Zuckerkandl and Pauling demonstrated that molecular sequences provide sets of characters that carry large amounts of evolutionary information. Assuming that these sequences have descended from some common ancestral gene in a common ancestral species, a likely phylogeny of species or genes can be inferred given a set of molecular sequences from different species in question.

A phylogenetic tree consists of branch nodes connected by edges. Terminal nodes correspond to observed sequences and are called leaves. In this study, all trees were assumed to be rooted and therefore binary, i.e. an edge that branches splits into two daughter edges. The length of each edge of a tree is defined by some measure of distance between sequences. True biological phylogeny has an ultimate ancestor of all sequences called *root*. Some tree building algorithms provide information about the position of the root and others, like parsimony and probabilistic models, are uninformative about its location[241].

For simplicity, I used a more intuitive tree building method that starts with a set of distances d_{ij} between each pair i, j of sequences in a given data set. There are many different ways to define distances. A distance function needs to fulfill the definition of a metric d with the following properties[268]:

1. $d(i,j) > 0 \quad$ for $i \neq j$.
2. $d(i,j) = 0 \quad$ for $i = j$.
3. $d(i,j) = d(j,i) \quad \forall\ i$ and j (symmetry).
4. $d(i,k) \leq d(i,j) + d(j,k) \quad \forall\ i$, j, and k (triangle inequality).

Any set of distances satisfying all four metric properties will produce an additive tree. Moreover, a tree can be ultrametric if $d(i,k) \leq \max\{d(i,j), d(j,k)\}$ ($\forall\ i$, j, and k) holds.

*http://rna.tbi.univie.ac.at/cgi-bin/RNAalifold.cgi; accessed June 18, 2013

In this study, I used the score of the pairwise sequence structure alignment computed by LocaARNA for each individual miRNA family and transformed them into distances. Thus, the distances correspond directly to LocARNA-scores. As successfully applied in the LocARNA-based clustering approach[262], the distance $d(i, j)$ between a pair of RNA sequences i and j were defined by

$$d(i, j) = \max\{0, q - score(i, j)\},$$

where $score(i, j)$ is the LocARNA-score of i and j, and q is the 99%-quantile of all pairwise scores. The resulting N x N distance matrix, where N is the number of sequences, can be used to derive a rooted tree T by hierarchical clustering.

To infer phylogenetic trees for each individual miRNA gene family, I applied the unweighted pair group method using arithmetic averages (UPGMA)[269]. Note that UPGMA produces ultrametric trees that assume a constant rate of evolution in which edge lengths can be viewed as times measured by the molecular clock. This means that at all points in the tree the divergence of sequences is assumed to occur at the same constant rate. Hence, if the molecular clock property fails, the resulting tree may be reconstructed incorrectly[241]. In this case, there are other algorithms like *neighbour-joining*[270] that reconstruct the tree correctly. Nevertheless, UPGMA is a simple, fast, and intuitive clustering method for reconstructing the topology of phylogenetic relationships. In the following, I will give an overview about different hierarchical clustering methods.

Hierarchical Clustering

Agglomerative hierarchical clustering merges clusters iteratively using a bottom-up approach. The basic procedure for clustering a set of N sequences given a N x N distance matrix is as follows[241]:

1. Assign each sequence i to its own cluster C_i. The distances between two clusters C_i and C_j will be the distances d_{ij} defined in the distance matrix. For each sequence one leaf of tree T will be defined and placed at height zero.

2. Merge the most similar pair of clusters C_i and C_j into a new cluster C_k. Create a node k of T with daughter nodes i and j at height $d_{ij}/2$.

3. Compute the distances d_{kl} between C_k and all other cluster C_l.

4. Add C_k to the current clusters and remove C_i and C_j. The number of total clusters will be reduced by one.

5. Repeat steps 2-4. When only two clusters C_i and C_j remain, place the root at height $d_{ij}/2$.

Distances between clusters can be calculated in different ways (step 3). Frequently used techniques are single-linkage, complete-linkage, and average-linkage clustering[271]:

Single-linkage The distance d_{kl} between two clusters C_k and C_l is equal to the shortest distance from any sequence i of C_k and any sequence j of C_l:

$$d_{kl}(C_k, C_l) = \min_{\substack{i \in C_k \\ j \in C_l}} d_{ij}.$$

Complete-linkage The distance d_{kl} between two clusters C_k and C_l is equal to the largest distance from any sequence i of C_k and any sequence j of C_l:

$$d_{kl}(C_k, C_l) = \max_{\substack{i \in C_k \\ j \in C_l}} d_{ij}.$$

Average-linkage (UPGMA) The distance d_{kl} between two clusters C_k and C_l is equal to the average distance from any sequence i of C_k and any sequence j of C_l:

$$d_{kl}(C_k, C_l) = \frac{1}{N_{C_k} N_{C_l}} \sum_{i \in C_k} \sum_{j \in C_l} d_{ij},$$

where N_{C_k} and N_{C_l} denote the respective number of sequences in cluster C_k and C_l.

2.3.5.4 Performance Evaluation

To evaluate the performance of my miRNA homology assignment strategy, particularly the functionality of two features has to be demonstrated: (i) correct grouping of miRNAs into families and (ii) correct ordering of the multiple sequence-structure alignment of a known miRNA family.

To test the functionality of the grouping of miRNAs into distinct families, I compiled a set of well-known *let-7* family miRNAs from miRBase (v20) as control data, combined these with 50 randomly generated miRNAs based on the sequences in the control set, and inferred families from this compiled data set according to Section 2.3.5.1. Initially, a set of 52.000 random di-nucleotide shuffled miRNAs (1000 randomly generated miRNAs for each of the 52 *let-7* miRNA from eight animal clades) were generated using uShuffle* with the following parameters: '-n 1000 -k 2' and default otherwise[272]. Fifty miRNAs were then selected randomly. The eight animal clades human (*hsa*), chimpanzee (*ptr*), mouse (*mmu*), rat (*rno*), fruit fly (*dme*), nematode (*cel*), planarian (*sme*), and sea urchin (*spu*) were chosen (see miRBase database* for three-letter code information of species).

Moreover, to demonstrate that this method produces well annotated multiple alignments reflecting a correct phylogenetic relationship, I computed a multiple sequence-structure alignment on the same compiled data set as described in Section 2.3.5.2. The results of the evaluation are presented in Section 4.5.

2.3.6 Single-Mutation Seed Network

To investigate the seed neighborhood of identified candidate regulators for potential conservation of expression signatures, I inferred a single-mutation seed network from all seed sequences of the miRNA sets in *C. elegans*, *P. pacificus*, and *S. ratti*. For this, I built an undirected graph with nodes representing seeds using the R package igraph (v0.6.2)[240]. Seed sequences that differ in a single nucleotide were connected by edges. The network was visualized using Cytoscape 2.8.2[273].

*http://digital.cs.usu.edu/~mjiang/ushuffle; downloaded August 7, 2013
*http://www.mirbase.org

Chapter 3

FLEXBAR - Flexible Barcode and Adapter Processing for Next-Generation Sequencing

The Flexible Barcode and Adapter Remover (FLEXBAR) originated from the Flexible Adapter Remover (FAR) and has been developed by Matthias Dodt in the bioinformatics group of Christoph Dieterich. As part of this project, I developed the adapter removal feature for SOLiD color space reads and focused on the application of small RNA-seq in letter and color space. Additionally, I was involved in the design of FAR and in the development of specific features of the subsequently added barcode detection function for demultiplexing. The final version of FLEXBAR (paper version) has been extensively revised and enhanced by Johannes Röhr through the introduction of novel and extended features, a cleanup in the source code, redesigned command-line interface, and optimized parameter settings.

In this chapter I will present the general concept of FLEXBAR with an emphasis on FLEXBAR's color space adapter removal mode for the application of small RNA-seq. Parts of this chapter have been published in 2012 in the special issue "Next-Generation Sequencing Approaches in Biology" in the journal *Biology*[1].

3.1 Background

Next-generation sequencing technologies, such as Illumina's GA/HiSeq, Applied Biosystems SOLiD and Roche 454, produce millions of short reads by massive parallel sequencing. All of these approaches introduce sequence tags that are typically ligated to the pool of target sequences. Sequence tags can be located anywhere in a given short sequencing read and often overlap with the sequenced region. These tags should be removed prior to downstream read processing. Failure to do so can result in a large amount of not mappable reads. Evidently, adapter sequences may confound any subsequent analysis step. A simple positional read trimming or quality-based read trimming is usually not sufficient to rule out mis-assemblies or low mapping rates.

Adapter sequence tags are inherently used in every sequencing platform to initiate sequencing or for other internal processing purposes (Figure 1.4B). Hence, short reads from any sequencing platform may contain adapters or fragments of adapters. Moreover, recent increases in sequencing capacity facilitate pooling of samples (multiplexing) in one sequencing reaction by the introduction of barcode sequences. Barcodes are used to tag a particular origin in a complex mixture of short reads. Several read processing scenarios emerge due to the use of adapter and barcode sequences. Our FLEXBAR software unifies high-processing speed, versatile approaches to basic filtering, quality trimming, barcode detection followed by demultiplexing, and adapter removal. It supports all current next-generation sequencing platforms, e.g. adapter sequences may be removed in letter or color space. FLEXBAR is not limited in read length and may be well suited for processing third-generation reads. In the following, I will discuss program features, implementation and usage, and compare them to other state-of-the-art software solutions.

3.2 Program Features

The rich feature set of FLEXBAR addresses many potential applications in single-end, paired-end, and mate-pair sequencing. As discussed in Materials & Methods 2.3.1, typical workflows involve a quality clipping step, demultiplexing, which potentially includes barcode removal, followed by a separate adapter removal step. All of these steps may be executed within one program call of FLEXBAR (Figure 3.1). The default parameters are optimized to deliver high quality results (especially Illumina and SOLiD)

for a large number of scenarios. However, customization of settings might improve results for specific applications.

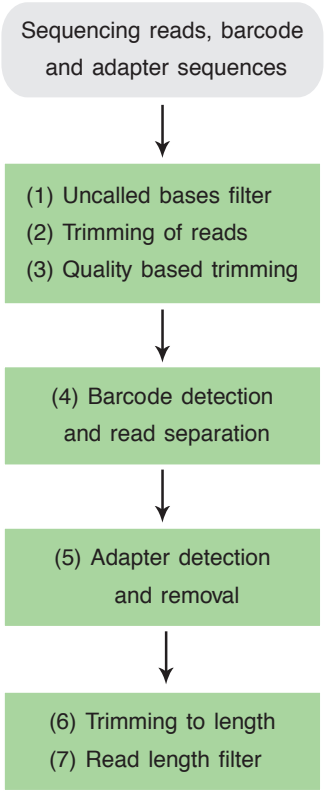

Figure 3.1: FLEXBAR's internal workflow
FLEXBAR takes sequencing reads, barcode and adapter sequences as input. All preprocessing steps emphasized in Materials & Methods 2.3.1 can be performed in one program call. In total, seven processing steps can be applied on the set of input sequences. These steps can be split into four categories: (i) basic clipping and quality filtering, (ii) barcode recognition and processing, (iii) adapter recognition and removal, and iv) output filtering.

FLEXBAR has been implemented in C++ using the Seqan library[274]. Multi-threading has been implemented with the Intel Threading Building Blocks library (TBB)*.
FLEXBAR detects target sequences by an overlap sequence alignment based on the Needleman-Wunsch algorithm[189]. An overlap (or semi-global) alignment uses the same recurrence relations as a global alignment but does not penalize gaps at both ends of the alignment (Figure 3.2; see Materials & Methods 3.2.1 for detailed description of an

*http://www.threadingbuildingblocks.org; accessed August 14, 2012

overlap alignment).

FLEXBAR offers maximal flexibility in target sequence recognition considering base substitutions, insertions and deletions. Moreover, the user is free to choose all alignment scoring parameters, the minimal overlap, and a threshold on sequence similarity. Default parameters are preset and were chosen to work best for Illumina GA II/HiSeq and AB SOLiD sequencing data. A simple perfect matching approach to expected sequence tags might not be adequate for sequencing platforms with elevated error rates. Furthermore, reads encoded in letter as well as in color space can be processed.

FLEXBAR supports five trim-end modes for sequences, which cover most sequencing applications: (i) ANY (where), (ii) LEFT, (iii) RIGHT, (iv) LEFT_TAIL or (v) RIGHT_TAIL trimming (Figure 3.3). These modes, which will be discussed in Section 3.2.2, can be independently combined for adapter and barcode sequence recognition in single- or paired-end data. Moreover, barcode reads might be separated from the actual single- or paired-end read set (as in Illumina TruSeq^{TM} sequencing). In the following, I will present the algorithm for an overlap alignment in detail.

3.2.1 Algorithmic Implementation

FLEXBAR uses the concept of an overlap alignment to detect target sequences, which is essentially a type of global alignment without penalizing gaps at the end. To this end, the first row and column of the dynamic programming matrix are initialized with zeros, and the alignment score maximum is searched in the last row and column of the alignment matrix. Hence, the dynamic programming matrix F indexed by i and j (one for each sequence x and y) for an overlap alignment is constructed recursively, where $F(i,j)$ is the score of an optimal alignment between the segments $x_{1..i}$ up to x_i and $y_{1..j}$ up to y_j. The matrix is initialized by $F(i,0) = 0$ for $i = 1,..,n$ and $F(0,j) = 0$ for $j = 1,..,m$, where n and m are the lengths of sequence x and y, respectively. The matrix is filled from top left to bottom right using the following recurrence relationship

$$F(i,j) = \max \begin{cases} F(i-1,j-1)+s(x_i,y_j), \\ F(i-1,j)-d, \\ F(i,j-1)-d; \end{cases}$$

where d denotes the cost for inserting a gap. While the values $F(i,j)$ are filled, pointers in each cell back to the cell from which its $F(i,j)$ was derived, has to be stored. To derive the optimal overlap alignment, the maximal score F_{max} is searched in the last row

$(i, m), i = 1, ..., n$ and the last column $(n, j), j = 1, ..., m$ of F. To find the alignment that leads to F_{max}, the path that led to F_{max} has to be found using the *traceback* procedure; i.e. the alignment is build reverse, starting from the cell F_{max} following the pointers until the top row or column is reached (Figure 3.2)[241]. End gaps should be added if applicable.

The algorithm takes $O(nm)$ CPU time and $O(nm)$ memory storage, where n and m are the lengths of the sequences, respectively. Because n and m are generally comparable, the algorithm is said to be $O(n)^2$.

3.2.2 Trim-end Modes

The user can choose among five different trim-end modes depending on the experimental setup (Figure 3.3). We will explain these trim-end modes based on a short read of length n and an adapter sequence of length m. All trim-end modes are available for barcode (option $--$barcode-trim-end) and adapter (option $--$adapter-trim-end) sequence recognition. We assume that $m < n$ holds.

1. **ANY:** The adapter sequence is searched anywhere within the given short read. In case of an adapter match, the longer non-matching substring of the read is retained.

2. **LEFT**: The matching adapter sequence is located in a prefix $p[1..k]$ of the short read with $k \leq n$. The corresponding short read prefix including the adapter sequence is removed.

3. **RIGHT**: The matching adapter sequence is located in a suffix $s[(n-k)..n]$ of the short read with $k < n$. The corresponding short read suffix including the adapter sequence is removed.

4. **LEFT_TAIL**: This is a special case of mode 2. We only consider the first m bases or colors of the short read. The read is trimmed from the 5' end.

5. **RIGHT_TAIL**: This is a special case of mode 3. We only consider the last m bases or colors of the short read. The read is trimmed from the 3' end.

The default parameters of FLEXBAR are set such that the ANY mode is used for barcode recognition and the RIGHT mode is used for adapter recognition.

3.2.3 Quality Clipping and Read Filtering

Many sequencing platforms provide phred-scaled quality scores for individual base calls (see Section 2.3.1.1 for details). FLEXBAR provides multiple options to filter reads and target read quality:

−−max-uncalled: Sets a threshold on the allowed number of unidentified bases within a given read. No uncalled bases are allowed as per default settings. All reads that exceed this threshold are excluded from barcode and adapter processing.

−−pre-trim-left: Allows trimming of a certain number of bases on the left end of short reads. Disabled by default.

−−pre-trim-right: Equivalent to −−pre-trim-left. This option allows trimming of a certain number of bases on the right end of short reads. Disabled by default.

−−pre-trim-phred: Trims all read positions from right to left up to the first base or color having a quality value larger or equal to the given quality score cutoff. Disabled per default.

−−post-trim-length: Specifies the number of bases to which reads are truncated from 3' end after all removal steps have been applied. This option is disabled by default.

−−min-readlength: Defines the minimal read length (18 nt by default). All reads that are equal or longer than the minimal read length are retained. All other reads are discarded or written to a special output file if requested.

3.3 Program Usage

A typical use case starts by defining the set of input reads. As default the read format is defined as fasta/q. If the read format is csfasta, the color space option −c has to be specified: flexbar −t <STRING> −r <FILE> [−p <FILE>] [−c]

Example: flexbar −t processed −r single_end_F3.csfasta −c

Option −t defines the prefix for the output filenames. Option −c specifies the color space read format csfasta. If a second read set is specified (option −p), paired-end

reads are processed, and read pairings are maintained in the output. Barcode and/or adapter sequences can be defined by the following options:

−a <FASTA FILE> or −as <STRING>
−b <FASTA FILE>

The option −as can be specified if only one adapter sequence needs to be detected. Depending on the sequencing setup, barcode and adapter sequences can be located in different or reside within the same read. For example, barcode reads in the Illumina TruSeq^{TM} system are represented by a second or third read set, which are sequenced independently from the actual single- or paired-end reads. Barcode detection precedes the adapter removal step in FLEXBAR. The user may specify the location of separate barcode reads by setting the option −br. If this option is not set FLEXBAR assumes that barcodes reside within the first read set (−r). Note that neither barcode detection nor adapter removal is a mandatory step in FLEXBAR. The available trim-end modes are selected by

−−adapter-trim-end <ANY | LEFT | LEFT_TAIL | **RIGHT** | RIGHT_TAIL>
and
−−barcode-trim-end <**ANY** | LEFT | LEFT_TAIL | RIGHT | RIGHT_TAIL>
(Default values are shown in bold characters.)

To optimize the barcode detection process for different experimental setups, the user can adjust alignment parameters as desired. Furthermore, the option −−barcode-keep determines if the barcode is removed or not from the assigned reads. Adapter removal is effected in a similar manner yet controlled by an entirely different set of parameters, the −−adapter-* options. We separated these parameters to allow use in a wide range of applications, for specific contexts it is desirable to apply distinct stringencies through parameter adjustments in the barcode and adapter detection process. For example, a particular situation may require asking for a higher specificity in assigning barcodes and, at the same time, require being more sensitive in adapter sequence recognition. Finally, all FLEXBAR processing steps outlined in Figure 3.1 and command line parameters including default settings are output in detail by calling the help page of the program (option −−help).

3.4 Program Validation

Numerous competing software solutions exist that emphasize barcode recognition and adapter trimming of short sequencing reads. The comparison of a number of program features in FLEXBAR and widely accepted (FASTX toolkit*) and recently published (CUTADAPT[275], BTRIM[276]) programs reveals that FLEXBAR is the only software with independent barcode and adapter processing, extensive verbose outputs (e.g. for read alignments), preservation of read pairs in paired-end or mate-pair mode, and separate barcode read support (e.g. Illumina TruSeq™; Table 3.1). The only program that is also capable of adapter sequence removal in color space is CUTADAPT.

Table 3.1: Comparison of FLEXBAR features with other software solutions

Feature	FLEXBAR	FASTX	BTRIM	CUTADAPT
Color space support	Yes	No	No	Yes
Simultaneous barcode & adapter processing	Yes	No	No	No
Preservation of read pairings[a]	Yes	No	No	No
Graphical alignment output[b]	Yes	No	No	No
Separate barcode reads[c]	Yes	No	No	No

[a] Read pairs are output in sync
[b] Log files with individual read alignments
[c] One or two read files (single- or paired-end) plus additional barcode read file (e.g. TrueSeq™)

The performance of FLEXBAR was evaluated by using five typical applications: (i) adapter removal from small RNA-seq data in letter space, (ii) processing of paired-end RNA-seq read sets in letter space (2 x 100 nt), (iii) barcode detection of *in silico* generated letter space barcode reads, (iv) simultaneous barcode and adapter recognition for splice leader detection in color space data, and (v) adapter removal from small RNA-seq data in color space [for benchmarks I-IV see Dodt *et al.* (2012)].

In the following, I demonstrate FLEXBAR's adapter removal functionality for small RNA-seq color space data, which was profiled during the microRNA study of free-living and parasitic nematodes presented in Chapter 4. The results are compared to CUTADAPT, the only program from our selected list of software solutions that can handle color space reads.

*http://hannonlab.cshl.edu/fastx_toolkit; accessed July 25, 2012

3.4.1 Adapter Removal from microRNA Short Reads in Color Space

Small RNA high-throughput sequencing is widely used to profile small ncRNAs, such as miRNA genes (see Introduction 1.4.3 for details). In small RNA-seq applications, the length of sequenced reads (∼35-50 nt) typically exceeds the length of small RNAs such as miRNAs (∼22 nt); therefore, adapter fragments that are not part of the biological molecule are oftentimes sequenced. Before reads are mapped to a reference genome, these fragments need to be detected and removed.

In order to test FLEXBAR's adapter removal functionality for color space reads and its performance, I detected and removed adapter sequences from small RNA-seq color space data and mapped the processed reads to the reference genome. As test data, I used a subset (one-quarter of a flow cell) of the *C. elegans* data set 2 (Table 2.1), which was profiled using the SREK Kit and ABI's SOLiD with a read length of 35 nt. To detect and remove adapter sequences, reads were processed by FLEXBAR v2.4 and CUTADAPT v1.21. To evaluate the performance of both programs, I mapped the processed reads (length ≥17 nt) to the *C. elegans* genome using the mapper SHRiMP2 v2.23[197]. For performance criteria, the number of uniquely mappable reads and the number of bases contained in these uniquely mappable reads were used (Figure 3.4). As a control, reads that were not processed by any adapter removal tool were also mapped to the *C. elegans* genome using SHRiMP2 with the same mapping parameters ('−o 10 −−max-alignments 10 −h 80%' and default otherwise).

For this benchmark, a set of 5,525,403 color space reads corresponding to 1,247,661 unique sequences were processed by FLEXBAR and CUTADAPT using a single core on a Dual Opteron 2356 (Quad-Core at 2.3 GHz) with 64 GB of memory. The following command line options were used:

flexbar −a abiAdapter.fa −r solid0174_ worms_1_F3.csfasta −c −t output −ae RIGHT −u 10

cutadapt −e 0.3 −m 17 −a CGCCTTGGCCGTACAGCAG solid0174_ worms_1_F3.csfasta −o output.csfasta

As expected, without a prior adapter detection and removal step only 3,903 reads corresponding to 136,033 bases could be mapped uniquely. The vast majority of reads (5,496,183) did not map to the genome. Comparing the mapping results of reads processed by FLEXBAR and CUTADAPT demonstrated that FLEXBAR performed better than CUTADAPT with ∼67 Mb vs. ∼63 Mb uniquely mappable bases (corresponding to ∼2.91 Mb vs. ∼2.75 Mb uniquely mapped reads). Although these numbers

do not differ much in this rather small data set, these differences may be much bigger when larger data sets are processed. Moreover, assuming that the average length of a miRNA is 22 nt, 4 million of uniquely mappable bases would correspond to an average of approximately 181,818 miRNA sequences, which may have been missed using CUTADAPT. Thus, the detection and removal of adapter sequences is an essential preprocessing step during the analysis of small RNA-seq data in color and letter space [for letter space see benchmark I in Dodt *et al.* (2012)].

The required runtime and memory consumption is another important aspect when evaluating the performance of software programs. To process these color space reads, FLEXBAR needed approximately one-third (19 s) of the compute time used by CUTADAPT (51.99 s) on a single core. Since FLEXBAR can compute in parallel using multiple threads, processing of larger data sets will be less problematic in runtime compared CUTADAPT. On modern compute systems the observed memory consumption is less of an issue with FLEXBAR using only ∼37 MB memory versus CUTADAPT ∼50 MB on a single core. Overall, these evaluations reveal that when processing small RNA-seq SOLiD color space reads, FLEXBAR performs better than CUTADAPT and will most likely outperform CUTADAPT when dealing with much larger data sets.

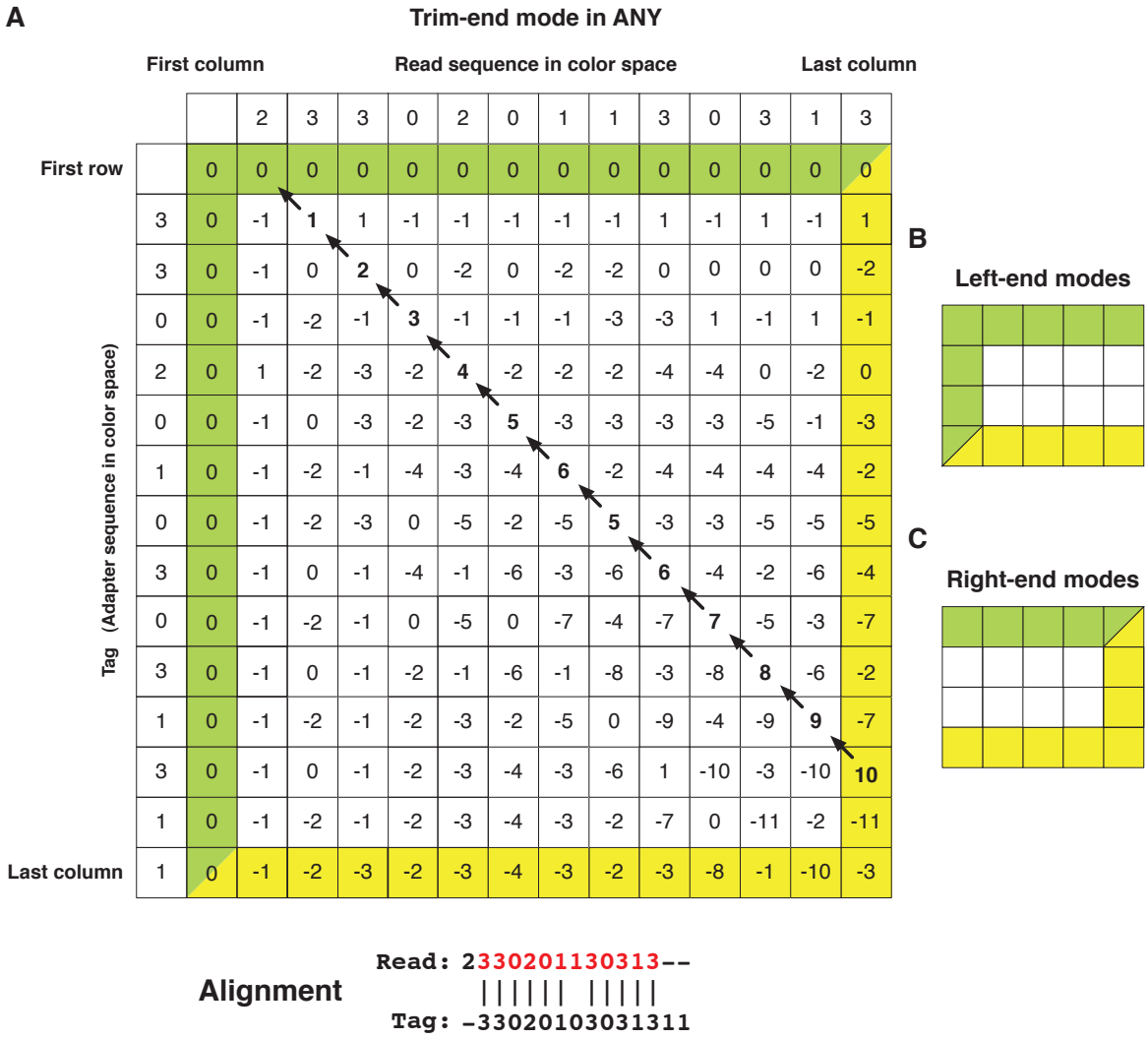

Figure 3.2: Sequence tag recognition with a dynamic programming matrix
(A) Example of dynamic programming matrix for trim-end mode ANY and corresponding overlap alignment in color space, with arrows indicating traceback pointers (default scoring parameters: match = 1, mismatch = −1, and gap = −7). Values on the optimal alignment path are shown in bold. Gaps at either end of any sequence are not penalized. The top-scoring alignment may start at any position in the read (first row; green area) or at any position in the sequence tag (first column; green area). The traceback of an alignment starts at the maximal score located at any position in the read (score maximum in last row; yellow area) or sequence tag (score maximum in last column; yellow area). **(B)** Left-end modes are represented. An alignment can start anywhere in the first row or column (green), but must end in the last row (yellow). **(C)** Right-end modes are represented. In contrast to left-end modes, an alignment must start in the first row (green) and end anywhere in the last row or column (yellow). Note that the read length is truncated to the sequence tag length in LEFT_TAIL and RIGHT_TAIL mode.

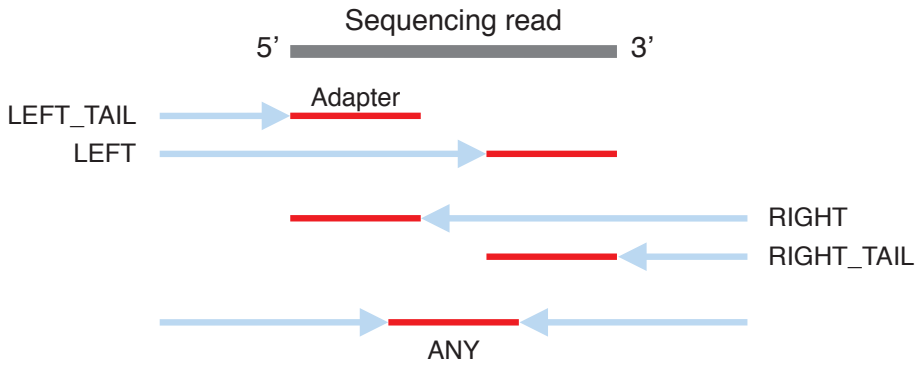

Figure 3.3: Graphical representation of FLEXBAR's sequence trim-end modes
The gray bar depicts the currently processed sequencing read (length n). The best alignment of an adapter sequence (length m; shown in red) can be located anywhere in the demarcated region (arrow + adapter region), which differs according to the selected trim-end mode (see main text). The name of the trim-end mode refers to the part of the short read that is removed. In the left modes, the 5' end is trimmed; in right modes, the 3' end, and otherwise the shorter end is removed.

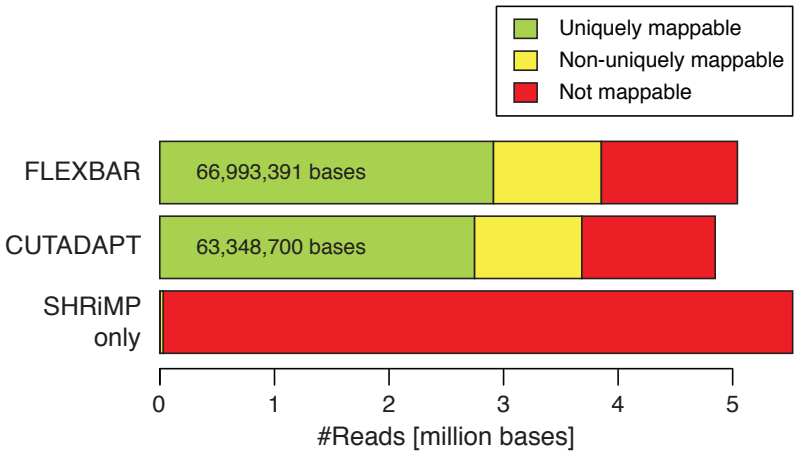

Figure 3.4: Benchmark V - Comparison of FLEXBAR and CUTADAPT
Mapping statistics of untreated reads of control (SHRiMP only) compared to FLEXBAR's and CUTADAPT's adapter removal functionality for color space reads. Significantly more reads (measured as uniquely mappable bases) processed by FLEXBAR mapped to the *C. elegans* genome (FLEXBAR: ~67 Mb vs. CUTADAPT: ~63 Mb).

Chapter 4

Conserved microRNAs are Candidate Post-Transcriptional Regulators of Developmental Arrest in Free-Living and Parasitic Nematodes

In this chapter I applied the bioinformatic approaches and computational strategies introduced in the previous two Chapters to address the question of whether miRNA genes impact developmental arrest and long-term survival in dauer and dauer-like stages, i.e. the infective stage of parasites. In particular, I am interested to discover whether conserved regulatory modules exist that would support the long-standing hypothesis that dauer and infective larvae share a common origin.

Parts of this chapter are based on the publication Ahmed *et al.*, which was published in July 2013 in the journal *Genome Biology and Evolution*[2]. As part of this collaborative project, I designed and performed all computational experiments.

4.1 Background

Nematodes inhabit a wide range of ecological niches encompassing free-living species as well as obligate parasites of plants and animals. Intriguingly, the basic life cycle of nematodes is well conserved across nematode clades and typically involves four larval molts[7]. The free-living nematode *C. elegans* has become the reference model to study developmental responses in the context of environmental changes and represents an ideal organism to study short RNA biology[277]. Under unfavorable conditions, such as starvation and crowding, *C. elegans* enter dauer diapause, a developmentally arrested, stress-resistant, and long-lived stage[20,21]. Dauer larvae share many traits with infective larvae of true parasites. Moreover, the dauer and infective larvae fate is determined by a conserved endocrine signaling mechanism[33,40]. Accordingly, dauer larvae have been suggested as an evolutionary precursor of infective larvae that facilitated the repeated evolution of parasitism (a preadaptation)[278].

I want to investigate what regulates these dauer and infective stages. In particular, I am interested in the post-transcriptional regulation of the dauer/infective larval fate and in the role of miRNAs in this context. Several lines of evidence suggest that post-transcriptional regulatory mechanisms dominate the transition from dauer back into the reproductive life cycle[41–44]. Moreover, recent studies demonstrate that miRNAs are involved in the regulation of lifespan as well as L1 and dauer diapause[120–126].

Here, I compare the small RNA complement and its expression changes in dauer/infective vs. nondauer samples of three nematode species with three different life styles: the free-living nematode *Caenorhabditis elegans*, the necromenic nematode *Pristionchus pacificus*, and the true parasite *Strongyloides ratti*. *Pristionchus pacificus* dauer larvae and no other larval stages have been observed on beetles yet do not harm their host. Upon death of the host, dauer larvae resume development by feeding on the beetle's carcass[29]. *Strongyloides ratti* is a true parasite of the rat with a direct and an indirect life cycle[36]. It is unlikely that a species evolves directly from a fully free-living to a parasitic life style[278]. We hypothesized that *S. ratti* still maintains the ancestral free-living life cycle along with the newly acquired parasitic life cycle.

In this study, I address the role of miRNAs in the dauer/infective larvae fate by comprehensive profiling of known and novel miRNA genes in *C. elegans*, *P. pacificus*, and *S. ratti*. Small RNAs were sequenced using a multiplatform sequencing approach (ABI SOLiD, Illumina GA II, and HiSeq). First, I developed a bioinformatics workflow that

identifies novel miRNAs from letter space (Illumina) and color space (SOLiD) data sets to extend the miRNA gene sets in these species. Then, I inferred miRNA families by sequence similarity followed by the identification of conserved candidate miRNA genes of the dauer and infective larvae fate. Finally, I investigated seed changes of miRNAs with emphasis on *mir-34*, a cross-species candidate regulator, using a single-mutation seed network.

4.2 Sequencing of microRNAs from Three Nematodes

Small RNA deep sequencing libraries from dauer and mixed-stage samples of *C. elegans* and *P. pacificus* and infective L3s and mixed-stage samples of *S. ratti* were generated using a multiplatform approach (Figure 2.1). More than 196 million sequencing reads for 10 small RNA libraries were obtained. As outlined in the bioinformatic methods section, I implemented short read processing, mapping, and miRNA gene inference in a custom bioinformatics pipeline (Figure 4.1; Materials & Methods 2.3). Before mapping to the respective reference genomes, poor quality reads were filtered (if the quality of each read was smaller than 10 in more than 10 positions) and corrected for read error using SAET 2.2* (SOLiD reads only). Barcode sequences were detected (HiSeq lane) and adapter sequences removed using FLEXBAR[1]. For subsequent analyses, reads were collapsed to non-redundant data sets. Approximately 120 million small RNAs ($\geq$ 18nt) mapped to their respective genomes (Table 4.1).

Our multiplatform small RNA-seq approach highly enriches for miRNAs relative to other types of small RNAs. For example, 88% of reads obtained by SOLiD sequencing (data set 1; Table 4.1) and 79% of reads obtained by Illumina sequencing (data set 2; Supplemental Table B.1) mapped to mature miRNAs in *C. elegans* (miRBase v18). In total, 193 out of 223 (87%) previously annotated *C. elegans* miRNAs could be identified (data set 1). In the remainder of reads, I detected sense and antisense hits to other small non-coding RNAs, including 21U-RNAs (the so-called 'pi-RNAs' in *C. elegans*) as well as protein-coding regions, 5' UTRs, and 3' UTRs (Supplemental Table B.1). In *P. pacificus*, I detected 123 out of 124 previously annotated miRNA genes (99%) in our small RNA data sets. No miRNA genes have been annotated for *S. ratti* so far. Based on the apparent quality of our data, I could exploit our high sequencing depth to extend, refine, and define the miRNA gene complements of *C. elegans*, *P. pacificus*,

*http://solidsoftwaretools.com/gf/project/saet; accessed December 2, 2009

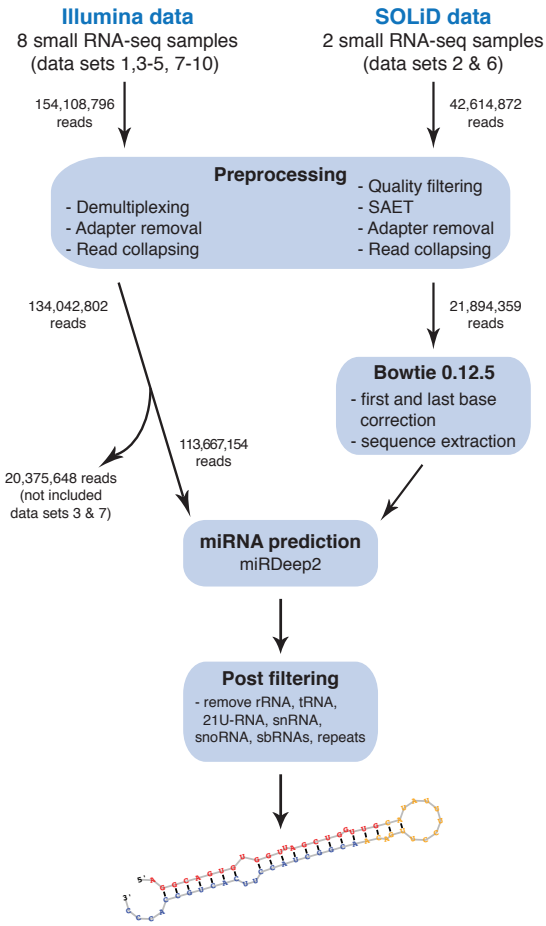

Figure 4.1: Bioinformatic analysis workflow for 10 small RNA data sets
This workflow involves the following computational steps: read quality filtering, barcode detection, error correction, adapter removal, and mapping to target genome sequences. Reads with less than 18nt were discarded from further analysis.

and *S. ratti*. Furthermore, I corrected or complemented sequence annotations for 5' or 3' arms of known miRNAs in both *C. elegans* and *P. pacificus* (a list of all miRNA sequences including sequence corrections are tabulated in Supplemental Table B.2).

In conclusion, our multiplatform deep sequencing approach is comprehensive enough to identify tissue and stage-specific miRNAs, such as *lsy-6*, a very rare miRNA, which is only expressed in less than 10 cells[279] and is hardly detected by qRT-PCR[125].

Table 4.1: Mapping statistics of 10 small RNA datasets
Mapping statistics of our 10 deep sequencing datasets derived from *C. elegans*, *P. pacificus*, and *S. ratti*.

Data set	Species	#Raw reads	#Preprocessed reads[†]	#Reads mapped	#Uniquely mapped reads
1	*C. elegans*	20,557,719	17,887,281	16,891,417[1]	15,753,555
2	*C. elegans*	21,307,436	14,342,977	11,173,018[1]	10,646,919
3	*C. elegans*	10,290,812	9,505,958	8,687,443[1]	8,233,943
4	*C. elegans*	10,349,552	10,042,941	9,450,442[1]	9,055,836
5	*P. pacificus*	27,208,332	23,579,694	11,947,471[2]	10,510,454
6	*P. pacificus*	25,717,306	7,551,382	3,039,155[2]	2,597,439
7	*P. pacificus*	11,347,692	10,869,690	5,859,988[2]	5,057,498
8	*P. pacificus*	10,382,820	9,412,219	3,685,876[2]	3,462,898
9	*S. ratti*	27,540,069	24,244,264	22,721,261[2]	6,881,764
10	*S. ratti*	32,021,930	28,500,755	26,779,392[2]	18,837,813
Total		**196,723,668**	**155,937,161**	**120,235,463**	**91,038,119**

[†]Reads after filtering and adapter removal
[1] at most 10 times
[2] at most 20 times

4.3 Unbiased Identification of Novel microRNA Genes

The miRDeep2 program[220] was applied to predict novel miRNA genes. This program uses a probabilistic model to discriminate miRNA candidate loci consistent with the expected processing of miRNA precursors by Dicer from other spurious candidate loci (a detailed description is given in Materials & Methods 2.3.3.2). I used individual small RNA data sets to predict miRNA genes for *C. elegans*, *P. pacificus*, and *S. ratti* in developmentally arrested (data sets 4, 8, and 10) and mixed-stage (data sets 1, 2, 5, 6, and 9) samples as summarized in Figure 4.2. To predict novel miRNAs based on the SOLiD data, I modified the miRDeep2 prediction pipeline (Figure 4.1; see Materials & Methods 2.3.3.2 for details). In total, I identified 33 novel *C. elegans* miRNA candidates (24 in mixed-stage, 8 in dauer, and 1 in both; Supplemental Table B.3.1), 230 novel *P. pacificus* miRNAs (91 in mixed-stage, 26 in dauer, and 113 in both; Supplemental Table B.3.2), and 106 miRNAs in *S. ratti* (18 in mixed-stage, 8 in iL3, and 80 in both; Supplemental Table B.3.3).

These results augment and complement the set of annotated miRNAs in all three species (Figure 4.3A). Despite that miRNAs have been extensively studied in *C. elegans* with 223 annotated to date (miRBase v18), I used our multiplatform approach to expand

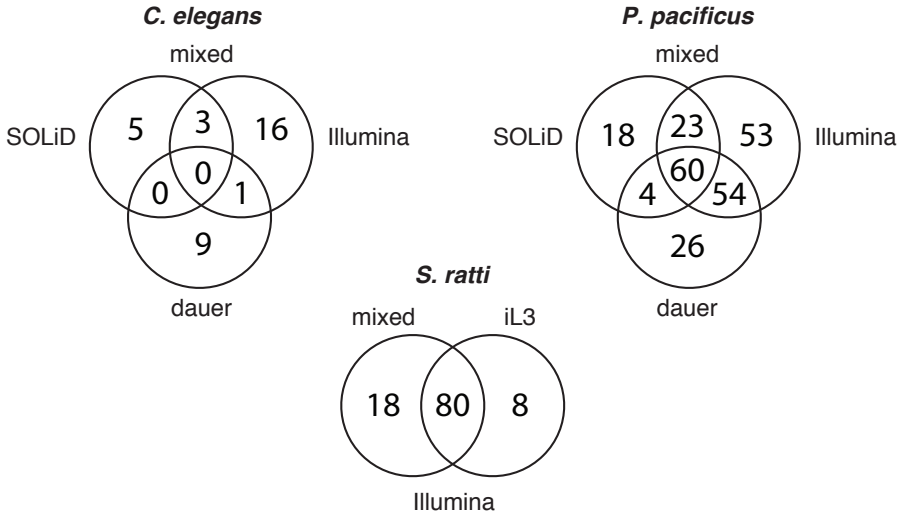

Figure 4.2: Identified miRNA genes by miRDeep2
Number of novel miRNAs predicted from small RNA sequencing data obtained from different sequencing platforms in *C. elegans*, *P. pacificus*, and *S. ratti*. I did not perform miRNA prediction on data set 3 and 7 (Table 4.1) because data set 1 and 5 represent the same samples but with a read output twice as high. Note: The number of miRNA genes corresponds to their occurrence in the respective genomes.

the set to 257 miRNA genes (one gene was duplicated on the genome). In contrast to *C. elegans*, only 124 miRNAs were annotated in *P. pacificus* (miRBase v18) based on a Roche 454 FLX sequencing run with a low sequencing depth of ∼160,000 reads[108]. Because *P. pacificus* has a significantly larger genome size compared to *C. elegans* and contains a higher number of protein-coding genes[28], I speculated that the sequencing depth to profile miRNAs was not sufficient to capture the full complement of miRNAs. Indeed, our data nearly triples the set of empirically supported miRNAs in *P. pacificus*, bringing the total to 362 (six genes occurred multiple times on the genome). Our data provides the first miRNA gene set annotation in *S. ratti*, which is the first annotation for any *Strongyloides* parasite. The size of the predicted miRNA gene complements correlates well with the species genome sizes (Figure 4.3B). In addition, I was able to resolve the 5' or 3' arms of known miRNA genes in *C. elegans* and *P. pacificus* that have not been annotated so far (Supplemental Table B.2).

These observations and the fact that I found both arms of most miRNAs covered by reads (up to 95% *S. ratti*), suggest that these candidates are bona fide miRNA genes. However, to further validate novel miRNAs experimentally, northern blot, *in*

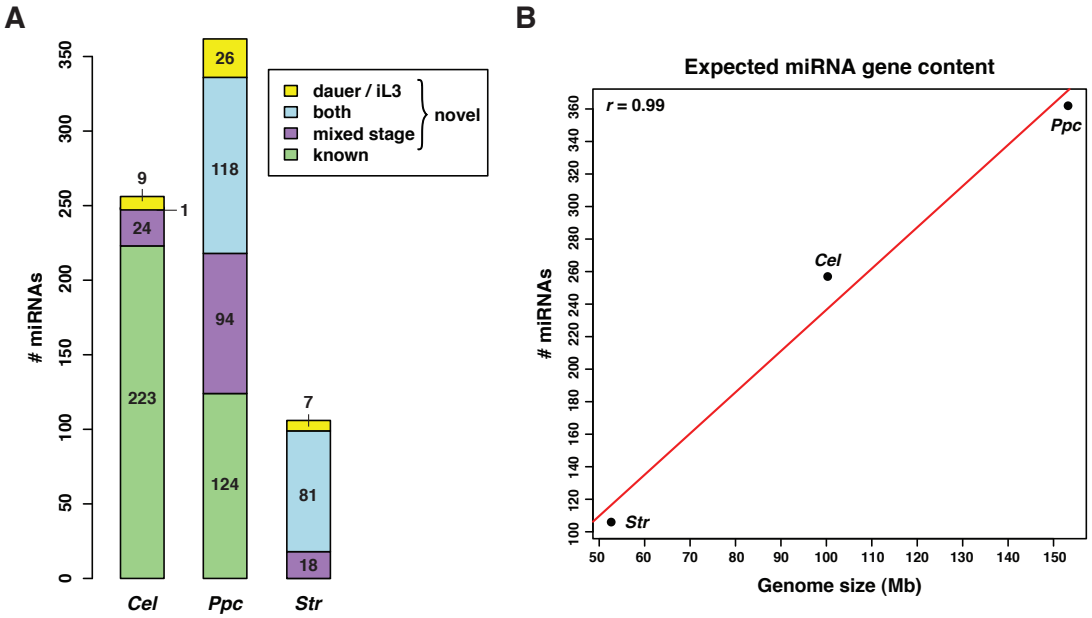

Figure 4.3: miRNA gene complement in *C. elegans*, *P. pacificus*, and *S. ratti*
(A) miRNA gene complement in all three species, including novel gene candidates. Note that several genes occur multiple times on the genome. **(B)** Expected miRNA gene content in relation to the genome size (Pearson's correlation, $r = 0.99$). The *S. ratti* genome, which is half the size of the *C. elegans* genome, contains roughly two times less miRNA genes. Additionally, the *P. pacificus* genome contains 40% more miRNA genes than the *C. elegans* genome. This is in accordance with the *P. pacificus* genome size, which is 50% larger than the *C. elegans* genome. The genome size is estimated based on the respective genome assemblies.

situ hybridization, or qRT-PCR could be applied.

4.4 Most microRNA Genes Are Not Conserved among Distantly Related Nematodes

Our deep miRNA profiling, the subsequent identification of novel miRNA candidates, and the revision of previous miRNA annotations in *C. elegans*, *P. pacificus*, and *S. ratti* provided the basis for a comprehensive phylogenetic investigation of miRNAs across these three nematode species. Typically, gene phylogenies are derived from sequence similarity (i.e. multiple sequence alignments) and a phylogenetic reconstruction method (see Materials & Method 2.3.5 for detail). For miRNAs, the seed region (position 2-8 from the 5' end) is the major determinant of target specificity and is widely used

to group miRNA genes into families[280,281]. However, the seed region is too small to distinguish cases of homoplasy from cases of common decent. To pinpoint potential cases of convergent evolution, I inferred miRNA families based on sequence similarity of the full miRNA 5' or 3' arm and contrasted them with miRNA sets, which were solely defined by seed identity.

Every precursor of a miRNA gene has the potential to generate two distinct regulatory RNAs derived from opposite strands of the stem (Figure 2.4). It is widely assumed that the more abundant sequence in small RNA sequencing data exclusively functions to suppress target transcripts (miRNA mature product), whereas its counterpart, the partially complementary sequence produced from the duplex stem, is non-functional (miRNA star product)[282]. However, experiments in *Drosophila melanogaster* have demonstrated that miRNA star species can be loaded into RISC and show regulatory activity[98–102]. Moreover, the dominate miRNA sequence in orthologous miRNAs can be processed from opposite arms as proposed in the arm-switching model in studies investigating miRNA evolution[107–110]. Since miRNA cloning involves amplification steps, sequence-specific biases arising from small RNA library preparation and sequencing technology cannot be excluded[188]. Therefore, I re-annotated all miRNA arms as '5p' or '3p' for subsequent analyses[283] and investigated 5p/3p read count ratios for *C. elegans* and *P. pacificus* miRNAs profiled from mixed-stages by different sequencing platforms (Supplemental Table B.4). Large variations in miRNA 5' to 3' arm read count ratios across sequencing platforms were present.

Because of the above arguments, I initially derived miRNA conservation levels by all 1335 miRNA 5' and 3' arms (corresponding to 725 precursors) based on sequence similarity of the full arm. Then I inferred gene families for free-living, necromenic and parasitic nematodes based on the most conserved arm of each precursor. In short, I retained the miRNA arm that belongs to the largest family or exhibited the highest seed conservation in the phyla Nematoda, Arthropoda, Lophotrochozoa, and Vertebrata and discarded the other arm (definition of miRNA age classes in Supplemental Methods). By this method, 725 precursors were grouped into 399 different gene families (Figure 4.4A). This analysis indicates that 63 (24.5%) precursors in *C. elegans*, 88 (24.3%) in *P. pacificus*, and 37 (34.9%) in *S. ratti* are conserved among all three species represented by 24 (6%) distinct families (Supplemental Table B.5). Moreover, 286 (39.4%) precursors were conserved between at least two species. Evidently, miRNA families could comprise of multiple precursor sequences. This analysis revealed that only a small fraction of miRNA families represent approximately one-quarter of all precursors

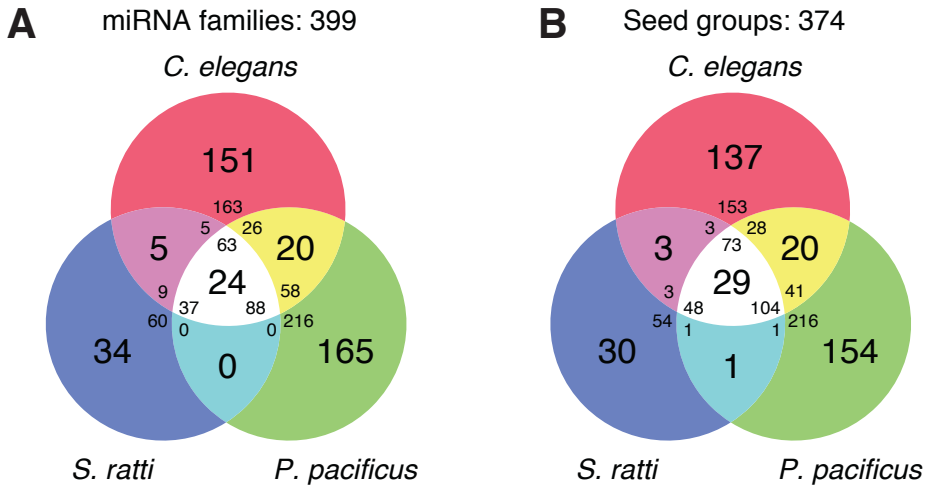

Figure 4.4: miRNA homology and seed conservation
(A) 725 precursors were stratified into 399 gene families by sequence similarity of the full miRNA arm based on the most conserved arm. If both arms of a miRNA were annotated, the arm contained in the largest group (inferred from all 1335 miRNA 5' and 3' arms) was considered. If group sizes were equal, the arm with the highest degree of conservation was considered (definition of miRNA age classes in Supplemental Methods). In case of equal conservation level, a miRNA arm was randomly chosen. **(B)** 725 precursors were stratified into 374 seed groups by perfect seed sequence identity (position 2-8) considering seeds based on the most conserved miRNA arm of a precursor. The miRNA arm was selected as explained above.

(188/725, 25.9%). Nevertheless, the majority of miRNA genes from three distantly related nematodes are not conserved. This finding is consistent with a previous study from de Wit *et al.* (2009). The authors presented the first experimental study on the evolution of miRNA genes in *C. elegans*, *C. briggsae*, *C. remanei*, and *P. pacificus* and concluded that the majority of miRNAs are conserved within the *Caenorhabditis* genus, with the notable exception of *P. pacificus* miRNAs.

The seed sequence of a miRNA is the major determinant of target specificity and represents the functional entity of a miRNA. Seed sequences are short (7 nt) and identical or almost identical seed sequences may have evolved through convergent evolution and are not conserved by descent. To investigate the impact of convergent evolution on miRNAs in our setting, I selected the most conserved arm of each miRNA as previously explained and classified these into 374 distinct groups based on perfect seed sequence identity (Figure 4.4B). Twenty-nine seed sequences (29/374, 7.8%) were conserved among all species representing 225 (31%) precursors. Thus, comparing the results of both classification procedures, i.e. full miRNA arm identity vs. seed identity,

revealed that five seed sequences are shared among all species exclusively using perfect seed similarity as classification criteria. These seed groups (1, 7, 12, 16, and 18) correspond to the following *C. elegans* precursors: *mir-34/-59/-228/-790/-791/-1820* (multiple alignments of precursors contained in seed groups are illustrated in Supplemental Figure B.1). Previous studies suggested that some of these genes are involved in developmental timing, embryogenesis, gonad migration, adult viability, and DNA damage response[284,285]. Interestingly, within all of these five seed groups the location of the seed sequence alternates between 5' and 3'. This suggests that some precursors most likely acquired a shared set of possible gene targets through convergent evolution.

4.5 Evaluation of microRNA Homology Assignment

To evaluate the performance of my miRNA homology assignment strategy, I investigated (i) the grouping of miRNAs into families using a test data set consisting of 52 miRNAs from the well-known *let-7* family from eight distinct animal clades (miRBase v20) and 50 randomly generated miRNAs; and (ii) the phylogenetic relationship of the multiple sequence-structure alignment on the same test data. As animals clades, human (*hsa*), chimpanzee (*ptr*), mouse (*mmu*), rat (*rno*), fruit fly (*dme*), nematode (*cel*), planarian (*sme*), and sea urchin (*spu*) were chosen (see miRBase database* for three-letter code information of species). Di-nucleotide shuffled miRNA sequences were generated based on the *let-7* family miRNAs using uShuffle[272] (Materials & Methods 2.3.5.4).

Since every miRNA gene has the potential to produce two distinct regulatory RNAs (Figure 2.4), I initially derived conservation levels for all 76 annotated miRNA 5' and 3' arms of the *let-7* family members for the eight animal clades and hundred 5' and 3' arms of the randomly generated precursor sequences. Then I selected the most conserved arm of each precursor (90 arms in total) and grouped all sequences into families. Essentially, all miRNAs that form a connected component in a graph, where miRNAs correspond to vertices, were assigned to a gene family (Figure 4.5). Vertices were connected by edges if the respective miRNAs were similar in sequence as defined by a valid pairwise alignment (see Materials & Methods 2.3.5.1 for detail). This graph clearly demonstrates that all *let-7* family members built one connected component (vertices presented in red), whereas all randomly generated miRNA sequences built singletons, i.e. vertices that are

*http://www.mirbase.org

not connected to any other vertex, and therefore gene families with only one member (vertices presented in green). This indicates that my strategy of grouping miRNAs into families distinguishes among true *let-7* family members and random precursor sequences because all precursors were classified into 51 distinct gene families.

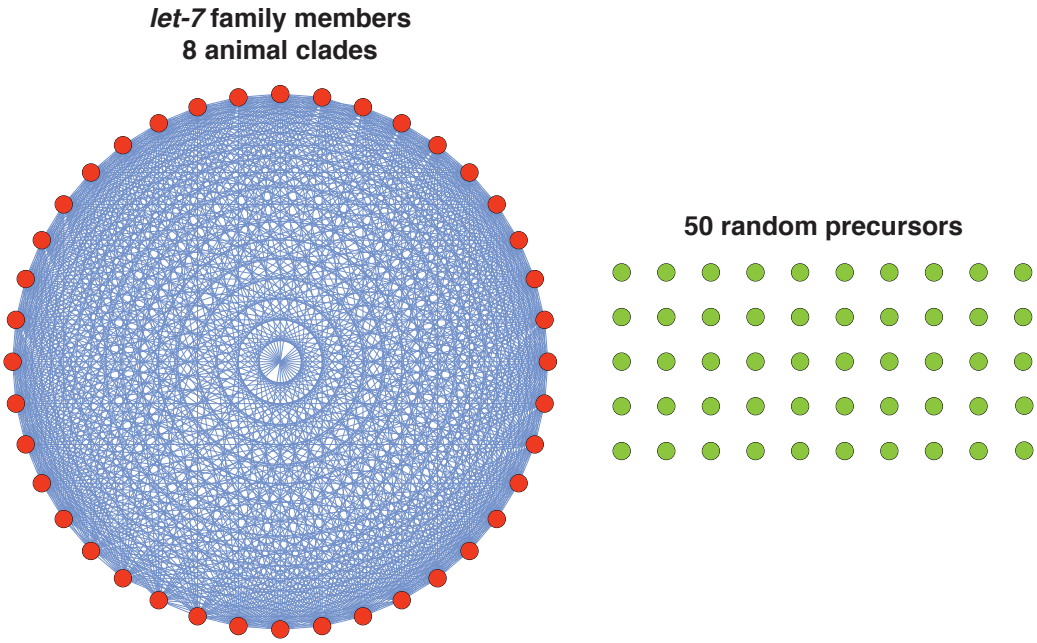

Figure 4.5: miRNA graph of 51 gene families
The most conserved arm of all miRNAs of the test data set were grouped into families as visualized by a graph with 51 connected components, where miRNAs correspond to vertices. Two miRNA vertices were connected by an edge if they were similar in sequence as defined by a valid pairwise alignment. This method clearly distinguishes between true *let-7* family members (vertices shown in red) and random precursor sequences (singleton vertices shown in green) by inferring 51 distinct miRNA families based on 90 miRNA arms.

To test the quality of computed multiple alignments and inferred phylogenetic relationships, I used the same test data as before. However, for visualization of the alignment and the phylogenetic tree, I selected five sequences from the randomly generated precursors, yet the results are comparable. The miRNA graph (Figure 4.5) illustrated that miRNAs from the test data were evidently grouped into 51 distinct gene families with all *let-7* family members being classified into the same gene family. However, to get an idea of the phylogenetic relationship of the *let-7* family members and five randomly generated precursors (*random-mir-1*, *random-mir-2*, etc.) for evaluation purposes of

my strategy, I computed a multiple alignment and inferred the phylogenetic tree (see Materials & Methods 2.3.5.2 and 2.3.5.3 for detail). The multiple alignment computed by LocARNA[262] clearly depicted and aligned the regulatory 5' arm of the *let-7* miRNAs indicated by a high consensus identity around the seed region and a high amount of nucleotides marked in blue (seed underlined in red; Supplemental Figure A.1). Moreover, the 3' arm, the arm that is usually not incorporated into RISC, displays a higher sequence similarity than the loop region (sequence between 5' and 3' arm), although generally less similarity than the 5' arm. Overall, the multiple sequence-structure alignment looks good with precursor sequences of the *let-7* family being aligned properly with a high sequence similarity as visualized by a large amount of nucleotides marked in blue and a generally high consensus identity in the precursor region. Notably, the five randomly shuffled precursor sequences are less similar and thus more divergent to all other *let-7* miRNAs. This is also reflected in the computed phylogenetic tree using UPGMA based on this MSA (Figure 4.6).

Overall, the phylogenetic tree illustrates that *let-7* miRNAs with the same lettered suffixes (e.g. *rno-let-7e*, *mmu-let-7e*, *ptr-let-7e*, and *hsa-let-7e*) are grouped together. The miRBase naming convention states that lettered suffixes denote closely related mature sequences*. Thus, this tree indicates that *let-7* miRNAs from distinct species with the same lettered suffix are more closely related to each other than *let-7* miRNAs from the same species, but having distinct lettered suffixes. In addition, *let-7* miRNAs from vertebrates are usually grouped together. However, two mouse miRNAs *mmu-let-7j* and *mmu-let-7k* are clustered in a group of planarian and fruit fly miRNAs and random generated precursors, respectively. Both mouse miRNAs are not listed as members of the *let-7* family on miRBase. The MSA (Supplemental Figure A.1) clearly illustrates that these sequence are more distinct from all other *let-7* precursor sequences indicated visually by a smaller amount of blue or light blue marked nucleotides.

In summary, grouping miRNAs into families and computing a multiple sequence alignment followed by the inference the phylogenetic relationships based on a test data set consisting of the well-known *let-7* family and randomly shuffled precursors sequences suggest that my strategy of miRNA homology assignment is reliable and robust method. However, the UPGMA method should only be employed for indicative purposes only and not as an estimate for phylogenetic time rates, since a constant rate of evolution can not be assumed.

*http://mirbase.org/help/nomenclature.shtml; accessed September 20, 2013

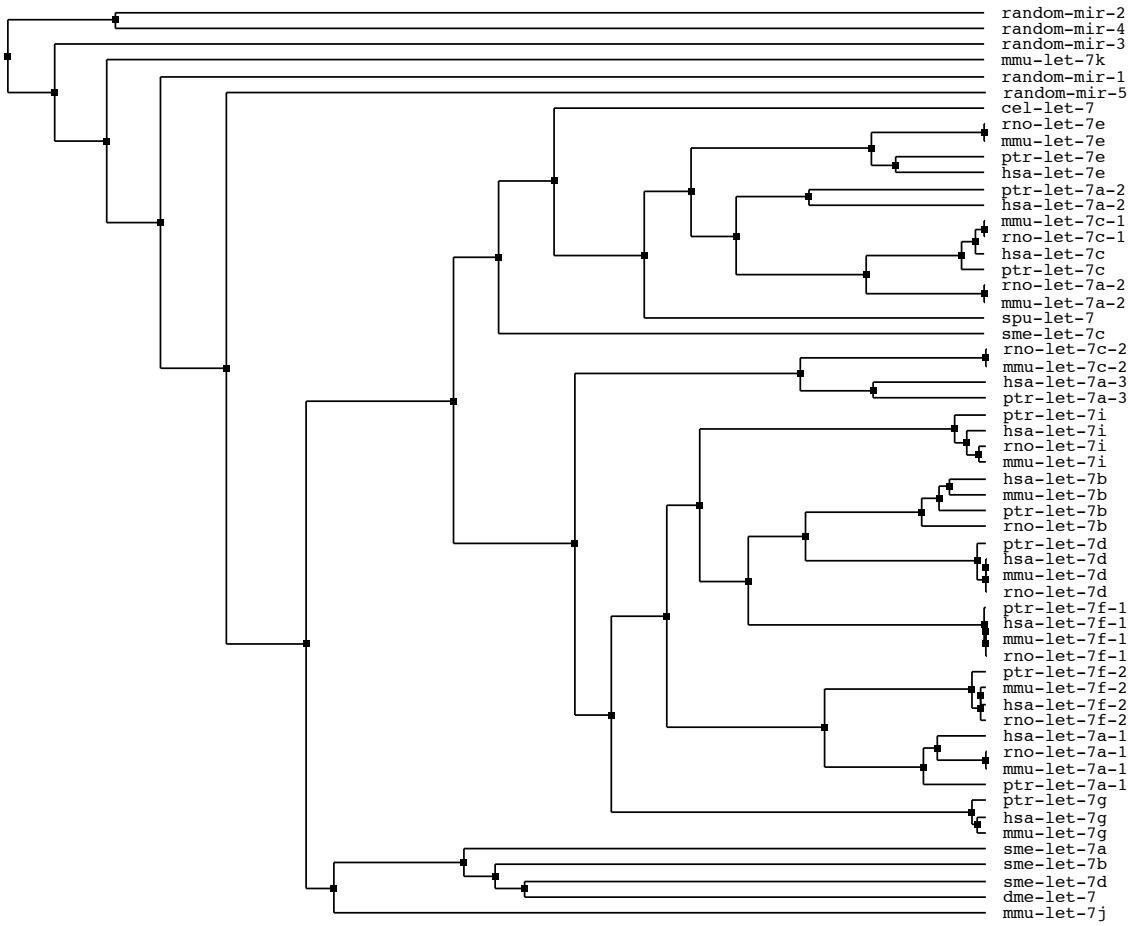

Figure 4.6: Phylogenetic tree of *let-7* family miRNAs and shuffled precursors
This tree was derived from a multiple alignment (Supplemental Figure A.1) computed on a test data set consisting of *let-7* family miRNAs from eight animal clades and five randomly generated precursor sequences. The tree was inferred using the UPGMA clustering method.

4.6 microRNA Expression Changes from Sequencing Data Agree with Published qRT-PCR Results

With the updated miRNA gene set at hand, I performed a stage-wise comparison of miRNA expression levels across species to identify miRNA genes that are not only conserved in sequence but also in expression pattern. To this end, expression changes of miRNAs in dauer/iL3 relative to mixed-stage samples in *C. elegans*, *P. pacificus*, and *S. ratti* were measured (data sets 3, 4, and 7-10; Table 4.1). As a quality control,

I directly compared this data to previously published miRNA expression changes in *C. elegans* measured by qRT-PCR[125].

To quantify miRNA expression changes, I normalized miRNA library read counts by reference based qq normalization[154], where mixed-stage libraries were selected as the reference. Expression changes could be estimated for 177 (69%) miRNAs between normalized dauer and mixed-stage read counts in *C. elegans* (Supplemental Table B.6). As a result, 71 (40%) *C. elegans* miRNAs were detected that exhibited differences in expression in the developmentally arrested stage compared with mixed-stage samples. Most miRNAs were downregulated (53, or 30%), whereas 18 (10%) miRNAs showed a relative increase in dauer expression (Table 4.2).

Table 4.2: Significantly upregulated miRNAs in *C. elegans* dauer larvae ($FDR < 0.05$)
Seed conservation column displays IDs of miRNAs that share a common seed within the phyla Nematoda. If the seed is not conserved in Nematoda miRNA IDs from the subsequent phyla are displayed (in the order of Nematoda, Arthropoda, Lochotrophozoa, and Vertebrata). Note: miRNA family ID is assigned if a miRNA is conserved at least once within *C. elegans*, *P. pacificus*, or *S. ratti*.

Rank	miRNA gene	miRNA family ID	Seed	Seed conservation[†]	Seed conservation profile[†]	Log_2 fold changes	Observed function
1	miR-797	miRNA family 50	AUCACAG	*mir-2/-43/-250*	+ / + / + / -	4.44	Gonad migration [285]
2	miR-4809	miRNA family 37	UAAGUUC	*mir-1018/-4809/-4810*	- / - / - / -	3.52	-
3	miR-2210	-	GGCAGAU	*mir-72*	+ / - / - / +	3.34	-
4	miR-1824	-	GGCAGUG	*mir-34*	+ / + / + / +	3.16	DNA damage response [284]
5	miR-4807	-	UGAGUUC	*mir-983*	- / + / - / -	2.76	-
6	miR-248	-	UACACGU	*mir-248*	+ / - / - / -	2.75	-
7	novel-miR-V_24974	-	GGCUCAA	-	- / - / - / -	2.19	-
8	novel-miR-I_285	-	GCGGGAC	-	- / - / - / -	2.10	-
9	miR-247	miRNA family 26	GACUAGA	*mir-44/-61/-247/-279*	+ / + / + / -	1.94	Gonad migration [285]
10	miR-34	miRNA family 31	GGCAGUG	*mir-34/-1824/-2227/-2239/-4933*	+ / + / + / +	1.89	DNA damage response [284]
11	miR-1	miRNA family 4	GGAAUGU	*mir-1/-796*	+ / + / + / +	1.85	Synaptic transmission [286]
12	miR-1820	-	UUUGAUU	*mir-315*	+ / + / + / +	1.59	-
13	miR-791	-	UUGGCAC	*mir-791*	+ / + / + / +	1.56	-
14	miR-54	miRNA family 40	ACCCGUA	*mir-51/-52/-53/-54/-55/-56/-2233/-2237/-2271/-2274*	+ / + / + / +	1.35	Embryogenesis, pharynx attachment, developmental timing [285,287,288]
15	miR-254	-	GCAAAUC	*mir-254*	+ / - / - / -	1.32	-
16	miR-71	miRNA family 30	AUCACUA	*mir-34/-71/-2953*	+ / - / - / +	1.26	Lifespan, AWC L/R neuron fate specification [123,289,290]
17	miR-84	miRNA family 1	GAGGUAG	*let-7, mir-48/-241/-795*	+ / + / + / +	1.23	Developmental timing, vulval cell fate specification [65,291–295]
18	miR-794	miRNA family 1	GAGGUAA	-	- / + / - / -	1.14	-

[†]Nematoda/Arthropoda/Lophotrochozoa/Vertebrata

Karp *et al.* monitored life history related expression level changes for 107 miRNAs in *C. elegans* using qRT-PCR. For a direct comparison of our dauer vs. mixed-stage expression change data with their dauer vs. L2m (late L2 - mid-L3) expression changes, I discretized miRNA expression changes into three categories: (i) upregulated, (ii) downregulated, and (iii) unaffected. This comparison was performed for 93 miRNA genes. Thirteen miRNAs were not measured in the qRT-PCR experiment (L2m or dauer), and miR-798 was not detected in our small RNA-seq data. Both methods indicate a good agreement of expression change classes (Figure 4.7A; $P = 1.1 \times 10^{-5}$, χ^2 test). However, 34% of miRNAs were classified into different expression categories. In particular, a few individual miRNAs were downregulated in dauer in our small RNA-seq data but unaffected in the qRT-PCR data, including members of the co-transcribed *mir-35-41* cluster and *mir-246*, which are known to be specifically enriched in *C. elegans* embryos[67,285,296] (*mir-41* was exclusively detected in our deep sequencing experiment). Such discrepancies could be explained by the developmental specific expression of these miRNAs, since I compared dauer with mixed-stages samples instead of L2m. Figure 4.7B depicts the small RNA-seq $\log_2$ ratios plotted against $-\Delta\Delta C_T$ values of the qRT-PCR experiment from Karp *et al.* (2011). Only three miRNAs (*mir-34/-71/-248*) are reported as upregulated in dauer relative to L2m. I observed the same expression pattern for all of these genes in my *C. elegans* dauer to mixed-stage comparison. Four genes, *mir-230/-241/-788/-795*, are consistently downregulated in both studies. Note that all miRNAs in the upper left quadrant that appear to be upregulated in dauer in the qRT-PCR experiment were classified as unaffected due to a non-significant *t*-test or inability to reproduce results by Karp and colleagues[125].

Whereas Karp *et al.* chose a targeted approach to measure expression changes of 107 selected miRNA genes, I was able to detect an unrestricted set of differentially expressed miRNAs because an unbiased strategy was applied in this study. As a result, I detected an additional set of 35 differentially expressed miRNAs that were not monitored in the qRT-PCR experiment. Eight of those were upregulated in dauer, including *mir-1820* and *mir-1824*, which seeds are identical to the highly conserved *mir-315* and *mir-34* family, respectively; 27 were downregulated, including *lsy-6*. Moreover, two of the novel miRNA candidates (*cel-mir-8193* and *cel-mir-8200*) were upregulated in dauer and three novel miRNAs (*cel-mir-8191*, *cel-mir-8208*, and *cel-mir-8190*) downregulated.

Overall, this data is in good agreement with reported qRT-PCR fold changes from Karp and colleagues. The observed discrepancies could be explained by: i) differences in experimental design (dauer/mixed-stages and dauer/L2m), ii) differences in assay

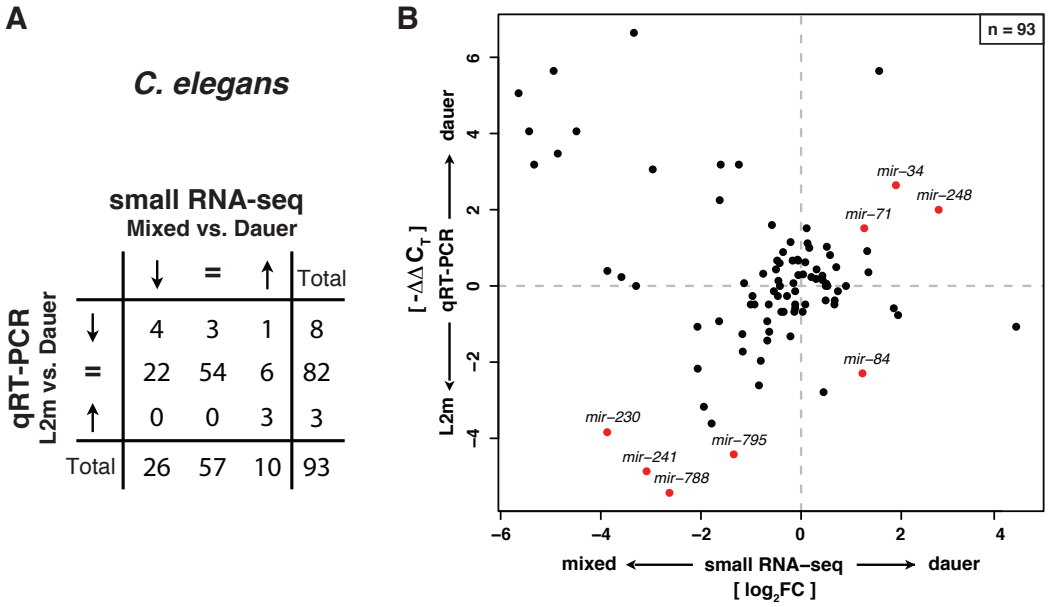

Figure 4.7: Small RNA-seq expression profiles in *C. elegans* agree with qRT-PCR data
(A) Contingency table of expression fold changes of *C. elegans* dauer vs. mixed-stage obtained by Illumina small RNA deep sequencing compared with qRT-PCR data of dauer vs. L2m from Karp *et al.* (2011) classified according to three categories (upregulated, downregulated, and unaffected). Expression fold changes of both data sets are significantly correlated ($P = 1.1 \times 10^{-5}$, χ^2 test). **(B)** Quantitative comparison of expression fold changes obtained by small RNA-seq and qRT-PCR experiments in *C. elegans*. Names of all miRNAs with a significant expression change of at least 2-fold in both experiments are displayed. Significance of differential miRNA levels in small RNA-seq data between mixed-stage and dauer/iL3 was determined by a two-sided binomial test constrained on the total library sizes followed by correction for multiple testing (FDR <0.05).

biases[297,298], and iii) asynchronous sampling across experiments[125].

4.7 Differential Expression Analysis Identifies Cross-Species Candidate Regulators

To begin to understand if the set of deeply conserved miRNAs may control aspects of developmental arrest in free-living and parasitic nematodes, I examined relative expression changes of developmentally arrested stages (dauer/iL3) to mixed-stage populations in the necromenic nematode *P. pacificus* and the parasite *S. ratti* (Supplemental Ta-

ble B.6). Significant changes in expression levels were observed for 40% (71/177) of miRNA genes in *C. elegans*, 60% (198/331) in *P. pacificus*, and 35% (37/106) in *S. ratti* (Figure 4.8). The majority of miRNAs that were differentially expressed in *P. pacificus* and *S. ratti* demonstrated an increase in expression in developmentally arrested stages (113/331 [34.1%] and 21/106 [19.8%], respectively). In contrast, most miRNAs in *C. elegans* for which we observed expression changes were downregulated in *C. elegans* dauer larvae (53/177, 29.9%) (18 miRNAs upregulated). Furthermore, one-quarter of *P. pacificus* miRNAs and 15% of *S. ratti* miRNAs were detected to be downregulated.

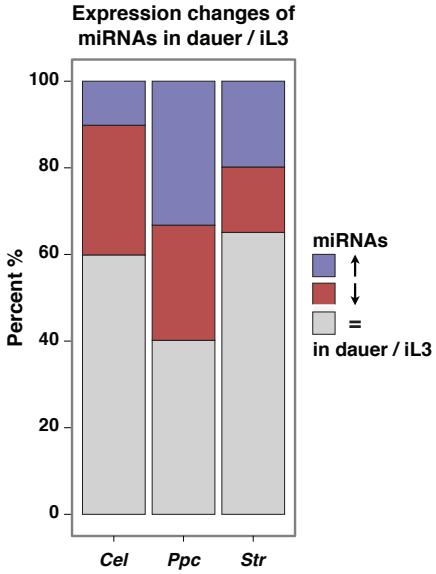

Figure 4.8: Proportion of expression changes between developmentally arrested stages
Significant expression changes were observed for 40% of miRNAs in *C. elegans*, 60% in *P. pacificus*, and 35% in *S. ratti*. Reads were normalized by reference based qq normalization[154]. Log_2 fold changes were computed between mixed-stage and dauer/iL3 samples. All miRNA genes with absolute fold change >1 and two-sided binomial test with p-value cutoff corresponding to FDR <0.05 were defined as differentially expressed ($\log_2$ fold change >1: upregulated (↑) and $\log_2$ fold change < -1: downregulated (↓) in dauer/iL3).

I wanted to address the long-standing hypothesis that dauer and infective larvae share a common origin. Our data demonstrated that 190 (26%) miRNAs are shared among *C. elegans* (64/257, 24.5%), *P. pacificus* (89/362, 24.6%), and *S. ratti* (37/106, 34.9%) based on sequence similarity of the miRNA 5' or 3' arm, respectively. If dauer and infective larvae have a common origin, I would expect to find conserved miRNAs in

dauer and iL3 that show a coherent expression signature. I tested this hypothesis by constructing seed-constrained multiple sequence alignments for each individual miRNA family as defined by sequence identity (Supplemental Figure B.2). These alignments were used to infer phylogenetic trees. To investigate expression signatures of miRNA families and detect possible conserved expression signatures, I combined this phylogenetic information with the derived miRNA $\log_2$ expression fold changes (Supplemental Figure B.3).

Overall, four miRNA gene families with homologs in all three species show a coherent expression pattern (i.e. at least one family member from each species is differentially expressed as the majority of family members): two families are upregulated (the *mir-1* and *mir-71* families) and two families are downregulated (the *mir-240* and *mir-35* families; see Supplemental Table A.1).

In the following, I will focus on *mir-71* and *mir-34*, two miRNA candidates that were upregulated in our small RNA-seq data and also in the published qRT-PCR data. The *mir-71* family is conserved across all three species and shows a coherent expression pattern whereas the *mir-34* family could represent a case of convergent evolution in *P. pacificus*.

The *mir-71* family includes one *S. ratti* gene (*mir-71*) and two genes in *C. elegans* (*mir-71/-2953*) and *P. pacificus* (*mir-71/-71b*) (Figure 4.9). Interestingly, the majority of miRNA genes of the *mir-71* family were increased in expression in dauer and iL3. This conserved expression signature indicates the potential importance of *mir-71* family members for developmentally arrested stages in free-living and parasitic nematodes.

Investigating the *mir-34* family revealed that this family contains one miRNA gene from *C. elegans* (*mir-34*) and three *S. ratti* genes that are clustered on the genome (located within 10Kb of distance on the same contig) demonstrating an expansion of the *mir-34* miRNA repertoire in *S. ratti* (Figure 4.10). Strikingly, I did not detect a *mir-34* precursor in *P. pacificus*, despite *mir-34* being highly conserved from various nematodes to vertebrates including humans (Figure 4.11). For the identified family members, I observed a conserved expression signature: All *mir-34* family miRNAs are upregulated in dauer and iL3, suggesting that they may be important for developmental arrest in free-living and parasitic nematodes. It is rather unlikely that *mir-34* was not profiled in our data in *P. pacificus* given the high sequencing depth. Therefore, I conclude that *mir-34* was lost in the lineage leading to *P. pacificus*.

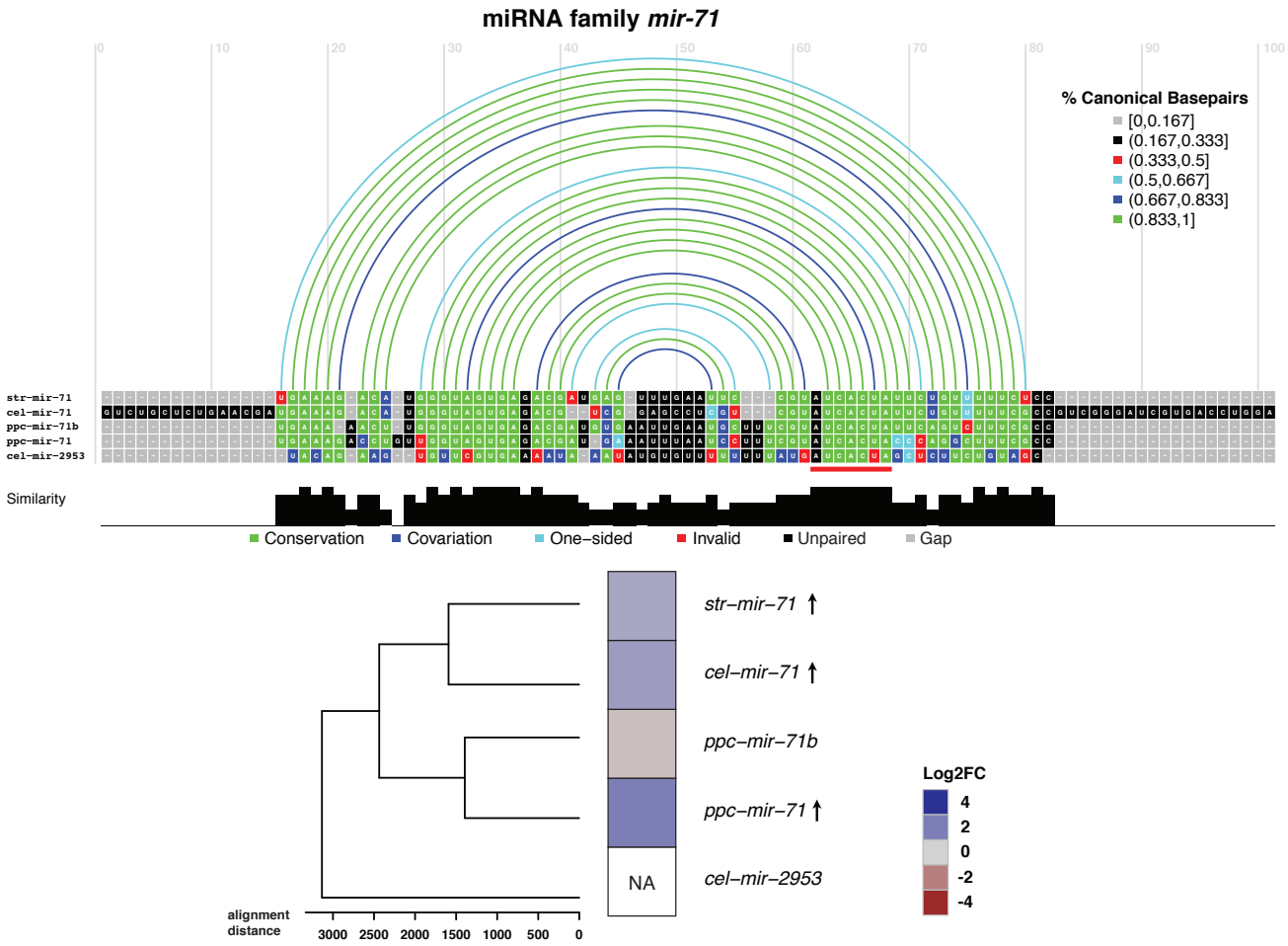

Figure 4.9: *mir-71* family miRNAs as cross-species candidate regulators in developmental arrest
Multiple sequence alignment for miRNA family *mir-71* computed by LocARNA[262]. The multiple alignment was constrained to align at the seed sequence position of each individual miRNA. The seed (position 2-8) of miR-71 is marked with a red line. Arcs above the alignment represent secondary structure information. Arc colors encode the fraction of canonical paired bases. Alignment colors are annotated according to their agreement with the predicted secondary structure. Nucleotides that are base-paired according to the structure are colored in green and unpaired bases in red. If mutations have occurred but basepairing potential is preserved, nucleotides are displayed in blue (dark blue for mutations in both bases and light blue for single-sided mutations). Unpaired nucleotides are colored in black and gaps in grey. The heatmap represents miRNA gene expression by color where heatmap rows are ordered by the inferred phylogeny from the alignment. Arrows next to the miRNA in the heatmap plot denote significantly up- (↑) or downregulation (↓) in dauer/iL3.

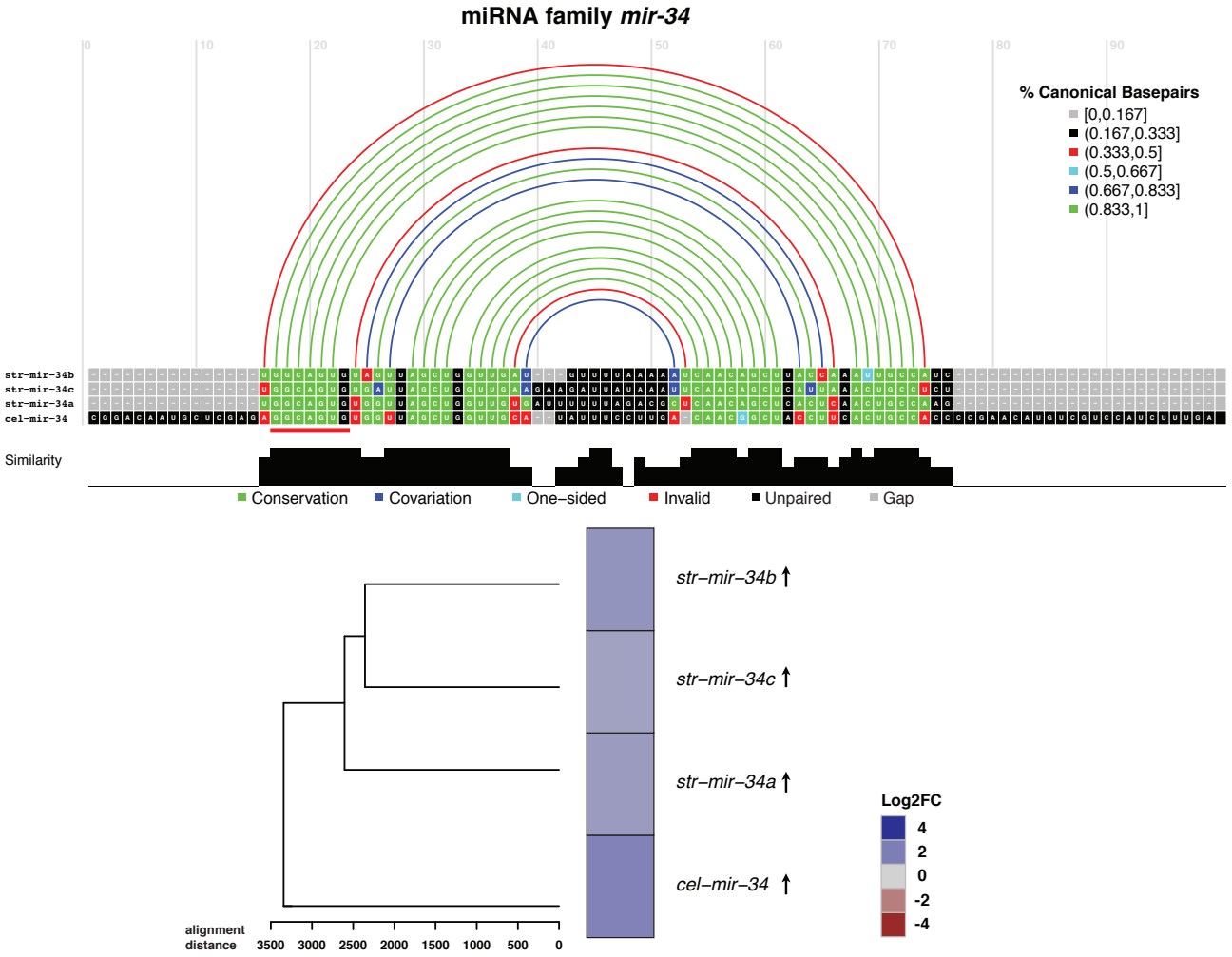

Figure 4.10: *mir-34* family miRNAs as cross-species candidate regulators in developmental arrest
Multiple sequence alignment for miRNA family *mir-34* was computed and visualized as described in Figure 4.9. As before, the heatmap represents miRNA gene expression by color where heatmap rows are ordered by the inferred phylogeny from the alignment. Arrows next to the miRNA in the heatmap plot denote significantly up- (↑) or downregulation (↓) in dauer/iL3.

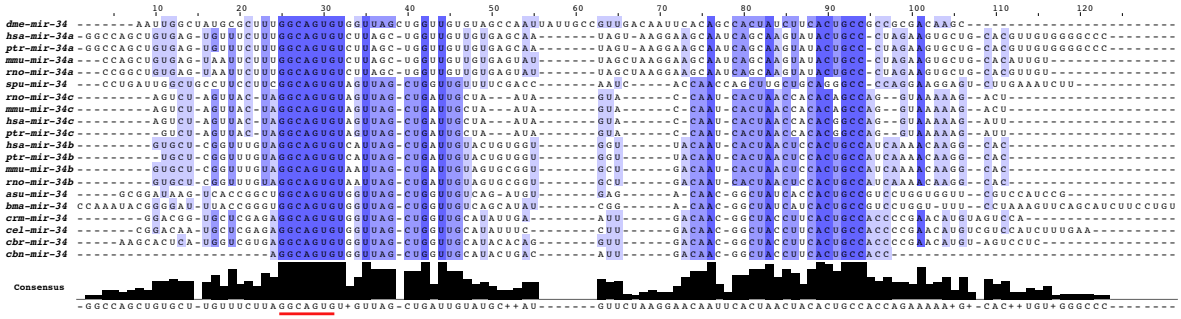

Figure 4.11: MSA of *mir-34* family miRNAs from seven animal clades
The multiple alignment was computed by LocARNA [262] and constrained to align at the seed (defined as nucleotide position 2-8) of each individual miRNA. As animal clades, human (*hsa*), chimpanzee (*ptr*), mouse (*mmu*), rat (*rno*), fruit fly (*dme*), nematode (*cel, cbr, crm, cbn, bma, asu*), and sea urchin were chosen (*spu*) [see miRBase database† for three-letter code information of species]. Nucleotides are colored based on a percentage identity threshold; i.e. nucleotides that occur in a particular column more than 80% are colored in mid blue, more than 60% in light blue, more than 40% in light grey, and white otherwise. The seed sequence is marked with a red underline.

Interestingly, my seed conservation analysis detected two miRNA genes in *P. pacificus* (*mir-2239-1/-2*) that give rise to miRNA arms with seed sequences identical to the miR-34 seed 'GGCAGUG'. Both miRNAs could potentially regulate similar target sets (seed group 58; Supplemental Figure B.1). However, these miRNA genes did not show an upregulation in dauer larvae (Supplemental Table B.5).

Moreover, I considered additional miRNA candidates in *P. pacificus* that could compensate for the 'loss' of *mir-34*. For this, I identified likely candidates by collecting miRNA genes whose seed sequence differs by one nucleotide from the miR-34 seed 'GGCAGUG' and examined their expression in dauer larvae.

4.8 *P. pacificus* miR-34 Seed Neighbors are Upregulated in Dauer Larvae

To assign functional conservation by 7-nt seed sequence identity is a conservative strategy. Bartel (2009) discusses different modes of canonical target recognition: 7mer-A1 sites, 7mer-m8 sites (our seed classification), and 8mer sites. Other miRNA genes might exist that regulate similar target sets like *mir-34* but have been missed by the stringent seed classification method I have chosen. To overcome this problem, I examined seed

changes of miRNA arms annotated in all three species. In order to do so in a systematic way, I generated a network in which nodes represent seed sequences. Nodes are connected if the corresponding seed sequences differ by one nucleotide. The resulting network consists of 742 nodes connected with 837 edges and 200 singletons (seeds not connected to any other seed). It contains 71 connected components with a maximum of 534 seeds (57%) in the largest component (Figure 4.12A and B).

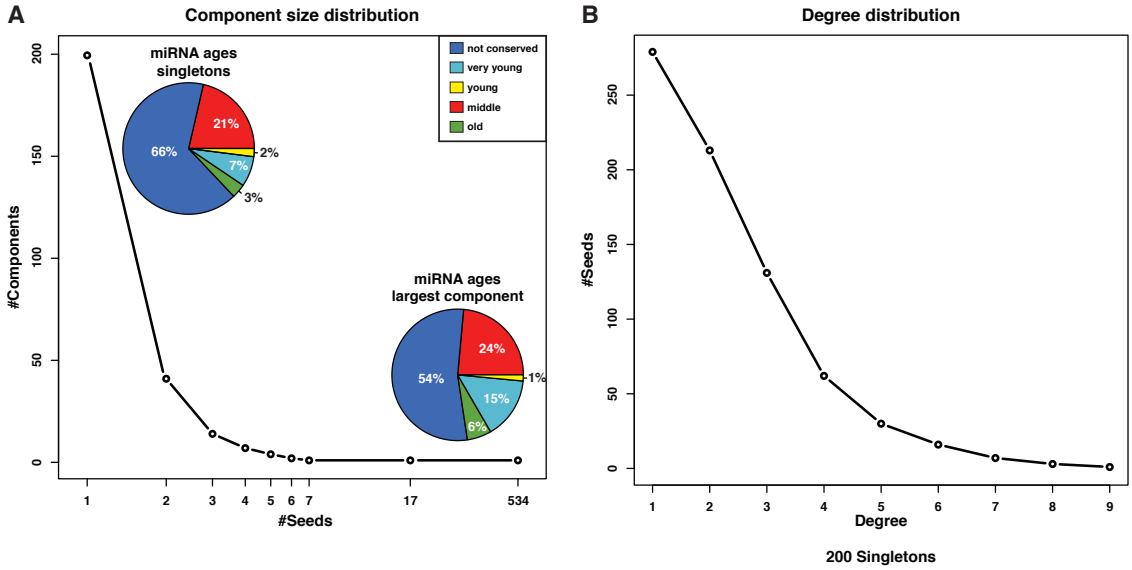

Figure 4.12: Properties of single-mutation seed network
(A) Distribution of component sizes with a maximum of 534 seeds being contained in the largest component. Pie charts illustrate miRNA age distributions for singletons and largest component. **(B)** Degree distribution of single-mutation network. It depicts to how many seeds a single seed is connected. One seed 'AUGACAG' that originate from seven *P. pacificus* miRNAs was at most connected to nine other seeds.

To identify substitutes for *mir-34* with an identical expression pattern, I examined the neighborhood of the *mir-34* family in the seed network (Figure 4.13). Three out of the four neighbors of the conserved miR-34 seed node ('GGCAGUG') originate from *P. pacificus* miRNA genes. All of those genes were upregulated in dauer. One seed neighbor ('GCCAGUG') that did not change its expression in developmental arrest originated from a miRNA in *S. ratti*. However, none of the three identified *P. pacificus* seed neighbors bear any sequence resemblance to the *mir-34* family. This finding strengthens our hypothesis of *mir-34* being lost in the lineage to *P. pacificus*. Three miRNA genes with conserved expression yet distinct seed sequences could act compen-

satory in the context of dauer development.

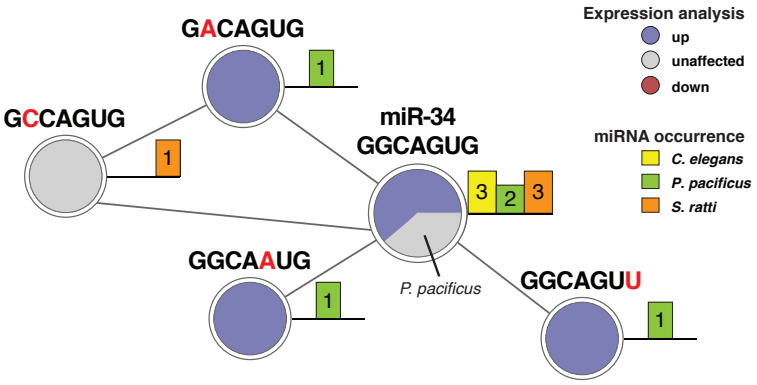

Figure 4.13: Expression conservation of miR-34 seed neighbors
The neighborhood of miRNA seeds was analyzed regarding expression changes in dauer and infective larvae. The neighborhood sub-network of the miR-34 seed reveals conserved upregulation of all *P. pacificus* seed neighbors. Node color represents expression changes classified into upregulated (blue), downregulated (red), and unaffected (light grey). Barplot next to a node represents the number of times a miRNA seed was identified in a specific nematode. Note that *mir-4933* of *C. elegans* was not measured in our data and is represented in the light grey pie.

Chapter 5

Discussion

In this work I present a systematic analysis of miRNA genes in free-living and parasitic nematodes using a multiplatform sequencing approach (ABI SOLiD, Illumina GA II, and HiSeq). The goal of this project was to analyze whether miRNA genes impact developmental arrest and long-term survival in dauer and dauer-like stages, i.e. the infective stage of parasites, and to address the long-standing hypothesis that dauer and infective larvae share a common origin. This investigation was specifically focused on determining whether shared 'dauer-infective' miRNA expression signatures exist. To this end, I developed a bioinformatics workflow that involves the following six distinct computational steps: (i) preprocessing (quality filtering, barcode detection, and adapter removal) of small RNA-seq data produced by NGS (Illumina and SOLiD), (ii) mapping to a reference genome, (iii) identification of known and novel miRNA genes in nematodes (*C. elegans*, *P. pacificus*, and *S. ratti*), (iv) identification of differentially expressed miRNAs in developmentally arrested stages, (v) inference of miRNA gene families and their phylogenetic relationships, and (vi) integration of observed phylogenetic relationships with expression level changes. This study identifies and extends miRNA gene sets in *C. elegans* and *P. pacificus* and reports the first coherent data on any *Strongyloides* parasite. The inference of miRNA families by sequence similarity revealed that miRNA gene sets diverge rapidly in nematodes. However, a small core set of conserved miRNA families exists, and some families even show conserved expression patterns. The comparison of miRNAs expressed in dauer and infective stages yielded candidate miRNAs that might serve as conserved post-transcriptional regulators of the dauer and infective larvae fate, supporting the hypothesis that dauer formation and parasitic life style share the same origin. Notably, the single-mutation seed network

of all miRNAs revealed that convergent evolution of seed sequences has taken place. This work constitutes a valuable resource to researchers studying miRNA evolution in general and in particular, aspects in developmental arrested in nematodes.

In the first part of this work I presented the bioinformatics methods and the computational strategies that I applied in order to accomplish the computational steps outlined above. Moreover, I was involved in the implementation of FLEXBAR, the flexible barcode and adapter removal tool, which I introduced in Chapter 3. As part of this project, I developed the adapter removal feature for SOLiD color space data and focused on the application of small RNA-seq in letter and color space. Additionally, I was involved in the design of the original program FAR and in the development of specific features of the subsequently added barcode detection function for demultiplexing.

5.1 FLEXBAR - Leading Solution in Barcode and Adapter Processing

FLEXBAR is a versatile solution for three critical preprocessing steps in any next-generation processing pipeline: (i) basic clipping and quality filtering, (ii) barcode recognition and processing, and (iii) adapter recognition and removal. Importantly, all of these steps can be performed in one program call and executed in parallel. FLEXBAR covers a larger range of sequencing platform applications, formats, and features than other tested solutions. Furthermore, it provides detailed output statistics and, if desired, extensive verbose output, such as graphical output of read alignments.

FLEXBAR performed slightly better than FASTX, which is widely considered to be the best of all (selected) competitors in removing adapters from an Illumina short read data set (benchmark I), as measured by the number of uniquely mappable reads and bases [for details on benchmark I-IV see Dodt *et al.* (2012)]. While consuming only slightly more runtime on one processor core than FASTX, FLEXBAR scales favorably when using multiple threads. As pinpointed by a paired-end RNA-seq example application (benchmark II), FLEXBAR handles four processing steps in one program call and requires almost 50% less runtime than FASTX. Of course, FLEXBAR preserves read pairings in all output files. Note that these benchmarks were computed using an older version of FLEXBAR (v2). The computation time of the current version 2.4 has been increased significantly due to an updated version of the Seqan library[274]. In benchmark III, we demonstrated how faithfully our software recognizes barcodes

and avoids false assignments. In addition, we could show that FLEXBAR is also useful for unconventional applications, such as identifying trans-splicing events in a color space transcriptome data set from *C. elegans* (benchmark IV). Finally, FLEXBAR is the leading solution compared to CUTADAPT when removing adapter sequences from SOLiD short color space reads (benchmark V; Figure 3.4).

In summary, FLEXBAR has been successfully applied in various genomics applications, such as small RNA-seq[2], PAR-CLIP[299,300], and poly-A tail detection in 3' UTR data (benchmark II and unpublished data). Moreover, to make FLEXBAR accessible to a wider community, e.g. biologists, it has been integrated into GALAXY[301–303]. With FLEXBAR at hand preprocessing of small RNA-seq data investigated in this study could be performed easily.

In the second part of my work I applied the bioinformatics workflow presented in the first part to address the question whether miRNAs may impact developmental arrest and long term survival in dauer and dauer-like stages. In particular, I was examining whether shared 'dauer-infective' miRNA expression signatures exist that may support the long-standing hypothesis that dauer and infective larvae share a common origin.

5.2 Comprehensive Bioinformatic Analysis Identifies Cross-Species Candidate Regulators in Nematodes

In harsh conditions, such as low food supply and stress, many nematodes are able to form dauer larvae, a developmentally arrested, stress-resistant, and long-lived state[20,22]. Infective larvae of parasitic nematodes share many morphological, behavioral, and physiological traits with dauer larvae of free-living nematodes[37–39]. Accordingly, dauer larvae have been suggested as an evolutionary precursor of infective larvae that facilitated the repeated evolution of parasitism (a pre-adaptation)[278]. We hypothesized that regulatory modules exist that play similar roles in regulating environmentally triggered alternative life styles across distantly related species in the nematode phylum. Consistent with this hypothesis, the results from this study demonstrate that conserved 'dauer-infective' miRNA expression signatures are present.

miRNA genes have been associated with signaling pathways that regulate the dauer fate[121,122,126] . However, a systematic assessment of the roles of miRNA genes as post-transcriptional regulators in dauer fate decisions and their conservation in parasitic

nematodes has been elusive. The miRNA gene discovery and expression profiling in developmentally arrested stages of three representative species (*C. elegans*, *P. pacificus*, and *S. ratti*) presented in this work reveal substantial regulation of miRNA genes in developmental arrest in free-living and parasitic nematodes. Moreover, my analyses of these data sets provide several implications:

First, my study of expression changes in *C. elegans* demonstrates that 71 miRNAs exhibit expression differences in dauer compared to non-dauer stages. A subset of 18 miRNAs is significantly upregulated in this comparison (Table 4.2). As a quality control, I intersected this data with recently published qRT-PCR data[125] (comparison of dauer and L2 larvae) and identified miR-34/-71/-248 to be upregulated in both data sets. So far, miR-34 and miR-71 have been assigned roles in longevity and stress response[120,123,124,284,289,304]. Overall, the data profiled in this study is in good agreement with reported qRT-PCR expression level changes from Karp and colleagues (2011). However, several miRNAs were classified as downregulated in dauer in my DE analysis but reported to be unaffected in the qRT-PCR experiments, e.g. members of the co-transcribed *mir-35-41* cluster and *mir-246* (note that *mir-41* was exclusively detected in the small RNA-seq data). This discrepancy could be explained by differences in experimental design (dauer/mixed-stages and dauer/L2m), since these miRNAs are known to be specifically enriched in *C. elegans* embryos[67,285]. Furthermore, differences in assay biases[297,298] and asynchronous sampling across experiments[125] could also explain observed variations. Nevertheless, by using an undirected sequencing approach I was able to identify a number of differentially expressed miRNAs (8 up- and 27 downregulated in dauer) that were not reported in the qRT-PCR experiment.

Second, I presented the first profiling and comparison of miRNA genes in nematodes from different life styles with emphasis on developmental arrested stages. Notably, we were the first to profile miRNAs in any *Strongyloides* parasite. The reported miRNA gene complements in *C. elegans*, *P. pacificus*, and *S. ratti* presented in this study are likely to be complete due to the high recovery rate of known miRNA genes in *C. elegans* (87%) and *P. pacificus* (99%) by our multiplatform sequencing strategy. Furthermore, the size of predicted miRNA gene sets correlates well with the species genome sizes (Figure 4.3B). Additionally, our NGS approach is comprehensive enough to identify tissue and stage-specific miRNAs, such as *lsy-6*, a very rare miRNA, which is only expressed in less than 10 cells[279] and is hardly detected by qRT-PCR[125]. It is frequently assumed that the mature miRNA (guide strand), the sequences that is loaded into RISC, is more abundant in sequencing data than the star sequence (passenger strand).

However, growing evidence suggests that both arms produced from a miRNA hairpin may be biologically functional[98–104] and that the dominate strand can vary in a cell-context and tissue-dependent fashion or between orthologous miRNAs[105–110]. In line with this, I observed miRNA sequences that are produced from both strands at similar frequencies. Moreover, I noticed large variations in 5' arm to 3' arm read count ratios depending on the sequencing platform employed (Supplemental Table B.4). In essence, I see a strong platform dependency of read count patterns across miRNA arms and therefore refrain from assigning mature and star sequences. Instead, I assigned names of the form ppc-miR-71-5p and ppc-miR-71-3p for sequences derived from the 5' and 3' arm, respectively. Thus, I reannotated, extended, and defined all miRNAs following this new nomenclature. This is in agreement with the revised naming guidelines described in miRBase. Kozomara and Griffiths-Jones announced in their latest publication (2014) that they recently started to replace the old nomenclature (i.e. ppc-miR-71/ppc-miR-71*) with this new one[75].

Finally, by examining sequence identity of miRNAs among free-living and parasitic nematodes, I identified a small core set of 24 miRNA families that are conserved among all three species. Importantly, despite rapid miRNA evolution in nematodes, homologous gene families with conserved 'dauer-infective' expression signatures are present. In particular, I find two miRNA gene families with homologs in all three species that demonstrate coherent upregulation and two families with coherent downregulation in developmental arrest (Supplemental Table A.1). Consistent with qRT-PCR data, I detected three miRNA genes (miR-34/-71/-248) to be upregulated in *C. elegans* dauer. While I did not detect any miRNA in *P. pacificus* or *S. ratti* that is homologous to *mir-248*, I found *mir-34* to be conserved in *S. ratti* and *mir-71* in both species. Although *mir-34* is not conserved in *P. pacificus*, the same seed sequence ('GGCAGUG'; position 2-8) is found in two apparently non-conserved *P. pacificus* miRNAs: miR-2239-1 and miR-2239-2. Both miRNAs are non-differential in the dauer fate. A careful inspection of the single-mutation seed network uncovered expression conservation of all *P. pacificus* miR-34 seed neighbors (i.e. upregulation in the dauer fate), providing evidence for convergent gene evolution (Figure 4.13).

This work is based on the development and application of bioinformatics methods to analyze digital gene expression data which were profiled using a multiplatform NGS approach. Thus, one of the challenges was to integrate data sets from Illumina and SOLiD small RNA sequencing. While the Illumina system had been widely applied in miRNA-profiling studies when I started this work, to my knowledge, only a few re-

search groups employed the SOLiD system[109,154]. In line with this, software solutions that could handle color space data were limited, e.g. software for data preprocessing, mappers, or miRNA prediction tools. Notably, direct translation of color space into letter space is error prone due to the characteristics of the di-base dependent sequencing strategy[134]. Hence, I had to develop and implement strategies, e.g. the adapter removal functionality of FLEXBAR, to solve the bioinformatics problems stated above. Nevertheless, several studies that applied the SOLiD system for miRNA profiling have been published recently[146,155–157].

By integrating phylogenetic information with gene expression profiles of miRNAs, I was able to identify conserved miRNA expression signatures between free-living nematodes and parasites. Here, the inference of miRNA gene families was a very important analysis step. A couple of strategies to derive miRNA families exist: (i) the Rfam database generates families of various ncRNA classes using covariance models[169], (ii) oftentimes miRNA conservation is defined simply based on seed similarity[108,305,306], which in result does not distinguish between homology and functional conservation (i.e. convergence), and (iii) miRBase[75], the main repository for miRNAs, provides gene families of miRBase deposited miRNAs defined by conservation across precursors and manual curation, which makes it impossible to recreate these families using a set of rules (Sam Griffith-Jones - pers. comm.). Novel miRNAs, as discovered in this study, required regrouping of known families and/or definition of novel families. Since, no consistent rule exists of how to categorize miRNAs into families, I developed a novel method for an automated inference of phylogenetic relationships among miRNAs. This method is able to differentiate between homology and convergent evolution as demonstrated and visualized with a single-mutation seed network of the *mir-34* family (Figure 4.13). Note that an all-against-all blast[193] approach on precursors as applied by Meunier *et al.* (2013) did not provide satisfying results, because *C. elegans*, *P. pacificus*, and *S. ratti* are distantly related species. In fact, *Caenorhabditis elegans* and *P. pacificus* diverged 280-430 million years ago and *S. ratti* even belongs to a distinct nematode class (Figure 1.1)[4,28]. Overall, my analysis of conserved miRNA expression signatures provides interesting strategies of how to integrate expression profiles with phylogenetic information.

Our multiplatform deep sequencing approach is comprehensive enough to identify known and novel miRNAs genes. It is rational to think that our data is reliable due to several reasons. First, our data is in good agreement with qRT-PCR data. Second, our approach demonstrates a high sensitivity in miRNA-profiling because tissue and stage-

specific miRNAs like *lsy-6* are identified, while being hardly detected in qRT-PCR data[125]. In line with this, Knutsen *et al.* concluded that NGS platforms offer a higher sensitivity than qRT-PCR, based on a comparative investigation of miRNA profiling strategies in human breast cancer cell lines[146]. Finally, similar expression intensities of all members of the co-transcribed *mir-35-41* cluster indicate a high accuracy of our data, because these miRNAs are controlled by one promoter[67,296]. Nevertheless, this study lacks biological replication, which could significantly improve statistical detection power of differentially expressed genes[308]. However, there is a trade-off between biological replication and sequencing depth because experimental budget is usually limited. Liu and colleagues demonstrate that beyond a certain sequencing depth the power to detect DE genes is generally more improving by sampling additional biological replicates than by deeper sequencing[309]. Thus, the authors conclude that in most scenarios sampling additional biological replicates should be favored. Regardless, the results of my DE analysis provide valid candidate regulators, such as *mir-34* and *mir-71*. This finding is consistent with results from Karp *et al.*, who include biological replicates to measured expression level changes using qRT-PCR[125].

In this study I applied a simple two-sample comparison using binomial testing conditioned on the library sizes because no biological replicates were sampled. However, it has been argued that the library-to-library variability is not well captured by a binomial or Poisson distribution, because the mean-variance relationship of these models might not provide enough flexibility. A characteristic of the Poisson model is that the mean and the variance are assumed to be equal. However, if the variance is greater than the mean, overdispersion occurs[310]. Popular models that account for the problem of overdispersion include the negative binomial (gamma-Poisson)[311], beta-binomial[312], or two-stage Poisson models[313]. In particular, the negative binomial distribution is used in the implementation of the R Bioconductor packages edgeR[235] and DESeq[314], two frequently used methods in RNA-seq DGE studies that include replicates in their statistical model.

Next-generation sequencing methods, as applied here, can only measure relative quantification levels due to sequence-specific biases. Although it has been suggested that it might be possible to measure absolute quantification through calibration using spike-ins, i.e. a pool of concentration-defined input oligonucleotide standards[188,315], a recent review stated that NGS is not able to perform absolute quantification[208]. Single-molecular sequencers (or 3rd generation sequencers) may potentially solve this problem in the future[135]. However, 3rd generation sequencers are currently very expensive

in usage, provide higher error rates, are not widely accessible, and a single-molecule real-time approach has yet to be demonstrated for miRNA profiling[208].

5.3 Future Directions

Nematodes parasitism is a worldwide health problem with over 1 billion people being infected. According to the World Health Organization (WHO) statistics*, parasites are the cause of more human death than anything else apart from HIV/AIDS and Tuberculosis[34]. As a result, nematode plant and animal parasites are of great medical and economic importance[35]. However, the molecular mechanisms controlling the infections with parasites are poorly understood. Progress has been made and an increasing number of draft genomes of numerous free-living and parasitic nematodes has been published in recent years[316]. This is potentially due to the decreasing sequencing costs and the wide availability of deep sequencing methods combined with the small genome sizes of nematodes. Prospectively, the genome and transcriptome of many more species will be sequenced in the future with manageable effort.

This work contributes to this effort and presents a comparative genome-wide investigation of the miRNA transcriptome in dauer and infective larvae of nematodes. Studying the mechanisms that control nematode life cycles is an attractive approach to identify new therapeutic targets. Here, I present cross-species candidate regulators that may be important for developmental arrest and long-term survival in free-living and parasitic nematodes. Although these miRNAs may need further experimental validation through e.g. northern blot or *in situ* hybridization, they constitute interesting target genes for potential genetic engineering in free-living nematodes and parasites. In particular, it will be important to determine the target genes that are regulated by these miRNAs. Over the years, a couple of methods have been developed to identify miRNA targets: (i) small-scale genetic methods using a miRNA mutant strain , (ii) computational prediction tools, and (iii) high-throughput biochemical approaches (e.g. PAR-CLIP)[114,116]. While computational prediction of miRNA target genes is oftentimes not very specific and identifies a large number of potential targets, experimental approaches are generally time-consuming and complicated (if even possible) to apply. However, the ability to predict miRNA targets with high confidence is still a reminding challenge in the field[114]. Nevertheless, to understand the regulatory mechanisms underlying develop-

*http://www.who.int, accessed July, 2014

mental arrest and long-term survival in dauer and dauer-like stages in detail, the entire regulatory network has to be revealed in future studies, i.e. miRNA genes as well as their targets and the regulatory feedback loops involved.

5.4 Concluding Remarks

Taken together, this thesis describes an extensive set of tools and strategies for the analysis of post-transcriptional gene regulators in free-living and parasitic nematodes. The goal of this project was to analyze whether miRNA genes impact developmental arrest and long-term survival in dauer and dauer-like stages. In particular, I wanted to address the long-standing hypothesis that dauer and infective larvae share a common origin. The starting point of this work was the identification of miRNAs in high-throughput small RNA sequencing data profiled by two distinct sequencing platforms. In this context, I provided sophisticated bioinformatics solutions to analyze these small RNA-seq data sets and to address the aforementioned questions computationally.

Although our data suggests that miRNA gene sets diverge rapidly in nematodes, my in-depth assessment of miRNAs in free-living and parasitic nematodes reveals conserved post-transcriptional regulators with similar expression signatures in dauer vs. non-dauer fates. I highlighted the case of miR-34 and miR-71, two miRNAs that are both important regulators of stress response and aging not only in worms, but also in flies and mammals[139,317–319]. While the *mir-71* family is a well-conserved post-transcriptional regulator with coherent expression across all three species, the *mir-34* family could constitute a case of convergent gene evolution in *P. pacificus*. Herein, unrelated miRNA precursors with identical or almost identical (off by one substitution) seed sequences show similar expression patterns in the dauer fate as the reference family. This study reports the first coherent data on any *Strongyloides* parasite and provides a valuable resource to researchers studying miRNA genes and their evolution and specifically aspects in developmental arrest in free-living and parasitic nematodes.

References

[1] M. Dodt, J. T. Röhr, R. Ahmed, and C. Dieterich. FLEXBAR - Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms. *Biology*, 1(3):895–905, 2012.

[2] R. Ahmed, Z. Chang, A. E. Younis, *et al.* Conserved miRNAs Are Candidate Post-Transcriptional Regulators of Developmental Arrest in Free-Living and Parasitic Nematodes. *Genome Biol Evol*, 5(7):1246–1260, 2013.

[3] P. J. Lambshead. Recent developments in marine benthic biodiversity research. *Oceanis*, 19:5, 1993.

[4] M. L. Blaxter, P. D. Ley, J. R. Garey, *et al.* A molecular evolutionary framework for the phylum Nematoda. *Nature*, 392(6671):71–75, Mar 1998.

[5] D. L. Lee. *The Biology of Nematodes.* Taylor and Francis, New York, 2002.

[6] G. O. Poinar. *The evolutionary history of nematodes: as revealed in stone, amber and mummies.* Leiden (The Netherlands). Brill., 2011.

[7] A. F. Bird and J. Bird. *The structure of nematodes.* Academic Press, 1991.

[8] W. J. Kent, C. W. Sugnet, T. S. Furey, *et al.* The human genome browser at UCSC. *Genome Res*, 12(6):996–1006, Jun 2002.

[9] S. Leroy, C. Duperray, and S. Morand. Flow cytometry for parasite nematode genome size measurement. *Mol Biochem Parasitol*, 128(1):91–93, Apr 2003.

[10] T. Chiche. *The biology and genome of Heterorhabditis bacteriophora.* WormBook, ed. The *C. elegans* Research Community, Feb 2007. URL `http://dx.doi.org/10.1895/wormbook.1.135.1`.

[11] R. J. Sommer and A. Streit. Comparative genetics and genomics of nematodes: genome structure, development, and lifestyle. *Annu Rev Genet*, 45:1–20, Dec 2011.

[12] M. Blaxter. *Caenorhabditis elegans* is a nematode. *Science*, 282(5396):2041–2046, Dec 1998.

[13] S. Brenner. The genetics of behaviour. *Br Med Bull*, 29(3):269–271, Sep 1973.

[14] D. Hirsh and R. Vanderslice. Temperature-sensitive developmental mutants of *Caenorhabditis* elegans. *Dev Biol*, 49(1):220–235, Mar 1976.

[15] J. E. Sulston, E. Schierenberg, J. G. White, and J. N. Thomson. The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Dev Biol*, 100(1):64–119, Nov 1983.

[16] L. Byerly, R. C. Cassada, and R. L. Russell. The life cycle of the nematode *Caenorhabditis elegans*. I. Wild-type growth and reproduction. *Dev Biol*, 51(1):23–33, Jul 1976.

[17] The *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, 282(5396):2012–2018, Dec 1998.

[18] T. W. Harris, I. Antoshechkin, T. Bieri, *et al.* WormBase: a comprehensive resource for nematode research. *Nucleic Acids Res*, 38(Database issue):D463–D467, Jan 2010.

[19] J. W. Golden and D. L. Riddle. The *Caenorhabditis elegans* dauer larva: developmental effects of pheromone, food, and temperature. *Dev Biol*, 102(2):368–378, Apr 1984.

[20] R. C. Cassada and R. L. Russell. The dauer larva, a post-embryonic developmental variant of the nematode *Caenorhabditis elegans*. *Dev Biol*, 46(2):326–342, Oct 1975.

[21] D. L. Riddle and P. S. Albert. *Genetic and Environmental Regulation of Dauer Larva Development.* Cold Spring Harbor Laboratory Press, 1997.

[22] M. Klass and D. Hirsh. Non-ageing developmental variant of *Caenorhabditis elegans*. *Nature*, 260(5551):523–525, Apr 1976.

[23] K. Kiontke and W. Sudhaus. *Ecology of Caenorhabditis species.* WormBook, ed. The *C. elegans* Research Community, WormBook, 2006.

[24] D. L. Motola, C. L. Cummins, V. Rottiers, *et al.* Identification of ligands for DAF-12 that govern dauer formation and reproduction in *C. elegans*. *Cell*, 124(6):1209–1223, Mar 2006.

[25] A. H. Ludewig, C. Kober-Eisermann, C. Weitzel, *et al.* A novel nuclear receptor/coregulator complex controls *C. elegans* lipid metabolism, larval development, and aging. *Genes Dev*, 18(17):2120–2133, Sep 2004.

[26] D. L. Riddle, M. M. Swanson, and P. S. Albert. Interacting genes in nematode dauer larva formation. *Nature*, 290(5808):668–671, Apr 1981.

[27] R. Sommer. *Pristionchus pacificus*. *WormBook*, pages 1–8, 2006. URL `http://dx.doi.org/10.1895/wormbook.1.102.1`.

[28] C. Dieterich, S. Clifton, L. Schuster, *et al.* The *Pristionchus pacificus* genome provides a unique perspective on nematode lifestyle and parasitism. *Nat Genet*, 40(10):1193–1198, Oct 2008.

[29] R. L. Hong and R. J. Sommer. Chemoattraction in *Pristionchus* nematodes and implications for insect recognition. *Curr Biol*, 16(23):2359–2365, Dec 2006.

[30] R. C. Anderson. The origins of zooparasitic nematodes. *Canadian Journal of Zoology*, 62 (3):317–328, 1984.

[31] B. Weischer and D. J. F. Brown. *An Introduction to Nematodes: General Nematology.* Pensoft Publishers, Sofia, Bulgaria, 2000.

[32] R. Poulin. *Evolutionary ecology of parasites*. Princeton University Press, second edition edition, 2007.

[33] A. Ogawa, A. Streit, A. Antebi, and R. J. Sommer. A conserved endocrine mechanism controls the formation of dauer and infective larvae in nematodes. *Curr Biol*, 19(1):67–71, Jan 2009.

[34] S. I. Hirst and L. A. Stapley. Parasitology: the dawn of a new millennium. *Parasitol Today*, 16(1):1–3, Jan 2000.

[35] D. P. Jasmer, A. Goverse, and G. Smant. Parasitic nematode interactions with mammals and plants. *Annu Rev Phytopathol*, 41:245–270, 2003.

[36] M. E. Viney. Exploiting the life cycle of *Strongyloides ratti*. *Parasitol Today*, 15(6):231–235, Jun 1999.

[37] M. Blaxter and D. Bird. Parasitic nematodes. *Cold Spring Harbor Monograph Series; C. elegans II*, 33:851–878, 1997.

[38] T. R. Buerglin, E. Lobos, and M. L. Blaxter. *Caenorhabditis elegans* as a model for parasitic nematodes. *International Journal for Parasitology*, 28:395–411, 1998.

[39] R. J. Sommer and A. Ogawa. Hormone signaling and phenotypic plasticity in nematode development and evolution. *Curr Biol*, 21(18):R758–R766, Sep 2011.

[40] Z. Wang, X. E. Zhou, D. L. Motola, *et al.* Identification of the nuclear receptor DAF-12 as a therapeutic target in parasitic nematodes. *Proc Natl Acad Sci U S A*, 106(23):9138–9143, Jun 2009.

[41] B. K. Dalley and M. Golomb. Gene expression in the *Caenorhabditis elegans* dauer larva: developmental regulation of Hsp90 and other genes. *Dev Biol*, 151(1):80–90, May 1992.

[42] T. J. Reape and A. M. Burnell. Dauer larva recovery in the nematode *Caenorhabditis elegans* - II. The effect of inhibitors of protein synthesis on recovery, growth and pharyngeal pumping. *Comparative Biochemistry and Physiology*, 98:245–252, 1991.

[43] T. J. Reape and A. M. Burnell. Dauer larva recovery in the nematode *Caenorhabditis elegans* - III. The effect of inhibitors of protein and mRNA synthesis on the activity of the enzymes of intermediary. *Comparative Biochemistry and Physiology B*, 102:241–245, 1992.

[44] T. J. Reape and A. M. Burnell. Dauer larva recovery in *Caenorhabditis elegans* - I. The effect of mRNA synthesis inhibitors on recovery, growth and pharyngeal pumping. *Comparative Biochemistry and Physiology*, 98:239–243, 1991.

[45] F. Crick. On protein synthesis. *Symp Soc Exp Biol*, 12:138–163, 1958.

[46] F. Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, Aug 1970.

[47] J. M. Vaquerizas, S. K. Kummerfeld, S. A. Teichmann, and N. M. Luscombe. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet*, 10(4): 252–263, Apr 2009.

[48] S. L. Berger. The complex language of chromatin regulation during transcription. *Nature*, 447(7143):407–412, May 2007.

[49] M. J. Moore. From birth to death: the complex lives of eukaryotic mRNAs. *Science*, 309 (5740):1514–1518, Sep 2005.

[50] J. D. Keene. RNA regulons: coordination of post-transcriptional events. *Nat Rev Genet*, 8(7):533–543, Jul 2007.

[51] W. Filipowicz, S. N. Bhattacharyya, and N. Sonenberg. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat Rev Genet*, 9(2):102–114, Feb 2008.

[52] E. S. Lander, L. M. Linton, B. Birren, *et al.* Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb 2001.

[53] S. Ohno. So much 'junk' DNA in our genome. *Brookhaven Symp Biol*, 23:366–370, 1972.

[54] The ENCODE Project Consortium, E. Birney, J. A. Stamatoyannopoulos, *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816, Jun 2007.

[55] C. P. Ponting and T. G. Belgard. Transcribed dark matter: meaning or myth? *Hum Mol Genet*, 19(R2):R162–R168, Oct 2010.

[56] R. J. Taft, M. Pheasant, and J. S. Mattick. The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays*, 29(3):288–299, Mar 2007.

[57] The ENCODE Project Consortium, B. E. Bernstein, E. Birney, *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, Sep 2012.

[58] M. Ghildiyal and P. Zamore. Small silencing RNAs: an expanding universe. *Nat Rev Genet*, 10(2):94–108, Feb 2009.

[59] J. L. Rinn and H. Y. Chang. Genome regulation by long noncoding RNAs. *Annu Rev Biochem*, 81:145–166, 2012.

[60] W. P. Kloosterman and R. H. A. Plasterk. The diverse functions of microRNAs in animal development and disease. *Dev Cell*, 11(4):441–450, Oct 2006.

[61] L. He and G. J. Hannon. MicroRNAs: small RNAs with a big role in gene regulation. *Nat Rev Genet*, 5(7):522–531, Jul 2004.

[62] R. C. Lee, R. L. Feinbaum, and V. Ambros. *The C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):843–854, Dec 1993.

[63] B. Wightman, I. Ha, and G. Ruvkun. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in *C. elegans*. *Cell*, 75(5):855–862, Dec 1993.

[64] M. Chalfie, H. R. Horvitz, and J. E. Sulston. Mutations that lead to reiterations in the cell lineages of C. elegans. *Cell*, 24(1):59–69, Apr 1981.

[65] B. J. Reinhart, F. J. Slack, M. Basson, *et al.* The 21-nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature*, 403(6772):901–906, Feb 2000.

[66] A. E. Pasquinelli, B. J. Reinhart, F. Slack, *et al.* Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, 408(6808):86–89, Nov 2000.

[67] N. C. Lau, L. P. Lim, E. G. Weinstein, and D. P. Bartel. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science*, 294(5543):858–862, Oct 2001.

[68] R. C. Lee and V. Ambros. An extensive class of small RNAs in *Caenorhabditis elegans*. *Science*, 294(5543):862–864, Oct 2001.

[69] M. Lagos-Quintana, R. Rauhut, W. Lendeckel, and T. Tuschl. Identification of novel genes coding for small expressed RNAs. *Science*, 294(5543):853–858, Oct 2001.

[70] B. J. Reinhart, E. G. Weinstein, M. W. Rhoades, B. Bartel, and D. P. Bartel. MicroRNAs in plants. *Genes Dev*, 16(13):1616–1626, Jul 2002.

[71] W. Park, J. Li, R. Song, J. Messing, and X. Chen. CARPEL FACTORY, a Dicer homolog, and HEN1, a novel protein, act in microRNA metabolism in *Arabidopsis thaliana*. *Curr Biol*, 12(17):1484–1495, Sep 2002.

[72] E. Berezikov, E. Cuppen, and R. H. A. Plasterk. Approaches to microRNA discovery. *Nat Genet*, 38 Suppl:S2–S7, Jun 2006.

[73] D. J. Studholme. Deep sequencing of small RNAs in plants: applied bioinformatics. *Brief Funct Genomics*, 11(1):71–85, Jan 2012.

[74] A. Kozomara and S. Griffiths-Jones. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res*, 39(Database issue):D152–D157, Jan 2011.

[75] A. Kozomara and S. Griffiths-Jones. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res*, 42(Database issue):D68–D73, Jan 2014.

[76] T.-C. Chang and J. T. Mendell. microRNAs in vertebrate physiology and human disease. *Annu Rev Genomics Hum Genet*, 8:215–239, 2007.

[77] A. Lujambio and S. W. Lowe. The microcosmos of cancer. *Nature*, 482(7385):347–355, Feb 2012.

[78] X. Cai, C. H. Hagedorn, and B. R. Cullen. Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA*, 10(12):1957–1966, Dec 2004.

[79] Y. Lee, M. Kim, J. Han, *et al.* MicroRNA genes are transcribed by RNA polymerase II. *EMBO J*, 23(20):4051–4060, Oct 2004.

[80] V. N. Kim, J. Han, and M. C. Siomi. Biogenesis of small RNAs in animals. *Nat Rev Mol Cell Biol*, 10(2):126–139, Feb 2009.

[81] B. Czech and G. J. Hannon. Small RNA sorting: matchmaking for Argonautes. *Nat Rev Genet*, 12(1):19–31, Jan 2011.

[82] R. Yi, Y. Qin, I. G. Macara, and B. R. Cullen. Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes Dev*, 17(24):3011–3016, Dec 2003.

[83] M. T. Bohnsack, K. Czaplinski, and D. Gorlich. Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. *RNA*, 10(2):185–191, Feb 2004.

[84] E. Lund, S. Gttinger, A. Calado, J. E. Dahlberg, and U. Kutay. Nuclear export of microRNA precursors. *Science*, 303(5654):95–98, Jan 2004.

[85] D. Murphy, B. Dancis, and J. R. Brown. The evolution of core proteins involved in microRNA biogenesis. *BMC Evol Biol*, 8:92, 2008.

[86] E. Bernstein, A. A. Caudy, S. M. Hammond, and G. J. Hannon. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature*, 409(6818):363–366, Jan 2001.

[87] A. Grishok, A. E. Pasquinelli, D. Conte, *et al.* Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control C. elegans developmental timing. *Cell*, 106(1):23–34, Jul 2001.

[88] G. Hutvgner, J. McLachlan, A. E. Pasquinelli, *et al.* A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science*, 293(5531):834–838, Aug 2001.

[89] R. F. Ketting, S. E. Fischer, E. Bernstein, *et al.* Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C. elegans. *Genes Dev*, 15 (20):2654–2659, Oct 2001.

[90] K. Okamura, J. W. Hagen, H. Duan, D. M. Tyler, and E. C. Lai. The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila. *Cell*, 130(1):89–100, Jul 2007.

[91] J. G. Ruby, C. H. Jan, and D. P. Bartel. Intronic microRNA precursors that bypass Drosha processing. *Nature*, 448(7149):83–86, Jul 2007.

[92] E. Berezikov, W.-J. Chung, J. Willis, E. Cuppen, and E. C. Lai. Mammalian mirtron genes. *Mol Cell*, 28(2):328–336, Oct 2007.

[93] J. E. Babiarz, J. G. Ruby, Y. Wang, D. P. Bartel, and R. Blelloch. Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs. *Genes Dev*, 22(20):2773–2785, Oct 2008.

[94] W.-J. Chung, P. Agius, J. O. Westholm, *et al.* Computational and experimental identification of mirtrons in Drosophila melanogaster and Caenorhabditis elegans. *Genome Res*, 21(2):286–300, Feb 2011.

[95] A. S. Flynt, J. C. Greimann, W.-J. Chung, C. D. Lima, and E. C. Lai. MicroRNA biogenesis via splicing and exosome-mediated trimming in Drosophila. *Mol Cell*, 38(6):900–907, Jun 2010.

[96] D. S. Schwarz, G. Hutvgner, T. Du, *et al.* Asymmetry in the assembly of the RNAi enzyme complex. *Cell*, 115(2):199–208, Oct 2003.

[97] A. Khvorova, A. Reynolds, and S. D. Jayasena. Functional siRNAs and miRNAs exhibit strand bias. *Cell*, 115(2):209–216, Oct 2003.

[98] A. Stark, P. Kheradpour, L. Parts, *et al.* Systematic discovery and characterization of fly microRNAs using 12 Drosophila genomes. *Genome Res*, 17(12):1865–1879, Dec 2007.

[99] W. Bender. MicroRNAs in the *Drosophila* bithorax complex. *Genes Dev*, 22(1):14–19, Jan 2008.

[100] K. Okamura, M. Phillips, D. Tyler, *et al.* The regulatory activity of microRNA* species has substantial influence on microRNA and 3' UTR evolution. *Nat Struct Mol Biol*, 15(4): 354–363, Apr 2008.

[101] A. Stark, N. Bushati, C. H. Jan, *et al.* A single Hox locus in *Drosophila* produces functional microRNAs from opposite DNA strands. *Genes Dev*, 22(1):8–13, Jan 2008.

[102] D. M. Tyler, K. Okamura, W.-J. Chung, *et al.* Functionally distinct regulatory RNAs generated by bidirectional transcription and processing of microRNA loci. *Genes Dev*, 22 (1):26–36, Jan 2008.

[103] J.-S. Yang, M. D. Phillips, D. Betel, *et al.* Widespread regulatory activity of vertebrate microRNA* species. *RNA*, 17(2):312–326, Feb 2011.

[104] B. Guennewig, M. Roos, A. M. Dogar, *et al.* Synthetic pre-microRNAs reveal dual-strand activity of miR-34a on TNF-α. *RNA*, 20(1):61–75, Jan 2014.

[105] S. Ro, C. Park, D. Young, K. M. Sanders, and W. Yan. Tissue-dependent paired expression of miRNAs. *Nucleic Acids Res*, 35(17):5944–5953, 2007.

[106] H. R. Chiang, L. W. Schoenfeld, J. G. Ruby, *et al.* Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. *Genes Dev*, 24(10):992–1009, May 2010.

[107] J. G. Ruby, A. Stark, W. K. Johnston, *et al.* Evolution, biogenesis, expression, and target predictions of a substantially expanded set of *Drosophila* microRNAs. *Genome Res*, 17 (12):1850–1864, Dec 2007.

[108] E. de Wit, S. E. V. Linsen, E. Cuppen, and E. Berezikov. Repertoire and evolution of miRNA genes in four divergent nematode species. *Genome Res*, 19(11):2064–2074, Nov 2009.

[109] A. Marco, J. H. L. Hui, M. Ronshaugen, and S. Griffiths-Jones. Functional shifts in insect microRNA evolution. *Genome Biol Evol*, 2:686–696, 2010.

[110] S. Griffiths-Jones, J. H. L. Hui, A. Marco, and M. Ronshaugen. MicroRNA evolution by arm switching. *EMBO Rep*, 12(2):172–177, Feb 2011.

[111] G. Mathonnet, M. R. Fabian, Y. V. Svitkin, *et al.* MicroRNA inhibition of translation initiation in vitro by targeting the cap-binding complex eIF4F. *Science*, 317(5845):1764–1767, Sep 2007.

[112] S. Djuranovic, A. Nahvi, and R. Green. miRNA-mediated gene silencing by translational repression followed by mRNA deadenylation and decay. *Science*, 336(6078):237–240, Apr 2012.

[113] D. Bartel. MicroRNAs: target recognition and regulatory functions. *Cell*, 136(2):215–233, Jan 2009.

[114] A. E. Pasquinelli. MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. *Nat Rev Genet*, 13(4):271–282, Apr 2012.

[115] S. W. Chi, J. B. Zang, A. Mele, and R. B. Darnell. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, 460(7254):479–486, Jul 2009.

[116] M. Hafner, M. Landthaler, L. Burger, *et al.* Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, 141(1):129–141, Apr 2010.

[117] D. G. Zisoulis, M. T. Lovci, M. L. Wilbert, *et al.* Comprehensive discovery of endogenous Argonaute binding sites in Caenorhabditis elegans. *Nat Struct Mol Biol*, 17(2):173–179, Feb 2010.

[118] A. K. L. Leung, A. G. Young, A. Bhutkar, *et al.* Genome-wide identification of Ago2 binding sites from mouse embryonic stem cells with and without mature microRNAs. *Nat Struct Mol Biol*, 18(2):237–244, Feb 2011.

[119] D. P. Bartel. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 116(2): 281–297, Jan 2004.

[120] C. Ibanez-Ventoso, M. Yang, S. Guo, *et al.* Modulated microRNA expression during adult lifespan in *Caenorhabditis elegans*. *Aging Cell*, 5(3):235–246, Jun 2006.

[121] A. Bethke, N. Fielenbach, Z. Wang, D. J. Mangelsdorf, and A. Antebi. Nuclear hormone receptor regulation of microRNAs controls developmental progression. *Science*, 324(5923): 95–98, Apr 2009.

[122] C. M. Hammell, X. Karp, and V. Ambros. A feedback circuit involving let-7-family miRNAs and DAF-12 integrates environmental signals and developmental timing in *Caenorhabditis elegans*. *Proc Natl Acad Sci U S A*, 106(44):18668–18673, Nov 2009.

[123] A. de Lencastre, Z. Pincus, K. Zhou, *et al.* MicroRNAs both promote and antagonize longevity in *C. elegans*. *Curr Biol*, 20(24):2159–2168, Dec 2010.

[124] X. Zhang, R. Zabinsky, Y. Teng, M. Cui, and M. Han. microRNAs play critical roles in the survival and recovery of *Caenorhabditis elegans* from starvation-induced L1 diapause. *Proc Natl Acad Sci U S A*, 108(44):17997–18002, Nov 2011.

[125] X. Karp, M. Hammell, M. C. Ow, and V. Ambros. Effect of life history on microRNA expression during *C. elegans* development. *RNA*, 17(4):639–651, Apr 2011.

[126] X. Karp and V. Ambros. Dauer larva quiescence alters the circuitry of microRNA pathways regulating cell fate progression in *C. elegans*. *Development*, 139(12):2177–2186, Jun 2012.

[127] J. D. McPherson, M. Marra, L. Hillier, *et al.* A physical map of the human genome. *Nature*, 409(6822):934–941, Feb 2001.

[128] J. C. Venter, M. D. Adams, E. W. Myers, *et al.* The sequence of the human genome. *Science*, 291(5507):1304–1351, Feb 2001.

[129] F. S. Collins, M. Morgan, and A. Patrinos. The Human Genome Project: lessons from large-scale biology. *Science*, 300(5617):286–290, Apr 2003.

[130] E. R. Mardis. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet*, 9:387–402, 2008.

[131] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, 74(12):5463–5467, Dec 1977.

[132] M. Margulies, M. Egholm, W. E. Altman, *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, Sep 2005.

[133] S. Bao, R. Jiang, W. Kwan, *et al.* Evaluation of next-generation sequencing software in mapping and assembly. *J Hum Genet*, 56(6):406–414, Jun 2011.

[134] M. L. Metzker. Sequencing technologies - the next generation. *Nat Rev Genet*, 11(1): 31–46, Jan 2010.

[135] E. E. Schadt, S. Turner, and A. Kasarskis. A window into third-generation sequencing. *Hum Mol Genet*, 19(R2):R227–R240, Oct 2010.

[136] C. S. Pareek, R. Smoczynski, and A. Tretyn. Sequencing technologies and genome sequencing. *J Appl Genet*, 52(4):413–435, Nov 2011.

[137] R. D. Mitra and G. M. Church. In situ localized amplification and contact replication of many individual DNA molecules. *Nucleic Acids Res*, 27(24):e34, Dec 1999.

[138] J. Shendure and H. Ji. Next-generation DNA sequencing. *Nat Biotechnol*, 26(10):1135–1145, Oct 2008.

[139] N. Liu, M. Landreh, K. Cao, *et al.* The microRNA miR-34 modulates ageing and neurodegeneration in *Drosophila*. *Nature*, 482(7386):519–523, Feb 2012.

[140] L. Liu, Y. Li, S. Li, *et al.* Comparison of next-generation sequencing systems. *J Biomed Biotechnol*, 2012:251364, 2012.

[141] M. Jessri and C. S. Farah. Next generation sequencing and its application in deciphering head and neck cancer. *Oral Oncol*, 50(4):247–253, Apr 2014.

[142] M. Fedurco, A. Romieu, S. Williams, I. Lawrence, and G. Turcatti. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res*, 34(3):e22, 2006.

[143] E. R. Mardis. Next-generation sequencing platforms. *Annu Rev Anal Chem (Palo Alto Calif)*, 6:287–303, 2013.

[144] D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, Nov 2008.

[145] K. Nakamura, T. Oshima, T. Morimoto, *et al.* Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res*, 39(13):e90, Jul 2011.

[146] E. Knutsen, T. Fiskaa, A. Ursvik, *et al.* Performance comparison of digital microRNA profiling technologies applied on human breast cancer cell lines. *PLoS One*, 8(10):e75813, 2013.

[147] S. E. Celniker, L. A. L. Dillon, M. B. Gerstein, *et al.* Unlocking the secrets of the genome. *Nature*, 459(7249):927–930, Jun 2009.

[148] J. Shendure, G. J. Porreca, N. B. Reppas, *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, 309(5741):1728–1732, Sep 2005.

[149] K. McKernan, A. Blanchard, L. Kotler, and G. Costa. Reagents, methods, and libraries for bead-based sequencing. 2006. US patent application 20080003571.

[150] A. E. Tomkinson, S. Vijayakumar, J. M. Pascal, and T. Ellenberger. DNA ligases: structure, reaction mechanism, and function. *Chem Rev*, 106(2):687–699, Feb 2006.

[151] J. N. Housby and E. M. Southern. Fidelity of DNA ligation: a novel experimental approach based on the polymerisation of libraries of oligonucleotides. *Nucleic Acids Res*, 26(18):4259–4266, Sep 1998.

[152] U. Landegren, R. Kaiser, J. Sanders, and L. Hood. A ligase-mediated gene detection technique. *Science*, 241(4869):1077–1080, Aug 1988.

[153] O. Harismendy, P. C. Ng, R. L. Strausberg, *et al.* Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol*, 10(3):R32, 2009.

[154] J. H. Schulte, T. Marschall, M. Martin, *et al.* Deep sequencing reveals differential expression of microRNAs in favorable versus unfavorable neuroblastoma. *Nucleic Acids Res*, 38 (17):5919–5928, Sep 2010.

[155] T. T. Bizuayehu, C. F. C. Lanes, T. Furmanek, *et al.* Differential expression patterns of conserved miRNAs and isomiRs during Atlantic halibut development. *BMC Genomics*, 13:11, 2012.

[156] T. T. Bizuayehu, J. Babiak, B. Norberg, *et al.* Sex-biased miRNA expression in Atlantic halibut (*Hippoglossus hippoglossus*) brain and gonads. *Sex Dev*, 6(5):257–266, 2012.

[157] G. Liang, J. Li, B. Sun, *et al.* Deep sequencing reveals complex mechanisms of microRNA deregulation in colorectal cancer. *Int J Oncol*, 45(2):603–610, Aug 2014.

[158] F. Ozsolak and P. M. Milos. RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet*, 12(2):87–98, Feb 2011.

[159] C. Lu, B. C. Meyers, and P. J. Green. Construction of small RNA cDNA libraries for deep sequencing. *Methods*, 43(2):110–117, Oct 2007.

[160] S. Brenner. The Genetics of Caenorhabditis elegans. *Genetics*, 77(1):71–94, May 1974. ISSN 1943-2631.

[161] M. E. Viney and J. B. Lok. Strongyloides spp. *WormBook*, pages 1–15, 2007. URL `http://dx.doi.org/10.1895/wormbook.1.141.1`.

[162] J. Keiser, K. Thiemann, Y. Endriss, and J. Utzinger. *Strongyloides ratti*: in vitro and in vivo activity of tribendimidine. *PLoS Negl Trop Dis*, 2(1):e136, 2008.

[163] J. A. Lewis and J. T. Fleming. Basic culture methods. *Methods Cell Biol*, 48:3–29, 1995. ISSN 0091-679X.

[164] J. B. Lok. Strongyloides stercoralis: a model for translational research on parasitic nematode biology. *WormBook*, pages 1–18, 2007.

[165] H. Soblik, A. E. Younis, M. Mitreva, *et al.* Life cycle stage-resolved proteomic analysis of the excretome/secretome from Strongyloides ratti–identification of stage-specific proteases. *Mol Cell Proteomics*, 10(12):M111.010157, Dec 2011.

[166] R. Edgar, M. Domrachev, and A. E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*, 30(1):207–210, Jan 2002.

[167] J. G. Ruby, C. Jan, C. Player, *et al.* Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell*, 127(6):1193–1207, Dec 2006.

[168] P. J. Batista, J. G. Ruby, J. M. Claycomb, *et al.* PRG-1 and 21U-RNAs interact to form the piRNA complex required for fertility in *C. elegans*. *Mol Cell*, 31(1):67–78, Jul 2008.

[169] P. P. Gardner, J. Daub, J. Tate, *et al.* Rfam: Wikipedia, clans and the "decimalrelease. *Nucleic Acids Res*, 39(Database issue):D141–D145, Jan 2011.

[170] R. C. Gentleman, V. J. Carey, D. M. Bates, *et al.* Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80, 2004.

[171] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0.

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

[173] B. Ewing and P. Green. Base-calling of automated sequencer traces using *Phred*. II. Error probabilities. *Genome Res*, 8(3):186–194, Mar 1998.

[174] D. S. Horner, G. Pavesi, T. Castrignan, *et al.* Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. *Brief Bioinform*, 11(2): 181–197, Mar 2010.

[175] A. Goffeau, B. G. Barrell, H. Bussey, *et al.* Life with 6000 genes. *Science*, 274(5287):546, 563–546, 567, Oct 1996.

[176] M. D. Adams, S. E. Celniker, R. A. Holt, *et al.* The genome sequence of *Drosophila melanogaster*. *Science*, 287(5461):2185–2195, Mar 2000.

[177] G. M. Church and S. Kieffer-Higgins. Multiplex DNA sequencing. *Science*, 240(4849): 185–188, Apr 1988.

[178] M. Meyer, U. Stenzel, S. Myles, K. Prfer, and M. Hofreiter. Targeted high-throughput sequencing of tagged nucleic acid samples. *Nucleic Acids Res*, 35(15):e97, 2007.

[179] P. Parameswaran, R. Jalili, L. Tao, *et al.* A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. *Nucleic Acids Res*, 35 (19):e130, 2007.

[180] D. W. Craig, J. V. Pearson, S. Szelinger, *et al.* Identification of genetic variants using bar-coded multiplexed sequencing. *Nat Methods*, 5(10):887–893, Oct 2008.

[181] F. Vigneault, A. M. Sismour, and G. M. Church. Efficient microRNA capture and barcoding via enzymatic oligonucleotide adenylation. *Nat Methods*, 5(9):777–779, Sep 2008.

[182] A. M. Smith, L. E. Heisler, J. Mellor, *et al.* Quantitative phenotyping via deep barcode sequencing. *Genome Res*, 19(10):1836–1842, Oct 2009.

[183] H. P. J. Buermans, Y. Ariyurek, G. van Ommen, J. T. den Dunnen, and P. A. C. 't Hoen. New methods for next generation sequencing based microRNA expression profiling. *BMC Genomics*, 11:716, 2010.

[184] I. J. Nijman, M. Mokry, R. van Boxtel, *et al.* Mutation discovery by targeted genomic enrichment of multiplexed barcoded samples. *Nat Methods*, 7(11):913–915, Nov 2010.

[185] A. M. Smith, L. E. Heisler, R. P. S. Onge, *et al.* Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. *Nucleic Acids Res*, 38(13):e142, Jul 2010.

[186] M. Hafner, N. Renwick, T. A. Farazi, *et al.* Barcoded cDNA library preparation for small RNA profiling by next-generation sequencing. *Methods*, 58(2):164–170, Oct 2012.

[187] S. Alon, F. Vigneault, S. Eminaga, *et al.* Barcoding bias in high-throughput multiplex sequencing of miRNA. *Genome Res*, 21(9):1506–1511, Sep 2011.

[188] M. Hafner, N. Renwick, M. Brown, *et al.* RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. *RNA*, 17(9):1697–1712, Sep 2011.

[189] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48(3):443–453, Mar 1970.

[190] C. Trapnell and S. L. Salzberg. How to map billions of short reads onto genomes. *Nat Biotechnol*, 27(5):455–457, May 2009.

[191] M. Ruffalo, T. LaFramboise, and M. Koyutrk. Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics*, 27(20):2790–2796, Oct 2011.

[192] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *J Mol Biol*, 147(1):195–197, Mar 1981.

[193] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, Oct 1990.

[194] W. J. Kent. BLAT–the BLAST-like alignment tool. *Genome Res*, 12(4):656–664, Apr 2002.

[195] J. A. Cox. ELAND: Efficient large-scale alignment of nucleotide database. *Illumina, San Diego*, 2007.

[196] H. Li, J. Ruan, and R. Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res*, 18(11):1851–1858, Nov 2008.

[197] M. David, M. Dzamba, D. Lister, L. Ilie, and M. Brudno. SHRiMP2: sensitive yet practical SHort Read Mapping. *Bioinformatics*, 27(7):1011–1012, Apr 2011.

[198] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 10(3):R25, 2009.

[199] B. Langmead and S. L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nat Methods*, 9(4):357–359, Apr 2012.

[200] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, Jul 2009.

[201] R. Li, C. Yu, Y. Li, *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15):1966–1967, Aug 2009.

[202] M. Burrows and D. Wheeler. Technical report 124. Palo Alto, CA: Digital Equipment Corporation. *A block-sorting lossless data compression algorithm*, 1994.

[203] P. Ferragina and G. Manzini. Opportunistic data structures with applications. In *Proceedings of the 41st Symposium on Foundation of Computer Science (FOCS 2000)*, pages 390–398. IEEE Computer Society, 2000.

[204] P. Flicek and E. Birney. Sense from sequence reads: methods for alignment and assembly. *Nat Methods*, 6(11 Suppl):S6–S12, Nov 2009.

[205] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*, 5(7):621–628, Jul 2008.

[206] S. E. V. Linsen, E. de Wit, G. Janssens, *et al.* Limitations and possibilities of small RNA digital gene expression profiling. *Nat Methods*, 6(7):474–476, Jul 2009.

[207] D. Leshkowitz, S. Horn-Saban, Y. Parmet, and E. Feldmesser. Differences in microRNA detection levels are technology and sequence dependent. *RNA*, 19(4):527–538, Apr 2013.

[208] C. C. Pritchard, H. H. Cheng, and M. Tewari. MicroRNA profiling: approaches and considerations. *Nat Rev Genet*, 13(5):358–369, May 2012.

[209] P. Landgraf, M. Rusu, R. Sheridan, *et al.* A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*, 129(7):1401–1414, Jun 2007.

[210] M. R. Friedländer, W. Chen, C. Adamidi, *et al.* Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotechnol*, 26(4):407–415, Apr 2008.

[211] Y. Li, Z. Zhang, F. Liu, *et al.* Performance comparison and evaluation of software tools for microRNA deep-sequencing data analysis. *Nucleic Acids Res*, 40(10):4298–4305, May 2012.

[212] V. Williamson, A. Kim, B. Xie, *et al.* Detecting miRNAs in deep-sequencing data: a software performance comparison and evaluation. *Brief Bioinform*, 14(1):36–45, Jan 2013.

[213] M. Hackenberg, M. Sturm, D. Langenberger, J. M. Falcn-Prez, and A. M. Aransay. miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Res*, 37(Web Server issue):W68–W76, Jul 2009.

[214] W.-C. Wang, F.-M. Lin, W.-C. Chang, *et al.* miRExpress: analyzing high-throughput sequencing data for profiling microRNA expression. *BMC Bioinformatics*, 10:328, 2009.

[215] D. Hendrix, M. Levine, and W. Shi. miRTRAP, a computational method for the systematic identification of miRNAs from high throughput sequencing data. *Genome Biol*, 11(4): R39, 2010.

[216] P.-J. Huang, Y.-C. Liu, C.-C. Lee, *et al.* DSAP: deep-sequencing small RNA analysis pipeline. *Nucleic Acids Res*, 38(Web Server issue):W385–W391, Jul 2010.

[217] E. Zhu, F. Zhao, G. Xu, *et al.* mirTools: microRNA profiling and discovery based on high-throughput sequencing. *Nucleic Acids Res*, 38(Web Server issue):W392–W397, Jul 2010.

[218] A. Mathelier and A. Carbone. MIReNA: finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data. *Bioinformatics*, 26(18):2226–2234, Sep 2010.

[219] R. Ronen, I. Gan, S. Modai, *et al.* miRNAkey: a software for microRNA deep sequencing analysis. *Bioinformatics*, 26(20):2615–2616, Oct 2010.

[220] M. R. Friedländer, S. D. Mackowiak, N. Li, W. Chen, and N. Rajewsky. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res*, 40(1):37–52, Jan 2012.

[221] I. L. Hofacker. Vienna RNA secondary structure server. *Nucleic Acids Res*, 31(13): 3429–3431, Jul 2003.

[222] E. Bonnet, J. Wuyts, P. Rouz, and Y. V. de Peer. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics*, 20(17):2911–2917, Nov 2004.

[223] S. W. Burge, J. Daub, R. Eberhardt, *et al.* Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res*, 41(Database issue):D226–D232, Jan 2013.

[224] C. J. Creighton, J. G. Reid, and P. H. Gunaratne. Expression profiling of microRNAs by deep sequencing. *Brief Bioinform*, 10(5):490–497, Sep 2009.

[225] K. P. McCormick, M. R. Willmann, and B. C. Meyers. Experimental design, preprocessing, normalization and differential expression analysis of small RNA sequencing experiments. *Silence*, 2(1):2, 2011.

[226] P. Chugh and D. P. Dittmer. Potential pitfalls in microRNA profiling. *Wiley Interdiscip Rev RNA*, 3(5):601–616, 2012.

[227] L. X. Garmire and S. Subramaniam. Evaluation of normalization methods in mammalian microRNA-Seq data. *RNA*, 18(6):1279–1288, Jun 2012.

[228] J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, 11:94, 2010.

[229] M. D. Robinson and A. Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*, 11(3):R25, 2010.

[230] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, Jan 2003.

[231] M.-A. Dillies, A. Rau, J. Aubert, *et al.* A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform*, 14(6):671–683, Nov 2013.

[232] S. Dudoit, Y. H. Yang, M. J. Callow, and T. P. Speed. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica sinica*, 12(1):111–140, 2002.

[233] A. J. Kal, A. J. van Zonneveld, V. Benes, *et al.* Dynamics of gene expression revealed by comparison of serial analysis of gene expression transcript profiles from yeast grown on two different carbon sources. *Mol Biol Cell*, 10(6):1859–1872, Jun 1999.

[234] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*, 18(9):1509–1517, Sep 2008.

[235] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1): 139–140, Jan 2010.

[236] Y. Ge, S. Dudoit, and T. P. Speed. Resampling-based multiple testing for microarray data analysis. *Test*, 12(1):1–77, 2003.

[237] R. R. Delongchamp, J. F. Bowyer, J. J. Chen, and R. L. Kodell. Multiple-testing strategy for analyzing cDNA array data on gene expression. *Biometrics*, 60(3):774–782, Sep 2004.

[238] Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. ISSN 00359246.

[239] R. C. Edgar. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19):2460–2461, Oct 2010.

[240] G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal Complex Systems*, 1695:1–9, 2006. URL `http://igraph.sf.net`.

[241] R. Durbin, S. Eddy, K. A., and M. G. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press, 1998.

[242] R. C. Edgar and S. Batzoglou. Multiple sequence alignment. *Current Opinion in Structural Biology*, 16(3):368–373, 2006. ISSN 0959-440X.

[243] L. Wang and T. Jiang. On the complexity of multiple sequence alignment. *J Comput Biol*, 1(4):337–348, 1994.

[244] P. Hogeweg and B. Hesper. The alignment of sets of sequences and the construction of phyletic trees: an integrated method. *J Mol Evol*, 20(2):175–186, 1984.

[245] C. Kemena and C. Notredame. Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics*, 25(19):2455–2465, Oct 2009.

[246] J.-F. Taly, C. Magis, G. Bussotti, *et al.* Using the T-Coffee package to build multiple sequence alignments of protein, RNA, DNA sequences and 3D structures. *Nat Protoc*, 6 (11):1669–1682, Nov 2011.

[247] M. A. Larkin, G. Blackshields, N. P. Brown, *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21):2947–2948, Nov 2007.

[248] R. C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, 32(5):1792–1797, 2004.

[249] C. Notredame, D. G. Higgins, and J. Heringa. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, 302(1):205–217, Sep 2000.

[250] K. Katoh and D. M. Standley. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*, 30(4):772–780, Apr 2013.

[251] C. B. Do, M. S. P. Mahabhashyam, M. Brudno, and S. Batzoglou. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res*, 15(2):330–340, Feb 2005.

[252] T. Lassmann and E. L. L. Sonnhammer. Kalign, Kalignvu and Mumsa: web servers for multiple sequence alignment. *Nucleic Acids Res*, 34(Web Server issue):W596–W599, Jul 2006.

[253] J. T. Madison, G. A. Everett, and H. Kung. Nucleotide sequence of a yeast tyrosine transfer RNA. *Science*, 153(3735):531–534, Jul 1966.

[254] R. R. Gutell. Comparative studies of RNA: inferring higher-order structure from patterns of sequence variation. *Current Opinion in Structural Biology*, 3(3):313–322, 1993.

[255] S. R. Eddy and R. Durbin. RNA sequence analysis using covariance models. *Nucleic Acids Res*, 22(11):2079–2088, Jun 1994.

[256] S. R. Eddy. A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC Bioinformatics*, 3:18, Jul 2002.

[257] B. Knudsen and J. Hein. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res*, 31(13):3423–3428, Jul 2003.

[258] M. Höchsmann, T. Tller, R. Giegerich, and S. Kurtz. Local similarity in RNA secondary structures. *Proc IEEE Comput Soc Bioinform Conf*, 2:159–168, 2003.

[259] S. Siebert and R. Backofen. MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. *Bioinformatics*, 21(16):3352–3359, Aug 2005.

[260] I. L. Hofacker, S. H. F. Bernhart, and P. F. Stadler. Alignment of RNA base pairing probability matrices. *Bioinformatics*, 20(14):2222–2227, Sep 2004.

[261] D. Sankoff. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM Journal on Applied Mathematics*, 45(5):810–825, 1985.

[262] S. Will, K. Reiche, I. L. Hofacker, P. F. Stadler, and R. Backofen. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput Biol*, 3(4):e65, Apr 2007.

[263] J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7):1105–1119, 1990.

[264] D. Lai, J. R. Proctor, J. Y. A. Zhu, and I. M. Meyer. R-CHIE: a web server and R package for visualizing RNA secondary structures. *Nucleic Acids Res*, 40(12):e95, Jul 2012.

[265] K. P. Schliep. phangorn: phylogenetic analysis in R. *Bioinformatics*, 27(4):592–593, Feb 2011.

[266] S. H. Bernhart, I. L. Hofacker, S. Will, A. R. Gruber, and P. F. Stadler. RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, 9:474, 2008.

[267] E. Zuckerkandl and L. Pauling. Molecular disease, evolution and genetic heterogeneity. *Horizons in Biochemistry, Academic Press, New York*, pages 189–225, 1962.

[268] M. Waterman, T. Smith, M. Singh, and W. Beyer. Additive evolutionary trees. *Journal of Theoretical Biology*, 64(2):199 – 213, 1977. ISSN 0022-5193.

[269] R. R. Sokal and C. D. Michener. *A statistical method for evaluating systematic relationships*. University of Kansas Science Bulletin, 1958.

[270] N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4(4):406–425, Jul 1987.

[271] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. New York: Springer, 2nd edition, 2009.

[272] M. Jiang, J. Anderson, J. Gillespie, and M. Mayne. uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts. *BMC Bioinformatics*, 9:192, 2008.

[273] P. Shannon, A. Markiel, O. Ozier, *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11):2498–2504, Nov 2003.

[274] A. Döring, D. Weese, T. Rausch, and K. Reinert. SeqAn an efficient, generic C++ library for sequence analysis. *BMC Bioinformatics*, 9:11, 2008.

[275] M. Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10–12, 2011. Date accessed: 08 Jul. 2013.

[276] Y. Kong. Btrim: a fast, lightweight adapter and quality trimming program for next-generation sequencing technologies. *Genomics*, 98(2):152–153, 2011.

[277] A. Grishok. Biology and Mechanisms of Short RNAs in Caenorhabditis elegans. *Adv Genet*, 83:1–69, 2013.

[278] C. Dieterich and R. J. Sommer. How to become a parasite - lessons from the genomes of nematodes. *Trends Genet*, 25(5):203–209, May 2009.

[279] R. J. Johnston and O. Hobert. A microRNA controlling left/right neuronal asymmetry in *Caenorhabditis elegans*. *Nature*, 426(6968):845–849, Dec 2003.

[280] N. Liu, K. Okamura, D. Tyler, *et al.* The evolution and functional diversification of animal microRNA genes. *Cell Res*, 18(10):985–996, Oct 2008.

[281] J. Krol, I. Loedige, and W. Filipowicz. The widespread regulation of microRNA biogenesis, function and decay. *Nat Rev Genet*, 11(9):597–610, Sep 2010.

[282] V. Ambros, B. Bartel, D. P. Bartel, *et al.* A uniform system for microRNA annotation. *RNA*, 9(3):277–279, Mar 2003.

[283] M. van Kouwenhove, M. Kedde, and R. Agami. MicroRNA regulation by RNA-binding proteins and its implications for cancer. *Nat Rev Cancer*, 11(9):644–656, Sep 2011.

[284] M. Kato, A. de Lencastre, Z. Pincus, and F. J. Slack. Dynamic expression of small non-coding RNAs, including novel microRNAs and piRNAs/21U-RNAs, during *Caenorhabditis elegans* development. *Genome Biol*, 10(5):R54, 2009.

[285] J. L. Brenner, K. L. Jasiewicz, A. F. Fahley, B. J. Kemp, and A. L. Abbott. Loss of individual microRNAs causes mutant phenotypes in sensitized genetic backgrounds in *C. elegans*. *Curr Biol*, 20(14):1321–1325, Jul 2010.

[286] D. J. Simon, J. M. Madison, A. L. Conery, *et al.* The microRNA miR-1 regulates a MEF-2-dependent retrograde signal at neuromuscular junctions. *Cell*, 133(5):903–915, May 2008.

[287] E. Alvarez-Saavedra and H. R. Horvitz. Many families of *C. elegans* microRNAs are not essential for development or viability. *Curr Biol*, 20(4):367–373, Feb 2010.

[288] W. R. Shaw, J. Armisen, N. J. Lehrbach, and E. A. Miska. The conserved miR-51 microRNA family is redundantly required for embryonic development and pharynx attachment in *Caenorhabditis elegans*. *Genetics*, 185(3):897–905, Jul 2010.

[289] K. Boulias and H. R. Horvitz. The *C. elegans* microRNA *mir-71* acts in neurons to promote germline-mediated longevity through regulation of DAF-16/FOXO. *Cell Metab*, 15(4):439–450, Apr 2012.

[290] Y.-W. Hsieh, C. Chang, and C.-F. Chuang. The microRNA *mir-71* inhibits calcium signaling by targeting the TIR-1/Sarm1 adaptor protein to control stochastic L/R neuronal asymmetry in *C. elegans*. *PLoS Genet*, 8(8):e1002864, Aug 2012.

[291] J. E. Abrahante, A. L. Daul, M. Li, *et al.* The *Caenorhabditis elegans hunchback*-like gene *lin-57/hbl-1* controls developmental time and is regulated by microRNAs. *Dev Cell*, 4(5):625–637, May 2003.

[292] S.-Y. Lin, S. M. Johnson, M. Abraham, *et al.* The *C elegans hunchback* homolog, *hbl-1*, controls temporal patterning and is a probable microRNA target. *Dev Cell*, 4(5):639–650, May 2003.

[293] A. L. Abbott, E. Alvarez-Saavedra, E. A. Miska, *et al.* The *let-7* MicroRNA family members *mir-48*, *mir-84*, and *mir-241* function together to regulate developmental timing in *Caenorhabditis elegans*. *Dev Cell*, 9(3):403–414, Sep 2005.

[294] H. Grosshans, T. Johnson, K. L. Reinert, M. Gerstein, and F. J. Slack. The temporal patterning microRNA *let-7* regulates several transcription factors at the larval to adult transition in *C. elegans*. *Dev Cell*, 8(3):321–330, Mar 2005.

[295] S. M. Johnson, H. Grosshans, J. Shingara, *et al.* *RAS* is regulated by the *let-7* microRNA family. *Cell*, 120(5):635–647, Mar 2005.

[296] N. Martinez, M. Ow, J. Reece-Hoyes, *et al.* Genome-scale spatiotemporal analysis of *Caenorhabditis elegans* microRNA promoter activity. *Genome Res*, 18(12):2005–2015, Dec 2008.

[297] H. Willenbrock, J. Salomon, R. Skilde, *et al.* Quantitative miRNA expression analysis: comparing microarrays with next-generation sequencing. *RNA*, 15(11):2028–2034, Nov 2009.

[298] A. Git, H. Dvinge, M. Salmon-Divon, *et al.* Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression. *RNA*, 16(5):991–1006, May 2010.

[299] S. Lebedeva, M. Jens, K. Theil, *et al.* Transcriptome-wide analysis of regulatory interactions of the RNA-binding protein HuR. *Mol Cell*, 43(3):340–352, Aug 2011.

[300] A. G. Baltz, M. Munschauer, B. Schwanhäusser, *et al.* The mRNA-Bound Proteome and Its Global Occupancy Profile on Protein-Coding Transcripts. *Mol Cell*, 46(5):674–690, Jun 2012.

[301] J. Goecks, A. Nekrutenko, J. Taylor, and G. Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*, 11(8):R86, 2010.

[302] D. Blankenberg, G. V. Kuster, N. Coraor, *et al.* Galaxy: a web-based genome analysis tool for experimentalists. *Curr Protoc Mol Biol*, Chapter 19:Unit 19.10.1–Unit 19.1021, Jan 2010.

[303] B. Giardine, C. Riemer, R. C. Hardison, *et al.* Galaxy: a platform for interactive large-scale genome analysis. *Genome Res*, 15(10):1451–1455, Oct 2005.

[304] Z. Pincus, T. Smith-Vikos, and F. J. Slack. MicroRNA predictors of longevity in *Caenorhabditis elegans*. *PLoS Genet*, 7(9):e1002306, Sep 2011.

[305] J. A. Guerra-Assuno and A. J. Enright. Large-scale analysis of microRNA evolution. *BMC Genomics*, 13:218, 2012.

[306] T. K. K. Kamanu, A. Radovanovic, J. A. C. Archer, and V. B. Bajic. Exploration of miRNA families for hypotheses generation. *Sci Rep*, 3:2940, 2013.

[307] J. Meunier, F. Lemoine, M. Soumillon, *et al.* Birth and expression evolution of mammalian microRNA genes. *Genome Res*, 23(1):34–45, Jan 2013.

[308] F. Rapaport, R. Khanin, Y. Liang, *et al.* Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol*, 14(9):R95, 2013.

[309] Y. Liu, J. Zhou, and K. P. White. RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics*, 30(3):301–304, Feb 2014.

[310] Z. Fang, J. Martin, and Z. Wang. Statistical methods for identifying differentially expressed genes in RNA-Seq experiments. *Cell Biosci*, 2(1):26, 2012.

[311] M. D. Robinson and G. K. Smyth. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, 9(2):321–332, Apr 2008.

[312] Y.-H. Zhou, K. Xia, and F. A. Wright. A powerful and flexible approach to the analysis of RNA sequence count data. *Bioinformatics*, 27(19):2672–2678, Oct 2011.

[313] P. L. Auer and R. W. Doerge. A two-stage Poisson model for testing RNA-seq data. *Statistical applications in genetics and molecular biology*, 10(1):1–26, 2011.

[314] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biol*, 11(10):R106, Oct 2010.

[315] U. Bissels, S. Wild, S. Tomiuk, *et al.* Absolute quantification of microRNAs by using a universal reference. *RNA*, 15(12):2375–2384, Dec 2009.

[316] C. Rödelsperger, A. Streit, and R. J. Sommer. *Structure, Function and Evolution of The Nematode Genome*. John Wiley & Sons, Ltd, Feb 2013. ISBN 9780470015902.

[317] X. Li, J. J. Cassidy, C. A. Reinke, S. Fischboeck, and R. W. Carthew. A microRNA imparts robustness against environmental fluctuation during development. *Cell*, 137(2): 273–282, Apr 2009.

[318] K. S. Kosik. MicroRNAs and cellular phenotypy. *Cell*, 143(1):21–26, Oct 2010.

[319] A. K. L. Leung and P. A. Sharp. MicroRNA functions in stress responses. *Mol Cell*, 40 (2):205–215, Oct 2010.

Abbreviations

E. coli	*Escherichia coli*
P. pacificus	*Pristionchus pacificus*
S. papillosus	*Strongyloides papillosus*
S. ratti	*Strongyloides ratti*
Ago	Argonaute protein
BAC	Bacterial Artificial Chromosome
BWT	Burrows-Wheeler transformation
CLIP	Crosslinking and immunoprecipitation
DA	Steroid hormone dafachronic acid
DE	Differential Expression
DGE	Digital gene expression
DNA	Deoxyribonucleic acid
Daf-c	Dauer-constitutive phenotype
Daf-d	Dauer-defective phenotype
Exp5	Exportin-5
FAR	Flexible Adapter Remover
FDR	False discovery rate
FLEXBAR	Flexible Barcode and Adapter Remover

GA	Genome Analyzer
HGP	Human Genome Project
Kb	Kilobases
LocARNA	Local alignment of RNA
MSA	Multiple sequence alignment
Mb	Megabases
NGS	Next-Generation Sequencing
PCR	Polymerase chain reaction
RBP	RNA-binding protein
RISC	RNA-induced silencing complex
RNA	Ribonucleic acid
RNA-seq	High-throughput RNA sequencing
RNP	Ribonucleoprotein particle
SCFG	Stochastic context-free grammars
SNP	Single Nucleotide Polymorphism
SOLiD	Support Oligonucleotide Ligation Detection
UPGMA	Unweighted pair group method with arithmetic averages
UTR	Untranslated region
WGS	Whole-Genome Sequencing
bp	Base pairs
ddNTPs	Dideoxynucleotides
dsRNA	Double-stranded RNA
iL3	Infective L3 larvae
mRNA	messenger RNA

miRISC	miRNA-induced silencing complex
ncRNA	Non-coding RNA
pre-miRNA	Precursor miRNA
pri-miRNA	Primary miRNA
profile HMM	Profile hidden Markov model
qRT-PCR	Quantitative reverse transcription PCR
rRNA	Ribosomal RNA
small RNA-seq	Small RNA high-throughput sequencing
small ncRNA	Small non-coding RNA
tRNA	Transfer RNA

Summary

The bioinformatics side has become the 'bottleneck' of all high-throughput based biological studies. Next-generation sequencers (NGS) produce millions of sequences (reads) in a short amount of time at low costs. A major problem is the handling and analysis of these large-scale data sets in an efficient and systematic way. Bioinformatics methods can be applied to analyze generated high-throughput sequencing data computationally and therefore help to address biological questions.

This thesis approaches computational challenges and biological questions that arise when investigating microRNA genes (miRNAs) in nematodes using NGS technologies (ABI SOLiD, Illumina GA II, and HiSeq). On the one hand, bioinformatics methods and computational strategies were identified and developed to analyze experimental large-scale small RNA data. These data sets were generated in-house and by collaborators as well as publicly available.
On the other hand, this work addresses the question whether miRNA genes impact developmental arrest and long-term survival in dauer larvae of two free-living nematodes (*Caenorhabditis elegans* (*C. elegans*) and *Pristionchus pacificus* (*P. pacificus*)) and the infective stage of parasites (*Strongyloides ratti* (*S. ratti*)). In particular, I address the long-standing hypothesis that dauer and infective larvae share a common origin. This investigation is specifically focused on determining whether these two larval stages exhibit similar miRNA expression signatures.

In the first part of this study I developed a bioinformatics workflow that characterizes the miRNA gene complement in *C. elegans*, *P. pacificus*, and *S. ratti* and investigates their expression levels. Additionally, this workflow infers miRNA gene families and integrates the observed phylogenetic relationships with measured expression level changes. As part of this study, I was involved in the development of FLEXBAR (published 2012 in the special issue "Next-Generation Sequencing Approaches in Biology", *Biology*[1]),

a program that I applied to preprocess our small RNA sequencing data.

FLEXBAR is a versatile solution for three critical preprocessing steps in any next-generation processing pipeline: (i) basic clipping and quality filtering, (ii) barcode recognition and processing, and (iii) adapter recognition and removal. Importantly, all of these steps can be performed in one program call and executed in parallel. FLEXBAR performs slightly better than FASTX, which is widely considered to be the best of all (selected) competitors in removing adapters from an Illumina read (benchmark I). Furthermore, FLEXBAR covers a large range of sequencing platform applications, formats, and features and provides detailed output statistics, e.g. graphical output of read alignments.

In the second part of this study I applied the bioinformatics workflow to address the question whether miRNAs impact developmental arrest and long term survival in dauer and infective larvae of nematodes (published 2013 in *Genome Biology and Evolution*[2]). This study identifies and extends the number of described miRNA genes to 257 for *C. elegans*, tripled the known gene set for *P. pacificus* to 362 miRNAs, and reports the first miRNAs in a *Strongyloides* parasite, i.e. 106 miRNAs in *S. ratti*. Although our data suggests that miRNA gene sets diverged rapidly in nematodes, my in-depth assessment of miRNAs in free-living and parasitic nematodes revealed conserved miRNA gene families with similar expression signatures in dauer and infective larvae. This finding suggests that common post-transcriptional regulatory mechanisms are at work and that the same miRNA families play important roles in developmental arrest and long-term survival in free-living and parasitic nematodes. Moreover, this result supports the hypothesis that dauer and infective larvae share a common origin.

Taken together, this thesis describes an extensive set of bioinformatic tools and strategies for the analysis of miRNA genes in free-living and parasitic nematodes and constitutes a valuable resource to researchers studying miRNA evolution and in particular, any aspects of developmental arrest. The starting point of this work was the identification of miRNAs in high-throughput small RNA sequencing data profiled by two distinct sequencing platforms. In this context, I provided sophisticated bioinformatic solutions to analyze small RNA sequencing data sets and to address the aforementioned questions computationally.

Zusammenfassung

Seit der Einführung und Etablierung von Next-Generation-Sequenzierern (NGS) hat die Bioinformatik auf dem Gebiet der Genomforschung entscheidend an Bedeutung gewonnen. Mit Hilfe von NGS werden Millionen von DNS-Fragmenten (Reads) innerhalb kürzester Zeit mit sehr geringen Kosten ausgelesen. Das Handling, sowie eine effiziente und systematische Auswertung dieser Hochdurchsatz-Daten, stellt jede biologische Studie vor große Herausforderungen. Durch bioinformatische Methoden wird es möglich gemacht, Hochdurchsatz-Sequenzierungsdaten computergestützt zu analysieren und auszuwerten und somit biologischen Fragestellungen zugänglich zu machen.

Diese Dissertation beschäftigt sich mit den bioinformatischen und biologischen Fragestellungen, die sich bei der Untersuchung von microRNA Genen (miRNAs) in Nematoden mit Hilfe von NGS-Technologien (ABI SOLiD, Illumina GA II, and HiSeq) ergeben. Einerseits wurden moderne computergestützte Ansätze und Methoden aus der Bioinformatik und Statistik angewendet oder eigens entwickelt, um experimentell generierte Hochdurchsatz-Daten von kleinen RNA-Sequenzen auszuwerten. Diese wurden innerhalb der Arbeitsgruppe und von Projektmitarbeitern gemessen oder öffentlich zugänglichen Datensätzen entnommen.
Andererseits wurde der Einfluss von miRNAs auf den Entwicklungsstillstand in Nematoden und auf das langfristige Überleben von Larven im Dauerstadium zweier frei lebender Nematoden (*Caenorhabditis elegans* (*C. elegans*) und *Pristionchus pacificus* (*P. pacificus*)) und Larven im infektiösen Stadium eines Parasiten (*Strongyloides ratti* (*S. ratti*)) untersucht. Ziel war es die langjährige Hypothese zu überprüfen, dass Dauerlarven und infektiöse Larven dieselbe Abstammung hätten. Im Speziellen wurde zu diesem Zweck untersucht, ob diese beiden Larvenstadien ähnliche miRNA Expressionsmuster aufweisen.

Im ersten Teil dieser Studie habe ich einen Ansatz für eine rechnergestützt systematische Auswertung entwickelt, mit dessen Hilfe das miRNA Repertoire von *C. elegans*, *P. pacificus*, und *S. ratti* bestimmt und ergänzt wurde und deren Expression ausgewertet werden konnte. Außerdem wurden auf diese Weise miRNA-Genfamilien hergeleitet und deren phylogenetische Abhängigkeiten mit den gemessenen Genexpressionsveränderungen in Zusammenhang gebracht. Im Rahmen dieser Studie war ich an der Entwicklung von FLEXBAR (veröffentlicht 2012 in einer Spezialausgabe von „Next-Generation Sequencing Approaches in Biology", *Biology*[1]) beteiligt, ein Programm, das ich zum Vorverarbeiten von unseren NGS-Datensätzen eingesetzt habe.

FLEXBAR ist ein vielseitiges Programm, das für drei wichtige Vorverarbeitungsschritte in NGS-Experimenten angewandt werden kann: einfaches Kürzen von NGS-Reads und Qualitätskontrolle, Barcodeerkennung und -verarbeitung, Adaptererkennung und -entfernung. Eine wesentliche Eigenschaft von FLEXBAR ist es, all diese Verarbeitungsschritte in einem Programmaufruf und parallelisiert auszuführen. Die Benchmark-Tests zeigen, dass FLEXBAR etwas bessere Ergebnisse liefert als FASTX, ein häufig angewendetes Programm zum Entfernen von Adaptersequenzen in Illumina-Reads (Benchmark-Test I). Darüber hinaus kann FLEXBAR mit den verschiedensten Sequenziertechnologie-Anwendungen, Dateiformaten und Eigenschaften umgehen und liefert zudem detaillierte Ausgabestatistiken wie beispielsweise eine grafische Ausgabe von Sequenzalignments.

Im zweiten Teil dieser Studie wende ich die zuvor entwickelten bioinformatischen Methoden und Strategien an, um meine biologischen Fragen hinsichtlich der Auswirkung von miRNAs in Dauer und in infektiösen Larvenstadien von Nematoden zu untersuchen (veröffentlicht 2013 in *Genome Biology and Evolution*[2]). Die Auswertung unserer Hochdurchsatz-Daten zeigt, dass die bereits bekannten miRNA Gensätze in *C. elegans* und *P. pacificus* zuverlässig identifiziert und mit neuen zuvor unbekannten Genen ergänzt werden konnten. Die Anzahl der bereits beschriebenen Gene von *C. elegans* wurde auf insgesamt 257 miRNAs erhöht und diejenigen von *P. pacificus* auf 362 miRNAs verdreifacht. Außerdem konnten mit der Untersuchung von *S. ratti* erstmals 106 miRNAs eines *Strongyloides* Parasiten veröffentlicht werden. Obwohl unsere Daten darauf hinweisen, dass miRNA Gene in Nematoden evolutiv schnell divergieren, konnte meine tiefgehende Analyse von miRNAs in frei lebenden und parasitären Nematoden konservierte miRNA-Genfamilien mit ähnlichen Expressionsmustern in Dauer und in infektiösen Larven aufdecken. Dieses Ergebnis weist darauf hin, dass ähnliche post-transkriptionelle regulatorische Mechanismen in Dauer und in infektiösen Larven wirken

und dass dieselben Genfamilien für deren Entwicklungsstillstand und langfristiges Überleben eine wichtige Rolle spielen. Zudem stützt dieses Resultat die oben genannte Hypothese, dass Dauerlarven und infektiöse Larven möglicherweise dieselbe Abstammung haben.

Zusammenfassend liefert diese Dissertation eine umfangreiche Darstellung von bioinformatischen Analysewerkzeugen und Strategien für die Auswertung von miRNAs in frei lebenden und parasitären Nematoden. Sie stellt somit eine wertvoll Quelle dar für Forscher, die sich mit miRNA-Evolution und speziell mit allen Aspekten des Entwicklungsstillstandes beschäftigen. Der Ausgangspunkt dieser Arbeit war die Identifikation von miRNAs in Hochdurchsatz-Sequenzierdaten, die mittels zwei verschiedenen NGS-Technologien erzeugt wurden. In diesem Zusammenhang habe ich bioinformatische Analysestrategien entwickelt, um die Sequenzierdaten von kleinen RNAs auszuwerten und die bereits erwähnten biologischen Fragen rechnergestützt zu untersuchen.

Appendix A - Supplemental Material

Supplemental Material includes:

Supplemental Methods

Supplemental Figures A.1

Supplemental Tables A.1

Supplemental Methods

miRNA Age Classes

miRNA genes were grouped into five different age classes based on all miRNAs in miRBase (v18) that were annotated as Nematoda, Arthropoda, Lochotrophozoa or Vertebrata. For example, if a seed (position 2-8) of miRNA X perfectly matched a seed from miRNA Y classified as Vertebrata, miRNA X would be defined as being conserved in Vertebrata. Assuming this miRNA X is only conserved in Vertebrata, the conservation signature '- / - / - / +' (Nematoda / Arthropoda / Lochotrophozoa / Vertebrata) would be assigned to miRNA X. miRNA age categories were defined as follows: (i) Not conserved: '- / - / - / -'; (ii) very young: '+ / - / - / -', '- / + / - / -'; (iii) young: '+ / + / - / -'; (iv) middle: '- / + / - / +', '+ / - / - / +', '- / + / + / -', '+ / - / + / -', '- / - / + / +', '- / - / + / -', '- / - / - / +'; (v) old: '+ / + / + / -', '- / + / + / +', '+ / - / + / +' , '+ / + / - / +', '+ / + / + / +'.

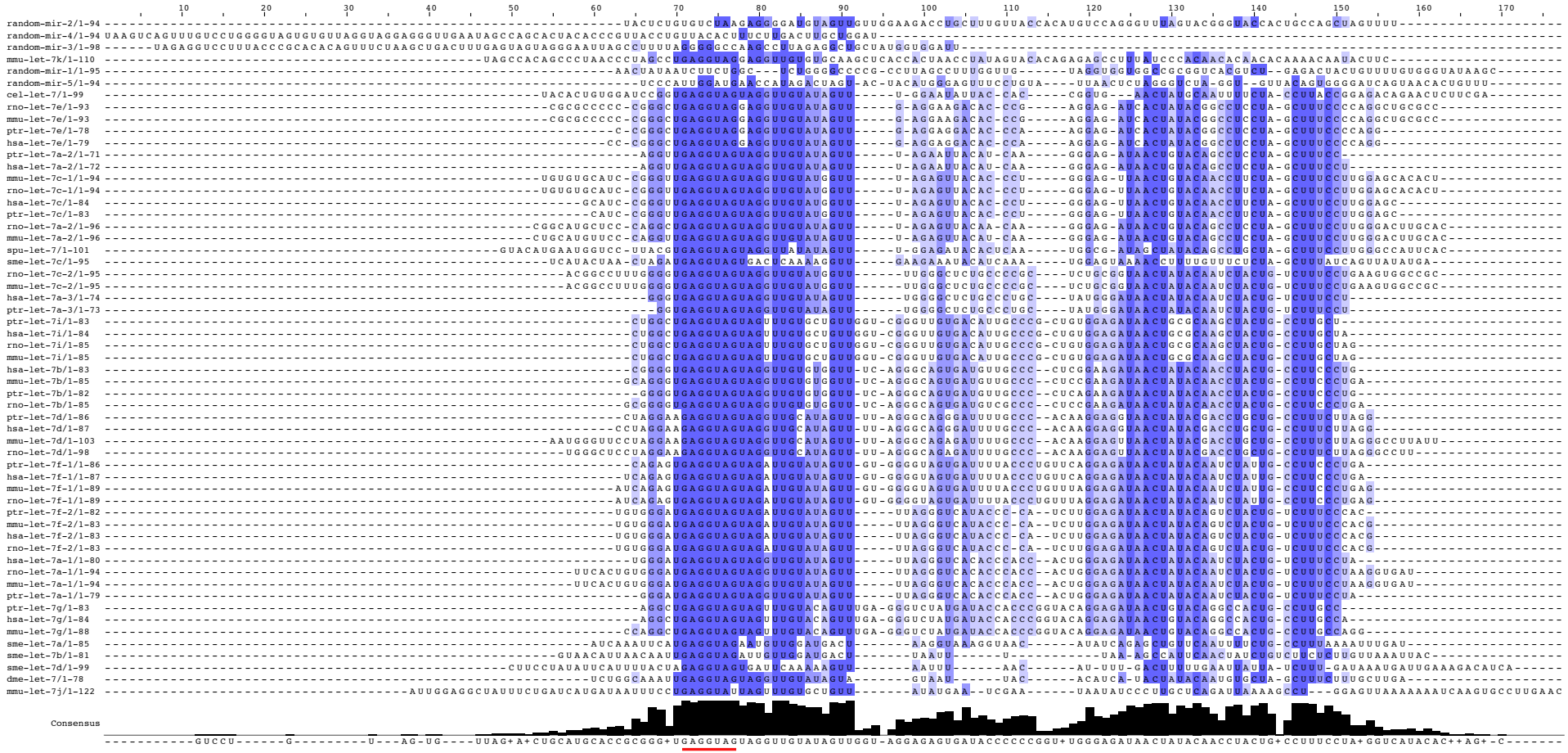

Figure A.1: MSA of *let-7* family miRNAs from eight animal clades and five random generated precursors
The multiple alignment was computed by LocARNA [262] and constrained to align at the seed (defined as nucleotide position 2-8) of each individual miRNA. Nucleotides are colored based on a percentage identity threshold; i.e. nucleotides that occur in a particular column more than 80% are colored in mid blue, more than 60% in light blue, more than 40% in light grey, and white otherwise. The seed sequence is marked with a red underline.

Table A.1: Conserved miRNAs with coherent expression signature in developmentally arrested stages
Four gene families that contain miRNAs conserved among all three nematode species with coherent expression signature in developmentally arrested stages (upregulation and downregulation). A miRNA family displays a coherent expression signature if at least one family member from each species is differentially expressed as the majority of family members. The seed conservation column depicts IDs of miRNAs that share a common seed within the Nematoda taxa. If the seed is not conserved in Nematoda the ID from the next taxa is shown (in the order of Nematoda, Arthropoda, Lochotrophozoa, and Vertebrata). Note that relative expression changes of some miRNAs of family 4 and 30 were not measured.

miRNA family ID	**Corresponding seed**	**Seed conservation**[†]	**Seed conservation profile**[†]	**Up**	**Down**	**#Precursor**	**Observed function**
miRNA family 4	GGAAUGU	*mir-1/-796*	+ / + / + / +	4	0	6	Synaptic transmission [286]
miRNA family 30	AUCACUA	*mir-71*	+ / - / - / +	3	0	5	Lifespan, AWC L/R neuron fate specification [123,289,290]
miRNA family 32	CACCGGG	*mir-35-42/-2235/-2240/-2251/-8232/-8243/-8283/-8393*	+ / + / + / +	0	32	36	Embryogenesis [287]
miRNA family 22	ACUGGCC	*mir-240*	+ / + / + / +	0	3	3	Defecation cycling, fertility [285]

[†]Nematoda/Arthropoda/Lophotrochozoa/Vertebrata

Appendix B - Supplemental CD

Table of Contents of CD

Supplemental Figures B.1 − B.3

Supplemental Tables B.1 − B.6

Figure Legends for Supplemental Figures

Figure B.1: Multiple sequence alignments for miRNAs grouped by perfect seed sequence similarity (position 2-8) computed by LocARNA[262]. These precursors correspond to miRNA arms that were selected based on largest group size or degree of conservation, respectively. All alignments were constrained to align at the seed sequence position of each miRNA.

Figure B.2: Multiple sequence alignments for miRNA families (at least two miRNA genes) computed by LocARNA[262]. These precursors correspond to miRNA arms that were selected based on largest group size or degree of conservation, respectively. All alignments were constrained to align at the seed sequence position of each miRNA.

Figure B.3: Visualization of miRNA expression fold changes for individual miRNA families that contained at least two precursor sequences. Heatmaps represent miRNA gene expression by color, and heatmap rows are ordered by the inferred phylogeny from multiple sequence alignments. The asterisk next to the miRNA in the heatmap plot denotes significant up- or downregulation in dauer or iL3.

Table Legends for Supplemental Tables

Table B.1: The two worksheets summarize the number of reads that mapped to various genomic feature annotations of the small RNA deep sequencing libraries from *C. elegans* and *P. pacificus* analyzed in this study.

Table B.2: Annotation of known miRNAs in *C. elegans* and *P. pacificus*. The first worksheet tabulates all miRNAs in *C. elegans* (miRBase v18) including seed and conservation information for both miRNA arms. The second worksheet includes the equivalent information for *P. pacificus*.

Table B.3: Novel miRNA gene candidates in *C. elegans*, *P. pacificus*, and *S. ratti*.

Table B.4: The first worksheet tabulates miRNA read counts and ratios of *C. elegans*. The second worksheet includes the equivalent information for *P. pacificus*. (Note: A pseudocount of one was added to all miRNA arms.)

Table B.5: miRNA families that are conserved among all three nematode species. These homology relationships were established based on all-against-all sequence similarity searches using USEARCH[239].

Table B.6: miRNA gene expression in *C. elegans*, *P. pacificus*, and *S. ratti*. The three different worksheets present read counts of miRNAs detected at least five times in both stages (mixed-stage/dauer or mixed-stage/iL3), including normalized read counts (reference based quantile normalization), $\log_2$ fold expression changes, FDR, and differential expression categories.

Curriculum Vitae

For reasons of data protection, the curriculum vitae is not included in the online version.

For reasons of data protection, the curriculum vitae is not included in the online version.

For reasons of data protection, the curriculum vitae is not included in the online version.

For reasons of data protection, the curriculum vitae is not included in the online version.

Selbständigkeitserklärung

Hiermit erkläre ich, dass ich diese Arbeit selbstständig verfasst habe und keine anderen als die angegebenen Quellen und Hilfsmittel in Anspruch genommen habe. Ich versichere, dass diese Arbeit in dieser oder anderer Form keiner anderen Prüfungsbehörde vorgelegt wurde.

Berlin, November 2014