JUSTUS-LIEBIG-
UNIVERSITÄT
GIESSEN

Elucidating the potential of microRNAs

-

towards a functional landscape of microRNAs
in the model organisms *Tribolium castaneum* and *Galleria mellonella*

INAUGURALDISSERTATION

zur Erlangung des akademischen Grades

- Dr. rer. nat. -

der Naturwissenschaftlichen Fachbereiche
der Justus-Liebig-Universität Gießen

vorgelegt von

Daniel Amsel

Master of Science Bioinformatik

2020

Die vorliegende Arbeit wurde von April 2015 bis April 2018 im Fraunhofer-Institut für Molekularbiologie und Angewandte Ökologie (IME-Br) in Gießen unter der Leitung von Prof. Dr. Andreas Vilcinskas und Prof. Dr. Alexander Goesmann angefertigt. Die Promotion wurde in diesem Zeitraum durch das LOEWE-Zentrum für Insektenbiotechnologie & Bioressourcen, sowie durch Dow AgroSciences finanziert.

Erster Gutachter
Prof. Dr. Andreas Vilcinskas

Fachbereich 09 - Agrarwissenschaften, Ökotrophologie und Umweltmanagement

Institut für Insektenbiotechnologie
AG Angewandte Entomologie

Justus-Liebig-Universität Gießen

Zweiter Gutachter
Prof. Dr. Alexander Goesmann

Fachbereich 08 - Biologie und Chemie

Institut für Systembiologie

Justus-Liebig-Universität Gießen

to my grandfather

Hans Marschall

LIST OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

- antimicrobial peptides (AMPs)
- argonaute-crosslinking and immunoprecipitation (AGO-CLIP)
- asymptomatic bacteriuria (ABU)
- coding sequence (CDS)
- counts per million (CPM)
- cross-linking immunoprecipitation (CLIP)
- cross-linking immunoprecipitation with high-throughput sequencing (CLIP-seq)
- false negatives (FN)
- false positives (FP)
- fragments per kilobase million (FPKM)
- high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation (HITS-CLIP)
- histone acetyltransferases (HATs)
- histone deacetylases (HDACs)
- individual-nucleotide resolution crosslinking immunoprecipitation (iCLIP)
- LPS-induced tumor necrosis alpha factor (LITAF)
- messenger RNAs (mRNAs)
- microRNA (miRNA)
- miRNA isoforms (isomiRs)
- National Center for Biotechnology Information (NCBI)
- photoactivatable ribonucleoside-enhanced crosslinking immunoprecipitation (PAR-CLIP)
- precursor-miRNA (pre-miRNA)
- primary-miRNA (pri-miRNA)
- reads per kilobase million (RPKM)
- reads per million (RPM)
- RNA induced silencing complex (RISC)
- RNA interference (RNAi)
- single nucleotide polymorphism (SNP)
- small interfering RNA (siRNA)

- Toll-like receptor 4 (TLR4)
- true negatives (TN)
- true positives (TP)
- untranslated region (UTR)
- urinary tract infections (UTIs)
- uropathogenic *Escherichia coli* (UPEC)
- zero-truncated negative binomial (ZTNB)

ABSTRACT

[ENGLISH]

Insects have been utilized by humans for thousands of years. At first in a very primitive way as direct food provider for example for honey or as material source like lac dye or silk. Nowadays, with emerging technical possibilities, resources and needs, the usage of insects, as the world's largest animal class, is getting much more skillful.

Technologies like Next Generation Sequencing emerged in the past decades, enabling a broad scientific community to study organisms, like insects, at a molecular genetic level for a comparable cheap price. This technological jump led to findings that promoted insects as valuable model organisms and resource of molecules for various applications. For example, the study of effects of pathogens to an insect organism and a potential transfer of knowledge to humans.

One of the emerging fields in the molecular genetic research is the investigation of gene expression regulation by microRNAs. Since their misregulation can lead to serious malformations of body parts, cancer and even death, the large research community is diversified and the numbers of microRNA publications is increasing to nearly 18,000 per year. Despite their popular impact, the *in silico* analysis of microRNAs needs the knowledge of a broad portfolio of task-specific tools and settings in order to compute results. Furthermore, due to their small size of only around 21 nucleotides, the assignment of a corresponding mRNA is a non-trivial task, leading to a broad variety of approaches in the laboratory and *in silico*. On the one hand, the target prediction algorithms mostly try to generalize microRNA behavior and filter for microRNA-mRNA bindings with these features. This leads to a large number of false-positives. On the other hand, many of the laboratory approaches are not high-throughput, making a validation of microRNA targets time consuming. With Next Generation Sequencing, certain methods (like CLIP-sequencing) are able to raise signals of microRNA binding sites on mRNAs in a high-throughput manner. However, these methods have the drawback that they are very difficult to perform. Datasets are therefore very rare in public databases.

Within this thesis I present the assessment of a best practice workflow for microRNA analysis, providing a guideline for other researchers. My benchmarking of microRNA isoform tools not only identified the most suitable tool for high-throughput pipelines, but also highlighted the broader impact on microRNA isoforms in *Tribolium castaneum* early development. Outgoing from the assessment, I created a scripted workflow which was then translated into a novel pipeline, called `microPIECE`, covering the widely used microRNA analysis tasks, including microRNA isoform detection, combined with a novel microRNA target prediction approach that relies on transferred and evolutionary conserved CLIP-sequencing data from closely related species. This approach makes the previously mentioned rare data available to other species.

Finally, I applied my pipeline exemplarily to *Galleria mellonella* in order to identify the impact of microRNAs to the immune response against pathogenic *Escherichia coli* strains, indicating a benefit for human pathogen investigations and shedding light on a potential insect utilization for humans.

The results of my research are publicly available in three different scientific journals (BMC Bioinformatics, JOSS- Journal of Open Source Software and Nature Scientific Reports). The source code of the `microPIECE` pipeline, as well as a `Docker` environment is available via `GitHub.com`.

[DEUTSCH]

Insekten wurden bereits vor Tausenden von Jahren vom Menschen genutzt. Zuerst nur auf eine sehr primitive Art und Weise als Nahrungsquelle für beispielsweise Honig oder als Quelle für Material, wie beispielsweise Schellack oder Seide. Heutzutage, mit voranschreitenden technischen Möglichkeiten, Ressourcen und Methoden, wird die Nutzbarkeit von Insekten, als weltweit größte Tierklasse, um einiges anspruchsvoller.

Technologien wie Next Generation Sequencing kamen im Laufe der letzten Dekaden hinzu und ermöglichten einer breiten wissenschaftlichen Community die Erforschung von Organismen, wie beispielsweise Insekten, auf einer molekulargenetischen Ebene, zu einem vergleichsweisen günstigen Preis. Dies führte zu Entdeckungen, die Insekten als wertvolle Modell-Organismen qualifizierte. Beispielsweise konnte die molekulargenetische Auswirkung von Pathogenen auf den Organismus von Insekten untersucht werden, deren Ergebnisse Rückschlüsse auf die Funktion im Menschen ermöglichen.

Eines der aufstrebenden Bereiche in der Molekulargenetik ist die Erforschung von Genexpressions-Regulation durch microRNAs. Eine Fehlregulation dieser microRNAs kann zu schwerwiegenden Fehlbildungen von Körperteilen, Krebs und sogar zum Tode führen. Dementsprechend divers ist auch die Forschungs-Community und die Zahl der microRNA Publikationen steigt auf fast 18.000 Artikel pro Jahr. Trotz ihres populären Einflusses ist die *in silico* Analyse von microRNAs immer noch anspruchsvoll und benötigt eine große Anzahl an spezifischen Programmen und Einstellungen zur Auswertung der Rohdaten. Weiterhin ist die Bestimmung von mRNA Zielgenen aufgrund der recht kleinen Größe von durchschnittlich 21 Nukleotiden der microRNAs keine einfache Aufgabe. Dies führte zu einer Vielzahl an Lösungsansätzen, sowohl im Labor, als auch *in silico*. Auf der einen Seite versuchen sogenannte Target-Prediction-Algorithmen das Verhalten von microRNAs zu generalisieren und nach microRNA-mRNA Bindungen mit diesen Eigenschaften zu filtern. Das führt zu einer großen Zahl von Falsch-Positiven. Auf der anderen Seite sind viele der im Labor eingesetzten Methoden zur Validierung nicht für den Hochdurchsatz geeignet, was die Validierung der Vorhersagen zeitaufwändig macht. Mit Hilfe von Next Generation Sequencing wurden Methoden (wie bspw. CLIP-Sequenzierung) etabliert, die es ermöglichen microRNA Bindestellen im Hochdurchsatz zu identifizieren. Jedoch haben diese Methoden den Nachteil, dass sie schwierig umzusetzen sind und ihre Datensätze deshalb auch selten in öffentlich Datenbanken zu finden sind.

In dieser Thesis präsentiere ich das Assessment eines Best-Practice Workflows für die microRNA Analyse, die als Leitfaden für andere Wissenschaftler (m/w/d) dienen kann. Mein Benchmark von microRNA Isoform Analyse-Programmen hat nicht nur das beste Tool für die Hochdurchsatz Analyse identifiziert, es hat auch den Einfluss von microRNA Isoformen im frühen Entwicklungsstadium von *Tribolium castaneum* beleuchtet. Ausgehend von diesem Assessment habe ich einen Workflow mit einzelnen Skripten entwickelt, der wiederum in eine zusammenhängende Pipeline mit dem Namen `microPIECE` konvertiert wurde. Sie deckt die am weitesten verwendeten microRNA Analysen ab, inklusive microRNA Isoform Detektion, in Kombination mit meinem neu entwickelten Zielvorhersage-Ansatz, welcher auf den Transfer von evolutionär konservierten CLIP-Sequenzierdaten von nah verwandten Spezies setzt. Diese Methode macht die vorher erwähnten, seltenen Datensätze auch für andere Spezies nutzbar.

Schließlich habe ich meine Pipeline exemplarisch an *Galleria mellonella* angewendet, um den Einfluss von microRNAs auf die Immunantwort gegenüber pathogenen *Escherichia coli* Stämmen zu identifizieren und eine mögliche Übertragbarkeit auf den Menschen herzustellen.

Die Ergebnisse meiner Forschungen sind öffentlich verfügbar und in drei verschiedenen wissenschaftlichen Journalen erschienen (BMC Bioinformatics, JOSS – Journal of Open Source Software und Nature Scientific Reports). Der Quellcode der `microPIECE` Pipeline, sowie die `Docker`-Umgebung sind auf `Github.com` verfügbar.

PREFACE

This thesis contains three already published articles in three different journals where I am two times the first author and one time a shared first author.

These publications mirror the main results of my work. In order to highlight the articles in the text, I reformatted them to have a two-column manuscript style. Furthermore, each publication is written in an individual chapter, namely chapter 4, chapter 6 and chapter 8.

In order to evaluate my achievements in each article, I included my individual contributions and gathered the signatures of all my co-authors mirroring their agreement on those contributions. These reports are appended in chapter 14.

1

INTRODUCTION

When we take a look at our human history, we notice that our species started to observe insects and tried to utilize them for the production of consumable and non-consumable goods a long time ago. One very old contemporary witness is a cave painting of a collector for wild honey in Spain that is estimated to be 6,000 to 10,000 years old. It shows a human being, climbing a tree, having one hand in a tree hole and the other hand holds a receiver (Crane 2005). Findings from ancient cultures of South Asia report the usage of silk that was harvested from caterpillars about 5,000 years ago (I. Good, Kenoyer, and Meadow 2008). Around 3,000 years ago in India, people wrote about a louse, the *Kerria lacca*, which produces a medical substance, known as lac. Nowadays, lac is mostly known as dye or as the ancestor of the vinyl record, the gramophone record. Mankind further learned that insects are important for pollination of plants. Discoveries a few years ago showed that a diverse bandwidth of different pollination insects leads to larger fruits compared to wind and self-pollination (Klatt et al. 2014; Abrol et al. 2017). Furthermore, the increasing world population and the need for food leads to the shift in viewing insects also as a common source of protein (Dossey, Tatum, and McGill 2016).

But besides those more or less obvious insights, the perspective of how mankind looks at insects has expanded from an observer and harvester to an investigator and developer. Recent developments of novel technologies, like Next Generation Sequencing and accompanying findings have enabled previously unknown possibilities for the beneficial use of insects. Foremost, one can name the possibility for the exploration of the molecular genetic repertoire of an animal in a large and cost-efficient manner and the transfer of genetic knowledge to other species. For example, a recent study uses insects as a bioresource for plant protection by transgenic expression of insect peptides in crops, leading to an increased fungal and bacterial resistance (Vilcinskas and Gross 2005). Another major research focus is human medicine,

where antimicrobial peptides from *Lucilia sericata* were found to show a therapeutic effect on wounds (Pöppel et al. 2015).

Most importantly, the number of insects that can be investigated is not limited anymore to the small group of obviously beneficial insects, but is now open to every insect that can be investigated in a laboratory. It is widely accepted that the insects are the most diverse animal group on earth and that many of them are still unknown. Nevertheless, the estimations of sizes vary greatly. One of the opinions stated that there are around one million species belonging to the class of insects (Baillie, Hilton-Taylor, and Stuart 2004). Other estimations range up to eight million species (Groombridge et al. 2002). This indicates that there are still many previously disregarded insects that could harbor useful molecules that can be investigated with the novel techniques and methods under the scope of molecular genetics.

1.1 Molecular Genetics And Epigenetic Gene Regulation In Animals

Due to the technical improvements in the last decades, the field of molecular genetics is rapidly developing and the number of animals whose genetic makeup is decoded and publicly available is increasing constantly. Researchers all over the world investigate the function of the diverse molecules that evolved during evolution. Some of these functions can be highly specific to certain species or clades in combination with their habitat. For example, the burying beetle can produce a secretion that preserves cadaver from decay (Heise et al. 2016). Additionally, it is important to know, how such extraordinary skills work on a molecular level and how the genes are regulated in order to identify molecular pathways that lead to discovery of new model organisms or drugs for medical purposes.

Gene regulation summarizes the mechanisms in the cells that control the production and amount of gene products. In unicellular organisms, gene regulation is essential to adapt to external factors, like food availability or temperature changes and is mainly regulated on the transcriptional level. In 1961, François Jacob and Jacques Monod discovered the operon mechanism in *Escherichia coli*. These operons can be categorized into positive and negative regulating operons. For the positive regulation, a certain activator promotes the regulated genes to be transcribed, whereas for the negative regulation, the binding of a repressor prevents the expression of the regulated genes. The operon model is robustly established in prokaryotes and seems efficient for the survival of a single cell.

In contrast, the eukaryotic cell needs far more regulating variability, since the gene regulation in eukaryotes is crucial for the differentiation of cells into highly specific compartments of the body with different gene expression profiles. Therefore, the eukaryotic gene expression is

controlled at many different timepoints during the processing from the gene to the gene product, which is in most cases, and in the following example, a protein. It begins with the chromatin accessibility, which means that specific genomic regions are more or less easily accessible for the transcriptional machinery than others (Klemm, Shipony, and Greenleaf 2019). The transcription itself is regulated by transcription factors that bind to regulatory DNA sequences next to a certain gene to enhance or repress the transcription (Villar, Flicek, and Odom 2014). The transcribed gene is then processed by splicing, capping and poly-A tailing, resulting in a mature messenger RNA (mRNA). In this step, different transcripts can be produced from the same locus for example by alternative splicing, resulting in different transcript concentrations in different cells. Furthermore, the stability of the transcript plays an important role in gene regulation, as well as its accessibility for the translational machinery. A recently emerging molecule, regulating the stability and accessibility of transcripts, is a small non-coding RNA called microRNA. The family of microRNAs plays a major role in the efficient and fast-responding gene regulation in cells. Moreover, it can be diagnostically used as a disease-, drug- or infection-indicating biomarker or even as a medical therapy. For these issues, insects could function as model organisms.

1.2 The Discovery Of microRNAs

The first microRNA (miRNA) was discovered in 1993 in *Caenorhabditis elegans* as postembryonic downregulation mechanism of LIN-14, targeting the 3' untranslated region (UTR) on the mRNA with a base pair binding of the first eight nucleotides (R. C. Lee, Feinbaum, and Ambros 1993). Seven years later, the let-7 miRNA was found in *C. elegans* (Ruvkun et al. 2000), but it took another year until those short sequences were initially termed "microRNAs" (Lau et al. 2001; Lagos-Quintana et al. 2001). Nevertheless, the importance of miRNAs in the organism was already observed, since a dysregulation of the let-7 miRNA in *C. elegans* led to lethal effects in the worm (Reinhart et al. 2000). In December 2002, `miRbase.org v1` started as a central resource for known miRNAs with a total of 218 entries. This included only *Homo sapiens* (56 entries), *Caenorhabditis elegans* (59 entries), *Mus musculus* (41 entries), *Drosophila melanogaster* (21 entries) and *Arabidopsis thaliana* (41 entries). The recent major release describes almost 40,000 miRNA entries on `miRBase.org v22.1` (released October, 2018) for a total of 285 organisms, including animals, plants and viruses. In the early phase of miRNA research, novel sequences were identified by explicit testing with dedicated methods, like the northern blot (Lagos-Quintana et al. 2001). The comparative expression levels of miRNAs were examined by micro-array approaches (C.-G.

Liu et al. 2008). Then in 2009, a novel technique emerged in the field, the small RNA sequencing and the identification of novel miRNAs and their individual expression in certain species or conditions became high-throughput (Wyman et al. 2009; H. Zhang et al. 2009). This enabled a rapid growth of miRNA entries in `miRBase.org` for a broad variety of species, providing a large landscape of miRNA evolution and conservation between species.

The identification of microRNAs with sequencing data is error prone, but one layer of confidence is gained from criteria derived from microRNA biogenesis. The maturation of a microRNA is different, compared to other ncRNAs and can therefore be used as classification. This helps to distinguish microRNAs from other small non-coding RNAs or fragments of larger non-coding RNAs.

1.3 The microRNA Biogenesis

A mature miRNA usually consists of about 21 nucleotides of single-stranded RNA that interacts with an Argonaut protein to form an RNA induced silencing complex (RISC), which then targets a cytoplasmic mRNA by sequence specific binding (Bartel 2004).

The mature miRNA sequence is derived from the stem of a hairpin structure, called the precursor-miRNA (pre-miRNA) (Bartel 2004). In some certain cases it was reported that one precursor can have adjacent mature miRNAs, next to the two regular mature miRNAs, the so-called offset-miRNAs (Shi et al. 2009). The precursor miRNAs are usually part of the primary-miRNA (pri-miRNA), a several kilobase long RNA transcript that can consist of at least one pre-miRNA (Figure 1 (1)) (Y. Lee et al. 2004; Cai, Hagedorn, and Cullen 2004). Some pre-miRNAs are then transcribed together as a so-called polycistronic cluster (Y. Lee et al. 2002). In some literature, the maximum allowed cluster distance between pre-miRNAs is defined by 1 kb whereas elsewhere, a distance of 50 kb is still considered as a cluster (Chan et al. 2012). The transcription is similar to mRNAs regarding the RNA polymerase II usage, the polyadenylation of the 3'-end and the 5'-end cap structure (Y. Lee et al. 2004). The regulation of miRNAs is in many cases controlled by promoters, as it is already known for protein-coding genes (Y. Lee et al. 2004). In one study, 175 human miRNAs were identified by chromatin structure analysis that were having proximal promoters (Ozsolak et al. 2008).

The miRNA precursors are recognized by their hairpin structure and are sliced out from the pri-miRNA by a protein called Drosha, an RNase III enzyme, together with DGCR8 (known as Pasha in *D. melanogaster* or *C. elegans*), a microprocessor subunit, within the nucleus (Figure 1 (2)). Initially observed in *D. melanogaster* (Ruby, Jan, and Bartel 2007), pre-miRNAs can also be derived directly from protein-coding introns by splicing activity, instead of being

transcribed separately and processed by Drosha, known as mirtrons (Wen et al. 2015). The processed pre-miRNA then has a hairpin structure with a length of about 70 bp (Y. Lee et al. 2003). The pre-miRNA is afterwards exported into the cytoplasm by a complex of RAN-GTP and Exportin 5 (Figure 1 (3)) (Bohnsack, Czaplinski, and Gorlich 2004). Once in the cytoplasm, the head of the hairpin pre-miRNA is cut off by a protein called Dicer (Figure 1 (4)) (Hutvágner et al. 2001). The stem decays into the mature miRNA sequences that are then distinguished as miR-3p, marking the 3p-end of the pre-miRNA, and miR-5p, marking the 5p-end of the pre-miRNA (Figure 1 (5)). From those two, mostly only one specific arm is used, whereas the other one, also called the miR* sequence, is degraded (Sam Griffiths-Jones 2004). In early investigations it was assumed that the selection of an arm is determined by thermodynamics or structural characteristics (Schwarz et al. 2003; Khvorova, Reynolds, and Jayasena 2003). But it has been shown that even the miR* sequence may have regulatory effects (Okamura et al. 2008). Further studies from large miRNA sequencing approaches revealed that the preference for one arm from a hairpin is dependent on the tissue, development stage or switched during evolution, like the miR-10 sequences from *T. castaneum* and *D. melanogaster* (Sam Griffiths-Jones et al. 2011; Jagadeeswaran et al. 2010; Ro et al. 2007).The Argonaut protein, as a main actor, binds to the mature miRNA sequence and forms a RISC (Figure 1 (6)). This complex then targets mRNA sequences in the cytoplasm (Figure 1 (7)) (V. N. Kim, Han, and Siomi 2009; Bernstein et al. 2001).

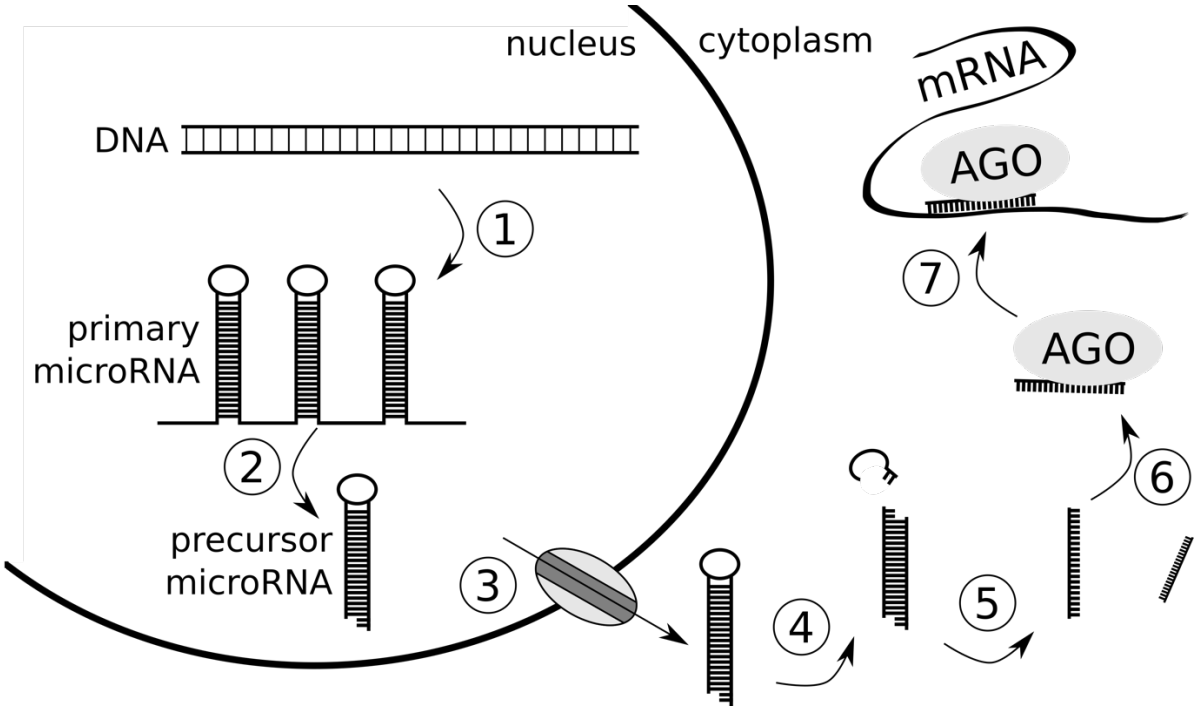


Figure 1 **The microRNA biogenesis** The primary microRNA is transcribed from the DNA (1), from where the precursor microRNA is cut out by Drosha and DGCR8/Pasha (2). The hairpin structured precursor microRNA is then exported into the cytoplasm by Exportin-5 and Ran-GTP (3). In the cytoplasm, the precursor is cut into a double stranded RNA by Dicer (4), followed by an unwinding of the double strand, leading to single strands (5). In most cases, one strand is now degraded and the other one binds to an AGO protein, forming a RISC (6). This complex then binds mRNA sequences via complementary base pairing of the miRNA and regulates the translation (7).

The cleavage process of the miRNA is believed to be a source for miRNA isoforms (isomiRs) (Morin et al. 2008), where the canonical mature miRNA contains one or more modifications due to imperfect cleavage (Kuchenbauer et al. 2008). A 5'-end modification that elongates or shortens the canonical mature sequence may result in a shift of the seed region and therefore in a change of the mRNA target (Tan et al. 2014a). Further types of isomiRs are derived from nucleotidyltransferases that append non-templated nucleotides to the 3'-end of the mature miRNA (Wyman et al. 2011), but also from nucleotide editing (Luciano et al. 2004; G. Sun et al. 2009). In early studies, those isoforms were seen as sequencing artefacts, but with increased technical credibility of minor expressed sequences, they emerged to a separate functional product with specific tasks within the cell (L. W. Lee et al. 2010). In the scope of evolution and diversification, it has been observed that miRNAs are more often gained than lost (Chang et al. 2016). This evolutionary concept is picked up in many algorithms, in order to assign mRNA targets to the miRNAs in order to elucidate their function in the organism. The determination

of the miRNA targets is one of the most important, but also one of the most difficult tasks, when working with microRNAs.

1.4 Identification of microRNA targets

Starting with the investigations of lin-4 and let-7 in 1993 and 2000 (R. C. Lee, Feinbaum, and Ambros 1993; Ruvkun et al. 2000), a demand arose to find the purpose behind those short sequences, by identifying interaction partners. Early studies were focusing only on very few miRNAs at once, validating mostly pairs of combinations experimentally, e.g. by Luciferase Reporter Assays (Yamakuchi, Ferlito, and Lowenstein 2008; Cordes et al. 2009).

Several research groups started to develop miRNA target prediction algorithms, like `TargetScan` (B. P. Lewis et al. 2003) or `PicTar` (Krek et al. 2005). It turned out that the miRNA-mRNA binding preferences were far more complex than initially assumed and that a perfect matching 5' seed region of the miRNA did not guarantee a miRNA-mRNA interaction (Didiano and Hobert 2006). There were various binding types other than the seed region, detected by several studies of which some are described in the following (Figure 2). Furthermore, miRNAs can behave like siRNAs e.g. in *H. sapiens*, when bound to AGO2 and showing a binding over the total length of the miRNA (Hammond et al. 2000). Studies also revealed the "G-bulge" binding type, containing a G nucleotide in the "seed" region of the mRNA between nucleotide five and six (Chi, Hannon, and Darnell 2012). Another study expanded the binding possibilities of miRNAs by a "perfect centered" pairing (Shin et al. 2010), as well as the "imperfect centered" pairing (H. C. Martin et al. 2014). Other investigations showed a "3' supplementary" binding in addition to a shorter "seed" binding (Brennecke et al. 2005) and a "3' compensatory" binding with relaxed "seed" bindings (Yekta, Shih, and Bartel 2004). But also the "seed" region can vary within the canonical 2-8 bp pairing, having also atypical or marginal sites (Bartel 2009).

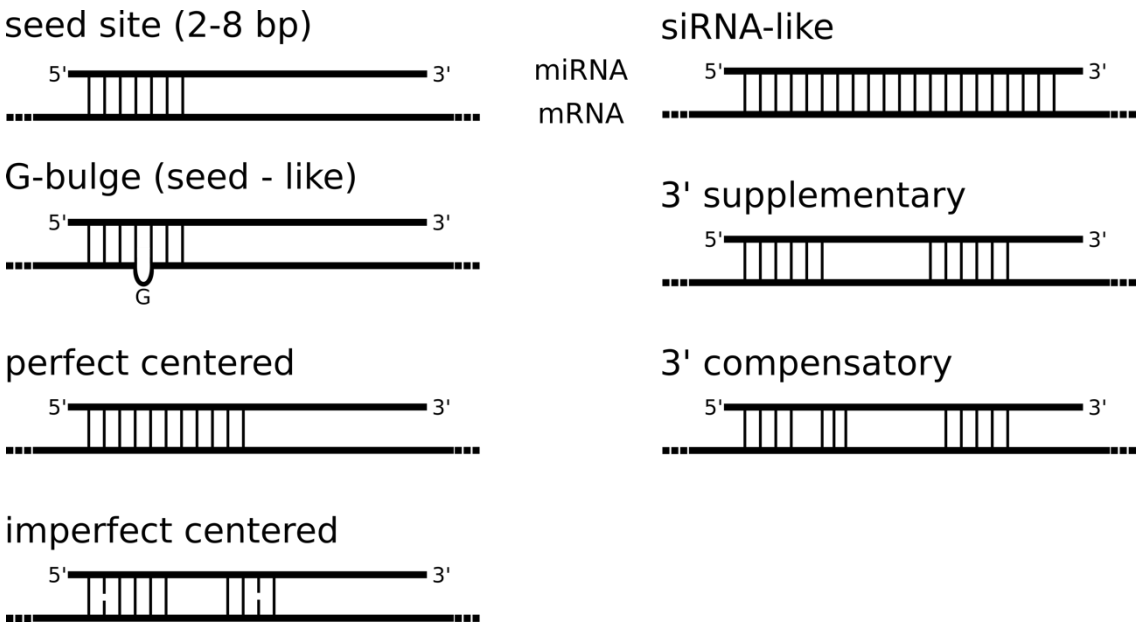


Figure 2 **The microRNA-mRNA binding variation** - Figure adapted from Cloonan 2015

Furthermore, the binding region on the mRNA is not limited to the 3' UTR but can also be in the coding sequence (CDS) (Hausser et al. 2013). These findings led to further developments of target prediction tools. Each of them tried to minimize the number of false positive predictions, using different approaches. For example, the previously mentioned `TargetScan` algorithm takes seed sequences from miRNAs and 3' UTRs as input and tries to find possible target sites, taking also evolutionary conserved target sites into account (Agarwal et al. 2015). In contrast, `PicTar` also uses the seed region of the miRNAs, but combines it with the mRNA-miRNA duplex free energy from `RNAhybrid` (Rehmsmeier et al. 2004) and evolutionary conservation (Krek et al. 2005). The standalone program `RNAhybrid` relies only on the criteria of minimal free energy hybridization of the mRNA and miRNA sequences, searching for the energetically most favorable intermolecular binding sites (Rehmsmeier et al. 2004). `MicroTar` also uses thermodynamic results from the miRNA-mRNA complex, but it considers the binding site accessibility for the target prediction by previously computing the unbound mRNA free energy (Thadani and Tammi 2006). The `miRanda` algorithm tries to identify highly complementary regions between the miRNA and the mRNA, ranking complementary bases at the 5' end of the miRNA higher than the 3' end and checks for thermodynamic stability (Betel et al. 2007a). There are many more target prediction tools available but some of them are only usable online, like `RNA22` (Miranda et al. 2006), making them inadequate for an universal local high-throughput pipeline. Moreover, some web-tools like `miRDB` (Wong and Wang 2015) or `ElMMo` (Gaidatzis et al. 2007) are focusing only on

very prominent species, like human or mouse, providing a curated set of miRNA-mRNA interaction patterns for target prediction.

Still, the target prediction suffers from a high false positive rate of around 40% to 66% (B. P. Lewis et al. 2003; Chi, Hannon, and Darnell 2012) and false negative rate of around 50% to 70% (Chi, Hannon, and Darnell 2012). Some researchers tried to overcome this by combining different target prediction tools and considering only intersecting predictions (A. C. Oliveira et al. 2017). This resulted in a reduced number of false positives, but on the same time in an increased number of false negatives. As an alternative, researchers recommended the use of an union approach of certain target prediction tools to decrease the number of undetected targets. A shortcoming of this study might be the usage of data from `miRTarBase` (Chou et al. 2018), whose miRNA-mRNA target interaction is in many cases validated by experiments, but relies on previous target predictions partly provided by the tested tools.

Besides limiting the surrounding parameters for predicting miRNAs on the whole transcriptome or 3' UTR, it is also possible to shrink the search space of putative miRNA binding sites to sequenced regions of mRNAs where a miRNA binding was detected. This technique is called cross-linking immunoprecipitation (CLIP). Investigations on miRNA and AGO CLIP techniques showed an improvement of false positive and false negative rates, with estimated values of 13% to 27% of false positives and 15% to 25% of false negatives (Chi et al. 2009).

1.5 Cross-linking immunoprecipitation sequencing

The development of RNA-sequencing technologies into a cheap and fast analysis method enabled the breakthrough of novel technologies, taking advantage of these nucleotide scale resolution results. Among others, cross-linking immunoprecipitation with high-throughput sequencing (CLIP-seq) technologies emerged. Originally invented for the identification of target interactions of a single protein, in order to enable a knock-out of all protein targets (Ule et al. 2005), the CLIP technique was enhanced by Next Generation Sequencing and then termed HITS-CLIP (High-throughput sequencing of RNA isolated by crosslinking immunoprecipitation) (Licatalosi et al. 2008). Other approaches were invented, like the photoactivatable ribonucleoside-enhanced crosslinking immunoprecipitation (PAR-CLIP) (Hafner et al. 2010) or the individual-nucleotide resolution crosslinking immunoprecipitation (iCLIP) (König et al. 2010).

Technically, the principle of those three mentioned methods is similar and described in the following (Figure 3) (König et al. 2012). The protein-mRNA interaction in the cell is permanently fixed by UV-light and the binding regions are identified by sequencing

mismatches or ends, respectively. For HITS-CLIP and iCLIP, living cells can be used as samples and are treated with 254nm UV light, whereas for PAR-CLIP, UV-light with a wavelength of 365nm is used. In the latter method, provided 4-thiouridine is incorporated into RNA by the cell, causing a mutation at binding sites in the later procedure. The UV treatment causes a fixation of the interacting sites of RNA and protein. Afterwards the cells are lysed and the RNA is partly digested. The fixed complex of RNA and protein is subsequently immunoprecipitated, using a protein-specific antibody. The proteinase K digests the attached protein and leaves a polypeptide at the crosslinking site. In the case of iCLIP, the following reverse transcription stops at the point of binding. Whereas the other two methods, HITS- and PAR-CLIP have adapter ligations to the 3'- and 5'-end, followed by the reverse transcription, iCLIP is missing the 5' adapter. Therefore, the 3' adapter contains also the 5' adapter together with a cleavage site in between. The transcript is circularized and then cleaved, followed by the PCR amplification. For HITS- and PAR-CLIP the reverse transcription reveals a mismatch at the binding site in the case of PAR-CLIP and a deletion or mutation at the binding site for HITS-CLIP. These conspicuous regions are then used for the identification of the binding regions via bioinformatics analysis.

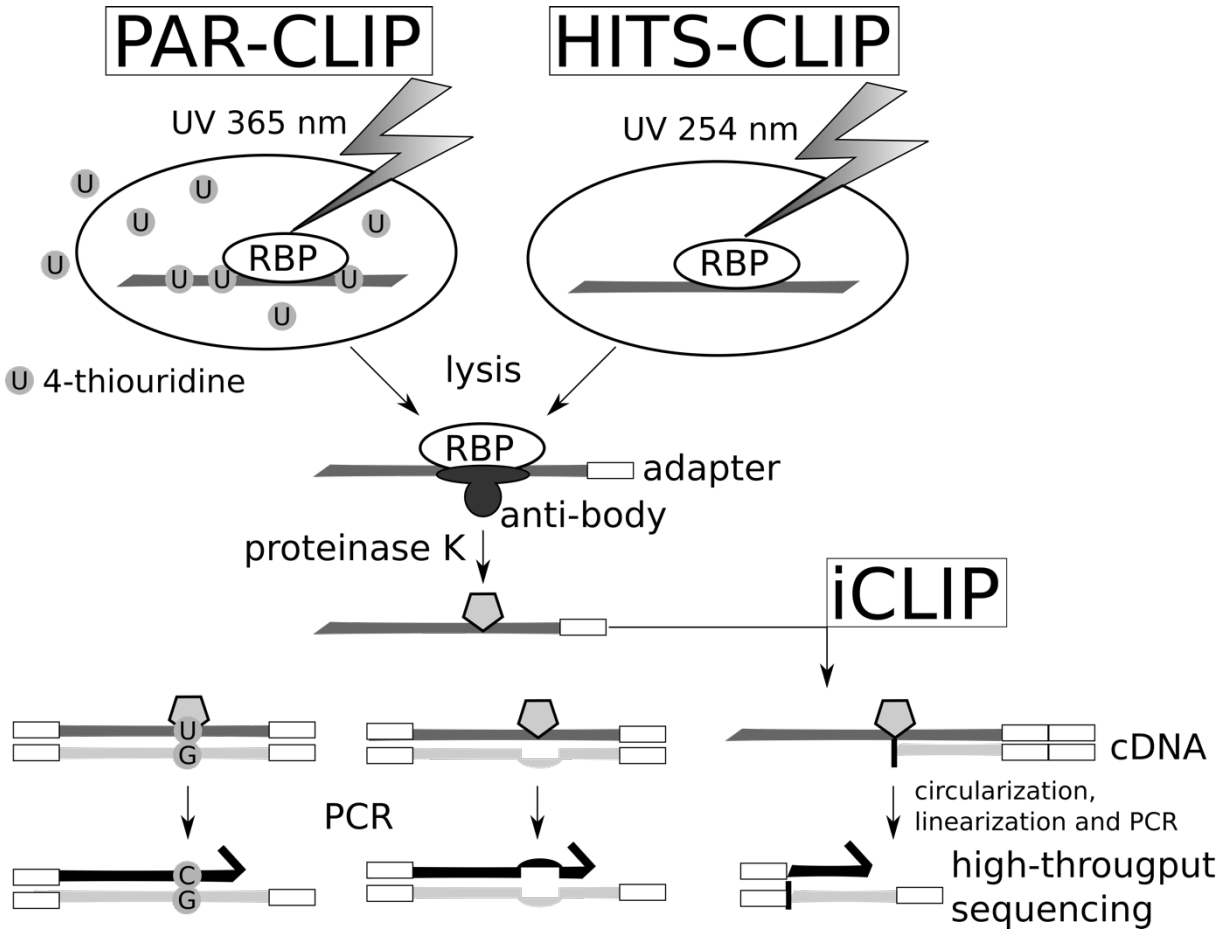


Figure 3 **Differences in CLIP variants** PAR-CLIP uses 4-thiouridine as mismatch inducer for cytosine, HITS-CLIP generates a mismatch or deletion at the binding positions and iCLIP stops the reverse transcription at the binding position. Figure adapted from König et al. 2012.

Outgoing from the predicted relation between miRNAs and mRNAs, inferences can be made about the miRNA influence on mRNAs and the effect of a regulation or even a misregulation.

1.6 Regulatory effects of microRNAs

The regulatory effect of miRNAs can be divided into two opposing models. On the one side, miRNAs are assumed to have fine-tuning effects on the mRNA expression level for whole expression networks, whereas the other model considers miRNAs as repressors of very few specific crucial mRNAs (Lai 2015).

Nevertheless, all models agree that miRNAs regulate the protein production process, there are various observations how this is achieved. One way is the direct interference during the initiation of translation by inhibiting the binding of the cap-binding complex (Humphreys et al. 2005; Mathonnet et al. 2007). Another way is the blockage of translating polyribosomes during the elongation process (Nottrott, Simard, and Richter 2006). The miRNA binding causes a

destabilization of the mRNA via deadenylation and decapping, followed by exonucleolytic digestion (Eulalio et al. 2008). In *H. sapiens*, it has been further demonstrated that miRNAs, in combination with AGO2, can behave like small interfering RNA (siRNA) (Hammond et al. 2000) by specifically cleaving the mRNA sequence, causing its degradation (Meister et al. 2004).

Independently from the type of regulation, namely the inhibition of translation or the destabilization of the mRNA sequence, the influence of miRNAs and the possible effects of their dysregulation have been broadly investigated in various research fields. For example, the LIM only protein, dLMO in *D. melanogaster*, is an essential transcription cofactor that inhibits Apterous (Biryukova et al. 2009). The cofactor is required for the proper wing development, by inhibiting the Apterous factor (Biryukova et al. 2009). Increased or decreased concentration of dLMO will lead to malformed wings that lack margins (Biryukova et al. 2009). Expression studies revealed that an overexpression of miR-9a results in a suppression of dLMO and vice versa, suggesting that miR-9a is fine-tuning the dLMO transcripts, ensuring an appropriate concentration of dLMO during the development of the wings and therefore a proper regulation of Apterous (Biryukova et al. 2009).

Besides those fine-tuning consequences with vital but malformed organisms, misregulated miRNAs can also have fatal effects on the whole organism. For instance a change of let-7 expression in *C. elegans* culminates in a lethal phenotype (Ruvkun et al. 2000).

In addition, miRNAs are also related to diseases, like cancer. A study from 2002 revealed the impact of miR-15 and miR-16 downregulation in lymphocytic leukemia (George Adrian Calin et al. 2002). Further investigations showed that those two miRNAs act as tumor suppressors that induce apoptosis in malignant nondividing B-cells and several solid malignancies, by repressing the anti-apoptotic protein Bcl-2 (Cimmino et al. 2005; George A. Calin et al. 2008).

But not only endogenous miRNAs may have an effect on the organism. Viruses, for example, can contain miRNAs that target the host immune system which enables them to settle more easily in the host organism (Stern-Ginossar et al. 2007).

To investigate those important molecules, a large variety of special tools are of need in a certain order. This workflow needs a well annotated species with a good-quality genome, as well as available miRNA and sequencing datasets to test different tools and methods for the individual steps properly. These criteria are fulfilled with the red flour beetle, *Tribolium castaneum*.

1.7 *Tribolium castaneum* - red flour beetle

The red flour beetle, *Tribolium castaneum* (named by Johann Friedrich Wilhelm Herbst in 1797), is a storage pest of food grains. During the averagely 22 weeks long oviposition period, the female animal lays two to three eggs per day, from which young larvae may hatch within one to two weeks (N. E. Good 1936). The larval condition has a median duration of one to two months until the pupation commences, which then again takes another one to two weeks until the adult animal ecloses and may expect to live around 547 days as males or around 226 days as females (N. E. Good 1936). The time periods may vary, depending on temperature, humidity and availability of food (N. E. Good 1936). Originally domiciled in the Indo-Australian region (E. H. Smith and Whitman 1992), *T. castaneum* can nowadays be found in nearly all temperate regions (Hill 1983). His kidney-like cryptonephridial organ allows him to populate even very dry areas (Richards et al. 2008). It further shows a broad resistance against pesticides (Richards et al. 2008). Those abilities, the rather quick alteration of generations, the possibility to occupy dry areas and the ineffective industrial control, promoted *T. castaneum* to a demanding target for scientific pesticide development in the scope of ecological and economic issues. But *T. castaneum* has also other beneficial properties, making him an ideal candidate for research studies. Besides of being cheap and easy to keep, the presence of a fully sequenced genome released in 2008 also facilitates investigations and led to its status of a model organism for beetles (Richards et al. 2008). The current genome version 5.2 from the Georgia GA2 strain was published around eight years later in 2016. This assembly consists of eleven chromosomes, 2,149 scaffolds, 7,059 contigs and has a total length of 165,944 Mbp with a protein count of 22,611 and a GC percentage of 35.1887 (information taken from National Center for Biotechnology Information (NCBI) - `Assembly GCA_000002335.3`).

Outgoing from the first established genome in 2008, early comparisons to the human genome already revealed more than 120 orthologous gene groups, previously not observed in all other sequenced insect genomes (Richards et al. 2008). This enabled investigations for the test of *T. castaneum* antimicrobial peptides against multidrug resistant bacteria in humans (Rajamuthiah et al. 2015). The beetle *T. castaneum* has also been established as screening model for transgenerational epigenetic side effects that were caused by pharmaceuticals (Bingsohn, Knorr, and Vilcinskas 2016). Another study revealed that RNA interference (RNAi) screenings, also known as RNA-silencing, for targetable lethal sequences can lead to the development of transgenic plants that produce this artificial RNA sequence as intrinsic pesticide (Knorr et al. 2013).

Further investigations may lead to screenings and regulatory manipulation techniques, exploiting natural and intrinsic mechanisms, like microRNAs.

Outgoing from this well annotated organism, one can transfer the techniques and methods to a species with no or nearly no information on miRNAs and their potential impact, but with promising indications for an impact of miRNAs and a putative transfer of knowledge towards human. Such an organism is the greater wax moth, *Galleria mellonella*.

1.8 *Galleria mellonella* - greater wax moth

The greater wax moth, *Galleria mellonella* (Linnaeus, 1758), belongs to the order of the butterflies. Originally domiciled in Asia (Paddock 1918), *G. mellonella* can nowadays be found all around the world (Kwadha et al. 2017). The adult animal is mostly active during dawn and night, where it is attracted by light and sugar. The females lay their eggs in the hives of bees, where the *G. mellonella* larvae use the wax, honey and pollen as nutriment (R. A. Nielsen and Brister 1979). The larvae then pupate within the honeycomb. For beekeepers this can cause serious problems, not only due to the physical damage the larvae cause, but also due to the potential of transmitting viral pathogens of bee diseases (Charriere and Imdorf 1999).

Furthermore, *G. mellonella* has emerged to an important host model organism for human pathogens (Cook and McArthur 2013). Studies demonstrated the antimicrobial activity within *G. mellonella* larvae, when infected with pathogenic and nonpathogenic *Listeria* strains (Mukherjee et al. 2010) or uropathogenic and commensal-like *Escherichia coli* strains (Heitmueller et al. 2017).

Unlike *T. castaneum*, the genome of *G. mellonella* was not available until a few months ago, when the first version was published as genome announcement without any annotation of genes (Lange et al. 2018). Previously, there were only *de novo* transcriptomic information available (Vogel et al. 2011), but no small RNA-seq derived microRNAs. The `miRBase.org` database held no entry on *G. mellonella* at that time.

2

MOTIVATION AND SCIENTIFIC AIMS

Recent technical advances have enabled mankind to utilize insects as model organisms for human research on a molecular level. Investigations of *Tribolium castaneum* revealed more than 120 orthologous gene groups to *Homo sapiens* and highlighted *Galleria mellonella* as model organism for human pathogens. Just like in humans, microRNAs were also found in insects, which leads to the assumption that one can use insects also as model organism for the investigation of microRNAs. Since their discovery in 1993 and first naming in 2001, miRNAs moved into the focus of broad research studies. The number of published manuscripts arose within the past decades to nearly 18,000 per year (from 2001 to 2019 - PubMed) (Figure 4a). By computing a word count on the abstracts of these publications, it becomes evident that most miRNA research is related to investigations covering mainly the topics “genes”, “expression”, “regulation”, “development”, “cancer”, “treatment” and “patients” (Figure 4b).

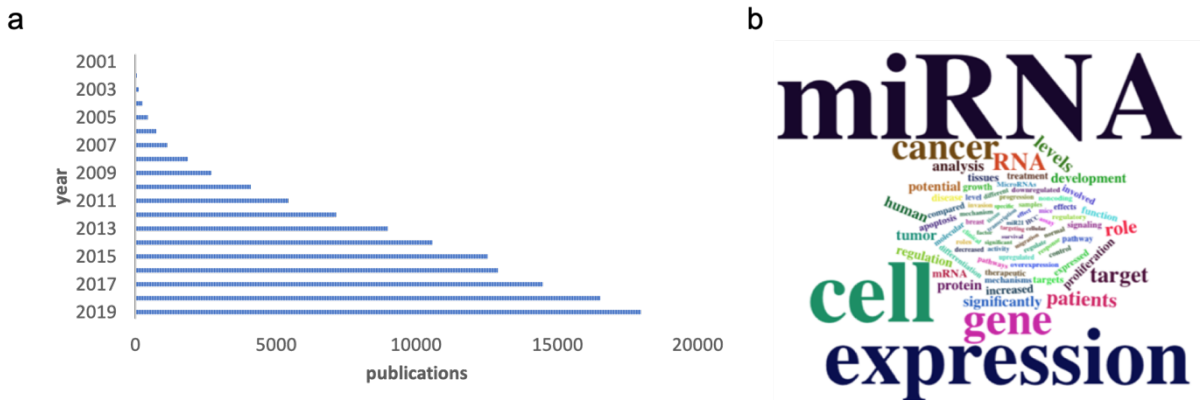


Figure 4 **a) Published manuscripts per year, since the first mention of “microRNA” or “miRNA” in 2001 until the end of 2019 on PubMed. (b) Frequently used words in the abstract of manuscripts about miRNAs.**

With the expanding number of identified miRNAs, the need for more sophisticated techniques to elucidate the identity and the role of all miRNAs in an organism became evident. For instance, the identification of novel miRNA targets was rendered more reliable by combining

laboratory methods, such as CLIP-seq together with bioinformatics tools (Chi et al. 2009). However, many different tools emerged, targeting various individual steps of the microRNA analysis. Hence, microRNA investigations are highly complex bioinformatics tasks that involve a great number of specialized tools and settings, treating large datasets. It is therefore necessary to be aware of proper tools with good performance, parameters for optimizing the results, but also of potential pitfalls.

To address the uncertainty regarding the analysis, I did a best practice assessment for CLIP-seq and miRNA analysis. The assessment was guided by the main question on which steps are needed for a standardized CLIP-seq and miRNA analysis workflow, when using CLIP-seq data as wet lab confirmation of miRNA binding sites on homologous mRNAs of closely related species. Several other considerations before starting the assessment had to be made: Which operating system and programming/scripting language/s are suitable? Within the assessment, I answered the questions of reliable tools and which parameters need to be adapted for the best performance.

Among the determined tools, needed for miRNA analysis, the miRNA isoform detection is an especially demanding topic. A recent study on *T. castaneum* early development stages showed the importance of poly-A-tails in microRNAs (Ninova, Ronshaugen, and Griffiths-Jones 2016), indicating the impact of such an analysis for certain research questions. The question on how miRNA isoforms can be identified was tried to be answered by multiple tools. However, a clear comparison of the different tools was needed to elucidate the advantages and disadvantages of each tool and choose the most appropriate one for further miRNA analysis. To determine a suitable tool, I created artificial test-sets, mirroring biological and technical variants in mature microRNAs. The individual performance of each tool in terms of true positives, false positives and false negatives in combination with an overall scoring highlighted strengths and weaknesses in the different isoforms. The most fitting tool for high-throughput analysis was then applied on the published data of *T. castaneum* early development stages and analyzed for further isoform types.

To meet the need for analyzing one or many huge datasets at once, the single tools identified in the best practice assessment and the benchmarking of miRNA isoform detection had to be combined into a workflow. This workflow was initially scripted in small, problem-oriented subsets and then transferred into an all-in-one pipeline. The main aim of the pipeline was the creation of a novel high-throughput pipeline that is capable of analyzing CLIP-seq data in the original species and automatically transfers these regions to homologous genes in the species of interest in order to shrink the search space for the final microRNA target prediction.

As a practical application of the pipeline, the role of microRNAs in the immune response against UPEC and ABU infected *E. coli* strains in *Galleria mellonella* was investigated, by firstly determining the microRNAome and identifying differentially regulated microRNAs. Recent findings demonstrated a differential epigenetic regulation when infected with different strains of *E. coli* that usually infect the human urinary tract (Heitmueller et al. 2017). Insects have an antimicrobial defense system in their fat body that is comparable to the mammalian liver (Lemaitre 2004). Therefore, insights of the antimicrobial peptide-microRNA connection could then be transferred to human in later investigations. Furthermore, the complex landscape of microRNAs and their regulatory functions are enhanced in general.

For the application of the pipeline to *G. mellonella*, I annotated the, at that time, newly released, but unannotated genome of *G. mellonella* and also assembled a reference-based transcriptome and created the first smallRNA-seq derived microRNAome with the aid of this new genome.

3

BEST PRACTICE ASSESSMENT RESULTS FOR CLIP-SEQ TRANSFER AND MICRORNA ANALYSIS

Outgoing from the assessment of a broad variety of investigations with microRNAs, I observed several computations that were performed in most of the studies (see chapter 3.3) and I discovered the need of an easy to use pipeline that solves the common tasks in a high-throughput manner. I further combined those commonly demanded tasks with some analysis that I encountered useful, like the determination of microRNA isoforms. Some tasks can also be valuable for some analysis, but do not necessarily require a fixed pipeline position, like arm-switching events or polycistronic clusters and are better implemented in database queries. Nevertheless, I tested them with dedicated scripts for their usability in advance.

An important point, I always encountered, was the determination of microRNA binding partners. This task is difficult due to the nature of microRNAs. Their small size makes it computationally challenging to determine correct binding solely based on nucleotide matches. As previously stated, several approaches have been developed to shrink search spaces and to limit the number of potential false positives. Some were trying to tackle this problem in an algorithmic fashion and others were trying to create a biological consent, like the limitation to the 3' UTR. To my understanding, one should include all transcript regions, but limit the region on the mRNA to the ones that give a biological signal of microRNA binding activity. I therefore planned to use CLIP-sequencing data of AGO-binding regions and transform it to other species (see chapter 3.2). This should retain evolutionary conserved regions and highlight the microRNA active sites, shrinking the search space to a less large area. In combination with an unbiased target prediction tool, the number of potential targets is more likely to be handled in wet-lab validation procedures. Unbiased in this case means that there is no limitation towards UTR or CDS regions.

Some general demands apply to the basis system on which the pipeline is developed. They are discussed in the following chapter.

3.1 Pipeline Basics

One of the first questions that arise when developing a pipeline is the destined operating system. Now for sake of bioinformatics, the answer is clearly LINUX, since a workflow that is treating large data, needs an operating system that is able to handle such an enormous amount of data. Furthermore, most bioinformatics tools are designed for a LINUX environment and of course, LINUX is mostly free of charge.

The very basic languages used, is LINUX SHELL. With SHELL, one can connect the different tools and custom scripts on a LINUX environment to powerful pipelines.

The scripting language PERL, initially released 1987, is a mighty language to evaluate and treat large text data, especially in the world of Bioinformatics. Therefore, many modules have been developed by the PERL community and made publicly available, e.g., via the Comprehensive Perl Archive Network (CPAN). For the microPIECE pipeline, a method for creating temporary files to save interim results at various timepoints during the analysis is needed. Therefore, PERL comes with a module, called File::Temp::tempnam that does exactly this. Besides the tools, mentioned in the following, there are some others that do not have that much alternative programs, because they are simply the standard application in this field. As there is for example the transformation, sorting, indexing or filtering of .sam and .bam files. This was performed with SAMtools (see supplemental material 13.1).

Another tool of this kind is BEDtools, which can be used for merging .bed files, extracting a .fastq or .fasta from .bed files (see supplemental material 13.2).

For the tool assessment, I considered only the ones that were open source and freely available in order to stay transparent in the analysis.

3.2 Best Practice CLIP-seq Analysis

As previously described, the CLIP technique is a promising way for identifying microRNA-mRNA pairs, but the availability of datasets and the possibility to create this data is rare. I therefore decided to use existing CLIP data and transfer the resulting signals to other species that are as closely related by evolution as possible. For this procedure, I investigated a workflow that fits the needs of such a project (Figure 5). The process starts with the raw CLIP .fastq files (0) and the need of removing the artificial adapter sequences (1). The CLIP reads are then

mapped to the reference genome (2). Outgoing from those results the signals of binding regions are identified via the so-called peak calling (3). In order to use the CLIP data for transfer, the homologous transcripts of another species need to be identified (4). Finally, the signaling peak regions need to be transferred to these homologous transcripts (5).

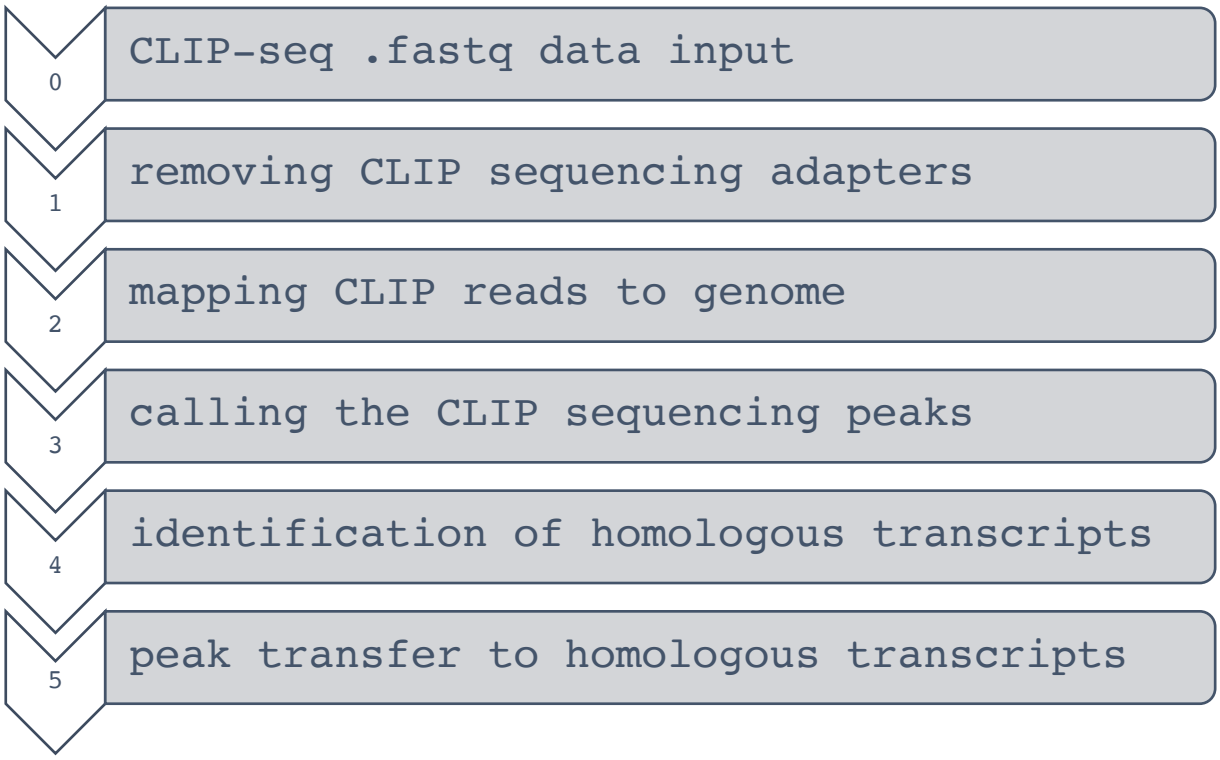


Figure 5 **Workflow for CLIP analysis and transfer.**

3.2.1 Removing CLIP Sequencing Adapters: Cutadapt

The Next Generation Sequencing techniques nowadays reach sequencing lengths of 2x300 bp (MiSeq V3 chemistry) for example. Although there are chemistries, suitable for shorter fragments, the variable size of those fragments may lead to the problem that the sequencer detects the appended technical adapters from the laboratory preparation of the sample. These adapters need to be removed in advance to further investigations. For the purpose of trimming, I chose the commonly used program `Cutadapt` (M. Martin 2011) as it is a fast and simple tool for this task and is not depending on third party software, like `Trimmomatic` (Bolger, Lohse, and Usadel 2014) that has an additional `Java` dependency. `Cutadapt`, written in `Python` and `C`, has a runtime of $O(nk)$ with n being the number of characters in all reads and k is the sum of the adapter length, leading to a treatment of 1 million reads per minute in case of 35 bp reads and an adapter of 18 bp (M. Martin 2011).

The algorithm of Cutadapt computes alignments between the given read and the provided adapters. The adapters can be parametrized as a 3' or 5' ligated adapter. In case of the 3' adapter, a random matching semi-global alignment between the 3' end of the adapter and the 5' end of the read could lead to a wrongly depletion of the read. Therefore, Cutadapt penalizes leading gaps in the read which forces the alignment of the adapter to start together with the read in this setup. Finally, all characters after the last non-adapter character are removed. In the case of the 5' adapter, the penalizing step is not included and all characters before the first non-adapter character are removed.

For the use on CLIP sequencing data, Cutadapt has been used with `--minimum_length 20` and `--trim-n` parameters.

3.2.2 Mapping CLIP Reads To The Genome: gsnap

The gsnap algorithm (Genomic Short-read Nucleotide Alignment Program) is able to align reads as short as 14 nt and was developed to account for the biological variance, not properly tackled by existing tools, like bwa or bowtie (T. D. Wu and Nacu 2010). The latter mentioned tools use the Burrows-Wheeler Transformation (see 8.4.2.2) to map reads quickly to a reference genome. This technique can account for very few mismatches between read and reference. The authors explain that nature and biological processes do not only consist of a single nucleotide polymorphism or a single mutation compared to the reference (T. D. Wu and Nacu 2010). Different studies discovered that indels in human play a role of 7-8% of all polymorphisms and that 25% of the indels in coding sequences are longer than 3nt (Weber et al. 2002; Bhangale et al. 2005). Therefore, they developed an algorithm that is more tolerant to sequence variation.

The bwa algorithm for example, has a faster runtime, because it uses a fixed starting seed and allows a certain number of mismatches, whereas gsnap uses all possible seeds over the entire length of the read. This method is therefore not biasing a certain read region.

The gsnap algorithm is used in the microPIECE pipeline for the mapping of CLIP-seq reads. As previously explained, the specialty of CLIP-seq reads is a mismatch signal in binding regions (see chapter 1.5). A mismatch tolerant and seed region flexible alignment program creates a large benefit in this case. I therefore used gsnap with the parameters `gsnap -N 1 -B 5 -O`. The used parameter `-N` includes splicing possibilities for the alignment. The `-B` parameter controls the batch mode. That means which data structures are allowed to allocate memory. The output shall be printed in the same order as the input with the `-O` parameter.

3.2.3 Calling The CLIP Sequencing Peaks: Piranha

Once the mapping of the CLIP-seq reads is finished, the resulting mapping file needs to be processed further to extract the location of the binding regions according to the specific technical signal. Therefore, a tool is needed that is tolerant for the specialties of different CLIP methods and the corresponding signals.

The `Piranha` tool is specifically designed to account for all these issues (Uren et al. 2012). The algorithm estimates the background noise of the signal by using the read-count distribution. Next, it searches for genomic spots that spike out from this background.

In the first place, it bins all reads together that start at the same nucleotide. Since data from CLIP-seq experiments are Poisson over-dispersed and the algorithm only retains the bins with at least one mapping read, the Piranha authors chose the `zero-truncated negative binomial` (`ZTNB`) for evaluation purposes. According to the authors, a p-value is computed for each bin, by subtracting the sum of the densities for all values below the read count in that bin.

In the `microPIECE` pipeline, a bin size of 30 is used.

3.2.4 Identification Of Homologous Transcripts: ProteinOrtho

The concept of orthology assumes a last common ancestral sequence, shared by two different species via a speciation event (Fitch 1970). This leads to the assumption that there should be no other sequence in the genomes of the two species that is more similar than the sequence of origin, enabling a reciprocal best hit search approach (Bork et al. 1998; Rivera et al. 1998; Hirsh and Fraser 2001; Remm, Storm, and Sonnhammer 2001). Although several databases, like `InParanoid` (Sonnhammer and Östlund 2015), `OrthoMCL-DB` (L. Li, Stoeckert, and Roos 2003) or `OMA Browser` (Altenhoff et al. 2018) already exist that cover the orthology information, I decided to compute this information during the runtime of the pipeline. This decision enables the use of species that are either not yet included in the databases or are even not yet published, because the data was created in-house.

The `ProteinOrtho` (Lechner et al. 2011) algorithm is an implementation of the above-mentioned reciprocal best hit or reciprocal best alignment heuristic. Depending on nature and biology, it could be that there is a pair of co-orthologs (`x1, y1, x1'` and `y1'`) and a pair of orthologs (`x2, y2, x3` and `y3`) (Figure 6a). The usual reciprocal best alignment hit would not identify the group of co-orthologs here, if only the best hit is considered (`n=1`). A solution could be to increase the number of best hits (`n=2`) (Figure 6b). This would lead to a detection of the co-orthologs (blue arrows), but also to the false positive detection of `x2-y3`

and x3-y2 (blue arrows). ProteinOrtho aims to solve this problem adaptively (Figure 6c), by introducing a similarity cut-off. This cut-off depends on the matching quality. In the case of BLAST (Altschul et al. 1990), the cut-off would be calculated according to the e-value of the best match, multiplied with a certain factor < 1.

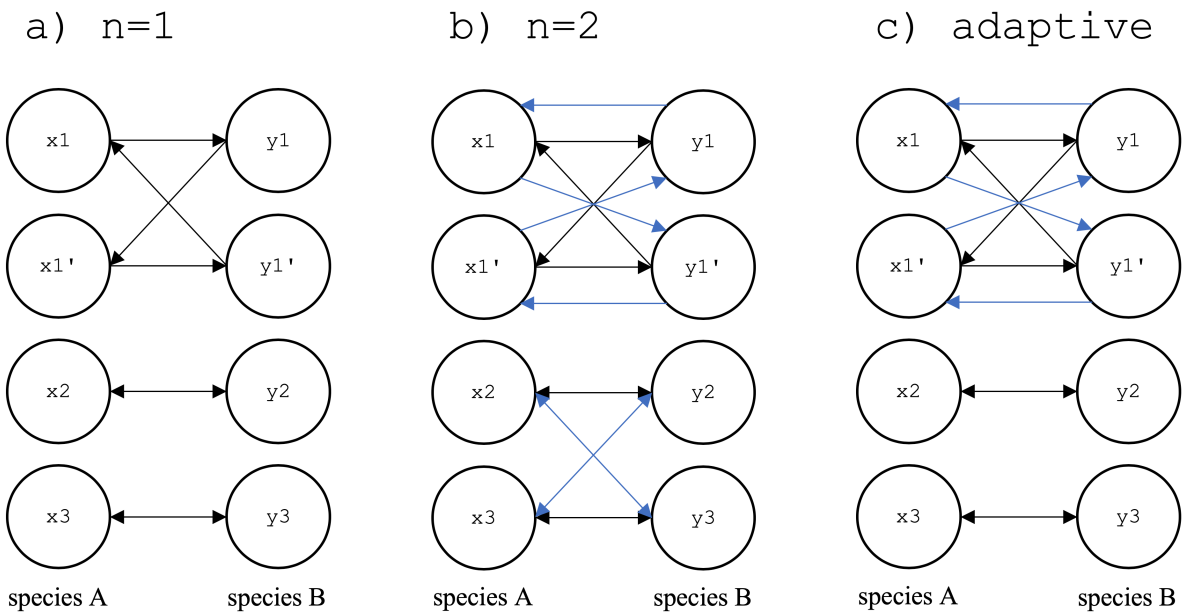


Figure 6 **Reciprocal best alignment heuristic issues** Adapted from (Lechner et al. 2011).

For the task of identifying homologs with the pipeline, I used ProteinOrtho without any parameter modifications, except the -clean for removing unnecessary files after the run automatically.

3.2.5 Peak Transfer To Homologous Transcripts: needle (EMBOSS)

The needle algorithm is an EMBOSS implementation of the Needleman-Wunsch global alignment (Needleman and Wunsch 1970). This algorithm is widely used in the bioinformatics field for the alignment of two nucleotide or two amino acid sequences. Exemplarily we want to find the optimal global alignment of the imaginary nucleotide sequences GTATC and CTA. The final outcome is now largely dependent on the used scoring system. For the sake of this example, we evaluate a matching between two characters with a score of 1 and a mismatch or an insertion/deletion with -1. For a better survey, the two sequences are written in a matrix *D* (Table 1a). Outgoing from the cell in the second row and second column, the algorithm recursively goes its way through the matrix, by evaluating the letter combination of the current

cell, taking the information of the putative previous cells into account (formula below – adapted from (Raden et al. 2018)).

$$D_{i,j} = \max \begin{cases} D_{i-1,j-1} + s(a_i, b_j) \\ D_{i-1,j} \quad + s(a_i, -) \\ D_{i,j-1} \quad + s(-, b_j) \end{cases} = max \begin{cases} D_{i-1,j-1} \quad + \quad 1 \;\; a_i = b_j \\ D_{i-1,j-1} \quad + -1 \;\; a_i \neq b_j \\ D_{i-1,j} \quad + -1 \;\; b_j = - \\ D_{i,j-1} \quad + -1 \;\; a_i = - \end{cases}$$

A path from left to right or from top to bottom would represent a gap (insertion / deletion) and would therefore be evaluated with -1 ($b_j = -$ and $a_i = -$), whereas a diagonal path is equivalent to a match ($a_i = b_j$) or mismatch ($a_i \neq b_j$). After computing each cell, the optimal alignment path, according to the score of each cell, is traced back from the lower right corner to the upper left corner, resulting in the final global alignment (Table 1b).

Table 1 **Needleman-Wunsch example alignment**

a)

		G	**T**	**A**	**T**	**C**
	0	-1	-2	-3	-4	-5
C	-1	-1	-2	-3	-4	-3
T	-2	-2	0	-1	-2	-3
A	-3	-3	-1	1	0	-1

b)

```
C  T  A  -  -
|  *  *
C  T  A  T  C
```

For the final application in the pipeline, I used several further settings compared to the custom program call in order to inveigle the algorithm to prefer a mismatch compared to a gap with the command `needle-endweight Y -gapopen 5 -gapextend 2`.

In contrast to the example, I used a gap opening penalty of five with `gapopen` and an extension penalty of two with `gapextend`. This leads to a less favorable result of stretched gapped alignments. Importantly, I enabled `endweight`, which applies penalties for gaps that were shifted outside of the other sequence. For the case of mismatching nucleotides in the alignment, the trivial assumption of a simple Boolean value, mirroring a match or no-match, is not sufficient. To account for the various possible combinations and putative effects of nucleotide changes, scoring matrices were developed. By default, `needle` uses the `EDNAFULL` matrix (Table 2). I customized this matrix slightly (Table 3) and named it `EDNACUSTOM`. The motivation here was to reduce the penalty for a mismatch, compared to a gap. Therefore, matching nucleotides scores, like `A` to `A` were changed from +5 to +2. Mismatching nucleotides, like `A` to `G` where scored with -3 instead of -4. Other scores were kept original. The matrix

consists of more than the commonly known four nucleotides. This is due to the fact that different nucleotides can be grouped into a single synonym character (Table 4).

Table 2 **EDNAFULL matrix** The original scoring matrix from needle (EMBOSS). Lowest score = -4 ; highest score = 5

	A	T	G	C	S	W	R	Y	K	M	B	V	H	D	N	U
A	5	-4	-4	-4	-4	1	1	-4	-4	1	-4	-1	-1	-1	-2	-4
T	-4	5	-4	-4	-4	1	-4	1	1	-4	-1	-4	-1	-1	-2	5
G	-4	-4	5	-4	1	-4	1	-4	1	-4	-1	-1	-4	-1	-2	-4
C	-4	-4	-4	5	1	-4	-4	1	-4	1	-1	-1	-1	-4	-2	-4
S	-4	-4	1	1	-1	-4	-2	-2	-2	-2	-1	-1	-3	-3	-1	-4
W	1	1	-4	-4	-4	-1	-2	-2	-2	-2	-3	-3	-1	-1	-1	1
R	1	-4	1	-4	-2	-2	-1	-4	-2	-2	-3	-1	-3	-1	-1	-4
Y	-4	1	-4	1	-2	-2	-4	-1	-2	-2	-1	-3	-1	-3	-1	1
K	-4	1	1	-4	-2	-2	-2	-2	-1	-4	-1	-3	-3	-1	-1	1
M	1	-4	-4	1	-2	-2	-2	-2	-4	-1	-3	-1	-1	-3	-1	-4
B	-4	-1	-1	-1	-1	-3	-3	-1	-1	-3	-1	-2	-2	-2	-1	-1
V	-1	-4	-1	-1	-1	-3	-1	-3	-3	-1	-2	-1	-2	-2	-1	-4
H	-1	-1	-4	-1	-3	-1	-3	-1	-3	-1	-2	-2	-1	-2	-1	-1
D	-1	-1	-1	-4	-3	-1	-1	-3	-1	-3	-2	-2	-2	-1	-1	-1
N	-2	-2	-2	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2
U	-4	5	-4	-4	-4	1	-4	1	1	-4	-1	-4	-1	-1	-2	5

Table 3 **EDNACUSTOM matrix** Modified version of the EDNAFULL matrix, delivered by default with the EMBOSS package. Lowest score = -3 ; highest score = 2

	A	T	G	C	S	W	R	Y	K	M	B	V	H	D	N	U
A	2	-3	-3	-3	-3	1	1	-3	-3	1	-3	-1	-1	-1	-2	-3
T	-3	2	-3	-3	-3	1	-3	1	1	-3	-1	-3	-1	-1	-2	2
G	-3	-3	2	-3	1	-3	1	-3	1	-3	-1	-1	-3	-1	-2	-3
C	-3	-3	-3	2	1	-3	-3	1	-3	1	-1	-1	-1	-3	-2	-3
S	-3	-3	1	1	-1	-3	-2	-2	-2	-2	-1	-1	-3	-3	-1	-3
W	1	1	-3	-3	-3	-1	-2	-2	-2	-2	-3	-3	-1	-1	-1	1
R	1	-3	1	-3	-2	-2	-1	-3	-2	-2	-3	-1	-3	-1	-1	-3
Y	-3	1	-3	1	-2	-2	-3	-1	-2	-2	-1	-3	-1	-3	-1	1
K	-3	1	1	-3	-2	-2	-2	-2	-1	-3	-1	-3	-3	-1	-1	1
M	1	-3	-3	1	-2	-2	-2	-2	-3	-1	-3	-1	-1	-3	-1	-3
B	-3	-1	-1	-1	-1	-3	-3	-1	-1	-3	-1	-2	-2	-2	-1	-1
V	-1	-3	-1	-1	-1	-3	-1	-3	-3	-1	-2	-1	-2	-2	-1	-3
H	-1	-1	-3	-1	-3	-1	-3	-1	-3	-1	-2	-2	-1	-2	-1	-1
D	-1	-1	-1	-3	-3	-1	-1	-3	-1	-3	-2	-2	-2	-1	-1	-1
N	-2	-2	-2	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2
U	-3	2	-3	-3	-3	1	-3	1	1	-3	-1	-3	-1	-1	-2	2

Table 4 **IUPAC nucleotide code** The code summarizes not only the characters for the single nucleotides, but also groups different bases together into one synonym.

IUPAC nucleotide code	Base
A	Adenine
C	Cytosine
G	Guanine
T (or U)	Thymine (or Uracil)
R	A or G
Y	C or T
S	G or C
W	A or T
K	G or T
M	A or C
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
N	any base

3.3 Best Practice microRNA Analysis

Common analysis methods and scientific questions accompanying miRNAs and their application by literature research led to the following best practice workflow. The workflow for microRNAs can be drafted by describing the, to my understanding, "must-have" parts as follows (Figure 7). The small RNA sequencing data (0) needs to be trimmed (1) and filtered (2) from unwanted ncRNA. Then a mining for novel miRNAs is performed (3). Having now the set of known and novel microRNAs, the expression can be measured (4). This is followed by the detection of homologous miRNAs in other species (5) and the target prediction (6) and isoform detection (7).

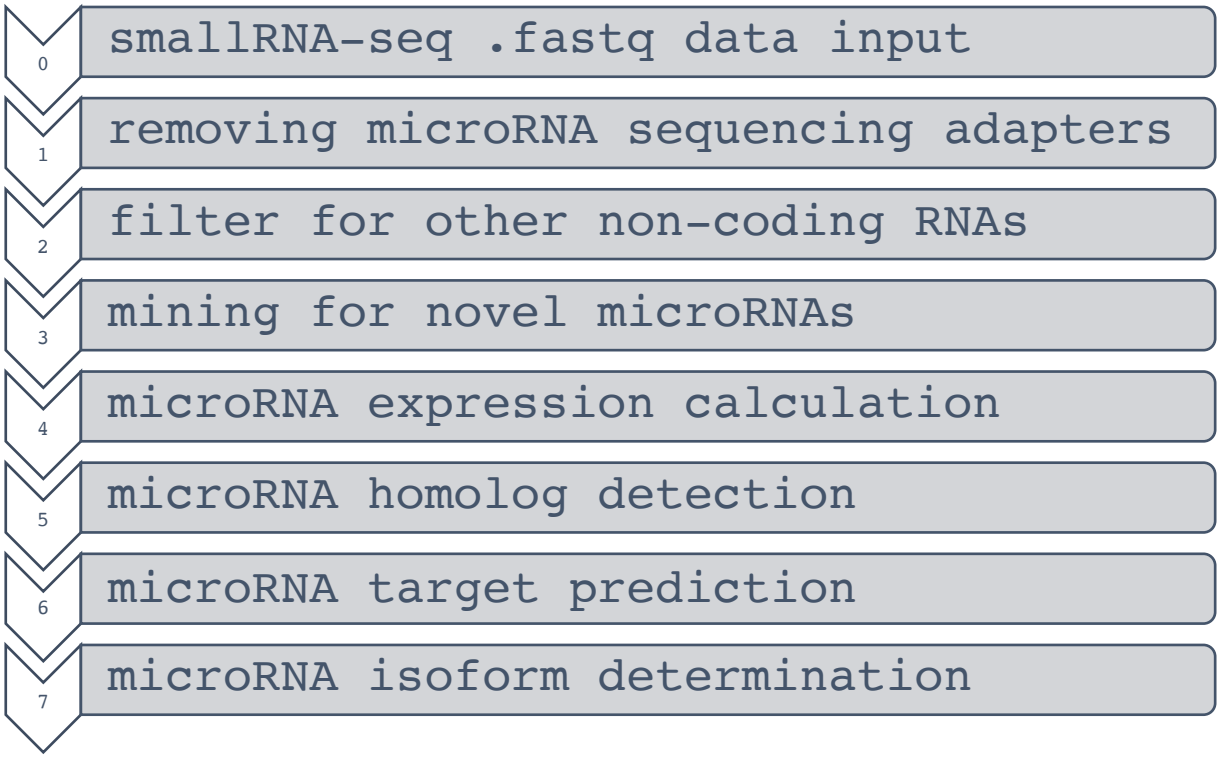


Figure 7 **Workflow in microRNA analysis.**

3.3.1 Removing miRNA Sequencing Adapters: Cutadapt

The `Cutadapt` tool has been described already in detail for the adapter trimming of the CLIP sequences. Here, for the miRNA context, the tool fits as well. The setup simply changes slightly to account for the special length of miRNAs. Outgoing from `miRBase.org`, the average length of all stored miRNAs is between 21 and 22. I observed 16 miRNAs having a length of 16 and three miRNAs having a length of 15 nucleotides, whereas the other 48.866 miRNAs are longer or equal to 17 nucleotides. A shorter read length cutoff would lead to a higher possibility of mapping reads to random regions simply by chance, due to the small alphabet of only four

characters and the comparably large reference genomes. A minimal length that is set too high would miss the smaller miRNAs and isomiRs with shortened ends. I therefore chose 17 nucleotides as minimal read length. Previous studies demonstrated that with a read length of 17, almost half of the reads can be mapped correctly, whereas the other half was not mappable with a broad range of programs (Ziemann, Kaspi, and El-Osta 2016). More dramatically is the mapping proportion of reads with a length of 16 nucleotides. Here only around a quarter of the reads were mappable. The numbers increase drastically to around 80 %, when reads have a length of 19 nucleotides. The size of 17 therefore appears to be a good tradeoff. I further used `Cutadapt` to discard reads with terminal nucleotides that are undetermined.

3.3.2 Filter For Other non-coding RNAs: bwa

The next step after trimming is the filtering against unwanted ncRNA fragments within the data. This is mostly achieved by mapping the reads against a set of ncRNAs, retaining only those reads that did not match. This step reduces the possibility to annotate artificial miRNA hairpins in the later process, but also the following computational time for further data processing.

For the mapping of miRNA reads from small RNA sequencing data, a variety of commonly used tools are available, like `bowtie1` (B Langmead et al. 2009), `bowtie2` (Ben Langmead and Salzberg 2012) and `bwa` (H. Li and Durbin 2009). The difficulties, compared to common RNA-sequence alignments, are the short read lengths in single-end mode in combination with the large reference genomes that in addition may contain complex or repetitive regions (Ziemann, Kaspi, and El-Osta 2016). Isoforms of miRNAs are increasing the difficulties, as well as reads with more than one best mapping locus, putatively derived from precursor homologs.

For this step, a precise mapping and a short runtime with a minimal amount of hardware consumption is desired. I applied `bwa` (Burrows-Wheeler Alignment tool) (H. Li and Durbin 2009) for this task. Benchmarking literature for read mapping software for miRNA reads showed that `bwa` appears to be very well suited and recommended for aligning small RNA sequencing data under the scope of miRNAs, allowing one mismatch in the read alignment which can be also located in the alignment seed region, but no gaps or gap extension (`bwa aln -n 1 -o 0 -e 0 -k 1`) (Tam, Tsao, and McPherson 2015).

The algorithm uses a backward search with the `BWT` (Burrows-Wheeler transformation) (Burrows, Burrows, and Wheeler 1994). The aim of the `BWT` is to construct a so called `BWT` string that can further be used for data compression for example. `BWT` takes an input string X and creates a circulating list, where the first letter of each line moves to the last position of the

string in the following line. This list is then lexicographically sorted to a sorted suffix array. The BWT output string is created by reading the last character in each line of the array.

If a substring *W* of *X* exists, each occurrence of *W* in *X* is within an interval in the suffix array. From this interval, the position is derived and therefor equivalent to a sequence alignment. The exact matching of a substring *W* to a string *X*, results in a single interval, whereas an inexact match leads to many potential intervals. Identical repeats in the string are collapsed in the same paths of the trie, returning all positions of these regions if a search query hits this branch. For the inexact match, a certain number of allowed mismatches and gaps is defined in advance. The traversal of the trie provides all possible results fulfilling the pre-defined criteria. The algorithm also uses a seed technique to restrict the number of mismatches in the first nucleotides. Exemplarily, the authors explain that for an alignment with 70 bp reads, a maximum of two differences in the 32 bp seed region, leads to a 2.5x faster runtime, compared to a no-seed approach.

3.3.3 Mining For Novel microRNAs: miRDeep2

After sequencing smallRNAs, one of the first questions that arise is whether there are novel microRNAs in the sequencing data. Therefore, a special tool is needed to identify possible new miRNAs. The miRDeep2 algorithm is specifically designed to extract novel microRNAs from small RNA sequencing data (Friedländer et al. 2012) It is widely used and appears to be the standard tool for this task. I considered it as an advantage over the Marco et al. method (Marco et al. 2010) that miRDeep2 is already available and does not need to be re-programmed. Furthermore, miRDeep2 is designed on Illumina reads and not on SOLiD, like the method described in Marco et al..

The core algorithm of miRDeep2 takes a reference genome as input, as well as small RNA sequencing data, together with a file with the positions of the mapped reads. It is also possible to include sets of known microRNAs. They can be from the current species, but also from related species or a combination of both in separate files.

The algorithm is separated into three parts. The already mentioned core algorithm, the mapper and the quantifier.

In the first place, a test is performed that checks if all provided files and parameters are correct. If the mature microRNAs from the species are provided as reference, this file is used as input for the quantifier. This enables miRDeep2 to report information for all already known microRNAs and the ones that are newly discovered in this analysis, even if the small RNA dataset does not include information on certain microRNAs. In the further progress, possible

sequences of microRNA precursors are identified, by using the provided mapping file. Here, putative stacks of reads that could be derived from microRNA biogenesis are considered further and the putative precursor is sliced out from the genome. Against this sequence, the reads are mapped, by using bowtie (B Langmead et al. 2009). The resulting putative precursor is afterwards checked for the secondary structure stability and shape, by using `RNAfold` (Hofacker 2003). The stability of the hairpin is predicted by the use of `randfold` (Bonnet et al. 2004). Next, the precursor is tested, if it fits a common structure of a hairpin and if the reads map to the stem regions, indicating a proper biogenesis. Finally, the precursor candidate is scored according to the previous measurements. If the potential precursor fails a test, it is discarded. The final results are summarized in a `.csv` file.

The `mapper` module processes the reads according to the provided parameters of minimal read length and collapses identical reads. It generates two output files, one is the processed read file and the other one holds the information of the mapping positions on the genome.

In the `quantifier` algorithm, the read counts are summed up. It uses the provided reference miRNAs and maps the reads against them. The correct mappings are then counted and exported into an output file. The used parameter `-P` specifies the use of the 5p and 3p notation from `miRBase` v18 and above.

3.3.4 microRNA expression calculation

Having now the novel miRNAs identified, the expression is calculated by counting the mapped reads for each mature miRNA. Therefore, `bwa` is used as previously explained with the miRNA optimized settings against the set of known and novel mature miRNAs.

In order to achieve a comparable statement of expression between different conditions, a normalization is needed. The normalization method for RNA-seq paired-end sequencing, Fragments Per Kilobase Million (FPKM) and Reads Per Kilobase Million (RPKM) for single-end reads are unsuitable for miRNA library normalization. In addition to the library size, both values take the fragment length into account, which is not necessary in the case of miRNAs, because they are only as long as one read and usually sequenced in single-end manner. I therefore used Reads Per Million (RPM) / Counts Per Million (CPM), representing a simple read count normalized by the library size.

3.3.5 microRNA homolog detection: BLASTN

In order to find homologous microRNAs between species, `BLASTN` is used (Camacho et al. 2009) since it is maybe the most prominent tool for homologous sequence search. `BLASTN` is the nucleotide part of the `BLAST` (`Basic Local Alignment Search Tool`) compendium and is specifically designed to compare nucleotide sequences with a nucleotide database (Altschul et al. 1990). As described by the authors, the algorithm searches for high scoring pairs between the query and the database, by using a heuristic version of the `Smith-Waterman` algorithm in order to gain computational speed compared to the standard variant. These short high scoring pairs are treated as seeds. The seeds are extended upstream and downstream, having an impact to the overall score of the alignment. If the overall score is above a predefined threshold, the alignment is reported in the results. In the pipeline, the command `blastn -outfmt '6 qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue itscore qseq sseq qlen slen' -word_size 4 -evalue 10000 -strand plus` was used. The output format was changed to tab separation (format 6) with the above specified columns. The word size for the seed was set to 4, due to the small sequence length of microRNAs. Because of the same reason, the evalue was change to 1.000, since the algorithm could evaluate the hit as a by chance event. The strand was set to plus, because only the matching strands were of interest and not the reverse strands.

3.3.6 microRNA target prediction: miRanda

Having the known and novel miRNAs analyzed, the question of mRNA targets arises. Here, one of the most prominent tools is `miRanda`. It has the advantage over other tools that it does not bias the results according to the search region, like `TargetScan` (B. P. Lewis et al. 2003). Algorithms focusing solely on the 3' UTR regions, may exclude potential binding regions in 5' UTRs and CDS, as demonstrated in *D. melanogaster* (Schnall-Levin et al. 2010). The mentioned publication also revealed that compared to *H. sapiens*, *D. melanogaster* may have more conserved CDS and 5' UTR regulating regions that appear to be regulatory active. Therefore, a limitation on the 3' UTR prediction by using the tools, mainly developed for *H. sapiens*, would bias the results by excluding more than 50 % of all possible binding regions (Hafner et al. 2010). Since the previous result is based on a HEK293 cell line, it can be assumed that the number for insects could be even higher, which was already indicated by a miRNA CLIP experiment of *A. aegypti* (X. Zhang et al. 2017).

The first phase of the miRanda target prediction algorithm is comparable to the Smith-Waterman algorithm (T. F. Smith and Waterman 1981). The Smith-Waterman algorithm tries to identify the part of a sequence with the highest identity when comparing to a larger sequence as optimal local alignment. Like in the previous chapter, we want to compare the two artificial sequences GTATC and CTA with each other. This time, we do not want the optimal global alignment, but the optimal local one. In contrast to the Needleman-Wunsch algorithm, the Smith-Waterman algorithm does not end in the lower right corner of the matrix, but in any cell of the matrix with the highest score of all cells in the matrix (Table 5). The individual scores are computed according to the recursion below (adapted from (Raden et al. 2018)). For the example, I used a score of 1 for matches and −1 for mismatches and gaps.

$$S_{i,j} = max\begin{cases} S_{i-1,j-1} + s(a_i, b_j) \\ S_{i-1,j} + s(a_i, -) \\ S_{i,j-1} + s(-, b_j) \\ 0 \end{cases} = max\begin{cases} S_{i-1,j-1} & + & 1 & a_i = b_j \\ S_{i-1,j-1} & + & -1 & a_i \neq b_j \\ S_{i-1,j} & + & -1 & b_j = - \\ S_{i,j-1} & + & -1 & a_i = - \\ & & 0 & \end{cases}$$

Table 5 **Smith-Waterman example alignment** The score is 2

a)

		G	**T**	**A**	**T**	**C**
	0	0	0	0	0	0
C	0	0	0	0	0	1
T	0	0	1	0	1	0
A	0	0	0	2	1	0

b)

```
T  A
*  *
T  A
```

The difference of the miRanda variant is that this version of the Smith-Waterman algorithm does not aim to align two similar sequences, but tries to find complementary nucleotides (A=U or G=C) with respect to wobble base pairs (G=U) (Enright et al. 2003). The developers set the parameters for matching complementary bases to +5 (A=U and G=C) and to +2 for G=U. All other combinations were set to -3. Gap opening penalties were set to -8 and gap extensions were fixed to -2. Furthermore, the alignment algorithm weights matching complementary nucleotides at the miRNA 5' end higher than at the 3' end, by multiplying the first eleven nucleotides from the 5' end with a scaling factor of 2. Finally, several empirical rules were applied by the developers. They forbid a mismatch at position 2 to 4, when counting from the 5' end of the miRNA. They allowed less than 5 mismatches between the nucleotide 3

and the nucleotide 12, but at least one mismatch between nucleotide 9 and the end of the alignment minus five positions. In those last five positions, less than two mismatches are allowed. The default cutoff for the overall score was set to 80 by the developers.
In the second phase, a thermodynamic analysis of the binding free energy (ΔG) of the miRNA-mRNA binding is computed by using the Vienna package (Wuchty et al. 1999).
Finally, all results are scored and reported according to their performance in the two phases.
The microPIECE pipeline uses the miRanda algorithm without any modifications of parameters.

3.3.7 microRNA isoform determination

Outgoing from various investigations, e.g., showing the importance of isomiRs for *D. melanogaster* development (Fernandez-Valverde, Taft, and Mattick 2010) and 5' miRNA modifications that lead to a target switch of a specific miRNA in *H. sapiens* (Tan et al. 2014b), I wanted to include this analysis into my pipeline. Hence, I encountered the problem that there are no comparative evaluations of the existing programs available. To choose the most suitable tool for this task, I did my own investigations and benchmarked available high-throughput tools, namely isomiR-SEA, isomiRID and miraligner. Besides, there were several other tools available, but they were only available via web-interface or with a desktop GUI and were therefore discarded. The mentioned desktop GUI tool is called IsomiRage (Muller, Marzi, and Nicassio 2014) and is a Java application with graphical frontend that aligns the reads to a so-called custom genome that includes a predefined set of potential microRNA isoforms. This method therefore could exclude potential isoforms that have not been considered previously.
The DeAnnIso tool is one of the web-based algorithms that allows only few species to be analyzed online. It uses the bowtie1 mapping algorithm and allows 0-3 mismatches for at least the first 5 bases of the 5p end. By default, 2 mismatches in the first 10 bases are allowed.
The program isomiRex is another web-based tool that also uses bowtie1 for mapping purposes. It also allows the user to visually inspect the alignment of the different isoforms underneath the precursor sequence. For miR-isomiRexp it was announced in the manuscript that there should be a web-interface, but this appeared offline to me at the time of investigation. The tool uses bowtie1 as well, but its focus is more tailored towards cancer datasets, for example from TCGA.
My investigations concerning isomiR tool performances and their application on an example dataset of *Tribolium castaneum* early development stages result in the following publication.

4

PUBLICATION I: BENCHMARKING OF MICRORNA ISOFORM DETECTION TOOLS

After the microRNA assessment of frequent scientific questions, analysis methods and tools, the need of a comparison between the different microRNA isoform detection tools arose.
My results are published in BMC Bioinformatics as research article, which was cited seven times at the time of writing.

Daniel Amsel, Andreas Vilcinskas, and André Billion.
"Evaluation of high-throughput isomiR identification tools: illuminating the early isomiRome of *Tribolium castaneum*."
BMC bioinformatics 18.1 (2017): 359.
https://doi.org/10.1186/s12859-017-1772-z

For this publication I designed the evaluation, chose the programs to be evaluated, designed the experiments, analyzed the results, created the draft manuscript and revised the manuscript. I further read and approved the paper.

The publicly available datasets analyzed in this study are available from the NCBI GEO repository:
`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63770`
The scripts are available via GitHub:
`https://github.com/DanielAmsel/isomiRBenchmark.git`

Amsel et al. BMC Bioinformatics (2017) 18:359
DOI 10.1186/s12859-017-1772-z

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Evaluation of high-throughput isomiR identification tools: illuminating the early isomiRome of *Tribolium castaneum*

Daniel Amsel[1*], Andreas Vilcinskas[1,2] and André Billion[1]

Abstract

Background: MicroRNAs carry out post-transcriptional gene regulation in animals by binding to the 3' untranslated regions of mRNAs, causing their degradation or translational repression. MicroRNAs influence many biological functions, and dysregulation can therefore disrupt development or even cause death. High-throughput sequencing and the mining of animal small RNA data has shown that microRNA genes can yield differentially expressed isoforms, known as isomiRs. Such isoforms are particularly relevant during early development, and the extension or truncation of the 5' end can change the profile of mRNA targets compared to the original mature sequence. We used the publicly available small RNA dataset of the model beetle *Tribolium castaneum* to create the first comparative isomiRome of early developmental stages in this species. Standard microRNA analysis software does not specifically account for isomiRs. We therefore carried out the first comparative evaluation of the specialized tools isomiRID, isomiR-SEA and miraligner, which can be downloaded for local use and can handle next generation sequencing data.

Results: We compared the performance of isomiRID, isomiR-SEA and miraligner using simulated Illumina HiSeq2000 and MiSeq data to test the impact of technical errors. We also created artificial microRNA isoforms to determine the effect of biological variants on the performance of each algorithm. We found that isomiRID achieved the best true positive rate among the three algorithms, but only accounted for one mutation at a time. In contrast, miraligner reported all variations simultaneously but with 78% sensitivity, yielding isomiRs with 3' or 5' deletions. Finally, isomiR-SEA achieved a sensitivity of 25–33% when the seed region was mutated or partly deleted, but was the only tool that could accommodate more than one mismatch. Using the best tool, we performed a complete isomiRome analysis of the early developmental stages of *T. castaneum*.

Conclusions: Our findings will help researchers to select the most suitable isomiR analysis tools for their experiments. We confirmed the dynamic expression of 3′ non-template isomiRs and expanded the isomiRome by all known isomiR modifications during the early development of *T. castaneum*.

Keywords: Insectomics, microRNA, Small RNA sequencing, isomiRID, isomiR-SEA, Miraligner

4.1 Background

MicroRNAs (miRNAs) are post-transcriptional regulators of gene expression that influence a wide range of biological processes (Bartel 2004). In insects, the dysregulation of miRNA expression during metamorphosis is often lethal (Agrawal et al. 2013; Y. L. Zhang et al. 2015; Jayachandran, Hussain, and Asgari 2013). Mature miRNAs are ~22 nucleotides in length and the 3′ end binds to a member of the Argonaute protein family to form an RNA-induced silencing complex (RISC) (Bernstein et al. 2001). The RISC binds target mRNAs within the 3′ untranslated region (UTR) or in the coding sequence via complementary base pairing with the miRNA seed region (nucleotides 1–8) and in some cases also the compensatory region (nucleotides 13–16) (Bartel 2009). RISC binding inhibits further processing of the mRNA, thus blocking translation or promoting degradation (Bartel 2004).

The biogenesis of miRNAs can involve the production of isoforms known as isomiRs (Morin et al. 2008). These are thought to be produced deliberately as separate products with defined roles in the cell, and do not represent errors of transcription or errors of sequencing (L. W. Lee et al. 2010). The isomiRs may be extended or truncated at either end compared to the mature miRNA, presumably due to imperfect cleavage by Drosha or Dicer (Kuchenbauer et al. 2008). Recent studies indicate that 5′ isomiRs undergo a seed region shift which changes the set of target mRNAs compared to the original miRNA (Tan et al. 2014a). The set of target mRNAs can also be changed by nucleotide editing (Luciano et al. 2004; G. Sun et al. 2009). Mature miRNAs may also acquire non-templated polynucleotide 3′ tails generated by nucleotidyltransferases (Wyman et al. 2011). This phenomenon has been observed during early insect development as part of maternal transcriptome regulation (Fernandez-Valverde, Taft, and Mattick 2010; Ninova, Ronshaugen, and Griffiths-Jones 2016).

The results described above show that miRNAs and isomiRs play important roles during animal development, especially insect morphogenesis. To gain more insight into the prevalence of isomiRs in insects we screened the publicly available small RNA dataset of the model beetle *Tribolium castaneum* originally focusing exclusively on 3′ non-templated isomiRs in the early development stages (Ninova, Ronshaugen, and Griffiths-Jones 2016). The data had already undergone a conservative form of isomiR investigation by iteratively truncating the non- templated 3′ ends until a certain minimal length was reached or the sequence perfectly matched a known miRNA. We investigated the performance of tools for isomiR identification that

account for more than non- templated 3′ tails. Several such tools have been developed but no comparative benchmarks are available. We selected a set of three candidate tools that are suitable for the analysis of high-throughput sequencing data and compared their performance to identify the best software. Using a simulated test set of Illumina reads and a set of artificial isomiRs, we investigated the influence of technical errors and biological variations on each type of software and determined the sensitivity and specificity for each case. From these values, we calculated a final weighted performance score for each tool. Taken individually, the two cases also provide detail information on the eventual need of post system error correction, considering the system error test case and possible detection leaks of isomiR types, uncovered by the biological variant test set.

4.2 Methods

4.2.1 IsomiR analysis software

Seven isomiR mining and alignment tools are currently available as non-proprietary software (Table 6). Three of them are command line tools that can be downloaded and integrated into high-throughput pipelines, and these are described in more detail below. We used these three methods for a comparative benchmark of their individual performance on simulated reads. If adjustable, we used the default settings in each tool without read abundance cutoffs. We wanted each tool to utilize its entire search space and therefore did not set the parameters to a common minimum in the case of mismatches, additions and deletions.

Table 6 **List of non-proprietary isomiR alignment programs**. The three command line tools were used for our comparative evaluation. The others were discarded because they were incompatible with local high-throughput pipelines.

Program	Usage	Alignment method	Publisher
isomiR-SEA 1.60	Command line	User-defined seed size (default 6)	(Urgese et al. 2016)
isomiRID	Command line	bowtie1	(L. F. V. de Oliveira, Christoff, and Margis 2013)
miraligner	Command line	8nt seed	(Pantano, Estivill, and Martí 2010)
IsomiRage	Desktop GUI	bowtie1	(Muller, Marzi, and Nicassio 2014)
DeAnnIso	Webapp	bowtie1 and BLAST	(Y. Zhang et al. 2016)
isomiRex	Webapp	bowtie1	(Sablok et al. 2013)
miR-isomiRExp	Webapp – offline	bowtie1	(Guo et al. 2016)

4.2.1.1 isomiR-SEA

The C++ program `isomiR-SEA` focuses on the seed region of miRNAs. It is a standalone executable file without dependencies and can be run with parameters in the command line. It requires the mature miRNA file from `miRBase` and the sequence reads. The reads must be collapsed and reformatted with the unique read and its abundance in one line. The algorithm extracts the seed regions from the mature miRNAs and groups them together. At first, the reads are screened for seed regions. When found, the seed region is extended without gaps in both directions and the correct position of the seed block is checked. The algorithm continues the extension towards the 3′ end and allows a second mismatch if the distance between the two mismatches falls within a user-defined threshold. The alignment is then extended further until either the third mismatch or the end of the read is encountered. Then the scores for each aligned read are computed. The output files are grouped into unique mapping reads, ambiguous reads that map more than once, and ambiguous selected reads that also map to various miRNAs but can be assigned to a unique one due to an internal scoring function (Table 7). There are also "unique", "ambiguous" and "ambiguous selected" output files, referring to the miRNA instead of the read.

Table 7 **Result files generated by isomiR-SEA**. The tag files focus on the read, whereas the others report the variants of the miRNA.

Unique	Tag_unique
Unique_ambigue_selected	
Ambigue	Tag_ambigue
Ambigue_ambigue_selected	Tag_ambigue_selected

4.2.1.2 isomiRID

The Python 2.7 script `isomiRID` uses bowtie (B Langmead et al. 2009) to map small RNA sequencing reads against reference precursor miRNAs. The script uses a configuration file in which the user can specify the paths of the executables, the data and the parameters. In the first round, perfect matches against the precursors are identified. An optional filtering step of the unaligned reads against the corresponding transcriptome or genome can be performed to filter reads not from miRNAs. In the second step, reads with one mismatch are taken into account. Iterative trimming of the 5′ and 3′ ends is used to seek potential non-templated miRNA isoforms. The findings are filtered according to user-defined abundance cutoffs and the results are concatenated into output files, allowing for reads with more than one mapping location. The output is a tab separated file in which every mapped read is aligned under the assigned precursor sequence together with the identified type of isoform and the abundance of the read.

4.2.1.3 miraligner

The Java tool `miraligner`, originally from the `SeqBuster` package but now independent, is a single jar file without dependencies. It uses a collapsed read file and the miRNA hairpin `FASTA` file from `miRBase` (S. Griffiths-Jones et al. 2006) together with the hairpin secondary structure file. The reads are mapped to the hairpin sequences via seeds of eight nucleotides, allowing one mismatch within the sequence. It allows up to three non-templated nucleotide additions at the 3′ end, as well as up to three nucleotides that differ from the mature 3′ or 5′ ends. This allows a slight shift of the precursor compared to the annotated position in the hairpin secondary structure file from `miRBase`. We used the default settings with a maximum substitution of one and a trimming/adding of three. The output is a tab separated file. It shows a result for each mutation type, the read sequence together with the number of its assignments, as well as the names of the miRNA.

4.2.2 Technical error simulation

We evaluated the effect of Illumina sequencing errors on the accuracy of isomiR identification by each tool. The small RNA sequencing data were simulated using `ART` (Huang et al. 2012) (version Mount Rainier 2016–06-05) with the Illumina HiSeq2000 and MiSeq-v1 sequencing system in single-strand mode: `art_illumina -c 1000 -ss [HS20|MSv1] -i < pattern_file_with_miR_length_X > -l < miR_length_X > -o < output>`. We grouped all miRNAs with the same length into one file and ran the command for each file separately. Afterwards, the files were merged into one. These sequencing systems are widely used for small RNA sequencing and mirror the most recently analyzed biological data. To ensure traceability, the simulated sequences must be uniquely assignable to their source. In case of `isomiRID` and `miraligner`, this can be achieved by the sequence header. The results of `isomiR-SEA` lack this header and a traceability can only be provided by sequence identity. Therefore, we had to ensure a uniqueness of miRNAs and their reads. We used the 430 *T. castaneum* mature miRNAs from `miRBase v21` and merged identical sequences. This new set of 422 sequences was then used as the pattern for the two simulations, with a coverage of 1000 reads per sequence. Due to the nature of the simulation program, about half of the 422,000 reads were sequenced as a reverse complement and were therefore omitted from further analysis. The remaining reads, 210,753 for HiSeq2000 and 210,961 for MiSeq-v1, were then filtered for redundancy. This resulted in 13,850 unique

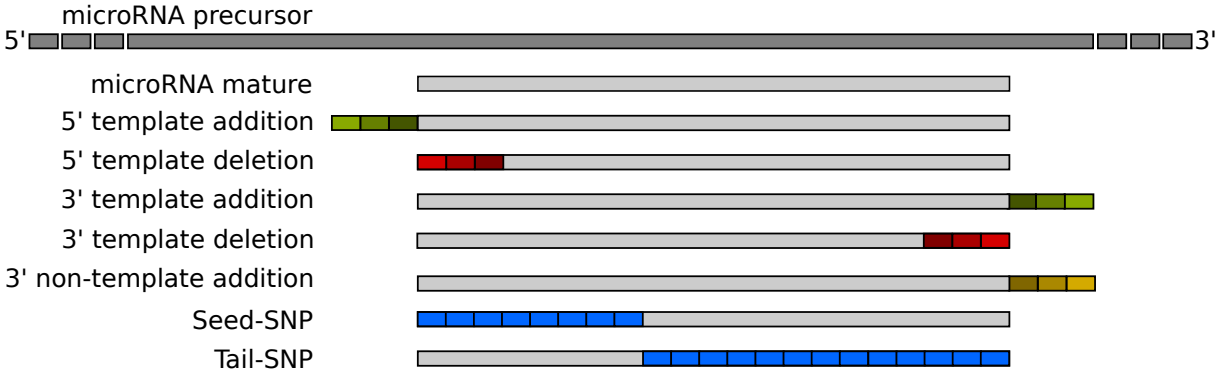


Figure 8 **The seven types of isomiR custom mutations.** The green boxes represent nucleotide additions. The red boxes represent nucleotide deletions. The yellow boxes represent non-template additions. The blue boxes show the positions of SNPs

reads for HiSeq2000 and 5964 unique reads for MiSeq-v1. This ensured a coverage of 14–32 read variants per original miRNA and therefore a broad variety of technical errors. The correct assignment of erroneous reads to its source was treated as true positive because the tools cannot distinguish between error and mutation. An additional analysis after the identification step might be of use, depending on the investigation.

4.2.3 Biological variant simulation

In order to evaluate the isomiR programs comprehensively using biological data, we created custom sequences based on the mature *T. castaneum* miRNAs from `miRBase v21`. This mirrored seven different types of isoforms (Figure 8). Both the 5′ and 3′ template isoforms were divided into truncated and extended variants. For the truncated variants, we created three different 5′ and three different 3′ isomiRs per mature microRNA, by iteratively trimming one nucleotide from the 5′ or from the 3′ end respectively. For the three 5′ and three 3′ extended variants, we added one nucleotide to the particular end of the mature miRNA, using the precursor miRNA as the template, until a maximum of three additions was reached. The 12 3′ non-templated isoforms per mature miRNA were created by adding one nucleotide of the same type to the mature miRNA, until a total of three nucleotides were added. We divided the single nucleotide polymorphism (SNP) isoforms into two distinct classes: the seed-SNPs and the tail-SNPs. We replaced each nucleotide from position 1 to 8 with the remaining three nucleotides for the seed-SNPs dataset and from position 9 to the end for the tail-SNPs dataset, resulting in three SNP isoforms per miRNA nucleotide position. This allowed us to distinguish the performance of seed-based search algorithms between seed and tail SNPs. We again kept the created reads non-redundant to ensure the traceability of the mapped reads by sequence identity. Our resulting test set finally mirrored each

possible variation and therefore provided a general unbiased condition.

4.2.4 Performance evaluation

We evaluated each algorithm using the simulated technical and biological *T. castaneum* reads. The results were classified as true positives (TP), false positives (FP) and false negatives (FN). True negatives (TN) were excluded because they were not needed for further calculations. Correctly assigned reads were treated as true positives. A wrongly assigned read was treated as false positive and a missing assignment to the correct miRNA was treated as false negative. We also calculated the sensitivity (TP/(TP + FN)) and the specificity (TP/(TP + FP)) of each isomiR software. Three possible approaches can be used to evaluate small RNA sequencing reads with more than one mapping location. One is to ignore multi-mapping reads completely and focus on distinct results. The second option is to group the miRNAs with the same read together. The third is to distribute the abundance of the read among the number of mapped miRNAs (Landgraf et al. 2007). We decided to use the third approach because the other options would modify the isomiRome.

4.2.5 *Tribolium castaneum* small RNA sequencing data

Recent studies have indicated the presence of abundant non-templated 3′ isomiRs during the early development stages of *T. castaneum* and *Drosophila melanogaster* (Fernandez-Valverde, Taft, and Mattick 2010; Ninova, Ronshaugen, and Griffiths-Jones 2015). We used the publicly available *T. castaneum* small RNA sequencing data from the `GSE63770` project (Table 8) for our analysis. Those datasets monitor the development of *T. castaneum* from the egg (including the switch from maternal to zygotic transcription after 5 h) until hatching (144 h) (Ninova, Ronshaugen, and Griffiths-Jones 2015).

Table 8 **List of publicly available *T. castaneum* small RNA datasets representing different developmental stages.** After ~5 h, the maternal transcription phase ends and zygotic transcription commences (Ninova, Ronshaugen, and Griffiths-Jones 2015).

ID	Sample	Transcription
GSM1556886	Oocyte small RNA replicate 1	Maternal
GSM1556887	Oocyte small RNA replicate 2	Maternal
GSM1556888	Embryo small RNA 0–5 h replicate 1	Maternal
GSM1556889	Embryo small RNA 0–5 h replicate 2	Maternal
GSM1556890	Embryo small RNA 8–16 h	Zygotic
GSM1556891	Embryo small RNA 16–20 h	Zygotic
GSM1556892	Embryo small RNA 20–24 h	Zygotic
GSM1556893	Embryo small RNA 24–34 h	Zygotic
GSM1556894	Embryo small RNA 34–48 h	Zygotic
GSM1556895	Embryo small RNA 48–144 h	Zygotic

4.2.6 Adapter trimming and quality filter

The *T. castaneum* small RNA sequencing data was trimmed with `Cutadapt` (M. Martin 2011) v1.8.3, using `-m 17` as the minimum read length, `-M 30` as the maximum read length and `--trim-n`, to trim potential N characters at the ends of the reads. We excluded reads with at least one N character in their sequence.

4.3 Results

We selected three high-throughput isomiR analysis tools suitable for command line use and investigated the effects of biological variation and sequencing-derived errors on the results produced by each tool (Figure S1). The technical test sets were created with `ART`, using a copy rate of 1000 reads per miRNA. We additionally created biological test sets geared to known miRNA isoforms and again reduced them to a non-redundant set, allowing us to measure the effects of biological variation on the results produced by each tool. We finally generated scores for each tool and selected the appropriate software for the analysis of the *T. castaneum* isomiRome.

4.3.1 Effect of technical errors on isomiR analysis

We created simulated HiSeq2000 and MiSeq-v1 reads based on mature miRNA templates from `miRBase v21` with `ART` (Huang et al. 2012). The multiple `isomiR-SEA` result files were divided into two distinct evaluations. We distinguished between the total results reported by `isomiR-SEA` (unique - reads that mapped only once and ambigue - reads that mapped more than once) on one hand and the selected results, already filtered by `isomiR-SEA` (unique - reads that mapped only once and ambigue_selected - reads that mapped more than once but were disambiguated through `isomiR-SEA` internal scorings) on the other. The number of `isomiR-SEA` false positives was lower in the selected set compared to the total results, falling by more than 15% for MiSeq-v1 and more than 18% for HiSeq2000 (Figure 9a). However, the false negative rate increased by nearly 7% for both HiSeq2000 and MiSeq-v1 in the selected set. This is also reflected in the increased specificity (+23.15% for HiSeq2000 and +21.97% for MiSeq-v1) and weaker sensitivity (−1.95% for HiSeq2000 and −1.37% for MiSeq-v1) (Figure 9b). The results produced by `miraligner` and `IsomiRID` were almost identical for this benchmark: `miraligner` achieved ~1.60% and ~0.78% more true positives than `IsomiRID` for the HiSeq2000 and MiSeq-v1 data, respectively, ~0.50% fewer false positives for both HiSeq2000 and MiSeq-v1, as well as 1.13% and 0.21% fewer false

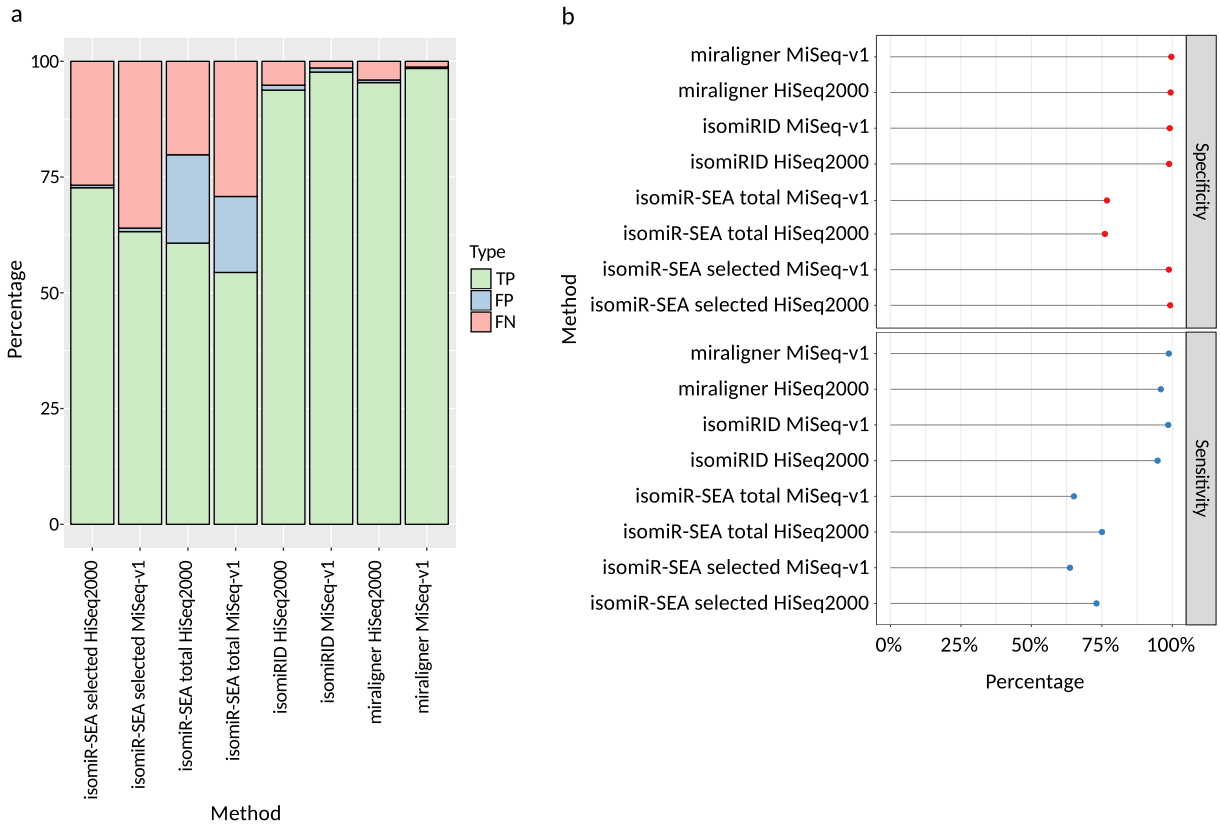


Figure 9 **Technical error benchmarking of the isomiR analysis tools.** Each algorithm was applied to the simulated sequencing error test set. (a) Plot of the true positive, false positive and false negative values from the mapping of erroneous reads against miRNAs.

negatives for HiSeq2000 and MiSeq-v1, respectively.

4.3.2 Effect of biological variant on isomiR analysis

We tested the three tools for their ability to process artificially mutated miRNAs representing isomiR variations. Although `isomiRID` achieved a true positive rate of at least 98.4%, the false positive rate was 0.7–1.6% for every variant, except 3′ additions with 0.08% false positives (Figure 10a). In contrast, `miraligner` achieved a true positive rate of >99.5% and a false negative rate of ≤0.5% for all variants except 3′ and 5′ deletions, where the false negative rate was ~21% (Figure 10b). We again distinguished between total and selected `isomiR-SEA` results, attempting to eliminate multi-mapping reads. For the total results (Figure 10c) we observed for nearly every type of mutation a false positive rate of ~25%, with the exception of seed-SNPs and 5′ deletions where the false positive rates ranged from ~7% to ~10%. We also observed false negative rates of 60% and 70% in these two variants. For the selected results (Figure 10d) the false positive rate ranged from 0% for 3′ non-templated additions to 1.5% for 5′ deletions. The false negative rates for 3′ and 5′ template additions, 3′-non-templated

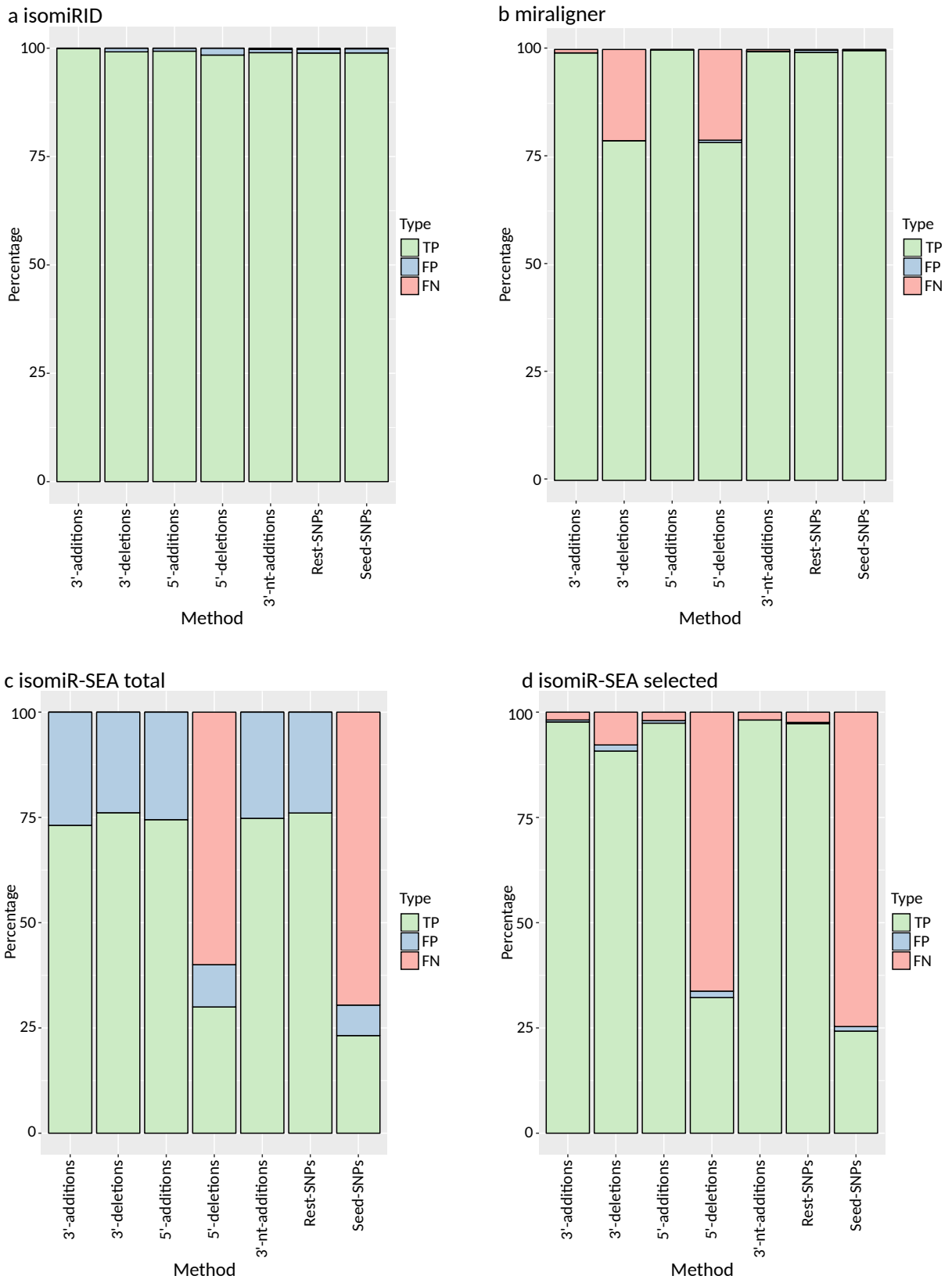


Figure 10 **True positive, false positive and false negative results generated by isomiR analysis tools.** The algorithms isomiRID (a), miraligner (b), isomiR-SEA total (c) and isomiR-SEA selected (d), were applied to the simulated biological variation test set

additions and variants covering mutations outside the seed region were all approximately 2%. However, the false negative rate increased to 7.8% for 3′

truncations, 66% for 5′ truncations and 77% for seed-SNPs.

The sensitivity of isomiRID was >99% for every variant and 100% for truncations and extensions at either end of the sequence (Figure 11a). In contrast, the sensitivity of `miraligner` for deletion variants was 79% and ~99% for every other variant (Figure 11b). When considering the total results, the sensitivity of `isomiR-SEA` was 100% for every variant except seed-SNPs and 5′ deletions, where the sensitivity fell to 33% and 25%, respectively (Figure 11c). When considering the filtered results, the sensitivity of `isomiR-SEA` ranged from 92% to 98% for most variants but again showed a lower sensitivity for seed-SNPs and 5′ deletions, with values almost identical to the total results (Figure 11d). The specificity of `isomiRID` ranged from 98% for 5′ truncations to 99% for 3′ templated additions (Figure 11a). The specificity of `miraligner` was 100% for templated 3′ and 5′ additions and 3′ truncations, and 99% for 5′ truncations (Figure 11b). The specificity of `isomiR-SEA` (total results) was 73–76% (Figure 11c) whereas the selected results improved the specificity to 95–98% (Figure 11d).

In order to exclude a possible influence of the read length to the result, we tested the effect of artificial read lengths on the method detection efficiency (Figures S2 and S3). `IsomiRID` had a weak anti-correlation between read length and false positive rate of −0.36. Its highest false negative rate was at the length of 18 nt. `Miraligner` had a moderate anti-correlation between read length and false negative rate of −0.53. This was mainly caused by read lengths between 15 and 17 nt. The two variations of `isomiR-SEA` performed equally, concerning the correlations. They show an anti-correlating value of −0.24 and −0.22 for false negatives, caused by read lengths between 18 and 26 nt.

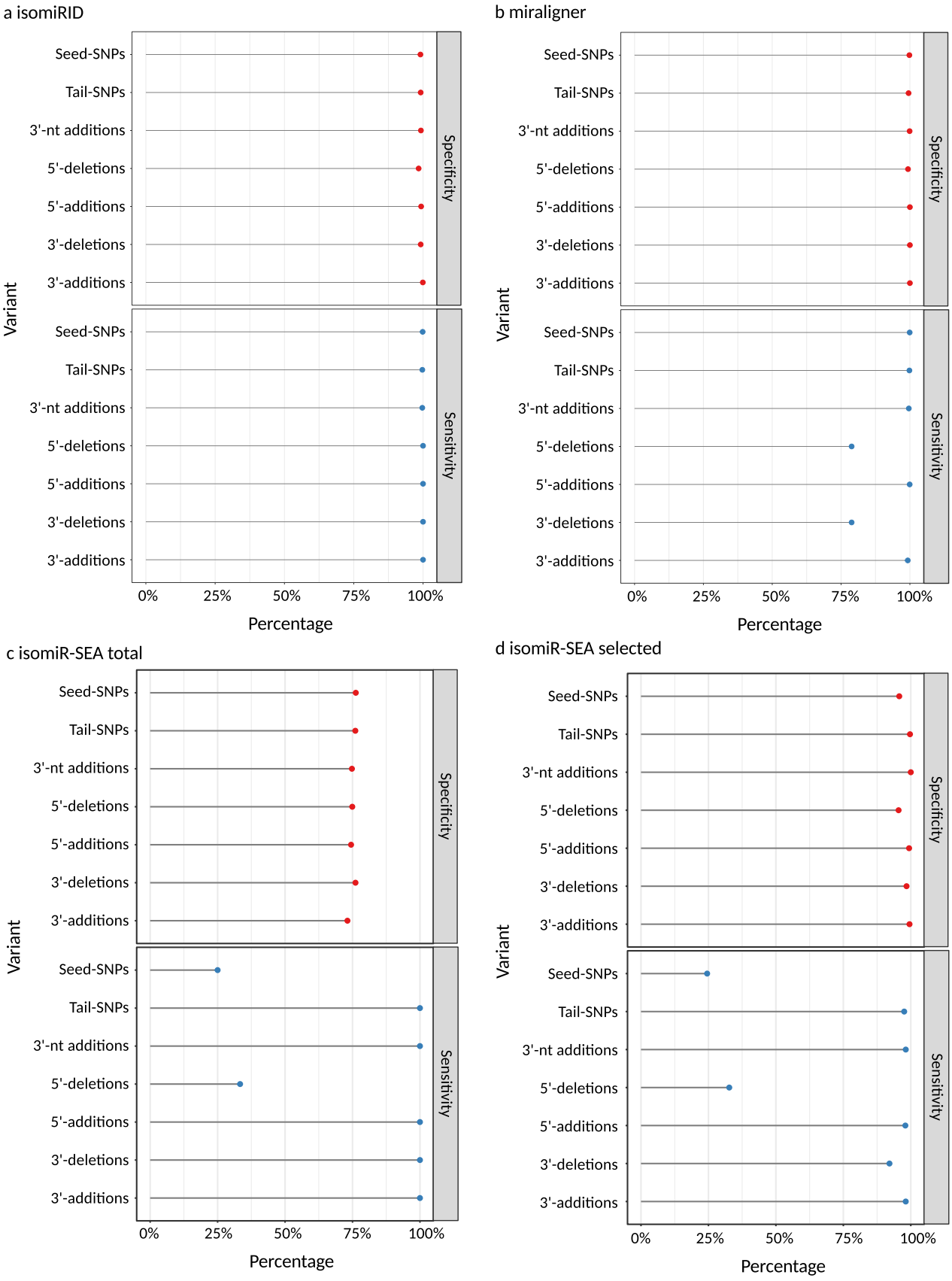


Figure 11 **Sensitivity and specificity of the isomiR analysis tools isomiRID (a), miraligner (b), isomiR-SEA total (c) and isomiR-SEA selected (d).** The values were calculated using the TP, FP and FN metrics from the analysis of the biological variation test set

4.3.3 Overall performance scores for isomiR analysis software

Each of the analysis tools was scored according to its performance when handling technical errors and biological variations as described above, resulting in the overall ranking presented in Figure 12. We calculated the f-scores for each tool and weighted them depending on their impact on real data. The highest score of 12.90 points was achieved by `isomiRID`, followed by `miraligner` with 12.59 points and `isomiR-SEA` with 9.13 and 10.25 points for the total and selected data, respectively.

We calculated the f-scores for each testing variant. Then each f-score was weighted regarding to its impact on the targeting mechanism of the miRNA isoform. We assigned a weighting of 1 to the templated 3′ additions and truncations as well as the tail-SNPs because these do not affect the seed region and therefore the range of mRNA targets is unchanged. However, variants that affect the seed region such as seed-SNPs and 5′ additions and truncations were weighted with a multiplier of 2, because changes in this region can modify the mRNA target range and are more biologically significant. We also assigned a multiplier of 2 to the 3′ non-templated additions because of their impact during early development. Finally, every score was summed up for each tool and set as final score for the evaluation.

In selecting a method for analysis of the *T. castaneum* isomiRome, we also considered

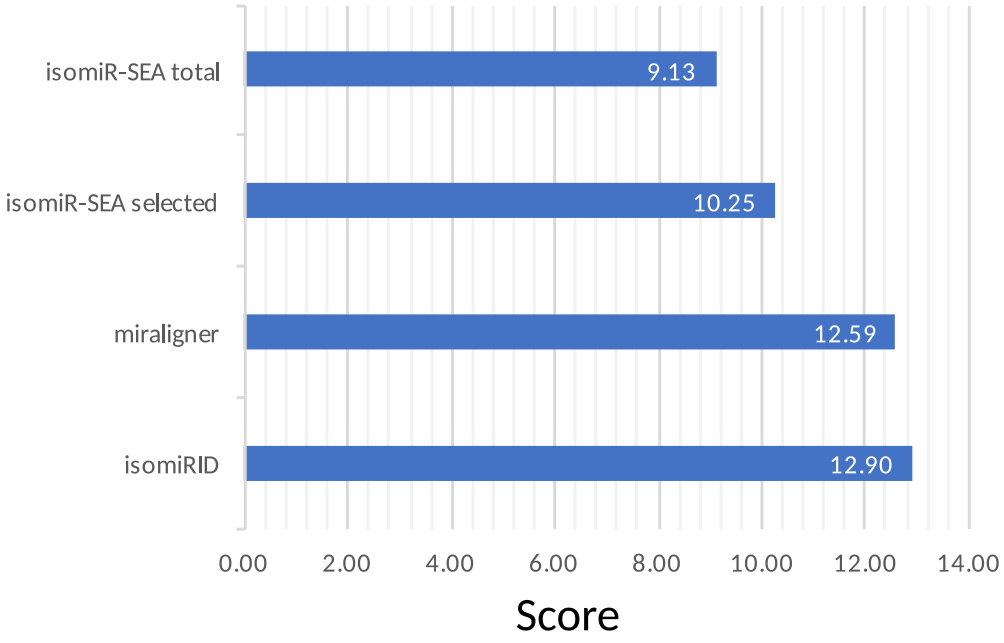


Figure 12 **Overall ranking of the isomiR analysis tools**. The points were calculated by weighting true positives, false positives and false negatives together with the impact on the seed region

aspects of general usability. For example, `isomiRID` uses precursor sequences and calculates a dot alignment for every matching read, but the number of dots is sometimes incorrect. This results in a visually shifted mature sequence alignment. Furthermore, `isomiRID` also reports only one mutation at a time and does not mark 5p and 3p miRNAs. In contrast, `miraligner` can report all isoforms simultaneously but replaces reads with the same name. We also observed that the precursors tca-miR-3811c-1 and tca-miR-3851a-1 were not reported in the test output even though they were provided in the input file, whereas the precursors tca-miR-3811c-2 and tca-miR-3851a-2 were present. We compared each pair and found that those precursors share the same mature sequence.

We nevertheless selected `miraligner` for the further analysis of the *T. castaneum* isomiRome, using the same settings as in the test cases. It scored 0.31 fewer points than `isomiRID` but 2.34 more than `isomiR-SEA` using the filtered data. It reported all variations for each read and generated fewer false positives than `isomiRID`, which reports only one mutation at a time and therefore cannot be

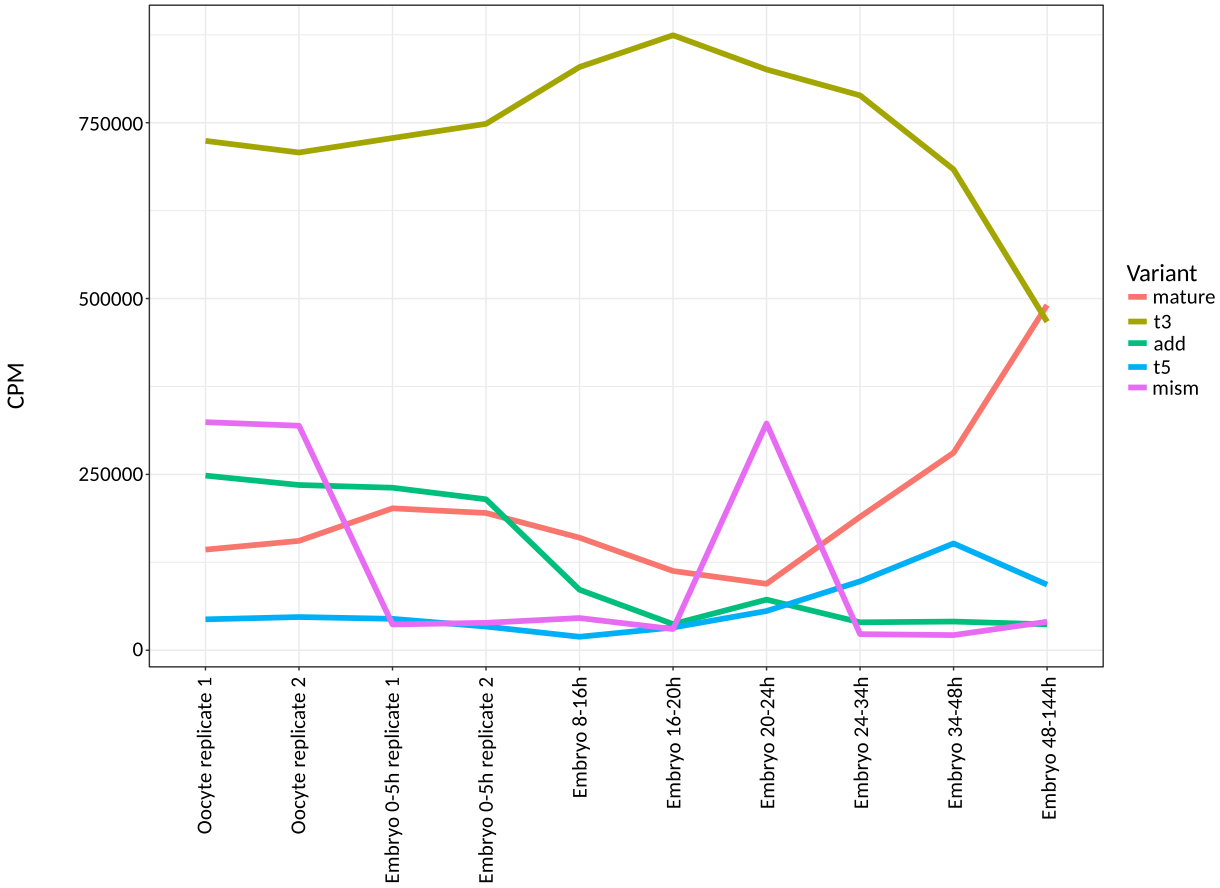


Figure 13 **Counts per million reads per condition, normalized by the number of multi-mapping reads.** This shows the 3' non-templated additions (add), the mature sequence (mature), the mismatches (mism), templated 3' additions and deletions (t3) and templated 5' add.

used for comprehensive isomiRome profiling. Precursor overwriting was ignored because we focused on the mature sequences.

4.3.4 The isomiRome of TCA

We calculated the number of reads that matched each type of isomiR variant in counts per million (CPM). The multi-mapping reads were normalized by the number of assigned microRNAs to avoid overrepresentation (Figure 13). We observed an increase in the number of 3′ non-templated additions (add) during the maternal transcription phase (oocyte replicates 1 and 2, embryo 0–5 h replicates 1 and 2) which agreed with previous studies in *T. castaneum* (Ninova, Ronshaugen, and Griffiths-Jones 2015) and *D. melanogaster* (Fernandez-Valverde, Taft, and Mattick 2010). We also observed an initial increase in the number of templated 3′ additions (t3) peaking during the embryonic phase 16–20 h and declining thereafter. The mature sequences showed an opposing expression profile, with the lowest point at 16–20 h and an increase thereafter. The final phase had a higher CPM than the templated 3′ additions. The 5′ templated additions (t5) were present at constantly low levels with the exception of the 34–48 h phase. The SNP isoforms (mism) ranked second highest in expression value in the oocytes, which is even higher than previously reported for non-templated 3′ additions (Ninova, Ronshaugen, and Griffiths-Jones 2015). The expression of SNP isoforms dropped to one of the lowest values of all variants in the post-oocyte phases although there was a second significant peak during the 20–24 h phase before falling to minimal levels thereafter.

We next scanned for all non-templated nucleotide additions at the 3′ end. We confirmed that isomiRs with polyadenylate tails are strongly expressed in the oocyte and during the first embryonic stage; then expression weakens at the beginning of the first zygotic transcription phase (8 h). This reproduced the findings of the original study using the same dataset (Ninova, Ronshaugen, and Griffiths-Jones 2015) (Figure S4). Templated 3′ additions and deletions occurred very frequently in these datasets, although the expression level dropped below that of the unmodified mature microRNA in the final phase (48–144 h). In most cases, the 3′ end was shortened by two or three nucleotides compared to the original miRNA, but we also observed isomiRs that were elongated by two or three nucleotides during the 8–16 h and 24–34 h phases (Figure 14). We observed a steady low level of 5′ isomiR expression with the exception of the penultimate and antepenultimate phases, where a single nucleotide 5′ extension was prevalent.

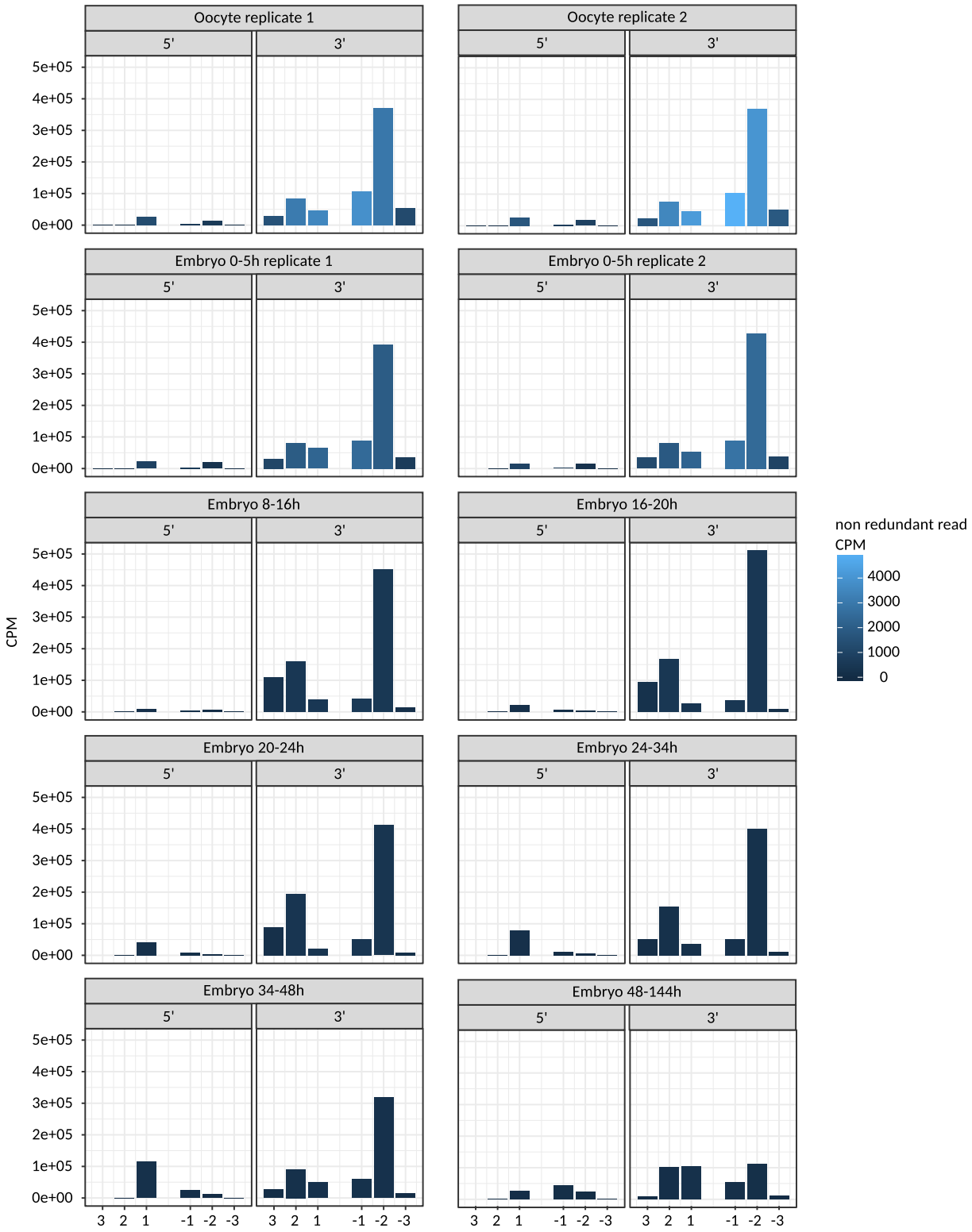


Figure 14 **Templated 3' and 5' additions and deletions.** The x-axis shows truncation in −1 steps and elongation in +1 steps and the y-axis shows the counts per million reads. The bar color displays the counts per million values of non-redundant reads supporting each miRNA variant

During embryonic development, we observed a significant increase in the abundance of single-nucleotide mismatches during the 20–24 h stage, with a rapid decline immediately afterwards. We therefore characterized this phase in more detail, revealing frequent A-to-C mutations especially at position 5–7 in the microRNA seed region, and at positions 10 and 17–21 (Figure 15). The latter segment lies directly behind the 3′ compensatory region (nucleotides 13–16) of the microRNA (Bartel 2009). In addition, we observed an increase in T-to-C, T-to-A and G-to-T transitions before the compensatory region, spanning positions 10–13.

We observed an increase in the expression of mature microRNAs during the last four phases, including tca-miR-10-5p (Figure S5). Furthermore, we observed an abrupt increase in the expression of tca-miR-376-3p, tca-bantam-3p and tca-miR-281-5p (among others) between the 34-48 h and 48-144 h phases. We observed an increase in the number of different mature miRNAs accumulating during each successive phase.

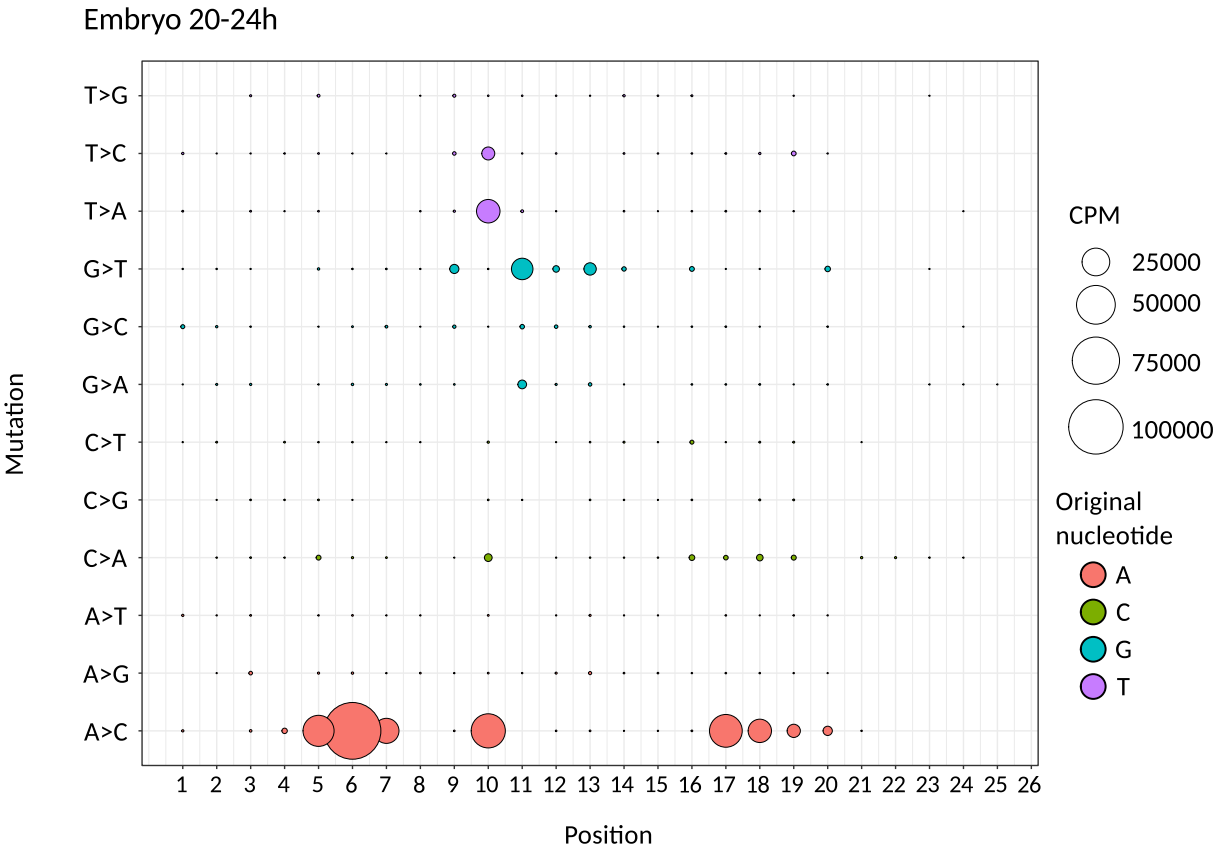


Figure 15 **Detailed characterization of miRNA SNP expression in the embryo during the 20-24 h phase.**

4.4 Discussion

We evaluated the performance of three algorithms for the identification of isomiRs in small RNA sequencing data (`isomiR-SEA`, `isomiRID` and `miraligner`) and used the most suitable of the three (`miraligner`) to generate an overview of the isomiRome of the red flour beetle *Tribolium castaneum*. All three tools found it difficult to process technical errors, probably because we clustered the identical reads. This step reduced the number of correct reads to single copies, shrinking the majority of reads. All the unique mutations and mutations with few copies were also reduced to a non-redundant set. Therefore, only one copy of each original miRNA remained in the data along with multiple variants with one or more sequencing errors. This may have increased the number of false negatives because the missed sequences presumably lay outside the scope of the algorithms due to the higher error rate as expected from isomiRs. False negatives were therefore weighted as neutral for the scoring process. Although a sequencing error can mislead the results of the study, we considered is a benefit, when the tools were able to assign it. Later analysis may then filter out possible erroneous reads to improve the investigation results. The evaluation of biological variants characterized the partially strong effects of sequence variations on the accuracy of isomiR identification. Both `isomiRID` and `miraligner` performed well, although `miraligner` was unable to identify all isomiRs with 3′ and 5′ deletions probably reflecting the seed-based search method. In contrast, `isomiR-SEA` performed poorly when mapping 5′ deletions and seed-mutated isoforms, but this was expected because the algorithm uses seed-based clustering for every miRNA and builds its entire analysis on these sets. Each of the algorithms demonstrated particular strengths for specific applications. Although `isomiR-SEA` achieved the weakest overall evaluation score, it is likely to be the most promising tool to screen for diverse and highly mutated isomiRs because it is the only software that supports more than one mismatch. It is also the only tool that uses just the read sequences and a single sequence file with all already known mature microRNAs. This makes it ideal for non-model organisms, especially compared to `isomiRID`, which requires a genome file in addition to the files from `miRBase`. We assume that the visual output of `isomiRID` is designed for the manual evaluation of a small set of microRNAs. Because it is based on the bowtie1 aligner, it can only report one type of isoform per read and will not recognize combined mutations such as a mismatch combined with a templated 3′ addition. This can be checked visually but such combinations are

not easily parsed by a pipeline. Finally, `miraligner` offered the best features of the other algorithms. It had a structured output comparable to `isomiR-SEA`, and scored nearly as much as `isomiRID` in terms of performance. It also makes use of `miRBase` files, but does not need a genome reference like `isomiRID`. Having evaluated and compared all three algorithms, we then used `miraligner` to characterize the *T. castaneum* isomiRome during embryonic development. Our analysis revealed that the isomiRome is more diverse and dynamic than previously reported. We were able to reproduce earlier reports that polyadenylated miRNAs are expressed in the oocyte and during the first embryonic phase. We found that the number of isomiRs with 5′ extensions increases during the 24–34 h and 34–48 h phases, which may cause a seed shift in the miRNAs and therefore modify the range of mRNA targets. We also observed a high mutation rate within the seed region during the 20–24 h phase which would also have a strong effect on the range of mRNA targets. Many miRNAs showed a surge in expression during the last four phases, suggesting a greater need for those miRNAs before hatching. Those observations would now need to be investigated by target verification methods such as cross-linking immunoprecipitation.

4.5 Conclusion

We evaluated the isomiR detection algorithms `isomiR-SEA`, `isomiRID` and `miraligner`, which are freely available and suitable for integration with local pipelines. We found that each program has advantages and disadvantages. Although `isomiRID` achieved the best performance against our evaluation criteria, the detailed visual output is more suitable for smaller datasets or the selected analysis of a few miRNAs. In contrast, `isomiR-SEA` gained a low score overall, but it allows the analysis of diverse mutations in large datasets because it accounts for more than one mutation in each miRNA, and because it can be run with only one file of mature miRNAs it is ideal for non-model organisms. Finally, we selected `miraligner` because it achieved a high-performance score and its clear output is ideal for pipeline integration. We used `miraligner` to screen the publicly available small RNA dataset of early development stages from *T. castaneum*, revealing the dynamic expression of isomiRs at each phase. These isomiRs must now be investigated in more detail to determine their biological functions.

5

DEVELOPMENT OF THE MICRORNA ANALYSIS PIPELINE

Before releasing the actual microPIECE pipeline, I considered the findings I made during the best practice investigation and my previous benchmarking of microRNA isoform detection tools, to develop a scripted workflow in an evolutionary prototyping manner. The evolutionary prototyping programming approach is defined as a method in software development that quickly leads to first results and feedback concerning the used techniques and solutions. It is constantly evolved further, based on experiences and feedback that arose through usage and testing.

The initial workflow was a compilation of LINUX shell and PERL scripts consecutively calling the different tools and performing the different analysis steps (Figure 16). For each script, I wrote a markdown file, to explain its operation and settings. The workflow at that early stage consists of 20 LINUX shell scripts, calling open source tools and 20 custom PERL scripts, which I will describe in the following (see supplemental material for pseudo codes). The custom formatting scripts were also tested by usual input-output test-cases to ensure a stable behavior (scripts not shown).
The final workflow results, as well as the later pipeline results, were then reformatted and imported into a MySQL database. From there it is also possible to answer further questions more easily, compared to the use of flat files. This also enables the quick analysis of microRNA arm-switching events and a range-flexible polycistronic cluster identification.

Outgoing from this stage, the actual microPIECE pipeline was developed in order to account for more species and an easier installation and usability via Docker.

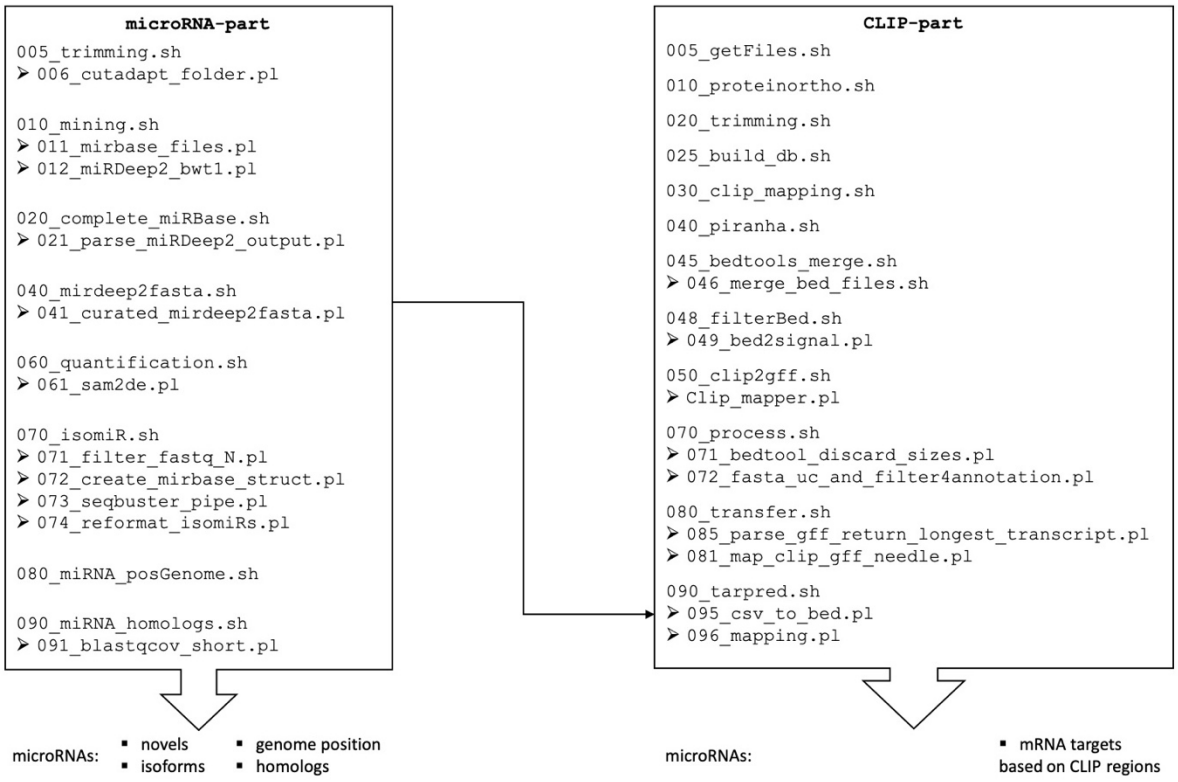


Figure 16 **Overview of scripted analysis workflow** The microRNA part runs several scripts to draw out information from the small RNA sequencing datasets and tries to complete existing information from public databases. The final set of microRNAs is used for mRNA target prediction in the CLIP-part that previously shrinks the search space of possible binding sites to CLIP-seq approved sites from related species.

5.1 Scripted workflow

As primary input, the workflow expects raw `FASTQ` files created with a small RNA sequencing protocol on an Illumina sequencing machine, with or without replicates from at least one condition.

In the first step, the workflow performs an adapter trimming with `Cutadapt` and filtering against unwanted non-coding RNA in the small RNA sequencing data with `bwa aln` (`005_trimming.sh` calling `006_cutadapt_folder.pl`).

The workflow uses `cutadapt` with the parameters `--trim-n`, to remove trailing n characters and `--minimum_length 17`, to filter for a minimal read length of 17. For the filtering step of ncRNAs, I used a database that included all known rRNAs for each species (Quast et al. 2013), but the workflow is designed in a manner that the user is able to provide a multi-fasta file of unwanted ncRNAs at own liking. The filtering of the mapped reads was performed, using `samtools` (H. Li, Handsaker, Wysoker, Fennell, Ruan, Homer, Marth,

Abecasis, Durbin, et al. 2009), retaining only the reads that did not map to the non-coding RNA data.

Coming now to the question of the detection of novel microRNAs in the small RNA sequencing data (`010_mining.sh`), the workflow calls `miRDeep2`. In advance to the `miRDeep2` run (`012_miRDeep2_bwt1.pl`), the script downloads the current `mature.fa`, `hairpin.fa` and `organism.txt` files from `miRBase.org` as input. Together with the user provided species 3-letter code, it generates the mature and precursor microRNA reference files for the species, as well as the mature microRNA sequences of all other species. Only metazoans are retained in the reference sets (`011_mirbase_files.pl`). This data is used as input for the following script, together with the concatenated `.fastq` file of all small RNA sequencing datasets and the reference genome. The result is the `miRDeep2` output, containing potentially novel microRNAs.

Since `miRBase.org` is a constantly evolving database, some entries may be incomplete. This means that for some microRNA precursors, only one mature sequence is annotated so far. The mature microRNA is then mostly called `miR-1` instead of `miR-1-3p` or `miR-1-5p`. Within the following step (`020_complete_miRBase.sh`), I try to close this gap with the provided small RNA sequencing data, resulting in a novel annotation that can be submitted to `miRBase.org`. Another problem is the notation of identical precursor copies (for example `mir-2a-1` and `mir-2a-2`). The resulting mature sequences would be named `mir-2a-3p/-5p`. Since this could lead to complications, the user can specify those microRNAs and the mature would be renamed accordingly.

In the following script, the previously mined microRNAs are parsed from the `miRDeep2.csv` output and merged into the modified mature and precursor microRNA file. The user may define a cutoff value for the score of novel `miRDeep2` microRNAs, alternatively a score of `10` is assumed by default.

After the enhancement of the microRNA data for the species with novel and re-annotated microRNAs, the workflow now performs a quantification of expression for the microRNAs (`060_quantification.sh`). As previously stated, the mapping of the reads is done by `bwa`. To avoid errors due to an RNA alphabet, the mature microRNA dataset is converted into DNA code in advance. Afterwards, the resulting file is filtered for mapping reads with `samtools view` and each read that has multiple potential mapping locations is written into a separate line by `xa2multi.pl` (taken from

https://github.com/lh3/bwa/blob/master/xa2multi.pl). The calculation for ReadsPerMillion (RPM) is performed by the `061_sam2de.pl` script. The script uses a tab-separated config file as input, to merge replicates and to report the expression per microRNA per condition (Table 9).

Table 9 **Config file for RPM calculation** Tab-separated list of files and the according condition

`path/to/condition1_rep1_smRNA.fastq`	`condition1`
`path/to/condition1_rep2_smRNA.fastq`	`condition1`
`path/to/condition2_rep1_smRNA.fastq`	`condition2`
…	…

The microRNA isoforms were detected, using the previously benchmarked `miraligner` tool in the `070_isomiR.sh` script. In advance to the analysis, remaining undetermined nucleotides have to be removed from the small RNA sequencing reads (`071_filter_fastq_N.pl`). Tailing undetermined reads were already removed in the trimming step. Since `miraligner` needs the hairpin structure from the `miRNA.str` file from `miRBase.org`, the information had to be pre-computed for the potentially new detected microRNAs and appended to the `miRNA.str` file (`072_create_mirbase_struct.pl`). The microRNA precursor is folded by `RNAfold` and the simulated binding free energy is written to the header of the structure, together with the positions of the mature microRNA sequences in the precursor. The mature microRNA sequences further had to be written in upper case letters in the secondary structure, whereas all other nucleotides are in lower case letters. Then the microRNA isoforms are identified by running `073_seqbuster_pipe.pl`. The script runs each small RNA sequencing file individually. Afterwards, all results from one condition are reformatted, taking the ID of the condition as further input (`074_reformat_isomiRs.pl`). The resulting reformatted output has expression values in ReadsPerMillion that are averaged over the small RNA sequencing replicates of the condition.

It is also of interest where the precursor sequences are located in the genomic landscape. Especially for the determination of polycistronic clusters. This is therefore computed in the `080_miRNA_posGenome.sh` script. Depending on the literature, the used distances between two microRNA precursors varies greatly. Therefore, we do not predefine this value and let the user decide a credible distance. First, a `BLAST` database is created from the genome

and a BLASTN search is run with the microRNA precursors as query. Afterwards, the resulting file is filtered via a LINUX shell script. Due to the inhomogeneous scientific opinions about the polycistronic cluster sizes, no evaluation is performed at this time point. Instead, it is intended to query the clusters dynamically, when the results are loaded into a database.
Finally, homologous microRNAs are determined, by comparing the set of mature microRNAs from the species with the mature microRNAs from miRBase.org via a BLASTN search in the 090_miRNA_homologs.sh script. This also includes the novel microRNAs, detected previously.
The microRNA analysis, based on the small RNA sequencing files, is finished now. Nevertheless, it is still necessary to identify potential mRNA targets of those microRNAs.

To account for this very challenging task with many approaches and a large number of false-positives, I present here a novel and conservative way of determining microRNA targets with a CLIP Analysis workflow part. This new method transfers the information of laboratory verified AGO binding sites from one species to the species of interest. Based on these transferred locations, a target prediction is performed.
The first script (005_getFiles.sh) helps to download all necessary data, using the LINUX wget command and the sra-toolkit from NCBI. If the data is already available, it is not necessary to run the script.
In order to identify homologous proteins between the species, donating the AGO-CLIP information and the species of interest, a ProteinOrtho run is performed within the script 010_proteinortho.sh. The previously downloaded multi-fasta files, containing the protein sequences of both species, are used as input. In advance, a BLAST database is calculated from both of the files.
For the analysis of the CLIP data, the CLIP-seq files first need to be trimmed from the artificial adapters in the script 020_trimming.sh. Here, cutadapt is used. The minimal length of the retaining reads was set to 20 and the trimming of terminal undetermined nucleotides was enabled.
Afterwards, the reference genome of the AGO CLIP-seq data needs to be prepared. Therefore, an indexing database is built, using the gsnap database builder gmap_build in the 025_build_db.sh script. The -g parameter allows the usage of a .gzip compressed genome and the -k 15 parameter sets the kmer size to 15.
After the successful build of the reference index database, the actual mapping of the AGO CLIP-seq reads is performed in the script 030_clip_mapping.sh. This is done by using

gsnap with the parameters -N 1 -B 5 and --speed 1. The -N parameter allows novel splice sites with the value 1, -B uses the batch mode 5 and --speed 1 ensures the highest accuracy by trading in computational speed. The output of gsnap is on the fly transformed into a sorted .bam format with samtools. Afterwards, it is converted into the .bed format by bedtools bamtobed.

The mapped reads have to be evaluated for their signal of potential binding regions (040_piranha.sh). Therefore, they are provided to Piranha (Uren et al. 2012), a so-called peak-calling algorithm that identifies the binding regions based on the read specialties. I chose Piranha because it has the ability to treat HITS-CLIP, PAR-CLIP and iCLIP sequencing data as well. The output is afterwards sorted with the LINUX shell command sort in order to be compatible to the following analysis and treatments.

The following 045_bedtools_merge.sh script merges the CLIP signal from the .bed files from each dataset into a single .bed file (046_merge_bed_files.pl). The script loops through all CLIP signal .bed files and regions. It remembers the positions from the regions and incrementally counts how often a certain nucleotide has been observed. This information is then appended to one new .bed file.

Initially, at an early developing stage, all .bed files were simply merged and signals in the new .bed file were raised from all CLIP dataset files that have been used.

The next script (048_filterBed.sh) produces multiple .bed files and each of them stands for one support level. The support level reflects the number of datasets that prove a certain signal position in the .bed file. Having six CLIP dataset files, the script will produce six .bed files with support level 1-6. The support level 1 means that at least one of the CLIP datasets raised the signal for a certain region. In support level 6, all regions in the .bed file were supported from all six CLIP datasets. The support levels increase or decrease with the number of supplied CLIP datasets.

The identified AGO binding regions further need to be assigned to a transcript for further processing (050_clip2gff.sh). Therefore, the .GFF annotation file from the donor species is parsed. Then, the script loops through the .bed file and checks if at least half of the region is hitting a mRNA transcript. If so, the ID of the mRNA is appended to the line in the .bed file.

Now that the regions are annotated, the following 070_process.sh script filters the regions by their length, retaining only those with a length between a minimal and maximal length, to account for plausible binding region sizes (071_bedtool_discard_sizes.pl). A

measurement showed that the vast majority of the binding regions had a size between 22 and 50 (data not shown). Next, the `.bed` file is sorted and is used for the extraction of the `.fasta` sequences from the `.bed` regions, by using `bedtools getfasta`. The resulting multi-fasta file is then filtered for those regions that map a mRNA gene (`072_fasta_uc_and_filter4annotation.pl`). Furthermore, all nucleotides are rewritten in upper case.

In the `080_transfer.sh` script, the actual transfer of the previously identified AGO CLIP binding region is computed using `needle` with the parameters `-endweight Y -gapopen 5 -gapextend 2` (`081_map_clip_gff_needle.pl`). In advance, the longest mRNA transcript isoform is identified in both species, by parsing the `.GFF` files (`085_parse_gff_return_longest_transcript.pl`).

The resulting needle output has a custom `.CSV` format, consisting of several mapping statistics, like coverage and identity, but also mapping positions. From them, the following script (`090_tarpred.sh`) creates a `.bed` file for further processing (`095_csv_to_bed.pl`). Then `bedtools merge` is used to merge transferred clip regions together. With `bedtools fasta`, the nucleotide sequences from the transcriptome and the `.bed` file are extracted into a `.fasta` file. Afterwards, `miranda` is used to predict the targets of the previously created microRNA dataset from the transferred sequences (`096_mapping.pl`). The target prediction is performed by `miRanda`, because it considers binding energies, but also specifically weights the binding of the 5' end of the miRNA and has no restriction towards certain parts of the mRNA, like `TargetScan` only accepts 3' UTR regions.

For further storage and easier visualization of results, the resulting flat files can now be loaded into a database, whose concept is explained in the following chapter. Depending on the scientific question and extent of datasets, this might also not be necessary and the analysis of the output files and forwarding into individual statistic calculations is sufficient. A possible upload script in pseudocode can be seen in the supplemental material chapter 13.7 .

5.2 Database

As enhancement to the microPIECE pipeline, I developed a MySQL (see supplemental material 13.7) database (Figure S 6) that can hold the information, generated by the pipeline and also synergistically can use the provided information to generate novel results and insights, like polycistronic cluster information or microRNA arm switching events from one condition to another. Without the database structure, one would need further tools or scripts in order to investigate those issues and therefore create static results that cannot be changed easily when a threshold, like the polycistronic cluster distance, is not fitting properly or when the research question slightly changes. A re-run of the software would be necessary. Especially for the above-mentioned analysis, a database that stores unbiased results, is very useful. As previously stated, the polycistronic cluster distance is not defined uniformly in the literature. It can therefore be necessary to search pre-miRNA distances that fit for the individual organism or condition. In case of arm switching events, it is maybe not only important to see a real switch, but it could also be of interest, when the less expressed arm is higher expressed than usual. The database structure can also hold the information of different genome versions or different animals with the same condition that can be compared with each other. With a growing database with many different results of various conditions and animals, one would also be able to investigate correlations between conditions and species. Besides, the database also helps to organize the data and keeps the information easily available, even when the analysis has been performed some time ago.

In the scheme (Figure 17), tables are colored differently, according to the type of data they contain. Green contains information about microRNAs, blue is for mRNAs and yellow has the information of the CLIP data and transfer. Orange tables contain additional information, like species, conditions or annotations for mRNAs, like GO classes. The grey tables are supporting tables to account for multiple relations. The dotted table, mRNA expression values, is currently not filled by the microPIECE pipeline, but I considered it a presumably useful information, if available in the research context and therefore included it.

The database holds the central table of microRNAs (table_miRNA – dark green). It holds the nucleotide sequence and potential miRBase.org IDs, together with the identifier for its type. This information is held in the table_miRNA_type. With this construct, the database is not only able to store the usual 5p/3p sequences, but also potential novel types, like offset microRNAs (elongated precursors with longer arms, producing more than one mature sequence per arm). The table_miRNA_homologs contains the information and BLASTN results of the search for homologous sequences in the microPIECE pipeline. Through the

NM_mature_precursor table, the mature sequences are connected to the table_miRNA_precursor. This NM_table accounts for the possibility that a mature microRNA can occur in more than one precursor, but also that one precursor has naturally more than one mature sequence. In a direct connection to the precursor table, I placed the table_species, holding the species information of the 3-letter code and complete name. Outgoing from the table_miRNA, the table_miRNA_expression is attached. It contains the previously computed expression value and is directly connected to the table_condition. This enables the storage of various expression values of the same microRNA, but for different conditions. Similarly, the table_isomiR is connected to the table_miRNA and also has an additional table_isomiR_expression that hold the expression values of the microRNA isoforms, being also connected to the table_condition. This again enables the storage of various microRNA isoforms in different conditions. Also connected to the table_miRNA is the table_targetprediction. This table contains the information, derived from miranda's target prediction run in the microPIECE pipeline. It is further connected to the table_clip, holding the information from the CLIP transfer. The table_targetprediction is also connected to the table_mRNA, having the mRNA IDs from NCBI, but more importantly the coding sequence start and stop positions, as well as the strand information. This would also enable a SQL query if the miRNA bound at the 3' or 5' UTR or to the coding sequence. Through the NM_mRNA_annotation table, the table_mRNA is connected to table_annotation and table_annotation_type. Those tables can hold further information on the mRNA, like GO classes or other sources.

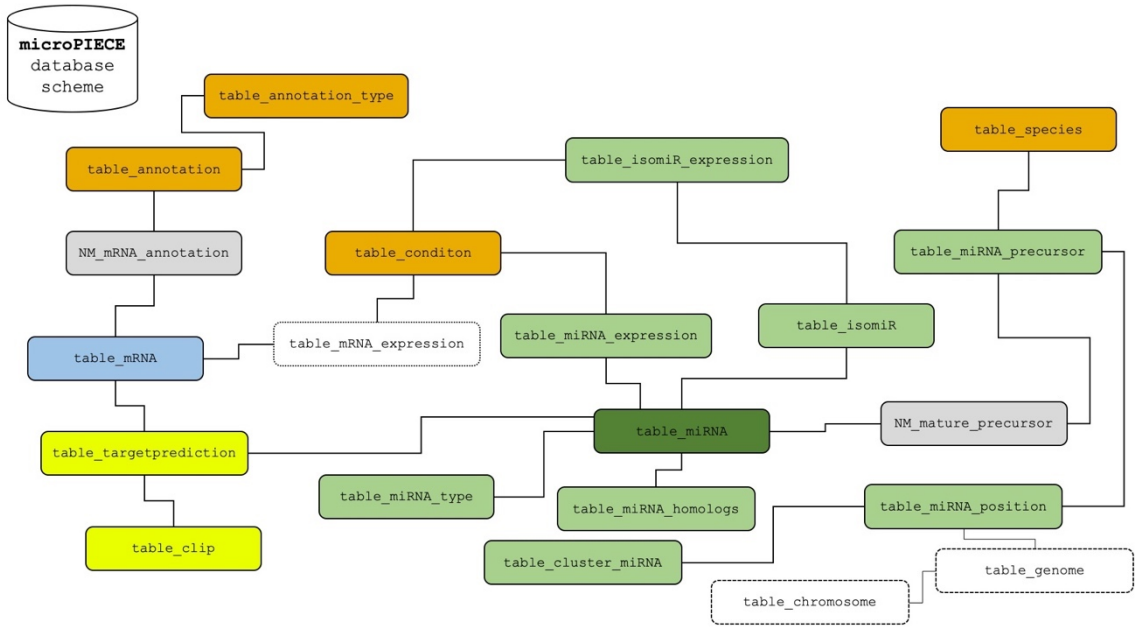


Figure 17 **Database scheme** The central table is the table_miRNA in dark green. It is connected, mainly with other tables, holding information about the miRNAs derived from the `microPIECE` pipeline in light green. Orange indicate tables that are enhancing the information, like species, further mRNA annotations or GO classes. Blue is for the mRNA information and yellow shows the information from the CLIP experiment. The dot-bordered tables (table_mRNA_expression, table_genome, table_chromosome) are included, but not officially considered in the `microPIECE` pipeline, yet.

The database is filled by a custom `PERL` script (see supplemental material chapter 13.7 for pseudo codes) that takes the data from the `microPIECE` pipeline and further sources. The script transforms this data into `SQL` statements and a final output for pushing it into the previously described database skeleton. It takes a config file as input, which includes the species name and its 3-letter code, as well as the genome name and download source. Further it has the microRNA arm types and conditions included. From this, the `SQL` statements for `table_genome`, `table_species`, `table_miRNA_type` and `table_condition` are created. With `table_genome`, putative versioning could be established, if an analysis is run again on another genome version or source. The `table_chromosome` stores information about the chromosome name and length and could be used for further studies on miRNA location statistics.

The script also parses the genome file to generate chromosome-name and chromosome-length tuples and derives the foreign key from the genome name of the `table_genome`. The next part, creates `SQL` statements for the microRNA tables and its connecting `NM_mature_precursor` table. First, the `table_miRNA_precursor` is filled and takes the species 3-letter code from `table_species` as foreign key. Then the `table_miRNA`

statements are created by filling the sequence and microRNA miRBase.org information into the SQL statements. The microRNA type IDs are derived from table_miRNA_type. For the NM_mature_precursor table, the IDs of table_miRNA and table_miRNA_precursor are pairwise assigned as foreign keys. The table_miRNA_position statements are created by parsing the position file and deriving the genome ID from table_genome as foreign key. The SQL statements in table_miRNA_expression are genereated by parsing the expression file and deriving the miRNA ID from table_miRNA as foreign key, as well as the condition ID from table_condition as foreign key. In the case of table_isomiR and table_isomiR_expression, the microRNA isoforms file is parsed and microRNA ID from table_miRNA is used as foreign key in the table_isomiR, whereas the condition ID from table_condition is used as foreign key in table_isomiR_expression.
The table_homologs is filled by the information, included in the homologs output file. It takes the miRNA ID from table_miRNA as foreign key. The table_mRNA data is taken from the .gff file, calculating the relative coding sequence start and stop positions from the parsed exon-CDS relationship. The target prediction file is parsed and for the table_targetprediction, the miRNA and mRNA IDs are taken from table_miRNA and table_mRNA as foreign keys. The table_clip foreign key of table_targetprediction is created during the statement creation dynamically. The comma-separated annotation file (Protein ID, annotation ID, annotation detail, annotation source, annotation source ID), previously prepared by the user with desired mRNA annotations, is parsed at the end and converted into SQL statements. The NM_mRNA_annotation table is filled with foreign keys from table_mRNA and table_annotation.

6

PUBLICATION II: THE MICRORNA PIPELINE microPIECE

Outgoing from the previously described scripted workflow, the script collection was transferred into the actual `microPIECE` pipeline with software-testing, example datasets and dockerized environment, representing a state-of-the art software publishing. An overview is described in my following manuscript, published in the Journal of Open Source Software (JOSS). This chapter is then followed by a detailed description of the individual steps of the current pipeline version. The pseudocode scripts are available in the supplemental material 13.8.

Daniel Amsel, André Billion, Andreas Vilcinskas and Frank Förster.
"microPIECE – microRNA pipeline enhanced by CLIP experiments."
JOSS - Journal of Open Source Software, 3(24), 616.
https://doi.org/10.21105/joss.00616

For this publication I created the basic scripts as scripted workflow and participated in re-writing the code into the actual pipeline. I furthermore created testcases for program modules, wrote code documentations, GitHub repository information. I also wrote the manuscript and draw the figure.

The pipeline is available via GitHub:
`https://github.com/microPIECE-team/microPIECE`

The Journal of Open Source Software

microPIECE - microRNA pipeline enhanced by CLIP experiments

Daniel Amsel[1], **André Billion**[1], **Andreas Vilcinskas**[1, 2], **and Frank Förster**[1]

1 Fraunhofer Institute for Molecular Biology and Applied Ecology, Department of Bioresources, Winchester Str. 2, 35394 Giessen, Germany **2** Institute for Insect Biotechnology, Heinrich-Buff-Ring 26-32, 35392 Giessen, Germany

DOI: 10.21105/joss.00616

Software

- Review
- Repository
- Archive

Submitted: 05 March 2018
Published: 10 April 2018

6.1 Summary

All microRNAs are assumed to be post-transcriptional fine-regulators. With a length of around 21 nucleotides, they form a RNA-induced silencing complex (RISC) complex with a protein of the Argonaute family. This complex then binds to the messengerRNA untranslated regions and coding sequence regions and in general promotes degradation or translational inhibition. It is now important to know the microRNA-mRNA pairs in order to infer dysregulating effects on the organism. In order to assign a microRNA to a mRNA target, various tools with different technical approaches were developed. They are mostly based on the assumption that the first eight nucleotides of the microRNA (seed region) determine the binding region on the mRNA. Some approaches also include supporting bindings in the rear part of the microRNA, others take secondary structures of the mRNA or binding energies of the mRNA-miRNA complex into account. Nevertheless, they all suffer from the statistical problem that such short target regions, often occur simply by chance in transcript sequences. This results in a huge amount of false positive predictions. A target prediction of all 590 *Tribolium castaneum* mature microRNAs from `miRBase.org v22` (Kozomara and Griffiths-Jones 2014) against all 18.534 protein coding cDNA sequences from `Ensembl.org` (Ensembl Genomes release 38 – December 2017) (Kinsella et al. 2011) results in 2.948.255 possible microRNA-target interactions, predicted by the commonly used tool `miRanda` (Betel et al. 2007a) with standard parameters. To increase the credibility, wet lab validation methods like luciferase reporter assays are required. The disadvantage here is that this workflow is not applicable for high-throughput analysis, as it can only treat small subsets of sequence combinations. Another, more scalable method is cross-linking immunoprecipitation-high-throughput sequencing (CLIP-seq). Here, binding regions of the RISC show a specific signal in the sequencing reads that can be used to shrink the search space of miRNA target predictions, when mapping them to the transcriptome. The limitation here is the difficult technical treatment in the laboratory. This is the reason why there are only a few datasets available for human, mouse, worm and mosquito. It would now be useful, if we could simply transfer the information of a binding region, already identified by CLIP-seq, to another species. This is what our **micro**RNA **pi**peline **e**nhanced by **C**LIP **e**xperiments `microPIECE` is about.

The pipeline (Figure 18) takes the AGO-CLIP data from a *speciesA* and transfers it

to a *speciesB*. Given a set of miRNAs from *speciesB* it then predicts their targets on the transferred CLIP regions.

For the *minimal workflow* it needs a genome file, as well as its annotation file in GFF format for *speciesA* and *speciesB*. For *speciesA* at least one AGO-CLIP dataset is needed and *speciesB* needs a set of miRNAs for the target prediction. For the *full workflow*, a set of smallRNA-sequencing data is additionally needed and a set of non-coding RNAs can be provided as filter. The pipeline uses the smallRNA data for the mining of novel microRNAs and the completion of the given miRNA dataset, if needed. It further performs expression calculation, isoform detection, genomic loci identification and orthology determination. In case of a provided smallRNA dataset, the pipeline starts with the miRNA analysis. It uses `Cutadapt` (M. Martin 2011) to trim the adapter sequences from the small RNA sequencing libraries from *speciesB*. If provided, the trimmed libraries are filtered for ncRNAs using `bwa` (H. Li and Durbin 2009). The resulting files are merged into a pooled set and used for mining of novel microRNAs with `miRDeep2` (Friedländer et al. 2012). The pipeline then parses the result file and tries to add missing entries from the `miRBase.org` database, e.g. if only one arm was previously annotated and the mining discovers the exact position of the particular arm. The novel miRNAs and completed entries are merged to the existing miRNA set and used as reference for the following analysis. The expression of each miRNA is calculated in RPM, outgoing from the non-pooled trimmed and filtered libraries. The pipeline also accounts for miRNA isoforms, by removing all trimmed reads, containing undetermined nucleotides. Orthologous miRNAs in other

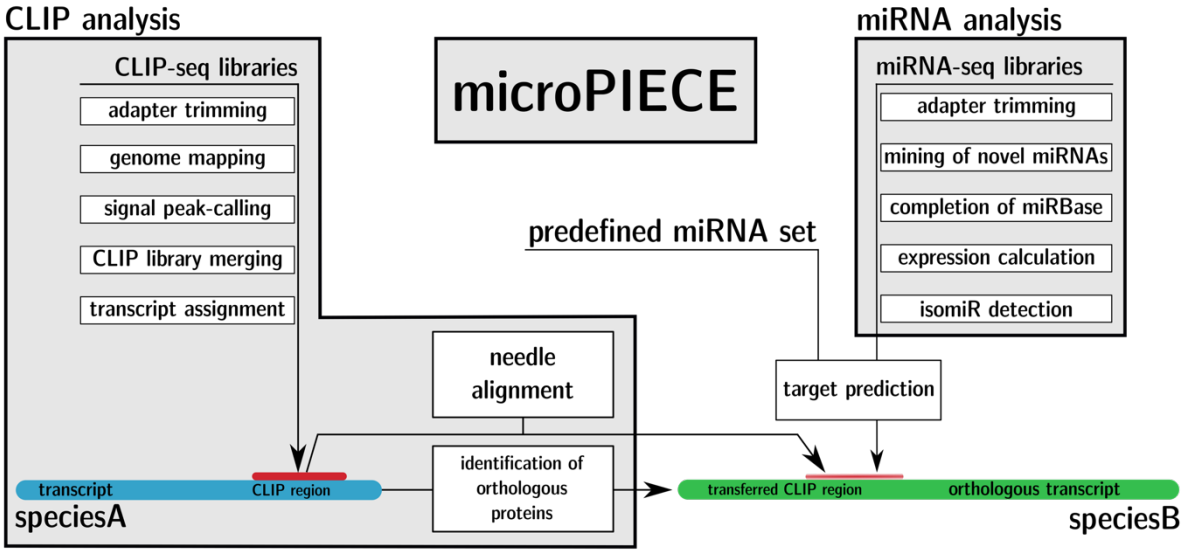


Figure 18 **Scheme of the microPIECE pipeline.** The user can choose either to provide smallRNA sequencing libraries or solely a set of known microRNAs in addition to the CLIP-seq libraries.

species were determined by a BLASTN (Camacho et al. 2009) search against all metazoan miRNAs from miRBase.org. Finally, the genomic regions for the miRNAs were also identified by a BLASTN search against the genome of *speciesB*.

If no smallRNA dataset is provided, the pipeline directly jumps to the CLIP analysis. There it starts with the *speciesA* CLIP-seq library trimming, using Cutadapt (M. Martin 2011). Trimmed reads are then mapped to the genome with gsnap (T. D. Wu and Nacu 2010) and the results are evaluated by Piranha (Uren et al. 2012). Then the libraries are merged into the BED file format. We further used SAMtools (H. Li, Handsaker, Wysoker, Fennell, Ruan, Homer, Marth, Abecasis, and Durbin 2009) and BEDtools (Quinlan and Hall 2010) for file conversions during the pipeline. The BED file includes a column that displays how many libraries support each genomic position. Next, a file for each library-support-level is created, so that the user can in the end decide how many CLIP libraries are necessary to account this region as binding region. Now for each library-support-level, an assignment of each sequence to the transcriptome is performed. Outgoing from the transcript, the corresponding protein is used to discover the orthologous protein in the *speciesB* by Proteinortho (Lechner et al. 2011). This information is used as criteria to align the CLIP region from *speciesA* to the orthologous transcript in *speciesB* with EMBOSS Needle (Rice, Longden, and Bleasby 2000).

Finally, the pipeline passes the miRNA set (either from the full or minimal workflow) to miraligner from the seqbuster package (Pantano, Estivill, and Martí 2010). Then a target prediction with miRanda (Betel et al. 2007a) on the previously transferred orthologous CLIP regions is performed.

Depending on the provided data, the minimal output of the pipeline consists of a target prediction output from miRanda for each library-support-level, based on the transferred CLIP regions. In case the pipeline additionally received smallRNA data, a microRNA set with known and novel miRNAs together with an expression file is saved, as well as the orthologs to other species, the genomic loci of the miRNAs and the identified isoforms.

As an example case, we used microPIECE on the AGO-CLIP data from *Aedes aegypti* (X. Zhang et al. 2017) and the 590 *Tribolium castaneum* miRNAs from miRBase.org v22. The target prediction resulted in 17692 miRNA-target interactions with a one of six library-support-level. Outgoing from the previously reported 2.948.255 possible microRNA-target interactions, we reduced the results to a very conservative set.

7

THE microPIECE PIPELINE IN DETAIL

The microPIECE pipeline, introduced in the previous chapter, is used on the LINUX/UNIX command line with the call of the identically named PERL script microPIECE.pl (see supplemental material chapter 13.8 for pseudo codes). This script consists to a large extent of Plain Old Documentation (pod), describing the pipeline, but also the different arguments and example runs.

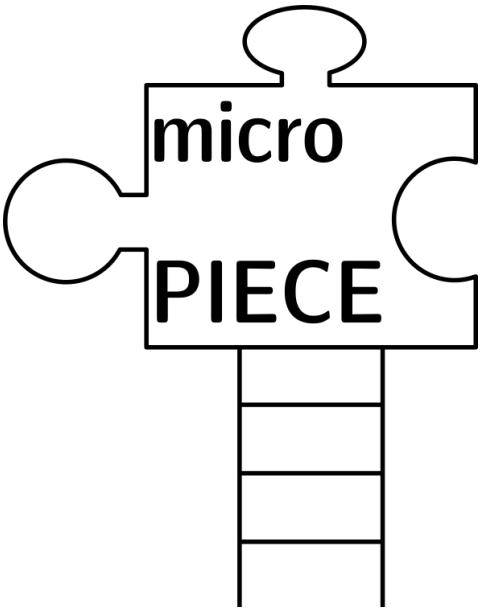


Figure 19 **microPIECE logo** The logo has roughly the shape of a microRNA hairpin, mixed with a piece of a puzzle.

The execution part of the script takes all the parameters as input that are needed for the pipeline via the GetOpt PERL module. This enables the possibility to use options with starting dashes to define the arguments, instead of simply parsing the arguments in a previously expected order after the script-call in the BASH command line. The previous scripts from the workflow were grouped together in PERL modules, called microPIECE.pm and mining.pm (see supplemental material chapter 13.8 for pseudo codes). These modules are imported at the very beginning and the individual routines are called via the custom PERL module together with the provided option parameters.

In the following, the pipeline is described in a more detailed fashion.

7.1 Algorithm Description

The microPIECE pipeline is designed very flexible, depending on the data that is provided, when calling the algorithm. At the very beginning, the scripts checks what kind of data is provided to run the script, by calling microPIECE::check_requirements subroutine from the microPIECE perl module (microPIECE.pm). At first, the subroutine calls the check_files subroutine, which closely inspects if the stated files physically exist on the hard disk. Next, the check_requirements subroutine checks for the types of arguments that are given. Depending on that, different branches of the pipeline are used. In detail, the pipeline will terminate, if no CLIP data is provided. In case of the miRNA data, it differentiates between a miRNA .fasta data set and small RNA sequencing data. In the latter case, the entire microRNA analysis procedure, including mining steps, are involved and in the other case, the mining part is skipped, due to the lack of data. Furthermore, tool dependencies are checked and the pipeline terminates, if they are not fulfilled. This was implemented in case of a local installation, instead of a dockerized environment which always contains the entire dependency tree. Finally, the provided output folder is checked and the pipeline terminates, if it is already occupied, except the --overwrite parameter was set. After completing the requirements check, the defined settings are printed into a log-file.

If the small RNA sequencing data was provided and detected by the pipeline in the check_requirements subroutine, the microRNA mining part is initiated within the microPIECE::run_mining subroutine. This subroutine is then calling various subroutines again for the different tasks, accompanying with this part of the pipeline. It starts with the creation of the output folders, by calling the create_folder subroutine. Afterwards, the actual treatment of the files starts.

The small RNA sequencing files are first cleansed from the artificial adapter sequences by run_mining_clipping. This subroutine calls the external program cutadapt, which removes the adapter sequences and retains only reads with a minimal length of 17 after trimming. It further removes undetermined nucleotides (N characters) from the ends of each read. Finally, the data from each small RNA sequencing file is written into an output file.

Next, the subroutine run_mining_filtering maps the reads to a user-provided set of non-coding RNA sequences, to keep only those reads that do not map to the contamination. This mapping is done with bwa and the filtering is performed with samtools. Bedtools is used to transfer the mapping results of the unmapped reads back to the .fastq file format. With the run_mining_downloads subroutine, the pipeline verifies that the

miRBase.org files organism.txt.gz, mature.fa.gz, hairpin.fa.gz and miRNA.dat.gz are either available or downloaded via the pipeline subroutine get_mirbase_download_or_local_copy. Outgoing from the miRBase.org files, the reference files for the microRNA mining tool miRDeep2 are created via run_mining_mirbase_files and the MINING_split_mirbase_files.pl script. Initially, the organisms.txt file is parsed for metazoan species identifiers. These are used to filter the mature.fa and hairpin.fa for metazoan species. Furthermore, the actual species of interest, described via a user-provided 3-letter code, is also selected and pasted into separate files for mature and precursor sequence. Now that the required input files for miRDeep2 are gathered, the mining process is started from the pipeline with the subroutine run_mining_mirdeep2. This step starts with the mandatory criteria from miRDeep2, namely the removal of whitespaces in the headers of the sequence files with the miRDeep2 PERL script remove_white_space_in_id.pl. The miRBase.org files are also translated from the RNA to the DNA alphabet via the subroutine run_mining_rna2dna. From the genome .fasta file, a bowtie index is created with bowtie-build. If more than one small RNA sequencing file is provided, all files are merged into one file and translated into .fasta format, using fastq2fasta.pl from miRDeep2. Here, the whitespaces in the headers are also removed with remove_white_space_in_id.pl from the miRDeep2 package. In the following, identical reads are collapsed with the miRDeep2 script collapse_reads_md.pl. Finally, the reads are mapped to the genome, using bowtie within the miRDeep2 script mapper.pl. Afterwards, the mapping results are scanned for novel microRNAs with the actual miRDeep2 script miRDeep2.pl.

In order to account for lacking information from miRBase.org, the pipeline now tries to complement microRNA precursors that lack a second mature strand annotation with run_mining_complete and the MINING_complete_mirbase_by_miRDeep2_output.pl script. The missing annotation is derived from the parsing of the miRNA.dat file from miRBase.org, using the mining::parse_mirbase_data subroutine. The precursors without two mature sequences are identified with mining::parse_mirdeep_known. If a potential mature microRNA was found that would complete a precursor sequence, it has to fulfill the high-confidence criteria, defined by miRBase.org (http://www.mirbase.org/blog/2014/07/high-confidence-mirna-set-available-for-mirbase-21/ ; published July 3, 2014 ; last access September 1st, 2019). Namely, either each mature sequence has >= 10 reads support or the sum

of both mature sequences is >= 100 and each as >= 5 reads support. The mature microRNAs are exported into .fasta format with mining::export_fasta and the novel .dat file is written by mining::export_mirbase_data.

The novel microRNAs are combined with the already known ones from miRBase.org by running the subroutine run_mining_mirdeep2fasta and MINING_curate_mirdeep2fasta.pl.

Now with the complete set of novel and known microRNAs, the expression per condition is computed via run_mining_quantifiaction. The bwa index command is run on the mature microRNA .fasta data and a bwa aln alignment with optimized parameters for microRNA detection is performed. The bwa aln output is transformed into .sam format, using bwa samse. The unmapped reads are removed with samtools view. Multimapping reads are transferred from a one-line notation into a multi-line notation, representing one line for each hit. Finally, the Reads Per Million values are calculated for each condition with MINING_sam2de.pl, also accounting for replicates.

Outgoing from the complete set of microRNAs, novel and known ones, the pipeline searches for microRNA isoforms, called isomiRs, using run_mining_isomirs. In order to include the novel microRNAs, the .str file from miRBase.org also needs to be extended with the novel microRNAs. This is done by ISOMIR_create_mirbase_struct.pl. Within this PERL script, mining::parse_mirbase_dat is used to parse the .dat file for mature and precursor sequence. RNAfold, an external tool from the ViennaPackage (Hofacker et al. 1994; Lorenz et al. 2011), is used to compute the free energy and dot-bracket fold. The RNAfold algorithm is a scanning tool that is used on long RNA sequences, to calculate the local stable substructures. RNAfold defines a maximum spanning distance between two mating base-pairs, computes the pairing probabilities of the nucleotides and averages over all binding probabilities in the sequence-windows, including this putative pair. It hereby calculates the secondary structure and total binding free energies that are used for the inclusion of the novel microRNAs to the structure file and microRNA isoform calculation.

The resulting dot-bracket fold is then used as input for the RNA::HairpinFigure module, an external PERL module is used to draw this secondary structure fold in a text format. Afterwards, the text format is checked for correct output, by comparing it again to the precursor sequence and the result is appended to the existing miRNA.str file from miRBase.org. Before the actual isoform mining, the reads with internal undetermined nucleotides need to be removed, since they would cause a misleading result. This is done by

filter_for_N_and_collapse_reads. Furthermore, the reads are collapsed and the information of the copy number is appended to the header. Finally, miraligner is run, accounting for one substitution and three nucleotides to be added or removed. Afterwards, the output of each condition is gathered, the average number of reads per isoform type in the condition is calculated and finally normalized with ReadsPerMillion, using the ISOMIR_reformat_isomirs.pl Perl script.

The genomic positions of microRNA precursors are determined by run_mining_genomicposition, using makeblastdb and blastn. The results are filtered for 100% identity and coverage.

For the identification of orthologous microRNAs, run_mining_orthologs with the PERL script MINING_ortholog_blast.pl is used to create a database with makeblastdb from the mature microRNA sequences from miRBase.org. A blastn search is performed afterwards, followed by several filtering steps. The minimal alignment length has to be 10. Query and subject need to have the same start. The first 10 basepairs are not allowed to have any gaps or mismatches. The remaining basepairs may have one gap or mismatch. Finally, the query and subject coverage is computed.

As previously said, this part is run, if, and only if small RNA sequencing data is provided. Otherwise the pipeline would directly jump to the now following CLIP part.

In the first step of the CLIP part, run_proteinortho determines the homologous protein sequences between the two species, the CLIP donor species and the CLIP receiver species, where we want to predict the microRNAs. For proteinortho, a blast database needs to be created for both protein datasets, using makeblastdb. The protein datasets are derived from the .gff files of each species, using gffread (see supplemental material 13.3).

The CLIP sequencing data needs to be trimmed, like every other sequencing data, to remove the artificial sequences. Therefore, the pipeline launches run_CLIP_adapter_trimming. It uses cutadapt with a minimal retaining read length of 20 and trimmed terminal undetermined nucleotides.

Afterwards, run_CLIP_build_db is used to create a database of the CLIP donor species genome, using gmap build.

Afterwards, run_CLIP_mapping is used to map the CLIP sequencing reads to the genome, using gsnap. The results are transformed into a sorted .bam format, followed by an indexing with samtools index.

The resulting mapping output is then used by PIRANHA, to compute the so-called peak-calling, a determination signaling regions from the read mapping. This is done within the subroutine run_CLIP_piranha. This subroutine uses the information of supplied CPU threads, in order to start a Piranha subroutine for each thread. The threaded subroutine is called run_CLIP_piranha_working_threads and calls the PERL script CLIP_binned_bed_from_bam_and_transcripts_for_piranha.pl in order to create pre-binned .bed files from the mapped reads first. Those pre-binned files can be supplied with pseudo-counts at transcript locations, in order to highlight exon regions for Piranha.

Outgoing from the situation, that more than one CLIP sequencing file was submitted, the run_CLIP_bedtools_merge subroutine merges the resulting .bed files into one, but retains the information, which condition gave rise to a signaling region, using CLIP_merge_bed_files.pl. The information is inserted into the fourth column of the .bed file. Whereas usual .bed files contain the „chromosome start stop“ information, the script inserts a row that contains „length=X;counts=A/B/C,D/E/F“. Here, A and D are the total counts of supporting positions, whereas B and E, as well as C and F represent the count for two conditions at position A/B/C and position D/E/F.

The prepared .bed file is then split into .bed files that represent different supporting strength, using run_CLIP_filterbed and CLIP_bed2signal.pl. Given exemplarily two conditions, a supporting strength of 1 would mean that at least one condition supports the region in the .bed file, whereas a supporting strength of 2 includes only .bed regions, where both regions agree.

The resulting .bed files are then filtered for regions that map at least with 50% to a coding transcript, using run_CLIP_mapper and CLIP_mapper.pl

In the following part, run_CLIP_process filteres for regions between 22 and 50 (default values) and discards them if they do not fit, using CLIP_bedtool_discard_sizes. The resulting .bed file is sorted and a .fasta file is created by using bedtools getfasta.

Afterwards, the run_CLIP_transfer subroutine uses the PERL script CLIP_parse_gff_return_longest_transcript.pl, to filter for the longest transcript of each gene. Afterwards, the exons are extracted, by using gffread. Then the peak regions are extracted and the homologous information is used to transfer this region to the other species, by using the CLIP_map_clip_gff_needle.pl Perl script and needle.

Finally, the output is translated into .bed format and overlapping regions are merged. The resulting regions are then used to extract the .fasta sequences of the final target regions.
The prediction of microRNA targets is computed in the subroutine run_targetprediction with the Targetprediction.pl PERL script. This script uses miranda to search potential targets for each microRNA sequence. The miranda output is condensed to the essential part with the target and its binding information.
After all, the resulting files are transferred to the final output folder via microPIECE::transfer_resultfiles.

The PERL scripts also include a PERL module, containing subroutines that are used frequently, like parsing some special data types. The module parse_mirbase_dat is designed to draw out several information from the miRBase.org .dat file for a given species in 3-letter code. Whereas parse_fasta simply creates a Perl hash structure from a .fasta sequence, with the header as key and the sequence as value. The modules parse_mirdeep, parse_mirdeep_novels and parse_mirdeep_known are designed to treat the .csv output file of miRDeep2 to get novel or known miRNA sequences. The subroutine export_fasta writes the mature and precursor sequences of the miRNAs into two .fasta files. A little different is the export_mirbase_data subroutine, which exports the miRNA information in the .dat format of miRBase.org. In the subroutine fix_hairpin, the hairpin structure of an external tool is checked and in case of need, a known, but unfixed bugged output, is corrected.

7.2 Comparing microPIECE to other tools and methods

The previously explained microPIECE pipeline comes with a large variety of features and algorithms that are combined into an all-in-one solution for microRNA analysis, which is not publicly and freely available so far. The selected tools for the pipeline have been tested and compared as best-performers for microRNA tasks and it can be assumed that their combination leads to very good results.
The highlight of microPIECE is not only the fact that it can be run locally in high-throughput mode and delivers all common results one can have with a microRNA smallRNA-seq dataset. With the CLIP-seq binding region transfer, it also has a novel approach in target prediction, which no other tool has used so far. In general, most comprehensive approaches are developed for human or human research related species, like mice and rats (Sticht et al. 2018). Here, the

microPIECE pipeline also clearly stands out, as it is not limited to a predefined species, but can be used for every species that has a sequenced and annotated genome.

With the microPIECE pipeline I use the accepted scientific opinion of conserved miRNAs and miRNA targets, their evolutionary conserved binding regions and the conserved regulatory processes (Enright et al. 2003) in order to transfer identified binding regions to evolutionary closely related species. Even invertebrates share around 10% of their miRNAs with mammals, which means that the regulation of those genes is presumably conserved, even though those species split evolutionary several million years ago (Enright et al. 2003). The strong conservation of those mature miRNA sequences leads to the assumption that one miRNA regulates more than one mRNA, explaining the very unlikely compensatory base-pair mutation (Enright et al. 2003).

I combined this doctrine with the knowledge about the CLIP technique for miRNA target prediction and search space shrinking, like previously mentioned (Chou et al. 2013). In contrast to many established methods, focusing the 3' UTR and the seed region of the miRNA, the CLIP approach in the microPIECE pipeline allows an unbiased target prediction in UTR and CDS regions. Already published CLIP experiments for miRNA target prediction highlighted the presence of miRNA binding positions in the coding regions and untranslated regions of a mRNA (Chi et al. 2009; Hafner et al. 2010). There are already existing tools and databases, tackling the miRNA prediction on CLIP areas, but they are also always limited to the species that was sequenced with the CLIP technique and do not contain further microRNA analysis options, like microPIECE does. There is for example miRTarCLIP, that has a locally usable pipeline for automatically analyzing CLIP data and perform target prediction steps (Chou et al. 2013). Here, microPIECE fills a gap by using those CLIP datasets, making them available for other species than the originally sequenced one.

Other approaches also make use of experimentally validated miRNA targets. Here, luciferase reporter assays are mainly used, by genetically inserting the possible binding region into a visual reporter in the laboratory. A successful binding prevents the luminescent reporter from being transcribed and no light signal is observed. Nevertheless, these results are generated more or less in a one-by-one manner and are therefore also very limited.

To my knowledge, there is no all-in-one microRNA analysis pipeline available, especially no pipeline that is comparable to the features and methods of microPIECE. The search-space-shrinking CLIP transfer method of the microPIECE pipeline, as well as the completion of miRBase.org by using small RNA sequencing data is novel in this field.

8

PUBLICATION III: APPLICATION OF microPIECE TO *GALLERIA MELLONELLA*

The release of the microPIECE pipeline itself enables the public community to use this novel target prediction technique in combination with automated smallRNA-seq analysis that accounts for the most questions in microRNA research. Subsequent to the development, I also applied my pipeline to various other projects. Amongst them was the investigation of microRNA immune response of *Galleria mellonella* to uropathogenic and commensal-like *Escherichia coli* strains. Within the following publication in Nature Scientific Reports, where I am the equally contributing first-author, I demonstrate the usability of microPIECE, identifying important immune response microRNAs with an outlook to transfer this information to human.

Krishnendu Mukherjee†, Daniel Amsel†, Miriam Kalsy, André Billion, Ulrich Dobrindt and Andreas Vilcinskas.
"MicroRNAs regulate innate immunity against uropathogenic and commensal-like *Escherichia coli* infections in the surrogate insect model *Galleria mellonella*."
†These authors contributed equally.

For this publication, I computed a reference-based assembly of the *Galleria mellonella* transcriptome with the newly published, but unannotated genome, described in the following chapter 8.5.4. For the annotation of the different isoforms, I used TransDecoder version 5.0.2 (Haas et al. 2013). The analysis was enhanced with the common analysis workflow, namely the BlastP (version 2.6.0+) (Camacho et al. 2009) search against SwissProt (evalue 1e-5, max_target_seqs 1; version from February, 28th 2018) (UniProt Consortium 2018) and a hmmer (Johnson, Eddy, and Portugaly 2010) search against Pfam-A (standard parameters; version from

February, 24th 2017) (Finn et al. 2016). The assembly evaluation was performed with rnaQUAST (Bushmanova et al. 2016). The transcriptome contains 18,620 sequences and has a total length of 21,599,142 nucleotides with a N50 of 1,578 and a N90 of 726.
I then annotated the transcripts using InterproScan (version 5.27-66.0) (Jones et al. 2014) in combination with CDD-3.16 (Marchler-Bauer et al. 2017), Coils-2.2.1 (Lupas, Van Dyke, and Stock 1991), Gene3D-4.1.0 (T. E. Lewis et al. 2018), Hamap-2017-10 (Pedruzzi et al. 2015), MobiDBLite-1.0 (Piovesan et al. 2018), Panther-12.0 (Mi et al. 2017), Pfam-31.0 (Finn et al. 2016), Phobius-1.01 (Käll, Krogh, and Sonnhammer 2004), PIRSF-3.02 (C. H. Wu et al. 2004), PRINTS-42.0 (Attwood et al. 2003), ProDom-2006.1 (Bru et al. 2005), ProSitePatterns-2017_09/ProSiteProfiles-2017_09 (Hulo et al. 2006), SFLD-3 (Akiva et al. 2014), SignalP 4.1 (H. Nielsen 2017), SMART-7.1 (Letunic, Doerks, and Bork 2012), SUPERFAMILY-1.75 (Wilson et al. 2009), TIGRFAM-15.0 (Haft, Selengut, and White 2003) and THMM-2.0c (Krogh et al. 2001).

In addition, I analyzed the microRNA data with my microPIECE pipeline, created the microRNA-target groups and verified them with RNAhybrid and RNA22. I created the figures 18 and 19, as well as supplemental figures S7 to S12 and participated in writing the manuscript.

The datasets are available at NCBI GEO repository:
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE123965

SCIENTIFIC REPORTS
nature research

OPEN

MicroRNAs regulate innate immunity against uropathogenic and commensal-like *Escherichia coli* infections in the surrogate insect model *Galleria mellonella*

Krishnendu Mukherjee[1,3,4], Daniel Amsel[1,4], Miriam Kalsy[1], Andre Billion[1], Ulrich Dobrindt[3] & Andreas Vilcinskas[2*]

Uropathogenic *Escherichia coli* (UPEC) strains cause symptomatic urinary tract infections in humans whereas commensal-like *E. coli* strains in the urinary bladder cause long-term asymptomatic bacteriuria (ABU). We previously reported that UPEC and ABU strains differentially regulate key DNA methylation and histone acetylation components in the surrogate insect host *Galleria mellonella* to epigenetically modulate innate immunity-related gene expression, which in turn controls bacterial growth. In this follow-up study, we infected *G. mellonella* larvae with UPEC strain CFT073 or ABU strain 83972 to identify differences in the expression of microRNAs (miRNAs), a class of non-coding RNAs that regulate gene expression at the post-transcriptional level. Our small RNA sequencing analysis showed that UPEC and ABU infections caused significant changes in the abundance of miRNAs in the larvae, and highlighted the differential expression of 147 conserved miRNAs and 95 novel miRNA candidates. We annotated the *G. mellonella* genome sequence to investigate the miRNA-regulated expression of genes encoding antimicrobial peptides, signaling proteins, and enzymatic regulators of DNA methylation and histone acetylation in infected larvae. Our results indicate that miRNAs play a role in the epigenetic reprograming of innate immunity in *G. mellonella* larvae to distinguish between pathogenic and commensal strains of *E. coli*.

8.1 Introduction

Urinary tract infections (UTIs) are a global public health problem, with 50% of all women experiencing a symptomatic UTI episode at least once in their lifetime. This results in 11 million medical visits and 100,000 hospital admissions in the United States every year (Russo and Johnson 2003; Foxman 2014). Uropathogenic *Escherichia coli* (UPEC) strains cause 70–90% of all UTIs in humans, and antibiotics are the front-line treatment option despite growing resistance among the target strains. UPEC strains infect the urinary bladder through the urethra (cystitis), and if they remain untreated, the infection can spread to the kidneys (pyelonephritis) leading to renal failure and sepsis. Unlike UPEC strains, commensal-like *E. coli* strains can colonize the urinary bladder in large numbers without symptoms. Such asymptomatic bacteriuria (ABU) strains have evolved from UPEC strains by losing the ability to express functional virulence factors (J. Zdziarski et al. 2008; Jaroslaw Zdziarski et al. 2010; Leimbach, Hacker, and Dobrindt 2013; Dobrindt, Wullt, and Svanborg 2016). The ABU *E. coli* strain 83972 achieves long-term growth in the urinary bladder by adopting a commensal-like lifestyle. It blocks disease-associated signaling pathways and prevents symptomatic UTIs caused by more virulent UPEC strains (Lutay et al. 2013; Wullt and Svanborg 2016; Ambite et al. 2016).

Innate immunity-related gene expression distinguishes between infections caused by ABU and UPEC strains in the urinary bladder. Bacterial molecular recognition patterns frequently expressed by bacterial pathogens activate different signaling pathways involved in innate immune response. Toll-like receptor (TLR) 4-mediated signaling distinguishes pathogenic from commensal strains and controls the downstream signaling pathways thus maintaining pathogen specificity (Hagberg et al. 1984; Ragnarsdóttir et al. 2007; Godaly, Ambite, and Svanborg 2015). Additionally, the secreted TIR domain homologue TcpC is expressed by many UPEC strains and inhibits MyD88 and inflammasome activation (Cirl et al. 2008; Waldhuber et al. 2016). ABU strains have also been shown to modulate host gene expression by suppressing RNA polymerase II (Lutay et al. 2013; Ambite et al. 2016). Surprisingly, the discriminatory host response is not restricted to humans and also occurs in the greater wax moth *Galleria mellonella*, which has been established as a surrogate insect model host to study human pathogens, including UPEC (Vilcinskas 2011; Mukherjee, Fischer, and Vilcinskas 2012; Mukherjee et al. 2013; Alghoribi et al. 2014; Williamson et al. 2014;

Ciesielczuk et al. 2015; Mukherjee, Twyman, and Vilcinskas 2015; Vilcinskas 2016; Heitmueller et al. 2017). The infection of *G. mellonella* larvae with UPEC strain CFT073 or ABU strain 83972 at 37°C resulted in the differential expression of genes encoding TLRs, cytokine-like proteins and antimicrobial peptides (AMPs) (Heitmueller et al. 2017). In eukaryotes, gene expression is regulated by epigenetic mechanisms resulting in heritable phenotypes without mutation. We previously found that DNA methylation and histone acetylation were differentially regulated in larvae infected with UPEC and ABU strains, underpinning the reprogramming of innate immunity at the level of transcriptional initiation (Heitmueller et al. 2017).

In this follow-up study, we investigated the role of microRNAs (miRNAs), which have the potential to regulate innate immunity at the post-transcriptional level. These non-coding RNAs are 18–24 nucleotides long and are conserved in most eukaryotes. They bind to the 3′ and 5′ untranslated regions (UTRs) of target messenger RNAs (mRNAs), causing translational repression and mRNA decay (Asgari 2013). They play important role in various infectious diseases, and facilitate the immune response to bacterial infection in insects (Mukherjee and Vilcinskas 2014; Das, Garnica, and Dhandayuthapani 2016; Mannala et al. 2017). We have constructed microarrays to analyze the expression of conserved miRNAs in *G. mellonella* larvae during infection with the entomopathogenic bacterium *Bacillus thuringiensis* and the entomopathogenic fungus *Metarhizium robertsii* (Mukherjee et al. 2017; 2019). Here, we carried out miRNA sequencing in

Table 10 **Length distribution of mappable reads (≥17 nt to ≤30 nt) obtained from UPEC and ABU infected *G. mellonella* deep sequencing**

Length (nt)	Number of Reads		
	CFT073	83972	Control
17	1080	304	424
18	3973	1075	1079
19	20574	5240	5460
20	61939	12931	18266
21	148923	33054	39247
22	522305	90841	111739
23	258561	43589	39245
24	186038	32528	43156
25	3906	414	539
30	0	2	0

G. mellonella larvae infected with UPEC strain CFT073 or ABU strain 83972 to investigate strain-dependent expression of novel and conserved miRNAs, to identify the mRNA targets of these miRNAs, and to analyze co-expression of miRNAs and their mRNA targets in infected larvae.

8.2 Results

8.2.1 Small RNA deep sequencing of *G. mellonella* larvae infected with UPEC/ABU strains

The miRNAs expressed in *G. mellonella* in response to UPEC and ABU infections were identified by high-throughput sequencing of whole-larvae samples 24 h after infection. The number of raw sequence reads was 62,511,810 for larvae infected with the UPEC strain (hereafter described as *UPEC larvae*), 53,675,182 for larvae infected with the ABU strain (hereafter described as *ABU larvae*) and 75,401,198 for the mock-injected controls. The size distribution of the trimmed, high-quality reads ranged from 17 to 30 nucleotides (Table 10) with a peak at 22 nucleotides. We identified 141 unique precursor hairpins and 257 unique mature miRNA sequences, with the greatest number detected in the UPEC larvae, followed by the control larvae and finally the ABU larvae (Table 10). Among the 257 mature miRNAs, 95 appeared to be novel, 148 were conserved and 14 included up to three mismatching nucleotides (Table S3).

8.2.2 Expression analysis of miRNAs in *G. mellonella* larvae infected with UPEC/ABU strains

We were able to classify 213 of the 257 mature miRNAs based on their comparative expression profiles in the UPEC/ABU larvae and uninfected controls. We found that 147 of the miRNAs were expressed consistently in all three groups, but 26 miRNAs were modulated in a single group and 40 miRNAs showed differential expression in pairwise comparisons (Figure 20, Table S4). Specifically, we identified 5, 18 and 3 miRNAs that were specifically modulated in ABU larvae, UPEC larvae or controls, respectively, and we identified 3, 19 and 18 miRNAs that showed differential expression between ABU larvae and controls, between UPEC larvae and

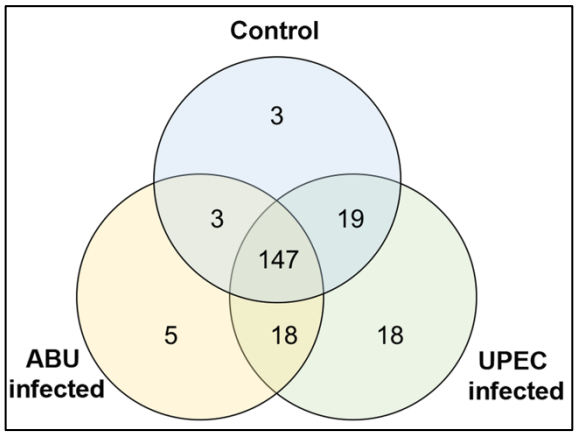


Figure 20 **Venn diagram showing the differential expression of miRNAs in ABU and UPEC infected, and mock injected whole animal *G. mellonella* larvae.** The miRNA sequences were obtained from small RNA sequencing of ABU and UPEC larvae, and mock injected control larva

controls, and between UPEC and ABU larvae, respectively.

We selected several miRNAs for classification based on their expression profile in UPEC and ABU larvae, describing them as conserved if they could be named according to current miRBase conventions and introducing the designation new in the name if they were novel. We found that the novel miRNAs gme-new-63-5p, gme-new-120-3p and gme-new-119-3p were downregulated and conserved miRNAs gme-miR-316-3p, gme-miR-2c-5p, gme-miR-1-5p were upregulated in UPEC and ABU larvae compared to controls, indicating a class of miRNAs expressed generally upon bacterial infection (Figure 21A).

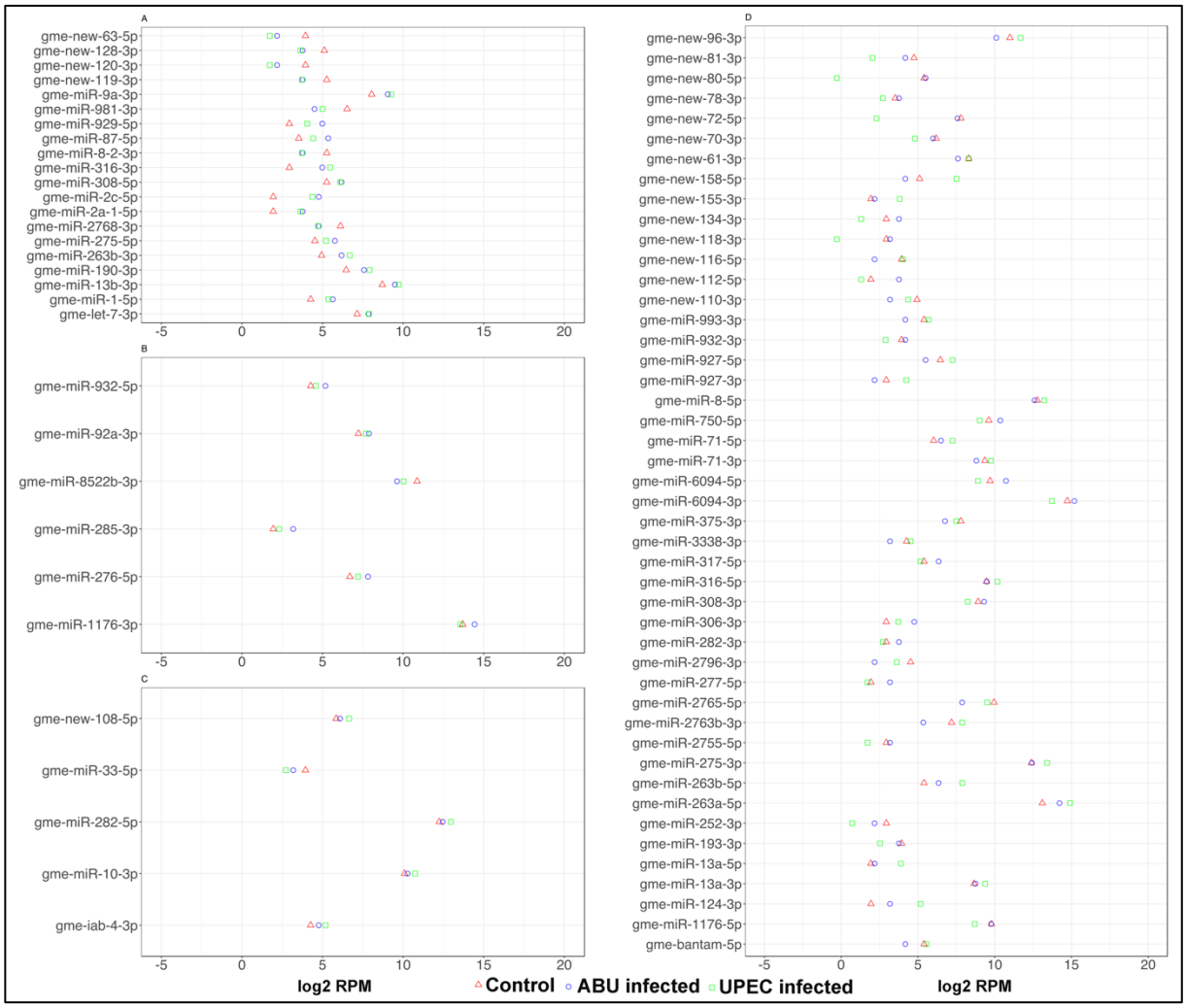


Figure 21 **Distribution of expressed miRNAs in ABU and UPEC infected, and mock injected whole animal *G. mellonella* larvae**. The miRNA sequences were obtained from small RNA sequencing of ABU and UPEC larvae, and mock injected control larvae. The figure (A-D) represent significantly expressed miRNAs in ABU and UPEC infected and mock injected larvae. The log expression levels were calculated in `reads per million` (`RPM`).

In contrast, the conserved miRNAs gme-miR-285-3p, gme-miR-1176-3p and gme-miR-276-5p were upregulated in ABU larvae compared to UPEC larvae and controls (Figure 21B) whereas novel miRNAs such as gme-new-108-5p and

conserved miRNAs such as gme-miR-10-3p and gme-miR-282-5p were upregulated in UPEC larvae compared to ABU larvae and controls (Figure 21C). Similarly, the novel miRNAs gme-new-81-3p, gme-new-80-5p and gme-new-72-5p were downregulated in UPEC larvae compared to ABU larvae and controls (Figure 21D) whereas gme-new-116-5p, gme-new-110-3p, gme-miR-993-3p and gme-miR-2765-5p were downregulated in ABU larvae compared to UPEC larvae and controls. These groupings indicated miRNAs that were specifically induced or repressed by one or other bacterial strain.

8.2.3 Identification and expression analysis of miRNA targets in *G. mellonella* larvae infected with UPEC/ABU strains

The targets of 257 mature miRNAs were predicted using our `microPIECE` (microRNA pipeline enhanced by CLIP experiments) pipeline (Amsel et al. 2018). This is based on argonaute-crosslinking and immunoprecipitation (AGO-CLIP) reference datasets from other insect species, allowing us to transfer the identified binding regions to orthologous transcripts in UPEC and ABU larvae. The target predictions were then mapped to conserved regions using the recently sequenced *G. mellonella* genome and transcriptome (Lange et al. 2018) and annotated as described in the methods section. We accepted a miRNA-binding region if one of the six CLIP-seq files was positive. We identified 6979 miRNA–target interactions comprising 1898 unique target mRNAs for 257 unique miRNAs. For comparison, target prediction without CLIP-seq containment of the search space resulted in 735,748 potential miRNA–mRNA interactions. We correlated the expression of miRNAs and target mRNAs in the UPEC and ABU larvae, focusing on miRNAs targeting genes regulating innate immunity and epigenetic mechanisms in *G. mellonella*. The selected miRNAs were found to bind mRNAs encoding proteins such as AMP-binding enzyme, phosphatidylinositol 3-kinase, AMP-dependent synthetase/ligase, lysozyme, histone deacetylases (HDACs), histone acetyltransferases (HATs), and methyltransferases (Table 11, Table S5). For example, gme-new-70-3p (target gene 1), gme-new-40-3p (target gene 2), gme-new-135-3p (target gene 13) and gme-new-160-5p (target gene 19) were downregulated in UPEC larvae compared to ABU larvae and controls, and their corresponding target mRNAs encoding AMP-binding enzyme, phosphatidylinositol 3-kinase, AGC-kinase C-terminal domain, and lipopolysaccharide-induced tumor necrosis alpha factor (LITAF) were upregulated

Table 11 **Annotation of miRNA targets**

miRNA	Target mRNA	mRNA Annotation
gme-new-70-3p	Gene 1	AMP-binding enzyme
gme-new-40-3p	Gene 2	Phosphatidylinositol 3-kinase, C2 domain
gme-new-138-3p	Gene 3	AMP-dependent synthetase/ligase
gme-new-4-5p	Gene 4	AMP-dependent synthetase/ligase
gme-new-135-5p	Gene 5	Invertebrate-type lysozyme
gme-new-121-3p	Gene 6	Acetyltransferase (GNAT) family
gme-new-70-3p	Gene 7	Aldolase-type TIM barrel
gme-new-147-3p	Gene 8	Histone deacetylase superfamily
gme-new-160-5p	Gene 9	Ubiquitin-activating enzyme
gme-new-106-5p	Gene 10	S-adenosyl-L-methionine-dependent methyltransferase
gme-new-147-3p	Gene 11	Histone deacetylase superfamily
gme-new-135-3p	Gene 12	AMP-dependent synthetase
gme-new-135-5p	Gene 13	AGC-kinase C-terminal domain
gme-new-161-3p	Gene 14	AMP-dependent synthetase/ligase
gme-new-122-3p	Gene 15	Ubiquitin carboxyl-terminal hydrolase superfamily
gme-new-82-5p	Gene 16	HECT, E3 ligase catalytic domain
gme-new-117-5p	Gene 17	Ubiquitin-like domain superfamily
gme-new-160-5p	Gene 18	LPS-induced tumor necrosis factor alpha factor
gme-new-160-5p	Gene 19	LITAF domain containing protein
gme-new-136-3p	Gene 20	Histone-lysine N-methyltransferase
gme-new-106-5p	Gene 21	Acetyltransferase (GNAT) domain

(Figure 22). Conserved and novel miRNAs targeting mRNAs encoding immunity-related proteins such as TNF-8-like or zf-LITAF-like were either upregulated (gme-miR-274-3p and gme-miR-8-5p) or downregulated (gme-new-135-3p, gme-new-161-3p and gme-new-160-5p) in UPEC larvae compared to ABU larvae, or uniformly expressed (gme-miR-124-5p, gme-miR-2a-3p and gme-miR-13b-3p) (Figure S7). Furthermore, miRNAs targeting mRNAs encoding invertebrate-type lysozyme were either upregulated (gme-miR-263a-5p) or downregulated (gme-new-135-5p and gme-miR-263b-5p) in UPEC larvae compared to ABU larvae, or uniformly expressed (gme-miR-2a-2-5p) (Figure S8).Novel miRNAs targeting genes expressing liner gramicidin synthase subunit D, long-chain-fatty acid- ligase--CoA ligase, Ras guanine-nucleotide exchange factor were either upregulated (gme-new-138-3p, gme-new-54-3p, gme-new-4-5p) or downregulated (gme-new-40-3p, gme-new-30-3p, gme-new-70-3p) in UPEC larvae compared to ABU larvae, or uniformly expressed in both conditions (gme-new-72-3p) (Figure S9). Similarly, gme-new-106-5p (target gene 10), gme-new-147-3p (target gene 11), gme-new-122-3p (target gene 15), gme-new-82-5p (target gene 16), gme-new-117-5p (target gene 17) and gme-new-136-3p (target gene 20) were downregulated in ABU larvae compared to UPEC larvae and controls, and their corresponding target mRNAs encoding methyltransferases, HDACs and hydrolases were upregulated (Figure 22). The miRNAs targeting mRNAs encoding HATs were either upregulated (gme-miR-184-5p and gme-miR-13a-3p) or downregulated (gme-new-70-3p and gme-new-135-3p) in UPEC larvae compared to ABU larvae, or uniformly expressed (gme-miR-970-3p and gme-new-108-3p) (Figure S10). The miRNAs targeting mRNAs

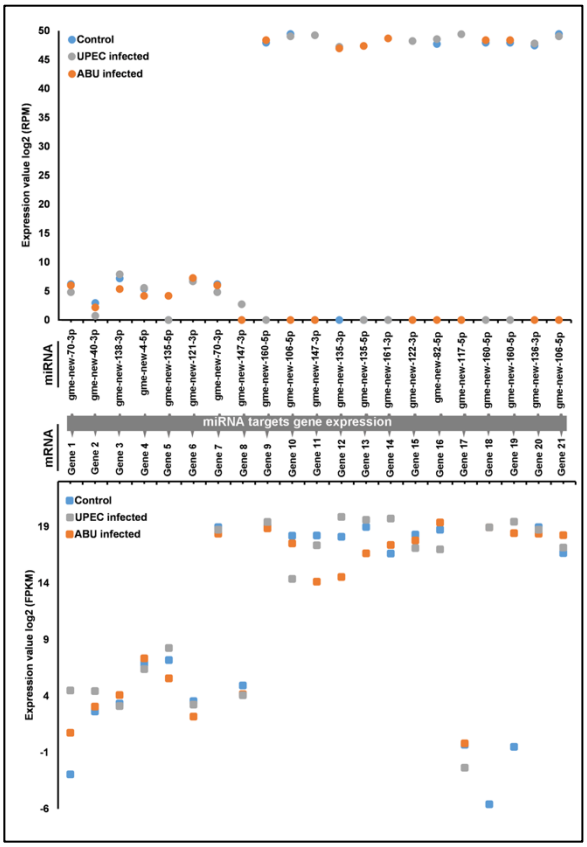


Figure 22 **Differential expression of miRNAs and predicted target mRNAs in ABU and UPEC infected, and mock injected whole animal *G. mellonella* larvae.** The novel miRNA sequences were obtained from small-RNA sequencing and their predicted mRNA targets were validated by RT-PCR to confirm differential expression in ABU and UPEC infected *G. mellonella* larvae: (A) gme-new-160-5p, gme-new-106-5p, gme-new-147-3p; (B) gene 9, gene 10, gene 11. The relative fold differences indicated for the miRNAs and mRNAs are normalized against gme-miR-133 and 18S rRNA as the internal reference control. (* $p < 0.05$– fold expression in ABU larvae were compared with fold expression in UPEC larvae).

encoding HDACs were either upregulated (gme-new-147-3p, gme-miR-71-3p and gme-miR-316-5p) or downregulated (gme-new-134-3p, gme-new-81-5p and gme-new-112-5p) in UPEC larvae compared to ABU larvae, or uniformly expressed (gme-miR-2766-3p, gme-new-80-3p and gme-miR-33-3p) (Figure S11). Novel miRNAs targeting mRNAs encoding methyltransferases were either upregulated (gme-new-136-3p, gme-new-136-3p and gme-new-61-5p), or downregulated (gme-new-160-5p, gme-new-135-3p and gme-new-123-3p) in UPEC larvae compared to ABU larvae, or uniformly expressed (gme-new-139-5p, gme-new-128-5p and gme-new108-3p) (Figure S12). The numbers of miRNA targets identified by `microPIECE` in figure 22 were validated by `RNAhybrid` (3′ UTR_Fly) and `RNA22` (sensitivity 63%, specificity 61%, minimal number of paired-up bases in heteroduplexes 10, max fold energy -5). With `RNAhybrid`, we confirmed all 21 predicted targets, whereas with `RNA22 v2` 16 of 21 microRNA-mRNA pairs were validated (Table 12).

Table 12 **Validation of miRNA target prediction by microPIECE from table 11.**

miRNA – target mRNA	miRanda	RNAhybrid	RNA22
gme-new-70-3p – Gene 1			
gme-new-40-3p – Gene 2			
gme-new-138-3p – Gene 3			
gme-new-4-5p – Gene 4			
gme-new-135-5p – Gene 5			
gme-new-121-3p – Gene 6			
gme-new-70-3p – Gene 7			
gme-new-147-3p – Gene 8			
gme-new-160-5p – Gene 9			
gme-new-106-5p – Gene 10			
gme-new-147-3p – Gene 11			
gme-new-135-3p – Gene 12			
gme-new-135-5p – Gene 13			
gme-new-161-3p – Gene 14			
gme-new-122-3p – Gene 15			
gme-new-82-5p – Gene 16			
gme-new-117-5p – Gene 17			
gme-new-160-5p – Gene 18			
gme-new-160-5p – Gene 19			
gme-new-136-3p – Gene 20			
gme-new-106-5p – Gene 21			
Successful prediction	**No prediction**		

The expression of miRNA candidates and their mRNA targets was further experimentally verified by RT-PCR. We selected gme-new-160-5p, gme-new-106-5p, gme-new-147-3p, gme-miR-929-5p, gme-miR-932-5p and gme-miR-let-5p because of their relatively high expression levels in UPEC/ABU larvae and confirmed their upregulation (gme-new-160-5p, gme-new-106-5p, gme-new-147-3p) or downregulation (gme-miR-929-5p) in UPEC larvae relative to ABU larvae 24 h after infection, as predicted by miRNA sequencing (Figure 23A, Figure S13). The expression pattern of gme-miR-932-5p by RT-PCR in UPEC and ABU larvae 24 h after infection was however different from sequencing. No differences in the expression levels of these miRNAs were observed 72 h after infection. The upregulation of gme-new-147-3p in UPEC larvae compared to ABU larvae resulted in the downregulation of its predicted target mRNA, for gene 11 (Figure 23B).

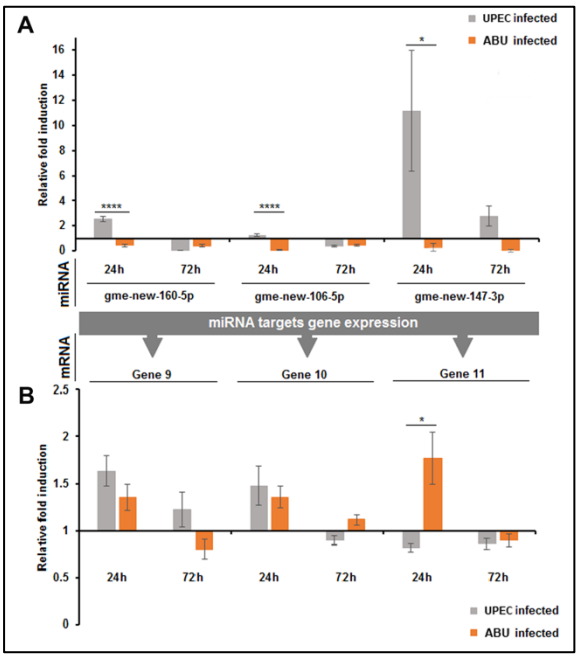


Figure 23 **Differential expression of miRNAs and predicted target mRNAs in G. mellonella larvae infected with ABU and UPEC strains, and in mock-injected controls.** The log expression levels of novel miRNAs identified by small RNA sequencing and their predicted mRNA targets were calculated in reads per million (RPM).

8.3 Discussion

Host susceptibility and innate immunity-related gene expression help determine the outcome of infections caused by UPEC and ABU strains in the urinary tract. In this study, we used the surrogate insect host *G. mellonella* to show that the differential innate immune response to UPEC and ABU infections is regulated at least in part by non-coding miRNAs. Following miRNA sequencing in ABU and UPEC larvae, we analyzed the strain-dependent expression of novel and conserved miRNAs, and using our `microPIECE` pipeline based on AGO-CLIP reference datasets from other insect species we predicted miRNA targets (Amsel et al. 2018). We found that ABU and UPEC infections can trigger the expression of novel and conserved miRNAs to modulate the expression of innate immunity-related genes (Figure 22, Figure S7–S9). ABU infection induces miRNAs that suppress genes related to cell signaling and innate immunity. For example, the expression of gme-new-160-5p inhibits LITAF like immunity related proteins in larvae and reduces the lipopolysaccharide-induced innate immune activation, resulting in improved host survival (Srinivasan, Leeman, and Amar 2010; Niu et al. 2011). The post-transcriptional suppression of LITAF may thus favor the long-term survival of ABU in larvae by attenuating the immune response, whereas the expression of LITAF in UPEC larvae encourages strong immune response (Heitmueller et al. 2017). Inhibition of LITAF provides resistance to systemic *E. coli* LPS-induced lethality in mammals. Thus, LITAF is a promising therapeutic target for the treatment of TNF-mediated inflammatory diseases, and we identified a miRNA in *G. mellonella* that inhibits its expression (Niu et al. 2011). ABU and UPEC infections also modulate miRNAs that affect the production of AMPs. For example, UPEC infections induce the expression of gme-138-3p, gme-new-4-5p and gme-new-135-3p while suppressing gme-new-70-3p and gme-new-161-3p, which has the net effect of increasing the synthesis of the most potent AMPs (lysozymes, cecropins, gloverin, galiomycin and moricin), whereas weaker AMPs (anionic peptides, apolipophoricin) are induced following ABU infection (Heitmueller et al. 2017). ABU *E. coli* strain 83972 differs from UPEC strain CFT073 in terms of virulence gene expression (Dobrindt, Wullt, and Svanborg 2016). Colonization and long-term survival of ABU *E. coli* strain 83972 in *G. mellonella* larvae and most likely also in the human urinary tract is achieved by miRNA-mediated suppression of gene expression constituting strong antimicrobial response. On the other hand, UPEC infection provokes expression of these AMPs by downregulating miRNAs that

specifically inhibit their expression. The expression of miRNAs targeting AMP synthesis has also been shown following infection in the insect host *Plutella xylostella*, and in some human diseases (P. T. Liu et al. 2012; Etebari and Asgari 2013). We used the *G. mellonella* system to discover new miRNAs that target innate immune related proteins important for UPEC infection implicating scope for therapeutic application to treat UTI in humans.

Several human pathogens can manipulate host cell antimicrobial responses and evade the immune system by influencing epigenetic mechanisms such as histone acetylation and DNA/RNA methylation (Bierne, Hamon, and Cossart 2012). Qualitative and quantitative differences in the responses to UPEC and ABU strains in *G. mellonella* are also epigenetically regulated in this manner (Heitmueller et al. 2017). While many target genes of miRNAs are known, even less information exists as to how miRNAs cooperate with histone acetylation and DNA methylation in the context of host-pathogen interaction. There is limited evidence suggesting that negative correlation between the expression of miRNAs and HDACs or HATs has been associated with infection by human pathogens (Xing et al. 2019). Interestingly, our analysis of miRNA targets revealed that both novel and conserved miRNAs can target genes encoding methyltransferases, HATs and HDACs, which are key regulators of DNA methylation and histone modifications. In UPEC larvae, the induction of miRNAs gme-new-106-5p, gme-new-184-5p and gme-new-13a-3p correlated with the downregulation of target genes encoding HATs, which acetylate histones to form open chromatin that favors the expression of immunity-relate genes. In mammals, HATs are degraded by a zinc-dependent metalloproteinase from enteropathogenic and enterohemorrhagic *E. coli* in order to dampen inflammatory responses (Shames et al. 2011; Grabiec and Potempa 2018). We suggest that UPEC also follow an alternative strategy of miRNA-mediated suppression of HAT expression in an infected host. The activity of HATs is also opposed by HDACs, which deacetylate histones and form condensed chromatin that suppresses gene expression. The HDAC sap18 subunit was downregulated in ABU larvae but upregulated in UPEC larvae and here we identified a miRNA (gme-new-134-3p) that targets the mRNA encoding this protein subunit (Mukherjee, Twyman, and Vilcinskas 2015). The induction of this novel miRNA in ABU larvae suggests a post-transcriptional mechanism for the suppression of HDAC sap18. The novel miRNA gme-new-106-5p was upregulated in UPEC larvae, and this targets the mRNA for S-adenosyl-L-methionine-dependent

methyltransferase, which is a methyl donor for DNA methyltransferases. We identified miRNAs that target methyltransferase mRNAs in UPEC and ABU larvae, but surprisingly none of them were the DNA methyltransferases responsible for epigenetic regulation. However, insect genomes are sparsely methylated compared to mammals, and the role of DNA methylation in innate immunity is not well understood. The lack of miRNAs targeting mRNAs encoding maintenance or *de novo* methyltransferases in UPEC and ABU larvae may indicate the limited significance of DNA methyltransferases in the regulation of innate immunity in *G. mellonella* (Glastad et al. 2011).

In addition to the strain-specific regulation of miRNAs, we also identified a large number of miRNA candidates that were commonly modulated in UPEC and ABU larvae. The majority of these miRNAs (such as mir-8 family) are conserved among other eukaryotes, for example let-7 and mir-124 have homologous targets in *G. mellonella* and humans. Gme-let-7-5p targets the HAT KAT2A, and both gme-miR-283-5p and gme-miR-33-3p target HDAC8, indicating the existence of cross-talk between miRNAs and other epigenetic mechanisms that regulate chromatin structure and gene expression (Bianchi et al. 2017).

8.4 Conclusion

Here we show that UPEC strain CFT073 and ABU strain 83972 trigger the modulation of host miRNAs in *G. mellonella* to epigenetically regulate the innate immune response. Many novel miRNAs showed strain-dependent expression in the infected larvae, whereas others were modulated similarly regardless of the infection status or in a similar manner during both infections. We argue that miRNAs determine the different pathogenic potential of the ABU and UPEC strains in *G. mellonella*, and may regulate different behavior of these strains in the human urinary tract. Taken together, our results emphasize the importance of *G. mellonella* miRNAs in the regulation of host innate immunity to distinguish between pathogenic and commensal-like *E. coli* strains.

8.5 Methods

8.5.1 Bacterial strains, insects, and culture media

Cultures of UPEC strain CFT073 and ABU strain 83972 were maintained aerobically in lysogeny broth (LB) at 37°C and on LB agar plates (Carl Roth, Karlsruhe, Germany). For long-term storage, bacteria were frozen at −80°C in LB supplemented with 30% (v/v) glycerol. *G. mellonella* larvae were

obtained from Fauna Topics Zoobedarf Zucht und Handels GmbH, Marbach am Neckar, Germany. The larvae were reared on an artificial diet (22% maize meal, 22% wheat germ, 11% dry yeast, 17.5% beeswax, 11% honey and 11% glycerin) at 32°C in darkness. Larvae at their sixth instar stage, each weighing 250–350 mg, were used in all experiments.

8.5.2 *G. mellonella* injection

Injection experiments were carried out using logarithmic growth-phase bacterial cultures in 10 ml LB. The bacteria were washed and serially diluted in 0.9% NaCl, and 10-µl aliquots (10^5 colony forming units/ml) were injected into larvae through the left proleg using 1-ml disposable syringes and 20-mm 0.4 gauge needles mounted on a microapplicator, as previously described (Mukherjee et al. 2010). Mock injections with an empty needle were carried out as controls. Larvae were considered dead after incubation at 37°C when they showed no movement in response to touch.

8.5.3 Small RNA isolation, library construction, sequencing and analysis

Total RNA was extracted using the miRNeasy Mini Kit (Qiagen, Hilden, Germany). We used the following criteria to design experiments and select miRNA sequencing technique, to minimize sequencing error rates and limitations for not performing cost-intensive multiple miRNA sequencing experiments. First, considering the importance of biological variations in the identification of conserved and highly expressed miRNA, RNA samples were pooled from three independent biological experiments comprising five larvae per experiment, and miRNA sequencing was performed by LC Sciences (Houston, Texas, USA). Briefly, total RNA was analyzed using a Nandrop spectrophotometer (Thermo Fischer Scientific, Waltham, Massachusetts, USA) and a 2100 Bioanalyzer (Agilent Technologies, Santa Clara, California, USA) and samples with an RNA integrity number (RIN) > 7 as well as 260/280 and 260/230 absorbance ratios > 2.0 were used for cDNA library preparation based on the TruSeq Small RNA Sample Preparation Protocol (Illumina, San Diego, California, USA). Next, the cDNA libraries were purified and sequenced on an Illumina HiSeq 2500 in the 50-bp single-end configuration, which has the least error rate (below 0.1%) compared to other sequencing techniques. *G. mellonella* miRNAs were identified using `miRDeep2` (Friedländer et al. 2012) `v2.0.0.8` to screen the recently published *G. mellonella* genome sequence

against miRbase (Kozomara and Griffiths-Jones 2014) release 22. We designated miRNAs from all animals including related insect species *Bombyx mori*, *Manduca sexta*, *Plutella xylostella*, *Spodoptera frugiperda* (order Lepidoptera), *Aedes aegypti*, *Anopheles gambiae*, *Tribolium castaneum*, *Drosophila melanogaster* (order Diptera), *Apis mellifera*, *Nasonia vitripennis* (order Hymenoptera), and *Acyrthosiphon pisum* (order Homoptera) for the purpose of miRNAs identification. The miRNA hits were temporarily numbered in ascending order and identical mature sequences were merged to reduce redundancy. The miRNAs with up to three nucleotide mismatches were renamed according to the best match from miRBase. Next, to reduce false positives, we discarded all reads that mapped to rRNAs from further analysis. Briefly, reads with a minimal length of 17 nucleotides were trimmed using Cutadapt v1.8.3 and rRNA sequences were filtered out by screening against the Silva rRNA database (Quast et al. 2013). Mature miRNAs were mapped using bwa v0.7.10-r789 and we calculated the expression level in reads per million (RPM) (H. Li and Durbin 2009). We used Deseq v1.32.0 to compute an indicative fold change between samples (Anders and Huber 2010) and the stem loop structures of 37 novel miRNAs were determined (Table S6). For the expression calculation, we allowed one measured miRNA read, because adapter ligations in library preparation steps, as well as unwanted intra- and inter-RNA bindings of miRNAs can disfavor certain mature miRNAs and may lead to underrepresented sequencing results. We considered results with a fold change of ≤ 0.5 and ≥ 1.5 as differentially expressed between two groups of larvae.

8.5.4 Annotation of *G. mellonella* transcriptome and miRNA target prediction

For miRNA target prediction, the publicly available paired-end RNA sequencing libraries from UPEC and ABU infected *G. mellonella* larvae (SRX2727976 and SRX2727977) were compared with non-infected larvae (SRX371340). We then aligned the reads from the transcriptome sequencing against the recently sequenced *G. mellonella* genome (ASM258982v1) (Lange et al. 2018) using HiSat2 (D. Kim, Langmead, and Salzberg 2015) v2.1.0 and transformed and sorted the results using samtools (H. Li, Handsaker, Wysoker, Fennell, Ruan, Homer, Marth, Abecasis, Durbin, et al. 2009) v1.6. Assembly was carried out using cufflinks v2.2.1 with standard parameters (Trapnell et al. 2010) and TransDecoder v5.0.2 analysis for the annotation of different isoforms

(Haas et al. 2013). The analysis was enhanced with a BlastP v2.6.0+ (Camacho et al. 2009) search against SwissProt (UniProt Consortium 2018) and a hmmer (Johnson, Eddy, and Portugaly 2010) search against Pfam-A with standard parameters (Finn et al. 2016). The assembly was evaluated using rnaQUAST (Bushmanova et al. 2016). We then annotated the transcripts using InterproScan v5.27-66.0 (Jones et al. 2014).

Targets were predicted using the microPIECE pipeline (Amsel et al. 2018), which uses AGO-CLIP libraries from other species and performs a trimming, mapping and peak calling on the datasets, like a typical CLIP analysis workflow. Briefly, the microPIECE takes the AGO-CLIP data from a species A and transfers it to a species B. That means it transfer the verified miRNA binding region of a mRNA to the homologous transcript of another species. Given a set of miRNAs from species B the pipeline then predicts their targets on the transferred CLIP regions with a much lower false-positive rate, since it relies on evolutionary conserved and experimentally verified binding regions. Identification of orthologous proteins and mapping of the signaling regions of miRNA binding sites to the orthologous transcripts in the species of interest were computed and used for target predictions with miRanda v3.3a (Betel et al. 2007b). The transfer of known miRNA binding sites was achieved using a set of six publicly available *Aedes aegypti* AGO-CLIP libraries, which is the only AGO-CLIP dataset available in the clade of insects (X. Zhang et al. 2017). From those six libraries, there are three replicates for two conditions. By using *A.aegypti* as AGO-CLIP data donor for *G. mellonella* the resulting miRNA target sites would be rather small compared to lepidopteran insects, but nevertheless the technique ensures that those target sites are already present between *A. aegypti* and *G. mellonella*. In addition to the miRanda tool, we used the prediction algorithms RNAhybrid (Rehmsmeier et al. 2004) and RNA22 v2 (Miranda et al. 2006) to validate the miRNA target prediction by microPIECE. We followed union over intersection when combining several tools for miRNA-mRNA predictions (A. C. Oliveira et al. 2017).

8.5.5 RT-PCR analysis

Relative miRNA and mRNA expression levels were determined by RT-PCR as previously described (Mukherjee and Vilcinskas 2014). For the analysis of miRNAs, cDNA was synthesized using the miScript II miRNA first-strand synthesis and qPCR kit (Qiagen, Hilden, Germany).

Small RNA-enriched total RNA was reverse transcribed using miScript HiSpec buffer, modified oligo-dT primers with 3′ degenerate anchors and 5′ universal tag sequence for the specific synthesis of mature miRNAs. The combination of polyadenylation and the universal tag ensures that miScript primer assays do not detect genomic DNA. Primers for the selected miRNAs were designed using the miScript miRNA product design webpage (Qiagen, Hilden, Germany). Candidate miRNA expression levels were normalized against gme-miR-133, which showed uniform expression across all samples. Real-time RT-PCR was carried out using the CFX 96 Mx3000P system (Bio-Rad, Hercules, California, USA), starting with a 15-min incubation at 95°C to activate the Hot Start Polymerase followed by 40 cycles at 94°C for 15 s, 55°C for 30 s and 70°C for 30 s. The following miRNA sequences were used for primer design: gme-new-160-5p, 5′- GTC ATT CAG CCT GCC AGC ATT GCT-3′; gme-new-106-5p, 5′-CCT TGT CAT TCT TCT TGC CCA GT-3′; gme-new-147-3p, 5′-ATT TGG TTC TCT CTA ATA GCA AT-3′; gme-miR-929-5p, 5′-AAA TTG ACT CTA GTA GGG AGT-3′; gme-miR-932-5p, 5′-TCA ATT CCG TAG TGC ATT GCA GT-3′; gme-miR-let-7-5p, 5′-TGA GGT AGT AGG TTG TAT AG -3′. The control miRNA sequence was gme-miR-133, 5′-AAG TTT TCC GTG ACG ATA TAA GGG GGC TCC-3′. The amplification of specific target mRNAs by RT-PCR was carried out as previously described (Mukherjee et al. 2013) using the following primer sequences: Gene 9-fwd 5′-CTA CAC TCG TCG CAG CAC AT-3′ and -rev 5′-GTG TTA CGG TGC ATT GTT GG-3′; Gene 10-fwd 5′-CAC CGC CTG GTA AAG AAC TC-3′ and -rev 5′-CCA TTT GAA TCC CAA GTG GA-3′; Gene 11-fwd 5′-GGC CGA TGT GTG GAG TTA GT-3′; and -rev 5′-TGC TGG GTG ATA TGT GCA GT-3′; and the housekeeping gene elongation factor 1-fwd 5′-ATG GTT GCA AAG CTG AAA CT-3′ and -rev 5′-TCC CGT GTT GAG TCA AAT TA-3′.

8.5.6 Data availability

The smallRNA sequencing data is available at NCBI via the accession number GSE123965. All data are accessible in the Supplementary Information.

8.5.7 Data Analysis

Data in Figure 22 and 23 were analyzed using Microsoft Excel 2013 (Microsoft Corp., USA). All experiments except small-RNA sequencing were performed a minimum of three times. Significant differences between pairs of values were compared using one-way analysis of variance and Holm-Šídák test

9

DISCUSSION AND CONCLUSION

The expanding number of identified microRNAs and the growing knowledge about their important role in gene regulation led to the development of a broad variety of programs. Researches are confronted with many possible ways of analyzing their data with a large number of parameters to tune the individual programs for the task. Many of the programs are standard in this field of study and their parameters have been tested to give the best results at that time. Some other programs are in strong competition with others for the same task and it is a matter of taste or detailed analysis aim which to favor. A few other tools tackle novel issues that were recently gaining importance, like the identification of microRNA isoforms and lack an independent performance comparison. To overcome this uncertainty, I did a best practice assessment for a microRNA high-throughput analysis pipeline and presented a novel method to shrink the search space of microRNA target predictions by using wet lab confirmed microRNA binding regions from CLIP-seq data of closely related species. The final pipeline was applied to analyze the microRNAome of *Galleria mellonella* under the scope off immune response behavior when infected with UPEC and ABU *E. coli* strains.

The best practice assessment was divided into two categories, the CLIP and the microRNA analysis and is focusing on the application on strong Linux machines or similar operating systems.

The CLIP workflow includes the trimming of artificial adapter sequences with `cutadapt`. As previously stated, some tools could be easily replaced by others, but it is a matter of taste which one to use. Nevertheless, `cutadapt` is very well established, lightweight and simple to use. It does not need further software, like `Java` and is comparably fast. The mapping algorithm `gsnap` was chosen because of its ability to tolerate SNPs everywhere in the read, making him the ideal tool for CLIP sequencing, where the SNPs can occur at every position and a sensitive

detection is necessary. Other tools, like `bwa` or `bowtie` would be faster, but would also ignore some of the SNPs, because of their mapping technique. I here clearly decided for the slower, but more sensitive algorithm in order to detect as many SNPs as possible. In case of the peak calling, I preferred `Piranha`, because it is designed to call the signals from all the different CLIP techniques. Since those datasets are rare in public databases, it would not make sense to focus on a mapper for a certain technique. In some future development stages, one could think of a user selection of peak calling tools, depending on the provided dataset type. Nevertheless, this has the tradeoff that users have to provide more information to the pipeline, making it more complex to use. The identification of homologous proteins was performed by `ProteinOrtho`. The analysis of homology can grow to a very complex issue, depending on the needed resolution of relationships. In the here presented case, a simple relationship between genes is sufficient and `ProteinOrtho` uses a commonly accepted way of determining this information. The reciprocal best alignment heuristic assumes that two genes of two species that share a common ancestor should find each other reciprocally as best hits in an alignment due to their ancestry. The transfer of a microRNA target region from one species to another is computed by `needle`. The alignment matrix was modified that a gap is very unfavorable compared to a mismatch. Since the CLIP derived region is not tightly embracing the binding region, some flanking nucleotides are transferred as well. As it is unclear if they have an evolutionary pressure to retain the nucleotide, a potential SNP should be allowed, whereas an indel is potentially not favorable, since this could shift the entire reading frame, for example.

The assessment of the microRNA analysis also started with the trimming of the artificial adapters from the sequencing reads. Here, `cutadapt` was preferred as well for the same reasons as previously stated, but also to reduce the number of dependency tools. The read mapping is done by `bwa`. In contrast to CLIP-seq and the flexible SNP tolerant mapping of `gsnap`, small RNA-seq and the seed region approach of `bwa` is favorable. As already described, independent benchmarks favor `bwa` over `bowtie` for miRNA mappings. The filtering against a noncoding RNA dataset is important, in order to reduce the number of reads from other sources than microRNAs. According to a broad number of investigations, the mining for novel microRNAs is mostly done by miRDeep2, which needs a reference genome. There are tools that can work without reference genomes, like `MiReader` (Jha and Shankar 2013), but implementing only this tool instead of a reference-based one, would lead to a loss of information that would be available. Of course, one discards a large group of species from the usage of the pipeline, but the remaining ones can be analyzed with an additional layer of information, when supported with genome information. In further versions of the pipeline one

could implement a variant with a reference free microRNA mining software. Anyhow this would also lead to a more complex user intervention. The expression of microRNAs is computed via mapping the reads and normalizing the counts to reads-per-million. Other normalizations, like fragments-per-kilobase-million would work as well, but since one read represents a transcript, a normalization over the length is not necessary. By using the maybe most prominent tool in bioinformatics, `BLAST`, a fast and easy search for homologous microRNAs can be achieved. Due to its famous status and long maintenance time, it is presumably very unlikely that errors occur, compared to any other tool that would compute a similar result. The target prediction is done via `miRanda` and compared to other tools, like `TargetScan`, it is open for all kinds of potential binding sites that are found via CLIP-seq transfer, making it the ideal tool for this task. Nevertheless, one could think of a pipeline variant with more than one target prediction tool or a possibility to choose many prediction tools at once, as it was done for the *Galleria mellonella* publication.

For the benchmarking of microRNA isoform tools, published in BMC Bioinformatics, I included all programs that were open-source, freely available and suitable for high-throughput pipelines. The considered tools were `isomiR-SEA`, `isomiRID`, and `miraligner`. The programs `IsomiRage`, `DeAnnIso`, `isomiRex` and `miR-isomiRExp` were excluded, because they were not fulfilling the pipeline-criteria. The remaining tools were tested under the scope of how well they deal with technical errors that were induced by the Illumina HiSeq and MiSeq sequencing machines and their limits in detecting theoretical biological variants. Of course, there are several more sequencing machines from other brands on the market available. Nevertheless, Illumina is widely accepted as the technique with the best per-base accuracy. The MiSeq system for example can deliver 2x25 bp reads with >90% of the bases yielding a quality higher than 30, which means that 1 of 1000 bases has a technical error. Now 2x25 bp is a rather rare read length, but even with a more common read length of 2x75 bp, the quality of 30 or higher is reached for over 85% of the bases. For the biological variance, seven types of theoretical microRNA isoforms were simulated. This means that the mature microRNA can be altered at the 5’ or 3’ end, by having a shorter or longer sequence than usual. Those variants are assumed to be templated, which means that they are matching the precursor of the microRNA. The 3’ end can also consist of a poly-A tail or in a more general way, a non-templated addition of nucleotides. Besides the addition or deletion of tailing nucleotides, the microRNA itself can contain single nucleotide mutations in the commonly accepted seed region (the first eight nucleotides) or in the remaining region.

Although all of the three tested tools struggled in mapping reads with simulated technical errors, isomiRID and miraligner performed better in terms of sensitivity and specificity than isomiR-SEA. It nevertheless showed that the internal re-allocation of multi-mapping reads of isomiR-SEA seems to reduce the number of false positives dramatically, but increasing the number of false negatives. Still the number of true positives is higher in the filtered results than in the raw results of isomiR-SEA. For the final evaluation, the number of false negatives was weighted as neutral event, because it is of course suboptimal, if possible results are missed, but it does not raise wrong results. Therefore, the number of false positives was weighted negatively and the number of true positives as positive impact within this study.

The test for biological variation was performed to identify potential weak points in the detection of isomiRs. In the case of isomiRID, the program found nearly all isomiRs, whereas for miraligner, I uncovered a weakness when detecting 3' or 5' deletions. This could potentially be due to the seed-based search technique of miraligner. Even worse was the performance with 5' deletions or seed-SNPs of isomiR-SEA that is using a seed-based clustering as basis for its analysis. That means that the alteration of the first eight nucleotides seems to annul the vast majority of this algorithm. Although this appears to be a large disadvantage in comparison to the other two tools, this technique allows isomiR-SEA to operate only with reference microRNAs, making it the ideal tool for non-model organisms. In contrast, isomiRID requires a genome file and additional files from miRBase.org. Furthermore, isomiRID seems to be designed for a smaller number of microRNAs, due to the visual output, indirectly suggesting a manual curation of isoforms or some independently developed parser. One shortcoming is that only one mutation at a time can be reported. Mixed mutations, like a SNP together with a poly-A-tail is visible in the alignment, but not stated together in the final result. A solely usage of the result column is therefore not sufficient and an automatic treatment of the analysis would need a special extraction parser. In contrast, miraligner comes with a structured output that makes it suitable for high-throughput pipelines and database connections. In terms of performance, it showed slightly weaker results than isomiRID, but does not need a reference genome file. The reference is taken from miRBase.org, which can be downloaded freely or created according to the miRBase.org archetype.

For the final application of an isomiR detection tool to a public dataset of *T. castaneum*, miraligner was used. The existing findings that in early development stages, namely the oocyte and first embryonic phase, miRNAs are more frequently polyadenylated than in the other stages, were verified by this tool. The analysis results further revealed a more diverse and

dynamic miRNA isoform environment during the early embryonic stages, which needs to be investigated further. A number of isomiRs had 5' extensions during 24-38 hours. This could potentially lead to a shift of the seed region and potentially other targets or even an inactivation. Theoretically this could make sense as the resources of an egg are limited and every molecule can be estimated to be used at maximal efficiency. This alteration would make the microRNA applicable for maybe another pathway or could influence the expression of the corresponding mRNA differently.

Furthermore, during the 20-24 hours development phase, the seed regions showed an increased point mutation rate, compared to all other stages. Those mutations occurred mainly at the positions five to seven and an adenosine was replaced by a cytosine. This is special in the way that usually, A-to-I (adenosine to inosine) editing, mediated by proteins, like ADAR which is the most prominent way of RNA editing (Nishikura 2016). The inosine would be read and interpreted by the sequencing workflow as a guanine (G). These changes were rarely measured here. Further A-to-C editing events were observed in a higher frequency at positions 10, 17, 18, 19 and 20.

An increased expression of microRNAs was observed in later timepoints, indicating their need before hatching. This could be explained by the entirely different biological needs between an egg and a larva.

These findings suggest a strong influence of miRNAs and its isoforms, the isomiRs, in the early development stages of *Tribolium castaneum*. But isomiRs are not only there of great interest. Recent findings demonstrate that wild-type microRNA and isomiR can trigger different pathways, for example in ischemic blood vessels, where the isomiR of miR-411 negatively influences the migration of vascular cells (van der Kwast et al. 2019).

Outgoing from those results and potential impact on the animal development, I decided to include the isomiR analysis into my microRNA workflow that led to the final pipeline, `microPIECE`, which was published in JOSS – Journal of Open Source Software. This publication describes the all-in-one solution for species-independent microRNA research, covering a broad range of common scientific question related to microRNAs, like mining novel miRNAs, isomiR detection, expression calculation, target prediction, homology prediction and genomic position determination. In combination with the database structure, further analysis approaches are possible, like arm-switching events and polycistronic cluster identification. This is an advantage to other existing tools and workflows. Besides this, other tools are rather rare, they are mostly only treating sub-tasks of `microPIECE`, like `QuickMIRSeq` (Zhao et al.

2017). This tool for example is designed for quantifying microRNAs and isomiRs, but cannot be used for the identification of novel microRNAs or differential expression between conditions. `CAP-miRSeq` is identifying novel microRNAs, but lacks the isomiR identification and target prediction for example (Z. Sun et al. 2014).

My pipeline accounts for common research questions related to microRNAs, like expression values and normalization between various conditions, as well as the search for novel microRNAs. The latter case is covered by the standard tool in this field, `miRDeep2`. Nevertheless, many authors still use their own methods for miRNA mining, resulting in very high numbers of potentially novel microRNAs that need to be verified in further projects (W. Wu et al. 2017).

The strength of `microPIECE` lies in the application of a novel approach for microRNA target prediction. Where usual target predictions suffer from a high number of potential targets, the `microPIECE` pipeline makes use of the scientifically accepted theorem of conserved binding sites in combination with sequenced microRNA target regions (AGO CLIP-seq). This technique overcomes currently available tools that try to shrink the search space of microRNAs by measuring the secondary structure accessibility for RISC, like `PITA` (Kertesz et al. 2007) or `MicroTar` (Thadani and Tammi 2006), because AGO CLIP-seq measures the actual binding regions and gives a direct signal without prediction.

The development of the `microPIECE` pipeline was done in an `evolutionary prototyping` manner. I therefore started to program individual scripts for each task, as well as corresponding test cases for scripts, where I altered data files to another format. This ensured a pre-defined behavior of the scripts. During the development of the pipeline, I also had various scripts included that did not make it into the final workflow. For example, I developed a miRNA arm-switching script that directly plotted the miRNA expressions in a line-graph for all miRNAs that underwent a switch in the major expressed arm from one condition to another. I decided to leave this script out, since later queries in databases would make this step far more efficient and more customizable. Nevertheless, I would like to highlight the potential benefit to look at changing expression patters of miRNA arms in different conditions. As for example in a certain type of human brain cancer, the glioblastoma, mir-324 arm-switches were identified to have an impact on cell proliferation (H. Kim et al. 2020). These arm-switches were induced by uridylyl transferases, leading to a microRNA isoform with a U at one end. It therefore seems to be valuable to check for isoform and arm-switch combinations. With the `microPIECE` pipeline in combination with the here presented database, the analysis of such combined issues is no problem. One could check the isomiRs of the early development miRNAome of *T.*

castaneum, where my investigation revealed a high expression of microRNAs with non-templated U nucleotides in all early conditions.

The final `microPIECE` pipeline is, as previously stated, designed to treat all kinds of species, although it was initially designed for *Tribolium castaneum*. Of course, the pipeline is not usable for all possible species, because one needs a minimal amount of data. This then limits the pipeline to organisms with a sequenced and annotated genome. I of course tested approaches for miRNA mining in species without a sequenced genome, but those were relying much on predictions and references to existing data. This concept made it hard for an algorithm to find novel miRNAs in small RNA sequencing data. Therefore, I decided to restrict the pipeline to species with a sequenced genome.

The step of CLIP data transfer was picked by me to create a novel approach for target prediction that relies on very conservative and credible data from wet lab experiments. To my understanding, these conserved regions highlight very functional regulations of genes in the cell that have been developed during very early speciation events in the clade of animals and should therefore be of high importance for the animal. Potential artificial alterations that are induced by humans in order to change the fitness or phylogeny in the animal should presumably focus on especially those points. This hypothesis is also highlighted by my manuscript about immune response regulation in *Galleria mellonella*.

I also did not restrict the miRNA-mRNA binding to the many times postulated seed-3'UTR combination, but accepted the whole CDS and the whole miRNA as possible binding partners. During my research I gained the impression that many things in miRNA research simply developed without real biological background. As for example the many times reported "seed" region of the miRNA. To my understanding, this seed concept was derived from the early miRNAs that were shown to bind with this region in the wet lab trials (Cloonan 2015). In bioinformatics, such patterns are easily implemented and therefore were quickly published. In my opinion, these algorithms are not wrong, but they presumably only deliver a part of the whole picture.

During my studies, machine learning and deep learning grew to a promising approach for a wide field of applications. Nevertheless, I did not want to follow this path, because those algorithms have to learn with many datasets how to predict a true miRNA-mRNA binding. This means that those algorithms highly scale with the data they are fed with. In the case of miRNAs, this would be many datasets that have been derived from already existing target predictions. This would already bias the results to the behavior of the now available tools. Another data-source would be experimentally verified miRNA-mRNA binding partners. There I have mostly

seen luciferase reporter assays. This means that one inserts the 3'UTR into a reporter gene and if the miRNA binds this sequence, a reporter-light-gene cannot be expressed. Indeed, this would lead to biological highly credible data. Still this data is not so frequent and mostly, only 3'UTR regions are inserted due to the still profound established opinion that the miRNA only binds in the 3'UTR. This would also bias the final results from artificial intelligence algorithms. Furthermore, the *in vitro* structure is not mirrored by this luciferase assay, potentially leading to signals that are not possible under real circumstances.
The shortcoming of my CLIP-based method is the lack of a possibility to verify the performance. This is due to the very rare data sets for only a handful of species which are not well comparable to each other. Another reason is that the field of microRNA stills lacks a gold standard set of targets against which one can benchmark independently. Nevertheless, as previously described, my method bases on the accepted hypothesis of evolutionary conserved binding sites and mechanisms.

In order to enhance the results that are computed by the pipeline, many research questions can be solved by comparing data. One can either compute these results each time, or construct a database with which one can query different scenarios multiple times. Therefore, I designed a database that takes the output of `microPIECE` as input and enhances its results. The database enables also further analysis steps, like polycistronic cluster determination and arm-switching events between conditions, just by querying the database. The miRNA table is a central instance in my scheme. It stores the actual sequence, together with an internal ID and a `miRBase` reference ID. The type of the mature microRNA is stored in a separate table, to be open for other types than 3p and 5p, like offset microRNAs, where a precursor contains more than one 3p and 5p sequence. I also included a species table to store more than one species in the database. This allows cross-species comparisons and analysis. The mRNA table is designed to include mRNA information, like genomic positions and reference IDs. It is directly linked to the miRNA table via the target prediction table. This table leads to a CLIP table, containing the information derived from the transfer of CLIP information. Here one can filter for values like gaps or mismatches within the transferred region. Both tables, miRNA and mRNA, share a condition table to which they are connected via an expression table. This allows the expression comparison of both sequences in various conditions. The mRNA table also has connections to an annotation table in order to provide information about GO terms or other data like this. Outgoing from the miRNA table, the miRNA precursor table leads to the genomic position table which itself is connected to the genome and chromosome table. By using this architecture,

it is also possible to compare different genome versions with each other. For example, in case the newer version comes with a better assembly, one could compare polycistronic clusters of miRNAs and how they may have changed with possible assumptions on their combined expression and targets.
According to my experience, this database covers all of the frequent scientific questions that came up during my investigations on miRNA related research, but also enables users to investigate miRNA studies on a meta level, like genome version differences, evolutionary conservation of miRNAs, miRNA-target pairs, but also on CLIP region conservation between species.
In combination with an attractive front end, this database would also be usable for a broad community of non-computer-scientists.

After the tool evaluation and pipeline development, I applied my tool to projects. One of them was for the detection and characterization of microRNAs that are responsible for the regulation of immune system related genes in *Galleria mellonella*. These results and findings were published in Nature Scientific Reports.
As previously mentioned, the `microPIECE` pipeline needs a reference genome for the reliable detection of novel microRNAs. At the very beginning of the project, *G. mellonella* only had a newly submitted assembled genome, without any annotation and only de-novo assemblies of the transcriptome were available. I therefor had to annotate this genome first by myself, but was also able to create the, to my knowledge, first reference based assembled transcriptome of *G. mellonella*. Meanwhile, in 2020, several versions of annotations are freely available. Since the available data at that early time was limited and re-sequencing of potentially weak regions was no option, I had to use the data and results as they were. Nevertheless, I was able to create an annotated genome and reference-based assembled transcriptome that was usable for the microRNA project.
Due to the fact that the assembled genome was newly published, I was able to perform the first mining of microRNAs in *G. mellonella*, detecting potentially novel microRNAs and providing the first landscape of smallRNA-seq derived microRNAs in the genome. Previous studies were only able to map their samples to already known microRNAs, without the possibility of identifying novel microRNAs (Mannala et al. 2017). These approaches also have to deal with the uncertainty of only nearly-identical sequences and their assignment to an existing miRNA or miRNA family. Whereas my approach was able to identify such nearly identical microRNAs as individuals or homologous sequences or isomiRs. In total, I identified 141 unique precursor

miRNAs and 257 unique mature miRNAs, whose peak length is 22 nucleotides, perfectly matching the common average of mature microRNAs (O'Brien et al. 2018). From those 257 miRNAs, 95 were reported as novel and 148 were reported as conserved. For 14 microRNAs the results were not entirely clear. Three mature microRNAs were identical to a reference, but the opposite mature strand showed one mismatch (gme-new-41-5p and gme-new-42-5p) or even three mismatches, in case of gme-new-47-3p. Two mature microRNAs were anti-strand orientated compared to the reference miR-8. These results demonstrate that the identification of novel microRNAs is still a non-trivial task that contains more than simply checking if two reads are forming a hairpin sequence. It is furthermore at some point very subjective. The `miRBase.org` FAQ advices to name novel miRNAs with lexicographically increasing characters if they are highly similar and with increasing numberings if they are identical but derived from different precursors. For identical miRNAs this task can be comparably easy, but at some point, one has to decide how many alterations are tolerated until the miRNA is considered as independent or still belonging to the family and receives a character extension (like miR-121a and miR-121b). The change-log of `miRBase.org` visualizes this challenge quite nicely. Here one can observe several microRNAs that have been reported but were then erased from the database again or modified in another way, like the re-naming or sequence alteration. Exemplarily, from the last version 21 to the current version 22, 582 re-naming, 805 sequence changes and 179 deletions were reported.

With the aid of my pipeline, I identified the expression of the miRNAs in all three conditions and was able to compare these expressions. For 147 miRNAs, the expression between all three conditions was constant, leading to the assumption that those could be somehow related to house-keeping genes or at least infection-independent genes. Five microRNAs showed a specific expression in ABU infected larvae and 18 microRNAs were specific to UPEC larvae, whereas three microRNAs were expression only in the control. Furthermore, three miRNAs were differentially expressed between ABU larvae and control larvae, 19 miRNAs were differentially expressed between UPEC larvae and controls and 18 miRNAs were differentially expressed between UPEC and ABU infected larvae. A combination of ABU and UPEC expressed miRNAs revealed several infection specific miRNAs. These specific findings lead to the suggestion that miRNAs play a role in the different pathogenic reactions of an organism to an infection with ABU or UPEC bacteria.

Further insights of miRNA impact to the immune response are given by the miRNA target prediction via the `microPIECE` pipeline. In order to account for a putative bias of the target prediction algorithm used in the pipeline, I additionally verified the miRNAs that target genes,

regulating the innate immunity and epigenetic mechanisms with `RNAhybrid` and `RNA22-v2`. The `RNAhybrid` algorithm agrees with the standard `miranda` prediction. `RNA22-v2` was not able to verify five of the 21 miRNA-mRNA targets. This is mostly due to its pattern-based approach. It only accepts bindings that match its own salient sequence feature list, derived from known mature miRNAs. Rare or novel bindings are therefore excluded from the report.

For the annotation of miRNA binding sites and the transfer of AGO-CLIP regions I used the *Aedes aegypti* dataset, because it was the closest available species with such a dataset. In addition, the AGO-CLIP dataset was created from the fat body, which is, as previously reported, related to the insect immune system. It can therefore be expected to serve as a good source for immunity related miRNA regulation annotations. The AGO-CLIP dataset itself separates into two conditions, one is the pre- and the other is the post-blood-meal data. Especially in the post-blood-meal dataset one can expect an immune response pathways activation and regulation by miRNAs (Kumar et al. 2018).

The target prediction with the CLIP approach resulted in around 7.000 miRNA-mRNA target pairs, whereas a target prediction on the entire mRNA set would result in more than 100-times as much miRNA-mRNA target pairs. Therefore, the number of potentially interesting pairings is much smaller with my pipeline than with a regular prediction, making it much more comfortable, time efficient and straight forward to validate the combinations in the wet lab, like it was done in this project with qPCR.

The `microPIECE` target prediction revealed mRNA targets that belong to innate immunity and epigenetic mechanisms. The algorithm identified AMP-binding enzymes, phosphatidylinositol 3-kinase, AMP-dependent synthetase/ligase, lysozyme, histone deacetylase (HDACs), histone acetyltransferase (HATs) and methyltransferases. For some miRNA-mRNA pairs, a mRNA expression change can be observed with qPCR, whereas for others not. This can mainly be explained via the commonly accepted behavior of miRNAs where some bindings lead to degradation of the mRNA and some to a blocking of translation without a change in expression (Thermann and Hentze 2007).

In general, my pipeline revealed several microRNAs that are expressed upon UPEC or ABU infection or even in both. Several of those microRNAs are as well conserved in mammals and especially in human, like mir-8 family, let-7 and mir-124 and show homologous targets genes, like HAT KAT2A and HDAC8. This is an indication for a relation between miRNAs and further epigenetic regulations, like the chromatin structure accessibility.

In conclusion, I provided a best-practice workflow for the main research questions concerning microRNAs, together with a fully functional pipeline for the all-in-one analysis of microRNAs, not limited to a certain species. This was achieved by investigating the best settings for the best programs for the individual tasks. For the microRNA isoform detection, no benchmark of the tools was available, so I measured the performance of several high-throughput open source tools on an artificial dataset and applied the best suitable tool to an early development dataset of *T. castaneum*, highlighting the importance of miRNA post-transcriptional modifications and leading to the conclusion that they might be of more importance than previously thought.
Within the scope of my pipeline, I developed a novel technique for microRNA target prediction, the AGO CLIP-seq transfer. This approach relies on evolutionary conserved and experimentally verified data and shrinks the search space for the target prediction, making it faster and more conservative. In addition to this pipeline, I here presented my database scheme that allows a structured analysis and deposition of multiple pipeline results from different species, enabling various comparative analysis approaches. Besides several projects, I applied my pipeline to *G. mellonella*, where I additionally made the first reference-based assembly of the transcriptome with annotation. Furthermore, I published the first miRNAome dataset derived from smallRNA-seq data for this species. Finally, I showed the importance of microRNAs for the immune response to UPEC and ABU infections in *G. mellonella*, using the `microPIECE` pipeline with a potential outlook on the usability towards human medicine.

Finally, I can state that due to my improvements on the functional landscapes of microRNAs, one can add *G. mellonella* to the list of usable insects for human beings in terms of microRNAs and with the help of my pipeline many others can be appended to the list in further projects.

10

REFERENCES

Abrol, Dharam Pal, Anil Kumar Gorka, Mohammad Javed Ansari, Ahmad Al-Ghamdi, and Saad Al-Kahtani. 2017. “Impact of Insect Pollinators on Yield and Fruit Quality of Strawberry.” *Saudi Journal of Biological Sciences*, August. https://doi.org/10.1016/J.SJBS.2017.08.003.

Agarwal, Vikram, George W Bell, Jin-Wu Nam, and David P Bartel. 2015. “Predicting Effective MicroRNA Target Sites in Mammalian MRNAs.” *ELife* 4 (August). https://doi.org/10.7554/eLife.05005.

Agrawal, Neema, Bindiya Sachdev, Janneth Rodrigues, K. Sowjanya Sree, and Raj K. Bhatnagar. 2013. “Development Associated Profiling of Chitinase and MicroRNA of Helicoverpa Armigera Identified Chitinase Repressive MicroRNA.” *Scientific Reports* 3 (July): 2292. https://doi.org/10.1038/srep02292.

Akiva, Eyal, Shoshana Brown, Daniel E. Almonacid, Alan E. Barber, Ashley F. Custer, Michael A. Hicks, Conrad C. Huang, et al. 2014. “The Structure–Function Linkage Database.” *Nucleic Acids Research* 42 (D1): D521–30. https://doi.org/10.1093/nar/gkt1130.

Alghoribi, Majed F, Tarek M Gibreel, Andrew R Dodgson, Scott A Beatson, and Mathew Upton. 2014. “Galleria Mellonella Infection Model Demonstrates High Lethality of ST69 and ST127 Uropathogenic E. Coli.” Edited by Dipshikha Chakravortty. *PloS One* 9 (7): e101547. https://doi.org/10.1371/journal.pone.0101547.

Altenhoff, Adrian M, Natasha M Glover, Clément-Marie Train, Klara Kaleb, Alex Warwick Vesztrocy, David Dylus, Tarcisio M de Farias, et al. 2018. “The OMA Orthology Database in 2018: Retrieving Evolutionary Relationships among All Domains of Life through Richer Web and Programmatic Interfaces.” *Nucleic Acids Research* 46 (D1): D477–85. https://doi.org/10.1093/nar/gkx1019.

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. 1990. “Basic Local Alignment Search Tool.” *Journal of Molecular Biology* 215 (3): 403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.

Ambite, Ines, Nataliya Lutay, Christoph Stork, Ulrich Dobrindt, Björn Wullt, and Catharina Svanborg. 2016. “Bacterial Suppression of RNA Polymerase II-Dependent Host Gene Expression.” *Pathogens* 5 (3): 49. https://doi.org/10.3390/pathogens5030049.

Amsel, Daniel, André Billion, Andreas Vilcinskas, and Frank Förster. 2018. “MicroPIECE - MicroRNA Pipeline Enhanced by CLIP Experiments.” *Journal of Open Source Software* 3 (24): 616. https://doi.org/10.21105/joss.00616.

Anders, Simon, and Wolfgang Huber. 2010. “Differential Expression Analysis for Sequence

Count Data." *Genome Biology* 11 (10): R106. https://doi.org/10.1186/gb-2010-11-10-r106.
Asgari, Sassan. 2013. "MicroRNA Functions in Insects." *Insect Biochemistry and Molecular Biology* 43 (4): 388–97. https://doi.org/10.1016/j.ibmb.2012.10.005.
Attwood, T K, P Bradley, D R Flower, A Gaulton, N Maudling, A L Mitchell, G Moulton, et al. 2003. "PRINTS and Its Automatic Supplement, PrePRINTS." *Nucleic Acids Research* 31 (1): 400–402. http://www.ncbi.nlm.nih.gov/pubmed/12520033.
Baillie, Jonathan E M, Craig Hilton-Taylor, and Simon N Stuart. 2004. "A Global Species Assessment."
Bartel, David P. 2009. "MicroRNAs: Target Recognition and Regulatory Functions." *Cell* 136 (2): 215–33. https://doi.org/10.1016/j.cell.2009.01.002.
Bartel, David P. 2004. "MicroRNAs: Genomics, Biogenesis, Mechanism, and Function." *Cell* 116 (2): 281–97. https://doi.org/10.1016/S0092-8674(04)00045-5.
Bernstein, Emily, Amy A. Caudy, Scott M. Hammond, and Gregory J. Hannon. 2001. "Role for a Bidentate Ribonuclease in the Initiation Step of RNA Interference." *Nature* 409 (6818): 363–66. https://doi.org/10.1038/35053110.
Betel, D., M. Wilson, A. Gabow, D. S. Marks, and C. Sander. 2007a. "The MicroRNA.Org Resource: Targets and Expression." *Nucleic Acids Research* 36 (Database): D149–53. https://doi.org/10.1093/nar/gkm995.
———. 2007b. "The MicroRNA.Org Resource: Targets and Expression." *Nucleic Acids Research* 36 (Database): D149–53. https://doi.org/10.1093/nar/gkm995.
Bhangale, Tushar R., Mark J. Rieder, Robert J. Livingston, and Deborah A. Nickerson. 2005. "Comprehensive Identification and Characterization of Diallelic Insertion–Deletion Polymorphisms in 330 Human Candidate Genes." *Human Molecular Genetics* 14 (1): 59–69. https://doi.org/10.1093/hmg/ddi006.
Bianchi, Marzia, Alessandra Renzini, Sergio Adamo, and Viviana Moresi. 2017. "Coordinated Actions of MicroRNAs with Other Epigenetic Factors Regulate Skeletal Muscle Development and Adaptation." *International Journal of Molecular Sciences* 18 (4). https://doi.org/10.3390/ijms18040840.
Bierne, Hélène, Mélanie Hamon, and Pascale Cossart. 2012. "Epigenetics and Bacterial Infections." *Cold Spring Harbor Perspectives in Medicine* 2 (12): a010272. https://doi.org/10.1101/cshperspect.a010272.
Bingsohn, Linda, Eileen Knorr, and Andreas Vilcinskas. 2016. "The Model Beetle Tribolium Castaneum Can Be Used as an Early Warning System for Transgenerational Epigenetic Side Effects Caused by Pharmaceuticals." *Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology* 185–186 (July): 57–64. https://doi.org/10.1016/J.CBPC.2016.03.002.
Biryukova, Inna, Joëlle Asmar, Houari Abdesselem, and Pascal Heitzler. 2009. "Drosophila Mir-9a Regulates Wing Development via Fine-Tuning Expression of the LIM Only Factor, DLMO." *Developmental Biology* 327 (2): 487–96. https://doi.org/10.1016/J.YDBIO.2008.12.036.
Bohnsack, Markus T, Kevin Czaplinski, and Dirk Gorlich. 2004. "Exportin 5 Is a RanGTP-Dependent DsRNA-Binding Protein That Mediates Nuclear Export of Pre-MiRNAs." *RNA (New York, N.Y.)* 10 (2): 185–91. https://doi.org/10.1261/RNA.5167604.
Bolger, Anthony M, Marc Lohse, and Bjoern Usadel. 2014. "Trimmomatic: A Flexible Trimmer for Illumina Sequence Data." *Bioinformatics (Oxford, England)* 30 (15): 2114–20. https://doi.org/10.1093/bioinformatics/btu170.
Bonnet, E., J. Wuyts, P. Rouze, and Y. Van de Peer. 2004. "Evidence That MicroRNA

Precursors, Unlike Other Non-Coding RNAs, Have Lower Folding Free Energies than Random Sequences." *Bioinformatics* 20 (17): 2911–17. https://doi.org/10.1093/bioinformatics/bth374.

Bork, Peer, Thomas Dandekar, Yolande Diaz-Lazcoz, Frank Eisenhaber, Martijn Huynen, and Yanping Yuan. 1998. "Predicting Function: From Genes to Genomes and Back." *Journal of Molecular Biology* 283 (4): 707–25. https://doi.org/10.1006/JMBI.1998.2144.

Brennecke, Julius, Alexander Stark, Robert B Russell, and Stephen M Cohen. 2005. "Principles of MicroRNA–Target Recognition." Edited by James C. Carrington. *PLoS Biology* 3 (3): e85. https://doi.org/10.1371/journal.pbio.0030085.

Bru, Catherine, Emmanuel Courcelle, Sébastien Carrère, Yoann Beausse, Sandrine Dalmar, and Daniel Kahn. 2005. "The ProDom Database of Protein Domain Families: More Emphasis on 3D." *Nucleic Acids Research* 33 (Database issue): D212-5. https://doi.org/10.1093/nar/gki034.

Burrows, M., M. Burrows, and D. J. Wheeler. 1994. "A Block-Sorting Lossless Data Compression Algorithm," 16. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.3.8069.

Bushmanova, Elena, Dmitry Antipov, Alla Lapidus, Vladimir Suvorov, and Andrey D. Prjibelski. 2016. "RnaQUAST: A Quality Assessment Tool for *de Novo* Transcriptome Assemblies: Table 1." *Bioinformatics* 32 (14): 2210–12. https://doi.org/10.1093/bioinformatics/btw218.

Cai, Xuezhong, Curt H Hagedorn, and Bryan R Cullen. 2004. "Human MicroRNAs Are Processed from Capped, Polyadenylated Transcripts That Can Also Function as MRNAs." *RNA (New York, N.Y.)* 10 (12): 1957–66. https://doi.org/10.1261/rna.7135204.

Calin, George A., Amelia Cimmino, Muller Fabbri, Manuela Ferracin, Sylwia E. Wojcik, Masayoshi Shimizu, Cristian Taccioli, et al. 2008. "MiR-15a and MiR-16-1 Cluster Functions in Human Leukemia." *Proceedings of the National Academy of Sciences* 105 (13): 5166–71. https://doi.org/10.1073/pnas.0800121105.

Calin, George Adrian, Calin Dan Dumitru, Masayoshi Shimizu, Roberta Bichi, Simona Zupo, Evan Noch, Hansjuerg Aldler, et al. 2002. "Frequent Deletions and Down-Regulation of Micro- RNA Genes MiR15 and MiR16 at 13q14 in Chronic Lymphocytic Leukemia." *Proceedings of the National Academy of Sciences of the United States of America* 99 (24): 15524–29. https://doi.org/10.1073/pnas.242606799.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. 2009. "BLAST+: Architecture and Applications." *BMC Bioinformatics* 10 (1): 421. https://doi.org/10.1186/1471-2105-10-421.

Chan, Wen-Ching, Meng-Ru Ho, Sung-Chou Li, Kuo-Wang Tsai, Chun-Hung Lai, Chun-Nan Hsu, and Wen-chang Lin. 2012. "MetaMirClust: Discovery of MiRNA Cluster Patterns Using a Data-Mining Approach." *Genomics* 100 (3): 141–48. https://doi.org/10.1016/J.YGENO.2012.06.007.

Chang, Zhao-Xia, Nan Tang, Lin Wang, Li-Qing Zhang, Ibukun A Akinyemi, and Qing-Fa Wu. 2016. "Identification and Characterization of MicroRNAs in the White-Backed Planthopper, Sogatella Furcifera." *Insect Science* 23: 452–68. https://doi.org/10.1111/1744-7917.12343.

Charriere, Jean-Daniel, and Anton Imdorf. 1999. "Protection of Honey Combs from Wax Moth Damage." *American Bee Journal*.

Chi, Sung Wook, Gregory J Hannon, and Robert B Darnell. 2012. "An Alternative Mode of MicroRNA Target Recognition." *Nature Structural & Molecular Biology* 19 (3): 321–27. https://doi.org/10.1038/nsmb.2230.

Chi, Sung Wook, Julie B Zang, Aldo Mele, and Robert B Darnell. 2009. “Argonaute HITS-CLIP Decodes MicroRNA-MRNA Interaction Maps.” *Nature* 460 (7254): 479–86. https://doi.org/10.1038/nature08170.

Chou, Chih-Hung, Feng-Mao Lin, Min-Te Chou, Sheng-Da Hsu, Tzu-Hao Chang, Shun-Long Weng, Sirjana Shrestha, Chiung-Chih Hsiao, Jui-Hung Hung, and Hsien-Da Huang. 2013. “A Computational Approach for Identifying MicroRNA-Target Interactions Using High-Throughput CLIP and PAR-CLIP Sequencing.” *BMC Genomics* 14 Suppl 1: S2. https://doi.org/10.1186/1471-2164-14-S1-S2.

Chou, Chih-Hung, Sirjana Shrestha, Chi-Dung Yang, Nai-Wen Chang, Yu-Ling Lin, Kuang-Wen Liao, Wei-Chi Huang, et al. 2018. “MiRTarBase Update 2018: A Resource for Experimentally Validated MicroRNA-Target Interactions.” *Nucleic Acids Research* 46 (D1): D296–302. https://doi.org/10.1093/nar/gkx1067.

Ciesielczuk, Holly, Jonathon Betts, Lynnette Phee, Michel Doumith, Russell Hope, Neil Woodford, and David W Wareham. 2015. “Comparative Virulence of Urinary and Bloodstream Isolates of Extra-Intestinal Pathogenic Escherichia Coli in a Galleria Mellonella Model.” *Virulence* 6 (2): 145–51. https://doi.org/10.4161/21505594.2014.988095.

Cimmino, A., G. A. Calin, M. Fabbri, M. V. Iorio, M. Ferracin, M. Shimizu, S. E. Wojcik, et al. 2005. “MiR-15 and MiR-16 Induce Apoptosis by Targeting BCL2.” *Proceedings of the National Academy of Sciences* 102 (39): 13944–49. https://doi.org/10.1073/pnas.0506654102.

Cirl, Christine, Andreas Wieser, Manisha Yadav, Susanne Duerr, Sören Schubert, Hans Fischer, Dominik Stappert, et al. 2008. “Subversion of Toll-like Receptor Signaling by a Unique Family of Bacterial Toll/Interleukin-1 Receptor Domain–Containing Proteins.” *Nature Medicine* 14 (4): 399–406. https://doi.org/10.1038/nm1734.

Cloonan, Nicole. 2015. “Re-Thinking MiRNA-MRNA Interactions: Intertwining Issues Confound Target Discovery.” *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology* 37 (4): 379–88. https://doi.org/10.1002/bies.201400191.

Cook, Simon M, and Jason D McArthur. 2013. “Developing Galleria Mellonella as a Model Host for Human Pathogens.” *Virulence* 4 (5): 350–53. https://doi.org/10.4161/viru.25240.

Cordes, Kimberly R., Neil T. Sheehy, Mark P. White, Emily C. Berry, Sarah U. Morton, Alecia N. Muth, Ting-Hein Lee, Joseph M. Miano, Kathryn N. Ivey, and Deepak Srivastava. 2009. “MiR-145 and MiR-143 Regulate Smooth Muscle Cell Fate and Plasticity.” *Nature* 460 (7256): 705. https://doi.org/10.1038/nature08195.

Crane, Eva. 2005. “The Rock Art of Honey Hunters.” *Bee World* 86 (1): 11–13. https://doi.org/10.1080/0005772X.2005.11099642.

Das, Kishore, Omar Garnica, and Subramanian Dhandayuthapani. 2016. “Modulation of Host MiRNAs by Intracellular Bacterial Pathogens.” *Frontiers in Cellular and Infection Microbiology* 6 (August): 79. https://doi.org/10.3389/fcimb.2016.00079.

Didiano, Dominic, and Oliver Hobert. 2006. “Perfect Seed Pairing Is Not a Generally Reliable Predictor for MiRNA-Target Interactions.” *Nature Structural & Molecular Biology* 13 (9): 849–51. https://doi.org/10.1038/nsmb1138.

Dobrindt, Ulrich, Björn Wullt, and Catharina Svanborg. 2016. “Asymtomatic Bacteriuria as a Model to Study the Coevolution of Hosts and Bacteria.” *Pathogens (Basel, Switzerland)* 5 (1): 21. https://doi.org/10.3390/pathogens5010021.

Dossey, A.T., J.T. Tatum, and W.L. McGill. 2016. “Modern Insect-Based Food Industry: Current Status, Insect Processing Technology, and Recommendations Moving Forward.”

Insects as Sustainable Food Ingredients, January, 113–52. https://doi.org/10.1016/B978-0-12-802856-8.00005-3.
Enright, Anton J, Bino John, Ulrike Gaul, Thomas Tuschl, Chris Sander, and Debora S Marks. 2003. "MicroRNA Targets in Drosophila." *Genome Biology* 5 (1): R1. https://doi.org/10.1186/gb-2003-5-1-r1.
Etebari, Kayvan, and Sassan Asgari. 2013. "Conserved MicroRNA MiR-8 Blocks Activation of the Toll Pathway by Upregulating Serpin 27 Transcripts." *RNA Biology* 10 (8): 1356–64. https://doi.org/10.4161/rna.25481.
Eulalio, A., E. Huntzinger, T. Nishihara, J. Rehwinkel, M. Fauser, and E. Izaurralde. 2008. "Deadenylation Is a Widespread Effect of MiRNA Regulation." *RNA* 15 (1): 21–32. https://doi.org/10.1261/rna.1399509.
Fernandez-Valverde, Selene L, Ryan J Taft, and John S Mattick. 2010. "Dynamic IsomiR Regulation in Drosophila Development." *RNA (New York, N.Y.)* 16 (10): 1881–88. https://doi.org/10.1261/rna.2379610.
Finn, Robert D., Penelope Coggill, Ruth Y. Eberhardt, Sean R. Eddy, Jaina Mistry, Alex L. Mitchell, Simon C. Potter, et al. 2016. "The Pfam Protein Families Database: Towards a More Sustainable Future." *Nucleic Acids Research* 44 (D1): D279–85. https://doi.org/10.1093/nar/gkv1344.
Fitch, W M. 1970. "Distinguishing Homologous from Analogous Proteins." *Systematic Zoology* 19 (2): 99–113. http://www.ncbi.nlm.nih.gov/pubmed/5449325.
Foxman, Betsy. 2014. "Urinary Tract Infection Syndromes: Occurrence, Recurrence, Bacteriology, Risk Factors, and Disease Burden." *Infectious Disease Clinics of North America* 28 (1): 1–13. https://doi.org/10.1016/j.idc.2013.09.003.
Friedländer, Marc R, Sebastian D Mackowiak, Na Li, Wei Chen, and Nikolaus Rajewsky. 2012. "MiRDeep2 Accurately Identifies Known and Hundreds of Novel MicroRNA Genes in Seven Animal Clades." *Nucleic Acids Research* 40 (1): 37–52. https://doi.org/10.1093/nar/gkr688.
Gaidatzis, Dimos, Erik van Nimwegen, Jean Hausser, and Mihaela Zavolan. 2007. "Inference of MiRNA Targets Using Evolutionary Conservation and Pathway Analysis." *BMC Bioinformatics* 8 (1): 69. https://doi.org/10.1186/1471-2105-8-69.
Glastad, K. M., B. G. Hunt, S. V. Yi, and M. A. D. Goodisman. 2011. "DNA Methylation in Insects: On the Brink of the Epigenomic Era." *Insect Molecular Biology* 20 (5): 553–65. https://doi.org/10.1111/j.1365-2583.2011.01092.x.
Godaly, Gabriela, Ines Ambite, and Catharina Svanborg. 2015. "Innate Immunity and Genetic Determinants of Urinary Tract Infection Susceptibility." *Current Opinion in Infectious Diseases* 28 (1): 1. https://doi.org/10.1097/QCO.0000000000000127.
Good, Irene, Jonathan Mark Kenoyer, and Richard Meadow. 2008. "New Evidence for Early Silk in the Indus Civilization." *Nature Precedings*, May, 1–1. https://doi.org/10.1038/npre.2008.1900.1.
Good, N E. 1936. "The Flour Beetles of the Genus Tribolium." *Technical Bulletin*, no. 498: 1–51. https://ideas.repec.org/p/ags/uerstb/164672.html.
Grabiec, Aleksander M, and Jan Potempa. 2018. "Epigenetic Regulation in Bacterial Infections: Targeting Histone Deacetylases." *Critical Reviews in Microbiology* 44 (3): 336–50. https://doi.org/10.1080/1040841X.2017.1373063.
Griffiths-Jones, S., Russell J. Grocock, Stijn van Dongen, Alex Bateman, and Anton J. Enright. 2006. "MiRBase: MicroRNA Sequences, Targets and Gene Nomenclature." *Nucleic Acids Research* 34 (90001): D140–44. https://doi.org/10.1093/nar/gkj112.
Griffiths-Jones, Sam. 2004. "The MicroRNA Registry." *Nucleic Acids Research* 32 (Database

issue): D109-11. https://doi.org/10.1093/nar/gkh023.
Griffiths-Jones, Sam, Jerome H L Hui, Antonio Marco, and Matthew Ronshaugen. 2011. "MicroRNA Evolution by Arm Switching." *EMBO Reports* 12 (2): 172–77. https://doi.org/10.1038/embor.2010.191.
Groombridge, Brian., Martin. Jenkins, World Conservation Monitoring Centre., and United Nations Environment Programme. 2002. *World Atlas of Biodiversity : Earth's Living Resources in the 21st Century*. University of California Press. https://archive.org/details/worldatlasofbiod02groo.
Guo, Li, Jiafeng Yu, Tingming Liang, and Quan Zou. 2016. "MiR-IsomiRExp: A Web-Server for the Analysis of Expression of MiRNA at the MiRNA/IsomiR Levels." *Scientific Reports* 6: 23700. https://doi.org/10.1038/srep23700.
Haas, Brian J, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D Blood, Joshua Bowden, Matthew Brian Couger, et al. 2013. "De Novo Transcript Sequence Reconstruction from RNA-Seq Using the Trinity Platform for Reference Generation and Analysis." *Nature Protocols* 8 (8): 1494–1512. https://doi.org/10.1038/nprot.2013.084.
Hafner, Markus, Markus Landthaler, Lukas Burger, Mohsen Khorshid, Jean Hausser, Philipp Berninger, Andrea Rothballer, et al. 2010. "Transcriptome-Wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP." *Cell* 141 (1): 129–41. https://doi.org/10.1016/j.cell.2010.03.009.
Haft, Daniel H, Jeremy D Selengut, and Owen White. 2003. "The TIGRFAMs Database of Protein Families." *Nucleic Acids Research* 31 (1): 371–73. http://www.ncbi.nlm.nih.gov/pubmed/12520025.
Hagberg, L, R Hull, S Hull, J R McGhee, S M Michalek, and C Svanborg Edén. 1984. "Difference in Susceptibility to Gram-Negative Urinary Tract Infection between C3H/HeJ and C3H/HeN Mice." *Infection and Immunity* 46 (3): 839–44. http://www.ncbi.nlm.nih.gov/pubmed/6389367.
Hammond, S M, E Bernstein, D Beach, and G J Hannon. 2000. "An RNA-Directed Nuclease Mediates Post-Transcriptional Gene Silencing in Drosophila Cells." *Nature* 404 (6775): 293–96. https://doi.org/10.1038/35005107.
Hausser, Jean, Afzal Pasha Syed, Biter Bilen, and Mihaela Zavolan. 2013. "Analysis of CDS-Located MiRNA Target Sites Suggests That They Can Effectively Inhibit Translation." *Genome Research* 23 (4): 604–15. https://doi.org/10.1101/gr.139758.112.
Heise, P, R Hirsch, H Vogel, and A Vilcinskas. 2016. "How to Fight against the Decay- the Burying Beetle and Its Gut Microbiota." *Planta Medica* 81 (S 01): S1–381. https://doi.org/10.1055/s-0036-1596811.
Heitmueller, Miriam, André Billion, Ulrich Dobrindt, Andreas Vilcinskas, and Krishnendu Mukherjee. 2017. "Epigenetic Mechanisms Regulate Innate Immunity against Uropathogenic and Commensal-Like Escherichia Coli in the Surrogate Insect Model Galleria Mellonella." *Infection and Immunity* 85 (10): e00336-17. https://doi.org/10.1128/IAI.00336-17.
Hill, Dennis S. 1983. *Agricultural Insect Pests of the Tropics and Their Control*. CUP Archive.
Hirsh, Aaron E., and Hunter B. Fraser. 2001. "Protein Dispensability and Rate of Evolution." *Nature* 411 (6841): 1046–49. https://doi.org/10.1038/35082561.
Hofacker, I. L. 2003. "Vienna RNA Secondary Structure Server." *Nucleic Acids Research* 31 (13): 3429–31. https://doi.org/10.1093/nar/gkg599.
Hofacker, I. L., W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster. 1994. "Fast Folding and Comparison of RNA Secondary Structures." *Monatshefte F□r Chemie Chemical Monthly* 125 (2): 167–88. https://doi.org/10.1007/BF00818163.

Huang, Weichun, Leping Li, Jason R Myers, and Gabor T Marth. 2012. “ART: A next-Generation Sequencing Read Simulator.” *Bioinformatics (Oxford, England)* 28 (4): 593–94. https://doi.org/10.1093/bioinformatics/btr708.

Hulo, Nicolas, Amos Bairoch, Virginie Bulliard, Lorenzo Cerutti, Edouard De Castro, Petra S Langendijk-Genevaux, Marco Pagni, and Christian J A Sigrist. 2006. “The PROSITE Database.” *Nucleic Acids Research* 34 (Database issue): D227-30. https://doi.org/10.1093/nar/gkj063.

Humphreys, D. T., B. J. Westman, D. I. K. Martin, and T. Preiss. 2005. “MicroRNAs Control Translation Initiation by Inhibiting Eukaryotic Initiation Factor 4E/Cap and Poly(A) Tail Function.” *Proceedings of the National Academy of Sciences* 102 (47): 16961–66. https://doi.org/10.1073/pnas.0506482102.

Hutvágner, G, J McLachlan, A E Pasquinelli, E Bálint, T Tuschl, and P D Zamore. 2001. “A Cellular Function for the RNA-Interference Enzyme Dicer in the Maturation of the Let-7 Small Temporal RNA.” *Science (New York, N.Y.)* 293 (5531): 834–38. https://doi.org/10.1126/science.1062961.

Jagadeeswaran, Guru, Yun Zheng, Niranji Sumathipala, Haobo Jiang, Estela L Arrese, Jose L Soulages, Weixiong Zhang, and Ramanjulu Sunkar. 2010. “Deep Sequencing of Small RNA Libraries Reveals Dynamic Regulation of Conserved and Novel MicroRNAs and MicroRNA-Stars during Silkworm Development.” *BMC Genomics* 11 (January): 52. https://doi.org/10.1186/1471-2164-11-52.

Jayachandran, Balachandran, Mazhar Hussain, and Sassan Asgari. 2013. “An Insect Trypsin-like Serine Protease as a Target of MicroRNA: Utilization of MicroRNA Mimics and Inhibitors by Oral Feeding.” *Insect Biochemistry and Molecular Biology* 43 (4): 398–406. https://doi.org/10.1016/j.ibmb.2012.10.004.

Jha, Ashwani, and Ravi Shankar. 2013. “MiReader: Discovering Novel MiRNAs in Species without Sequenced Genome.” *PLoS ONE* 8 (6). https://doi.org/10.1371/journal.pone.0066857.

Johnson, L Steven, Sean R Eddy, and Elon Portugaly. 2010. “Hidden Markov Model Speed Heuristic and Iterative HMM Search Procedure.” *BMC Bioinformatics* 11 (1): 431. https://doi.org/10.1186/1471-2105-11-431.

Jones, P., D. Binns, H.-Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, et al. 2014. “InterProScan 5: Genome-Scale Protein Function Classification.” *Bioinformatics* 30 (9): 1236–40. https://doi.org/10.1093/bioinformatics/btu031.

Käll, Lukas, Anders Krogh, and Erik L.L Sonnhammer. 2004. “A Combined Transmembrane Topology and Signal Peptide Prediction Method.” *Journal of Molecular Biology* 338 (5): 1027–36. https://doi.org/10.1016/J.JMB.2004.03.016.

Kent, W James, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. 2002. “The Human Genome Browser at UCSC.” *Genome Research* 12 (6): 996–1006. https://doi.org/10.1101/gr.229102.

Kertesz, Michael, Nicola Iovino, Ulrich Unnerstall, Ulrike Gaul, and Eran Segal. 2007. “The Role of Site Accessibility in MicroRNA Target Recognition.” *Nature Genetics* 39 (10): 1278–84. https://doi.org/10.1038/ng2135.

Khvorova, Anastasia, Angela Reynolds, and Sumedha D Jayasena. 2003. “Functional SiRNAs and MiRNAs Exhibit Strand Bias.” *Cell* 115 (2): 209–16. http://www.ncbi.nlm.nih.gov/pubmed/14567918.

Kim, Daehwan, Ben Langmead, and Steven L Salzberg. 2015. “HISAT: A Fast Spliced Aligner with Low Memory Requirements.” *Nature Methods* 12 (4): 357–60. https://doi.org/10.1038/nmeth.3317.

Kim, Haedong, Jimi Kim, Sha Yu, Young Yoon Lee, Junseong Park, Ran Joo Choi, Seon Jin Yoon, Seok Gu Kang, and V. Narry Kim. 2020. “A Mechanism for MicroRNA Arm Switching Regulated by Uridylation.” *Molecular Cell* 78 (6): 1224-1236.e5. https://doi.org/10.1016/j.molcel.2020.04.030.

Kim, V. Narry, Jinju Han, and Mikiko C. Siomi. 2009. “Biogenesis of Small RNAs in Animals.” *Nature Reviews Molecular Cell Biology* 10 (2): 126–39. https://doi.org/10.1038/nrm2632.

Kinsella, R. J., A. Kahari, S. Haider, J. Zamora, G. Proctor, G. Spudich, J. Almeida-King, et al. 2011. “Ensembl BioMarts: A Hub for Data Retrieval across Taxonomic Space.” *Database* 2011 (0): bar030–bar030. https://doi.org/10.1093/database/bar030.

Klatt, Björn K, Andrea Holzschuh, Catrin Westphal, Yann Clough, Inga Smit, Elke Pawelzik, and Teja Tscharntke. 2014. “Bee Pollination Improves Crop Quality, Shelf Life and Commercial Value.” *Proceedings. Biological Sciences* 281 (1775): 20132440. https://doi.org/10.1098/rspb.2013.2440.

Klemm, Sandy L., Zohar Shipony, and William J. Greenleaf. 2019. “Chromatin Accessibility and the Regulatory Epigenome.” *Nature Reviews Genetics*. Nature Publishing Group. https://doi.org/10.1038/s41576-018-0089-8.

Knorr, Eileen, Linda Bingsohn, Michael R. Kanost, and Andreas Vilcinskas. 2013. “Tribolium Castaneum as a Model for High-Throughput RNAi Screening.” In *Advances in Biochemical Engineering/Biotechnology*, 136:163–78. https://doi.org/10.1007/10_2013_208.

König, Julian, Kathi Zarnack, Nicholas M. Luscombe, and Jernej Ule. 2012. “Protein–RNA Interactions: New Genomic Technologies and Perspectives.” *Nature Reviews Genetics* 13 (2): 77–83. https://doi.org/10.1038/nrg3141.

König, Julian, Kathi Zarnack, Gregor Rot, Tomaž Curk, Melis Kayikci, Blaž Zupan, Daniel J Turner, Nicholas M Luscombe, and Jernej Ule. 2010. “ICLIP Reveals the Function of HnRNP Particles in Splicing at Individual Nucleotide Resolution.” *Nature Structural & Molecular Biology* 17 (7): 909–15. https://doi.org/10.1038/nsmb.1838.

Kozomara, Ana, and Sam Griffiths-Jones. 2014. “MiRBase: Annotating High Confidence MicroRNAs Using Deep Sequencing Data.” *Nucleic Acids Research* 42 (D1): D68–73. https://doi.org/10.1093/nar/gkt1181.

Krek, Azra, Dominic Grün, Matthew N Poy, Rachel Wolf, Lauren Rosenberg, Eric J Epstein, Philip MacMenamin, et al. 2005. “Combinatorial MicroRNA Target Predictions.” *Nature Genetics* 37 (5): 495–500. https://doi.org/10.1038/ng1536.

Krogh, Anders, Björn Larsson, Gunnar von Heijne, and Erik L.L Sonnhammer. 2001. “Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genomes.” *Journal of Molecular Biology* 305 (3): 567–80. https://doi.org/10.1006/JMBI.2000.4315.

Kuchenbauer, Florian, Ryan D Morin, Bob Argiropoulos, Oleh I Petriv, Malachi Griffith, Michael Heuser, Eric Yung, et al. 2008. “In-Depth Characterization of the MicroRNA Transcriptome in a Leukemia Progression Model.” *Genome Research* 18 (11): 1787–97. https://doi.org/10.1101/gr.077578.108.

Kumar, Ankit, Priyanshu Srivastava, P. D.N.N. Sirisena, Sunil Kumar Dubey, Ramesh Kumar, Jatin Shrinet, and Sujatha Sunil. 2018. “Mosquito Innate Immunity.” *Insects*. MDPI AG. https://doi.org/10.3390/insects9030095.

Kwadha, Charles A., George O. Ong’amo, Paul N. Ndegwa, Suresh K. Raina, Ayuka T. Fombong, Charles A. Kwadha, George O. Ong’amo, Paul N. Ndegwa, Suresh K. Raina, and Ayuka T. Fombong. 2017. “The Biology and Control of the Greater Wax Moth,

Galleria Mellonella." *Insects* 8 (2): 61. https://doi.org/10.3390/insects8020061.
Kwast, Reginald V.C.T. van der, Tamar Woudenberg, Paul H.A. Quax, and A. Yaël Nossent. 2019. "MicroRNA-411 and Its 5'-IsomiR Have Distinct Targets and Functions and Are Differentially Regulated in the Vasculature under Ischemia." *Molecular Therapy* 28 (1): 157–70. https://doi.org/10.1016/j.ymthe.2019.10.002.
Lagos-Quintana, M., R Rauhut, W Lendeckel, and T Tuschl. 2001. "Identification of Novel Genes Coding for Small Expressed RNAs." *Science* 294 (5543): 853–58. https://doi.org/10.1126/science.1064921.
Lagos-Quintana, Mariana, Reinhard Rauhut, Winfried Lendeckel, and Thomas Tuschl. 2001. "Identification of Novel Genes Coding for Small Expressed RNAs." *Science* 294 (5543).
Lai, Eric C. 2015. "Two Decades of MiRNA Biology: Lessons and Challenges." *RNA (New York, N.Y.)* 21 (4): 675–77. https://doi.org/10.1261/rna.051193.115.
Landgraf, Pablo, Mirabela Rusu, Robert Sheridan, Alain Sewer, Nicola Iovino, Alexei Aravin, Sébastien Pfeffer, et al. 2007. "A Mammalian MicroRNA Expression Atlas Based on Small RNA Library Sequencing." *Cell* 129 (7): 1401–14. https://doi.org/10.1016/j.cell.2007.04.040.
Lange, Anna, Sina Beier, Daniel H Huson, Raphael Parusel, Franz Iglauer, and Julia-Stefanie Frick. 2018. "Genome Sequence of Galleria Mellonella (Greater Wax Moth)." *Genome Announcements* 6 (2). https://doi.org/10.1128/genomeA.01220-17.
Langmead, B, C Trapnell, M Pop, and S L Salzberg. 2009. "Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome." *Genome Biol*, 1–10. https://doi.org/gb-2009-10-3-r25 [pii]\r10.1186/gb-2009-10-3-r25.
Langmead, Ben, and Steven L Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9 (4): 357–59. https://doi.org/10.1038/nmeth.1923.
Lau, N C, L P Lim, E G Weinstein, and D P Bartel. 2001. "An Abundant Class of Tiny RNAs with Probable Regulatory Roles in Caenorhabditis Elegans." *Science (New York, N.Y.)* 294 (5543): 858–62. https://doi.org/10.1126/science.1065062.
Lechner, Marcus, Sven Findeiß, Lydia Steiner, Manja Marz, Peter F. Stadler, and Sonja J. Prohaska. 2011. "Proteinortho: Detection of (Co-)Orthologs in Large-Scale Analysis." *BMC Bioinformatics* 12 (April). https://doi.org/10.1186/1471-2105-12-124.
Lee, Lik Wee, Shile Zhang, Alton Etheridge, Li Ma, Dan Martin, David Galas, and Kai Wang. 2010. "Complexity of the MicroRNA Repertoire Revealed by Next-Generation Sequencing." *RNA (New York, N.Y.)* 16 (11): 2170–80. https://doi.org/10.1261/rna.2225110.
Lee, R C, R L Feinbaum, and V Ambros. 1993. "The C. Elegans Heterochronic Gene Lin-4 Encodes Small RNAs with Antisense Complementarity to Lin-14." *Cell* 75 (5): 843–54. http://www.ncbi.nlm.nih.gov/pubmed/8252621.
Lee, Yoontae, Chiyoung Ahn, Jinju Han, Hyounjeong Choi, Jaekwang Kim, Jeongbin Yim, Junho Lee, et al. 2003. "The Nuclear RNase III Drosha Initiates MicroRNA Processing." *Nature* 425 (6956): 415–19. https://doi.org/10.1038/nature01957.
Lee, Yoontae, Kipyoung Jeon, Jun-Tae Lee, Sunyoung Kim, V Narry Kim, E. Bernstein, AA. Caudy, et al. 2002. "MicroRNA Maturation: Stepwise Processing and Subcellular Localization." *The EMBO Journal* 21 (17): 4663–70. https://doi.org/10.1093/EMBOJ/CDF476.
Lee, Yoontae, Minju Kim, Jinju Han, Kyu-Hyun Yeom, Sanghyuk Lee, Sung Hee Baek, V Narry Kim, et al. 2004. "MicroRNA Genes Are Transcribed by RNA Polymerase II." *The EMBO Journal* 23 (20): 4051–60. https://doi.org/10.1038/sj.emboj.7600385.
Leimbach, Andreas, Jörg Hacker, and Ulrich Dobrindt. 2013. "E. Coli as an All-Rounder: The

Thin Line between Commensalism and Pathogenicity." *Current Topics in Microbiology and Immunology* 358: 3–32. https://doi.org/10.1007/82_2012_303.
Lemaitre, Bruno. 2004. "The Road to Toll." *Nature Reviews Immunology*. Nature Publishing Group. https://doi.org/10.1038/nri1390.
Letunic, Ivica, Tobias Doerks, and Peer Bork. 2012. "SMART 7: Recent Updates to the Protein Domain Annotation Resource." *Nucleic Acids Research* 40 (Database issue): D302-5. https://doi.org/10.1093/nar/gkr931.
Lewis, Benjamin P., I-hung Shih, Matthew W. Jones-Rhoades, David P. Bartel, and Christopher B. Burge. 2003. "Prediction of Mammalian MicroRNA Targets." *Cell* 115 (7): 787–98. https://doi.org/10.1016/S0092-8674(03)01018-3.
Lewis, Tony E, Ian Sillitoe, Natalie Dawson, Su Datt Lam, Tristan Clarke, David Lee, Christine Orengo, and Jonathan Lees. 2018. "Gene3D: Extensive Prediction of Globular Domains in Proteins." *Nucleic Acids Research* 46 (D1): D435–39. https://doi.org/10.1093/nar/gkx1069.
Li, H., and R. Durbin. 2009. "Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform." *Bioinformatics* 25 (14): 1754–60. https://doi.org/10.1093/bioinformatics/btp324.
Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25 (16): 2078–79. https://doi.org/10.1093/bioinformatics/btp352.
Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25 (16): 2078–79. https://doi.org/10.1093/bioinformatics/btp352.
Li, Li, Christian J Stoeckert, and David S Roos. 2003. "OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes." *Genome Research* 13 (9): 2178–89. https://doi.org/10.1101/gr.1224503.
Licatalosi, Donny D., Aldo Mele, John J. Fak, Jernej Ule, Melis Kayikci, Sung Wook Chi, Tyson A. Clark, et al. 2008. "HITS-CLIP Yields Genome-Wide Insights into Brain Alternative RNA Processing." *Nature* 456 (7221): 464–69. https://doi.org/10.1038/nature07488.
Liu, Chang-Gong, George Adrian Calin, Stefano Volinia, and Carlo M Croce. 2008. "MicroRNA Expression Profiling Using Microarrays." *Nature Protocols* 3 (4): 563–78. https://doi.org/10.1038/nprot.2008.14.
Liu, Philip T, Matthew Wheelwright, Rosane Teles, Evangelia Komisopoulou, Kristina Edfeldt, Benjamin Ferguson, Manali D Mehta, et al. 2012. "MicroRNA-21 Targets the Vitamin D-Dependent Antimicrobial Pathway in Leprosy." *Nature Medicine* 18 (2): 267–73. https://doi.org/10.1038/nm.2584.
Lorenz, Ronny, Stephan H Bernhart, Christian Höner Zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. 2011. "ViennaRNA Package 2.0." http://www.tbi.univie.ac.at/RNA.
Luciano, Daniel J, Henry Mirsky, Nicholas J Vendetti, and Stefan Maas. 2004. "RNA Editing of a MiRNA Precursor." *RNA (New York, N.Y.)* 10 (8): 1174–77. https://doi.org/10.1261/rna.7350304.
Lupas, A., M. Van Dyke, and J. Stock. 1991. "Predicting Coiled Coils from Protein Sequences." *Science* 252 (5009): 1162–64. https://doi.org/10.1126/science.252.5009.1162.
Lutay, Nataliya, Ines Ambite, Jenny Grönberg Hernandez, Gustav Rydström, Bryndís Ragnarsdóttir, Manoj Puthia, Aftab Nadeem, et al. 2013. "Bacterial Control of Host Gene Expression through RNA Polymerase II." *Journal of Clinical Investigation* 123 (6):

2366–79. https://doi.org/10.1172/JCI66451.
Mannala, Gopala K, Benjamin Izar, Oliver Rupp, Tilman Schultze, Alexander Goesmann, Trinad Chakraborty, and Torsten Hain. 2017. “Listeria Monocytogenes Induces a Virulence-Dependent MicroRNA Signature That Regulates the Immune Response in Galleria Mellonella.” *Frontiers in Microbiology* 8 (December): 2463. https://doi.org/10.3389/fmicb.2017.02463.
Marchler-Bauer, Aron, Yu Bo, Lianyi Han, Jane He, Christopher J. Lanczycki, Shennan Lu, Farideh Chitsaz, et al. 2017. “CDD/SPARCLE: Functional Classification of Proteins via Subfamily Domain Architectures.” *Nucleic Acids Research* 45 (D1): D200–203. https://doi.org/10.1093/nar/gkw1129.
Marco, A., J. H. L. Hui, M. Ronshaugen, and S. Griffiths-Jones. 2010. “Functional Shifts in Insect MicroRNA Evolution.” *Genome Biology and Evolution* 304 (September): 594–96. https://doi.org/10.1093/gbe/evq053.
Martin, Hilary C, Shivangi Wani, Anita L Steptoe, Keerthana Krishnan, Katia Nones, Ehsan Nourbakhsh, Alexander Vlassov, Sean M Grimmond, and Nicole Cloonan. 2014. “Imperfect Centered MiRNA Binding Sites Are Common and Can Mediate Repression of Target MRNAs.” *Genome Biology* 15 (3): R51. https://doi.org/10.1186/gb-2014-15-3-r51.
Martin, Marcel. 2011. “Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads.” *EMBnet.Journal* 17 (1): 10. https://doi.org/10.14806/ej.17.1.200.
Mathonnet, Géraldine, Marc R Fabian, Yuri V Svitkin, Armen Parsyan, Laurent Huck, Takayuki Murata, Stefano Biffo, et al. 2007. “MicroRNA Inhibition of Translation Initiation in Vitro by Targeting the Cap-Binding Complex EIF4F.” *Science (New York, N.Y.)* 317 (5845): 1764–67. https://doi.org/10.1126/science.1146067.
Meister, Gunter, Markus Landthaler, Agnieszka Patkaniowska, Yair Dorsett, Grace Teng, and Thomas Tuschl. 2004. “Human Argonaute2 Mediates RNA Cleavage Targeted by MiRNAs and SiRNAs.” *Molecular Cell* 15 (2): 185–97. https://doi.org/10.1016/j.molcel.2004.07.007.
Mi, Huaiyu, Xiaosong Huang, Anushya Muruganujan, Haiming Tang, Caitlin Mills, Diane Kang, and Paul D. Thomas. 2017. “PANTHER Version 11: Expanded Annotation Data from Gene Ontology and Reactome Pathways, and Data Analysis Tool Enhancements.” *Nucleic Acids Research* 45 (D1): D183–89. https://doi.org/10.1093/nar/gkw1138.
Miranda, Kevin C., Tien Huynh, Yvonne Tay, Yen-Sin Ang, Wai-Leong Tam, Andrew M. Thomson, Bing Lim, and Isidore Rigoutsos. 2006. “A Pattern-Based Method for the Identification of MicroRNA Binding Sites and Their Corresponding Heteroduplexes.” *Cell* 126 (6): 1203–17. https://doi.org/10.1016/J.CELL.2006.07.031.
Morin, Ryan D, Michael D O’Connor, Malachi Griffith, Florian Kuchenbauer, Allen Delaney, Anna-Liisa Prabhu, Yongjun Zhao, et al. 2008. “Application of Massively Parallel Sequencing to MicroRNA Profiling and Discovery in Human Embryonic Stem Cells.” *Genome Research* 18 (4): 610–21. https://doi.org/10.1101/gr.7179508.
Mukherjee, Krishnendu, Boran Altincicek, Torsten Hain, Eugen Domann, Andreas Vilcinskas, and Trinad Chakraborty. 2010. “Galleria Mellonella as a Model System for Studying Listeria Pathogenesis.” *Applied and Environmental Microbiology* 76 (1): 310–17. https://doi.org/10.1128/AEM.01301-09.
Mukherjee, Krishnendu, Ivan Dubovskiy, Ekaterina Grizanova, Rüdiger Lehmann, and Andreas Vilcinskas. 2019. “Epigenetic Mechanisms Mediate the Experimental Evolution of Resistance against Parasitic Fungi in the Greater Wax Moth Galleria Mellonella.” *Scientific Reports* 9 (1): 1626. https://doi.org/10.1038/s41598-018-36829-8.

Mukherjee, Krishnendu, Rainer Fischer, and Andreas Vilcinskas. 2012. “Histone Acetylation Mediates Epigenetic Regulation of Transcriptional Reprogramming in Insects during Metamorphosis, Wounding and Infection.” *Frontiers in Zoology* 9 (1): 25. https://doi.org/10.1186/1742-9994-9-25.

Mukherjee, Krishnendu, Ekaterina Grizanova, Ekaterina Chertkova, Ruediger Lehmann, Ivan Dubovskiy, and Andreas Vilcinskas. 2017. “Experimental Evolution of Resistance against Bacillus Thuringiensis in the Insect Model Host Galleria Mellonella Results in Epigenetic Modifications.” *Virulence* 8 (8): 1618–30. https://doi.org/10.1080/21505594.2017.1325975.

Mukherjee, Krishnendu, Torsten Hain, Rainer Fischer, Trinad Chakraborty, and Andreas Vilcinskas. 2013. “Brain Infection and Activation of Neuronal Repair Mechanisms by the Human Pathogen Listeria Monocytogenes in the Lepidopteran Model Host Galleria Mellonella.” *Virulence* 4 (4): 324–32. https://doi.org/10.4161/viru.23629.

Mukherjee, Krishnendu, Richard M Twyman, and Andreas Vilcinskas. 2015. “Insects as Models to Study the Epigenetic Basis of Disease.” *Progress in Biophysics and Molecular Biology* 118 (1–2): 69–78. https://doi.org/10.1016/j.pbiomolbio.2015.02.009.

Mukherjee, Krishnendu, and Andreas Vilcinskas. 2014. “Development and Immunity-Related MicroRNAs of the Lepidopteran Model Host Galleria Mellonella.” *BMC Genomics* 15 (1): 705. https://doi.org/10.1186/1471-2164-15-705.

Muller, Heiko, Matteo Jacopo Marzi, and Francesco Nicassio. 2014. “IsomiRage: From Functional Classification to Differential Expression of MiRNA Isoforms.” *Frontiers in Bioengineering and Biotechnology* 2 (September): 38. https://doi.org/10.3389/fbioe.2014.00038.

Needleman, Saul B., and Christian D. Wunsch. 1970. “A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins.” *Journal of Molecular Biology* 48 (3): 443–53. https://doi.org/10.1016/0022-2836(70)90057-4.

Nielsen, Henrik. 2017. “Predicting Secretory Proteins with SignalP.” In , 59–73. Humana Press, New York, NY. https://doi.org/10.1007/978-1-4939-7015-5_6.

Nielsen, Ross A, and C D Brister. 1979. “Greater Wax Moth: Behavior of Larvae.” *Annals of the Entomological Society of America* 72 (6): 811–15. https://doi.org/10.1093/aesa/72.6.811.

Ninova, Maria, Matthew Ronshaugen, and Sam Griffiths-Jones. 2015. “MicroRNA Evolution, Expression and Function during Short Germband Development in Tribolium Castaneum.” *Genome Research*, October, gr.193367.115-. https://doi.org/10.1101/gr.193367.115.

———. 2016. “MicroRNA Evolution, Expression, and Function during Short Germband Development in Tribolium Castaneum.” *Genome Research* 26 (1): 85–96. https://doi.org/10.1101/gr.193367.115.

Nishikura, Kazuko. 2016. “A-to-I Editing of Coding and Non-Coding RNAs by ADARs.” *Nature Reviews Molecular Cell Biology*. Nature Publishing Group. https://doi.org/10.1038/nrm.2015.4.

Niu, Yuna, Delin Mo, Limei Qin, Chong Wang, Anning Li, Xiao Zhao, Xiaoying Wang, et al. 2011. “Lipopolysaccharide-Induced MiR-1224 Negatively Regulates Tumour Necrosis Factor-α Gene Expression by Modulating Sp1.” *Immunology* 133 (1): 8–20. https://doi.org/10.1111/j.1365-2567.2010.03374.x.

Nottrott, Stephanie, Martin J Simard, and Joel D Richter. 2006. “Human Let-7a MiRNA Blocks Protein Production on Actively Translating Polyribosomes.” *Nature Structural & Molecular Biology* 13 (12): 1108–14. https://doi.org/10.1038/nsmb1173.

O'Brien, Jacob, Heyam Hayder, Yara Zayed, and Chun Peng. 2018. "Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation." *Frontiers in Endocrinology*. Frontiers Media S.A. https://doi.org/10.3389/fendo.2018.00402.

Okamura, Katsutomo, Michael D Phillips, David M Tyler, Hong Duan, Yu-ting Chou, and Eric C Lai. 2008. "The Regulatory Activity of MicroRNA* Species Has Substantial Influence on MicroRNA and 3' UTR Evolution." *Nature Structural & Molecular Biology* 15 (4): 354–63. https://doi.org/10.1038/nsmb.1409.

Oliveira, Arthur C, Luiz A Bovolenta, Pedro G Nachtigall, Marcos E Herkenhoff, Ney Lemke, and Danillo Pinhal. 2017. "Combining Results from Distinct MicroRNA Target Prediction Tools Enhances the Performance of Analyses." *Frontiers in Genetics* 8: 59. https://doi.org/10.3389/fgene.2017.00059.

Oliveira, Luiz Felipe Valter de, Ana Paula Christoff, and Rogerio Margis. 2013. "IsomiRID: A Framework to Identify MicroRNA Isoforms." *Bioinformatics (Oxford, England)* 29 (20): 2521–23. https://doi.org/10.1093/bioinformatics/btt424.

Ozsolak, Fatih, Laura L Poling, Zhengxin Wang, Hui Liu, X Shirley Liu, Robert G Roeder, Xinmin Zhang, Jun S Song, and David E Fisher. 2008. "Chromatin Structure Analyses Identify MiRNA Promoters." *Genes & Development* 22 (22): 3172–83.

Paddock, Floyd B. 1918. *The Beemoth or Waxworm*. Texas Agricultural Experiment Stations.

Pantano, Lorena, Xavier Estivill, and Eulàlia Martí. 2010. "SeqBuster, a Bioinformatic Tool for the Processing and Analysis of Small RNAs Datasets, Reveals Ubiquitous MiRNA Modifications in Human Embryonic Cells." *Nucleic Acids Research* 38 (5): e34. https://doi.org/10.1093/nar/gkp1127.

Pedruzzi, Ivo, Catherine Rivoire, Andrea H. Auchincloss, Elisabeth Coudert, Guillaume Keller, Edouard de Castro, Delphine Baratin, et al. 2015. "HAMAP in 2015: Updates to the Protein Family Classification and Annotation System." *Nucleic Acids Research* 43 (D1): D1064–70. https://doi.org/10.1093/nar/gku1002.

Piovesan, Damiano, Francesco Tabaro, Lisanna Paladin, Marco Necci, Ivan Mičetić, Carlo Camilloni, Norman Davey, et al. 2018. "MobiDB 3.0: More Annotations for Intrinsic Disorder, Conformational Diversity and Interactions in Proteins." *Nucleic Acids Research* 46 (D1): D471–76. https://doi.org/10.1093/nar/gkx1071.

Pöppel, Anne-Kathrin, Heiko Vogel, Jochen Wiesner, and Andreas Vilcinskas. 2015. "Antimicrobial Peptides Expressed in Medicinal Maggots of the Blow Fly Lucilia Sericata Show Combinatorial Activity against Bacteria." *Antimicrobial Agents and Chemotherapy* 59 (5): 2508–14. https://doi.org/10.1128/AAC.05180-14.

Quast, C., E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glockner. 2013. "The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-Based Tools." *Nucleic Acids Research* 41 (D1): D590–96. https://doi.org/10.1093/nar/gks1219.

Quinlan, Aaron R., and Ira M. Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." *Bioinformatics* 26 (6): 841–42. https://doi.org/10.1093/bioinformatics/btq033.

Raden, Martin, Syed M Ali, Omer S Alkhnbashi, Anke Busch, Fabrizio Costa, Jason A Davis, Florian Eggenhofer, et al. 2018. "Freiburg RNA Tools: A Central Online Resource for RNA-Focused Research and Teaching." *Nucleic Acids Research*, no. 1. https://doi.org/10.1093/nar/gky329.

Ragnarsdóttir, Bryndís, Martin Samuelsson, Mattias C. U. Gustafsson, Irene Leijonhufvud, Diana Karpman, and Catharina Svanborg. 2007. "Reduced Toll-Like Receptor 4 Expression in Children with Asymptomatic Bacteriuria." *The Journal of Infectious*

Diseases 196 (3): 475–84. https://doi.org/10.1086/518893.
Rajamuthiah, Rajmohan, Elamparithi Jayamani, Annie L. Conery, Beth Burgwyn Fuchs, Wooseong Kim, Tatiana Johnston, Andreas Vilcinskas, Frederick M. Ausubel, and Eleftherios Mylonakis. 2015. "A Defensin from the Model Beetle Tribolium Castaneum Acts Synergistically with Telavancin and Daptomycin against Multidrug Resistant Staphylococcus Aureus." Edited by Surajit Bhattacharjya. *PLOS ONE* 10 (6): e0128576. https://doi.org/10.1371/journal.pone.0128576.
Rehmsmeier, Marc, Peter Steffen, Matthias Hochsmann, and Robert Giegerich. 2004. "Fast and Effective Prediction of MicroRNA/Target Duplexes." *RNA (New York, N.Y.)* 10 (10): 1507–17. https://doi.org/10.1261/rna.5248604.
Reinhart, Brenda J., Frank J. Slack, Michael Basson, Amy E. Pasquinelli, Jill C. Bettinger, Ann E. Rougvie, H. Robert Horvitz, and Gary Ruvkun. 2000. "The 21-Nucleotide Let-7 RNA Regulates Developmental Timing in Caenorhabditis Elegans." *Nature* 403 (6772): 901–6. https://doi.org/10.1038/35002607.
Remm, Maido, Christian E.V. Storm, and Erik L.L. Sonnhammer. 2001. "Automatic Clustering of Orthologs and In-Paralogs from Pairwise Species Comparisons." *Journal of Molecular Biology* 314 (5): 1041–52. https://doi.org/10.1006/jmbi.2000.5197.
Rice, Peter, Lan Longden, and Alan Bleasby. 2000. "EMBOSS: The European Molecular Biology Open Software Suite." *Trends in Genetics*. https://doi.org/10.1016/S0168-9525(00)02024-2.
Richards, Stephen, Richard A Gibbs, George M Weinstock, Susan J Brown, Robin Denell, Richard W Beeman, Richard Gibbs, et al. 2008. "The Genome of the Model Beetle and Pest Tribolium Castaneum." *Nature* 452 (7190): 949–55. https://doi.org/10.1038/nature06784.
Rivera, M. C., R. Jain, J. E. Moore, and J. A. Lake. 1998. "Genomic Evidence for Two Functionally Distinct Gene Classes." *Proceedings of the National Academy of Sciences* 95 (11): 6239–44. https://doi.org/10.1073/pnas.95.11.6239.
Ro, Seungil, Chanjae Park, David Young, Kenton M Sanders, and Wei Yan. 2007. "Tissue-Dependent Paired Expression of MiRNAs." *Nucleic Acids Research* 35 (17): 5944–53. https://doi.org/10.1093/nar/gkm641.
Ruby, J Graham, Calvin H Jan, and David P Bartel. 2007. "Intronic MicroRNA Precursors That Bypass Drosha Processing." *Nature* 448 (7149): 83–86. https://doi.org/10.1038/nature05983.
Russo, Thomas A, and James R Johnson. 2003. "Medical and Economic Impact of Extraintestinal Infections Due to Escherichia Coli: Focus on an Increasingly Important Endemic Problem." *Microbes and Infection* 5 (5): 449–56. http://www.ncbi.nlm.nih.gov/pubmed/12738001.
Ruvkun, Gary, Brenda J. Reinhart, Frank J. Slack, Michael Basson, Amy E. Pasquinelli, Jill C. Bettinger, Ann E. Rougvie, and H. Robert Horvitz. 2000. "The 21-Nucleotide Let-7 RNA Regulates Developmental Timing in Caenorhabditis Elegans." *Nature* 403 (6772): 901–6. https://doi.org/10.1038/35002607.
Sablok, Gaurav, Ivan Milev, Georgi Minkov, Ivan Minkov, Claudio Varotto, Galina Yahubyan, and Vesselin Baev. 2013. "IsomiRex: Web-Based Identification of MicroRNAs, IsomiR Variations and Differential Expression Using next-Generation Sequencing Datasets." *FEBS Letters* 587 (16): 2629–34. https://doi.org/10.1016/j.febslet.2013.06.047.
Schnall-Levin, Michael, Yong Zhao, Norbert Perrimon, and Bonnie Berger. 2010. "Conserved MicroRNA Targeting in Drosophila Is as Widespread in Coding Regions as in 3'UTRs." *Proceedings of the National Academy of Sciences of the United States of America* 107

(36): 15751–56. https://doi.org/10.1073/pnas.1006172107.
Schwarz, Dianne S, György Hutvágner, Tingting Du, Zuoshang Xu, Neil Aronin, and Phillip D Zamore. 2003. “Asymmetry in the Assembly of the RNAi Enzyme Complex.” *Cell* 115 (2): 199–208. http://www.ncbi.nlm.nih.gov/pubmed/14567917.
Shames, Stephanie R, Amit P Bhavsar, Matthew A Croxen, Robyn J Law, Stefanie H C Mak, Wanyin Deng, Yuling Li, et al. 2011. “The Pathogenic Escherichia Coli Type III Secreted Protease NleC Degrades the Host Acetyltransferase P300.” *Cellular Microbiology* 13 (10): 1542–57. https://doi.org/10.1111/j.1462-5822.2011.01640.x.
Shi, Weiyang, David Hendrix, Mike Levine, and Benjamin Haley. 2009. “A Distinct Class of Small RNAs Arises from Pre-MiRNA-Proximal Regions in a Simple Chordate.” *Nature Structural & Molecular Biology* 16 (2): 183–89. https://doi.org/10.1038/nsmb.1536.
Shin, Chanseok, Jin-Wu Nam, Kyle Kai-How Farh, H Rosaria Chiang, Alena Shkumatava, and David P Bartel. 2010. “Expanding the MicroRNA Targeting Code: Functional Sites with Centered Pairing.” *Molecular Cell* 38 (6): 789–802. https://doi.org/10.1016/j.molcel.2010.06.005.
Smith, E H, and R C Whitman. 1992. *NPCA Field Guide to Structural Pests*. NPMA.
Smith, T.F., and M.S. Waterman. 1981. “Identification of Common Molecular Subsequences.” *Journal of Molecular Biology* 147 (1): 195–97. https://doi.org/10.1016/0022-2836(81)90087-5.
Sonnhammer, Erik L.L., and Gabriel Östlund. 2015. “InParanoid 8: Orthology Analysis between 273 Proteomes, Mostly Eukaryotic.” *Nucleic Acids Research* 43 (D1): D234–39. https://doi.org/10.1093/nar/gku1203.
Srinivasan, Sreedevi, Susan E Leeman, and Salomon Amar. 2010. “Beneficial Dysregulation of the Time Course of Inflammatory Mediators in Lipopolysaccharide-Induced Tumor Necrosis Factor Alpha Factor-Deficient Mice.” *Clinical and Vaccine Immunology : CVI* 17 (5): 699–704. https://doi.org/10.1128/CVI.00510-09.
Stern-Ginossar, Noam, Naama Elefant, Albert Zimmermann, Dana G Wolf, Nivin Saleh, Moshe Biton, Elad Horwitz, et al. 2007. “Host Immune System Gene Targeting by a Viral MiRNA.” *Science* 317 (5836): 376–81.
Sticht, Carsten, Carolina De La Torre, Alisha Parveen, and Norbert Gretz. 2018. “MiRWalk: An Online Resource for Prediction of MicroRNA Binding Sites.” Edited by Moray Campbell. *PLOS ONE* 13 (10): e0206239. https://doi.org/10.1371/journal.pone.0206239.
Sun, Guihua, Jin Yan, Katie Noltner, Jinong Feng, Haitang Li, Daniel A Sarkis, Steve S Sommer, and John J Rossi. 2009. “SNPs in Human MiRNA Genes Affect Biogenesis and Function.” *RNA (New York, N.Y.)* 15 (9): 1640–51. https://doi.org/10.1261/rna.1560209.
Sun, Zhifu, Jared Evans, Aditya Bhagwate, Sumit Middha, Matthew Bockol, Huihuang Yan, and Jean Pierre Kocher. 2014. “CAP-MiRSeq: A Comprehensive Analysis Pipeline for MicroRNA Sequencing Data.” *BMC Genomics* 15 (1): 423. https://doi.org/10.1186/1471-2164-15-423.
Tam, S., M.-S. Tsao, and J. D. McPherson. 2015. “Optimization of MiRNA-Seq Data Preprocessing.” *Briefings in Bioinformatics* 16 (6): 950–63. https://doi.org/10.1093/bib/bbv019.
Tan, Geok Chin, Elcie Chan, Attila Molnar, Rupa Sarkar, Diana Alexieva, Ihsan Mad Isa, Sophie Robinson, et al. 2014a. “5’ IsomiR Variation Is of Functional and Evolutionary Importance.” *Nucleic Acids Research* 42 (14): 9424–35. https://doi.org/10.1093/nar/gku656.
———. 2014b. “5’ IsomiR Variation Is of Functional and Evolutionary Importance.” *Nucleic Acids Research* 42 (14): 9424–35. https://doi.org/10.1093/nar/gku656.

Thadani, Rahul, and Martti T Tammi. 2006. “MicroTar: Predicting MicroRNA Targets from RNA Duplexes.” *BMC Bioinformatics* 7 (Suppl 5): S20. https://doi.org/10.1186/1471-2105-7-S5-S20.

Thermann, Rolf, and Matthias W. Hentze. 2007. “Drosophila MiR2 Induces Pseudo-Polysomes and Inhibits Translation Initiation.” *Nature* 447 (7146): 875–78. https://doi.org/10.1038/nature05878.

Trapnell, Cole, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. 2010. “Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Transcripts and Isoform Switching during Cell Differentiation.” *Nature Biotechnology* 28 (5): 511–15. https://doi.org/10.1038/nbt.1621.

Ule, Jernej, Kirk Jensen, Aldo Mele, and Robert B. Darnell. 2005. “CLIP: A Method for Identifying Protein–RNA Interaction Sites in Living Cells.” *Methods* 37 (4): 376–86. https://doi.org/10.1016/j.ymeth.2005.07.018.

UniProt Consortium, The. 2018. “UniProt: The Universal Protein Knowledgebase.” *Nucleic Acids Research* 46 (5): 2699–2699. https://doi.org/10.1093/nar/gky092.

Uren, Philip J, Emad Bahrami-Samani, Suzanne C Burns, Mei Qiao, Fedor V Karginov, Emily Hodges, Gregory J Hannon, Jeremy R Sanford, Luiz O F Penalva, and Andrew D Smith. 2012. “Site Identification in High-Throughput RNA-Protein Interaction Data.” *Bioinformatics (Oxford, England)* 28 (23): 3013–20. https://doi.org/10.1093/bioinformatics/bts569.

Urgese, Gianvito, Giulia Paciello, Andrea Acquaviva, Elisa Ficarra, DP Bartel, DP Bartel, H Dong, et al. 2016. “IsomiR-SEA: An RNA-Seq Analysis Tool for MiRNAs/IsomiRs Expression Level Profiling and MiRNA-MRNA Interaction Sites Evaluation.” *BMC Bioinformatics* 17 (1): 148. https://doi.org/10.1186/s12859-016-0958-0.

Vilcinskas, Andreas. 2011. “Insects Emerge as Valuable Model Hosts to Explore Virulence.” *Virulence* 2 (5): 376–78. https://doi.org/10.4161/viru.2.5.18289.

———. 2016. “The Role of Epigenetics in Host-Parasite Coevolution: Lessons from the Model Host Insects Galleria Mellonella and Tribolium Castaneum.” *Zoology (Jena, Germany)* 119 (4): 273–80. https://doi.org/10.1016/j.zool.2016.05.004.

Vilcinskas, Andreas, and Jürgen Gross. 2005. “Drugs from Bugs: The Use of Insects as a Valuable Source of Transgenes with Potential in Modern Plant Protection Strategies.” *Journal of Pest Science* 78 (4): 187–91. https://doi.org/10.1007/s10340-005-0114-5.

Villar, Diego, Paul Flicek, and Duncan T. Odom. 2014. “Evolution of Transcription Factor Binding in Metazoans-Mechanisms and Functional Implications.” *Nature Reviews Genetics*. Nature Publishing Group. https://doi.org/10.1038/nrg3481.

Vogel, Heiko, Boran Altincicek, Gernot Glöckner, and Andreas Vilcinskas. 2011. “A Comprehensive Transcriptome and Immune-Gene Repertoire of the Lepidopteran Model Host Galleria Mellonella.” *BMC Genomics* 12 (June): 308. https://doi.org/10.1186/1471-2164-12-308.

Waldhuber, Anna, Manoj Puthia, Andreas Wieser, Christine Cirl, Susanne Dürr, Silke Neumann-Pfeifer, Simone Albrecht, et al. 2016. “Uropathogenic Escherichia Coli Strain CFT073 Disrupts NLRP3 Inflammasome Activation.” *The Journal of Clinical Investigation* 126 (7): 2425–36. https://doi.org/10.1172/JCI81916.

Weber, James L, Donna David, Jeremy Heil, Ying Fan, Chengfeng Zhao, and Gabor Marth. 2002. “Human Diallelic Insertion/Deletion Polymorphisms.” *American Journal of Human Genetics* 71 (4): 854–62. https://doi.org/10.1086/342727.

Wen, Jiayu, Erik Ladewig, Sol Shenker, Jaaved Mohammed, and Eric C. Lai. 2015. “Analysis of

Nearly One Thousand Mammalian Mirtrons Reveals Novel Features of Dicer Substrates." Edited by Isidore Rigoutsos. *PLOS Computational Biology* 11 (9): e1004441. https://doi.org/10.1371/journal.pcbi.1004441.

Williamson, Deborah A, Grant Mills, James R Johnson, Stephen Porter, and Siouxsie Wiles. 2014. "In Vivo Correlates of Molecularly Inferred Virulence among Extraintestinal Pathogenic Escherichia Coli (ExPEC) in the Wax Moth Galleria Mellonella Model System." *Virulence* 5 (3): 388–93. https://doi.org/10.4161/viru.27912.

Wilson, Derek, Ralph Pethica, Yiduo Zhou, Charles Talbot, Christine Vogel, Martin Madera, Cyrus Chothia, and Julian Gough. 2009. "SUPERFAMILY—Sophisticated Comparative Genomics, Data Mining, Visualization and Phylogeny." *Nucleic Acids Research* 37 (suppl_1): D380–86. https://doi.org/10.1093/nar/gkn762.

Wong, Nathan, and Xiaowei Wang. 2015. "MiRDB: An Online Resource for MicroRNA Target Prediction and Functional Annotations." *Nucleic Acids Research* 43 (D1): D146–52. https://doi.org/10.1093/nar/gku1104.

Wu, C. H., Anastasia Nikolskaya, Hongzhan Huang, Lai-Su L Yeh, Darren A Natale, C R Vinayaka, Zhang-Zhi Hu, et al. 2004. "PIRSF: Family Classification System at the Protein Information Resource." *Nucleic Acids Research* 32 (90001): 112D – 114. https://doi.org/10.1093/nar/gkh097.

Wu, T. D., and S. Nacu. 2010. "Fast and SNP-Tolerant Detection of Complex Variants and Splicing in Short Reads." *Bioinformatics* 26 (7): 873–81. https://doi.org/10.1093/bioinformatics/btq057.

Wu, Wei, Wenfeng Xiong, Chengjun Li, Mengfan Zhai, Yao Li, Fei Ma, and Bin Li. 2017. "MicroRNA-Dependent Regulation of Metamorphosis and Identification of MicroRNAs in the Red Flour Beetle, Tribolium Castaneum." *Genomics* 109 (5–6): 362–73. https://doi.org/10.1016/j.ygeno.2017.06.001.

Wuchty, Stefan, Walter Fontana, Ivo L. Hofacker, and Peter Schuster. 1999. "Complete Suboptimal Folding of RNA and the Stability of Secondary Structures." *Biopolymers* 49 (2): 145–65. https://doi.org/10.1002/(SICI)1097-0282(199902)49:2<145::AID-BIP4>3.0.CO;2-G.

Wullt, Björn, and Catharina Svanborg. 2016. "Deliberate Establishment of Asymptomatic Bacteriuria-A Novel Strategy to Prevent Recurrent UTI." *Pathogens (Basel, Switzerland)* 5 (3): 52. https://doi.org/10.3390/pathogens5030052.

Wyman, Stacia K., Rachael K. Parkin, Patrick S. Mitchell, Brian R. Fritz, Kathy O'Briant, Andrew K. Godwin, Nicole Urban, Charles W. Drescher, Beatrice S. Knudsen, and Muneesh Tewari. 2009. "Repertoire of MicroRNAs in Epithelial Ovarian Cancer as Determined by Next Generation Sequencing of Small RNA CDNA Libraries." Edited by Sudhansu Kumar Dey. *PLoS ONE* 4 (4): e5311. https://doi.org/10.1371/journal.pone.0005311.

Wyman, Stacia K, Emily C Knouf, Rachael K Parkin, Brian R Fritz, Daniel W Lin, Lucas M Dennis, Michael A Krouse, Philippa J Webster, and Muneesh Tewari. 2011. "Post-Transcriptional Generation of MiRNA Variants by Multiple Nucleotidyl Transferases Contributes to MiRNA Transcriptome Complexity." *Genome Research* 21 (9): 1450–61. https://doi.org/10.1101/gr.118059.110.

Xing, Tongjing, Jiansheng Zhu, Jianchun Xian, Ali Li, Xuequan Wang, Wei Wang, and Qian Zhang. 2019. "MiRNA-548ah Promotes the Replication and Expression of Hepatitis B Virus by Targeting Histone Deacetylase 4." *Life Sciences* 219 (December 2018): 199–208. https://doi.org/10.1016/j.lfs.2018.12.057.

Yamakuchi, Munekazu, Marcella Ferlito, and Charles J Lowenstein. 2008. "MiR-34a

Repression of SIRT1 Regulates Apoptosis." *Proceedings of the National Academy of Sciences of the United States of America* 105 (36): 13421–26. https://doi.org/10.1073/pnas.0801613105.

Yekta, S., I-Hung Shih, and David P Bartel. 2004. "MicroRNA-Directed Cleavage of HOXB8 MRNA." *Science* 304 (5670): 594–96. https://doi.org/10.1126/science.1097434.

Zdziarski, J., C. Svanborg, B. Wullt, J. Hacker, and U. Dobrindt. 2008. "Molecular Basis of Commensalism in the Urinary Tract: Low Virulence or Virulence Attenuation?" *Infection and Immunity* 76 (2): 695–703. https://doi.org/10.1128/IAI.01215-07.

Zdziarski, Jaroslaw, Elzbieta Brzuszkiewicz, Björn Wullt, Heiko Liesegang, Dvora Biran, Birgit Voigt, Jenny Grönberg-Hernandez, et al. 2010. "Host Imprints on Bacterial Genomes--Rapid, Divergent Evolution in Individual Patients." Edited by David S. Guttman. *PLoS Pathogens* 6 (8): e1001078. https://doi.org/10.1371/journal.ppat.1001078.

Zhang, Hua, Jian-Hua Yang, Yu-Sheng Zheng, Peng Zhang, Xiao Chen, Jun Wu, Ling Xu, et al. 2009. "Genome-Wide Analysis of Small RNA and Novel MicroRNA Discovery in Human Acute Lymphoblastic Leukemia Based on Extensive Sequencing Approach." Edited by Baohong Zhang. *PLoS ONE* 4 (9): e6849. https://doi.org/10.1371/journal.pone.0006849.

Zhang, Xiufeng, Emre Aksoy, Thomas Girke, Alexander S Raikhel, and Fedor V Karginov. 2017. "Transcriptome-Wide MicroRNA and Target Dynamics in the Fat Body during the Gonadotrophic Cycle of Aedes Aegypti." *Proceedings of the National Academy of Sciences of the United States of America* 114 (10): E1895–1903. https://doi.org/10.1073/pnas.1701474114.

Zhang, Yu Liang, Qi Xing Huang, Guo Hua Yin, Samantha Lee, Rui Zong Jia, Zhi Xin Liu, Nai Tong Yu, Kayla K. Pennerman, Xin Chen, and An Ping Guo. 2015. "Identification of MicroRNAs by Small RNA Deep Sequencing for Synthetic MicroRNA Mimics to Control Spodoptera Exigua." *Gene* 557 (2): 215–21. https://doi.org/10.1016/j.gene.2014.12.038.

Zhang, Yuanwei, Qiguang Zang, Huan Zhang, Rongjun Ban, Yifan Yang, Furhan Iqbal, Ao Li, and Qinghua Shi. 2016. "DeAnnIso: A Tool for Online Detection and Annotation of IsomiRs from Small RNA Sequencing Data." *Nucleic Acids Research* 44 (W1): W166–75. https://doi.org/10.1093/nar/gkw427.

Zhao, Shanrong, William Gordon, Sarah Du, Chi Zhang, Wen He, Li Xi, Sachin Mathur, et al. 2017. "QuickMIRSeq: A Pipeline for Quick and Accurate Quantification of Both Known MiRNAs and IsomiRs by Jointly Processing Multiple Samples from MicroRNA Sequencing." *BMC Bioinformatics* 18 (1): 180. https://doi.org/10.1186/s12859-017-1601-4.

Ziemann, Mark, Antony Kaspi, and Assam El-Osta. 2016. "Evaluation of MicroRNA Alignment Techniques." *RNA (New York, N.Y.)* 22 (8): 1120–38. https://doi.org/10.1261/rna.055509.115.

11

DECLARATION OF INDEPENDENCE

I declare that I have completed this dissertation single-handedly without the unauthorized help of a second party and only with the assistance acknowledged therein. I have appropriately acknowledged and cited all text passages that are derived verbatim from or are based on the content of published work of others, and all information relating to verbal communications. I consent to the use of an anti-plagiarism software to check my thesis. I have abided by the principles of good scientific conduct laid down in the charter of the Justus Liebig University Giessen „Satzung der Justus-Liebig-Universität Gießen zur Sicherung guter wissenschaftlicher Praxis“ in carrying out the investigations described in the dissertation.

Gießen, October, 5th 2020

Daniel Amsel

12

ACKNOWLEDGEMENT

First and foremost, I would like to thank Dow AgroSciences and the Hessen State Ministry of Higher Education, Research and the Arts (HMWK) via the LOEWE Center for Insect Biotechnology and Bioresources for funding.

I would also like to thank Prof. Dr. Andreas Vilcinskas for funding, for providing me the possibility to work on my thesis and for the good computational infrastructure. I also would like to thank Prof. Dr. Alexander Goesmann for the supervision of my work and the possibility to use his computational infrastructure.

Furthermore, I would like to express my gratitude to Dr. André Billion and Dr. Frank Förster for their mentoring of my projects.

My thanks also go out to my former colleagues Heiko Herrmann, Roman Szimanski, Niklas Pfeiffer, Christoph Maxeiner, Dr. Philipp Heise, Miriam Kalsy, Ricarda Döring and Dr. Ina Schüttmann for inspirational conversations and an enjoyable worktime.

I additionally thank my new colleagues in the Neuropathology, Carmen Selignow, Nadja Ritschel, Dr. Hildegard Dohmen, Kai Schmid and Jannik Sehring for motivating words in difficult times, but especially Prof. Dr. Till Acker for his support and understanding.

Finally, I thank my parents Silvia and Guido and my grandparents Maria and Hans for their love and support throughout my life, especially when the situation was difficult.

Last, I thank my wife Annika for her love, great support and sympathy, when things were not working as expected and hurdles seemed insurmountable.

13

SUPPLEMENTAL MATERIAL

13.1 SAMtools

SAMtools is a compilation of tools to manipulate the standard .sam (Sequence Alignment/Map) format from sequence alignments (H. Li, Handsaker, Wysoker, Fennell, Ruan, Homer, Marth, Abecasis, Durbin, et al. 2009) (Table S1).

Table S1 **SAM format** Taken from (H. Li, Handsaker, Wysoker, Fennell, Ruan, Homer, Marth, Abecasis, Durbin, et al. 2009)

No.	Name	Description
1	QNAME	Query NAME of the read or the read pair
2	FLAG	Bitwise FLAG (pairing, strand, mate strand, etc.)
3	RNAME	Reference sequence NAME
4	POS	1-Based leftmost POSition of clipped alignment
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIGAR	Extended CIGAR string (operations: MIDNSHP)
7	MRNM	Mate Reference NaMe ('=' if same as RNAME)
8	MPOS	1-Based leftmost Mate POSition
9	ISIZE	Inferred Insert SIZE
10	SEQ	Query SEQuence on the same strand as the reference
11	QUAL	Query QUALity (ASCII-33=Phred base quality)

Within the microPIECE pipeline, SAMtools sort, view, index.were used in the following manner: samtools view -F 4; samtools view -b -f 4; samtools view -Sb; samtools sort; samtools index. The view -F 4 excludes

unmapped reads, whereas `-f 4` only reports the unmapped reads. The parameter `-b` forces the output in `.bam` format, the binary version of `.sam`.

With the `sort` option, the algorithm sorts the alignments according to the leftmost coordinate. By using `index`, the `.bam` file is indexed with the UCSC binning scheme (Kent et al. 2002) and a linear index for a fast random access.

13.2 BEDtools

The toolset BEDtools covers a broad range of dedicated algorithms for genomic feature determination from `.bed` (`Browser Extensible Data`) files (Table S2).

Table S2 **BED file format** Adapted from https://genome.ucsc.edu/FAQ/FAQformat.html

No.	Name	Description
1	`chrom`	name of chromosome
2	`chromStart`	start position of feature
3	`chromEnd`	stop position of feature
4	`name`	name of BED line
5	`score`	score 1-1000
6	`strand`	. (no strand) or + or -
7	`thickStart`	start position of thick drawn
8	`thickEnd`	stop position of thick drawn
9	`itemRgb`	RGB value
10	`blockCount`	number of exons in BED line
11	`blockSizes`	comma list of block sizes
12	`blockStarts`	comma list of block starts

Within the `microPIECE` pipeline, `merge`, `getfasta` and `bamtofastq` are used. The `merge` tool combines overlapping or nearby sequences into one. The `getfasta` command extracts a `.fasta` sequence from a `.bed` file. With `bamtofastq`, a `BAM` record is converted into a `.fastq` file.

13.3 Minimizing The User-Provided Data: gffread

For the complex tasks, covered by the `microPIECE` pipeline, several reference datasets are needed. So for example, one needs a file with protein sequences and one with the transcripts,

the annotation file in .gff format and the reference genome. Downloading references datasets that match to each other can sometimes be difficult, depending on the species. In any case it is an avoidable task for the user. The microPIECE pipeline therefore creates the needed data out of a minimal set of user-provided reference data. For this task, gffread is used.

The gffread program is part of the Cufflinks (Trapnell et al. 2010) package and is used to create protein sequences (-y) and exon transcript sequences (-w) from the genome and the .gff data files in the microPIECE pipeline. In the context of the pipeline, this tool comes handy, because the user only needs to provide a .gff file together with the the genome and the pipeline extracts the information it needs. Otherwise the user would need to upload the protein- and the exon-based transcript-file additionally.

13.4 MySQL Database

The exemplarily used database is based upon the MySQL open-source relational database management system (RDBMS). A relational database has a database model that relies on tables (relations), holding the data information. Each row of a table is called a tuple, holding a dataset. Each position in the tuple equates to a column in the table. This is called an attribute. An attribute has a certain datatype and maximum size. The connection between tables is done by keys. Those keys are called primary keys, if they uniquely identify a specific tuple in a table. It is also sometimes useful to have tables that consist of primary keys from other tables. Those keys are then called foreign keys in that table. Each ID can have different relations to IDs. As for example, one ID from a table could be uniquely assigned to another ID from another table. This would be a 1-to-1 relation. Other relations would be a 1-to-many (1-to-N or N-to-1), where one ID from a table can have more than one relation to IDs from another table. It is also possible to have a many-to-many relationship (N-to-M). This would be a similar case like the example before, but with the exception that the IDs from the other table could also be assigned to many IDs from the first able.

13.5 Supplemental Material: Evaluation of high-throughput isomiR identification tools: illuminating the early isomiRome of *Tribolium castaneum*

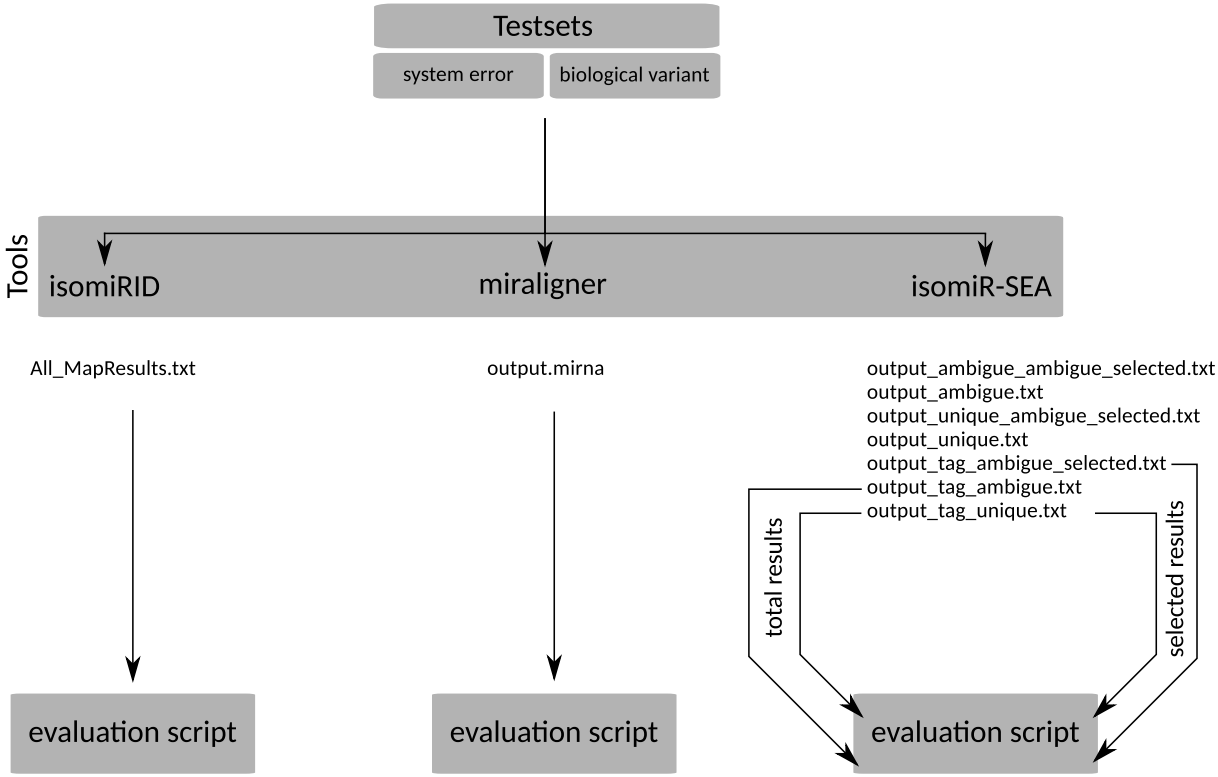


Figure S 1 **Analysis scheme for artificial test set evaluation**

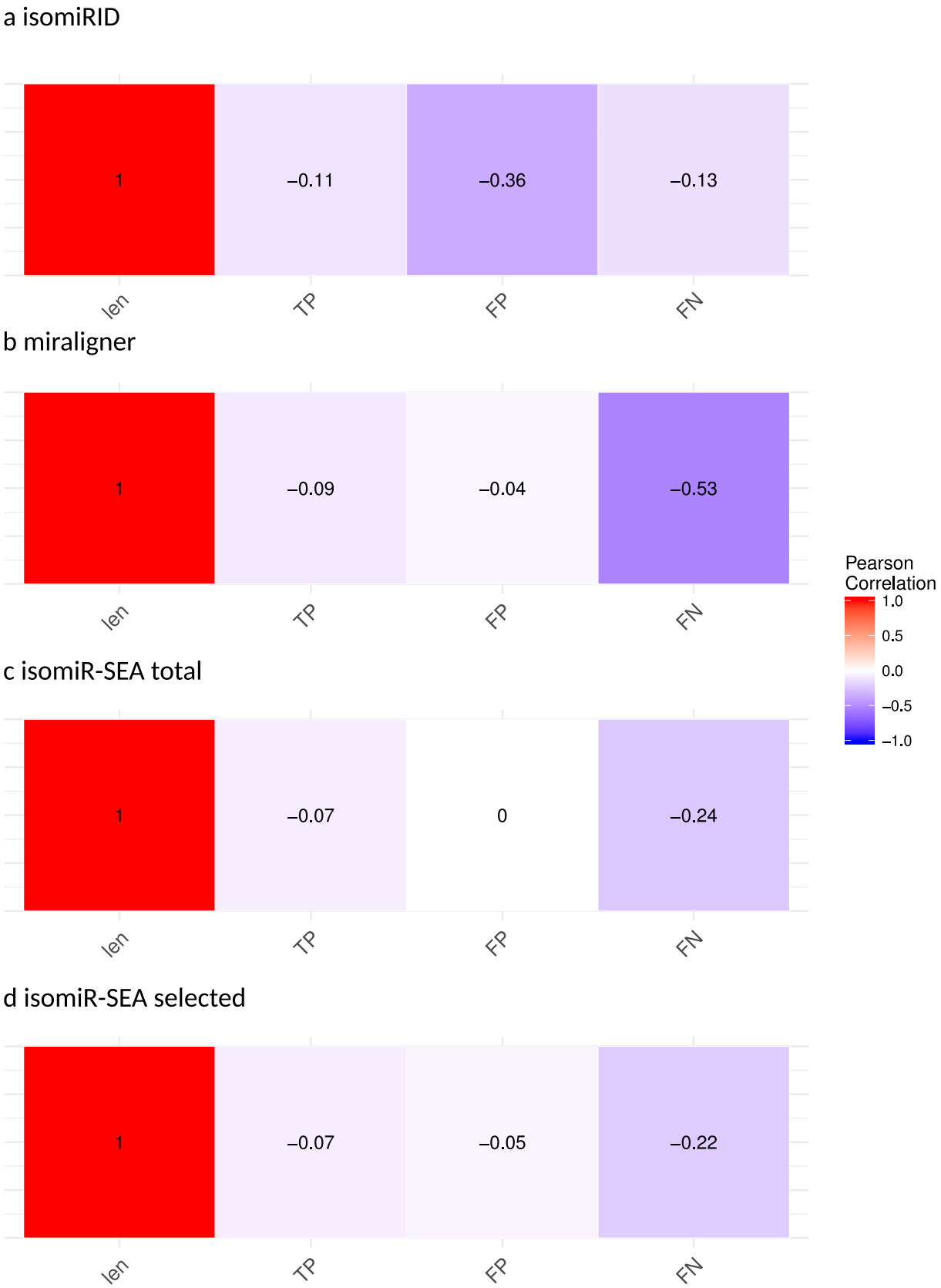


Figure S 2 **Pearson correlation of the length against the true positive, false positive and false negative rate.** IsomiRID has a weak anti-correlation of length and false positive rate. Miraligner has a moderate anti-correlation of length and false negative

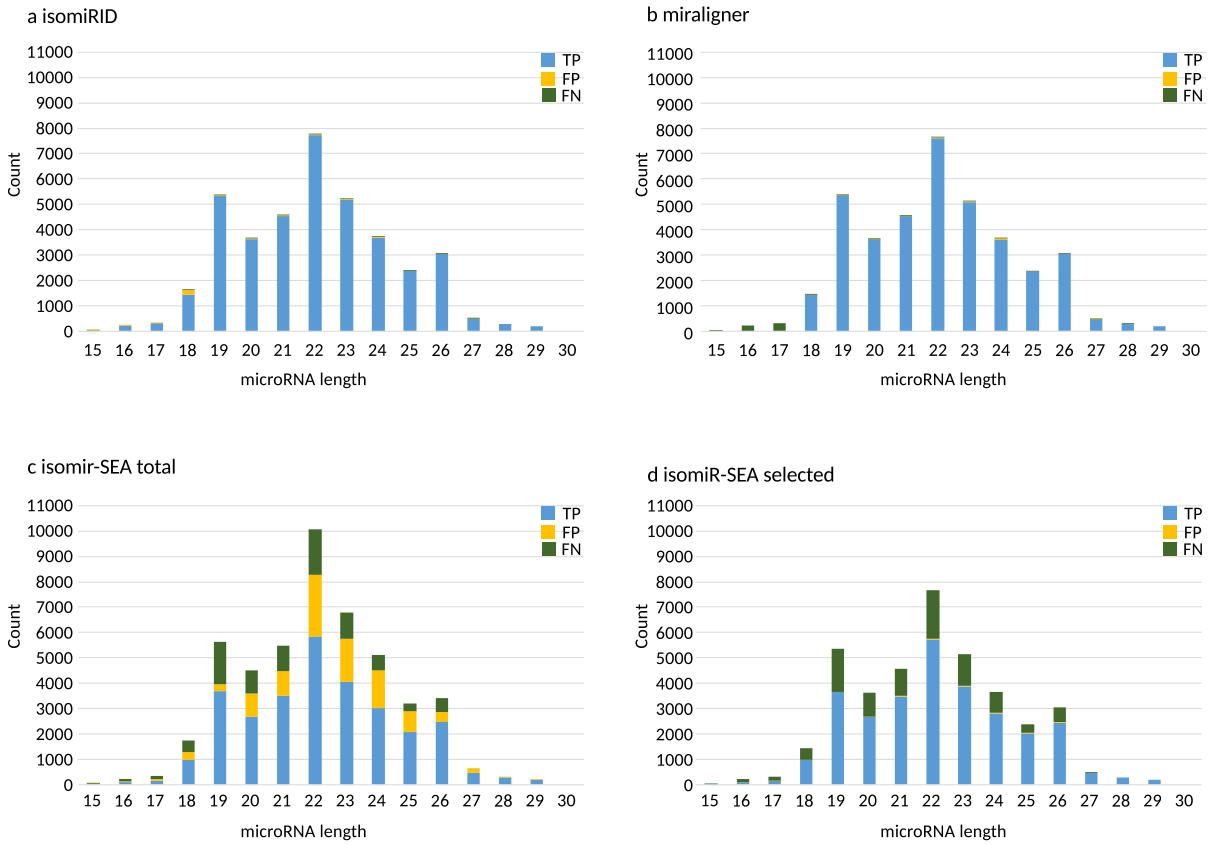


Figure S 3 **Detail view on the various lengths and their individual TP, FP and FN rates.**

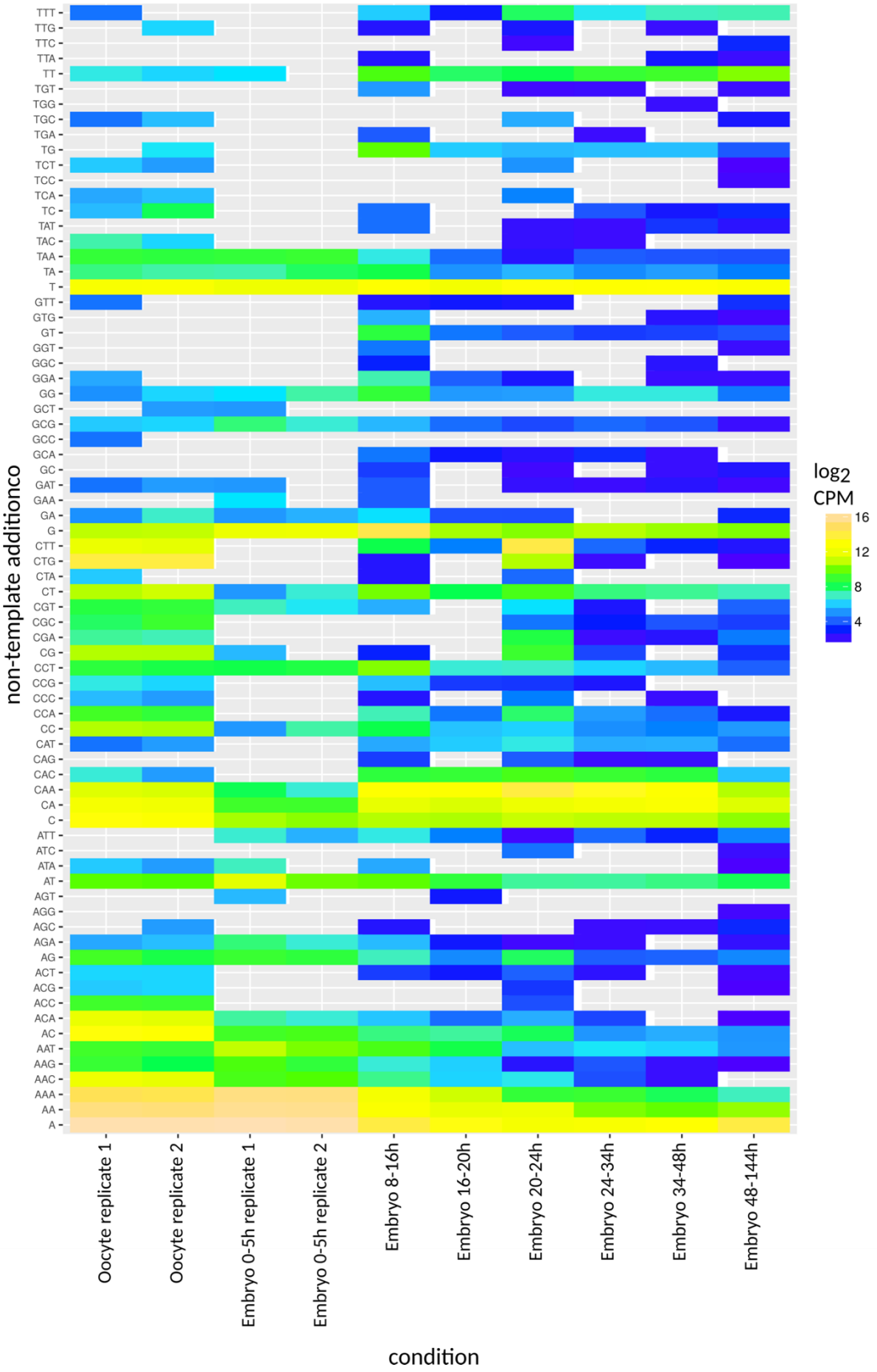


Figure S 4 **Non-templated 3' additions over all conditions. Strong expression of isomiRs with polyadenylate tails was observed in the oocyte and during the first embryonic phase.**

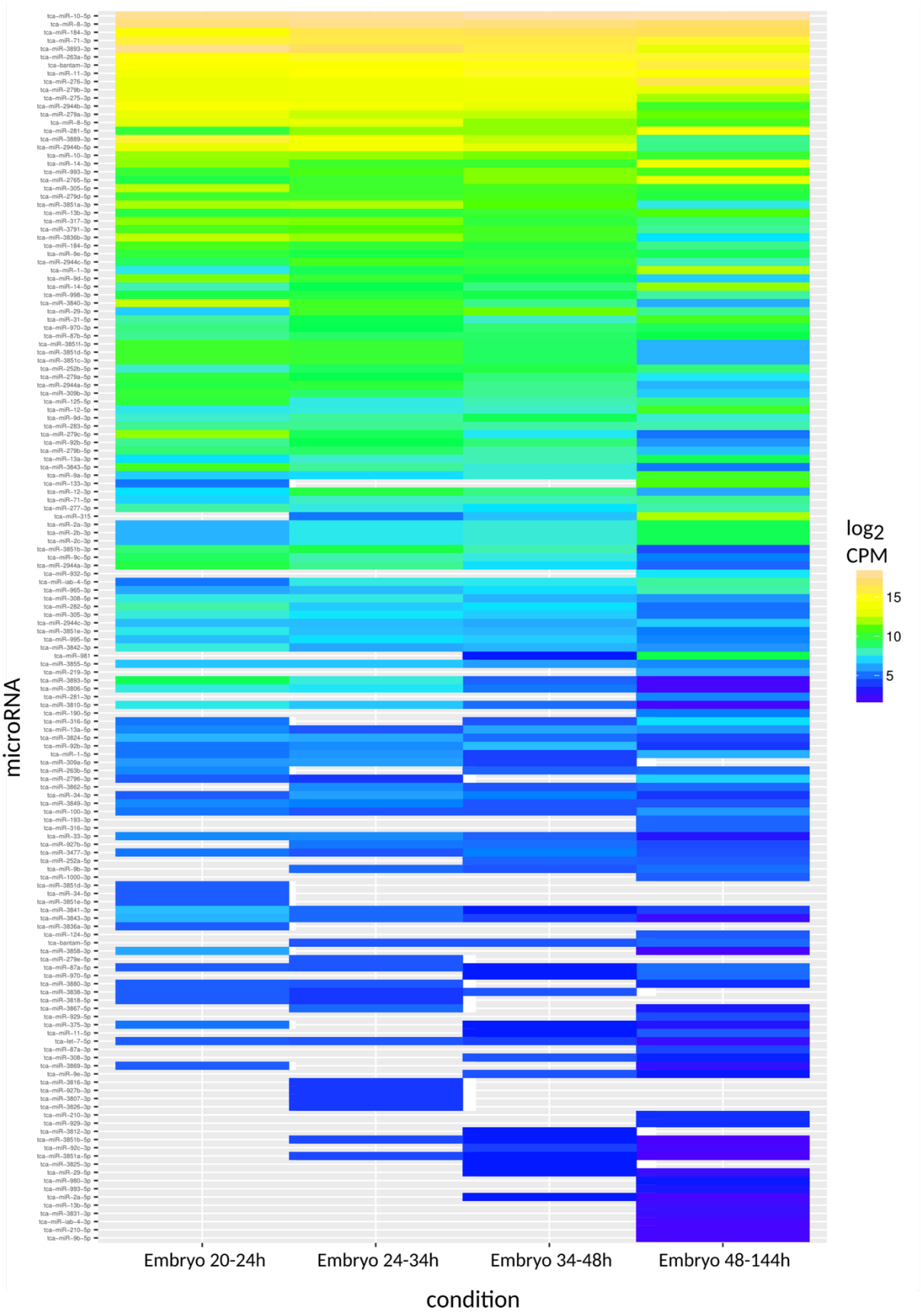


Figure S 5 **Expression of mature miRNAs during the last four embryonic phases. The number of mature miRNAs increases between the 20–24h and 48–144h phases.**

13.6 Supplemental Material: Scripted Workflow Code Documentation

The scripts included in the workflow are described in the following. Each pseudocode block is highlighted in boxes with the name of the current LINUX SHELL script in the top left corner. At first, I described the tool dependencies for the particular script, followed by an input and output definition. Comments to the pseudocode are written in italics and begin with a hash symbol (#). Lines of hashes are for formatting purposes. The lines starting with five hashes indicate a 'chapter' in the code, treating a sub-issue. The custom scripts in PERL are written in bold and the lines start with an arrow (→). Open source tools are also written in bold, but do not start with an arrow sign. Pseudocode, belonging to a certain block, like a PERL script, are indented. Special parameters for a tool are written as comments below the actual tool.

005_trimming.sh

```
# check for installed tools:
      # [cutadapt, bwa, samtools]

# dependencies [] have been checked earlier

#########################################################################
# INPUT: adapter sequences, small RNA raw reads, non-coding RNA reference
multi-fasta

# OUTPUT: trimmed and non-coding RNA filtered small RNA reads
#########################################################################


##### trim reads from small RNA fastq
# IN: adapter sequences, small RNA raw reads
# OUT: trimmed small RNA reads
→ 006_cutadapt_folder.pl
      cutadapt
            # minimal read length: 17
            # maximal read length: 40
            # remove terminal N characters


##### filter unwanted non-coding RNA
# IN: trimmed small RNA reads
# OUT: trimmed and non-coding RNA filtered small RNA reads
```

```
# index non-coding RNA data
bwa index

# map reads to non-coding RNAs
bwa aln

# transform bwa output format
bwa samse

# retain reads that did NOT map to non-coding RNAs
samtools view
```

010_mining.sh

```
# check for installed tools:
      # [miRDeep2.pl]

# dependencies [] have been checked earlier

#############################################################################

# INPUT: 3-letter code of species, small RNA sequencing file(s),
reference genome of species

# OUTPUT: miRDeep2 output in .csv format

#############################################################################


download current [mature.fa, hairpin.fa, organism.txt] from miRBase.org

##### prepare miRNA reference files for miRDeep2
# IN: [mature.fa, hairpin.fa, organism.txt] from miRBase.org, 3-letter
code of species
# OUT: xxx_mature_sequences.fa, all_other_mature_sequences.fa,
xxx_precursor_sequences.fa
→ 011_mirbase_files.pl
      parse miRBase.org files
      filter for metazoan data only
```

```
      generate microRNA mature and precursor file for species
      generate microRNA mature file for all other metazoans


concatenate all small RNA .fastq files into one

##### run miRDeep2 for mining novel microRNAs
# IN: concatenated_smallRNA.fastq, xxx_mature_sequences.fa,
all_other_mature_sequences.fa, xxx_precursor_sequences.fa, reference
genome
# OUT: miRDeep2 output in .csv format
→ 012_miRDeep2_bwt1.pl

      for [reference genome, xxx_mature_sequences.fa,
      all_other_mature_sequences.fa, xxx_precursor_sequences.fa]:

            remove whitespaces

      for [xxx_mature_sequences.fa, all_other_mature_sequences.fa,
      xxx_precursor_sequences.fa]:

            convert RNA to DNA

      create bowtie index for genome
      convert small RNA .fastq to .fasta
      remove whitespaces from small RNA .fasta
      collapse identical reads from small RNA .fasta
      bowtie1 map reads to reference genome
      miRDeep2 mining of novel reads
```

```
020_complete_miRBase.sh

# check for installed tools:
      # none

# dependencies [] have been checked earlier

#########################################################################
```

```
# INPUT: miRDeep2 result .csv file, xxx_mature_mirbase.fa, list of
genomic copy precursors

# OUTPUT: xxx_mature_mirbase_complete.fa file with additional annotations

#########################################################################



##### try to complete miRBase mature file for species
# IN: miRDeep2 result .csv file, xxx_mature_mirbase.fa, list of genomic
copy precursors
# OUT: xxx_mature_mirbase_complete.fa file
```

→**021_parse_miRDeep2_output.pl**

parse miRBase mature file for missing mature annotations
identify arm according to 5p/3p notation in miRDeep2 results
rename and copy mature sequences of precursor copies

040_mirdeep2fasta.sh

```
# check for installed tools:
      # none

# dependencies [] have been checked earlier

#########################################################################

# INPUT: miRDeep2 result .csv file, miRDeep2 cutoff for novel microRNAs,
xxx_mature_mirbase_complete.fa, xxx_precursor_mirbase.fa

# OUTPUT: xxx_mature_mirbase_complete_novel.fa and
xxx_precursor_mirbase_novel.fa with novel microRNAs

#########################################################################



##### parse novel microRNAs from miRDeep2 according to cutoff
# IN: miRDeep2 result .csv file, miRDeep2 cutoff for novel microRNAs,
xxx_mature_mirbase_complete.fa, xxx_precursor_mirbase.fa
```

```
# OUT: xxx_mature_mirbase_complete_novel.fa and
xxx_precursor_mirbase_novel.fa with novel microRNAs
→ 041_curated_mirdeep2fasta.pl
      parse miRDeep2 result .csv
      identify 5p and 3p arms according to their position in precursor
      name the novel miRNAs xxx-new-count

concatenate existing and novel mature sequences of species
concatenate existing and novel precursor sequences of species
```

```
060_quantification.sh

# check for installed tools:
      # none

# dependencies [miRDeep2, bwa, samtools] have been checked earlier

#########################################################################

# INPUT: xxx_mature_mirbase_complete_novel.fa, filtered and trimmed small
RNA read files

# OUTPUT: expression values for each microRNA in each condition

#########################################################################


##### map small RNA sequencing reads against the mature microRNAs
# convert xxx_mature_mirbase_complete_novel.fa from RNA to DNA with
script from miRDeep2
rna2dna.pl

# index xxx_mature_mirbase_complete_novel.fa
bwa index

# map filtered and trimmed small RNA reads to
xxx_mature_mirbase_complete_novel.fa
bwa aln

# convert output of bwa to sam
```

bwa samse

 # discard reads that did not map
samtools view

 # convert multimapping read notation from one-line to multiple lines per hit
xa2multi.pl

Calculate the normalized expression of each mature microRNA in ReadsPerMillion (RPM)
 # IN: config file that assigns each small RNA sequencing file a condition, xxx_mature_mirbase_complete_novel.fa
 # OUT: expression values for each microRNA in each condition
→ 061_sam2de.pl

parse xxx_mature_mirbase_complete_novel.fa for microRNA list
parse config file
save hash with conditions and array with replicates
loop through replicates
calculate RPM for each microRNA
average RPM over all conditions
report RPM and condition per microRNA from list

070_isomiR.sh

 # check for installed tools:
 # [RNAfold, miraligner]

 # dependencies [] have been checked earlier

##

 # INPUT: trimmed small RNA sequencing files, condition ID, miRNA.str file from miRBase.org, 3-letter code of species

 # OUTPUT: reformatted and normalized miraligner output according to replicates and expression in ReadsPerMillion

##

Remove undetermined (N) nucleotides within reads

 # IN: trimmed small RNA sequencing read files

 # OUT: trimmed small RNA seqeuencing read files without N's in reads

→ 071_filter_fastq_N.pl

remove 'N' characters within reads

calculates structure information for novel precursor microRNAs and adds them to the existing miRNA.str file from miRBase.org

 # IN: miRNA.str file from miRBase, xxx_mature_mirbase_complete_novel.fa, xxx_precursor_mirbase_novel.fa

 # OUT: custom.str file, like miRNA.str (miRBase), but for novel microRNAs

→ 072_create_mirbase_struct.pl

parse input files

identify positions of mature microRNA sequences in precursor

 # create secondary structure of precursor hairpin

RNAfold

create header for structure file with information:

- species 3-letter code
- folding free energy
- mature microRNA sequence positions in precursor

make mature microRNA parts uppercase in secondary structure

add secondary structure below header

append entry to miRNA.str

run miraligner to identify microRNA isoforms

 # IN: trimmed and N-filtered small RNA sequencing reads, species 3-letter code, modified miRNA.str structure file from miRBase.org

 # OUT: miraligner output file

→ 073_seqbuster_pipe.pl

miraligner

 # -sub 1

 # -trim 3

 # -add 3

reformat output of miraligner and normalize according to replicates and expression in ReadsPerMillion

```
# IN: all miraligner output files from one condition, condition ID
# OUT: reformatted and normalized miraligner output according to
replicates and expression in ReadsPerMillion
→ 074_reformat_isomiRs.pl
      parse miraligner output
      calculate ReadsPerMillion for expression
      average over number of replicates
```

```
080_miRNA_posGenome.sh

# check for installed tools:
      # none

# dependencies [blast] have been checked earlier

#############################################################################

# INPUT: xxx_precursor_mirbase_novel.fa, reference genome

# OUTPUT: tab separated list of precursor positions on genome

#############################################################################

##### create blast database for reference genome
# IN: reference genome
# OUT: blast database of reference genome
makeblastdb

##### microRNA precursors versus genome
# IN: blast database, xxx_precursor_mirbase_novel.fa
# OUT: tab separated list of precursor positions on genome

blastn
      # -dust no
      # -soft_masking false
      # -outfmt 6
            "qseqid sseqid pident length qlen mismatch gapopen qstart
            qend sstart send evalue"
```

```
##### filtering for 100% hits
# IN: tab separated list of precursor positions on genome
# OUT: filtered tab separated list of precursor positions on genome
awk
```

090_miRNA_homologs.sh

```
# check for installed tools:
      # none

# dependencies [blast] have been checked earlier

#############################################################################

# INPUT: mature microRNA sequences from miRBase.org without species,
xxx_mature_mirbase_novel_complete.fa

# OUTPUT: filtered list of homologous microRNAs in miRBase.org

#############################################################################


##### create blast database of miRBase.org mature microRNA sequences
# IN: mature microRNA sequences from miRBase.org without species
# OUT: blast database
makeblastdb

##### run blastn search of homologous microRNAs
# IN: blast database, xxx_mature_mirbase_novel_complete.fa
# OUT: filtered list of homologous microRNAs in miRBase.org
→ 091_blast_qcov_short.pl
      blastn
            # -word_size 4
            # -evalue 10000
            # -strand plus
            # -outfmt
                  "6 qseqid sseqid pident length mismatch gapopen qstart
                  qend sstart send evalue bitscore qseq sseq"
      filter result
      # - minimal hit length 10
```

```
      # - query and target start are identical
      # - from position 11 to end, one gap or one mismatch is allowed
```

005_getFiles.sh

```
# check for installed tools:
      # sra-toolkit

# dependencies [] have been checked earlier

################################################################################

# INPUT: CLIP-seq, genome, gff, rna and protein URLs

# OUTPUT: downloaded and unzipped data

################################################################################


download and unzip the data for the analysis
```

010_proteinortho.sh

```
# check for installed tools:
      # [proteinortho, blast]

# dependencies [] have been checked earlier

################################################################################

# INPUT: protein multi-fasta files for both species

# OUTPUT: ProteinOrtho output

################################################################################


# make blast databases for protein sets of both species
```

```
makeblastdb

# run proteinortho between the two species
proteinortho5.pl
```

020_trimming.sh

```
# check for installed tools:
      # [cutadapt]

# dependencies [] have been checked earlier

#############################################################################

# INPUT: CLIP-seq read files in fastq format, artificial adapter

# OUTPUT: trimmed CLIP-seq read files

#############################################################################

# trim artificial adapter sequences and remove terminal undetermined
nucleotides
cutadapt
```

025_build_db.sh

```
# check for installed tools:
      # [gmap_build]

# dependencies [] have been checked earlier

#############################################################################

# INPUT: reference genome of AGO CLIP-seq donor species

# OUTPUT: reference genome database for gsnap

#############################################################################
```

```
# build genome index for gsnap
gmap_build
      # -k 15
      # -g
```

030_clip_mapping.sh

```
# check for installed tools:
      # [gsnap, samtools, bedtools]

# dependencies [] have been checked earlier

#############################################################################

# INPUT: AGO CLIP-seq read files, indexed reference database

# OUTPUT: gsnap mapping output in .bed format

#############################################################################


##### perform mapping of the AGO CLIP-seq reads against the genome
# IN: AGO CLIP-seq read files, indexed reference database
# OUT: gsnap mapping output in unsorted .sam format
gsnap
      # -N 1
      # -B 5
      # --speed 1


##### reformat gsnap output to sorted bam format
# IN: gsnap output in .sam format
# OUT: gsnap output in sorted .bam format
samtools view
      # -Sb
samtools sort
```

```
##### reformat gsnap/samtools output from .bam to .bed
# IN: sorted .bam output from gsnap/samtools
# OUT: output in .bed format
bedtools bamtobed
```

040_piranha.sh

```
# check for installed tools:
      # Piranha

# dependencies [] have been checked earlier

################################################################################

# INPUT: gsnap output in bed format

# OUTPUT: sorted Piranha output in .bed format

################################################################################


##### run Piranha to call signaling peaks of putative AGO binding sites
Piranha

##### sort .bed output of Piranha
# IN: .bed file
# OUT: sorted .bed file
sort
      # -k1,1 -k2,2n
```

045_bedtools_merge.sh

```
# check for installed tools:
      # none

# dependencies [] have been checked earlier
```

```
#########################################################################

# INPUT: all CLIP .bed files

# OUTPUT: CLIP .bed file with support level for each position in region

#########################################################################

##### parse all .bed files and save the support level for each position
in the genome, then append the support info to reach region in the .bed
file
→ 046_merge_bed_files.pl
      loop through .bed files
      memorize chromosome, strand and condition
      incrementally count the support at each position
      append information to each line in .bed file
```

048_filterBED.sh

```
# check for installed tools:
      # none

# dependencies [] have been checked earlier

#########################################################################

# INPUT: custom .bed file with support level information

# OUTPUT: .bed files for each support level

#########################################################################

##### create .bed files for each support level
→ 049_bed2signal.pl
      loop through .bed file
      extract regions for each support level
```

050_clip2gff.sh

```
# check for installed tools:
      # none

# dependencies [] have been checked earlier

#############################################################################

# INPUT: .GFF annotation file from AGO CLIP donor species

# OUTPUT: .bed file with transcript annotation information

#############################################################################

##### Add the XM transcript IDs to the .bed file in case they overlap
with a transcript region on the genome
→ clip_mapper.pl
      parse .GFF file
      loop through .bed file
      append transcript ID to line,
            if bed region hits a transcript at least half
```

070_process.sh

```
# check for installed tools:
      # none

# dependencies [bedtools] have been checked earlier

#############################################################################

# INPUT: .bed file, minimal and maximal region values, reference genome

# OUTPUT: .bed file with regions between the minimal and maximal size,
that map a mRNA gene on the genome

#############################################################################
```

```
##### remove .bed entries that are too small or too large
# IN: .bed file, minimal and maximal value
# OUT: filtered .bed file
→ 071_bedtool_discard_sizes.pl
      loop through bed file and discard too small or too large regions

##### sort .bed file
# IN: .bed file
# OUT: sorted .bed file
sort
      -k1,1 -k2,2n


##### get fasta sequences from .bed file and reference genome
# IN: reference genome, .bed file
# OUT: .fasta file of .bed regions
bedtools getfasta


##### Filter those regions that mapped to a mRNA gene on the genome and
make all nucleotides in the .fasta file upper case
# IN: .fasta file of .bed regions
# OUT: upper case .fasta file, filtered for mRNA mapping
→ 072_fasta_uc_and_filter4annotations.pl
      loop through .fasta file
      parse header for annotation: next if not available
      make nucleotide sequence upper case
```

080_transfer.sh

```
# check for installed tools:
      # needle

# dependencies [] have been checked earlier

#########################################################################
```

 # INPUT: .GFF of both species, .fasta file with CLIP regions, ProteinOrtho output, transcriptome of species of interest

 # OUTPUT: needle output of transferred CLIP regions in .CSV format

##

loop through each .GFF file and write out mRNA ID and protein ID of each longest mRNA isoform
 # IN: .GFF file of both species
 # OUT: .CSV file with IDs of the longest mRNA isoform and protein

→ 085_parse_gff_return_longest_transcript.pl

loop through .GFF
next if not gene, mRNA, exon or CDS
create parent-child relation between gene-mRNA-exon-CDS
get protein ID from CDS entry
get mRNA ID from exon entry

transfer the CLIP region from the donor species to the species of interest
 # IN: gff-csv file of donor and species of interest, ProteinOrtho output, .fasta file with CLIP regions, transcriptome of species of interest
 # OUT: needle output of transferred CLIP regions

→ 081_map_clip_gff_needle.pl

parse gff-csv file for proteinID-mRNAID for donor species
parse ProteinOrtho output file
    for proteinID assignment between species
parse clip .fasta file
parse gff-csv file for proteinID-mRNAID for species of interest
parse transcriptome of species of interest

identify orthologous CLIP sequence from donor
    and mRNA from species of interest
 # run needle with those two sequences
needle
    # -endweight Y
    # -gapopen 5
    # -gapextend 2
parse needle output for mapping statistics and IDs

090_tarpred.sh

```
# check for installed tools:
      # miranda

# dependencies [bedtools] have been checked earlier

#############################################################################

# INPUT: microRNA .fasta file, .bed file of CLIP regions, transcriptome
of species of interest, needle .csv output file

# OUTPUT: miranda target prediction file
#############################################################################


# IN: needle output file in .csv format
# OUT: needle output file in .bed format
→ 095_csv_to_bed.pl
      parse needle .csv output file
      rearrange information to .bed format

##### merge transferred clip regions
# IN: .bed file of clip regions
# OUT: .bed file of merged clip regions
bedtools merge


##### get fasta file from transcriptome and .bed file
# IN: .bed file, transcriptome
# OUT: .fasta file of .bed regions
bedtools getfasta


##### perform target prediction of previously created mature microRNA set
against transferred CLIP regions
# IN: microRNA .fasta file, transferred target CLIP .fasta file
# OUT: targetprediction file from miranda
→ 096_mapping.pl
      parse .fasta file
```

```
create temporary file for each sequence
# run miranda with all microRNAs against the single sequence
miranda
parse miranda output and write to output file
```

13.7 Supplemental Material: Database

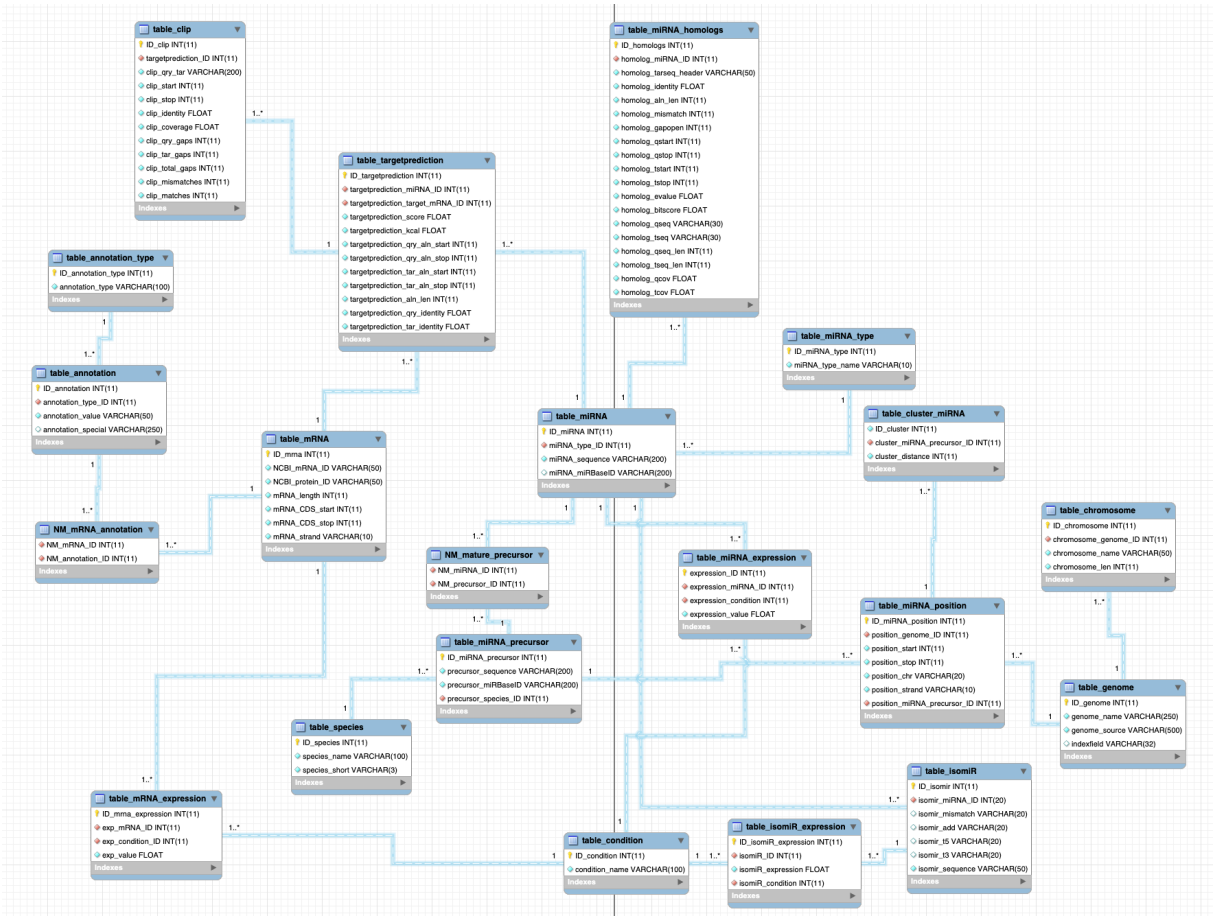


Figure S 6 **Database detail scheme**

Load_db.pl

```
# check for installed tools:
      # none


#############################################################################

# INPUT: config file, genome file, microRNA position file, microRNA
expression file, microRNA isoform file, microRNA homolog file, .gff file,
mRNA expression file, mRNA target prediction file, annotation file,
miRNA.dat file from miRBase with microPIECE extended results

# OUTPUT: .sql statement file for loading the database

#############################################################################


##### table_genome; table_species; table_miRNA_type; table_condition
parse config file with information about
      - species name
      - genome name and download link
      - microRNA arm types (like 3p;5p)
      - conditions list

parse genome file and save information about
      - chromosome name
      - chromosome length

create SQL statements

##### table_chromosome
parses the genome file to create chromosome-name;chromosome-length tuples
takes genome name from table_genome as foreign key

create SQL statements

##### table_miRNA; talbe_miRNA_precursor; NM_mature_precursor
# early version parsed .fasta file of mature and precursor
# newer version parses miRNA.dat file (microPIECE output / miRBase.org)
get sequence data and miRBase.org name for microRNA precursor
```

```
derive 3-letter code as foreign key from table_species

get sequence data and miRBase.org name for mature microRNA
get microRNA type ID dynamically from table_miRNA_type comparison

dynamically create IDs for the NM_mature_precursor table
create SQL statements


##### table_miRNA_position
parse miRNA position file to gain data for
      - start
      - stop
      - chromosome
      - strand
get genome ID from table_genome as foreign key
create SQL statements

##### table_miRNA_expression
parse miRNA expression file for information on expression
get microRNA from table_miRNA as foreign key
get condition ID from table_condition as foreign key
create SQL statements

##### table_isomiR; table_isomiR_expression
parse microRNA isoforms file
get microRNA ID from table_miRNA as foreign key for table_isomiR
get condition ID from table_condition
      as foreign key for table_isomiR_expression
create SQL statements


##### table_homologs
parse homologs file according to BLASTN output format
get miRNA ID from table_miRNA as foreign key
create SQL statements


##### table_mRNA
parse gff file
create parent-child dependencies for gene-exon-cds
```

```
calculate relative start and stop of UTR and CDS
extract information for strand, IDs and mRNA length
create SQL statements


##### table_targetprediction; table_clip
parse target prediction file
extract clip values from target column in result file
get miRNA and mRNA IDs from table_miRNA and table_mRNA as
      foreign keys for table_targetprediction
get dynamic ID from table_targetprediction for table_clip as foreign key
create SQL statements


##### table_annotation; table_annotation_type; NM_mRNA_annotation
parse .csv annotation file line by line and split at comma character
get foreign keys from table_mRNA and table_annotation
create SQL statements
```

13.8 Supplemental Material: The microPIECE Pipeline

microPIECE.pl

```
##########################################################################

# INPUT: speciesA genome, speciaA .gff, speciesA AGO-CLIP data, speciesB
genome, speciesB .gff, speciesB mature miRNA set (speciesB ncRNA set,
speciesB miRNA precursor set, speciesB smallRNA-seq data)

# OUTPUT: final_mirbase_microPIECEoutput.dat,
mature_combined_mirbase_novel.fa, hairpin_combined_mirbase_novel.fa,
miRNA_expression.csv, miRNA_orthologs.csv, mirdeep_output.html/csv,
isomir_output_CONDITION.csv, miRNA_genomic_position.csv, all library
support level target predictions, all library support level CLIP transfer
.bed files

##########################################################################

# print hello
microPIECE::hello

# run requirements check
microPIECE::check_requirements

# print settings
microPIECE::print_settings

# run microRNA part
microPIECE::run_mining

# run CLIP part
microPIECE::run_clip

# run targetprediction
microPIECE::run_targetprediction

# transfer resultfiles
microPIECE::transfer_resultfiles
```

sub create_folder

creates output folders

```
################################################################################
# HELLO PART
################################################################################
```

sub hello

prints a hello message with the current version of microPIECE

```
################################################################################
# CHECKING PART
################################################################################
```

sub check_files

tests if the provided files exist

sub check_requirements

→ **check_files**

 # check if mandatory files are provided and accessible

genomeA
genomeB
annotationA
annotationB

check if CLIP data is provided
    if yes: run at least minimal pipeline with region transfer
        check if data is accessible
            if yes: continue
            if no: terminate
        check if adapter sequence was provided &&
        check if adapter sequence is >0 &&
        check if adapter sequence contains only ACGT characters
            if yes: continue
            if no: terminate
    if no: terminate

```
      check if small RNA sequencing data is provided
            if yes: include mining procedure into pipeline
                  check if data is accessible
                        if yes: continue
                        if no: terminate
                  check if adapter sequence was provided
                        if yes: trimming will be performed
                              check if provided adapter has unexpected
                              characters
                                    if yes: terminate
                                    if no: continue
                        if no: adapter cannot be removed
                  → check_files # file for non-coding RNA filtering
                        check if 3-letter code was defined
                        if yes: continue
                        if no: terminate
            if no: skip mining branch

      check if microRNA data is provided
            if yes: skip mining, because .fasta file of microRNAs is
            provided
            if no: continue

      check if CLIP data and (microRNA data OR small RNA data) is
      provided
            if yes: include target prediction into pipeline
            if no:  continue

      check external tool dependencies are fulfilled
            if yes: continue
            if no: terminate

      check if output directory already exists
            if yes: terminate and suggest –overwrite parameter
            if no: continue


#########################################################################
# PRINT SETTINGS PART
#########################################################################
```

sub print_settings

prints all used settings into the log-file

```
################################################################################
# MINING PART
################################################################################
```

sub run_mining

 # calls the individual subroutines for each step

→ **create_folder**

→ **run_mining_clipping**

→ **run_mining_filtering**

→ **run_mining_downloads**

→ **run_mining_mirbase_files**

→ **run_mining_mirdeep2**

→ **run_mining_complete**

→ **run_mining_mirdeep2fasta**

→ **run_mining_quantification**

→ **run_mining_isomir**

→ **run_mining_genomicposition**

→ **run_mining_orthologs**

sub run_mining_clipping

cutadapt

set minimal read length to 17

remove undefined nucleotides at the ends

write trimmed file with "_trimmed.fq" suffix

sub run_mining_filtering

 # create BWA index database for non-coding RNA file

bwa index

 # map small RNA seq data against other non-coding RNAs

bwa aln

 # -n 1 := edit distance of 1

 # -o 0 := no gap opens

 # -e 0 := no gap extension

 # -k 1 := max 1 difference in seed region

```
      # convert bwa output to .sam format
      bwa samse

      # retain only unmapped results
      samtools view -f 4

      # sort retaining results
      samtools sort

      # convert .sam file back to .fastq reads
      bedtools bamtofastq


sub run_mining_downloads
      # call subroutine to either download or verify local files:
      # organisms.txt.gz, mature.fa.gz, hairpin.fa.gz, miRNA.dat.gz
      → sub get_mirbase_download_or_local_copy


sub get_mirbase_download_or_local_copy
      check if directory and files exist
      if yes: continue
      if no: download files from miRBase.org and unkompress them
            wget
            gunzip


sub run_mining_mirbase_files
# create miRDeep2 reference files from miRBase.org files
      → MINING_split_mirbase_files.pl
            # --species
            # --precursor-file
            # --mature_file
            # --organism
            # --outmature
            # --outprecursor
            # --outnonspeciesmature
            parse organisms.txt file and select Metazoan species only
            parse mature.fa from miRBase.org
                  filter for Metazoan species
```

```
                  divide sequences into 3-letter-code species and other
            parse hairpin.fa from miRBase.org
                  filter for Metazoan species
                  select sequences from 3-letter-code species

sub run_mining_mirdeep2
# remove whitespaces from .fasta headers with miRDeep2 included script
      → remove_white_space_in_id.pl (miRDeep2)
            # Input:
            # genome file
            # splitted miRBase.org files

      # convert RNA to DNA alphabet
      → run_mining_rna2dna
            # Input:
            # splitted miRBase.org files

      # create index file for genome to run with miRDeep2
      bowtie-build
            # Input:
            # genome .fasta file without whitespaces

      concatenate all small RNA sequencing files into one file
      # convert .fastq format to .fasta format with miRDeep2 included
      script
      → fastq2fasta.pl (miRDeep2)
            # concatenated small RNA sequencing files

      # remove whitespaces from .fasta headers with miRDeep2 included
      script
      → remove_white_space_in_id.pl (miRDeep2)
            # concatenated small RNA sequencing .fasta file

      # collapse reads in .fasta file with miRDeep2 included script
      → collapse_reads_md.pl (miRDeep2)
            # concatenated small RNA sequenceing .fasta file without
            whitespaces

      # map the concatenated and collapsed small RNA .fasta data to
      genome using the script from miRDeep2, generating the .arf input
      for miRDeep2
```

```
      → mapper.pl (miRDeep2)
            bowtie # minimal read length 17
      # run the actual miRDeep2 mining step
      → miRDeep2.pl (miRDeep2)
            # .arf file from mapper.pl
            # concatenated and collapsed small RNA fasta file
            # reference genome
            # mature miRBase.org reference
            # precursor miRBase.org reference
            # other mature miRBase.org reference species


sub run_mining_rna2dna
      parse file and convert nucleotide U characters to T characters


# complete (if possible) missing entries from miRBase.org for the species
sub run_mining_complete
      → MINING_complete_mirbase_by_miRDeep2_output.pl
            → mining::parse_mirbase_dat

            → mining:parse_mirdeep_known
            filter for precursors without two mature sequences
            if miRDeep2 found missing mature sequences check coverage
            criteria
            # EITHER each mature sequence has >= 10 reads support
            # OR sum of both matures is >= 100 and each has min. 5 reads
            add missing mature sequence
            check if new mature sequence is 5p or 3p

            → mining::export_fasta
            # write completed .fasta file with novel mature sequences
            → mining::export_mirbase_data
            # write completed .dat file with novel mature sequences


# get novel microRNA sequences and combine with known ones
sub run_mining_mirdeep2fasta
      → MINING_curate_mirdeep2fasta.pl
            # filter for score default cutoff 10
```

```
      merge novel and known mature microRNA sequences into .fasta file
      ensure DNA alphabet
      merge novel and known precursor microRNA sequences into .fasta file
      ensure DNA alphabet

# map reads to mature microRNAs to get expression values
sub run_mining_quantification
      # run mapping for each .fastq file
      bwa index
      # run alignment
      bwa aln
            # -n 1
            # -o 0
            # -e 0
            # -k 1
      # transform bwa output file to .sam format
      bwa samse
      # remove unmapped reads
      samtools view
      # transfer multimappings to own line
      xa2multi.pl # external tool
            parses .sam file
            checks for results with multi mapping reads
            creates a novel line for each mutli mapped position
      create config with files and corresponding condition
      MINING_sam2de.pl
            parse config
            parse mature microRNA .fasta file as reference
            count mapping reads for each mature microRNA
            count multimappings divided by their number of mappings
            normalize the read sum by the number of replicates
            # information was given in config file
            calculate Reads Per Million


# include novel microRNAs and run isoform mining for microRNAs
sub run_mining_isomir
      # create miRBase.org structure file .str with novel microRNAs
      → ISOMIR_create_mirbase_struct.pl
            # parse precursor and mature microRNA from .dat file
            mining:parse_mirbase_dat
```

```
            # compute free energy and dot-bracket fold
            RNAfold # external tool
            # plot secondary structure in text format
            RNA::HairpinFigure # external tool
                  generate secondary structure plot
            # check if plot is correct and repair otherwise
            mining::fix_hairpin
      # discard reads with unspecified nucleotides and run miraligner
      → sub filter_for_N_and_collapse_reads
      miraligner
            # -sub 1
            # -trim 3
            # -add 3
      # parse miraligner output files with condition information
      # and merge all isoform types and replicates into one file
      → ISOMIR_reformat_isomirs.pl
          # comma separated list of miraligner output (replicates)
          # condition ID
          parse miraligner output files
          count reads and normalize over replicates
          calculate ReadsPerMillion

# filter reads and collapse identical ones
sub filter_for_N_and_collapse_reads
      parse .fastq file
      discard read if it contains an undetermined nucleotide
      count identical reads and collapse them
      append read count to header of collapsed read


# get genomic positions of microRNA precursors
sub run_mining_genomicposition
      makeblastdb
      blastn
            # -dust no
            # -soft_masking false
            # -outfmt 6

      filter for 100% identity
      filter for 100% coverage
```

```
# determination of homologous sequences in miRBase.org
sub run_mining_orthologs
      → MINING_ortholog_blast.pl
            # create database for blastn search
            makeblastdb
            blastn
                  # -word_size 4
                  # -evalue 10000
                  # -strand plus
                  # -outfmt 6

            filter:
                  -  minimal alignment length of 10
                  -  query and subject having the same start
                  -  the first 10 bp having no gap or mismatch
                  -  the remaining bp have one gap or mismatch maximum
            calculate query and subject coverage



#########################################################################
# CLIP PART
#########################################################################



sub run_clip
      # calls the individual subroutines for each step
      → run_proteinortho
      → run_CLIP_adapter_trimming
      → run_CLIP_build_db
      → run_CLIP_mapping
      → run_CLIP_piranha
      → run_CLIP_bedtools_merge
      → run_CLIP_filterbed
      → run_CLIP_clip_mapper
      → run_CLIP_process
      → run_CLIP_transfer
```

```
# run proteinortho (external tool) for orthologous protein detection
sub run_proteinortho
      # extract protein sequences from .gff and genome files of both
      species
      gffread #external tool
            # -y := write .fasta with protein sequence of CDS
      #create blast database for both species protein sets
      makeblastdb
      # run orthology prediction
      proteinortho

# trim artificial adapters from sequencing reads of CLIP data
sub run_CLIP_adapter_trimming
      # run cutadapt and remove undetermined nucleotides from the end and
      filter for minimal remaining read length of 20 nucleotides
      cutadapt
            # -m 20
            # --trim-n

# create database for mapping the reads with gsnap lateron
sub run_CLIP_build_db
      gmap_build
            # -k 15


# map reads to genome with gsnap and post-process result with samtools
sub run_CLIP_mapping
      gsnap
            # -N 1 := look for splice sites
            # -B 5 := batch mode 5, allocate positions, genome and suffix
            array
            # -O := ordered output
      # convert from .sam to .bam format
      samtools view
      # sort .bam file
      samtools sort
      # index .bam file for further processing
      samtools index


# perform peak calling with mapping results
```

```
sub run_CLIP_piranha
      # if > 1 threads provided, run pre-binning and Piranha in parallel
      for each file
      # else, run pre-binning and Piranha for each file sequentially
      → run_CLIP_piranha_working_thread


# create pre-binned .bed file from alignment .bam file and run Piranha
sub run_CLIP_piranha_working_thread
      # run with default bin-sizes of 30
      # if transcripts are provided, a pseudo-count is added in
      transcript regions to highlight exons for the Piranha peak calling
      → CLIP_binned_bed_from_bam_and_trascripts_for_piranha.pl
          # --bam := .bam file from read alignment
          # --size := bin size
          # --transcripts := .gff file for transcript positions
          # --reqfeature := type of .gff file, default: exon
      # run Piranha to identify the signal regions via peak calling
      Piranha
      sort output

# merge the resulting .bed files from different conditions into one, by
merging overlapping regions and retaining the information of the
individual condition in the fourth .bed column
sub run_CLIP_bedtools_merge
      # call the actual script for merging .bed files into one file
      → CLIP_merge_bed_files.pl
      # input argument can occur more than one time, accounting for
      different conditions, e.g.,
      # --input 24h=file1.bed,file2.bed
      # --input 72h=file3.bed,file4.bed
      # --overwrite overwrites an existing output file
      sort each .bed file by chromosome, strand ,start and stop position
      get counts per position per condition per chromosome per strand
      insert information into the fourth column of the single .bed file


# split the single .bed file into .bed files of different supporting
strength, e.g., only .bed regions, supported by all or by one library
sub run_CLIP_filterbed
      # call the actual script for extracting signal strength positions
```

```
      → CLIP_bed2signal.pl
      # take .bed file and desired signal strength as input
      parse .bed file
      use coordinates and signal counts of fourth column
      keep stretches where signal count >= desired strength


# compare and filter .bed file with .gff file transcript coordinates
sub run_CLIP_clip_mapper
      # retain only .bed coordinates that map to a transcript
      → CLIP_mapper.pl
      parse .bed file
      parse .gff file
      compare .bed regions to transcript regions in .gff file
      retain .bed regions that map at least half to a transcript region

# filter for .bed region min and max length and extract .fasta sequence
sub run_CLIP_process
      # -min 22 (default)
      # -max 50 (default)
      check if min value is smaller than max value
      # filter for CLIP regions of certain length
      → CLIP_bedtool_discard_sizes
      sort .bed file according to chromosome and start
      # get .fasta file from .bed coordinates
      bedtools getfasta
      remove strand information from .fasta header
      convert .fasta sequence to upper case


# use defined minimal and maximal length of region to filter .bed file
sub CLIP_bedtool_discard_sizes
      parse .bed file
      calculate length of region
      compare if length is between minimal and maximal value
      if yes: keep entry
      if no: discard entry


# transfer the identified CLIP signal regions from one species to another
via orthology information
```

```
sub run_CLIP_transfer
      # filter .gff for longest transcripts
      → CLIP_parse_gff_return_longest_transcript.pl
            parse .gff file
            save lengths of each transcript
            select longest transcript per gene
      # extract transcript exons from genome with .gff coordinates
      gffread
            # -w := write .fasta with spliced exons for each transcript
      # extract peak-regions and use orthology information to transfer
      them to homologous transcript
      → CLIP_map_clip_gff_needle.pl
            parse .gff file of both species
            parse proteinortho output
            parse .fasta file of CLIP regions
            parse .fasta file of target transcripts
            extract peak region
            get homologous sequence and use both as input for needle
            needle
                # -endweight Y
                # -gapopen 5
                # -gapextend 2
                # -datafile EDNACUSTOM
                # -auto
                # -aformat markx3
            parse needle output and write to .fasta file

      # use .csv file to create a .bed file
      → CLIP_csv_to_bed.pl
            parse .csv file and rearrange data to .bed format
      # merge overlapping .bed regions
      bedtools merge
      # convert merged .bed file into transcript sequences
      bedtools getfasta



#########################################################################
# TARGET PREDICTION PART
#########################################################################
```

```
# run the target prediction on the transferred regions
sub run_targetprediction
      → Targetprediction.pl
      # provide mature microRNAs and potential target sequences
      parse mature microRNA sequence .fasta file
      write each sequence to temp file
      run miranda with each temp file as input
      miranda
      parse miranda output from each mature microRNA
      filter for lines with mapping details
      # this excludes, e.g., alignment views
      write to condensed output file


################################################################################
# TRANSFER PART
################################################################################

# transfer generated files to output destination
sub transfer_resultfiles
      → copy_final_files
            - miRBase.org .dat file
            - mature microRNA .fasta file
            - precursor microRNA .fasta file
            - mature microRNA expression .csv file
            - microRNA orthologs .csv file
            - miRDeep2 mining result .html file
            - isomiR microRNA output .csv file
            - microRNA genomic location .csv file
            - miranda target prediction .txt file
            - all library support level CLIP .bed files


# copy file from source to destination
sub copy_final_files

# used to unify system calls with log-files, paths,
# output files and error logs
sub run_cmd
```

```
# used to create one or more temporary files at one to save data on the
disk
sub create_tempfile
      use File::Temp::tempnam to create temporary file(s) for processing

# used to remove the temporary file(s) from the disk
sub clean_tempfiles
      unlink temporary file(s)
```

mining.pm

```
##### parses the .dat file from miRBase.org
sub parse_mirbase_dat
      # input file
      # 3-letter code species name
      parse file for blocks of microRNA information
      → _parse_mirbase_dat_block
            # submits the block to the function
            # each block starts with "ID"
            # each block ends with "//"
      filter for blocks that match the 3-letter code of the species
      generate mature sequences from the parsed information


##### function that parses the .dat file
sub _parse_mirbase_dat_block
      # microRNA block, parsed from miRNA.dat file
      parse lines, starting with "FT" to get mature microRNA coordinates
      parse lines, starting with "SQ" to get the precursor sequence
      return precursor ID, species 3-letter code, precursor length,
      mature coordinates and the precursor sequence

##### parse a .fasta file into a PERL hash structure
sub parse_fasta
      use header, starting with ">" as hash-key
      append all sequence lines per key as value to the hash


##### call the _parse_mirdeep function and select novels from output
sub parse_mirdeep_novels
```

```
      → _parse_mirdeep


##### call the _parse_mirdeep function and select known from output
sub parse_mirdeep_known
      → _parse_mirdeep


##### parse the .csv output of miRDeep2 for novel and known microRNAs
sub _parse_mirdeep
      parse block of novel microRNAs in .csv file
      assign 5p and 3p notation to mature/star notation of miRDeep2
      check for genomic copies via checksum
      parse block of known microRNAs in .csv file


##### write precursor and mature sequences into two .fasta files
sub export_fasta
      write precursor sequence with header to .fasta file
      write mature sequences with header to .fasta file


##### generate miRNA.dat entries
sub export_mirbase_data
      create ID line # ids, 3-letter code and basepair length
      create XX line
      create DE line # species and microRNA name
      create XX line
      create FT lines # mature information lines: start, stop, names
      create XX line
      create SQ line # precursor sequence, length, nucleotide counts
      create // line # indicates block ending
      append each line to file


##### controls the external folding algorithm for mistakes
sub fix_hairpin
      parse created structure and compare to sequence
      correct errors by replacing characters
```

13.9 Supplemental Material: MicroRNAs regulate innate immunity against uropathogenic and commensal-like *Escherichia coli* infections in the surrogate insect model *Galleria mellonella*

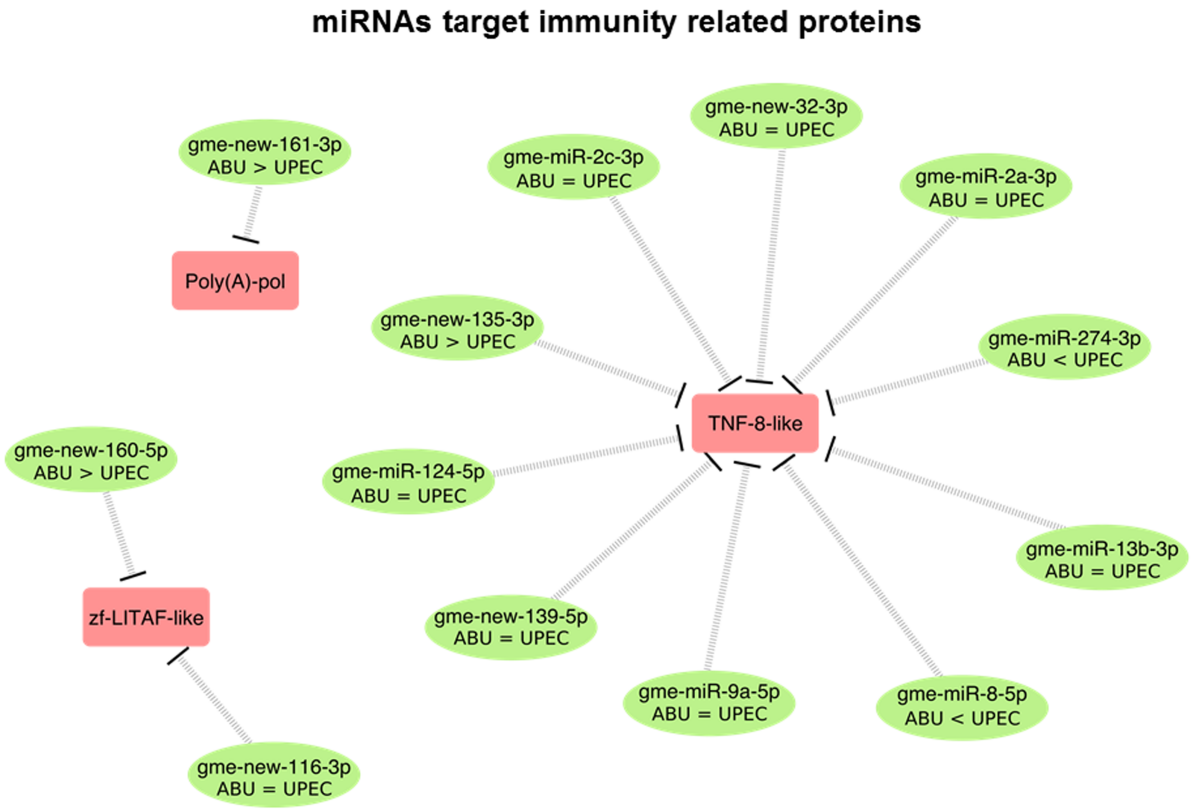


Figure S 7 **The immunity related mRNAs as target of miRNAs in ABU and UPEC infected *G. mellonella* larvae.** The network diagram generated with Cytoscape shows mRNAs related to innate immunity related proteins targeted by miRNAs in ABU and UPEC infected larvae. Target mRNAs are represented in different colors to maintain contrast with the miRNAs. Connecting lines (edges) are used to indicate miRNA targets identified in this study by consulting Gene Ontology terms related to *G. mellonella* genome and transcriptome sequences. ABU<UPEC means miRNA upregulation in UPEC compared to ABU, ABU>UPEC means miRNA upregulation in ABU compared to UPEC, and ABU=UPEC means no difference in miRNA expression.

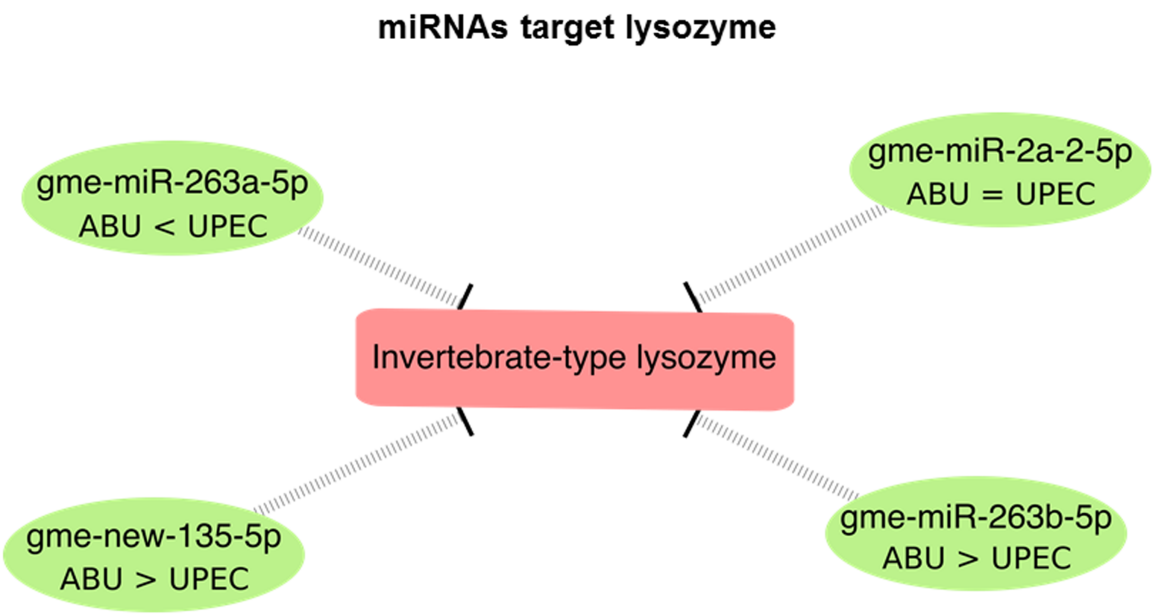


Figure S 8 **The lysozyme mRNA as targets of miRNAs in ABU and UPEC infected *G. mellonella* larvae.** The network diagram generated with Cytoscape shows mRNAs encoding lysozyme targeted by miRNAs in ABU and UPEC infected larvae. Target mRNAs are represented in different colors to maintain contrast with the miRNAs. Connecting lines (edges) are used to indicate miRNA targets identified in this study by consulting Gene Ontology terms related to *G. mellonella* genome and transcriptome sequences. ABU<UPEC means miRNA upregulation in UPEC compared to ABU, ABU>UPEC means miRNA upregulation in ABU compared to UPEC, and ABU=UPEC means no difference in miRNA expression.

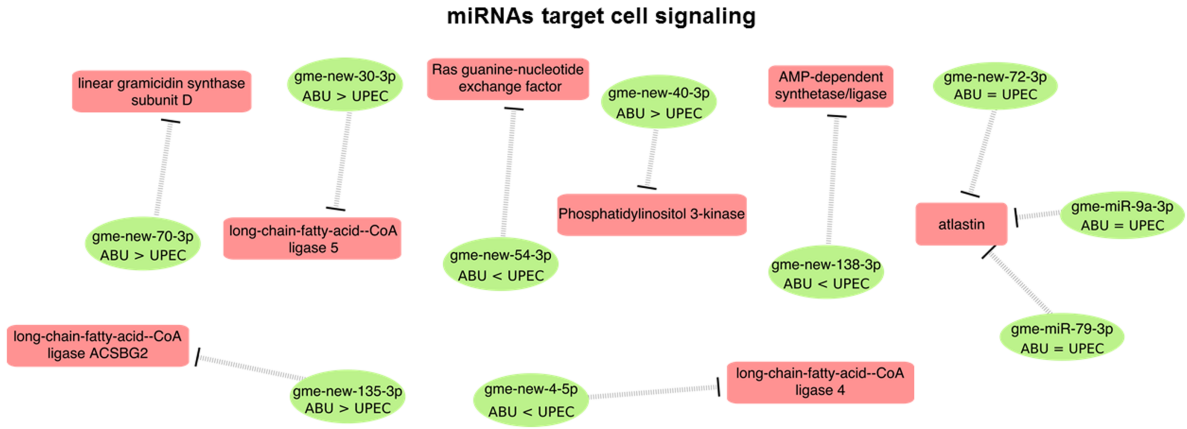


Figure S 9 **The mRNAs related to cell signaling as targets of miRNAs in ABU and UPEC infected *G. mellonella* larvae.** The network diagram generated with Cytoscape shows mRNAs related to cell signaling targeted by miRNAs in ABU and UPEC infected larvae. Target mRNAs are represented in different colors to maintain contrast with the miRNAs. Connecting lines (edges) are used to indicate miRNA targets identified in this study by consulting Gene Ontology terms related to *G. mellonella* genome and transcriptome sequences. ABU<UPEC means miRNA upregulation in UPEC compared to ABU, ABU>UPEC means miRNA upregulation in ABU compared to UPEC, and ABU=UPEC means no difference in miRNA expression.

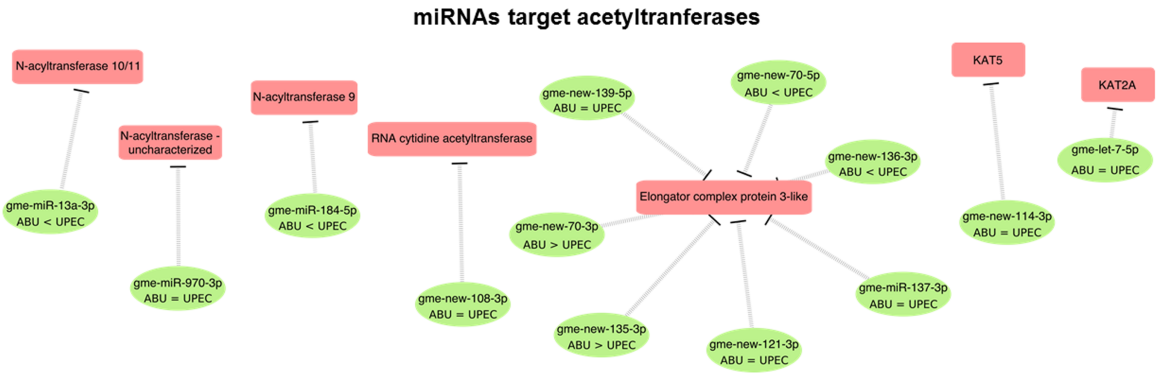


Figure S 10 **The enzyme acetyltransferase mRNAs as targets of miRNAs in ABU and UPEC infected *G. mellonella* larvae.** The network diagram generated with Cytoscape shows mRNAs encoding acetyltranferases targeted by miRNAs in ABU and UPEC infected larvae. Target mRNAs are represented in different colors to maintain contrast with the miRNAs. Connecting lines (edges) are used to indicate miRNA targets identified in this study by consulting Gene Ontology terms related to *G. mellonella* genome and transcriptome sequences. ABU<UPEC means miRNA upregulation in UPEC compared to ABU, ABU>UPEC means miRNA upregulation in ABU compared to UPEC, and ABU=UPEC means no difference in miRNA expression.

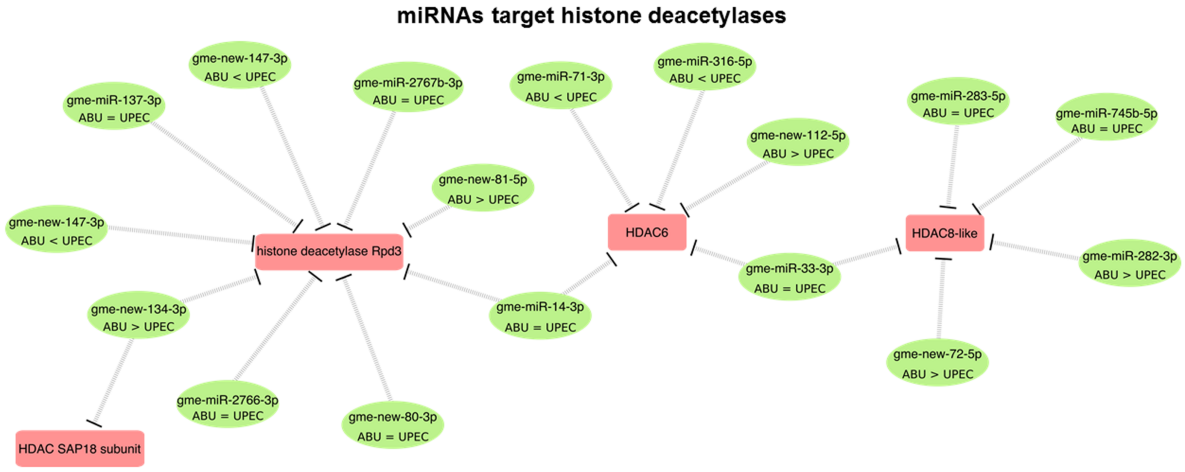


Figure S 11 **The enzyme histone deacetylases mRNAs as targets of miRNAs in ABU and UPEC infected *G. mellonella* larvae.** The network diagram generated with Cytoscape shows mRNAs encoding histone deacetylases targeted by miRNAs in ABU and UPEC infected larvae. Target mRNAs are represented in different colors to maintain contrast with the miRNAs. Connecting lines (edges) are used to indicate miRNA targets identified in this study by consulting Gene Ontology terms related to *G. mellonella* genome and transcriptome sequences. ABU<UPEC means miRNA upregulation in UPEC compared to ABU, ABU>UPEC means miRNA upregulation in ABU compared to UPEC, and ABU=UPEC means no difference in miRNA expression.

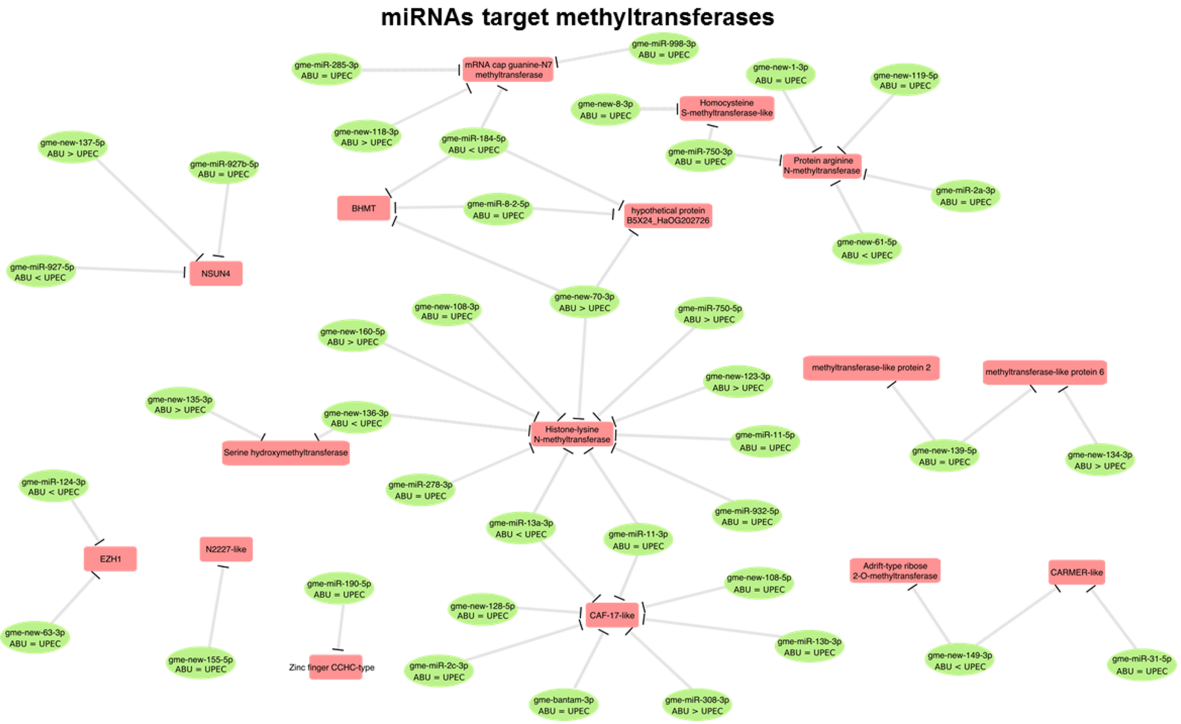


Figure S 12 **The enzyme methyltransferases mRNAs as targets of miRNAs in ABU and UPEC infected *G. mellonella* larvae.** The network diagram generated with Cytoscape shows mRNAs encoding methyltransferases targeted by miRNAs in ABU and UPEC infected larvae. Target mRNAs are represented in different colors to maintain contrast with the miRNAs. Connecting lines (edges) are used to indicate miRNA targets identified in this study by consulting Gene Ontology terms related to *G. mellonella* genome and transcriptome sequences. ABU<UPEC means miRNA upregulation in UPEC compared to ABU, ABU>UPEC means miRNA upregulation in ABU compared to UPEC, and ABU=UPEC means no difference in miRNA expression.

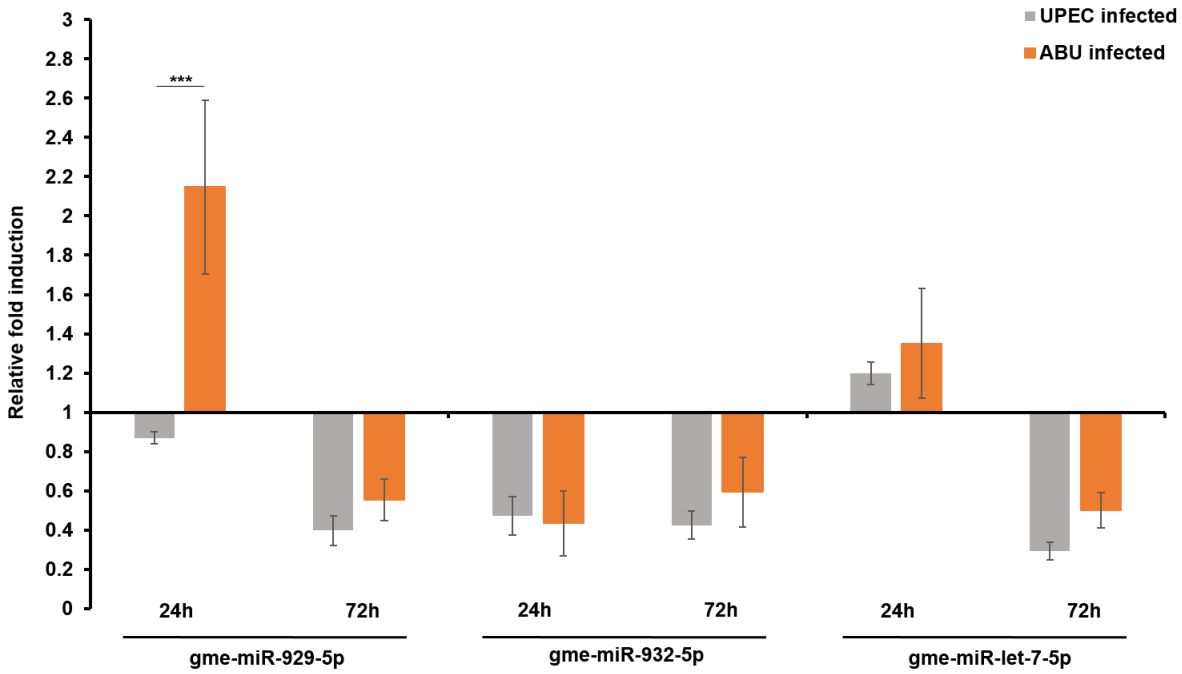


Figure S 13 **Confirmation of the differential expression of selected miRNAs by RT PCR in *G. mellonella* larvae infected with ABU and UPEC strains, and in mock-injected controls.** The conserved miRNA sequences were obtained from small RNA sequencing by RT-PCR to confirm the differential expression of gme-miR-929-5p, gme-miR-932-5p, gme-miR-let-7-5p. The relative fold differences indicated for the miRNAs are normalized against mock-injected control and gme-miR-133 as the internal reference control (****$p < 0.0001$).

Table S3 **Novel microRNAs in *G. mellonella*, identified by smallRNA sequencing.**

Novel miRNAs	As per miRBase record	Missmatch/ antistrand	Sequences (5'-3')
gme-new-107-3p			TATTCTATGATTACCATGCTGC
gme-new-107-5p			TGCACGGTTATTGTAGAATACA
gme-new-108-3p			CGTTTGTTCTACTAGTATACA
gme-new-108-5p			TACTACTAATAGGACACAACGGA
gme-new-110-3p			TAGAGGATAATGTAATAATGTA
gme-new-110-5p			CATTATTACATTATCCTCTACT
gme-new-112-3p			CTGGCTTATGTTAATAGCCACAG
gme-new-112-5p			GGAGCCATTAACAAAGCTGGAT
gme-new-114-3p			CACTCGTCTAGAGGGTAAGGACGG
gme-new-114-5p;>gme-new-161-5p			ATTCTATCTCACGGAGTCGC
gme-new-161-3p			TACTTGTCTAGAGGGCGTGA
gme-new-116-3p			GTCATTCATGTCTACTTACAA
gme-new-116-5p;>gme-new-118-5p			ATAAGTAGGTCATGTCTGACTT
gme-new-118-3p			AAGTCACTCGTATCTACTTAAAA
gme-new-117-3p			TGACGTCGTTGATGATAAATAAG
gme-new-117-5p			TTTTTCACCAGCGAAGTCAGA
gme-new-119-3p			TCTGACTTGCTTGGTGACTACC
gme-new-119-5p			TAATCACTAGCGAGTTAAATA

gme-new-120-3p			TTCTCCTGGAATACACTTATGA
gme-new-120-5p			ATAAGTGTAATCTAGAGTAAGT
gme-new-122-3p;>gme-new-127-3p			TAATGCCTTCCTCTAACTACA
gme-new-122-5p;>gme-new-127-5p			TAGTTAGAGGAAGGCATTGAA
gme-new-123-3p;>gme-new-124-3p;>gme-new-126-3p;>gme-new-133-3p			TACCCAGAGATCGCTGCACT
gme-new-123-5p;>gme-new-124-5p;>gme-new-126-5p;>gme-new-133-5p			CGCAGCGTATCATAGGTGGT
gme-new-128-3p			TCTAACTTGGTTGGTAGATGCC
gme-new-128-5p			CAACCACTAACAAAGTTAAACG
gme-new-134-3p			AATATACATGTAGTACTGTACT
gme-new-134-5p			CACAGTACTACATGTATATTCC
gme-new-135-3p			CTCATTATTTCGTCTGAAAAGAAA
gme-new-135-5p			AAGTTTTCCGTGACGATATAAGGGGGCTCC
gme-new-136-3p			TCAATTTGGTTTTAATCTGAAT
gme-new-136-5p			TGTGAATTAAGAATACT
gme-new-137-3p;>gme-new-139-3p			TCTCTACTGTGTTGAAATACATA
gme-new-139-5p			TGTATTTTAACACAGAAGAGATG
gme-new-137-5p			TGTATTTCAACACAGTAGAGATG
gme-new-140-3p;>gme-new-145-3p			TTTTCCATGAAGTCGCCATCC
gme-new-140-5p;>gme-new-145-5p			ATGGCGACTTCATGGAAAATA
gme-new-141-3p;>gme-new-142-3p;>gme-new-143-3p			AGTGTTCATCCACTTTTACACT
gme-new-141-5p;>gme-new-142-5p;>gme-new-143-5p			TGCAAAAGTCGGTGGACACTTT
gme-new-144-3p			TACCCTAAGGTCGTGTCAGTCC
gme-new-144-5p			ACTGACCCGACTTGGATTATA
gme-new-146-3p			CAGCGCGCACGCCGCCAGCAGC
gme-new-146-5p			GCGCGGCGGCGGGCGCGCGCG
gme-new-147-3p;>gme-new-148-3p			ATTTGGTTCTCTCTAATAGCAAT
gme-new-147-5p;>gme-new-148-5p			CGCTGTTAGTCTGACCAAATAT
gme-new-149-3p			TGTGTGAGAATTCACGATTGAAG
gme-new-149-5p			TCAATCGTGAATACTCACACACA
gme-new-150-3p			CAATACTATGGTCAAGTGAGAA
gme-new-150-5p			GTCACTTGACCATAGTATTGAG
gme-new-151-3p			TGAGCTATCTGCATCGACTGATT
gme-new-151-5p			TCAGTCGATGTAGGTGCCAGA
gme-new-153-3p;>gme-new-154-3p			TGTACACTGGTTCTACTGTAGG
gme-new-153-5p			TACAGTAGAACCAGCGTACATT
gme-new-154-5p			TACAGTAGAACCAGCATACATT
gme-new-155-3p			TTCGAAGTTCGTTAACGAAAACT
gme-new-155-5p			TCTTCGTTAACGAACTTTGAACA
gme-new-156-3p			TGAGGATGAAAACGGATCGATT
gme-new-156-5p			TCGATCCGTTTTCATTCTCACC
gme-new-157-3p			CAACGTTTAATACCACTTTGGA
gme-new-157-5p			CAATGTGGTATTAAACGTTGTA

gme-new-158-3p			CCGGCGCCGCTGGGGGAGG
gme-new-158-5p			TCGGATCGCGGCGACGGGG
gme-new-160-3p			TTGTGTATGCTGCTGGATGGGGT
gme-new-160-5p			GTCATTCAGCCTGCCAGCATTGCT
gme-new-52-3p			TAAATTCGTGTCATCGGACGTT
gme-new-52-5p			AGTTTGATTTCACGAATACGGC
gme-new-61-3p			AAATTAGTGGACGTTAAAGGAA
gme-new-61-5p			CCTTTAACGCCCACTAATTT
gme-new-63-3p			GATCCGTCAGTATTCTACGACT
gme-new-63-5p			TCATAGAATACTGAAGGATCTA
gme-new-70-3p			GATCAATTTGTTTTCTTTACAGC
gme-new-70-5p			GTAAGTAAAACACGTTGATTCA
gme-new-72-3p			TTATTAAAATTACTGTA
gme-new-72-5p			GCACAATAGTTTTTGATACA
gme-new-78-3p;>gme-new-79-3p;>gme-new-94-5p			TGTCGTGAAGTAGTAATGCTAC
gme-new-94-3p			AGCATTACTACTTTCACGACAGA
gme-new-78-5p;>gme-new-79-5p			TAGCATTACTACTTCACGAC
gme-new-80-3p			GGTACGTTCAGATGTTTGTGGTA
gme-new-80-5p			GTCCACATCCAGGCAGTGCCCTC
gme-new-81-3p			TAAGAGGCGACTAAGGGAATTT
gme-new-81-5p			ATTCACTTAGTCGCCTCCTATG
gme-new-82-3p			TATTTGGCTATTACTAACAGTAGA
gme-new-82-5p			TCTGTTCGTATTGTCAAGTATA
gme-new-88-3p			TATTCTATAATTTCTTTGCTGCT
gme-new-88-5p			TGTAGGGTTATTGTAGAATACA
gme-new-89-3p			CTCTCTCGCCTCCATGACCTGA
gme-new-89-5p			AGGCCATGCAGGCGAAGGAAT
gme-new-90-3p			CAAGTGATTGTTGGCACATAGT
gme-new-90-5p			TATGTGCCAACAATCACTTGTT
gme-new-93-3p			TTCTCGTATCGCACTGTCTACT
gme-new-93-5p			TAGACAGTGCGATACGAGAACG
gme-new-96-3p			ACGCTGGGAACCGAGTGACTAT
gme-new-96-5p			GCTCACTCGGTTCCCGGCGTGA
gme-new-97-3p			TAGCCGGTTCCGTTGCTGCATT
gme-new-97-5p			TGTAGTAACCGAACGGGCTGCC
gme-new-1-3p;>gme-new-2-3p	gme-miR-8-3p		TAATACTGTCAGGTAAAGATGTCG
gme-new-1-5p;>gme-new-2-5p	gme-miR-8-5p		CATCTTACCGGGCAGCATTAGA
gme-new-100-3p	gme-miR-2796-3p		GTAGGCCGGCGGAAACTACTTGC
gme-new-100-5p	gme-miR-2796-5p		AGGGGTTTCTTTCGGCCTTCAG
gme-new-101-3p	gme-miR-2766-3p		TCAGTCTTGTCGAATGGTG
gme-new-101-5p	gme-miR-2766-5p		CACCGTTCGTCTCGACTGG
gme-new-102-3p	gme-miR-283-3p		ACTACCAGATGGTATACAG
gme-new-102-5p	gme-miR-283-5p		AAATATCAGCTGGTAATTCT

gme-new-103-3p	gme-miR-307-3p		TCACAACCTCCTTGAGTGAGC
gme-new-103-5p	gme-miR-307-5p		ACTCACTCAACCTGGGTGTG
gme-new-104-3p	gme-miR-2765-3p		CAACGGAGGCAGAGTTCCCGTT
gme-new-104-5p	gme-miR-2765-5p		TGGTAACTCCACCACCGTTGGC
gme-new-105-3p	gme-iab-4-3p		TATACCTTCAGTATACGTAACA
gme-new-105-5p	gme-iab-4-5p		ACGTATACTGAATGTATCCTGA
gme-new-106-3p	gme-miR-184-3p		TGGACGGAGAACTGATAAGG
gme-new-106-5p	gme-miR-184-5p		CCTTGTCATTCTTCTTGCCCAGT
gme-new-109-3p	gme-miR-981-3p		TTCGTTGTCGACGAAACCTGCA
gme-new-109-5p	gme-miR-981-5p		CGGGTTTCGGGACGATGTAACC
gme-new-11-3p	gme-miR-14-3p		TCAGTCTTTTTCTCTCTCCTAT
gme-new-11-5p	gme-miR-14-5p		CGGGGGGAGAAATTGACTTGACT
gme-new-111-3p	gme-miR-124-3p		TAAGGCACGCGGTGAATGCC
gme-new-111-5p	gme-miR-124-5p		CGTTCACTGCCGGAGCCATTATG
gme-new-113-3p;>gme-new-115-3p	gme-miR-3338-3p		CATGTACTTACTTTGTTTGTTCT
gme-new-113-5p;>gme-new-115-5p	gme-miR-3338-5p		AGCAAACACAGTAGTGTGCAATGCT
gme-new-12-3p;>gme-new-13-3p	gme-miR-281-3p		CTGTCATGGAGTTGCTCTCTTT
gme-new-12-5p;>gme-new-13-5p	gme-miR-281-5p		AAGAGAGCTATCCGTCGACAGT
gme-new-121-3p	gme-miR-137-3p		TATTGCTTGAGAATACACGTAG
gme-new-121-5p	gme-miR-137-5p		ACGCGTATTCTTGGGGAATTAAC
gme-new-125-3p	gme-miR-285-3p		TAGCACCATTCGAATTCAGT
gme-new-125-5p	gme-miR-285-5p		TGCATTCGAGTGTGTGGATAGA
gme-new-129-3p	gme-miR-7-3p		GGAATCACTAATCTGCCTAC
gme-new-129-5p	gme-miR-7-5p		TGGAAGACTAGTGATTTTGT
gme-new-130-3p	gme-miR-1176-3p		TGAGATTCAACTCCTCCAACTT
gme-new-130-5p	gme-miR-1176-5p		AAGTGGAGGTGTGATCTCTTC
gme-new-14-3p	gme-miR-6094-3p		TATTCGAGACCTCTGCTGATCCT
gme-new-14-5p	gme-miR-6094-5p		ATCAGCGGTGGTCTGGGGTACC
gme-new-15-3p;>gme-new-16-3p	gme-miR-279c-3p		TGACTAGATTTTCACTTATCCT
gme-new-15-5p;>gme-new-16-5p	gme-miR-279c-5p		GATAAGTGAATTTCTAGTTCA
gme-new-152-3p	gme-miR-998-3p		TAGCACCATGGGATTCAGCTCA
gme-new-152-5p	gme-miR-998-5p		AGCTGCATCCCGTGCGGCTCCA
gme-new-17-3p;>gme-new-18-3p	gme-miR-9c-3p		TAAAGTTATGGTACCGAAGTTA
gme-new-17-5p;>gme-new-18-5p	gme-miR-9c-5p		TCTTTGGTATCCTAGCTGTAGG
gme-new-19-3p	gme-miR-275-3p		TCAGGTACCTGAAGTAGCGCGC
gme-new-19-5p	gme-miR-275-5p		CGCGCTACTCCGGCGCCAGGGCT
gme-new-20-3p	gme-miR-2755-3p		CACCCTGTCAGACCATACTTGTT
gme-new-20-5p	gme-miR-2755-5p		CAAGGTGGCCTAGCAGCGTGTT
gme-new-21-3p	gme-miR-278-3p		TCGGTGGGATCTTCGTCCGTTT
gme-new-21-5p	gme-miR-278-5p		CCGGATGAAATTCCCGTCGGCC
gme-new-22-3p	gme-miR-9a-3p		ATAAAGCTAGGTTACCGGAGTTA
gme-new-22-5p	gme-miR-9a-5p		TCTTTGGTTATCTAGCTGTATGA
gme-new-23-3p;>gme-new-24-3p	gme-miR-970-3p		TCATAAGACACACGCGGCTCT

gme-new-23-5p;>gme-new-24-5p	gme-miR-970-5p		AGCCTTGCGTGTGCTCTTATTGGTA
gme-new-25-3p	gme-miR-279b-3p		TGACTAGATCTACACTCATCGA
gme-new-25-5p	gme-miR-279b-5p		CATGGGTGTAAGTCTGGTAACACA
gme-new-26-3p	gme-miR-282-3p		GACATAGCCTGATAGAGGTTACG
gme-new-26-5p	gme-miR-282-5p		TAGCCTCTCCTAGGCTTTGTCT
gme-new-27-3p	gme-miR-317-3p		TGAACACAGCTGGTGGTATCTCAGT
gme-new-27-5p	gme-miR-317-5p		CGGGTGCCACGCTGTGCTCTCT
gme-new-28-3p	gme-miR-279-3p		TGACTAGATCCACACTCATCCA
gme-new-28-5p	gme-miR-279-5p		GATGAGTGGAGGTTTAGTGCATG
gme-new-29-3p	gme-miR-2c-3p		TATCACAGCCAGCTTTGTTGACT
gme-new-29-5p	gme-miR-2c-5p		CCAACAAAGTGGTTGTGGCGTG
gme-new-3-3p	gme-miR-276-3p		TAGGAACTTCATACCGTGCTCT
gme-new-3-5p	gme-miR-276-5p		AGCGAGGTATAGAGTTCCTAC
gme-new-30-3p;>gme-new-31-3p	gme-miR-306-3p		CAGAGCCGCCTCGTGCCTCACA
gme-new-30-5p;>gme-new-31-5p	gme-miR-306-5p		TCAGGTACTAGGTGACTCTGA
gme-new-32-3p;>gme-new-36-3p	gme-miR-2a-3p		TCACAGCCAGCTTTGATGAGC
gme-new-32-5p	gme-miR-2a-1-5p		CCATCAAAGTTGGTTTGTCATA
gme-new-33-3p	gme-miR-190-3p		CCCAGGAATCAAACATATTAC
gme-new-33-5p	gme-miR-190-5p		AGATATGTTTGATATTCTTGGTTG
gme-new-34-3p;>gme-new-35-3p	gme-miR-277-3p		TAAATGCACTATCTGGTACGACA
gme-new-34-5p;>gme-new-35-5p	gme-miR-277-5p		CGTGCCAGGAGTGCGTTTACA
gme-new-36-5p	gme-miR-2a-2-5p		CTCACAAAGTGGTTGTCATATG
gme-new-37-3p	gme-miR-274-3p		CTCGTTTTGACGATCGCAAAATG
gme-new-37-5p	gme-miR-274-5p		TTTGTGACCGTCACTAACGGGC
gme-new-38-3p;>gme-new-39-3p	gme-miR-34-3p		AGCCACTAACGCCACTGCCCCT
gme-new-38-5p;>gme-new-39-5p	gme-miR-34-5p		TGGCAGTGTGGTTAGCTGGTTGT
gme-new-4-3p	gme-bantam-3p		TGAGATCATTGTGAAAGCTGATT
gme-new-4-5p	gme-bantam-5p		TGGTTTTCATAATGATTTGACAGA
gme-new-40-3p	gme-miR-252-3p		CCTGCTGCTTAAGTGCTTATC
gme-new-40-5p	gme-miR-252-5p		CTAAGTACTAGTGCCGCAGGAG
gme-new-43-3p	gme-miR-279d-3p		TGACTAGATCCATACTCGTCTGC
gme-new-43-5p	gme-miR-279d-5p		GGCGAGTTTGCTTCTGGTTCATG
gme-new-44-3p	gme-miR-316-3p		ACAGCAAAGTGAAAAGGTCTCCT
gme-new-44-5p	gme-miR-316-5p		TGTCTTTTTCCGCTTTGCTGCT
gme-new-46-3p	gme-miR-133-3p		TTGGTCCCCTTCAACCAGCTGT
gme-new-46-5p	gme-miR-133-5p		AGCTGGTTGATTTCGGGTCAAAT
gme-new-48-3p	gme-miR-71-3p		TCTCACTACCTTGTCTTTCATG
gme-new-48-5p	gme-miR-71-5p		TGAAAGACATGGGTAGTGAGATG
gme-new-49-3p	gme-miR-13b-3p		TATCACAGCCATTTTTGACGAGTT
gme-new-49-5p	gme-miR-13b-5p		TCGTAAAAATGGCCGTGCCAT
gme-new-5-3p	gme-miR-10-3p		CAAATTCGGTTCTAGAGAGGTTT
gme-new-5-5p	gme-miR-10-5p		TACCCTGTAGATCCGAATTTGT
gme-new-50-3p;>gme-new-51-3p	gme-miR-79-3p		ATAAAGCTAGATTACCAAAGC

gme-new-50-5p;>gme-new-51-5p	gme-miR-13b-5p		CTTTGGCGATTTAGCTCCATGA
gme-new-53-3p	gme-miR-13a-3p		TATCACAGCCACTTTGATGTGGT
gme-new-53-5p	gme-miR-13a-5p		CTGTCAAAGCGGCGGTGAAATG
gme-new-54-3p	gme-miR-375-3p		TTTGTTCGCCCCGGCTCGTGTCG
gme-new-54-5p	gme-miR-375-5p		ACCCGAGCGGTATGAGCAAACT
gme-new-55-3p;>gme-new-56-3p	gme-miR-308-3p		AATCACAGGATAATACTGCGA
gme-new-55-5p;>gme-new-56-5p	gme-miR-308-5p		CGTGGTATTATACCTGTGAATGT
gme-new-57-3p	gme-miR-965-3p		TAAGCGTATAGCTTTTCCCCT
gme-new-57-5p	gme-miR-965-5p		CGGGAGAAGTTGTATCGCTTTATG
gme-new-58-3p;>gme-new-59-3p	gme-miR-87-3p		GTGAGCAAACTTTCAGGTGTGT
gme-new-58-5p;>gme-new-59-5p	gme-miR-87-5p		GGGCCTGAATCGTTGCTCTTACC
gme-new-6-3p	gme-miR-1-3p		TGGAATGTAAAGAAGTATGGAG
gme-new-6-5p	gme-miR-1-5p		CCGTGCTTCCTTACTTCCCATA
gme-new-60-3p	gme-miR-92b-3p		AATTGCACCAATCCCGGCCTGC
gme-new-60-5p	gme-miR-92b-5p		AGGACGCGATTTGGTGTAAACCTTG
gme-new-62-3p	gme-miR-263b-3p		CGTGAATTTCCTGATGCCTCA
gme-new-62-5p	gme-miR-263b-5p		CTTGGCACTGGGAGAATTCACAG
gme-new-64-3p	gme-miR-92a-3p		TATTGCACCAGTCCCGGCCTAT
gme-new-64-5p	gme-miR-92a-5p		GGGCGGTGACTGGTGCTATATT
gme-new-66-3p	gme-miR-927-3p		CAAAGCGTTTGGATTCTAAAAC
gme-new-66-5p	gme-miR-927-5p		TTTAGAATTCCTACGCTTTACC
gme-new-67-3p	gme-miR-993-3p		GAAGCTCGTCTCTACAGGTATCT
gme-new-67-5p	gme-miR-993-5p		TACCCTGTAGATCCGGGCTTTTGT
gme-new-68-3p;>gme-new-69-3p	gme-miR-932-3p		TGCAAGCAGTGCGGAAGTGAGG
gme-new-68-5p;>gme-new-69-5p	gme-miR-932-5p		TCAATTCCGTAGTGCATTGCAGT
gme-new-7-3p	gme-miR-11-3p		CATCACAGTCAGAGTTCTAGCT
gme-new-7-5p	gme-miR-11-5p		CTAGAACTCCGGCTGTGACTTGT
gme-new-71-3p	gme-miR-10485-3p		CCCTGGACGGACAGGCCGCTC
gme-new-71-5p	gme-miR-10485-5p		ACAGTCTACCCGGACAGGCCG
gme-new-73-3p	gme-miR-929-3p		CTCCCTAATCGAGTCAGGTTGA
gme-new-73-5p	gme-miR-929-5p		AAATTGACTCTAGTAGGGAGT
gme-new-74-3p;>gme-new-75-3p	gme-miR-33-3p		CAATATCACTACAAGGCAAATC
gme-new-74-5p;>gme-new-75-5p	gme-miR-33-5p		GTGCATTGTAGTTGCATTGC
gme-new-76-3p;>gme-new-77-3p	gme-miR-193-3p		TACTGGCCTGCTAAGTCCCAAG
gme-new-76-5p;>gme-new-77-5p	gme-miR-193-5p		AGGGACTTAGTGGTCTGGTGTG
gme-new-8-3p	gme-miR-750-3p		CCAGATCTATCTTTCCAGCTCA
gme-new-8-5p	gme-miR-750-5p		AGTTGGACAGGGGATCTTGACA
gme-new-83-3p;>gme-new-84-3p	gme-miR-927b-3p		CAAAACGAACTGATTCTTTAGT
gme-new-83-5p;>gme-new-84-5p	gme-miR-927b-5p		TTTAGAATCAGTACGCTTTGTC
gme-new-85-3p	gme-miR-1000-3p		CTGCTGCGTCCGGACAAGTTGG
gme-new-85-5p	gme-miR-1000-5p		ATATTGTCCTGTCACAGCAGTA
gme-new-86-3p;>gme-new-87-3p	gme-miR-3327-3p		TATGTAACGTTTTTGTTGTCTT
gme-new-86-5p;>gme-new-87-5p	gme-miR-3327-5p		AACAACAGGAATGTTATGTAC

gme-new-9-3p;>gme-new-10-3p	gme-miR-263a-3p		CGTGGTCTCTTAGTGGCATC
gme-new-9-5p;>gme-new-10-5p	gme-miR-263a-5p		AATGGCACTGGAAGAATTCACGG
gme-new-91-3p	gme-miR-31-1-3p		GGCTGTGTCACTTCGAGCCAGC
gme-new-91-5p;>gme-new-159-3p	gme-miR-31-5p		AGGCAAGAAGTCGGCATAGCT
gme-new-159-5p	gme-miR-31-2-5p		CTCCGCCGATTTCACCGTTGCAAAC
gme-new-92-3p	gme-miR-8506-3p		TTAGGTGTGAGGGTCACAGC
gme-new-92-5p	gme-miR-8506-5p		GCTGTGACTCTCACACTTAGTA
gme-new-95-3p;>gme-new-99-3p	gme-let-7-3p		CTGTATAGCCTGCTAACTTTCC
gme-new-95-5p;>gme-new-99-5p	gme-let-7-5p		TGAGGTAGTAGGTTGTATAG
gme-new-98-3p	gme-miR-2768-3p		ATTGGTTAAGATATTGCATCGT
gme-new-98-5p	gme-miR-2768-5p		GGTGCAATATTTTGACCAATTT
gme-new-131-3p;>gme-new-132-3p	gme-miR-8-3p	anti strand	TAATGCTGCCCGGTAAGATG
gme-new-131-5p;>gme-new-132-5p	gme-miR-8-5p	anti strand	CATCTTTACCTGACAGTATTAGA
gme-new-138-3p	gme-miR-2763-3p	2MM	ATATTATGCACATTACTATGGAT
gme-new-138-5p	gme-miR-2763-5p	2MM	CCAAAGTAGTGAGCATAAGTTATCA
gme-new-41-3p	gme-miR-745-3p	ident	CAGCTGCCTAGCGAAGGGCAAC
gme-new-41-5p	gme-miR-745-5p	1MM	CGGCTCATCGTATGGCAGTTTGCT
gme-new-42-3p	gme-miR-8522-3p	ident	CTTTGCCGAAGGTTCTGAGGTA
gme-new-42-5p	gme-miR-8522-5p	1MM	CCTCACAACCTTCGGCAAACGA
gme-new-45-3p	gme-miR-2756-3p	1MM	CCCCTACGCTGCTACATTGTAT
gme-new-45-5p	gme-miR-2756-5p	1MM	ACCCTGTAGCTGCTTAGGGGC
gme-new-47-3p	gme-miR-2767-5p	3MM	AGCCGTTCGAGATTCACTTAGA
gme-new-47-5p	gme-miR-2767-3p	ident	CAAGTAAATCTCGTGCGGCTTG
gme-new-65-3p	gme-miR-988-3p	1MM	CCCCTTGTTACAAACCTCACTT
gme-new-65-5p	gme-miR-988-5p	2MM	GTGTGCGTTGTGGCAAAGGAGAT

Table S4 **UPEC and ABU specific miRNAs**

UPEC specific	**ABU specific**
gme-new-10-3p	gme-new-78-5p
(5’-CGTGGTCTCTTAGTGGCATC-3’)	(5’-TAGCATTACTACTTCACGAC-3’)
gme-new-70-5p	gme-new-135-5p
(5’-GTAAGTAAAACACGTTGATTCA-3’)	(5’-CTCATTATTTCGTCTGAAAAGAAA-3’)
gme-new-87-5p	gme-new-136-5p
(5’-AACAACAGGAATGTTATGTAC-3’)	(5’-TGTGAATTAAGAATACT-3’)
gme-new-88-5p	gme-new-137-5p
(5’-TGTAGGGTTATTGTAGAATACA-3’)	(5’-TGTATTTCAACACAGTAGAGATG-3’)
gme-new-88-3p	gme-new-161-3p
(5’-TATTCTATAATTTCTTTGCTGCT-3’)	(5’-TACTTGTCTAGAGGGCGTGA-3’)
gme-new-89-5p	
(5’-AGGCCATGCAGGCGAAGGAAT-3’)	
gme-miR-92b-5p	
(5’-AGGACGCGATTTGGTGTAAACCTTG-3’)	
gme-new-117-5p	
(5’- TTTTTCACCAGCGAAGTCAGA-3’)	

gme-new-127-3p (5’-TAATGCCTTCCTCTAACTACA-3’) gme-new-145-3p (5’-TTTTCCATGAAGTCGCCATCC-3’) gme-new-148-3p (5’-ATTTGGTTCTCTCTAATAGCAAT-3’) gme-new-150-3p (5’-CAATACTATGGTCAAGTGAGAA-3’) gme-new-156-3p (5’-TGAGGATGAAAACGGATCGATT-3’) gme-new-157-3p (5’-CAACGTTTAATACCACTTTGGA-3’) gme-miR-274-3p (5’-CTCGTTTTGACGATCGCAAAATG-3’) gme-miR-307-5p (5’-ACTCACTCAACCTGGGTGTG-3’) gme-miR-929-3p (5’-CTCCCTAATCGAGTCAGGTTGA-3’) gme-miR-2756b-3p (5’-CCCCTACGCTGCTACATTGTAT-3’)	

Table S5 **MiRNA and mRNA sequences**

miRNA	**Target mRNA**
gme-new-70-3p 5’-GATCAATTTGTTTTCTTTACAGC-3’	Gene 1 5’- ATGGGCTCCCTGCCTCGCATGTCGGTGGTGAGCGGCGGGCGGGCGGTGGTA CCGGCTGCGCCGCTGCCGACCCACCTGGCACGGCTGGCCGCCAGCCAGGAG ACGGCTCTTGTATATTCAGATGAAACCTGCAATGCTCGTATCAGTTATGTGGA AATGGAAGCTCAGACGAATGCGATAGCGAAAGCGTTGTCTAAACGCGCCAG ACCGACTGGCGCCAACAGGGACGGTGACTATGTGATAGCTGTGTGCATGCA ACCTACACACAATACAGTGTTAACATTATTAGCGACTTGGAAGACGGGGGCA GCGTATGTGCCGATGGAACCCAGCTTCCCACAAGCGAGGATATCACACATAC TGCAAGACGCTGAGCCGGCCTTAGTTATTTATGATGATAGTGCAAATCCAGC TATGTTCGCCGGCAGTGGCATCCCATCGGTGTCTTTTGAAGAATTGATTCAAG AAGCCAGTGGACTATCCGCCGAGGAACTCAAAAGTCCGGAAGTGTTGGCTC ACGCCGGAACAGACAGTATTGCTATCGTACTGTACACATCTGGAAGCACGGG TGTGCCGAAAGGTGTCCGTCTCCCTTATTCGGCTATATGCAACCGACTCTGGT GGCAATTCCGGACCTTCCCCTATTCCAATACAGAGAAGACCTGTGTTTGGAA GACGGCTTTGACCTTTGTGGATTCAGTCTGTGAAATTTGGGGCCCTCTTCTAC ACGGTAGGACTCTGTTAATCCTATCGAAAGAGACGACTAGGGATCCACAAAA ATTGGTACGAGTTTTAGCTGAGAATCAGGTTGAAAGATTAGTACTAGTGCCG ACTCTCCTTCGTTCTATCCTAATGTATCTATCCCTCACACCTTCCGAAAGGCCT TTACAGTACCTAAAGCTTTGGGTCTGCTCAGGAGAGACCCTAAGTAAAGAGT TAACAACCCAGTTCTTCCGATACTTCGGTGACAATGGTGGATACAAACTGGC GAACTTCTACGGCAGTACCGAAGTTATGGGAGACGTCACTTACTATGTACTG GAGAAACTTAATCAGTTAGACGTATATAATACTGTTCCTATCGGTTCCCCATT AGACAACTGTGCAGTGTACCTCTTAGACGAGGAGATGAATCCAGCCCGTGA GACTGAACCAGGGGAAGTGTGGGTCGCTGGACATAACCTAGCAGCTGGGTA

	CGTGGGAGCTCAGGGTGCTGATAAATTCTGTGACAACCCGCATGCTGACCAT CCAGACTTTAGCCGTCTATATCGCACTGGCGACTTTGGGATCCTACAGAAGG GAGTGATCCTTTATGCTGGACGTACCGACTCTCAAGTCAAAATTAGAGGTCA CAGGGTCGATTTGCAAGAGGTTGATCGCGCCGTGACTGCAGTACCTGGCATT GAAAAATGCGTTGTACTATGCTACGGCTTGGAACGAGGGAACCCTGAAATCT TAGCATTTGTCACCATAGAACCAAGTGCACGCATCGCCGCGCATCATATTGA GGCCAGCTTGAAGAACTCTTTAACTAGCTACATGATACCGCAGGTAATTGTG ATAGAAAGCATCCCCTTGCTTGTGAATGGGAAAGTGGACCGGCAAGCATTGC TGAAGATGTACGAGAATACAAACAATAATGACGACTCTGCAATCGCATTAGA CTTTGACTATACAGGCGTAGATGAACAAGATAAAGAAGCGGCTAAAGTTCTC TTCGAAACAGTGGGAGAAGTGCTGGGGCGCGCCGCCAGAGGAACTTTGTCC GTAAGAGCCGGCTTCTACGAACTGGGAGGGAATTCTCTCAACTCCATCTACA CGATCACCAAGTTGAGAGAAAAAGGATACTATATCGAAATCAGCGAGTTCCT GGGCGCAGCCAACCTTGGTGAGGTACTGGCCAACATGAGTACGAGCCCCGA CAGTGGAGCGGACAGCAACGAGCCCAAGTTCGCAGCCGAGCTCATGAAGGA TGAGGACAAACAGCAGGTTATCGACATGATAGTATCATCATTCTACGAGAAA GCGGAGTTGGAACAGTTCCTGAAGCACGAGATTGACACCATGGATTACGCA CACTGCATAGACGCGTGCTGGGCTGCACTGCTACGAGCCAGGCTCAGCGTG GTGCTGAGAGATGGTACAAACACGCCGGTGGCAGTGGCTCTAAACTTCGAC GCTCGGGACGAACCTGAAATTGAATTGACCGGTGGACTTGCCAAGATAATG GGATTTTTAGAGTTTGTTGAGGGCTCTGTGAGAGATACGTTATTACCGGAAG GCAAAGGCACGATTCTCCACTCGTTCATGATGGCCACGAACGCGGAACTGTC TCCGCGAGACAACGTCGCCGCGATACGGGCGTTAGAGCATGCCACCATGAG GATTGCGAGGGACAGGCGGTTCAAAGGCGTGTTCACCACAAACACTAGTCC ACTTACACAGCAATTAGGAACCGATGTGCTCGGCTTCCAAACACTCCTCGACT ATCAGATCAATCAATACGTAGATTCCAATGGAGACAGGACCTTCGGGAAAGC ACCAGATGAC3’
gme-new-40-3p 5’-CCTGCTGCTTAAGTGCTTATC-3’	Gene 2 5’- ATGGTTCCTGCACCGAGCTGCCCGTCCACTTGGGACTACTGGACGACATCGC CGTCCGACTACGTGGAGCTGACATGCCTCCTGCCAAATACTATTTACATTCCT CTGAGAGTGAGTTGGGATGCTACTCTTCAGGATGTGAAAGAGGAGTTATGG GAGAAGGCGGCACATTATCCCTTATTTGGGGTGATGCATGAAATGTCAGGTT ATGTGTTCCAATTTGTCAACTCCCTGGCAGTTCTTGAGGAGGTGGATGATGA GAACAAAAGGCTGCGAGATATCAAACCAGTATGTGGAGTGCTTATGATCATA AAGAGGTCTGTGAAACCAGGAGAATATCTGCTGAACACACAAATAAGCCATT TAATTGGAAAAGGTTTAAACGAATTTGATAGTCTAAGAAGTAATGAAGTGAA TGATTTTCGGACACAAATGCACAATTTAGCTGAAGAGAGTTTATTAAGACGA ATGAAAAGTGATTGGCAAGAGAAACTGCGATATCACTACCCACCAAGACTG GCTGATCAACCTATACCTACCACACTCAAGAACCAACTCAATAGAAATAGTTT TATGCTTGTCACTAGATTTGCTAATTCTGAGTTTTCATCATCTCTTTGCGTACC ATTCACACAAACGCCGCAGCAGCACATCGAGATCATTCTCAGAAAGCAAGCG AAGTCACTGAACATACGCGGCGAGCACCCACACAATTATGTTTTGAAAGTGT GCGGCCGCGAGGAGTATTTATTCGGGGATTACCCGCTCATACAGTTCAAATA CGTCCAGGAGATGCTGTCTCGGGACTCCGTGCCCCAAGTCATGACTGTCAGT GTGGACAAATTGAGATTTTTAAATGCCGACCCCCAGCAGTACTATATACGTG AACAGAGGCGGCAGACCGCTGACTCGAATACAATGAAAAGACGGAAGAATG AACTCTCCTGGGATATAGACAAGTTATACTCTTGCATGGTGCTGAGTGTTGG CGGGTTGAACGTGGACCCTAATCGTGTTGTTGAGGTAATTTGTCAAGCTGGC GTATTTCACGGCGGTAAGCCTTTATGCGAGGCGCAGAAGACTCGAGCGGCG GCGGTGTCGTCGGAGGGCGTTGCGCAGTGGCAGCAAGAACTCAAGTTCCCG

	CTCAAGGTCTACAACATACCGAGAATGGCGAGACTATGCTTTGGGATATATG AAATTGAAATTAATAAGACTAAGAATAAGAAGAAGGGGAAGGACTCCGGCA AGGACTCAATAAACCGGCTGGCGTGGGCGAACACAATGATATTCGACTACA AGGATCAACTGAGGACGGATAAAGTGTCATTTTTCATGTCAACTCACGTGGC GGATGAAACGCAGGGCGACGATCAGCTGCTGCATCCCCTGGGGACAGTGTT CTCCAATTCGAATACAGACTCGTGTAGCGCCGTTTTACATGTGCAGTTCTCAA ATTATGATTGCCAATATCCTATTGTATTTCCGAAACAAGAAATGGTAAAAGCT TATGCGGAGCGCATTGAGAATGGGTCACCCGAAGTGATGTCGCGCCTTAAG AGAGATTTCGAGAAGCTTCGAGCGACCGCCGAAAAGGATCCCATGTATGAA ATGCATGAGCAAGACAAGAAGAATATATGGGCGTTGAGAAACGACTTCCGC TCACTGGCCCCGTGGTTGCTGCCGCGCCTGCTGTGCTGCGTGGAGTGGGGC GAGCGGGCGGAGGCGGCGGCGGTGGCGCGGCTGCTGGACGACTGGCCCAT CTCGCTGCCGGTCGAGTCCGCGCTCGAGTTACTCGACTACGCGTACGCGGAC GCCACCGTCCGGAGCTTCGCTGTTCGGTGTCTGCAGAAGATCAGTGACGAAG ACCTCCTGTTATATCTATTACAACTGGTTCAAGCGTTGAAACACGAGCCCTAC CTCATGTGTGATTTGTCGGTGTTTTTGTTACAACGCGCGTTTAAAAACATGAT CATTGGTCACTATCTTTTCTGGCATTTAAGATCGGAGATGCACATGCCGTCGG TGTCGGTGCGGTTTGGTCTACTGTTGGAGGCGTACTGCCGCGGCTGCCAAGA CCACATCAGTATTCTGCTGCGGCAAATCGCATGTCTCGACAAACTCAAGTGG GTGAGCCAGAATGTCCGCAAGAAGAAGGAAATATCAAAAGCGCGGGCAGC GCTGCAGCAGAATTTGCAACAGACGCATTGCATCGAAACGCTCTGCGACTTC GTATCGCCACTTAATCCGAGCTACCGCTGCAAACGGATACAGCCGGAGAAAT GCCGCGTTATGGACAGCAAGATGCGCCCTCTGATGGTGGACTTCGAGAATA GTGACCCGTTCGGCTCAGATATCCGGATCATACTGAAGATCGGCGACGACCT TCGTCAGGATATGTTCACATTGCAAATGCTCAGGATAATGGACAGGCTGTGG AAGAGTCACGGTTATGACTTTAGGTTAAGTCCATACAATTGTATTTCAATGGA GAACGAAGTGGGTATGATCGAGGTGGTGGAGGACGCGGAAACGGTTGCTA ATATACAAAAACAACCCGCCATGTTCCAAGCCGCCTCCACAATGTACAAAGG GACTTTGCTACAATGGCTAAAGAAGCAGACAGAGGACGAGTGCGGGCGTCC CAACGAGGCGGCGTTCAACAAAGCAGTGGACGAGTTCACAATGAGCTGCGC CGGCTACTGTGTCGCCACCTACGTCCTCGGTATCGCGGACCGACACCCGGAC AACATTATGGTCAAGAAAAGTGGACAGCTATTCCACATAGACTTCGGCCATT TCCTCGGCCACTTCAAACAGAAGTATGGGTTTAAGCGCGAACGTGTGCCGTT CGTTTTAACGCACGATTTTATACACGTGATCAACAAGGGGCAGCGGGGCTCG GGCGATAATGAGCCCATCGACTTCAAGATATTCAAAGAGCACTGTGATACGG CATTCAGAATACTCCGAAAGCATGGCCACCTCATCTTGTCTCTCTTCTCGATG ATGATCTCTACCGGCCTTCGGGAGCTGAGCTCCGAGAAGGACTTGCAGTACC TTAGAGAAACCCTCGTAATGGATTTATCCGAGGAGAAAGCCATGGAGCACTT CAGGTCGAAGTTCAGCGAGGCGATGAAGAACTCATGGACGACATCGCTCAA CTGGGCGTTCCATAATATCGACAAAAACAACTGA-3’
gme-new-138-3p 5’-ATATTATGCACATTACTATGGAT-3’	Gene 3 5’- ATGAAATTTACAATTGTCATAAATATTACAAATAAGGATGCAGCAACTACATC TGAGTTATTTAATAAAAAACCAGATAAGTCCGTAACCATTTTTGATTTGGTTA TACACTTTTATAATTTTGACTATCAGAATAGTACTACACATGTACCTGAACACT CATACATAGCAATGACATCTGGCAGCACTGGAGAACCAAAGCACATACAAGT GCCTGTGCAATGTATTCAACCAAATATAGATGATTTAACTAAATTATTTAATA TTACTGCTGATGATATAATATATTTCTCTACACCACTAACATTTGACCCATCCA TGATAGAGATACTGCTCGCCTGTATGAATGGAGCCTCTCTACTTATTGCACCT GAAAAAGCAGACATATTATTCCCCAACAACAAAGAGAATTCTGTGACATTTT GGCAACTTACACCATCACGATTTTTTCAGCATTCAAATTCTGATATCAAGAAT

	AAAATATTAAGTGCAAATTCAACATTGAAAATACTAGCTTTAGGTGGTGAGC CATTAAATGGTGTGAAGAGACTAAAGGAATTGAAAGATTGGGATAATAAAA CTAGAATATTCACATTGTATGGAGTAACTGAAATGTCATGCTGGGCTTGTGTT GCTGAGTTAGATCTTAACAAAATACTAACTGACAAAGAAATACCATTAGGTA ACTGTCTGTCAGAAACAGAATTACATGTGGAATCAAATGATGACAATAAAGA ATGTGGAAAAATTATTTTAGTAAGCAAAACAAGAAAGTGTGTTATATTAAAC AAAACCATTGGAAATGAAGATGAAAATTCTTTAAAATTTATTGACACTGGAG ATTTAGGTGAAGTAAGAAATGGCACTGTGTATTATCGGGGCCGCAAAGATG ATATTATTAAAAGATTTGGACACAAAATTAATTTACAGTTTATTGAATCAACT ATAATGCAATGTCCGAGTGTGAAAACAAGCTCCTGTATCTGGCTCCCAAAAT CATTACTTCTGATTGCCTATTTCTCATCAGAAACACTTAGCAGCCAAGAGTTG TCCAACTTTTTGAAATGTAAACTTGATGATAAGCACTGGCCAGATAAAATAAT TAGAGTTGACAATTTACCAACAAATCCTCATGGGAAGATATCTAAATTGATAT TATCTAAAATGTATGAAAAAACAATGAACACACCACAGACATTAGATTCCTTA AAAGTGAGTTTCTTAAAGGAACTTCAAGCTGTAATGGGTCAACACTTCACTTA TGATCAAATTAAAAGTAAAGACTTCTTTGCCATTGGTGGCACATCTTTTCTAG CGATATCTATGTGTAATAAGCTTTCACTACTTTGTCCAAAATTTGGTAAACTAA TTCTCCCTTACTTGATGTCTCAAAAAAATACTATAGATGATATTATGCAATTGG CATATAAAGAAATACATGTTGATGAAATAAAAGTTAAGAAAAAAATTAAAAG GTCACGGTCAGATGCAGGTGGTTATGTAGAGAGTCAGTCTTACAAAAGAACT AACACAAAAAGTCTGACAAATCCTGTAAAATTCATTGTGTTATGGTCATATGA TACTGGAAAATGTGTAGATGCTTCTCCATCTTTATTTCAAATAGGATTCAATTT ATATGTGACAGTCGGGAGTCATTCAGGGAAGATTATAGTCATGGATGCGATA TCTGGAATATTGCAGGGAATGGTAACAGTAAAATCACGTGTTGAAGCATCTG TATTTTGTTACCACAAGAGAGACATGTCGCCGTGCGGTGTGGTTGGCACTTA CGATGGCACAGTGGTATGCTTCCAATTAGAAACATGCAAAGAGTTGTGGAG AATCAACATTGGATCAATGATAAAAAGTAAAGCAACATGTTGCAATGATCTA CTCTACATTGCTTCCTATGATGGAAAGATAAGATGTATAGATATTGCGATAG GGGTAATCAAAGAGACTATATATGTGGCAGATCAAGGTATATCAGCTGATCT AGTACTTGCTAAAAACAAGTATGTGTTAACCAGTACGCTGTCTGGTGTGTGT GCAAGTATACATACTCTAACCAATACTGTGGCTTGGCGCTGCACACTGAGTA GTCCAGTATTTGCAAGTCCTGTGCTTTACGATGACGACAAGTATGTGGTATTC GCGGAAGTTAATGGTGAAATACATTGCAGGACCGTTGAAAAGGGTATTAAG ATATGGAATTATCAAGGAGCAAGAGGTAACATTTTTTCCTCCCTCTACATAAA AGAAGTTGATAAACTGAAATGGCAAATGGTTTTTGGCTGTCACGATAACAAA GTTTACAGCATTAATATCAAGAATTTCCAACCTAGCTTGAATTGGAAAGCACA ACTCACATCACCTGTATACTCCACTCCATGTGGTCTAAGTGACAAATTAATAC TTGCTGCCTCTAATAATGGCAGGTTATGCGTTATAGATGAAGAAAACGGGAT AATATTGGCAGAACATCATCTGCCAAATGAAACATTTTCATCACCAGCAGTTT ATGGAGATTACATATTCATTGGTTGCAGGAATGACCACCTTTACTCTTTGAAA TATATTTTAAATTTATAA-3’
gme-new-4-5p 5’-TGGTTTTCATAATGATTTGACAGA-3’	Gene 4 5’- TTGAAGCCTACAGAAATGTTTATAGACGAAAATGAGACGCCGGAGACATGG TGGGTGTCGGCGGTGCTCAAGACCATCAAGGCGGTGACGCTGGTCTTCGAC ATCCTCACCTTCCCCATTCACTTAATAGTGCAGCGACCATGGAGGAAGCGCG CCCTCTCGCGCCGGATTAAGGCGCGCATCACGCAGTCGTCGCCGGGCTGCGT GACGGTGCGGTCGGTGTCGTCGCCGGGCGAGCTGCACGTGCGGCTGGTGCG CGACGGCGTGCTCACCATGGAGAGCATGCTGCGCGCCGCCGCCCAGCGCTG GGGCTCGCGCCGCTGCCTCGGCACGCGCACCGTGCTCAGCGAGGAGGACGA GCCGCAGCCCAACGGCAGGCTCTTCAAGAAGTTCAAAATGGGCGATTACGT

	GTGGCGCACGTACACGGAGGTGGAGGCGGAGGCGCGGCAGTTCGCGAGCG GGCTACGCGCGCTGGGCTGCGCGCCGCGCGCCAACATCGCCATGTTCGCCG AGACGCGCGCCGAGTGGATGCTCGCCGCGCACGGCTGCTTCAAGCTCAGCA TCCCAGTGGTAACAATCTACGCGACGCTCGGCGACGAGGCCATAGCACACG GCATCAACGAGACGGAGGTGTCCACTGTCATCACCACTCACGACCTGCTGCC CAAGTTCAAGAAGATCCTCGCCAAGACGCCCAAAGTGGACACCATCATATAC ATGGAGGACCAGCTGCAGACCATCGACCGGGAGGGCTACAAGCCCGGCATC AGGATCGTTGGCTACAAGGAGGTCATACAGAAAGGAATAAACGCCAGCTTT GAGGCGGTGCCGCCGGCGCCGACGGACACGGCCATCATCATGTACACGTCG GGCTCGACGGGCGTGCCCAAGGGCGTCATCCTGTCGCACCGCAACATGCTG GCCACGCTCAAGGCGTTCGCGGACGTCGTGCCCATATACGAGGACGACATG CTCATGGGCTTCCTGCCGCTCGCGCACGTCTTCGAACTATTGGCCGAGAGTCT CTGCATCATTGGTGGTGTCCCCATCGGGTACTCGACGCCGCTGACGATGCTG GACTCCTCCAGCAAGATCATGAAGGGCACCAGCGGAGATGCCACCGTCCTCA AGCCCACCTGCATGACCACAGTGCCGTTGATAATGGACCGCATCAGCAAGGG CATCACGGACAAGGTGTCTCGCAGCGGGCCGTTCGCGAGCGCGTTCTTCCGC TGGGCGTACTCGTACAAGCAGACGTGGATGCGGCGCGGATACGACACGCCC ATACTTAACAGGATTATGTTCAGCAAGATCCTGGGCCTGCTGGGCGGGCGGC TGCGGCTGCTGCTGGCGGGCGGCGCGCCGCTGGCGCCGGACACGCACCAGC AGCTGCGCATCTGCCTGTGCTGCGACGTGGTGGCGGGCTACGGGCTCACCG AGACCACGTCGGCCGCCACCGTCATGGACGCGCACGACCGCTCCACCGGCC GCGTGGGCGCGCCCTCGCCCGGCACCGCGCTGCGCCTGCTCGACTGGGCCG AGGGCGGCTACCGCGTCGCCAACAGGCCCTTCCCGCAGGGCGAGATCGTGA TCGGAGGTGATTGCGTAGCGGAGGGCTACTACAAGAATCCAGAGAAGACCC GGGAGGAATTCATCGAGGAGGACGGCATTCGCTGGTTCAGGTCTGGGGACA TCGGCGAACTGCATCACGACGGCTGCATCAAGATCATCGACCGCAAGAAGG ACCTAGTGAAGCTGCAGGCCGGCGAGTACGTGTCCCTGGGCAAGGTGGAGG CGGAGCTGAAGACGTGCCCCATCGTGGAGAACATCTGCGTGTACGGCGACA GCTCCAAGACGTACACCGTGGCGCTGGTGGTGCCCAACCCGCGGCACCTGG CCGAGCTGGCCGCGCGGCTCGGCCTGCCCGACCGGGACTTCGACCAGCTCT GCCACAACACCGCCGTCGAGAAGGCCGTCGTCAAGGAGCTCGCCGACCACG CTAGGAAGTGTGGGCTGGAGAAATTCGAGGTTCCTGCTGCAGTGAAGCTGT GCACGGAGGTGTGGTCCCCCGACATGGGGCTAGTTACCGCCGCCTTCAAGAT CAAGCGGAAAGACATCCAGGAGCGGTACAAGGAAGACATCAAACGAATGTA CGCCTCCTGA-3’
gme-new-135-5p 5’-CTCATTATTTCGTCTGAAAAGAAA-3’	Gene 5 5’- CCATGTTCATTGCCAGTCGTTTGTGGTTGCGCGCGAACATACGTCGTCCATTT TATATTTAAGTTTTTTTTTTATATATTTTTTTTAATTTCCGCCACGATGGCGTCA GCCGTGATTAAGTTTAGTGCCATAATGATGCTAATCGGCGTGTGTGCAGCTG ATGTCTCTGAACTCCCCGAAGTCAAGGCAGCAGCGGCCGCAGCCCCGCCCGT GACTGAGGTTTGTCTTGGCTGCATCTGCCAAGCCGTGTCGGGGTGCAAGCAA GGAACTCAATGCGAAGGAGACCATTGTGGTCTATTCCACATCACTTGGCCAT ACTGGGCTGATGCCGGGAAACCAACGATTAATGGACTCTCACCTGACGATCC TAATGCGTACCCCAGTTGCACCAATGACCCGTACTGCGCTGCGCAGACCGTA CAAGGCTACATGAAGAGATACGCTCAGGACTGCAACGGCGATGGTCAGATA AACTGCTACGATTACATGGCCATCCACAAGAAGGGAGGGTACGGGTGCAGT GGGGAGCTGCCCTTCAATTACGTCAACACTTTCAACCAGTGTGTGGCAATTGT TGCTTCTCAG-3’
gme-new-121-3p 5’-TATTGCTTGAGAATACACGTAG-3’	Gene 6

	5’- ATGAGTGACATGTTGAATTATGGTTGCACTAGGCTTGAAATTGGTGTTCAAT CGGTTTATGAGGACATTGCTCGTGATACTAATAGGGGACACACAGTAAAAGC TGTCTGTGAGAATTTTAATTTGGCCAAGGATGCTGGATATAAGATTGTTGCG CATATGATGCCTGATTTACCCAATGTGGATTTTGAACGTGATGTGGAACAATT TATTGAATTCTTTGAAAATCCCGCATTTCGAGCTGACGGCCTTAAGATATACC CGACTTTAGTTATTAGAGGTACTGGTCTATATGAACTATGGAAGACTGGGCG ATACAGAAGTTATCCTCCATCAACTTTGGTTGATTTGATTGCAAAAATACTGG CATTGGTACCACCATGGACTAGGGTCTACAGGGTCCAACGTGACATTCCCAT GCCCCTTGTTTCATCCGGAGTGGAACATGGCAACTTGAGGGAGTTGGCGCTA GCTCGCATGGCTGACTTGGGTACAGATTGCAGAGACGTGAGGACCAGGGAA GTAGGGATACAGGAGATACATAACAGAGTTAGGCCATACGAGGTAGAGTTA ATAAGACGAGATTATGTTGCCAATGGTGGATGGGAGACATTCCTAGCATACG AGGATCCAGATCAGGATATATTGGTAGGCCTCCTGAGGCTCAGGAAATGTGC CTCGGACACTTACCGGCCAGAATTGAAACCTGGTCCAAATTCAAATTTCAAGC AATGTAGTATAGTCAGAGAACTGCATGTTTATGGATCAGTTGTACCTGTGAA TGCCCGCGACCCAACAAAATTCCAACACCAAGGCTTTGGGATGTTGCTAATG GAAGAAGCGGAGAGGATAGCTAAAGAGGAA-3’
gme-new-70-3p 5’- GATCAATTTGTTTTCTTTACAGC-3’	Gene 7 5’- ATGAGTGACATGTTGAATTATGGTTGCACTAGGCTTGAAATTGGTGTTCAAT CGGTTTATGAGGACATTGCTCGTGATACTAATAGGGGACACACAGTAAAAGC TGTCTGTGAGAATTTTAATTTGGCCAAGGATGCTGGATATAAGATTGTTGCG CATATGATGCCTGATTTACCCAATGTGGATTTTGAACGTGATGTGGAACAATT TATTGAATTCTTTGAAAATCCCGCATTTCGAGCTGACGGCCTTAAGATATACC CGACTTTAGTTATTAGAGGTACTGGTCTATATGAACTATGGAAGACTGGGCG ATACAGAAGTTATCCTCCATCAACTTTGGTTGATTTGATTGCAAAAATACTGG CATTGGTACCACCATGGACTAGGGTCTACAGGGTCCAACGTGACATTCCCAT GCCCCTTGTTTCATCCGGAGTGGAACATGGCAACTTGAGGGAGTTGGCGCTA GCTCGCATGGCTGACTTGGGTACAGATTGCAGAGACGTGAGGACCAGGGAA GTAGGGATACAGGAGATACATAACAGAGTTAGGCCATACGAGGTAGAGTTA ATAAGACGAGATTATGTTGCCAATGGTGGATGGGAGACATTCCTAGCATACG AGGATCCAGATCAGGATATATTGGTAGGCCTCCTGAGGCTCAGGAAATGTGC CTCGGACACTTACCGGCCAGAATTGAAACCTGGTCCAAATTCAAATTTCAAGC AATGTAGTATAGTCAGAGAACTGCATGTTTATGGATCAGTTGTACCTGTGAA TGCCCGCGACCCAACAAAATTCCAACACCAAGGCTTTGGGATGTTGCTAATG GAAGAAGCGGAGAGGATAGCTAAAGAGGAA-3’
gme-new-147-3p 5’- ATTTGGTTCTCTCTAATAGCAAT-3’	Gene 8 5’- ATGGCTATGCAACCACACAGTAAGAAAAGAGTCTGCTACTATTATGATAGTG ATATTGGGAATTACTATTATGGACAAGGTCATCCCATGAAACCTCATCGCATA CGCATGACACATAATTTACTTCTAAATTACGGCTTGTACAGAAAAATGGAAAT TTATAGACCACATAAGGCGACAGCTGATGAGATGACAAAGTTTCATTCGGAT GACTACATTCGCTTCCTGCGCTCCATCAGACCGGATAATATGTCTGAATATAA CAAACAAATGCAGAGATTCAATGTTGGTGAAGACTGTCCAGTGTTTGACGGC TTGTATGAGTTTTGTCAATTGTCTGCTGGAGGTTCTGTTGCCGCTGCTGTTAA ATTAAATAAACAGGCATCAGAAATCTGCATCAACTGGGGTGGTGGCCTTCAC CACGCAAAGAAGTCGGAAGCATCAGGTTTCTGCTATGTAAATGATATTGTAC TTGGCATATTGGAGTTACTGAAGTATCATCAAAGAGTACTGTATATTGATATT GACGTACACCACGGTGATGGGGTCGAGGAAGCTTTTTACACCACAGACAGA GTAATGACTGTCTCCTTCCATAAGTATGGGGAATACTTCCCTGGAACAGGTG

	ATCTCCGGGATATCGGTGCCGGCAAGGGCAAATACTATGCTGTGAATATTCC CTTGCGTGACGGTATGGACGATGAATCATACGAGTCAATTTTCGTACCCATCA TATCCAAAGTCATGGAGACCTTCCAGCCGAGTGCGGTGGTACTTCAATGTGG AGCTGATTCACTTACTGGTGATAGATTGGGCTGTTTTAATTTAACAGTCAGAG GTCACGGCCGATGTGTGGAGTTAGTAAAGCGATTTGGCCTACCTTTCCTTCTT GTGGGAGGTGGAGGATATACGATCCGCAACGTGTCACGGTGCTGGACATAC GAAACATCAGTAGCACTAGGCGTTGAGATCGCTAACGAGCTGCCCTACAATG ACTACTTTGAATACTTCGGTCCGGACTTCAAACTGCACATATCACCCAGCAAT ATGTCCAATCAGAACACTCCAGAGTACTTGGAGAAGATTAAAAATAGGCTCT TTGAGAATCTACGGATGTTGCCGCATGCACCTGGTGTACAGGTACAAGCCAT TCCAGAAGATGCGGTAAATGATGAGTCAGAAGACGAAGATAAGATTGACAA AGATGAAAGGCTGCCACAAAGTGAAAAGGATAAACGCATTACGGGTGACGG CGAGTTGTCAGACTCTGAAGACGAGGGTGAAGGCGGGCGGCGCGACAACC GCTCGTACCGCGCGCCGCAACGTAAGAGGCCCCGTCTCGACAAGGACGGCT CGCAGATCAAAGACGAAATTAAAACTGAAGATATAAAAGATGACGTGAAGA ATGTGAGCAGTGTAGAGGAACCAAAAAAGGAAGTGCCACCCAATCCCTGA- 3’
gme-new-160-5p 5’- GTCATTCAGCCTGCCAGCATTGCT-3’	Gene 9 5’- ATGCCAAATATTTTGTTAAACTTTGTTGTTCTTTCATTCGCTGATCTCAAGAAA TTTCATTACTACTACTGGTTTGCATTTCCCACTCCTAGTCAGCCCACAGTCTAT TTAAATGAAAAAGCAAAATTGATAACAACTCAGTTCAGTTTCGATCAACTTCA ATTATTAGTCCAAGGGTACAATTCATTAGATTCAAGTCAAAAATCATTCTTCA TTGTTACCAAAAGTGATGATAAGTTGTCAGTAATTCCATTGTCAAATATTTTA CAACCAAACTGCAATGAATTAAGTCTTGATTTATCAGATGTGTATTTTGTGTT TGCTGATCCAAGTAACACCGATAATCCTGGGTGGCCTTTAAGAATTTTCTTAG CAGCTTTACTTGATCATTGCACAAATCTTTCTGGGAAAAATATTCAAGTCATT GGTTTGAGGTGCAATGTTAAATATGAAATTGCAAACAGTCTTGTTTATTCTAT CTACATTTCTCAGGATATTCAATCAGCAGAGAATGCAGGTTGGGTTGGATGG GAGAGAAATGATAAAGGCAATTTTGGCCCTAAACTGGCTAATATGTCAGCGT CAATGGATCCTGTAGTTTTAGCTGATACGTCATCAGATTTGAACATCAAACTA ATGAAATGGCGTTTAGTACCTAATATAGATGTAGAGGTGATGAAGAGTACTA AATGTTTATTATTGGGTGCTGGTACGCTTGGATGTCATGTGGCTCGTAATCTG TTGGCTTGGGGATTTCGTCATATAACATTTATTGACAACGGTAAAGTGTCGTA CTCAAATCCGACGCGACAAGTACTCTTCAATTATCAAGATTGTCTAAATGGAG GACGTAAGAAAGCTGAAGCAGCGGCCGACAATCTTAAAAATATACTACCTAC TGTAACTGCCAAGGGTTTAGTAGCTCATATACCAATGCCTGGTCATCCTATTG GAGAATCTCTTAAGGCCGAAACTATATCCAATATCAAGACTATTACTGAAGCT ATCGCAGATCACGATGTGATATTTTTGTTACTCGATACTAGAGAGGCAAGAT GGTTGCCTACACTCGTCGCAGCACATTATGGAAAGATAGTAATCAATGCAGC GTTGGGTTTCGATAGTTACTTAGTGATGAGACATGGAATAGGTGGCACGCCA TCAGAGGGCGCGACACTCGTCAATGCCACACACATAGCAGGTGGACAACTG GGTTGCTATTTCTGTAATGACGTCACTGCTCCAGGGAATTCCCTTAAAGATCG CACTCTTGACCAACAATGCACCGTAACACGGCCAGGCGTGGCGGCAATAGCC GGGGCGCTTGCTGTGGAAATTCTAGTTGGATTGTTGCAGCATCCTCTAAGGG TAGAAGCACCAGCGATATATAACTTGAATCAAGAAATTGACACCATATCATC AGATATGCAAGGTGTTTTGGGCCCCGTACCACACTCTATCAGAGGATTTCTCC ATTCATACCAAACAGTAGTGCCAACATGCGGGAAATTCAAACAGTGCATAGC CTGTTCTGATAATGTTTTAAACAAGTACAAAGAAGCGGGAATGGAGTTCTTG TTCAATGTATTCAATAGTGGAAAATATTTGGAAGAAGTAACTGGATTAACAG

	AATTACAATTGTCTGCAGAAATGACCGATATATTGACATTTTCTGATGACGAC GATAATGAATAA-3’
gme-new-106-5p 5’- CCTTGTCATTCTTCTTGCCCAGT-3’	Gene 10 5’- ATGATGTCGACTTCAACAGATACTTTATCACAAAACTTAAAAAGTGATGAAA ATGGTGAATCTAAATCAAATGACACTGAACACCTTGAAGGAAATAAACTCTT GGTTAAAAGAAGAAATTGCCAAGACGAATTGGCGCCACCTTCAAAGTCTTTT AAAAATGATGAACATAGTACCGTAGTAGCAGCTCATTACAATCATCTTGAAG AAAAAGGACTAAAGGAAAGGTTCAAATCTCCCATATTTTACGTACGAAATTTT AATAACTGGGTAAAAAGTGTGCTCATTCAGGAGTACACAGATAAAGTGAGA GAAAAAGACTATGGTAAACCTATCATGGTGCTTGATATATGTTGCGGTAAAG GAGGAGACCTCAGCAAGTGGCAAAAAGCGCGTGTTGAAAAAGTGATATTTG CTGATATAGCGGATGTGTCTGTTCAGCAATGTAAAATTCGCTATGATGATTTA CATAAAAGATGTGGCAGACTTTACTCTGCTGAATTCATTGCAGCTGATTGTAC AAAAGAGACTCTGAGAGATAAATACTCAGACCCATCAATAAATTTTGATCTT GTAAGCTGTCAGTTTGGACTACATTATAGTTTTGAAAGTCTAGGCCAAGCTA GAAGAATGCTCACTAACATAACAGAGTGTCTCCGTCCAGGTGGATATTTTTTT GGTACTATTCCAGATGCATATGAAATTATTTCCCGAACTAAAAAGTCAGCTGA TGGGTCTTTTGGAAATAGAATCTACAATATTAAGTTATTATTTGATTCCAAAA CAGGTTATCCATTGTTTGGTGCAAAATATGATTTCCATTTAGAGGGAGTAGTA GATTGTCCTGAGTTCTTAGTTAATTTCGAACTATTTGTTAAACTAGCTGCTGA GTATGGACTTGAATTAGTATACAAAGCTAGGTTTTCAGATTTCTTCAAAGATC ATTCAGATAACTATAAACAGTTACTGCATAGAATCATTTGTTTTGAAAGTTAT CCAGCACCGCCTGGTAAAGAACTCATTGGAGATGAGGCAGAATATGAGCAT GCAAAACAATTTTGGGAAAATATGGAGAAGAAAAATGAACATGATCATATTG GAACAATGAGTATGTGTGAATGGGAAGTAGCCACTATCTATATGGCATTTGC ATTTAAGAAACAGAAGTCCACTTGGGATTCAAATGGAAAACCAGTATACAAA TTGCCTCAAGATGAGGAGAAAGCAGAGTGA-3’
gme-new-147-3p 5’- ATTTGGTTCTCTCTAATAGCAAT-3’	Gene 11 5’- ATGGCTATGCAACCACACAGTAAGAAAAGAGTCTGCTACTATTATGATAGTG ATATTGGGAATTACTATTATGGACAAGGTCATCCCATGAAACCTCATCGCATA CGCATGACACATAATTTACTTCTAAATTACGGCTTGTACAGAAAAATGGAAAT TTATAGACCACATAAGGCGACAGCTGATGAGATGACAAAGTTTCATTCGGAT GACTACATTCGCTTCCTGCGCTCCATCAGACCGGATAATATGTCTGAATATAA CAAACAAATGCAGAGATTCAATGTTGGTGAAGACTGTCCAGTGTTTGACGGC TTGTATGAGTTTTGTCAATTGTCTGCTGGAGGTTCTGTTGCCGCTGCTGTTAA ATTAAATAAACAGGCATCAGAAATCTGCATCAACTGGGGTGGTGGCCTTCAC CACGCAAAGAAGTCGGAAGCATCAGGTTTCTGCTATGTAAATGATATTGTAC TTGGCATATTGGAGTTACTGAAGTATCATCAAAGAGTACTGTATATTGATATT GACGTACACCACGGTGATGGGGTCGAGGAAGCTTTTTACACCACAGACAGA GTAATGACTGTCTCCTTCCATAAGTATGGGGAATACTTCCCTGGAACAGGTG ATCTCCGGGATATCGGTGCCGGCAAGGGCAAATACTATGCTGTGAATATTCC CTTGCGTGACGGTATGGACGATGAATCATACGAGTCAATTTTCGTACCCATCA TATCCAAAGTCATGGAGACCTTCCAGCCGAGTGCGGTGGTACTTCAATGTGG AGCTGATTCACTTACTGGTGATAGATTGGGCTGTTTTAATTTAACAGTCAGAG GTCACGGCCGATGTGTGGAGTTAGTAAAGCGATTTGGCCTACCTTTCCTTCTT GTGGGAGGTGGAGGATATACGATCCGCAACGTGTCACGGTGCTGGACATAC GAAACATCAGTAGCACTAGGCGTTGAGATCGCTAACGAGCTGCCCTACAATG ACTACTTTGAATACTTCGGTCCGGACTTCAAACTGCACATATCACCCAGCAAT ATGTCCAATCAGAACACTCCAGAGTACTTGGAGAAGATTAAAAATAGGCTCT

	TTGAGAATCTACGGATGTTGCCGCATGCACCTGGTGTACAGGTACAAGCCAT TCCAGAAGATGCGGTAAATGATGAGTCAGAAGACGAAGATAAGATTGACAA AGATGAAAGGCTGCCACAAAGTGAAAAGGATAAACGCATTACGGGTGACGG CGAGTTGTCAGACTCTGAAGACGAGGGTGAAGGCGGGCGGCGCGACAACC GCTCGTACCGCGCGCCGCAACGTAAGAGGCCCCGTCTCGACAAGGACGGCT CGCAGATCAAAGACGAAATTAAAACTGAAGATATAAAAGATGACGTGAAGA ATGTGAGCAGTGTAGAGGAACCAAAAAAGGAAGTGCCACCCAATCCCTGA- 3’
gme-new-135-3p 5’- AAGTTTTCCGTGACGATATAAGGGGGCTCC-3’	Gene 12 5’- TTGAAGCCTACAGAAATGTTTATAGACGAAAATGAGACGCCGGAGACATGG TGGGTGTCGGCGGTGCTCAAGACCATCAAGGCGGTGACGCTGGTCTTCGAC ATCCTCACCTTCCCCATTCACTTAATAGTGCAGCGACCATGGAGGAAGCGCG CCCTCTCGCGCCGGATTAAGGCGCGCATCACGCAGTCGTCGCCGGGCTGCGT GACGGTGCGGTCGGTGTCGTCGCCGGGCGAGCTGCACGTGCGGCTGGTGCG CGACGGCGTGCTCACCATGGAGAGCATGCTGCGCGCCGCCGCCCAGCGCTG GGGCTCGCGCCGCTGCCTCGGCACGCGCACCGTGCTCAGCGAGGAGGACGA GCCGCAGCCCAACGGCAGGCTCTTCAAGAAGTTCAAAATGGGCGATTACGT GTGGCGCACGTACACGGAGGTGGAGGCGGAGGCGCGGCAGTTCGCGAGCG GGCTACGCGCGCTGGGCTGCGCGCCGCGCGCCAACATCGCCATGTTCGCCG AGACGCGCGCCGAGTGGATGCTCGCCGCGCACGGCTGCTTCAAGCTCAGCA TCCCAGTGGTAACAATCTACGCGACGCTCGGCGACGAGGCCATAGCACACG GCATCAACGAGACGGAGGTGTCCACTGTCATCACCACTCACGACCTGCTGCC CAAGTTCAAGAAGATCCTCGCCAAGACGCCCAAAGTGGACACCATCATATAC ATGGAGGACCAGCTGCAGACCATCGACCGGGAGGGCTACAAGCCCGGCATC AGGATCGTTGGCTACAAGGAGGTCATACAGAAAGGAATAAACGCCAGCTTT GAGGCGGTGCCGCCGGCGCCGACGGACACGGCCATCATCATGTACACGTCG GGCTCGACGGGCGTGCCCAAGGGCGTCATCCTGTCGCACCGCAACATGCTG GCCACGCTCAAGGCGTTCGCGGACGTCGTGCCCATATACGAGGACGACATG CTCATGGGCTTCCTGCCGCTCGCGCACGTCTTCGAACTATTGGCCGAGAGTCT CTGCATCATTGGTGGTGTCCCCATCGGGTACTCGACGCCGCTGACGATGCTG GACTCCTCCAGCAAGATCATGAAGGGCACCAGCGGAGATGCCACCGTCCTCA AGCCCACCTGCATGACCACAGTGCCGTTGATAATGGACCGCATCAGCAAGGG CATCACGGACAAGGTGTCTCGCAGCGGGCCGTTCGCGAGCGCGTTCTTCCGC TGGGCGTACTCGTACAAGCAGACGTGGATGCGGCGCGGATACGACACGCCC ATACTTAACAGGATTATGTTCAGCAAGATCCTGGGCCTGCTGGGCGGGCGGC TGCGGCTGCTGCTGGCGGGCGGCGCGCCGCTGGCGCCGGACACGCACCAGC AGCTGCGCATCTGCCTGTGCTGCGACGTGGTGGCGGGCTACGGGCTCACCG AGACCACGTCGGCCGCCACCGTCATGGACGCGCACGACCGCTCCACCGGCC GCGTGGGCGCGCCCTCGCCCGGCACCGCGCTGCGCCTGCTCGACTGGGCCG AGGGCGGCTACCGCGTCGCCAACAGGCCCTTCCCGCAGGGCGAGATCGTGA TCGGAGGTGATTGCGTAGCGGAGGGCTACTACAAGAATCCAGAGAAGACCC GGGAGGAATTCATCGAGGAGGACGGCATTCGCTGGTTCAGGTCTGGGGACA TCGGCGAACTGCATCACGACGGCTGCATCAAGATCATCGACCGCAAGAAGG ACCTAGTGAAGCTGCAGGCCGGCGAGTACGTGTCCCTGGGCAAGGTGGAGG CGGAGCTGAAGACGTGCCCCATCGTGGAGAACATCTGCGTGTACGGCGACA GCTCCAAGACGTACACCGTGGCGCTGGTGGTGCCCAACCCGCGGCACCTGG CCGAGCTGGCCGCGCGGCTCGGCCTGCCCGACCGGGACTTCGACCAGCTCT GCCACAACACCGCCGTCGAGAAGGCCGTCGTCAAGGAGCTCGCCGACCACG CTAGGAAGTGTGGGCTGGAGAAATTCGAGGTTCCTGCTGCAGTGAAGCTGT GCACGGAGGTGTGGTCCCCCGACATGGGGCTAGTTACCGCCGCCTTCAAGAT

	CAAGCGGAAAGACATCCAGGAGCGGTACAAGGAAGACATCAAACGAATGTA CGCCTCCTGA-3’
gme-new-135-5p 5’- CTCATTATTTCGTCTGAAAAGAAA-3’	Gene 13 5’- ATGAAGAATCTGGAGATGACTCAGATCCGGGAAATAGTGGATTGCATGTATC CAGTAGAATATGCCGCCGGTAGCCTCATCATCAAAGAAGGAGATGTTGGCA GTATTGTATATGTTATGGAAGAGGGAAGAGTGGAAGTGTCTAGAGAGAACA AATACCTCAGCACAATGGCACCGGGCAAGGTGTTTGGTGAACTAGCCATTCT TTACAACTGCAAGAGAACGGCCACAATAAAAGCAGCAACTGATTGTCGGTTG TGGGCCATTGAACGTCAATGCTTCCAGACTATTATGATGAGAACTGGACTCA TAAGACAAGCGGAATACACTGATTTCTTGAAGAGTGTGCCGATCTTCAAAGA CCTTCCCGAAGACACGCTTATCAAAATTTCTGATGTTTTGGAAGAGACACATT ATCAGAACGGTGACTACATTATCAGGCAAGGAGCGCGTGGTGACACGTTCTT CATCATTTCCAAAGGACAGGTAAAAGTGACCCAGAAGCAACCAAACAGTAAC GATGAGAAATTCATTAGAACACTAACGAAAGGCGATTTCTTCGGAGAAAAA GCGTTACAAGGAGATGACCTTCGAACAGCCAACATCATCTGTGACTCACCAG AAGGTTGTACATGCCTTGTAATTGATCGGGAGACCTTCAACCAACTCATTTCG ACCCTAGATGAGATACGTACCAAATATAAAGACGAAGGCGATAGTAGACAG AGATTAAATGAAGAATTTGCCAATTTGCGTTTATCAGATCTTCGTATCATAGC CACCCTCGGTATCGGCGGTTTCGGAAGAGTGGAACTTGTGCAAATACAAGG AGATCCGAGTCGATCGTTCGCCTTGAAGCAGATGAAGAAAGCCCAAATCGTT GAAACGAGACAGCAGCAACATATTATGTCAGAAAAGGAGATAATGTCAGAA ATGAACTGCGAATTCATAGTGAAGCTATTTAAGACATTTAAAGATCGCAAAT ACTTGTATATGTTGATGGAGACATGCCTCGGAGGAGAGTTGTGGACTATTTT AAGAGACAGAGGCCAGTTTGATGATGCCACAACAAGGTTCTATACCGCTTGC GTTGTAGAAGCCTTCCATTATCTACATTCTAGGAATATCATTTACAGGGATCT CAAACCGGAAAACTTATTATTAGACTCCAAAGGTTATGTGAAATTAGTCGATT TCGGTTTCTCCAAGAAGCTGCAAGCGAGCCGTAAGACTTGGACATTCTGTGG TACTCCTGAGTATGTTGCACCCGAAGTCATTATGAATAGAGGTCATGATATCA GCGCAGACTATTGGTCATTAGGTGTGCTAATGTTCGAGCTGCTGACAGGATC ACCTCCATTCACCGGAGCTGACCCAATGAAAATTTATAACAAGATTCTCAAGG GTATTGATGCCGTGGAATTCCCCCGATCGATCACCAGAAACGCAGCTAATCT CATAAAGAAATTGTGTCGTGACAACCCTGCCGAACGGCTTGGGTATCAGAGA GGAGGCATCACTGAGATACAGAAACACAAATGGTTTGACGGCTTCAACTGG GAGGGCCTGGCTCAGCGCACCTTAGAGCCGCCGATCACACCAGTCGTAAAGT CCGCTGTCGATACGCACAACTTCGACCAGTATCCGCCTGATGCAGATGAACC GCCA- 3’
gme-new-161-3p 5’- TACTTGTCTAGAGGGCGTGA-3’	Gene 14 5’- ATGAAATTTACAATTGTCATAAATATTACAAATAAGGATGCAGCAACTACATC TGAGTTATTTAATAAAAAACCAGATAAGTCCGTAACCATTTTTGATTTGGTTA TACACTTTTATAATTTTGACTATCAGAATAGTACTACACATGTACCTGAACACT CATACATAGCAATGACATCTGGCAGCACTGGAGAACCAAAGCACATACAAGT GCCTGTGCAATGTATTCAACCAAATATAGATGATTTAACTAAATTATTTAATA TTACTGCTGATGATATAATATATTTCTCTACACCACTAACATTTGACCCATCCA TGATAGAGATACTGCTCGCCTGTATGAATGGAGCCTCTCTACTTATTGCACCT GAAAAAGCAGACATATTATTCCCCAACAACAAAGAGAATTCTGTGACATTTT GGCAACTTACACCATCACGATTTTTTCAGCATTCAAATTCTGATATCAAGAAT AAAATATTAAGTGCAAATTCAACATTGAAAATACTAGCTTTAGGTGGTGAGC CATTAAATGGTGTGAAGAGACTAAAGGAATTGAAAGATTGGGATAATAAAA

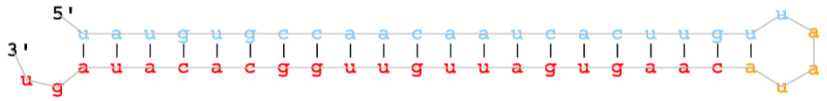
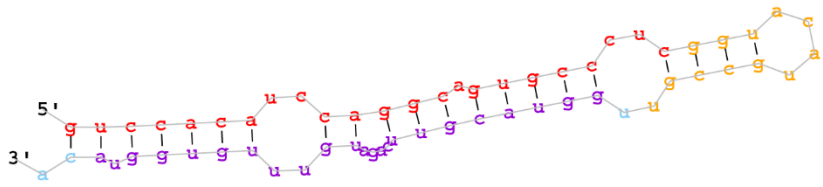
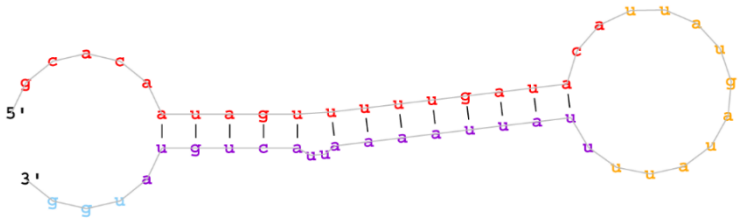
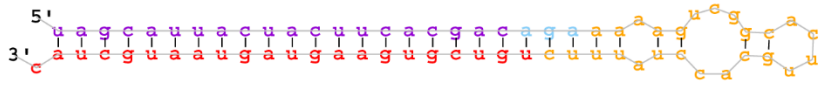
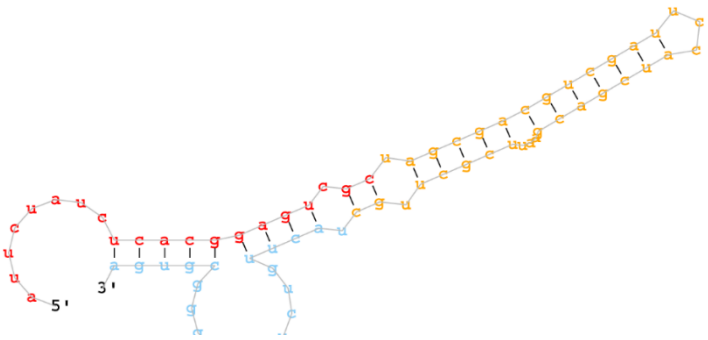
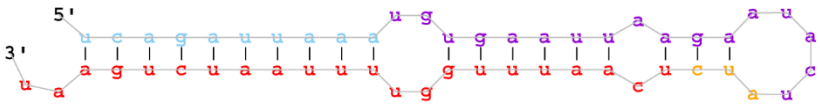
	CTAGAATATTCACATTGTATGGAGTAACTGAAATGTCATGCTGGGCTTGTGTT GCTGAGTTAGATCTTAACAAAATACTAACTGACAAAGAAATACCATTAGGTA ACTGTCTGTCAGAAACAGAATTACATGTGGAATCAAATGATGACAATAAAGA ATGTGGAAAAATTATTTTAGTAAGCAAAACAAGAAAGTGTGTTATATTAAAC AAAACCATTGGAAATGAAGATGAAAATTCTTTAAAATTTATTGACACTGGAG ATTTAGGTGAAGTAAGAAATGGCACTGTGTATTATCGGGGCCGCAAAGATG ATATTATTAAAAGATTTGGACACAAAATTAATTTACAGTTTATTGAATCAACT ATAATGCAATGTCCGAGTGTGAAAACAAGCTCCTGTATCTGGCTCCCAAAAT CATTACTTCTGATTGCCTATTTCTCATCAGAAACACTTAGCAGCCAAGAGTTG TCCAACTTTTTGAAATGTAAACTTGATGATAAGCACTGGCCAGATAAAATAAT TAGAGTTGACAATTTACCAACAAATCCTCATGGGAAGATATCTAAATTGATAT TATCTAAAATGTATGAAAAAACAATGAACACACCACAGACATTAGATTCCTTA AAAGTGAGTTTCTTAAAGGAACTTCAAGCTGTAATGGGTCAACACTTCACTTA TGATCAAATTAAAAGTAAAGACTTCTTTGCCATTGGTGGCACATCTTTTCTAG CGATATCTATGTGTAATAAGCTTTCACTACTTTGTCCAAAATTTGGTAAACTAA TTCTCCCTTACTTGATGTCTCAAAAAAATACTATAGATGATATTATGCAATTGG CATATAAAGAAATACATGTTGATGAAATAAAAGTTAAGAAAAAAATTAAAAG GTCACGGTCAGATGCAGGTGGTTATGTAGAGAGTCAGTCTTACAAAAGAACT AACACAAAAAGTCTGACAAATCCTGTAAAATTCATTGTGTTATGGTCATATGA TACTGGAAAATGTGTAGATGCTTCTCCATCTTTATTTCAAATAGGATTCAATTT ATATGTGACAGTCGGGAGTCATTCAGGGAAGATTATAGTCATGGATGCGATA TCTGGAATATTGCAGGGAATGGTAACAGTAAAATCACGTGTTGAAGCATCTG TATTTTGTTACCACAAGAGAGACATGTCGCCGTGCGGTGTGGTTGGCACTTA CGATGGCACAGTGGTATGCTTCCAATTAGAAACATGCAAAGAGTTGTGGAG AATCAACATTGGATCAATGATAAAAAGTAAAGCAACATGTTGCAATGATCTA CTCTACATTGCTTCCTATGATGGAAAGATAAGATGTATAGATATTGCGATAG GGGTAATCAAAGAGACTATATATGTGGCAGATCAAGGTATATCAGCTGATCT AGTACTTGCTAAAAACAAGTATGTGTTAACCAGTACGCTGTCTGGTGTGTGT GCAAGTATACATACTCTAACCAATACTGTGGCTTGGCGCTGCACACTGAGTA GTCCAGTATTTGCAAGTCCTGTGCTTTACGATGACGACAAGTATGTGGTATTC GCGGAAGTTAATGGTGAAATACATTGCAGGACCGTTGAAAAGGGTATTAAG ATATGGAATTATCAAGGAGCAAGAGGTAACATTTTTTCCTCCCTCTACATAAA AGAAGTTGATAAACTGAAATGGCAAATGGTTTTTGGCTGTCACGATAACAAA GTTTACAGCATTAATATCAAGAATTTCCAACCTAGCTTGAATTGGAAAGCACA ACTCACATCACCTGTATACTCCACTCCATGTGGTCTAAGTGACAAATTAATAC TTGCTGCCTCTAATAATGGCAGGTTATGCGTTATAGATGAAGAAAACGGGAT AATATTGGCAGAACATCATCTGCCAAATGAAACATTTTCATCACCAGCAGTTT ATGGAGATTACATATTCATTGGTTGCAGGAATGACCACCTTTACTCTTTGAAA TATATTTTAAATTTATAA-3’
gme-new-122-3p 5’- TAATGCCTTCCTCTAACTACA-3’	Gene 15 5’- ATGGCGTCTGATACCTTAGTACCTATAGAATCAAATCCTGAGGTTATGAATAA ATTCCTTCAAAAATTAGGTGTTCCATCTAATTGGAGCATAGTCGATGTAATGG GCTTAGATTCTGAGATGCTGTCTTGGGTTCCTCGTCCGACTATTTCTGTTATGC TGCTGTTTCCTGTATCTGCTGCATATGAAGATCATAAAAAGAAAGAGGAAAG TGAAATATTGGCTAAGGGCCAAGAAGTTTCAAGTGACATTTTTTATATGAAA CAAAATGTAAGTAATGCTTGTGGCACTGTAGCTCTGGTACACAGTGTCGCCA ACAATTATGATAAAATCCAGCTTTCTGATGGCCCTATGAAAAAATTTATAGAA GAAGCCAAACCATTAGATGCTGCTGCTCGAGGAACTCTGTTTGCAAAGACTG AAGGCATTATCAATGCTCATAAAGAATTGGCTCAAGAGGGTCAAACTAATAC CCCCAGTGCCGAAGAACCTGTTGATCATCATTTTGTAGCATTTGTACACAAAA

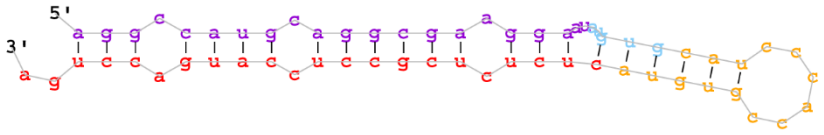
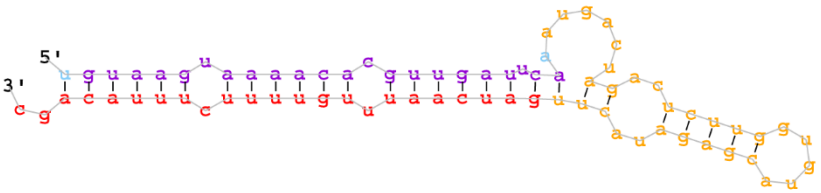
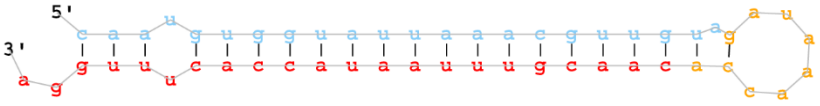
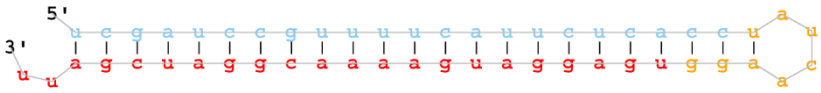
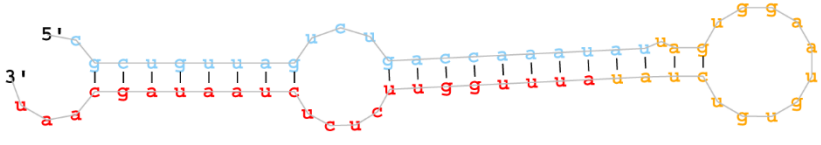
	ATGGAGCATTATATGAATTAGATGGCAGAAAGGCTTTCCCTATCAATCATGG ACCTACTACACCAGATAATCTATTAGAAGATGCTGCTAAAATTTGCAAAGAAT TCATGGCTCGTGATCCTAATGAAGTTCGTTTTACTGTGATGGCTTTAGCAGCT TCCAACTAA-3’
gme-new-82-5p 5’- TCTGTTCGTATTGTCAAGTATA-3’	Gene 16 5’- ATGAGAGATCCGCGCACGTTCGCAGTCATGCATAACGTTAGGCCACTGCAAT TTGTTACGGAGTCGGCCGGTGGGCACCGGAATCGTTCGATGTCATTGTCTAC ACACCACCACAATGGAGCGCCGAGACCGGAGCCTAGATATTCTTTGCAACTT CAGACTGATGAAAACAGCTACCTACTTCGTTTGAAGATAATAGGTGCTACGT CATTGGCCAAAAAAGATATATTCGGTGCTAGCGACCCCTATGTGCGTGTGGA ATTGCAGAAACTCGACAGCGACGTCACCCTCGAGACGTTTCTTACAAAAACC AAGAAAAAGACATTAAATCCAGTATGGAATCAGGAGTTTGTATTTAGGGTGA ACCCCCGCGAGCACAAGCTGCTGATCCAGGTGTTCGACGAGAACCGGCTGA CGCGTGACGACTTCCTCGGCATGGTGGAGCTGGCGCTGGGCGCCGTGCCCA CCGAGAGCGCGGCCGCCGCCCGCCCGCCGCCCCTCAAGTACCCGCTGCGCCC GCGCAGCGCGCGGTCTCGGGTCCGCGGCCATATCGAGGTGTACGCGGCGCT GGTGGGGCGGGTGGGCGAGCCGGGCGTGGCGGCGGCCGCCGAGCGCGGC GACGACTGGGAGCTGGTGCACGCCGCGCCCGCCGCCGGCGAGGTTCACTCG ACGGTGGTGGGTGATCCGTTACCGCCCGGCTGGGAGGAGCGACAGGACGG CAACGGGCGGACTTACTACGTGAACCACATCGAGCGGTCCACGCAGTGGGA GCGGCCTACATTCACCCGCAACCAAAGTGTGGAGTCGCAAGCTGAACGTATG GAGACGGCGGTCACCGAGTTCCAGCGGCGGTTCCACATCTCTGCGGACGAG GAGCACTCTAGTCCTGCGTCCTCGCAGCACCAGGACGAGACCGACGGGTCG ACCGAGAGCACGCGCAGCAGCAGCGCGAGTAGTACGCAGAGTCAGAATCTA CCGAACGCTGATGGATTACCACCGGGATGGACGATGCAGAAAGCGCCCAAC GGCAGGATATTCTTCATCGACCACAACCAGAAAACAACGACGTGGATAGACC CCAGAACAGGGTGCGCATCAAGCTTGCCGAGTGCGGCGGCGGCGACGGCC GGTGCTGAAGCGGACGAATTGGGAGCGCTGCCAGAAGGCTGGGAGGAGAG GGTCCACACCGACGGCAGGATCTTCTTCATCGACCACAACACGCGTACAACC CAGTGGGAGGATCCGCGGCTATCAAACCCTCAAATCGCGGGTCCCGCCGTGC CGTACTCGCGAGACTATAAGCGCAAATATGAGTACCTGAAGAGCCAGCTACG CAAGCCGAGCAACGTGCCCAACAAGTTCGAGATCAAAGTGCGTCGTAACTCG ATACTGGAGGACTCATACCGCATAATCACCTCCGTGAACCGCATCGAACTGC TGAAGACCAAGCTGTGGGTGGAGTTCGAGTCGGAAGTGGGACTAGATTACG GAGGTCTGGCGCGCGAGTGGTTCTTCCTTCTGTCCAAGGAGATGTTCAATCC TTACTACGGGTTGTTCGAGTACTCGGCGATGGACAACTACACGCTGCAGATC AACCCGAACAGCGGCGTCTGTAACGAGGAACACCTGAGCTACTTCAAGTTCA TCGGTCGCGTAGCCGGCATGGCGGTCTACCATGGGAAATTGCTCGACGCATT CTTCATCCGTCCATTCTACAAAATGATGCTGAGCAAGCCCATAGAGCTGCAG GACATGGAGTCGGTCGACTTGGAGTACTATAACTCGCTCATATACATCAAGG AAAATGATCCGTCGGAATTGTACCTAACGTTCTCGGTGGACGAGGAACAATT CGGCAAGACCATACAGAGGGACCTCAAACCAGGAGGTGCTAACATACCAGT CGATGAAGAGAACAAGGATGAGTATATCAAGCTGGTGATCCAATGGCGGTT CGTCAGTAGAGTCCAAGAGCAGATGTCATCGTTCCTCGAAGGGTTCGGGGC GCTGGTGCCGCTGAACCTGCTGAAGATCTTCGACGAGCACGAACTGGAGCT GTTGCTGTGTGGAATCCAGCATATAGACGTGCGAGATTGGCGCGCCAACACA CTGTACAAGGGAGACTATCATGCCAATCATATAGTGGTCCAATGGTTCTGGA GGGTAGTGCTGTCGTTTTCGAACGAGATGCGGTCGCGGCTGCTACAGTTCGT GACGGGTACGTCGCGGGTGCCCATGAACGGCTTCAAGGAGCTGTACGGCTC CAACGGGCCGCAGCTGTTCACCATCGAGCGCTGGGGCAGCCCCGACAACTA

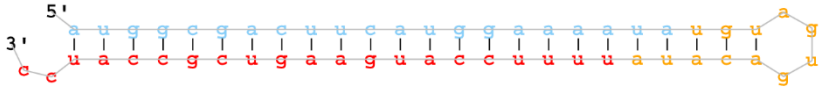
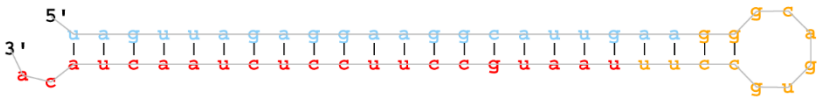
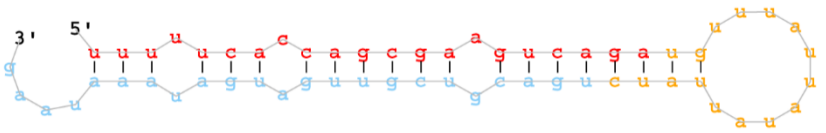
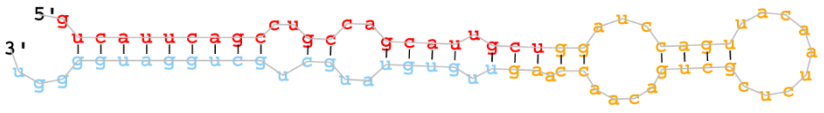
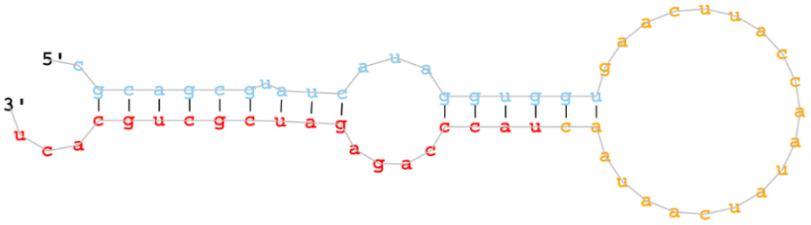
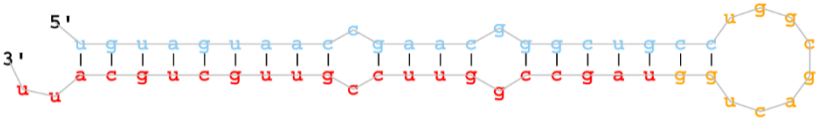
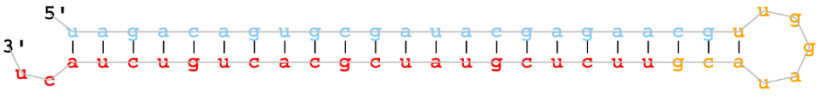
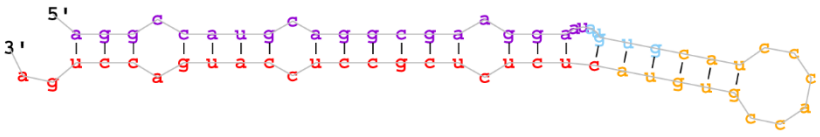
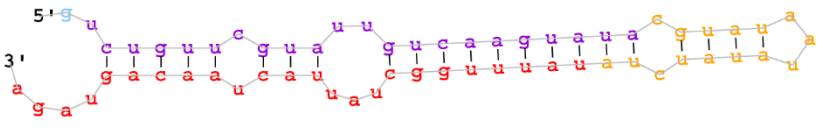
	CCCCAGGGCGCACACCTGTTTCAATCGAATCGACCTACCTCCATACGAGAGTT ATCAGCAGCTTCGCGAGAAGCTAGTCAAAGCGATCGAGGGCTCACAAGGCT TCGCCGGCGTCGACTGA-3’
gme-new-117-5p 5’- TTTTTCACCAGCGAAGTCAGA-3’	Gene 17 5’- ATGCAGATATTTGTCAAAACATTAACTGGGAAAACCATCACATTGGAAGTAG AACCATCGGATACTATTGAAAATGTGAAAGCCAAAATTCAAGACAAGGAAG GCATTCCCCCAGACCAGCAACGACTCATTTTTGCAGGCAAACAATTGGAGGA TGGCCGTACACTTTCAGATTACAACATCCAGAAGGAATCTACGTTGCACCTTG TTCTTCGTCTAAGAGGTGGTATGCAGATCTTTGTAAAGACATTAACAGGAAA AACTATCACCTTGGAGGTTGAACCTTCTGATACTATTGAAAATGTAAAGGCTA AGATTCAGGATAAAGAGGGCATTCCACCAGACCAACAACGTCTTATCTTTGC TGGCAAGCAGCTAGAAGATGGACGCACACTCTCTGATTATAACATCCAAAAA GAATCAACATTACATTTAGTATTACGACTTCGTGGTGGTATGCAAATTTTCGT AAAAACCTTGACTGGTAAAACAATCACATTGGAAGTTGAACCCTCTGATACT ATTGAGAATGTGAAAGCCAAAATCCAAGATAAAGAGGGTATTCCACCTGACC AACAACGTCTTATCTTTGCTGGCAAGCAGTTAGAAGATGGACGCACACTCTCT GATTATAACATTCAAAAAGAATCTACTTTACATTTAGTATTACGACTTCGTGG TGGTATGCAAATTTTCGTAAAAACCTTGACTGGTAAAACAATCACATTGGAA GTTGAGCCATCTGACACTATTGAGAATGTGAAGGCTAAAATTCAGGATAAAG AGGGCATTCCACCTGACCAACAACGTCTTATCTTTGCTGGCAAGCAGTTAGA AGATGGACGCACACTCTCTGATTATAACATTCAAAAAGAATCTACTTTACATT TAGTATTACGACTTCGTGGTGGTATGCAAATTTTCGTAAAAACCTTGACTGGT AAAACAATCACATTGGAAGTTGAACCCTCTGATACTATTGAGAATGTGAAAG CTAAAATCCAAGATAAAGAGGGTATTCCACCAGACCAGCAGCGACTTATCTT TGCTGGCAAGCAACTTGAAGATGGGCGTACACTTTCTGATTACAATATACAG AAGGAATCAACACTACATCTTGTATTACGTCTTCGTGGTGGTATGCAAATTTT CGTAAAAACCTTGACTGGTAAAACAATCACATTAGAAGTTGAACCCTCTGAT ACTATTGAGAATGTGAAAGCCAAAATCCAAGATAAAGAGGGAATTCCTCCAG ACCAGCAGCGACTTATCTTTGCTGGCAAGCAACTTGAAGATGGCCGTACACT TTCCGATTACAACATTCAAAAAGAATCTACTTTGCATCTTGTACTAAGGTTGA GA-3’
gme-new-160-5p 5’- GTCATTCAGCCTGCCAGCATTGCT-3’	Gene 18 5’- ATGACGGATTCTACAGTTGTGTTTGTCAATTCTGGACAGCCGCCCCCTCCATA TATTGCTCAGCCCGTGGCGCCCCCAGTCGCAGTGGTAATGACTGGGCCAGTG GGCTCAGAGCCTATAATCATGGCCTGCCCCTCGTGCCGTCACCAGATCGCAA CGAGAGTTGAAAGAGCAGCATCATCAAAAACTCATATCATAGCTTGCTTATT GTGTTTATTTGTATGCTGGCCATGCGTTTGTGTACCATACTGTGTGGATTCGT GCAACAATGCCAATCATTACTGCCCTAACTGTAACGCGTATATAGGCAGTTAT AATTTTTAA-3’
gme-new-160-5p 5’- GTCATTCAGCCTGCCAGCATTGCT-3’	Gene 19 5’- ATGACGGATTCTACAGTTGTGTTTGTCAATTCTGGACAGCCGCCCCCTCCATA TATTGCTCAGCCCGTAGCGCCCCCAGTTGCAGTGGTAATGACTGGGCCAGTG GGCTCAGAGCCTATAATCATGGCCTGTCCTTCATGCCGTCACCAGATCGCAAC GAGAGTTGAAAGAGCAGCATCATCAAAAACTCATATCATAGCTTGCTTATTG TGTTTATTTGTATGCTGGCCATGCGTTTGTGTACCATACTGTGTGGATTCGTG CAACAATGCCAATCATTACTGCCCTAACTGTAACGCGTATATAGGCAGTTATA ATTTTTAA-3’
gme-new-136-3p	Gene 20

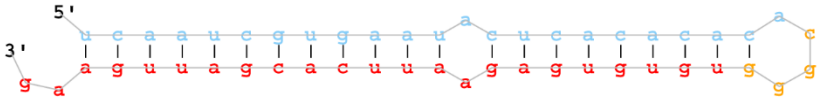
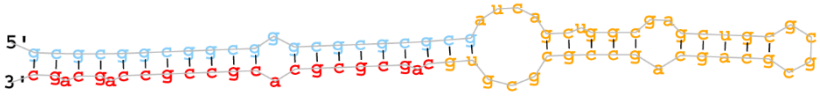
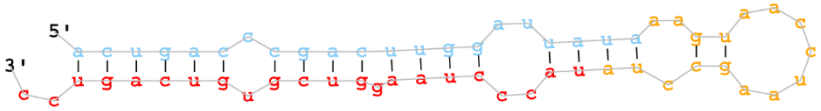
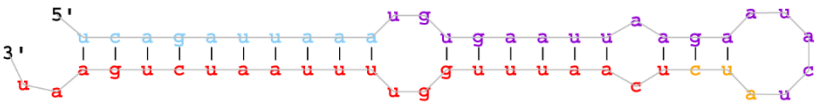
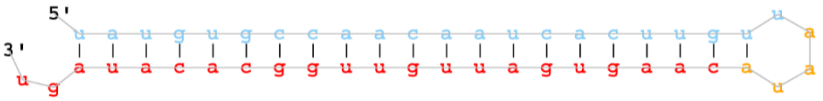
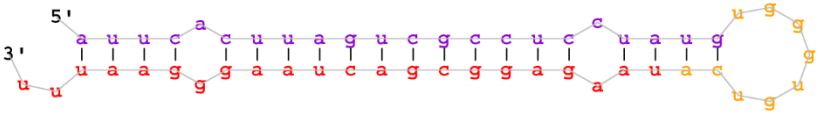
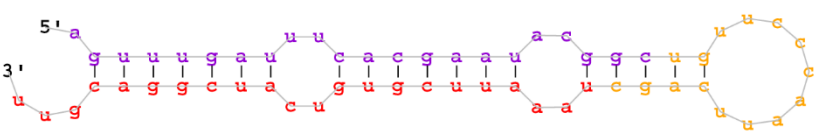
5’- TCAATTTGGTTTTAATCTGAAT-3’	5’- ATGGGAAGAAAATCACAGTCAAAACTAAGTTCTAAAAAGAATGCTAACAGG GAAAATAATAGAATAGTACAAAAGCGCAAAGAACTTGCGATTCTTGCGGATA AACTACTTCGATTAACAAGTATAATAACTCAAGTGTCCAACATTGGAAATAGC TGGGAACTTCACAAACAAATTGAAGCAGTCATCAAAGAAATATTAATTATTG AAGCACCATTCAATATTAAAACAAAACAAAATCCACGTCACTTAAACATAGAA AATTTCTTAAAATGGTTAAACGAAAATGTAGCTACATTTGAAGGTGTGGAAA TTGGTGAATTTGAAGGCTATGAATTTGGTTTAAAAGCAACAAAGAACTTTAA AGAGGGTTCATTATTGCTCACAGTGCCCAGTAAACTTATGTTGACTGTACAAA ATGCCAAGGAATCTGAACTATCGGATTTTATCAGTATGGATCCTCTTCTACAA AATATGCCCAACATAACATTGTCATTATTTTTATTATTGGAAAAAAATAATCCA GATTCTTTTTGGAAGCCATATATTGATATTCTGCCAGAAAAATATTCAACCAT CCTTTACTTCACCGCTGAAGAACTAGCTGAACTCAGGCCTTCTCCAGCTTTTG AGTCAGCTCTAAAACTTTATAAAAGTATTGCAAGACAATATGCATATTTCTAC AATAAAATACATACTTCGAATATACCAGTCTTGAAAAATCTCCAAGAAATTTT CACATTTGATAACTACAGGTGGGCTGTGTCCACTGTAATGACGCGACAGAAC ATGATACAGCTGGATGACTGTGACGTCACGTCATTTATACCCTTATGGGATAT GTGTAACCATGAACACGGAAAGATCACAACCGACTATAACAAGGAACTAAAT AGAGGCGAATGCTACGCACTGCGCGACTTCCAACAAGGCGAACAGATCTTCA TATTCTACGGTGCGAGATCCAACGTTGATCTGTTCTTGCACAATGGTTTTGTA TATCCAAAGAACCAATACGATAGTCTATCCCTATCTCTGGGTATAAGTTCAAG TGATCCACTGCGCGAAACGAAGCTATCCTTATTGAGCAAACTGGGTCTCGCC GGAGTGACACATTACAACCTATACCGTGGTGAAACACCCATCAGCGCTGAGC TGCTCGCCTTCATCAGGATATTTAATATGAATCAGGAGGAATTGGCAAAGTG GTCGAGCCAGGGTTTGCCCAGTGACCTGGTATCGTCCGAGTCCTCTAGCGTC GATGCAGTGGGCGCGGACGTCGACCGGCGCGCGTACAAGTACCTGCTAACC CGTTGCGAACTCATACGTGCTTCATACAACAAAACTAATAGCGATACAGAAC AAAACGGCCCTGAATCTACACACAGGAAAAATATCAAACTGTTGAAGGAATG TGAAGTACAAATATTGGAAGGTGCCATAAAATATCTTGAGAACGTTTTACAA AAACTACCCACCGTGAAATAA-3’
gme-new-106-5p 5’- CCTTGTCATTCTTCTTGCCCAGT-3’	Gene 21 5’- ATGAAACTTAATATTAACACACAAATTAATGGATTAAATATTACGTTAGTGCC TTACAGAAAACATCATGTACCAAAATATCATATGTGGATGAAATCAGAGGAA CTACAAAGACTCACTGCATCAGAACCTCTGTCCTTAGAACAAGAGTATGAAA TGCAAAAATCTTGGCAAGACGATGATGACAAATGTACGTTTATAATATTGCA GAAGAATACAGAAACAAATGAAATAGATTCAATGATAGGTGATACAAATATA TTTGTAACAGACAAAGAAAATTATATTGGAGAAATAGAAATAATGATAGCAG AAGAGTCTGCAAGAGGAAAGAAACTTGGTTGGGAAGCTGTAATTTTAATGTT GATTTATGGTATTAAATATATCAATCTTAAAATGTATGAAGCAAAAATATCAT TAAGTAATATTATAAGTATACAAATGTTTAAAAAGTTAGGTTTTAGTGAGAAA TCAAAAAGT-3’

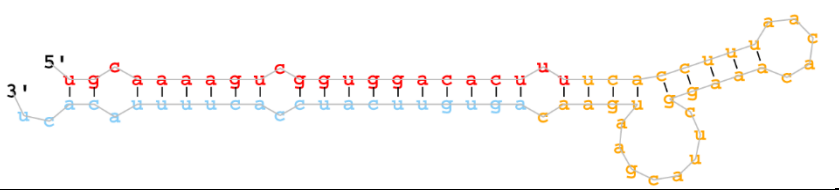
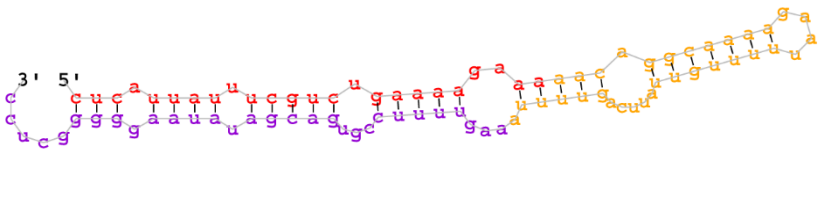
Table S 6 **Stem-loop structures of 37 novel *G. mellonella* miRNAs indicating mature miRNA sequence (red color), miRNA star strand sequences (violet color), loop (orange color), and star sequence (blue color)**

Novel miRNA	Structure
gme-new-90	
gme-new-80	
gme-new-72	
gme-new-78	
gme-new-161	
gme-new-137	
gme-new-136	

gme-new-135	
gme-new-89	
gme-new-88	
gme-new-70	
gme-new-157	
gme-new-156	
gme-new-150	
gme-new-147	
gme-new-140	

	
gme-new-122	
gme-new-117	
gme-new-160	
gme-new-123	
gme-new-97	
gme-new-93	
gme-new-89	
gme-new-82	
gme-new-61	

	
gme-new-149	
gme-new-146	
gme-new-144	
gme-new-136	
gme-new-107	
gme-new-90	
gme-new-81	
gme-new-52	
gme-new-153	

gme-new-141	
gme-new-137	
gme-new-135	

14

CONTRIBUTION REPORTS

The following chapter contains the contribution reports of my publications with signatures of all co-authors. The reports explain the individual tasks I performed in each publication, together with the percentage.

Report on Daniel Amsel's contributions in principal publications

The following research article will be submitted as part of Daniel Amsel's (DA) inaugural dissertation.

Amsel, D., Vilcinskas, A., & Billion, A. (2017). Evaluation of high-throughput isomir identification tools: illuminating the early isomirome of *Tribolium castaneum*. *Bmc Bioinformatics*, *18*(1), 359.

Coauthors full name and title

- Andreas Vilcinskas, Prof. Dr.[1,2]
- André Billion, Dr.[1]

Contact address

1. Fraunhofer Institute for Molecular Biology and Applied Ecology, Department of Bioresources, Winchester Str. 2, 35394, Giessen, Germany
2. Institute for Insect Biotechnology, Heinrich-Buff-Ring 26-32, 35392, Giessen, Germany

Contributions

DA's overall contributions is estimated at 95%. DA's main contributions are detailed below. Percentage values of the individual parts indicate the extent a task was performed by DA.

- Experiment design: 90%
- Investigation of potentially useful programs: 100%
- Selection of programs to be benchmarked: 100%
- Investigation of selected programs: 100%
- Investigation of possible microRNA isoform types: 100%
- Selection of tool for creating artificial sequencing data: 100%
- Artificial test-set creation and read simulation: 100%
- Performance calculation: 100%
- Performance evaluation: 100%
- *Tribolium castaneum* public smRNA-seq dataset selection: 100%
- *Tribolium castaneum* public smRNA-seq dataset analysis: 100%
- Analysis of microRNA isoforms of *Tribolium castaneum*: 100%
- Figure plotting: 100%
- Analysis, creation and plotting of supplemental material: 100%
- Manuscript writing: 90%
- Revision writing: 90%

Herewith I declare that specified qualitative and quantitative contributions by Daniel Amsel to this research article do not contradict with my own contributions and are to the best of my knowledge correct.
I used or plan to use part of this work also for my own dissertation, habilitation or other examination procedure.

Gießen 18.7.2020	Andreas Vilcinskas	
Place, date	Name	Signature

Justus-Liebig-Universität Gießen
Institut für Insektenbiotechnologie
Angewandte Entomologie
Heinrich-Buff-Ring 26-32
35392 Gießen

Report on Daniel Amsel's contributions in principal publications

The following research article will be submitted as part of Daniel Amsel's (DA) inaugural dissertation.

Amsel, D., Vilcinskas, A., & Billion, A. (2017). Evaluation of high-throughput isomir identification tools: illuminating the early isomirome of *Tribolium castaneum*. *Bmc Bioinformatics*, *18*(1), 359.

Coauthors full name and title

- Andreas Vilcinskas, Prof. Dr.[1,2]
- André Billion, Dr.[1]

Contact address

1. Fraunhofer Institute for Molecular Biology and Applied Ecology, Department of Bioresources, Winchester Str. 2, 35394, Giessen, Germany
2. Institute for Insect Biotechnology, Heinrich-Buff-Ring 26-32, 35392, Giessen, Germany

Contributions
DA's overall contributions is estimated at 95%. DA's main contributions are detailed below. Percentage values of the individual parts indicate the extent a task was performed by DA.

- Experiment design: 90%
- Investigation of potentially useful programs: 100%
- Selection of programs to be benchmarked: 100%
- Investigation of selected programs: 100%
- Investigation of possible microRNA isoform types: 100%
- Selection of tool for creating artificial sequencing data: 100%
- Artificial test-set creation and read simulation: 100%
- Performance calculation: 100%
- Performance evaluation: 100%
- *Tribolium castaneum* public smRNA-seq dataset selection: 100%
- *Tribolium castaneum* public smRNA-seq dataset analysis: 100%
- Analysis of microRNA isoforms of *Tribolium castaneum*: 100%
- Figure plotting: 100%
- Analysis, creation and plotting of supplemental material: 100%
- Manuscript writing: 90%
- Revision writing: 90%

Herewith I declare that specified qualitative and quantitative contributions by Daniel Amsel to this research article do not contradict with my own contributions and are to the best of my knowledge correct.
I used or plan to use part of this work also for my own dissertation, habilitation or other examination procedure.

Gießen, 14.07.'20	André Billion	
Place, date	Name	Signature

Report on Daniel Amsel's contributions in principal publications

The following research article will be submitted as part of Daniel Amsel's (DA) inaugural dissertation.

Amsel, D., Billion, A., Vilcinskas, A., & Förster, F. (2018). microPIECE-microRNA pipeline enhanced by CLIP experiments. Journal of Open Source Software, 3(24), 616.

Coauthors full name and title

- Andreas Vilcinskas, Prof. Dr.[1,2]
- André Billion, Dr.[1]
- Frank Förster, Dr.[1]

Contact address

1. Fraunhofer Institute for Molecular Biology and Applied Ecology, Department of Bioresources, Winchester Str. 2, 35394, Giessen, Germany
2. Institute for Insect Biotechnology, Heinrich-Buff-Ring 26-32, 35392, Giessen, Germany

Contributions

DA's overall contributions is estimated at 80%. DA's main contributions are detailed below. Percentage values of the individual parts indicate the extent a task was performed by DA.

- Experiment design: 95%
- Selection of programs/tools for the pipeline: 95%
- Investigation of settings and cutoffs for each tool: 95%
- Selection of the public *Aedes aegypti* data as reference CLIP dataset: 100%
- Selection of the public *Tribolium castaneum* miRNA, mRNA and ncRNA data: 100%
- *Tribolium castaneum* public dataset analysis: 100%
- Development of the pipeline: 75%
    - miRNA scripts: trimming; ncRNA filtering; reference indexing; read mapping; mining of novel miRNAs and result export to .fasta; miRNA reference completion; miRNA expression (RPM) calculation; isomiR input preparation; isomiR mining and RPM expression calculation; miRNA genome location mapping; search for homologous miRNAs in miRBase; miRNA target prediction on transferred CLIP sequences and output parsing
    - CLIP scripts: trimming; reference indexing; read mapping; file format conversion; peak-calling; filter CLIP peaks (length and transcript hit); transfer CLIP peak .bed coordinates to uppercase .fasta; identification of longest transcript isoforms from .GFFs of both species; blastdb creation; protein homology search between both species; mapping of CLIP regions between longest homologs; merge overlapping transferred CLIP regions; translate transferred CLIP regions to .fasta;
- Code Review of Co-Authors: 100%
- Comments/Documentation in pipeline: 50%
- Comments/Documentation on GitHub: 50%
- Generation of test-data and test-scripts: 50%
- Generation of example data set: 100%
- Figure plotting: 100%
- Manuscript writing: 95%

Herewith I declare that specified qualitative and quantitative contributions by Daniel Amsel to this research article do not contradict with my own contributions and are to the best of my knowledge correct. I used or plan to use part of this work also for my own dissertation, habilitation or other examination procedure.

Gießen, 14.7.2020	Vilcinskas	
Place, date	Name	Signature

Justus-Liebig-Universität Gießen
Institut für Insektenbiotechnologie
Angewandte Entomologie
Heinrich-Buff-Ring 26-32
35392 Gießen

Report on Daniel Amsel's contributions in principal publications

The following research article will be submitted as part of Daniel Amsel's (DA) inaugural dissertation.

Amsel, D., Billion, A., Vilcinskas, A., & Förster, F. (2018). microPIECE-microRNA pipeline enhanced by CLIP experiments. Journal of Open Source Software, 3(24), 616.

Coauthors full name and title

- Andreas Vilcinskas, Prof. Dr.[1,2]
- André Billion, Dr.[1]
- Frank Förster, Dr.[1]

Contact address

1. Fraunhofer Institute for Molecular Biology and Applied Ecology, Department of Bioresources, Winchester Str. 2, 35394, Giessen, Germany
2. Institute for Insect Biotechnology, Heinrich-Buff-Ring 26-32, 35392, Giessen, Germany

Contributions

DA's overall contributions is estimated at 80%. DA's main contributions are detailed below. Percentage values of the individual parts indicate the extent a task was performed by DA.

- Experiment design: 95%
- Selection of programs/tools for the pipeline: 95%
- Investigation of settings and cutoffs for each tool: 95%
- Selection of the public *Aedes aegypti* data as reference CLIP dataset: 100%
- Selection of the public *Tribolium castaneum* miRNA, mRNA and ncRNA data: 100%
- *Tribolium castaneum* public dataset analysis: 100%
- Development of the pipeline: 75%
    - miRNA scripts: trimming; ncRNA filtering; reference indexing; read mapping; mining of novel miRNAs and result export to .fasta; miRNA reference completion; miRNA expression (RPM) calculation; isomiR input preparation; isomiR mining and RPM expression calculation; miRNA genome location mapping; search for homologous miRNAs in miRBase; miRNA target prediction on transferred CLIP sequences and output parsing
    - CLIP scripts: trimming; reference indexing; read mapping; file format conversion; peak-calling; filter CLIP peaks (length and transcript hit); transfer CLIP peak .bed coordinates to uppercase .fasta; identification of longest transcript isoforms from .GFFs of both species; blastdb creation; protein homology search between both species; mapping of CLIP regions between longest homologs; merge overlapping transferred CLIP regions; translate transferred CLIP regions to .fasta;
- Code Review of Co-Authors: 100%
- Comments/Documentation in pipeline: 50%
- Comments/Documentation on GitHub: 50%
- Generation of test-data and test-scripts: 50%
- Generation of example data set: 100%
- Figure plotting: 100%
- Manuscript writing: 95%

Herewith I declare that specified qualitative and quantitative contributions by Daniel Amsel to this research article do not contradict with my own contributions and are to the best of my knowledge correct. I used or plan to use part of this work also for my own dissertation, habilitation or other examination procedure.

Gießen, 14.07.20	André Billion	André Bi.
Place, date	Name	Signature

Report on Daniel Amsel's contributions in principal publications

The following research article will be submitted as part of Daniel Amsel's (DA) inaugural dissertation.

Amsel, D., Billion, A., Vilcinskas, A., & Förster, F. (2018). microPIECE-microRNA pipeline enhanced by CLIP experiments. Journal of Open Source Software, 3(24), 616.

Coauthors full name and title

- Andreas Vilcinskas, Prof. Dr.[1,2]
- André Billion, Dr.[1]
- Frank Förster, Dr.[1]

Contact address

1. Fraunhofer Institute for Molecular Biology and Applied Ecology, Department of Bioresources, Winchester Str. 2, 35394, Giessen, Germany
2. Institute for Insect Biotechnology, Heinrich-Buff-Ring 26-32, 35392, Giessen, Germany

Contributions

DA's overall contributions is estimated at 80%. DA's main contributions are detailed below. Percentage values of the individual parts indicate the extent a task was performed by DA.

- Experiment design: 95%
- Selection of programs/tools for the pipeline: 95%
- Investigation of settings and cutoffs for each tool: 95%
- Selection of the public *Aedes aegypti* data as reference CLIP dataset: 100%
- Selection of the public *Tribolium castaneum* miRNA, mRNA and ncRNA data: 100%
- *Tribolium castaneum* public dataset analysis: 100%
- Development of the pipeline: 75%
    - miRNA scripts: trimming; ncRNA filtering; reference indexing; read mapping; mining of novel miRNAs and result export to .fasta; miRNA reference completion; miRNA expression (RPM) calculation; isomiR input preparation; isomiR mining and RPM expression calculation; miRNA genome location mapping; search for homologous miRNAs in miRBase; miRNA target prediction on transferred CLIP sequences and output parsing
    - CLIP scripts: trimming; reference indexing; read mapping; file format conversion; peak-calling; filter CLIP peaks (length and transcript hit); transfer CLIP peak .bed coordinates to uppercase .fasta; identification of longest transcript isoforms from .GFFs of both species; blastdb creation; protein homology search between both species; mapping of CLIP regions between longest homologs; merge overlapping transferred CLIP regions; translate transferred CLIP regions to .fasta;
- Code Review of Co-Authors: 100%
- Comments/Documentation in pipeline: 50%
- Comments/Documentation on GitHub: 50%
- Generation of test-data and test-scripts: 50%
- Generation of example data set: 100%
- Figure plotting: 100%
- Manuscript writing: 95%

Herewith I declare that specified qualitative and quantitative contributions by Daniel Amsel to this research article do not contradict with my own contributions and are to the best of my knowledge correct. I used or plan to use part of this work also for my own dissertation, habilitation or other examination procedure.

Gießen, 20.07.2020	Frank Förster	
Place, date	Name	Signature

Report on Daniel Amsel's contributions in principal publications

The following research article will be submitted as part of Daniel Amsel's (DA) inaugural dissertation.

Krishnendu, M*., **Amsel, D.***, Miriam, K., Andre, B., Ulrich, D., & Andreas, V. (2020). MicroRNAs regulate innate immunity against uropathogenic and commensal-like Escherichia coli infections in the surrogate insect model *Galleria mellonella*. *Scientific Reports (Nature Publisher Group), 10*(1).
* These authors contributed equally

Coauthors full name and title

- Andreas Vilcinskas, Prof. Dr.[1,2]
- Ulrich Dobrindt, Prof. Dr.[3]
- Krishnendu Mukherjee, Dr.[1,3]
- André Billion, Dr.[1]
- Miriam Kalsy, M.Sc.[1]

Contact address

1. Fraunhofer Institute for Molecular Biology and Applied Ecology, Department of Bioresources, Winchester Str. 2, 35394, Giessen, Germany
2. Institute for Insect Biotechnology, Heinrich-Buff-Ring 26-32, 35392, Giessen, Germany
3. Institute of Hygiene, University of Münster, Mendel Strasse 7, 48149, Münster, Germany

Contributions
DA's overall contributions is estimated at 45%. DA's main contributions are detailed below. Percentage values of the individual parts indicate the extent a task was performed by DA.

- Running the `microPIECE` pipeline on the smRNA-seq data: 100%
- Upload of the smRNA-seq data to NCBI: 100%
- Selection of the public *Aedes aegypti* data as reference CLIP dataset: 100%
- Selection of miRNA reference species from `miRBase.org`: 100%
- Identification and naming of all microRNAs and submission to miRBase.org: 100%
- Target prediction of microRNAs with `microPIECE`, `RNAhybrid` and `RNA22`: 100%
- Genome-reference based assembly and annotation of *Galleria mellonella* transcriptome: 100%
- Annotation of *Galleria mellonella* genome: 100%
- Calculation of differential expression: 100%
- Computing Statistics: 100%
- Plotting of figures: 80%
- Manuscript writing: 40%

Herewith I declare that specified qualitative and quantitative contributions by Daniel Amsel to this research article do not contradict with my own contributions and are to the best of my knowledge correct. I used or plan to use part of this work also for my own dissertation, habilitation or other examination procedure.

Gießen, 18.7.2020 Andreas Vilcinskas

Place, date Name Signature

Justus-Liebig-Universität Gießen
Institut für Insektenbiotechnologie
Angewandte Entomologie
Heinrich-Buff-Ring 26-32
35392 Gießen

Report on Daniel Amsel's contributions in principal publications

The following research article will be submitted as part of Daniel Amsel's (DA) inaugural dissertation.

Krishnendu, M*., Amsel, D.*, Miriam, K., Andre, B., Ulrich, D., & Andreas, V. (2020). MicroRNAs regulate innate immunity against uropathogenic and commensal-like Escherichia coli infections in the surrogate insect model *Galleria mellonella*. *Scientific Reports (Nature Publisher Group), 10*(1).
* These authors contributed equally

Coauthors full name and title

- Andreas Vilcinskas, Prof. Dr.[1,2]
- Ulrich Dobrindt, Prof. Dr.[3]
- Krishnendu Mukherjee, Dr.[1,3]
- André Billion, Dr.[1]
- Miriam Kalsy, M.Sc.[1]

Contact address

1. Fraunhofer Institute for Molecular Biology and Applied Ecology, Department of Bioresources, Winchester Str. 2, 35394, Giessen, Germany
2. Institute for Insect Biotechnology, Heinrich-Buff-Ring 26-32, 35392, Giessen, Germany
3. Institute of Hygiene, University of Münster, Mendel Strasse 7, 48149, Münster, Germany

Contributions
DA's overall contributions is estimated at 45%. DA's main contributions are detailed below. Percentage values of the individual parts indicate the extent a task was performed by DA.

- Running the `microPIECE` pipeline on the smRNA-seq data: 100%
- Upload of the smRNA-seq data to NCBI: 100%
- Selection of the public *Aedes aegypti* data as reference CLIP dataset: 100%
- Selection of miRNA reference species from `miRBase.org`: 100%
- Identification and naming of all microRNAs and submission to miRBase.org: 100%
- Target prediction of microRNAs with `microPIECE`, `RNAhybrid` and `RNA22`: 100%
- Genome-reference based assembly and annotation of *Galleria mellonella* transcriptome: 100%
- Annotation of *Galleria mellonella* genome: 100%
- Calculation of differential expression: 100%
- Computing Statistics: 100%
- Plotting of figures: 80%
- Manuscript writing: 40%

Herewith I declare that specified qualitative and quantitative contributions by Daniel Amsel to this research article do not contradict with my own contributions and are to the best of my knowledge correct. I used or plan to use part of this work also for my own dissertation, habilitation or other examination procedure.

Place, date	Name	Signature
Münster, 20.07.2020	Professor Dr. rer. nat. Ulrich Dobrindt	
	Dr. rer. nat. Krishnendu Mukherjee	

Universitätsklinikum Münster
Institut für Hygiene
Mikrobielle Genomplastizität
Direktor: Univ.-Prof. Dr. rer. nat. U. Dobrindt
Mendelstraße 7 - 48149 Münster

Report on Daniel Amsel's contributions in principal publications

The following research article will be submitted as part of Daniel Amsel's (DA) inaugural dissertation.

Krishnendu, M*., **Amsel, D.***, Miriam, K., Andre, B., Ulrich, D., & Andreas, V. (2020). MicroRNAs regulate innate immunity against uropathogenic and commensal-like Escherichia coli infections in the surrogate insect model *Galleria mellonella*. *Scientific Reports (Nature Publisher Group)*, *10*(1).
* These authors contributed equally

Coauthors full name and title

- Andreas Vilcinskas, Prof. Dr.[1,2]
- Ulrich Dobrindt, Prof. Dr.[3]
- Krishnendu Mukherjee, Dr.[1,3]
- André Billion, Dr.[1]
- Miriam Kalsy, M.Sc.[1]

Contact address

1. Fraunhofer Institute for Molecular Biology and Applied Ecology, Department of Bioresources, Winchester Str. 2, 35394, Giessen, Germany
2. Institute for Insect Biotechnology, Heinrich-Buff-Ring 26-32, 35392, Giessen, Germany
3. Institute of Hygiene, University of Münster, Mendel Strasse 7, 48149, Münster, Germany

Contributions
DA's overall contributions is estimated at 45%. DA's main contributions are detailed below. Percentage values of the individual parts indicate the extent a task was performed by DA.

- Running the `microPIECE` pipeline on the smRNA-seq data: 100%
- Upload of the smRNA-seq data to NCBI: 100%
- Selection of the public *Aedes aegypti* data as reference CLIP dataset: 100%
- Selection of miRNA reference species from `miRBase.org`: 100%
- Identification and naming of all microRNAs and submission to miRBase.org: 100%
- Target prediction of microRNAs with `microPIECE`, `RNAhybrid` and `RNA22`: 100%
- Genome-reference based assembly and annotation of *Galleria mellonella* transcriptome: 100%
- Annotation of *Galleria mellonella* genome: 100%
- Calculation of differential expression: 100%
- Computing Statistics: 100%
- Plotting of figures: 80%
- Manuscript writing: 40%

Herewith I declare that specified qualitative and quantitative contributions by Daniel Amsel to this research article do not contradict with my own contributions and are to the best of my knowledge correct. I used or plan to use part of this work also for my own dissertation, habilitation or other examination procedure.

Gießen, 14.07.'20	André Billion	André Bil.
Place, date	Name	Signature

Report on Daniel Amsel's contributions in principal publications

The following research article will be submitted as part of Daniel Amsel's (DA) inaugural dissertation.

Krishnendu, M*., **Amsel, D.***, Miriam, K., Andre, B., Ulrich, D., & Andreas, V. (2020). MicroRNAs regulate innate immunity against uropathogenic and commensal-like Escherichia coli infections in the surrogate insect model *Galleria mellonella*. *Scientific Reports (Nature Publisher Group), 10*(1).
* These authors contributed equally

Coauthors full name and title

- Andreas Vilcinskas, Prof. Dr.[1,2]
- Ulrich Dobrindt, Prof. Dr.[3]
- Krishnendu Mukherjee, Dr.[1,3]
- André Billion, Dr.[1]
- Miriam Kalsy, M.Sc.[1]

Contact address

1. Fraunhofer Institute for Molecular Biology and Applied Ecology, Department of Bioresources, Winchester Str. 2, 35394, Giessen, Germany
2. Institute for Insect Biotechnology, Heinrich-Buff-Ring 26-32, 35392, Giessen, Germany
3. Institute of Hygiene, University of Münster, Mendel Strasse 7, 48149, Münster, Germany

Contributions
DA's overall contributions is estimated at 45%. DA's main contributions are detailed below. Percentage values of the individual parts indicate the extent a task was performed by DA.

- Running the `microPIECE` pipeline on the smRNA-seq data: 100%
- Upload of the smRNA-seq data to NCBI: 100%
- Selection of the public *Aedes aegypti* data as reference CLIP dataset: 100%
- Selection of miRNA reference species from `miRBase.org`: 100%
- Identification and naming of all microRNAs and submission to miRBase.org: 100%
- Target prediction of microRNAs with `microPIECE`, `RNAhybrid` and `RNA22`: 100%
- Genome-reference based assembly and annotation of *Galleria mellonella* transcriptome: 100%
- Annotation of *Galleria mellonella* genome: 100%
- Calculation of differential expression: 100%
- Computing Statistics: 100%
- Plotting of figures: 80%
- Manuscript writing: 40%

Herewith I declare that specified qualitative and quantitative contributions by Daniel Amsel to this research article do not contradict with my own contributions and are to the best of my knowledge correct. I used or plan to use part of this work also for my own dissertation, habilitation or other examination procedure.

Gießen, 22.7.20	Miriam Kalsy	
Place, date	Name	Signature