Non-coding RNAs and Conserved Non-coding Elements in Insect Genomes

Dissertation

zur

Erlangung des Doktorgrades (Dr. rer. nat.)

der

Mathematisch-Naturwissenschaftlichen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Tanja Ziesmann

aus

Dinslaken

Bonn, November 2018

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn.

Angefertigt am Zoologischen Forschungsmuseum Alexander Koenig, Bonn.

Erstgutachter:	Prof. Dr. Bernhard Misof
Zweitgutachter:	PD Dr. Lars Podsiadlowski

Tag der Promotion: 18.03.2019

Erscheinungsjahr: 2019

Abstract

Insects are the largest group within arthropods and in this group various phenotypes and lifestyles can be found. To understand where how this diversity evolved insects are studied on both a morphological and genomic level. The focus of the genomic research lies on protein-coding genes.

Genomes, however, consist of different parts with different functions. Only a small fraction (∼2 % in humans) is made up of protein-coding genes, whereas the majority of the genome consists of functional parts such as non-coding RNAs (ncRNAs), or regulatory elements, and parts where first evidence shows function but is not yet known what it is, such as conserved non-coding elements (CNEs), transposable elements or repeats. ncRNAs are involved in a plethora of processes in an organism, such as gene regulation, RNA modification and processing, mRNA translation, RNA silencing, and defence against predatory genomic elements. CNEs have been shown to be involved in gene regulation, although the mechanism remains unclear. As stated lies the research focus on protein-coding genes, making most other genomic parts understudied, especially in non-model organisms. In chapter 1 I provide detailed information about the function of different ncRNA classes as well as their functions, and known presence in insects. Regarding the CNEs I also present their background as well as the current state of research.

Within this thesis I analyse different Hymenoptera genomes regarding their ncRNA and CNE repertoire. In chapters 2, 3, and 4 I focus on the two species *Athalia rosae* and *Orussus abietinus* and categorise their ncRNA repertoire through both homology

and *de novo* analysis. Using the ncRNAs known from other Hymenoptera and present in the databases Rfam and miRBase, I was able to identify a set of ncRNA families that is present in all analysed Hymenoptera. Further *de novo* analysis of these two genomes showed, that the ncRNA repertoire of miRNAs, tRNAs, lncRNAs, and snoRNAs is larger than shown through the homology prediction alone. This emphasises the importance of not only relying on data present in databases to predict the full ncRNA repertoire of a species, especially in not well studied lineages.

Chapters 5, 6, and 7 focus on the identification of CNEs in four Hymenoptera species (*Apis mellifera*, *Athalia rosae*, *Nasonia vitripennis*, and *Orussus abietinus*). Comparing the genomes using pairwise whole genome alignments I was able to identify numerous CNEs in these Hymenoptera. The CNEs were often found in cluster of at least two (between 76 % and 89 %). My search for genes that are likely associated with these CNE clusters identified a number of lncRNAs as potential interaction partners. Looking at the CNE clusters consisting of more than 10 CNEs and having an lncRNA as the interaction partner, I found these clusters conserved between at least two species. My analysis shows, that these conserved regions can still be identified in lineages with a long divergence time (over 240 million years) as well as a high sequence divergence. Furthermore, the focus of gene interaction partners should be broadened to include non-protein-coding genes.

The final chapter provides an overview of the results of this thesis as well as a discussion how my findings fit into the general context of theses fields of research.

Contents

1 Introduction **1**

1.1 Non-coding RNAs 1

1.1.1 Small RNAs 4

1.1.2 Transfer RNAs 17

1.1.3 Small nucleolar RNAs 18

1.1.4 Ribosomal RNAs 19

1.1.5 Long non-coding RNAs 20

1.1.6 Current state of non-coding RNA research 22

1.1.7 Non-coding RNAs in insects 25

1.2 Conserved non-coding elements 26

1.2.1 Characteristics of conserved non-coding elements 26

1.2.2 Function of conserved non-coding elements 28

1.2.3 Where are conserved non-coding elements known so far? 30

1.2.4 Conserved non-coding elements in insects 31

1.3 Aim of this thesis 31

2 Methods non-coding RNAs **33**

2.1 Genomic data 33

2.2 Homology prediction of non-coding RNAs 34

2.3 *de novo* prediction of non-coding RNAs 38

2.3.1 tRNAscan-SE 38

2.3.2 DARIO pipeline 38

2.3.3 RNAz 40
2.3.4 FEELnc 41

3 Results non-coding RNAs 45

3.1 Database curation 45
3.2 Results of the homology prediction 45
3.2.1 Predicted ncRNAs in *Athalia rosae* 46
3.2.2 Predicted ncRNAs in *Orussus abietinus* 48
3.3 Results of the *de novo* prediction 53
3.3.1 DARIO datasets 53
3.3.2 *de novo* Prediction of tRNAs in *Athalia rosae* 54
3.3.3 *de novo* Prediction of tRNAs in *Orussus abietinus* 59
3.3.4 *de novo* Prediction of miRNAs in *Athalia rosae* 62
3.3.5 *de novo* Prediction of miRNAs in *Orussus abietinus* 62
3.3.6 *de novo* Prediction of snoRNAs in *Athalia rosae* 62
3.3.7 *de novo* Prediction of snoRNAs in *Orussus abietinus* 63
3.3.8 RNAz 64
3.3.9 *de novo* Prediction of lncRNAs in *Athalia rosae* 64
3.3.10 *de novo* Prediction of lncRNAs in *Orussus abietinus* 65
3.3.11 lncRNA-protein-coding gene interaction in *Apis mellifera* and *Nasonia vitripennis* 66

4 Discussion non-coding RNAs 67

4.1 Database curation 67
4.2 Homology prediction of non-coding RNAs 69
4.3 *de novo* prediction of non-coding RNAs 71
4.4 Non-coding RNA repertoire of *Athalia rosae* and *Orussus abietinus* . . 73

5 Methods conserved non-coding elements 75

5.1 CNEr 75

5.2 CNE_gene_neighbourhood.pl 77
5.3 unique_cnes_in_cluster.pl 77
5.4 cne_gene_count.pl 78
5.5 cne_get_one_closest_gene.pl 78
5.6 cne_synteny.pl 79
5.7 cne_diff_species_ident.pl 79

6 Results conserved non-coding elements 81
6.1 CNE prediction 81
6.2 CNE cluster analysis 83
6.3 CNE gene interaction 85
6.4 CNE cluster synteny 93

7 Discussion conserved non-coding elements 95

8 Conclusions 99

References 113

A Appendix 115
A.1 Prediction of non-coding RNAs 115
A.2 Electronic supplement 132
A.2.1 Scripts 132
A.2.2 ncrna_results 133
A.2.3 CNE_results 133

Appendix 134

Acronyms 135

Acronyms 137

List of figures 137

List of tables **138**

Danksagung **139**

Electronic Supplement **141**

1. Introduction

Eukaryotic genomes consist of a lot more parts than just protein-coding regions. Because the protein-coding part of the human genome is only 1 %, the remaining parts of the genome were first called 'junk DNA' (Ohno, 1972) and thought to be without function. As research on genomes continued it became clear that this part consists of different elements most of which are necessary for the organism. Some examples for these elements are repeats, transposable elements, different classes of non-coding RNA, different regulatory elements, and otherwise conserved regions such as conserved non-coding elements.

1.1. Non-coding RNAs

Non-coding RNAs (ncRNA) are transcribed but not translated and are involved in the workings of the cells. The different classes have different functions. They interact directly with deoxyribonucleic acid (DNA) or ribonucleic acid (RNA), or are involved in cellular processes. Currently at least nine different classes of ncRNAs are known, with a varying number of members. The classes with most known members are the transfer RNAs (tRNAs), micro RNAs (miRNAs), small nuclear RNAs (snRNAs), ribosomal RNAs (rRNAs), and the long non-coding RNAs (lncRNAs). The snRNA class has several subtypes, with the most common subtypes being the small nucleolar RNAs (snoRNAs). Other classes are small interfering RNAs (siRNAs) and PIWI-interacting RNAs (piRNAs).

The length of these vary, but all have an important secondary structure (figure 1.1). Some of the shorter ncRNAs are often collected under the umbrella term 'small RNAs'. These include miRNAs, siRNAs, and piRNAs. The small RNAs are all around 20-30 base pair (bp) long and are associated with the Argonaute family proteins.
Another well studied class that consists of only a few families is rRNAs. They can be quite long compared to the other classes.
The number of ncRNA classes shifts overtime as more becomes known about ncRNAs. Some former classes get integrated into others (e.g., piRNAs now inlcude repeat-associated RNAs (rasiRNAs)) or they may be split as more becomes known about their function or biogenesis, and also completely new classes may be discovered. This creates problems with the comparison between different ncRNA annotations as they might use different categories and standards for their annotation.

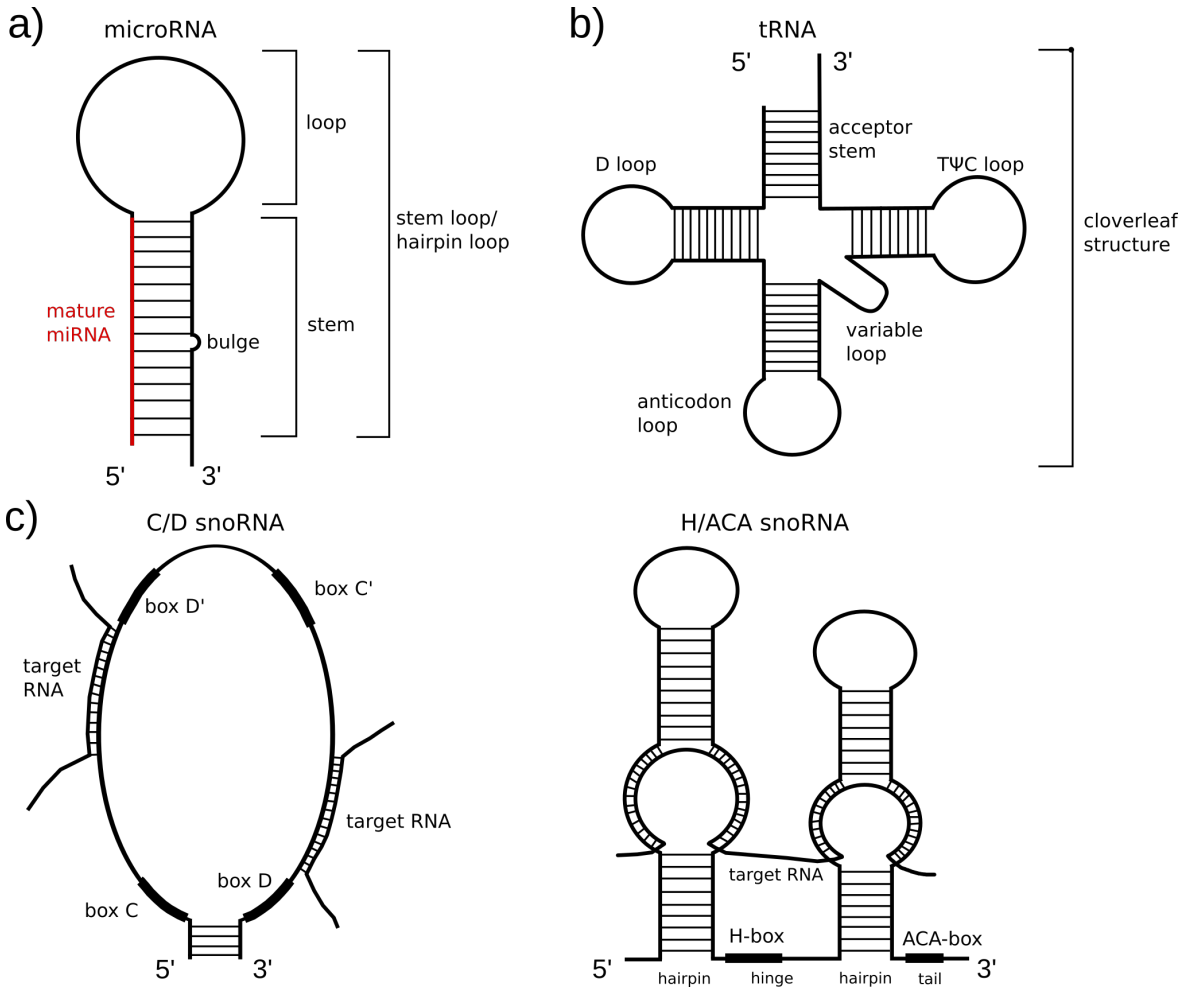


Figure 1.1.: Schematic overview of the secondary structures of four different non-coding RNAs. a) shows the typical hairpin structure of a pre-micro RNA in Metazoa. They consist of a stem and a loop, the combination of the two is called a hairpin loop or stem loop. The stem can contain unbound nucleotides which create so called 'bulges'. The mature miRNA (red) consists only of one half of the stem. b) shows a tRNA which folds into a cloverleaf structure. It consists of three hairpin loops, an additional stem, and another variable loop. The anticodon loop binds to the corresponding amino acids and transports it to the protein synthesis machinery. c) shows the secondary structures of the two most common types of snoRNA. Left a C/D snoRNAs is shown. They fold into a short stem and a big loop. Important are the boxes inside the loop as well as the regions that pair with the target RNA (see subsection 1.1.3 for further details). On the right side a H/ACA snoRNA is shown. They contain two hairpin loops that both have an internal loop separating the stem into an upper and a lower stem. This internal loop binds to the target RNA. Between the two hairpin loops the H-box can be found and at the 3' end of the snoRNA the ACA-box is present.

1.1.1. Small RNAs

Small RNAs, also sometimes called small silencing RNAs, include several different ncRNA classes that are associated with Argonaute proteins, are short (20-30 bp), and typically have a target gene of which they reduced the expression (Ghildiyal and Zamore, 2009; Kim and Pritchard, 2007). The actual mechanism of the gene regulation varies between the different RNA classes, as well as their biogenesis. The process of gene regulation these siRNAs are involved in is called RNA interference (RNAi). RNAi can be found in Metazoa as well as in plants (Ghildiyal and Zamore, 2009). Since the first discovery of RNAi in *Caenorhabditis elegans* the understanding of how this mechanism works has changed a lot.

In 1991 Fire et al. (1991) used some single-stranded antisense RNAs to disrupt the expression of genes responsible for the encoding of myofilament proteins. They showed that some kind of interference exists but the mechanisms was still unclear. Further studies showed that is doesn't matter whether a sense or antisense RNA was used for interference, and that the effects from the interference can be passed onto an offspring (Burton et al., 2011). The fact that sense as well as antisense RNA strands lead to a change in gene expression led to research on the involvement of double-stranded RNA (dsRNA). dsRNA can also interfere in protein expression, however it is highly specific (Fire et al., 1998). Also, it is possible for the dsRNAs to spread to other cells, crossing cellular boundaries. This discovery was a step in the right direction discovering the mechanism of RNAi, but still did not reveal the mechanism itself. However, Fire et al. (1998) proposed that a simple antisense model for RNAi is unlikely, and that the process of RNAi itself exists because it has a biological purpose. Another result was that a transport mechanism for dsRNA must be present to get these RNAs into other cells or even the germline.

Other studies in plants showed that dsRNAs are involved in the targeting or viral RNAs (Hannon, 2002). In these cases the interference works on the post-transcriptional level. But this is not the only level where RNAi is active. In plants it has been shown that

RNAi is involved in some methylation processes, and in *Drosophila* it has been found to regulate gene expression at the chromatin structure level (Hannon, 2002).
Through further research the nuclease complex that is responsible for the gene silencing has been discovered. It is called RNA-induced silencing complex (RISC) (Hammond et al., 2000). This complex identifies the target of the small RNA through sequence complementarity (Bartel, 2004). An important part of the RISC are members of the Argonaute protein family, which play a crucial role in the RNAi process (Bartel, 2004). The Argonaute protein family can be split into two groups. One is the Ago subfamily, the other is the Piwi subfamily. The latter gives the piRNAs their name, as they interact only with this subfamily, whereas both miRNAs and siRNAs interact with the Ago subfamily (Kim and Pritchard, 2007).
The three most prominent classes are further described in the following sections.

MicroRNAs

MicroRNAs (miRNAs) are single-stranded RNAs (ssRNAs), short (between 22-24 bp), and have a characteristic hairpin structure (figure 1.1 a). miRNAs have a short seed region of around 10 bp that is important for their interaction with a target gene. To get the seed out of the whole miRNA a complex machinery is involved, called the miRNA biogenesis machinery (figure 1.2). In this pathway the two RNase III enzymes Drosha and Dicer are involved. Before the mature miRNA is ready a primary miRNA transcript (pri-miRNA) is transcribed from the genome by RNA polymerase II (Pol II) (Lee et al., 2004) (figure 1.2). This single strand can consist of one or several neighbouring miRNA hairpin loops with flanking regions, and the length can vary between several hundred basepairs to kilobases (Denli et al., 2004). The whole pri-miRNA contains a cap structure on one end and a poly(A) tail on the other. Both cap and poly(A) tail are not present in further miRNA transcripts.
The next step is to cut out the pre-miRNAs which are one single hairpin loop without any tails, with a length of around 70 bp (Denli et al., 2004). This cleaving is done by Drosha, which is a nuclear RNase III-type protein and is still happening in the nucleus

(Denli et al., 2004; Kim et al., 2009b). Drosha interacts with another protein in this step that contains domains for dsRNA-binding. This protein is DGCR8 in humans (Han et al., 2004) or Pasha in *Drosophila* and *C. elegans* (Denli et al., 2004). From this loop a miRNA duplex is cut out by Dicer by removing the loop section. The duplex contains a miRNA and a miRNA*. To determine which of the two strands is loaded onto RISC the binding of the 5' end is evaluated. The one where this end is less tightly paired enters the RISC (Bartel, 2004). The miRNA is loaded into the RISC together with the target messenger RNA (mRNA), leading to a name change of the RISC. It is now called microRNA-induced silencing complex (miRISC) to show that it is loaded with a miRNA. A part of the RISC is the Argonaute protein (Bartel, 2004). The mature miRNA contains a seed region that directly interacts with the target gene. The level of regulation depends on the number of binding sites between miRNA and gene. Being part of the miRISC leads either to an endonucleolytic cleavage of the mRNA or an interference of the protein synthesis (Denli et al., 2004).

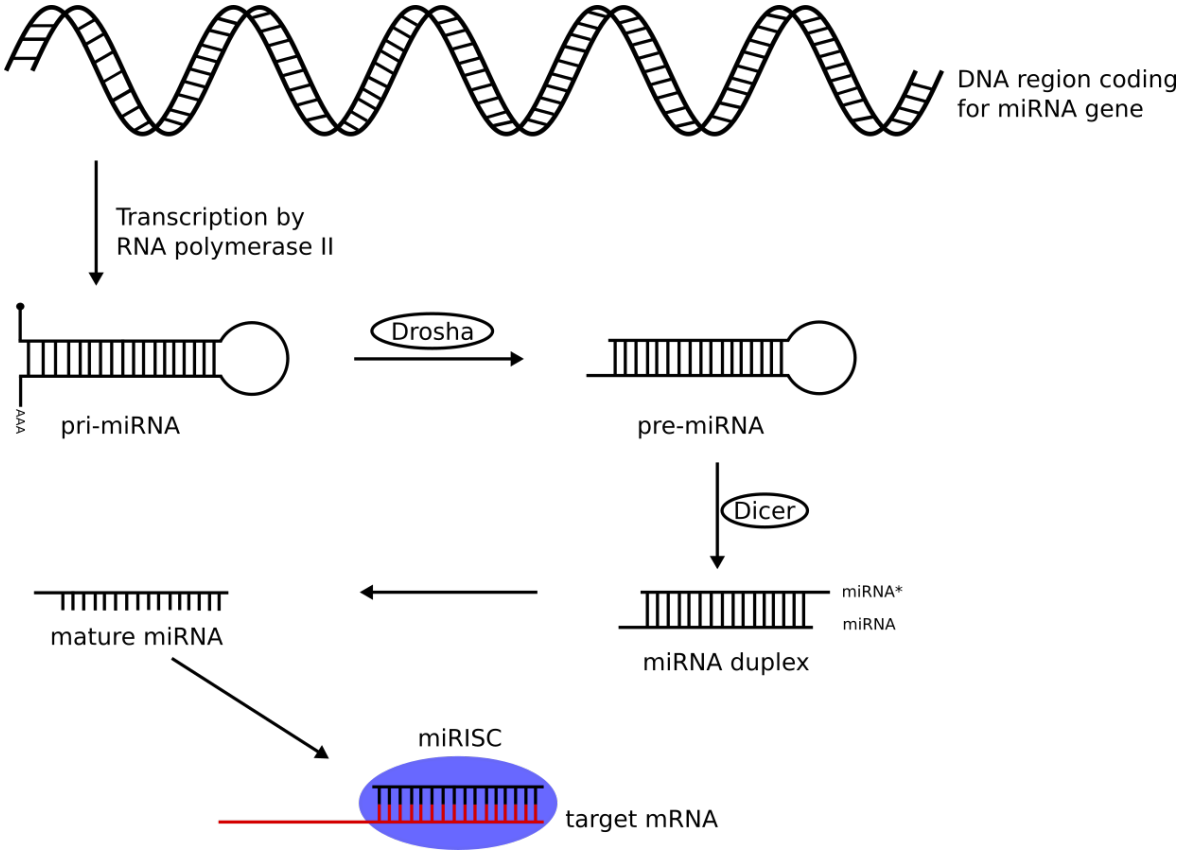


Figure 1.2.: Biogenesis pathway of a micro RNA. The miRNA gene is transcribed from theDNA inside the nucleus by Pol II. The transcription product is called pri-miRNA and consists of one or more hairpin loops containing a cap and poly(A) tail. The pri-miRNA is further processed by Drosha into a pre-miRNA. The pre-miRNA is one hairpin loop without any tails, where the stem contains the mature miRNA. The next step is to remove the loop section using Dicer. This results in a miRNA duplex. The duplex is separated into two single-stranded miRNAs, one called miRNA, the other miRNA*. Together with the target mRNA the mature miRNA is loaded onto the RISC.

The seed region of a miRNA is generally 6-8 bp long and binds to the mRNA of the target gene, most of the time in the 3' untranslated region (UTR) (Kim et al., 2009b). The more nucleotides of the seed are paired and bonded with a nucleotide of the mRNA without any bulges, the stronger the gene regulation through cleavage of the mRNA will be (Doench et al., 2003; Olsen and Ambros, 1999). If the miRNA binds only partially the mRNA is not cleaved, however it will not be translated (Doench et al., 2003; Olsen and Ambros, 1999).

The first discovered miRNA was called lin-4 (Lee et al., 1993). It was found in *C. elegans* and was originally thought to be a protein-coding gene. However, it was discovered to actually produce small RNAs. This miRNA is involved in a pathway that triggers the transition to the second larval stage. Since this first discovery it has become clear that miRNAs have many different functions and their expression can be developmental stage or tissue dependent. They are most researched in humans, followed by other model organisms.

miRNAs are grouped into families through their seed regions. miRNA families can be species/lineage specific or can be shared between different organisms (Ruby et al., 2007; Warren et al., 2008; Marco et al., 2012). The seed region is the most important part of a miRNA to identify homologs in other species. A miRNA family can be present with more than one member in a species. The sequence of different family members can vary in most of their sequence as long as the seed region, and therefore also the part that pairs with these nucleotides, is conserved. Mir-2 for example can be found with four copies in several species. They are often present in a cluster and that cluster is also preserved between species.

One such miRNA cluster consists of several members of the mir-2 family. This family is present in various invertebrates, but the copy number varies. Whereas *C. elegans* has only one mir-2 gene, *D. melanogaster* has eight, and most other insects have five copies (Marco et al., 2012). The eight mir-2 genes in *D. melanogaster* are organised in two clusters. The overall structure of mir-2 clusters varies in length in different species. The expansion of this family happened through several tandem duplications and deletions (Marco et al., 2012). One of these duplications happened in a common insect ancestor, but the split of the cluster into two happened in an *Drosophila* ancestor. After this split more duplications happened, explaining the difference in mir-2 gene number between insects. Through all duplications the seed sequence on the 3' arm has been conserved. The mir-2 cluster is in most organisms spatially linked to the mir-71 gene which is present in front of the cluster, but mir-71 was lost in the dipteran lineage

(Marco et al., 2012). Both gene families are evolutionary unrelated. Target prediction of the mir-2 family showed that this family targets genes involved in neural development in both *Drosophila* and *Caenorhabditis* (Marco et al., 2012). Expression data showed mir-2 products being highly expressed in the adult head of *Drosophila*. The split of the mir-2 cluster in *Drosophila* triggered a subfunctionalization event through decoupling of the transcription machinery leading to a change in the spatial expression patterns in the second cluster.

Due to their high specificity of the seed region miRNAs are under high selective pressure to keep their sequences conserved. In fact mutated miRNAs have been shown to be involved in different diseases. Mutations or change of expressions patterns of miRNAs have been linked to different types of cancer (Haller et al., 2010; Wu et al., 2008) or hearing loss in mice (Lewis et al., 2009).

miRNAs can be found in intergenic regions, as well as in introns. The majority of miRNA loci an be found in intronic regions, either of non-coding transcripts (∼40 % of known loci) or protein-coding transcripts (∼40 % of known loci) (Kim et al., 2009b). In mammals miRNAs tend to cluster with ≤ 10 kb distance together with other miRNAs (∼50 % of miRNAs can be found in close proximity to each other) (Kim et al., 2009b). These clusters are one transcriptional unit and are transcribed together (Lee et al., 2002). From these clusters the pri-miRNAs are formed and further processed. In humans miRNAs can be found on all chromosomes except the Y chromosome (Kim and Nam, 2006).

Mir-196 for example targets mRNAs from the homeobox gene (Hox gene) cluster and is located inside this gene cluster (Yekta et al., 2008). The Hox genes play a major role in vertebrate limb development and are highly conserved. Mir-196 is not the only miRNA found inside the Hox cluster. Both mir-196 and mir-10 are located inside one of the Hox gene clusters and regulate the expression of different Hox genes.

All of the above describes miRNAs in animals.

miRNAs also exist in plants. They are also ∼22 bp long, however the secondary structure of the precursor miRNAs is different, and the categorisation into families varies from animals. Plant miRNAs families are bigger than animal ones, and in contrast to animal miRNAs the whole mature sequence is conserved between members of the same family and not only the seed region as is often the case in animals (Jones-Rhoades et al., 2006; Bartel, 2004). The secondary structure and the sequence not belonging to the mature miRNA, including the loop region, however can vary between members of the same miRNA family (Jones-Rhoades et al., 2006). Mostly plant miRNAs are found in protein-coding genes lacking regions where they can form clusters and seem to have their own transcriptional units (Jones-Rhoades et al., 2006). However, miRNA clusters in plants are rarer than in animals.

Because miRNAs have a specific function and a limited target list the accepted standard for miRNA loss is that it rarely happens if they have accumulated a function (Tarver et al., 2013). Some recent studies challenging this view through the proposal of a loss of 80 % miRNA families, depending on the species, lead to a big analysis of microRNAomes by Tarver et al. (2018). They took a curated set of miRNA families present in Eumetazoa and analysed the miRNA families present in these lineages in combination with a phylogenetic analysis. Additionally they compared the results of miRNA diversification between their curated data set and an uncurated one. With this they showed that miRNAs are rarely lost, but that a small amount of families is responsible for nearly 50 % of the losses.

PIWI-interacting RNAs

PIWI-interacting RNAs (piRNAs) are important in the process of silencing transposable elements (TEs) (Kim et al., 2009b). The name comes from their interaction with the PIWI clade of the Argonaute protein family. This PIWI clade is present in all animals but is absent in plants and fungi (Grimson et al., 2008). It is however also present in ciliates and slime moulds, leading to the assumption that it is an ancient mechanism

(Aravin et al., 2007). The proteins in this protein clade, for example the name giving Piwi, Aubergine (Aub) and Ago3 in *Drosophila*, have been known longer than the small RNAs they interact with. In most animals the PIWI proteins are only expressed in germline cells. The PIWI proteins of mice and *Drosophila* are not orthologs to each other and are in fact more closely related within a species than between two species (Senti and Brennecke, 2010). In *Drosophila* the PIWI proteins are expressed in both male and female germline cells, in mice however the PIWI proteins MIWI, MILI, and MIWI2 are only expressed in male germline cells (Aravin et al., 2007). Individuals with mutated proteins of this clade show defects in their germ cell development (Aravin et al., 2007). In *Drosophila* it was shown that the expression of the three PIWI proteins varies between cells. Germline cells express Piwi, Aub, and Ago3 cytoplasmic, whereas somatic cells express only Piwi in their nucleus (Senti and Brennecke, 2010; Brennecke et al., 2007; Chambeyron and Seitz, 2014).

Aravin et al. (2001) first discovered a dsRNA associated with the silencing of the repeat locus *Stellate* in *Drosophila*. They called this dsRNA *Suppressor of Stellate*. Through further studies a new category of small RNAs was discovered called rasiRNAs, where RNAs involved in repeat silencing were categorised into (Aravin et al., 2003). Now the rasiRNAs are handled as a subcategory of piRNAs (Aravin et al., 2007). The rasiRNAs are not specific to a type of repeats but consist of sequences ofDNA transposons, satellites, retrotransposons, as well as complex repeats (Aravin et al., 2003). Transposons are mobile elements of the genome that can reproduce and insert themselves in the genome (Slotkin and Martienssen, 2007). They target protein-coding regions for their insertion into the genome and are therefore able to disrupt genes and the organisational structure of the genome. They are found all throughout eukaryotes (Huang et al., 2017).

piRNAs are ∼25-30 bp long (Grimson et al., 2008; Girard et al., 2006) and tend to be found in clusters in the genome (Girard et al., 2006; Chambeyron and Seitz, 2014). Unlike miRNAs their sequence is so unique for each piRNA that is is not possible to classify them into families (Huang et al., 2017). Of those piRNAs that bind to Piwi in the cell soma 75 % carry a uridine at the 5' end, and over 60 % can be mapped to multiple genomic loci (Senti and Brennecke, 2010). During the discovery of the siRNA pathway

it has been shown that both sense and antisense strands of these RNAs can induce gene silencing. For the piRNAs that are annotated as matching to a transposon, over 90 % of the transcripts are antisense to the active transposon. It is important for an organism to control transposons to increase their fitness (Hua-Van et al., 2011). It is possible to pass on an immunity to a specific transposable element, but only through the female germline (Bregliano et al., 1980). This allows a defence against new transposons where the matching piRNAs have not yet been included in the genome. piRNAs have slight variances in their length and sequence, which plays a role in their binding to a protein of the PIWI class. Those with a 5' terminal uridine tend to bind to Piwi and Aub, whereas the ones binding to Ago3 lack this terminal uracil most of the time (Brennecke et al., 2007). Also, the length of the piRNA is a deciding factor. piRNAs bound to Piwi are the largest with 25.7 bp mean length, and Ago3 bound are the smallest (24.1 bp) (Brennecke et al., 2007). The size difference however does not have an impact on the corresponding genomic elements. As stated before most piRNA-Aub complexes interact with sequences that are antisense to the transposable element. If the piRNA is part of the Ago3 complex however, a strong bias towards sense transposon strands (75 % of the transposon strands are sense) is observed (Brennecke et al., 2007).

piRNAs are not randomly distributed in the genome and can be found in clusters (Girard et al., 2006; Chambeyron and Seitz, 2014). Girard et al. (2006) identified piRNAs in mice and found clusters with 10-4,500 piRNAs spanning 10-83 kb. These clusters tend to occur in repeat- and gene-poor regions of the mice genome. piRNA clusters can form at any position of the genome. However, they show a preference to regions that contain remnants of TEs (Olovnikov et al., 2013). Experiments showed that artificial sequences inserted into a piRNA cluster were treated and expressed as piRNAs, showing that any sequence inside this cluster can act as a piRNA (Muerdter et al., 2012; Olovnikov et al., 2013). It is not yet known what conditions have to be met to create a cluster of piRNAs in the genome.

In *Drosophila* the 15 largest piRNA clusters are responsible for up to 70 % of all piRNAs and 57 % of the unique piRNAs (Brennecke et al., 2007). One well studied example of

piRNA cluster is the *flamenco* cluster present in the *Drosophila* genome. This locus can be found on the X chromosome of *Drosophila* and spans over 180 kb (Zanni et al., 2013). 87 % of the sequence of this cluster are transposable elements (Brennecke et al., 2007). *flamenco* has been shown to control three different retrotransposons: *ZAM*, *Idefix*, and *gypsy* (Prud'Homme et al., 1995; Desset et al., 2003). The sequences of the transposons are included in this cluster, some in multiple copies and additional fragments. Further analysis of this locus in several *Drosophila* species showed it acts as a trap for TEs that are transferred horizontally between species and contains also recent insertions (Zanni et al., 2013). The age of the different TE inserted into a piRNA cluster can differ, leading to the possible presence of both old and recent copies of them in the same genome.

Studying piRNAs in mice Girard et al. (2006) found that only 17 % of piRNAs mapped to repeats, whereas in *Drosophila* nearly 80 % of the piRNAs identified by Brennecke et al. (2007) could be classified as rasiRNAs due to their repeat association. Since the numbers for repeat association depend on the repeat annotation and especially the annotation of transposable elements this number is likely underestimated (Chambeyron and Seitz, 2014).

piRNAs are not only transcribed from piRNA clusters, but in some cases also directly from individual transposons or the 3' UTR of some genes (Huang et al., 2017). The different origins lead to slightly different biogenesis pathways, but the piRNAs are always processed from longer precursors. So far no common secondary structural motifs or sequences have been found in the piRNA precursors (Huang et al., 2017).

To transcribe piRNA clusters Pol II is involved, which transcribes them as long non-coding RNAs. The transcription happens even though the piRNA clusters are enriched with the histone 3 lysine 9 tri-methylation (H3K9me3) that usually is found on silenced, heterochromatic regions. Here, the mark does not suppress the transcription, but instead is a necessary requirement for the expression of piRNAs (Huang et al., 2017). Furthermore, in Diptera a specific set of proteins is bound to dual-strand clusters, but not uni-strand cluster or genic piRNAs. A dual-strand cluster has no distinct promotor, no splicing, and allows transcription from both strands. This protein set consists

of Rhino (Rhi), Deadlock (Del), and Cutoff (Cuff), and together they form the RDC complex, which is also necessary for transcription (Huang et al., 2017). In non-Diptera species the transcription initiation complex (TREX) is a requirement for piRNA biogenesis. The transcribed precursor piRNA is exported to the cytoplasm, where it is further processed into mature piRNAs.

The further processing of the piRNA precursors involves two pathways, the Zuc-dependent and ping-pong loop processing. The processing of mature piRNAs includes the formation of the 5' and 3' ends. As stated above a bias for an uridine exists at the 5' end of the mature piRNA. The cleavage of this end can be done in several ways. The first is Zucchini (Zuc) mediated processing in nurse and follicular cells (Huang et al., 2017), where the 3' end can be formed through cleavage with Zuc. Other ways are sliver cleavage (as part of the ping-pong loop) and further processing through other exonucleases. In all theses mechanisms the last step is the 2'OMe-modification of the last nucleotide by Hen1. This is probably stabilising the piRNA (Huang et al., 2017). This processed piRNA is loaded onto the Piwi protein and after methylation of the piRNA the Piwi-piRNA complex is transported into the nucleus, where the mature piRNA silences transposons (Ku and Lin, 2014).

The second processing pathway and an important defence against active transposons is the piRNA ping-pong loop. This is an auto-amplifying biogenesis pathway possible trough a sequence feature in piRNAs (figure 1.3) creating antisense piRNAs complementary to an expressed transposon, whereas sense piRNAs are transcribed from a piRNA cluster (Chambeyron and Seitz, 2014). The first 10 bp of the sense and antisense piRNAs are in general complementary to each other (Olovnikov et al., 2013). This feature makes it possible for one mature piRNA to guide the cleavage of a piRNA precursor that is complementary, leading to the maturation of this piRNA (Chambeyron and Seitz, 2014; Brennecke et al., 2007). This loop uses the proteins Aub and Ago3 which tend to have piRNAs bound that are complementary to each other. The binding is referred as the protein being loaded with the piRNA. As stated above the piRNAs loaded onto Aub often have a 5' uridine whereas the piRNA loaded onto Ago3 tend to have an adenine at position 10 (Chambeyron and Seitz, 2014). The loop does not only produce mature

piRNAs through auto-amplification, but also degrades mRNAs of TEs. The TE mRNAs are degraded through either Aub or Ago3, leading to a post-transcriptional repression. The TE are recognised by the proteins because the piRNAs are sense or antisense to the TE sequence (Chambeyron and Seitz, 2014). This process happens inside the nurse cells (Huang et al., 2017) and is also responsible for the sliver cleavage of the 3' end of the mature piRNA. So the ping-pong loop is an adaptive immune response that destroys active TEs through the amplification of piRNAs (Lau et al., 2009).

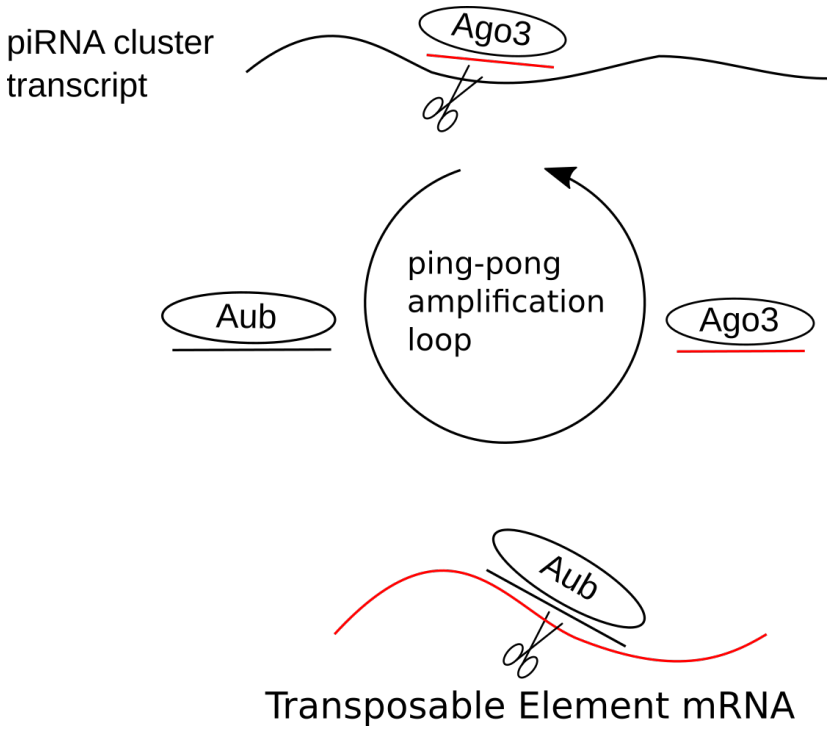


Figure 1.3.: Visualisation of the ping-pong loop that piRNAs and transposable elements are involved in. The process happens in the cytoplasm. The two PIWI proteins involved are Ago3 and Aub. Aub binds to a subsequence of a transposable element mRNA whereas Ago3 binds to a partial piRNA. These protein-sequence complexes start the so called ping-pong amplification loop. The Aub complex binds to a TE sequence that is complementary to the piRNA. The bound part of the TE sequence is cleaved and separated from the Aub complex.

The Piwi protein is not involved in the ping-pong loop. However, it plays a role in the deposition of heterochromatic marking of TEs (Chambeyron and Seitz, 2014). The

piRNA acts as a guide for Piwi by binding to the TE transcript, which triggers a local heterochromatinization of the target gene and its neighbours (Chambeyron and Seitz, 2014).

Even though a lot of the components of the piRNA biogenesis have been identified, there are still some open questions, such as how piRNA clusters are identified or which proteins are involved in this pathway, or how the first piRNA is selected that is required to recognise piRNA precursors (Huang et al., 2017).

Studies showed that piRNAs are present in *Drosophila* embryos (Huang et al., 2017). The mother deposits Piwi proteins loaded with piRNAs directly into the embryo, leading to an epigenetic inheritance of piRNAs through the maternal line (Chambeyron and Seitz, 2014).

Small interfering RNAs

Small interfering RNAs (siRNAs) are dsRNA that are involved in RNAi. They can be found in all lineages of eukaryotes (Zamore and Haley, 2005). Longer dsRNAs are cleaved by Dicer, an RNase III enzyme into a dsRNA duplex with a length of around 22 bp. Characteristic for this duplex is the symmetric 3' nucleotide overhang on each end of 2-3 bp as well as the 3'-hydroxyl and 5' phosphate groups (Dykxhoorn et al., 2003). The cleavage of the dsRNA happens in the cytoplasm. The mature siRNA is then incorporated in the RISC which requires the 5' phosphorylation. Only the antisense strand guides the RISC to the target. The target is identified through the sequence homologous to the siRNA and cleaved at a single centred site. This site is 10 bp away from the 5' end of the siRNA (Dykxhoorn et al., 2003).

For the function of siRNAs the complementarity between mRNA and siRNA is the most important part. A single nucleotide mutation at the wrong position can destroy the activity of the siRNA, whereas mutations at other positions can just lead to a down-regulation of activity (Bantounas et al., 2004).

Studies in *Schizosaccharomyces pombe* showed that siRNAs are not randomly distributed in the genome (Cam et al., 2005). They tend to cluster in heterochromatic domains as

well as in the vicinity of repeat elements that were corresponding with heterochromatic domains.

The distinguishing factor for miRNAs and siRNAs is not their function, but the origin of the transcripts. siRNAs derive from dsRNAs that are up to thousands of basepairs long, whereas miRNAs derive from the pre-miRNAs that are around 70 bp long and are ssRNAs (Zamore and Haley, 2005).

1.1.2. Transfer RNAs

Transfer RNAs (tRNAs) belong to the more commonly known types of ncRNAs. They tend to have a typical clover leaf secondary structure (fig. 1.1 b), however some exceptions miss one or more of the arms and just contain the anticodon loop. They are 75-90 bp long and are involved in mRNA translation. Each tRNA has an anticodon that interacts with a specific amino acid. They can be found as well in the mitochondrial genome in Metazoa (Ojala et al., 1981) and in plant chloroplasts (Leis and Keller, 1970).

The processing of tRNA includes the synthesis of a precursor tRNA that has a 5' leader sequence as well as a 3' ending (Phizicky and Hopper, 2010). The 5' end is clipped by RNase P, the 3' end by the endonuclease RNase Z as well as different exonucleases. At the 3' end a CCA is added after the trimming if not already present.

tRNA genes can contain introns. At least one tRNA family with introns is present in so far all sequenced archaea and eukaryotes, and in at least one tRNA family all members contain one, making splicing a necessity (Phizicky and Hopper, 2010). However, tRNAs with introns are the minority.

In yeast genomes the tRNAs are randomly distributed on the chromosomes. The transcription of these genes only happens in the nucleolus (Phizicky and Hopper, 2010). The location of the splicing machinery for tRNAs varies depending on the organism. In vertebrates the splicing happens in the nucleus, whereas in yeast it occurs in the

cytoplasm. If not already there the mature tRNA is exported to the cytoplasm, where it is either charged with an amino acid or, after some modifications, reimported into the nucleus. The reimported tRNA is charged with an amino acid inside the nucleus and then re-exported (Phizicky and Hopper, 2010).
Over 100 tRNA modifications are known so far, making tRNAs the most heavily modified cellular RNA (Vilardo et al., 2012). The modifications include changes in the anticodon region, different kinds of methylations, and pseudouridinylations. These changes can stabilise the 3D structure and expand the coding capacity of the anticodon. Even though not all modifications are completely understood they are often a necessity (Vilardo et al., 2012).

1.1.3. Small nucleolar RNAs

Small nucleolar RNAs (snoRNAs) are mostly involved in rRNA modification, but also in ribosomal RNA processing. They are ∼70-250 bp long and can be classified into families based on secondary structure, the two most prominent ones being H/ACA snoRNAs and C/D box snoRNAs (Maxwell and Fournier, 1995). In these two families the secondary structure is conserved, however the sequence can vary substantially. H/ACA snoRNAs consist of two stem loops that contain each one interior loop where the target area of the target rRNA is captured (figure 1.1 d). C/D snoRNAs have one big loop where the target region binds (figure 1.1 c). Both types have so called boxes which consist of a conserved nucleotide sequence and are needed for metabolic stability or help to fold the snoRNA into the correct secondary structure. In both types of snoRNAs the boxes function as measuring devices to get the specific position of the rRNA where the modification should happen.
C/D snoRNAs have a C box (UGAUGA) near the 5' and a D box (CUGA) near the 3' end, and if folded the boxes are near the stem but inside the loop (Eliceiri, 1999).
H/ACA snoRNAs have an ACA motif three nucleotide from the 3'end and an H box (containing the nucleotide pattern ANANNA) that is located in the hinge region between the two loops (Eliceiri, 1999).

snoRNAs are involved in different stages of rRNA pathways. Both C/D and H/ACA snoRNAs are responsible for cleavage of pre-rRNAs, folding and nuclear exportation (Eliceiri, 1999; Henras et al., 2015). C/D snoRNAs are also involved in pre-rRNA ribose methylation, specifically 2'-O-methylation. They use an antisense element to the rRNA target inside the loop in combination with the boxes. The H/ACA snoRNAs also function as pre-rRNA pseudouridylation guides that are site-specific. They also use an antisense element inside the loop to target the rRNA.
Even though there are only two families of snoRNAs there are around 200 different snoRNAs suspected to exist in a single vertebrate cell (Smith and Steitz, 1997).

1.1.4. Ribosomal RNAs

Ribosomal RNAs (rRNAs) are part of the ribosomal complex and are involved in protein synthesis. rRNAs are known from all organisms, however the types that are present vary. In eukaryotes 5.8S rRNA, 28S rRNA (in Metazoa)/26S rRNA (in plants), 18S rRNA, and 5S rRNA can be found. They are split into two transcriptional units, one containing only the 18S rRNA, called small transcriptional subunit (SSU) or 40 S subunit, and the other containing the 5S rRNA, 5.8S rRNA, and 28S rRNA, called large transcriptional subunit (LSU) or 60 S subunit (Srivastava and Schlessinger, 1991; Fatica and Tollervey, 2002).
The 5.8S, 18S, and 28S rRNAs are found in series in the genome and are also often referred to as the ribosomal DNA (rDNA) cluster. The cluster composition is highly conserved. The cluster starts with the 18S rRNA followed by the 5.8S rRNA and the 28S rRNA (figure 1.4). The three subunits are separated by internal transcribed spacer (ITS), in this case ITS1 between 18S and 5.8S, and ITS2 between 5.8S and 28S. The ITS are less conserved than the different subunits. The whole cluster can be found multiple times in a genome.

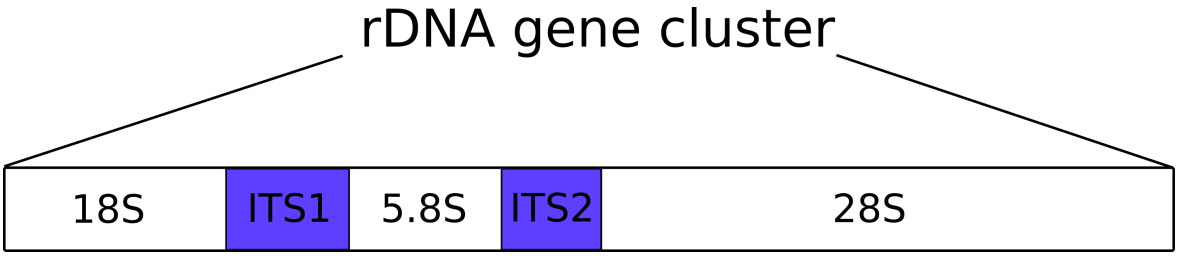


Figure 1.4.: Graphical overview of the rDNA cluster. It consists of an 18S rRNA gene, followed by the first internal transcribed spacer (ITS1), the 5.8S rRNA gene, the second internal transcribe spacer (ITS2), and the 28S rRNA gene.

The 5S rRNA is also found in multiple copies in the genome and those copies occur in several clusters of tandem repeats (Ciganda and Williams, 2011). The single genes have a length of about 120 bp. The sequence is highly conserved and often used as a phylogenetic marker. The secondary structure consists of four loops (two internal, two hairpin) and five stems. One internal loop connects to three stems and acts as a hinge (Ciganda and Williams, 2011).

In arthropods some repeat elements are known to insert themselves into rRNAs. One of these elements is the R2 element. The R2 element is a non-LTR retrotransposon that inserts itself at a specific position into the 28s rRNA (Burke et al., 1999). This insertion is known throughout arthropods. The R2 element is not necessarily present in all 28S rRNA copies in a species. Some copies may have no insertions, others may have an insertion of another R element (Jakubczak et al., 1991).

Due to the high copy number of rRNA genes their assembly is problematic and often only single copies of a gene can be found in a genome assembly. This also leads to problems in identifying whole rRNA clusters. However, recent developments in sequencing will change this. Using PacBio or Oxford NanoPore Technology machines it is now possible to sequence several thousand basepairs continuously, or even a whole DNA molecule.

1.1.5. Long non-coding RNAs

Even though lncRNAs are classified as non-coding RNAs they differ quite a lot from the other ncRNA types. They are defined as transcripts that are longer than 200 bp, but do

not code for proteins and have in contrast to other ncRNA transcripts no conserved secondary structure, and mostly have a poly(A) tail. Since the first discovery, the number of lncRNAs constantly increases and it is now thought that in humans the lncRNAs are more abundant than the protein-coding genes (Quinn and Chang, 2016). Originally, the transcripts of lncRNAs were thought to be just noise without a function. But further experiments lead to the identification of functions. Already in the early 1990s some transcript with functions that did not follow the typical protein-coding gene transcript patterns were identified (Brannan et al., 1990). It took until the 2000s for the term of lncRNA to come up and classifying and naming the first lncRNA HOTAIR (Rinn et al., 2007; Baker, 2011).

lncRNAs are involved in several different processes in an organism. They play a role in imprinting of genomic loci, dosage compensation, regulating enzyme activity, or coordinate cell differentiation and development. Also, quite some lncRNAs are involved in diseases through either a changed expression level or mutations of their sequence (Wapinski and Chang, 2011).

Even though the general functions of lncRNAs are known, only for a small number of lncRNA their specific function is known. In the human genome out of tens of thousands of known lncRNAs only for 299 a known specific function is listed (Jandura and Krause, 2017). The focus of databases is still heavily biased towards model organisms and vertebrates, especially mouse and human. In insects a bias towards certain species is also present.

The biogenesis of lncRNAs is similar to that of mRNAs as they are often 5'-capped, spliced, and polyadenylated, and contain exons. However, they have fewer, but longer exons compared to mRNAs, and they lack an open reading frame. They are also expressed at lower levels and have poor sequence conservation. This poor sequence conservation also creates problems when identifying homologous lncRNAs between different species.

While lncRNAs can be found in all tissue types their expression level varies. In *Drosophila* ∼70 % of the known lncRNAs are expressed in the testes, in humans 78 % (Jandura and Krause, 2017).

lncRNAs can be categorised into genic and intergenic ones, the latter often called long intergenic non-coding RNAs (lincRNAs). Genic lncRNAs overlap with a protein-coding gene, but are often found on the opposite strand in antisense to the gene.

Legeai and Derrien (2015) looked at the lncRNAs known in insects and at ways to identify them. Most research regarding lncRNAs is focused on *D. melanogaster* and *Apis mellifera*. Even if the lncRNA focus is on these two species, in their study Legeai and Derrien (2015) only list ten lncRNAs as well studied, i.e. for them the function is characterised. In other insect model organisms they could not find a detailed functional analysis of lncRNAs.

lncRNA annotation faces the problem that they are mostly conserved in function, but not their sequence or secondary structure. This means that the standard ways to identify homologous ncRNAs, where the sequence and structure of candidate hits between different species is compared, does not work here. To identify lncRNAs in genomes the coding potential of a candidate is calculated and combined with mapped reads.

1.1.6. Current state of non-coding RNA research

Not all known ncRNAs types can be found in all organisms. miRNAs can be found in animals as well as in plants. However, due to their differences in structure, biogenesis, and how they work on targets it is assumed that they have independent origins with similar functions (Grimson et al., 2008). The number of miRNAs increases with the complexity of organisms, leading to the idea that they played an important role in the increasing complexity (Grimson et al., 2008).

Several databases exist that only contain ncRNA. The most prominent ones are probably Rfam (Nawrocki et al., 2014) and miRBase (Kozomara and Griffiths-Jones, 2013). The Rfam is a general database for ncRNAs and contains all types of ncRNAs as well as

RNA elements and motifs that can be present in different ncRNA families. In version 12 the Rfam contains 2,450 different ncRNA families. With version 13 this number was updated to 2,686 ncRNA families. Rfam uses seed regions for each family in combination with covariance models to identify ncRNA candidates in a genome. The results are then manually curated to create a high quality sequence background for each family. This database contains sequences from all areas of life, however ∼60 % are bacterial. The rest are from viruses, Eukaryota, and archaea. With version 13 of the database the focus shifted to annotate full genomes and use those as reference genomes to reduce data redundancy (Kalvari et al., 2018).

The miRBase is a more specialised database, as it contains only miRNAs. Version 21 contains 1,983 different miRNA families. The miRNAs are mostly from Eukaryota and plants as well as some protists and viruses. The database collects the stem loop of a miRNA and marks the mature sequence. Ideally, experimental evidence is also provided. The shortness and relatively simple secondary structure of miRNAs lead to an inflation of false positives in the database (Kozomara and Griffiths-Jones, 2013; Ludwig et al., 2017).

Both databases do not contain all known ncRNAs (of the ones they curate) because they rely heavily on users to submit ncRNAs to be included in the database. They are still the most useful for comprehensive analyses because they provide family alignments and secondary structure information. There do exist other ncRNA databases, but they are smaller and often organism or ncRNA specific, i.e. snoRNAbase (Lestrade and Weber, 2006), tRNAdb (Jühling et al., 2008), noncode (Fang et al., 2017). Especially for lncRNAs quite a few different databases exist, i.e. lncrnadb (Quek et al., 2014), lincipedia (Volders et al., 2014). The most extensive databases for lncRNAs contain only information on human ones.

If one looks at the documented numbers of a certain ncRNA type in the databases a huge variety between the actual count of e.g. miRNAs can be found. In miRBase v21 the number of annotated precursor miRNAs in insects varies between 7 (*Locusta migratoria*) and 487 (*Bombyx mori*). In humans currently 1,881 precursor miRNAs are

annotated. The huge difference leads to the question if this number is real as a result of different complexity of the organisms and corresponds, e.g., to phenotype changes, or if it is a result of understudying, or a false annotation due to a lack of data. A study by Wang et al. (2015) reported 833 identified miRNAs in *L. migratoria* of which the miRBase only lists 7. The huge difference in numbers is most likely due to the limited data that was available during the first study, which was done using only transcriptome data as no genome was available at that time (Wei et al., 2009).

Of course, for most species it is impossible to identify all present miRNAs through homology prediction. In an understudied group all or most lineage specific miRNAs are most likely missing from the database. This highlights the importance of using a combination of homology and *de novo* prediction for all species to identify their ncRNA repertoire. This problem was enhanced in the past by the focus on model organisms. Although now the focus is shifting from model organisms to non-model organisms, a lot of genome projects still focus on the protein-coding part of the genome. This means that even though the genomes are available, they do not necessarily include ncRNAs annotations. Genomes that are submitted to the NCBI are run through their ncRNA pipeline (Thibaud-Nissen et al., 2013). This pipeline is supposed to identify ncRNAs in genomes. However, the pipeline is not well documented and the total number of ncRNAs types as well as ncRNA genes identified in the genomes is very low. Up to version 8.0 only miRNAs and tRNAs were annotated through this pipeline. After the release of version 8.0 in November 2017 rRNAs, snoRNAs, and snRNAs were added to the annotation pipeline. Another aspect is that these are generally not manually curated and not sent to the miRBase/Rfam to be included in the databases, however they rely on both miRBase and Rfam for their own annotations.

How well ncRNAs are studied varies between the different classes. Some, e.g. tRNAs and rRNAs, are quite regularly annotated in genome and other projects, making these type of data available for a variety of species. Others, like miRNAs or snoRNAs, are quite

often only annotated for specific questions, such as their involvement in a certain gene regulation, but rarely the complete repertoire of a species is catalogued. Contrasting to other ncRNAs, tRNAs are nearly always predicted and researched in genome projects.

1.1.7. Non-coding RNAs in insects

Most research of ncRNAs in insect has been done on *Drosophila* so far as it is a model organism with a well annotated genome and easy availability of specimens for further sequencing. Ylla et al. (2016) looked into the question if ncRNAs are just less studied or if a real difference exists between species, and tried to identify the miRNA toolkit of insects to answer it. In this case they were especially interested in the change between hemimetabolan and holometabolan insect species. They were able to identify 65 conserved miRNA families in the insect species they looked at. Also, they conclude that the variation in miRNA family number in insects is an artefact due to inaccurate annotation and poor sampling. However, they also suggest to broaden the scope to identify more lineage specific miRNAfamily gains and losses.

The focus of genome research in insects lies often on protein-coding genes. This leads to only a small number of the available insect genomes being annotated with ncRNAs, and even less are added to the specialised databases. For example the genomes of several ants do exist, but none of their sparsely annotated ncRNAs were added to the ncRNA databases. This reduces the available genomes for comparison drastically.

One of the larger comparative analyses of ncRNAs in closely related species happened as part of the 'Anopheles Genomes Cluster Consortium' where 16 *Anopheles* genomes were sequenced and analysed (Dritsou et al., 2014). Using computational approaches they focused on tRNAs, miRNAs, rRNAs, and snoRNAs, and also analysed their genomic context. For this study the species set was expanded to include 20 different *Anopheles* genomes as more genomes became available.

Using a combination of known sequences present in VectorBase and the Sequence Read Archive they identified the different rRNA genes in the Anophilids. The completeness and copy number of the different rRNA genes varied between the species. In both 18S

and 5.8S rRNA they were able to identify at least partial sequences in the majority of species. The 18S rRNA was found at least partially in 15 out of 17 species and not found in two, and the 5.8S rRNA was found as full sequences in 16 out of 19 species. In case of the 28S rRNA only in four out of 17 species a complete gene was identified, with one additional large partial gene. In nine species the found sequences were shorter than half of the expected length of the 28S rRNA and in three species no 28S was found. The partial genes may be the result of TE insertions, but this was not analysed in this study. The 5S rRNA was the only rRNA were for all 19 analysed species a complete sequence was identified.
The method they applied for the identification of snoRNAs produced only C/D snoRNAs. Even though for some species it was necessary to use the target rRNA of a closely related species for the analysis they were able to predict between 29 and 460 snoRNAs. Some of the snoRNAs are shared between distant related species with fully conserved sequences (Dritsou et al., 2014).

1.2. Conserved non-coding elements

Genomes contain a variety of different elements that are not coding for proteins, such as non-coding RNAs, transcription factors or repeats. During the first analyses of the human genomes some parts of the genome were identified that are not coding for above mentioned elements, but were more conserved than expected. This led to the assumption that these regions have some functions. They were called conserved non-coding elements.

1.2.1. Characteristics of conserved non-coding elements

Conserved non-coding elements (CNEs) are regions of genomes that are conserved between species and are not protein-coding, repeats or non-coding RNAs. They were first described in humans as sequences longer than 200 base pairs that are 100 % con-

served (Bejerano et al., 2004). There is no general consensus about the definition of the minimal length and sequence conservation of conserved non-coding elements (CNEs). Different studies use a different minimal length of CNEs such as 45 bp (Yue et al., 2016), 100 bp (Woolfe et al., 2004), or 200 bp (Bejerano et al., 2004), as well as different conservation over this length, such as 70 % (Woolfe et al., 2004), 90 % (Yue et al., 2016) or 100 % (Bejerano et al., 2004). CNEs with 100 % conservation are often classified into another category called ultraconserved elements (UCEs). For these, usually a shorter sequence length is assumed (often 50 bp) (Glazov et al., 2005).

These elements are often found in clusters as well as regulatory blocks with a gene (Polychronopoulos et al., 2017). The definition of the maximum distance of two CNEs for them to belong to a cluster varies. The gene a CNE regulates can be found in a distance up to 500 kb (Woolfe et al., 2004).

To identify conserved elements at least two different genomes of different species are compared. The first studies in humans started with the first draft of the human genome, where it was compared to mice genomes (Hardison, 2000). They studied only one locus which contained 90 conserved non-coding sequences (CNSs), but extrapolating from it they suspect 270,000 CNSs in the whole human genome. A later study looked for UCEs conserved between humans, mice, and rats, which identified 481 fully conserved sequences (Bejerano et al., 2004). They used whole genome alignments to identify the conserved regions. The last common ancestor of human and rodents existed ∼60 million years ago (mya), but still it was possible to identify over 400 fully conserved regions that are longer than 200 bp. Bejerano et al. (2004) also included the pufferfish to figure out if a time limit for the identification of CNEs exists. The last common ancestor between puffer fish and mammals was 430 mya (Aparicio et al., 1995). The puffer fish genome is a lot smaller than the human one, but it was possible to align 12 % of its genome to the human genome. Woolfe et al. (2004) were able to identify nearly 1,400 CNEs between the puffer fish and mammals. An observation they made was that the CNE set conserved between the mammals and the set conserved in the vertebrates overlapped

only partially. This led to the assumption that CNEs emerge over time and are not only an ancient remnant. Overall it has been shown that it is possible to identify non-coding sequence conservation after more than 400 million years (my) of evolution in vertebrates and that this conservation exceeds the conservation of protein-coding genes (Polychronopoulos et al., 2017).

Table 1.1.: A selection of different definitions of conserved non-coding elements and ultra-conserved elements.

Min. length	Min. conservation	Author	Element name
45 bp	90 %	Yue et al. (2016)	conserved non-coding elements
50 bp	100 %	Glazov et al. (2005)	ultraconserved elements
100 bp	74 %	Woolfe et al. (2004)	conserved non-coding sequences
200 bp	100 %	Bejerano et al. (2004)	ultraconserved elements

1.2.2. Function of conserved non-coding elements

Gene regulation

Because CNEs are, as the name says, not coding for anything it was questioned after their discovery if they have a function (Nobrega et al., 2003). The fact that these regions are more conserved than expected by random chance suggests that they are subject to fixating pressure. Studies that focused on the question of functionality showed that CNEs are involved in gene regulation (Glazov et al., 2005). They were identified as enhancers for developmental genes in *Fugu rubripes* (Aparicio et al., 1995) as well as general developmental gene regulation in flies (Warnefors et al., 2016), and it has been shown in humans that some CNEs regulate the expression of certain interleukins (Hardison, 2000). Through trans-mice it has been shown that the expression is downregulated if the CNEs are deleted from the genome.

In their study Warnefors et al. (2016) looked for UCEs and their possible relation to alternative splice site in flies. Focusing on a UCE that overlaps with a small exon in the Hox gene cluster they showed that a mutation in this conserved region leads to a reduced expression of mRNAs. This evidence of functions led to the theory that CNEs are *cis*-regulatory elements that are involved in the coordination of gene expression, especially for developmental genes (Polychronopoulos et al., 2017).
It has also been shown in humans that a disruption in a regulatory block involving CNEs can lead to developmental diseases or cancer (Calin et al., 2007). For the function of the CNE it is therefore important that the organisation of a CNE or a CNE cluster and the regulated gene together with the promotor architecture are conserved (Polychronopoulos et al., 2017). This should show in a synteny analysis of older CNE regulatory blocks in inter species comparisons.
It is very specific which gene is regulated by a CNE and so it can happen that genes are located inside a CNE cluster but are not affected by the regulation (Polychronopoulos et al., 2017). This shows that the position alone of a gene in relation to a CNE is not enough evidence for it to be a potential target.
For vertebrates some characteristics of these target genes have been described. They have longer CpG islands, a certain histone modification pattern, a different distribution of transcription start sites (TSSs) for alternative splicing, and a certain spatial organisation of transcription factor binding sitess (TFBSs) (Polychronopoulos et al., 2017). A closer look at the CNE target genes in *Drosophila* showed that they also have extensive Polycomb binding, and longer introns, that often have a CNE inside (Polychronopoulos et al., 2017).

Results of CNE loss

A loss of a CNE does not necessarily result in a non-viable organism, but can result in a change of phenotype. In snakes, for example, CNEs associated with limb development

genes are partially or fully deleted from the genome leading to the limblessness of snakes (Polychronopoulos et al., 2017).

In a study in mammals Marcovitz et al. (2016) predicted the function of CNEs through so called "reverse genomics". They compared morphological changes between lineages with the loss or gain of CNEs. Overall they identified 2,759 CNEs in humans associated with certain mammalian phenotypes, including an aquatic forelimb CNE, a pelvic CNE, a brain morphology element, and an ear element (Marcovitz et al., 2016). They also assume that the number of CNE and phenotype associations will rise with more sequences genomes and more trait annotations.

If a CNE becomes disease associated a single point mutation can already be enough to create a change in function (Polychronopoulos et al., 2017). Such single point mutations of CNEs are involved with Pierre Robin syndrome, cleft lip, but also in behavioural disorders such as autism or restless leg syndrome. But also complete deletion of CNEs or a duplication can lead to a disease. Diseases associated with a duplication event of a CNE include brachydactyly or syndactyly. CNE deletions can be associated with deafness, Leri-Weill dyschondrosteosis or blepharophimosis syndrome (Polychronopoulos et al., 2017). In all these listed diseases a change in CNEs leads to a phenotypic effect. However, there also exist cases were CNE deletions do not lead to a visible change in phenotype. This has been shown in knock-out mice, where CNE deletions did not lead to phenotype changes (Polychronopoulos et al., 2017). Still the results might differ in wild conditions.

1.2.3. Where are conserved non-coding elements known so far?

In Metazoa CNEs have been found in several lineages and are most studied in vertebrates. Starting with the discovery in humans and mice the research broadened to include several fish species, cephalochordates, and insects. They have not been identified in every lineage in Metazoa so far, but CNEs seem to be an ancient feature of metazoan genomes (Polychronopoulos et al., 2017).

Outside of Metazoa, CNEs are also known in higher plants. There they have been shown to be around genes involved in hormonal stimuli, regulation of organ development, and flowering time (Polychronopoulos et al., 2017). However, they are understudied in regards to their specific roles and the distribution in the genome. So far the assumption is that CNEs are an ancient part of multicellular eukaryotes. How they emerged, are maintained, or whether their function is conserved over all eukaryotic lineages still remains unclear (Polychronopoulos et al., 2017).

1.2.4. Conserved non-coding elements in insects

CNEs have been rarely studied in insects so far. The only group of insects where they have been studied are Drosophilids. But the focus lies on UCEs (Warnefors et al., 2016). UCEs are also used in hybrid enrichment as a targeting tactic. So far baits from UCEs have been created for Hymenoptera (Faircloth et al., 2015) and some other insect lineages (Faircloth, 2017). Still, these studies have a different focus than CNEs shared between species or the genes that they are associated with.
The availability of more insect genomes makes it likely that more research in this direction will be done in the future. The more fully sequenced genomes of a group exist, the better CNEs can be studied, as all methods rely on at least one full genome in combination with other genomes or transcriptomes.

1.3. Aim of this thesis

The focus of this thesis are non-protein-coding regions of insect genomes, especially non-coding RNAs and conserved non-coding elements.
We characterised the ncRNA repertoire of the two Hymenoptera species *Athalia rosae* (Scopoli, 1763) and *Orussus abietinus* (Linnaeus, 1758) through homology and *de novo* prediction.
So far only two Hymenoptera species have a more comprehensive repertoire of ncRNAs characterised: the jewel wasp *Nasonia vitripennis* (Walker, 1836) and the honeybee

Apis mellifera Linnaeus, 1758. Looking at the species richness of the Hymenoptera this number is too low to get a comprehensive overview of the ncRNA repertoire in the different Hymenoptera lineages or even of the Hymenoptera ancestral state of ncRNAs. With this study we add two more Hymenoptera genomes to the well annotated ones, the turnip sawfly *A. rosae* and the parasitic wood wasp *O. abietinus*.

We also identified CNEs conserved between four different Hymenoptera species (*A. rosae*, *O. abietinus*, *A. mellifera*, *N. vitripennis*). CNEs have not been studied in this group so far and also in insects no study with such a distance to the last common ancestor has been done. We chose this set to get a species set that includes as many different annotated gene features as possible. The gene features are necessary to exclude areas of the genomes that are conserved due to a gene function and would therefore not fall under the CNE definition.

2. Methods non-coding RNAs

2.1. Genomic data

The genome assemblies of *Apis mellifera* and *Nasonia vitripennis* were downloaded from GenBank. We used assembly version Nvit 2.1 of *N. vitripennis* (GenBank assembly accession GCA_000002325.2) (Werren et al., 2010) and Amel 4.5 for *A. mellifera* (GCA_000002195.1) (Elsik et al., 2014).

The genome assemblies of *Athalia rosae* and *Orussus abietinus* were downloaded from the i5k server (ftp://ftp.hgsc.bcm.edu/I5K-pilot/). We used assembly version Aros 1.0 of *A. rosae* and Oabi 1.0 of *O. abietinus*.

We sequenced RNA short reads for both *A. rosae* and *O. abietinus* using Illumina machines of the company StarSEQ. From samples preserved in RNAlater, short read libraries with different size ranges were prepared using the Illumina TrueSeq Small RNA kit. One range contained fragments of the size 18-30 bp, the other 30-200 bp, both without strand information. Separated by sex, two short read libraries for each sample were sequenced using Nextseq 500 machines. This resulted in four libraries for *A. rosae* (two different lengths for each sex) and two in *O. abietinus* (no females were available for sequencing).

All RNA short reads were clipped using the program Trimmomatic version 0.33 (Bolger et al., 2014) before further processing. We clipped the Illumina adaptors and kept reads with a minimal length of 18 bp.

2.2. Homology prediction of non-coding RNAs

We relied on two databases for the homology prediction of ncRNAs. The first one is the Rfam version 12 (Nawrocki et al., 2014), which is a database containing 2,450 different ncRNA families in this version, as well as a list of species were the family is identified so far. We excluded the annotated miRNAs and tRNAs because they were either analysed using a special database (miRNAs) or through *de novo* prediction (tRNAs).
The second database is the miRBase version 21 (Kozomara and Griffiths-Jones, 2013), which focuses on miRNAs and currently contains 1,983 different miRNA families.
Both databases contain ncRNAs from all domains of life.
The two databases were kept separate for the analysis, but the handling was the same. For our search we created subsets of the families listed in these databases. To this end, we first removed all ncRNA families that are known to only exist outside of Metazoa (miRBase) or outside of eukaryotes (Rfam). Afterwards we manually curated all remaining families and removed false-positive families. As false-positives we classified such ncRNAs that are majorly found outside of Metazoa but contained one or two hits for Metazoa, and which are also most likely a contamination of a sample. These lists were used for filtering steps later in the analyses.
We searched for ncRNAs in our genomes using the cmsearch script of the program Infernal version 1.1.1 (Nawrocki and Eddy, 2013). It requires a genome file in fasta-format as well as covariance models of the ncRNAs of interest as input. A covariance model is a multiple sequence alignment (MSA) with additional information on the secondary structure of the sequences. Rfam provides a file that contains a covariance model for each family present in the database. The miRBase only provides MSAs for each family. We used the script cmbuild from the program Infernal to create the miRNA-models from stockholm alignments. The stockholm format is a MSA format with a strict layout. The MSAs provided by miRBase are not in this format, so we used the script aln2sto.pl to translate the MSA to stockholm format and created the covariance models out of these alignments by using the Infernal scripts cmbuild, cmcalibrate, and cmpress.
Although we have already created filter lists, we ran the analysis on all families present in the databases and removed false-positives and hits in families outside of Metazoa

afterwards using the script cmsearch_analysis.pl. This stemmed from the experience that using the covariance model file containing all families provided by Rfam and filtering it afterwards is less time consuming. The covariance models were then used with the cmsearch function of the program Infernal to search for ncRNA candidates in the genomes of *A. rosae* and *O. abietinus*. This analysis was done on the complete dataset from the databases and the results were filtered.

Cmsearch returns hits in two confident settings depending on the e-value. All hits with an e-value ≤ 10 are marked with an '?' (weak hits) and all hits with an e-value ≤ 0.01 are marked with an '!' which indicates a reliable hit. All weak hits were filtered out, leaving only the reliable hits to be used in further analysis. Additionally, we removed all hits in ncRNA families not present in Metazoa and all hits on the false-positive list. We aligned the sequence of all remaining hits with the corresponding ncRNA family alignment to manually inspect the fit of the predicted ncRNA with the family. The alignments were created using the cmalign function of Infernal. Using the sequence information provided by each reliable hit we cut out this sequence from the corresponding genome with focus on the predicted directionality of the ncRNA (figure 2.1). The sequence was added to the covariance model file of the corresponding ncRNA family using the Infernal script cmalign. The resulting alignment was manually curated using the ralee mode of emacs (Griffiths-Jones, 2004; Stallman, 1981).

In case of the miRNAs, hits were excluded based on the alignments, if the loop region of the miRNAs was too long, the secondary structure did not fit the expected stem loop structure, or the base pair conservation was too low. Regarding the base pair conservation the focus lay on the seed region of the miRNA. If more than three nucleotides varied from the seed ,the conservation was deemed too low. For this we directly compared the sequences of *A. rosae* and *O. abietinus* with the phylogenetically closest species possible in the alignment. This was a Hymenoptera if present, otherwise another insect or arthropod. If none of these were present, the consensus sequence of the alignment was used.

In case of the other ncRNAs that were curated using the Rfam database a mismatch of the predicted secondary structure was evaluated depending on the type of ncRNA fam-

ily. For example, in case of H/ACA snoRNAs, we checked if two loops were predicted. Also, the sequence conservation was checked in the same way as for the miRNAs. In case of long ncRNAs such as rRNAs we also allowed partial matches.

We compared the ncRNAs predicted through this method with those found in other Hymenoptera in the miRBase and the Rfam. For this we used *Apis mellifera* and *Nasonia vitripennis*, however we included two additional Nasonia species *N. longicornis* and *N. giraulti* to get a more complete picture of the ncRNA distribution in Nasonia. As an outgroup we used *Tribolium castaneum*. For these comparisons we extracted the ncRNAs annotated for these species from the two databases. We then removed all ncRNAs present in our false-positive lists and compared all remaining ncRNAs to our results from *A. rosae* and *O. abietinus*. If a ncRNA was only present in one of those species we additionally did a search with Infernal to check whether it really cannot be found in the remaining species. For this we used the genomes of the Hymenoptera present in the databases (*A. mellifera*, *N. vitripennis*, *N. longicornis* (Nlon 1.0, GCA_000004759.1, (Werren et al., 2010)), and *N. giraulti* (Ngir 1.0, GCA_000004775.1, (Werren et al., 2010)), as well as the *T. castaneum* genome (Tcas 3.0, GCA_000002335.2, (Kim et al., 2009a)) and specifically did cmsearch with the ncRNA family in question.

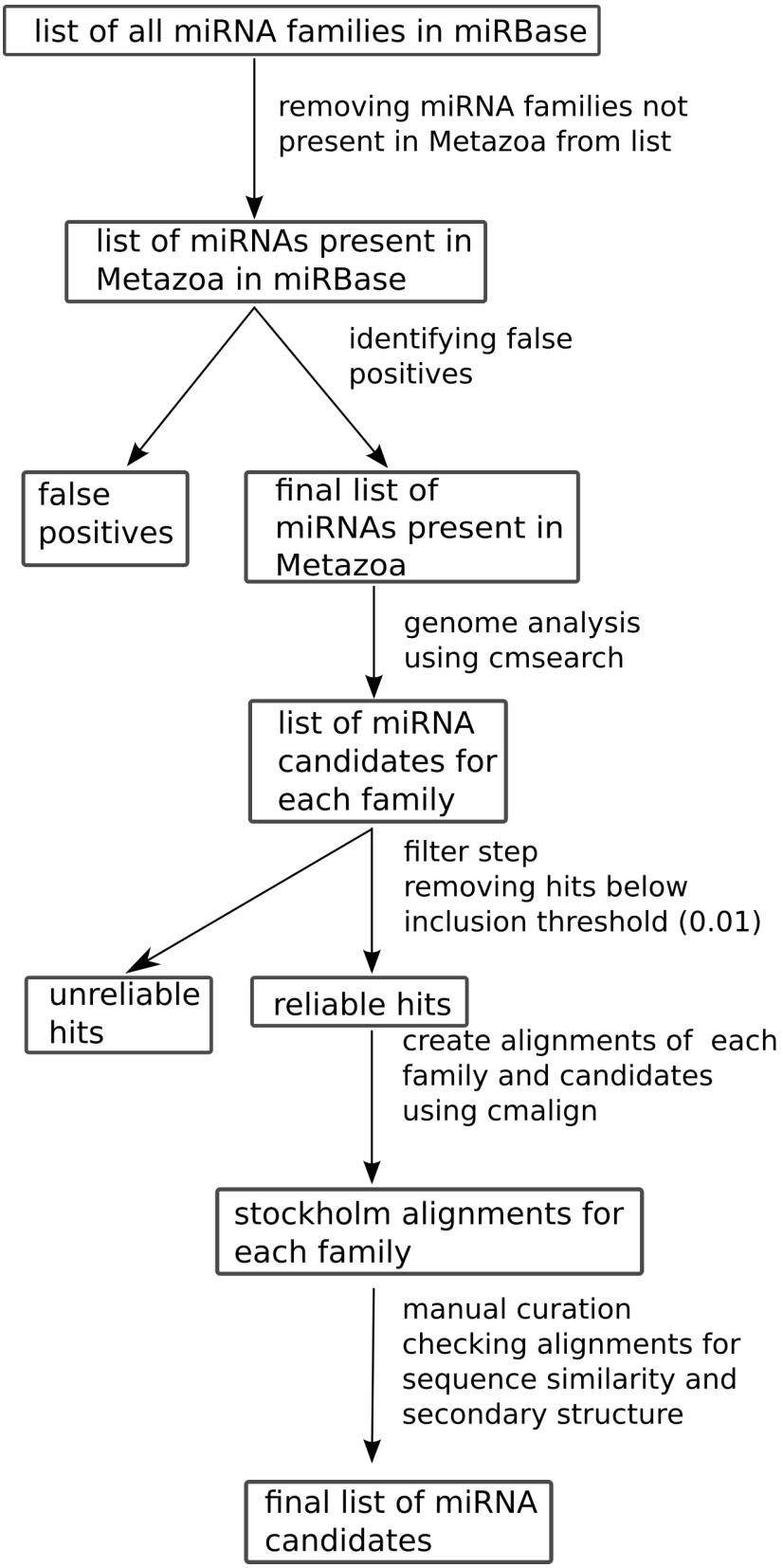


Figure 2.1.: Graphical overview of the steps in the pipeline used for homology prediction of miRNAs.

2.3. *de novo* prediction of non-coding RNAs

Using *de novo* prediction we annotated the tRNAs in *A. rosae* and *O. abietinus* and identified more ncRNAs than predicted through the homology analysis using the DARIO pipeline (Fasold et al., 2011).

2.3.1. tRNAscan-SE

To identify tRNAs we used the program tRNAscan-SE version 1.3.1 (Lowe and Eddy, 1997) with the settings -C -H -o. We discarded all tRNA candidates with a score below 55 and those that were classified as pseudo-tRNAs. This cut-off value removed pseudo-tRNAs, but kept those with introns. We kept tRNAs which were predicted as containing an intron.

2.3.2. DARIO pipeline

The software pipeline DARIO (Fasold et al., 2011) was used to do a *de novo* search for miRNAs, H/ACA snoRNAs, C/D snoRNAs, and tRNAs. DARIO uses an existing ncRNA annotation of these four classes and short reads to classify the read stack pattern of the different ncRNA classes. We did this analysis three times with three different read mapping strategies to identify the best treatment for short reads.

For all three strategies we used the reads were the adaptors were already clipped. We combined the male and female short reads for *A. rosae*. After this step the libraries for *O. abietinus* and *A. rosae* were treated the same.

Our first dataset consisted of only merged reads. For this we took those reads of the short read library with the smaller insert size (read length 20-40 bp) that were still paired after trimming and merged the paired-end reads for each species using bbmerge (version 8.0) of bbmap (version 35.14) (Bushnell, 2014)) with the minimal insert size set to 17 bp. This produced three output files, one containing all successfully merged reads and two with the non-merged reads separated by first or second mate. All reads

that were successfully merged were used in further analysis making up our first set, called 'merged'.

Using the reads of the short read libraries that were left after trimming we created two datasets were the reads were not merged. The one set, called 'paired', contained all reads that still had a partner after trimming and the second, called 'paired_unpaired', contained all reads left after trimming regardless whether they still had a partner or not.

The next step was mapping the reads onto the corresponding genome. Using segemehl (version 0.2.0) (Hoffmann et al., 2009)) with split read option but otherwise default settings, we mapped the three different read sets per species.

DARIO is only available as a web service with a preset of species. For our analyses the developer set up two special data sets with the genomes of *A. rosae* and *O. abietinus*. The mapped reads were used as input for the DARIO pipeline. Due to a restriction of DARIO all reads that were ≥ 54 bp were removed before the analysis.

We used the annotation of miRNAs and both types of snoRNAs from our homology prediction as well as the tRNAs from the *de novo* prediction as the basis for DARIO. DARIO uses the reads mapped onto these annotations to train a random forest classifier the stack pattern of ncRNAs to predict further ncRNAs in these classes.

The three different results, one per input file, per species were compared afterwards. The total amount of predicted ncRNAs were compared by type. We also checked whether the same ncRNAs were predicted, and if so, whether they have the same length to compare the accuracy between the sets. To separate the results by ncRNA type we used custom scripts. To visualise the results we created venn diagrams. The venn diagrams were either created using the R package vennDiagram (Chen et al., 2016) or the web service venny (Oliveros, 2015).

Two predictions were classified as overlapping if at least 70 % of the predictions overlapped and they were predicted on the same strand. For our final list of ncRNAs we excluded all predicted ncRNAs that overlapped with an exon from the official protein-

coding gene set, those that overlapped with an ncRNA prediction from the OGS, and all predicted genes from DARIO that overlapped with another DARIO prediction.

2.3.3. RNAz

We also used another method for the prediction of non-coding RNAs using the program RNAz (version 2.1) (Gruber et al., 2010)). It uses a secondary structure approach to predict novel ncRNAs. To predict the likelihood of a secondary structure we used RNAfold from the ViennaPackage (Lorenz et al., 2011). This calculates for each candidate the minimal free energy (MFE). The MFE measures the binding energy between the paired nucleotides predicted in the secondary structure. If this value is too low or too high it is unlikely that this structure is stable and real.

RNAz used whole genome alignments as input. These alignments between the four species *A. mellifera*, *N. vitripennis*, *O. abietinus*, and *A. rosae* were created using Progressive Cactus v0.0 (Date of Download: 30.03.2016) (Paten et al., 2011). Progressive Cactus needs a guide tree for the alignments, which was extracted as a subtree from the 1KITE tree (Misof et al., 2014). Because the 1KITE tree did not include *A. rosae*, *Tenthredo koehleri* was used as a substitute, because both belong to the Tenthredinidae. Using the perl-script 'rnazWindow.pl' provided by the RNAz suite we extracted all parts from this WGA that contained sequences in all four species. This filtered alignment was provided to RNAz, which was run with the –both-strands option but otherwise default settings. The default strand setting of RNAz searches for ncRNA candidates only on the +-strand. The results from this RNAz run were clustered by position using the script 'rnazFilter.pl'. We also removed all hits with a p-value below 0.9 in this step to remove unreliable hits. The clustered and filtered RNAz results were then compared with the existing annotations (official gene set (OGS), three different DARIO annotations, mapped reads) using 'rnazAnnotate.pl'. Using 'rnazIndex.pl –html' the annotated results were transformed into html format for visual inspection.

2.3.4. FEELnc

lncRNAs in *A. rosae* and *O. abietinus* were predicted using FEELnc (v0.1.0 pre-release) (Wucher et al., 2017). We followed the workflow described on the FEELnc github-page (`https://github.com/tderrien/FEELnc`, May 2018). The first step was to mask areas with protein-coding gene candidates. For this we first indexed the genome using bowtie (v.2.3.2) (Langmead et al., 2009) and then mapped RNA-Seq reads of protein-coding genes onto the respective genome. The reads were either downloaded from the RefSeq database or the i5k server. The mapped reads were assembled using Cufflinks (v.2.2.1) (Trapnell et al., 2010). The reference genomes and annotation combined with these transcript models were analysed with the FEELnc pipeline.This pipeline uses RNA-Seq reads to identify regions of a genome that might contain lncRNAs.

First all transcript that overlapped in sense with an exon of the reference annotation were filtered out. In this FEELnc_filter step we kept monoexonic hits which differs from the default setting.

The next step was to calculate the coding potential for these candidates using FEELnc_codpod. In this step a kmer-approach is used to asses the candidates. We tested different kmer-combinations to obtain optimal results. For *Orussus* we finally selected 1-2-3-4-6-7-12 and for *Athalia* it was 1-2-3-5-6-7-12. The kmers were used to simulate lncRNAs for training the models, because no lncRNA are available for *A. rosae* or *O.abietinus*. We kept all lncRNA predictions with at least one exon.

The final step was the classification of the predicted lncRNAs according to their localisation and transcription direction using FEELnc_classifier. The file with the classified lncRNAs included several lncRNA-gene interactions for each lncRNAs. For further analysis we excluded all those results that were not classified as 'isBest'. The interactions between lncRNA and gene were classified into two different types, each with several subtypes and locations (figure 2.2). The types were 'genic' and 'intergenic' depending on whether the lncRNA was found overlapping a known gene or not. lncRNAs of the 'intergenic' types are also called lincRNAs. The subtypes for 'intergenic' were 'divergent', 'convergent', 'same strand', and 'unknown strand'. 'Divergent' means the lncRNA is transcribed on the other strand as the gene is present on with a head to

head orientation, 'convergent' means the lncRNA is found on the other strand with a tail to tail orientation, and 'same strand' means both lncRNA and gene are found on the same strand but are not overlapping. 'Unknown strand' means it was not possible to categorise the interactions. The location for all these subtypes are either 'upstream' or 'downstream'. For 'genic' interactions the subtypes 'overlapping', 'containing', and 'nested' exist. 'Overlapping' means the lncRNA partially overlaps the gene, 'containing' means the gene is completely found inside the lncRNA prediction, and 'nested' means the lncRNA is completely found inside the gene. For all these cases the location 'exonic' and 'intronic' exist.

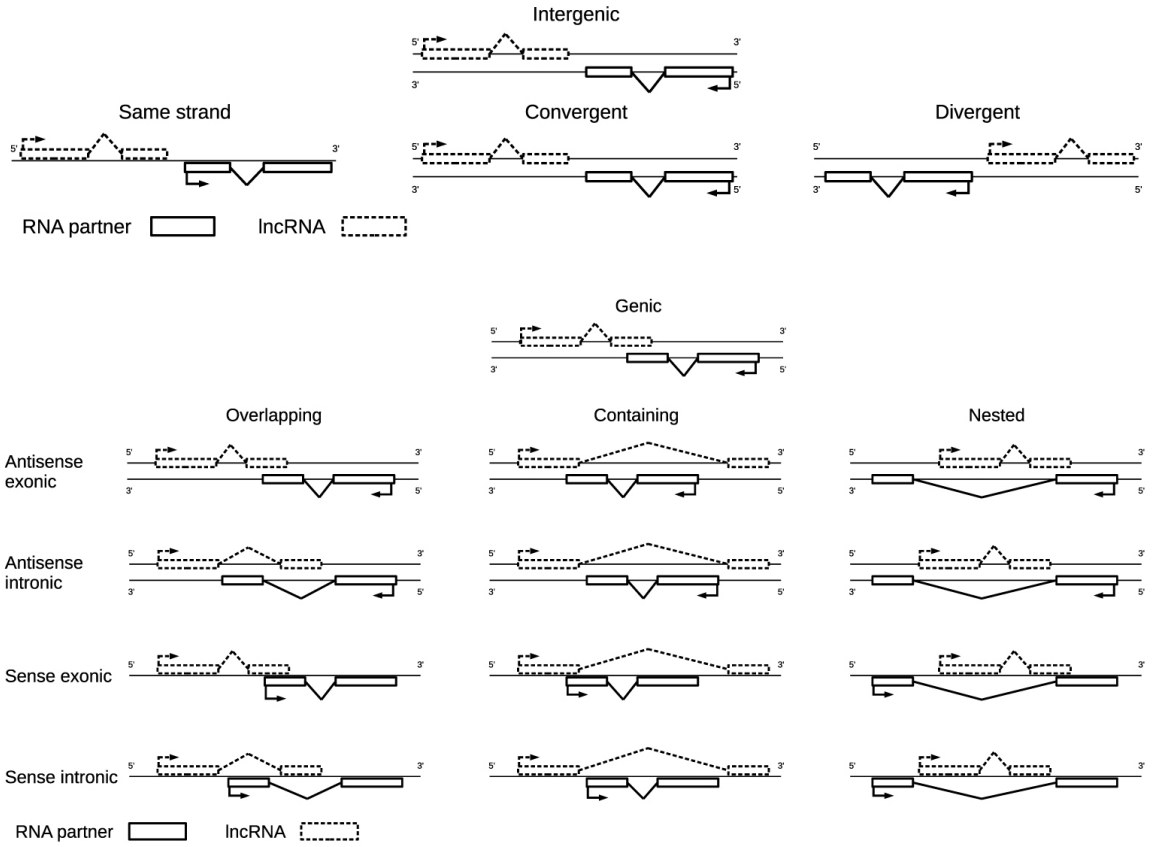


Figure 2.2.: FEELncclassifier description. Sub classification of intergenic and genic lncRNA/transcripts interactions by the FEELncclassifier module. Taken from Wucher et al. (2017).

Additionally we used the FEELnc_classifier script to classify the interaction between lncRNA and protein-coding genes of the lncRNAs present in the official gene set of *A.*

mellifera and *N. vitripennis*. The filtering was done the same way we used for *A. rosae* and *O. abietinus*.

3. Results non-coding RNAs

3.1. Database curation

The Rfam 12 contains 2,450 different ncRNA families. After filtering these families for sequences that are only present in eukaryotes, 1,661 families remain. However, this number still contains false positives. After we removed the 123 families we classified as false positives, 1,538 families remain, of which 1.107 are found only in Metazoa. This list still contains tRNAs and miRNAs, which were later removed during the analysis.
The miRBase v21 contains 1,983 different miRNA families. After manual curation we excluded 564 as non-metazoan miRNAs and used the remaining 1,419 miRNA families in our further analysis.

3.2. Results of the homology prediction

Through the homology prediction using the manually curated databases miRBase and Rfam we identified miRNAs and the different ncRNA types listed in Rfam. Because we either used a specialised database or a specialised prediction program, we excluded miRNAs and tRNAs that are listed in Rfam. The ncRNAs identified using the Rfam belong to the classes of snoRNAs (C/D, H/ACA, small Cajal body-specific RNA (scaRNA)), rRNA, mitochondrial RNA processings (MRP RNAs), snRNA, as well as some types of regulatory RNAs, such as 3'-UTR and RNA editing signal.

3.2.1. Predicted ncRNAs in *Athalia rosae*

microRNAs

Without any curation the Infernal search predicted 248 miRNA families in *A. rosae*. After removing all hits with an e-value ≥ 0.01, removing all families that are listed in our false positive list, and manually curating the MSA of each miRNA family we identified 80 miRNAs belonging to 62 different families (table 3.1, figure 3.1).

Some miRNA families are present with multiple copies in a genome. In these multiple copies the seed region is highly conserved. Of the miRNA families present with multiple copies in *A. rosae*, two were present with two copies (mir-67, mir-263), one with three copies (mir-25), one with four copies (mir-9), and two with five copies (mir-2, mir-279). All five copies of mir-2 occurred in a cluster on a single scaffold all oriented in the same strand direction (figure 3.2). The distance between the different copies of mir-2 in this cluster ranged between 130 bp and 269 bp. In front of this cluster with only 364 bp distance to mir-2c we found a miRNA belonging to the mir-71 family. In all other multi-copy cases only some of them were clustered on the same scaffol,d but never all copies.

The miRNA family mir-1923 was found only in *A. rosae* in our analysis. It is however also known from *Acyrthosiphon pisum* and *Bombyx mori*.

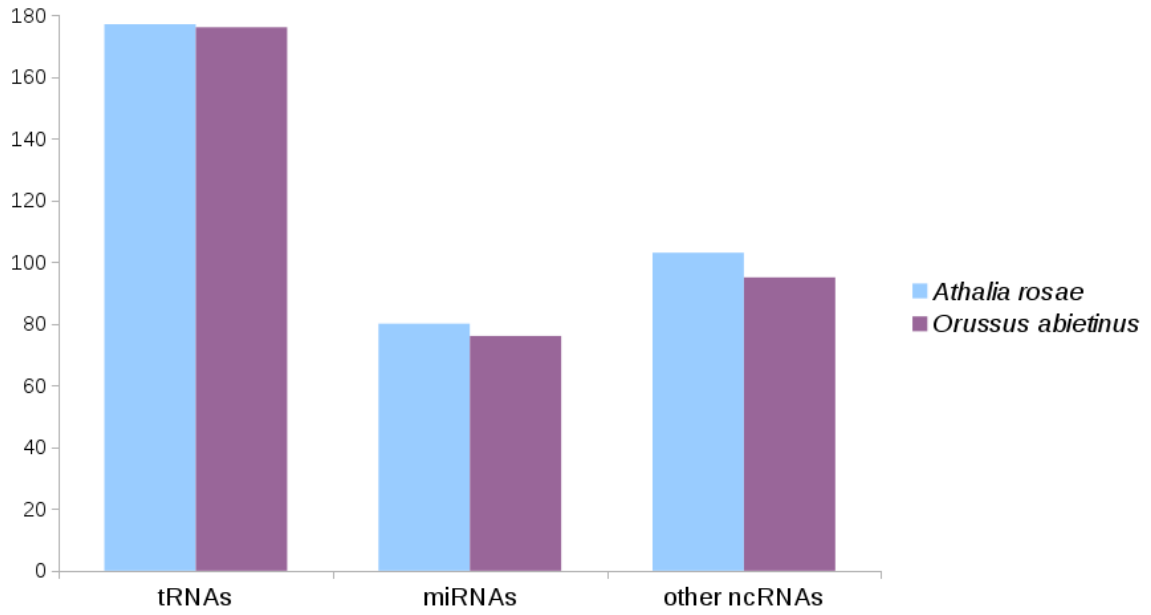


Figure 3.1.: Absolute numbers of the different types of ncRNAs found through homology analysis (miRNAs, other ncRNAs) or through *de novo* prediction (tRNAs) in *Athalia rosae* and *Orussus abietinus*. The number of tRNAs results from the analysis with tRNAscan-SE, the numbers of miRNAs through a homology search using the miRBase, and for all other ncRNAs the Rfam was used for the homology search.

Other ncRNA families

We predicted 103 other ncRNAs (excluding tRNAs) belonging to 35 families in *A. rosae* (figure 3.1). This includes 14 snoRNA families, two lncRNA families (Sphinx 1 and Sphinx 2), two RNase families, four rRNA families, one signal recognition particle RNA (SRP RNA), 10 snRNA families, and two *cis*-regulatory elements (potassium channel RNA editing signal (K_chan_res), histone 3' UTR stem-loop (Histone3)) (table 3.2). Of the snoRNA families five are present with multiple copies in the genome. SNORD31 is present with four copies, which all cluster on the same scaffold oriented in the same strand direction. Additionally, all five copies of snosnR60_Z15 are oriented in a cluster on the same scaffold.

The ten snRNA families were comprised of 29 snRNAs. With only one copy present we found five snRNAs (Arthropod_7SK, U4atac, U6atac, U11, U12). We found one famliy

present with two copies (U4), two with four copies (U5, U6), one with five copies (U2), and one with nine copies (U1).

We identified one complete rRNA gene cluster consisting of 28S, 18S, and 5.8S rRNA. The 28S rRNA was split in two parts through the insertion of an R2 element. The R2 element is a non-LTR retrotransposon which can be found as an insertion in the 28S rRNA throughout arthropods (Burke et al., 1999). Including the 18S that is part of the rRNA gene cluster we found four copies of this rRNA and four 5.8S rRNA copies. The 5.8S copies are all found on different scaffolds, whereas one copy of the 18S can be found next to the 18S that is part of the rRNA gene cluster. Of the 5S rRNA we identified 11 copies. The split 28S rRNA was the only copy found of this rRNA.

3.2.2. Predicted ncRNAs in *Orussus abietinus*

microRNAs

In *O. abietinus* we predicted miRNAs belonging to 380 families. After removing all families that are listed in our false positive list, manual curation of the sequence alignments of each miRNA family and removal of all hits with an e-value ≥ 0.01 76 miRNAs belonging to 60 different miRNA families remained (table 3.1, figure 3.1). Six families are present in multiple copies, with either two copies (mir-263), three copies (mir-25), four copies (mir-10, mir-279, mir-9) or five copies (mir-2). The five copies of mir-2 can be found in one cluster all in the same orientation on one scaffold (figure 3.2). Even if the name suggests otherwise, mir-13a does belong to the mir-2 family, as sometimes miRNA families are combined if new evidence is found without changing the names of the members. The distance between the miRNAs in this cluster varies between 82 and 326 bp. In front of this cluster mir-71 can be found in 276 bp distance. This miRNA does not belong to the mir-2 family, however the position in front of the mir-2 cluster is conserved between different species.

All of the miRNAs identified in *Orussus* were present in at least one other species in our analysis.

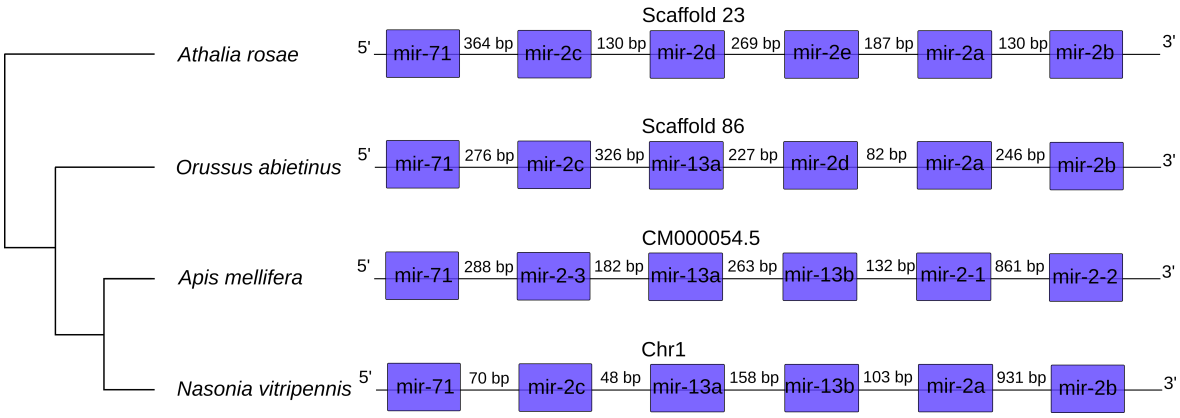


Figure 3.2.: Graphical overview of the mir-2 cluster in *Athalia rosae*, *Orussus abietinus*, *Apis mellifera*, and *Nasonia vitripennis*. In front of the cluster mir-71 is shown which is not part of the mir-2 family but has a conserved position next to the cluster. The data for *A. mellifera* and *N. vitripennis* were taken from the miRBase. Not shown is the mir-2b of *A. mellifera* as it is orientated differently and lies completely inside of mir-2-1.

Table 3.1.: List of all miRNAs present in the species *Tribolium castaneum* (Tcas), *Apis mellifera* (Amel), *Nasonia vitripennis* (Nvit), *N. longicornis* (Nlon), and *N. girault* (Ngir) as listed in miRBase, after manual curation, sorted by families and copy number per family. *Athalia rosae* (Aros) and *Orussus abietinus* (Oabi) are the results of our homology analysis.

miRNA	Tcas	Aros	Oabi	Amel	Nvit	Nlon	Ngir
bantam	1	1	1	1	1		
let-7	1	1	1	1	1		1
mir-1	1	1	1	2	1		1
mir-2	5	5	5	6	5	2	2
mir-7	1	1	1	1	1	1	
mir-8	1	1	1	1	1	1	
mir-9	4	4	4	4	1	1	1
mir-10	4	4	4	4	4	3	2
mir-11	1	1	1	1	2		
mir-12	1	1	1	1	1	1	
mir-14	1	1	1	1	1		1
mir-25	3	3	3	4	1	1	1
mir-29	1		1	1	1		1
mir-31	1	1	1	1	1	1	1
mir-33	1	1	1	1	1		
mir-34	1	1	1	1	1	1	1
mir-46	1	1	1	1	1	1	1
mir-67	1	2	1	1	1		

Continued on next page

Table 3.1.: Continued from previous page.

miRNA	Tcas	Aros	Oabi	Amel	Nvit	Nlon	Ngir
mir-71	1	1	1	1	1		
mir-87	2		1	2			
mir-124	1	1	1	1	1		
mir-133	1	1	1	1	1	1	1
mir-137	1	1	1	1	1	1	1
mir-184	1	1	1	1	1	1	1
mir-190	1	1	1	1	1		
mir-210	1	1	1	1	1		1
mir-216	2	1	1	1	1		1
mir-219	1	1	1	1	1	1	1
mir-252	1	1	1	1	1	1	1
mir-263	2	2	2	2	1	1	1
mir-275	1	1	1	1	1	1	1
mir-276	1	1	1	1	1	1	1
mir-277	1	1	1	1	1	1	1
mir-278		1	1	1			
mir-279	3	5	4	3	1		
mir-282	1	1	1	1	1	1	
mir-305	1	1	1	1	1	1	1
mir-315	1	1	1	1	1	1	
mir-316	1	1	1	1			
mir-317	1	1	1	1	1		1
mir-375	1	1	1	1	1		1
mir-750		1			1	1	1
mir-927	1	1	1	1	1	1	1
mir-928		1	1	1	1		
mir-929	1	1	1	1	1	1	1
mir-932	1	1	1	1	1	1	1
mir-965	1	1	1	1			
mir-970	1						
mir-971	2	1	1	1	1		
mir-980	1	1	1	1	1		
mir-981	1	1	1	1	1		
mir-1000	1	1	1	1			
mir-1175	1	1	1	1	1		
mir-1923		1					
mir-2765	1	1	1	1	1		
mir-2788	1	1	1	1	1		
mir-2796	1	1	1	1	1		
mir-2944	3	1	1	1	1		
mir-3477		1	1	1	1		
mir-3478		1	1	1	1		

Continued on next page

Table 3.1.: Continued from previous page.

miRNA	Tcas	Aros	Oabi	Amel	Nvit	Nlon	Ngir
mir-3718				2			
mir-3747				2			
mir-3804	2						
mir-3811	8						
mir-3817	2						
mir-3836	2						
mir-3851	7						
mir-6012	1	1	1	1	1		
mir-6497	1	1					1
mir-iab-4	1	1	1	1	1		1
mir-iab-8	1	1	1	2	1		1

Other ncRNA families

Using the Rfam as reference for the homology prediction of other ncRNAs we identified 95 ncRNAs (excluding tRNAs and miRNAs) belonging to 30 families (figure 3.1). Of the predicted ncRNAs 22 were classified as snRNAs, 11 as snoRNAs, 40 rRNAs, two lncRNAs (Sphinx 1 and Sphinx 2), two RNase families, one SRP RNA, and 17 *cis*-regulatory elements (10 histone 3' UTR stem-loops (Histone3), four potassium channel RNA editing signals (K_chan_res), three R2 RNA elements (R2_retro_el)) (table 3.2).

The 11 identified snoRNAs belong to 10 families and only snosnR60_Z15 was found with two copies. These two copies were found on the same scaffold with only 278 bp between them.

We found 22 snRNAs belonging to ten different families. Of those six were found with only one copy (Arthropod_7SK, U4, U4atac, U6atac, U11, U12), one with two copies (U6), two with four copies (U2, U5), and one with six copies (U1). In no multi copy case all copies were found on the same scaffold.

Of the four expected rRNA families we were only able to identify the 18S and 5S rRNA. We identified two copies of the 18S rRNA and 38 5S rRNA copies. No complete rRNA gene cluster was found due to the lack of 28S and 5.8S rRNAs. No R2 element was found, as the insertion site is missing, however in the Rfam an R2 RNA element is

listed, which was found with three copies.

Table 3.2.: List of all regulatory elements and ncRNAs, excluding miRNAs and tRNAs, present in the species *Tribolium castaneum* (Tcas), *Apis mellifera* (Amel), *Nasonia vitripennis* (Nvit), *N. longicornis* (Nlon), and *N. giraulti* (Ngir) and listed in Rfam, after manual curation. For *Athalia rosae* (Aros) and *Orussus abietinus* (Oabi) the results are from our homology analysis.

ncRNA	Tcas	Aros	Oabi	Amel	Nvit	Nlon	Ngir
RF00001 5S_rRNA	225	11	38	62	31	19	20
RF00002 5_8S_rRNA		4		1	1	1	3
RF00003 U1	5	9	6	7	8	8	7
RF00004 U2	5	5	4	7	5	3	4
RF00007 U12	1	1	1	2	2	2	2
RF00008 Hammerhead_3	1			3			
RF00009 RNaseP_nuc	2	2	1	2	1	1	1
RF00012 U3	2			3	4	5	4
RF00015 U4	2	2	1	2	3	2	2
RF00017 Metazoa_SRP	2	2	1	1	2	2	2
RF00020 U5	6	4	4	3	5	2	3
RF00026 U6	3	4	2	3	5	3	4
RF00030 RNase_MRP	1	1	1	1	2		
RF00032 Histone3	18	16	10	22	173	125	111
RF00049 SNORD36	1	2		1	1	1	1
RF00059 TPP					1		
RF00089 SNORD31	3	4	1	1	1	1	1
RF00093 SNORD18	1	2		1	1	1	1
RF00133 SNORD33		1					
RF00190 SNORA16		1		1	2	2	2
RF00191 SNORA57	1	1	1				
RF00274 SNORD57	1	2		1	2		
RF00277 SNORD49		1					
RF00286 SCARNA8	1	1	1		1	1	1
RF00291 snoR639				1			
RF00309 snosnR60_Z15	3	5	2	2	4	4	4
RF00334 SNORA3	1		1	1	1		
RF00377 snoU6-53				1	1	1	1
RF00476 snosnR61	1			1			
RF00485 K_chan_RES	5	5	4	5	4	4	4
RF00524 R2_retro_el			3		1	1	1
RF00533 snoMe18S-Gm1358					1	1	1
RF00535 snoMe28S-Am982			1	1	1		
RF00542 snopsi28S-1192			1		1	1	1
RF00543 snopsi18S-1377	5	1					
RF00548 U11	1	1	1	2	2	1	1

Continued on next page

Table 3.2.: Continued from previous page.

Rfam ID and name	Tcas	Aros	Oabi	Amel	Nvit	Nlon	Ngir
RF00563 SNORA53	1						
RF00575 SNORD70					1	1	1
RF00600 SNORA79	1	1	1	1	1		
RF00618 U4atac	1	1	1	1	1	1	1
RF00619 U6atac	2	1	1	1	1	2	3
RF01052 Arthropod_7SK	1	1	1	1	2	2	2
RF01159 snoU18				2	1	1	1
RF01174 snoU43	1	1	1	1	1	1	1
RF01848 ACEA_U3	1	2	1	1	1		
RF01960 SSU_rRNA_eukarya	21	4	2	1	1	3	3
RF01988 SECIS_2					1		
RF02046 Sphinx_1	1	1	1		1	17	15
RF02047 Sphinx_2	1	1	1		1	17	15
RF02253 IRE_II				1			
RF02543 LSU_rRNA_eukarya	28			6	3	17	15

3.3. Results of the *de novo* prediction

3.3.1. DARIO datasets

In addition to *de novo* identifying further ncRNAs using DARIO we also compared the prediction results based on differently prepared datasets. The difference between these datasets was how the paired-end reads were treated after adaptor trimming (see 2.3.2). In the first set called 'merged' the two mates of a pair were combined. The second set called 'paired' included those matepairs that still had a partner after trimming without combining the two. The third set called 'paired_unpaired' included all reads that were left after trimming, leading to a library containing both complete pairs as well as unpaired reads (table 3.3).

Table 3.3.: Read counts of the different datasets that were prepared for DARIO.

	merged	paired	paired_unpaired
A. rosae	18,385,448	20,320,551	20,320,551
O. abietinus	14,124,071	15,200,893	15,200,893

3.3.2. *de novo* Prediction of tRNAs in *Athalia rosae*

Our *de novo* tRNA prediction was done with two different programs, tRNAscan-SE and DARIO. DARIO used the tRNAs predicted with tRNAscan-SE to identify further tRNAs.

tRNAscan-SE

First we used tRNAscan-SE which resulted in 184 tRNA candidates. After removing pseudo-tRNAs and those with a score below 55, 177 tRNAs remained (table 3.4). Of these tRNAs nine contained one intron. The only tRNA types identified as containing introns were tRNA-Tyr, tRNA-Ile, and tRNA-Leu. In tRNA-Ile and tRNA-Leu we predicted tRNAs containing introns as well as without (table 3.5). All predicted tRNA-Tyr genes contained an intron.

Some tRNAs of the same type can be found in clusters with short distances between the single genes. The tRNA-Ala was found within one cluster containing four genes that had 81-90 bp between them. Other clusters were found with the tRNA-Asp (five genes, 77-763 bp distance), tRNA-Val (three genes, 93-307 bp distance), and tRNA-Gly (three genes, 116-119 bp distance). The genes in all these clusters were found in the same strand orientation (figure 3.3).

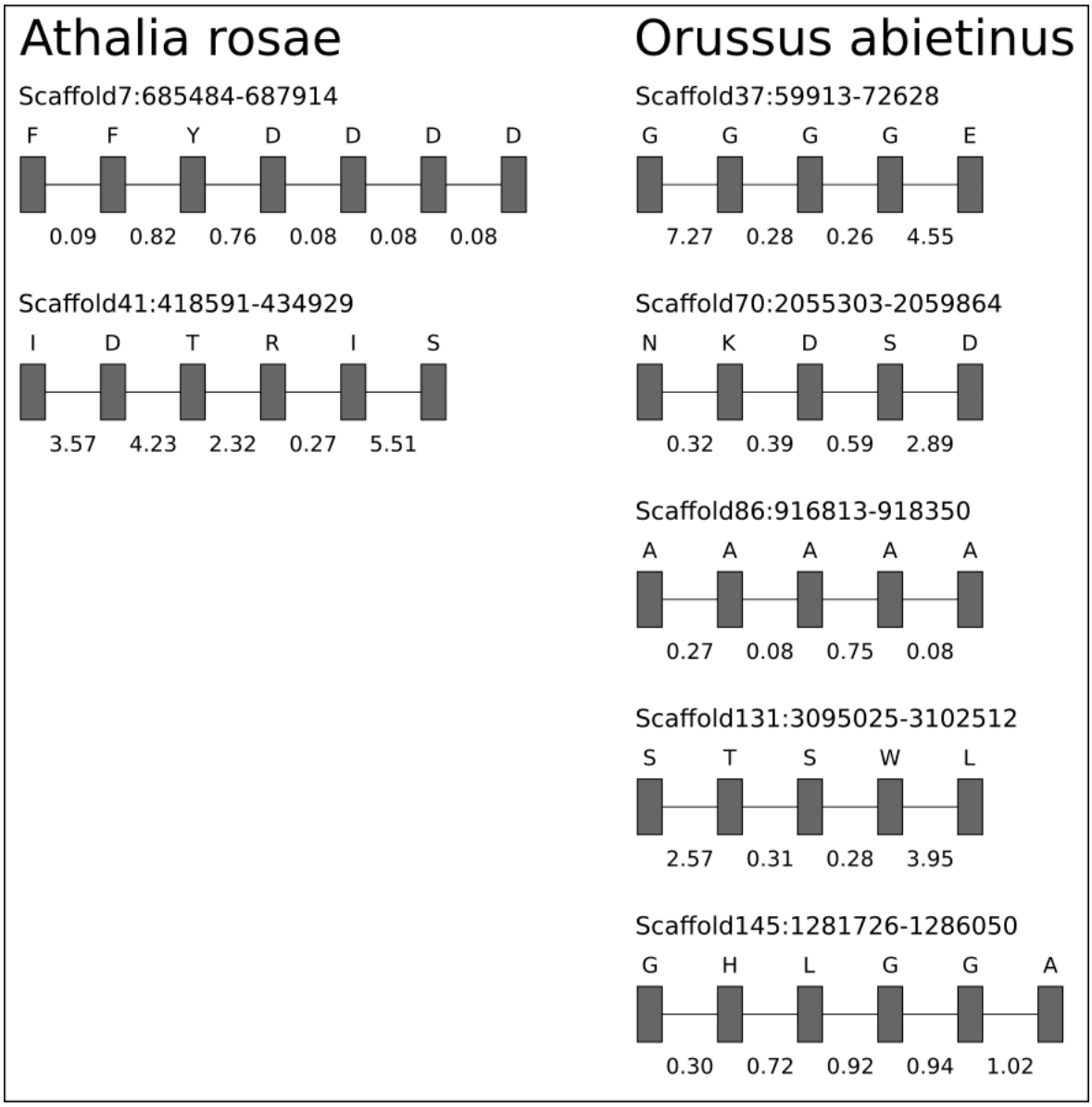
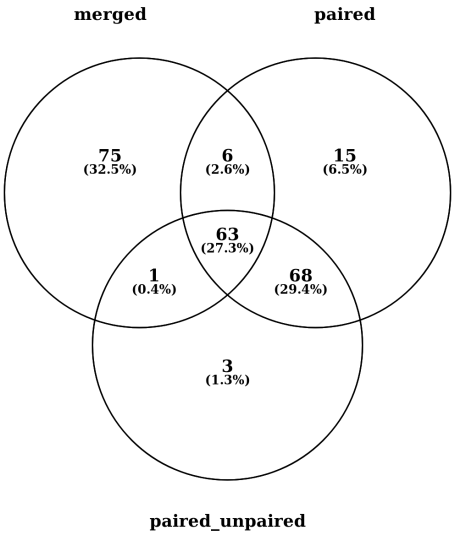


Figure 3.3.: Visualisation of tRNA clusters found in *A. rosae* and *O. abietinus*. Distance between tRNA genes is given in kb. A = tRNA-Ala, D = tRNA-Asp, E = tRNA-Glu, F = tRNA-Phe, G = tRNA-Gly, H = tRNA-His, I = tRNA-Ile, L = tRNA-Leu, N = tRNA-Asn, R = tRNA-Arg, S = tRNA-Ser, T = tRNA-Thr, W = tRNA-Trp, Y = tRNA-Tyr

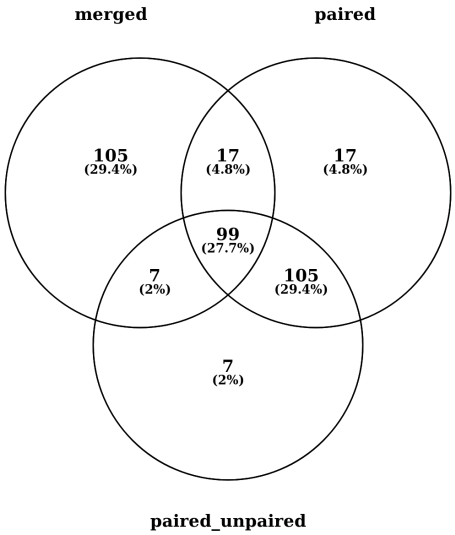
In some cases tRNA genes of different types were also found in close proximity to each other. Two more tRNAs were predicted next to the above mentioned tRNA-Asp cluster (821 bp distance to the cluster and 89 bp between them).

DARIO

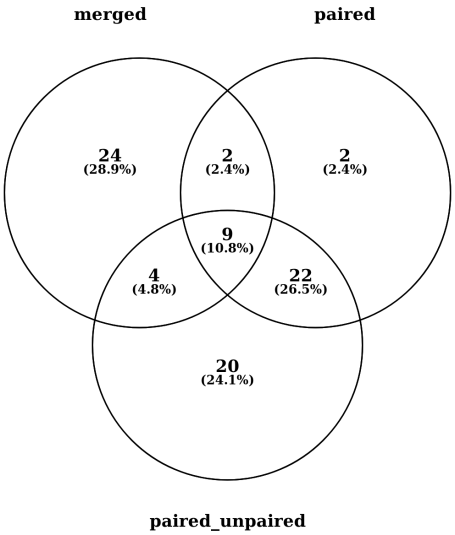
Additional tRNAs were predicted through the DARIO pipeline, which were not classified into the different tRNA types as this is not part of the DARIO pipeline. The numbers differ for our three different datasets (table 3.6). Our merged dataset had 245 tRNAs predicted, the paired set 276, and the paired_unpaired 254. After removing the tRNAs overlapping with exons from protein-coding genes, other ncRNA predictions, or other ncRNAs predicted by DARIO, 145 tRNAs remained for the merged set, 152 for the paired, and 135 for the paired_unpaired. Only 63 of the predicted tRNAs are present in all three datasets (figure 3.4a). The merged set shows the highest divergence to the other two sets. It has only seven predictions that overlap with one of the other datasets whereas 68 are shared between the paired and the paired_unpaired sets.



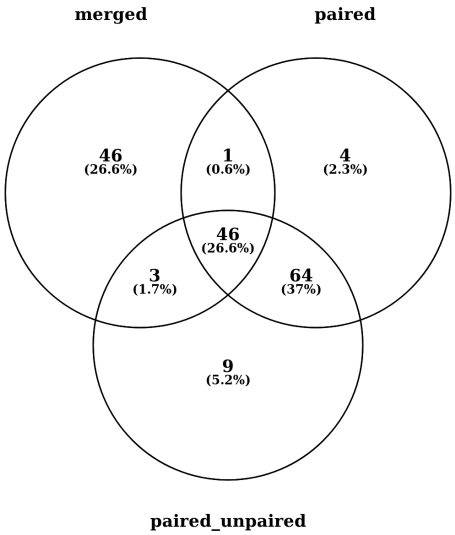
(a) tRNAs



(b) miRNAs



(c) H/ACA snoRNAs



(d) C/D snoRNAs

Figure 3.4.: Visualisation of the overlapping ncRNAs predicted by the DARIO pipeline in *Athalia rosae*, with the three different datasets sorted by how the read sets were constructed. ncRNAs were classified as being the same one if at least 70 % of the sequence overlapped.

Table 3.4.: Results of the *de novo* prediction of tRNAs in the genomes of *Athalia rosae* (Aros) and *Orussus abietinus* (Oabi) using tRNAscan-SE. Our results are compared to the predicted tRNA numbers in *Nasonia vitripennis* (Nvit) and *Apis mellifera* (Amel) from Behura et al. (2010). Modified after Behura et al. (2010).

tRNA gene	Aros	Oabi	Amel	Nvit
Ala	15	12	14	16
Arg	14	11	13	10
Asn	5	4	8	8
Asp	7	7	9	10
Cys	3	2	3	5
Gln	8	12	18	9
Glu	10	10	11	14
Gly	13	11	14	17
His	5	4	7	8
Ile	8	8	8	12
Leu	14	12	11	18
Lys	10	9	13	18
Met	9	8	7	7
Phe	4	3	6	7
Pro	12	8	12	14
Ser	12	10	15	12
Thr	11	9	10	10
Trp	9	3	4	4
Tyr	5	7	5	9
Val	9	8	11	13
Sum	177	158	199	221

Table 3.5.: List of tRNA families containing introns as they were identified in *Orussus abietinus* (Oabi) and *Athalia rosae* (Aros) compared to *Nasonia vitripennis* (Nvit) and *Apis mellifera* (Amel) (Behura et al., 2010). + indicates an intron present in this tRNA, - indicates the lack of an intron. Modified after Behura et al. (2010).

tRNA gene	Aros	Oabi	Amel	Nvit
Tyr (I+)	5	7	5	9
Tyr (I-)	0	0	0	0
Ile (I+)	2	2	2	3
Ile (I-)	6	6	6	9
Leu (I+)	2	2	3	3
Leu (I-)	12	10	8	15

Table 3.6.: Results of the DARIO pipeline for the *de novo* prediction of ncRNAs in *Athalia rosae* (Aros) and *Orussus abietinus* (Oabi). The four types predicted were tRNAs, miRNAs, H/ACA snoRNAs, and C/D snoRNAs. Included are the results for our three different read mapping datasets. Two numbers are shown for each type. The first is the number of ncRNAs DARIO predicted, the second one shows the final set after sorting out those predictions that overlapped known exons or with other DARIO predictions.

ncRNA type	Aros	Oabi	Aros	Oabi	Aros	Oabi
	merged		paired only		paired and unpaired	
miRNA predicted	400	1291	440	1380	401	1441
miRNA final	228	974	238	1061	218	1105
tRNA predicted	245	494	276	471	254	468
tRNA final	145	341	152	326	135	324
H/ACA snoRNA predicted	65	84	63	314	110	272
H/ACA snoRNA final	39	55	35	190	55	160
C/D snoRNA predicted	162	24	177	48	191	32
C/D snoRNA final	96	20	115	28	122	20

3.3.3. *de novo* Prediction of tRNAs in *Orussus abietinus*

tRNAscan-SE

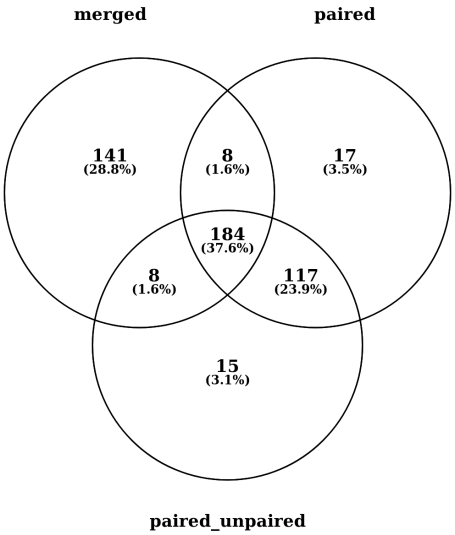
In *O. abietinus* tRNAscan-SE predicted 176 tRNAs. After removing all pseudo-tRNAs and those hits with an e-value below 55, 158 tRNAs remained (table 3.4). We identified 11 tRNAs containing an intron. All predicted tRNA-Tyr contain an intron, whereas in the cases of tRNA-Ile and tRNA-Leu we found some with introns as well as without (table 3.5).

Some of the same type of tRNA gene can be found in clusters. The cluster with the least distance between tRNA genes belonged to the tRNA-Gln type (four genes, 69 bp distance). Two more copies of the same tRNA were found next to this cluster (214 bp distance), however they were orientated in the other direction and showed a distance of 65,522 bp to the cluster. Next to these two tRNA-Gln genes we found two tRNA-Tyr genes with only 322 bp distance to the cluster and 102 bp between each other. Another cluster was composed of five tRNA-Ala genes (82-747 bp distance), another of three tRNA-Val genes (990-1235 bp distance), one of four tRNA-Ile genes (120-590 bp

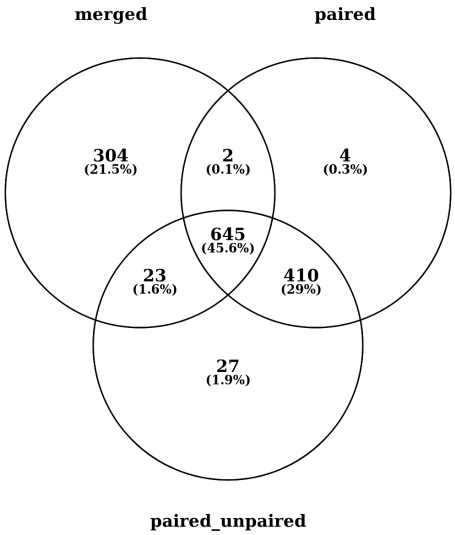
distance), and one of four tRNA-Gly genes (262-7273 bp distance).
We were also able to identify some clustered tRNA genes that belonged to different types (figure 3.3). In one case a tRNA-Glu gene and a tRNA-Leu gene were only separated by 168 bp, in another four different tRNA genes (Asn, Ly, Asp, Ser) were found with 323-591 bp between them. They were oriented into different strand directions. One cluster made up of five tRNA genes (two Ser, Thr, Trp, Leu; 275-3948 bp distance) was found with all tRNA genes orientated in the same direction. Another cluster with six tRNA genes contained three tRNA-Gly, of which two were neighbouring, together with one tRNA-His, one tRNA-Lys, and one tRNA-Ala (303-1,1018 bp distance) and had also different orientation of the genes.

DARIO

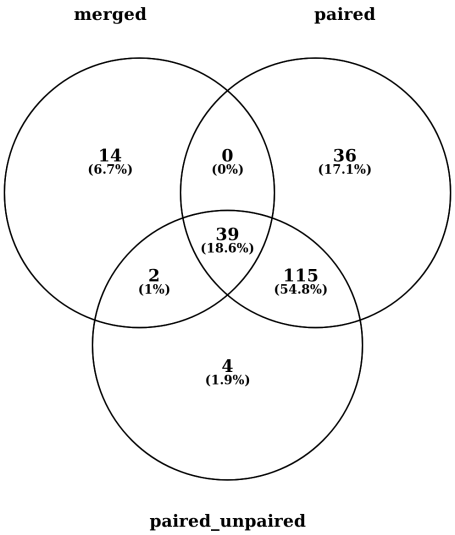
The DARIO pipeline predicted tRNAs that were not identified by tRNAscan-SE. The most tRNAs were predicted with the set merged (341), followed by paired (326), and paired_unpaired (324) (figure 3.5a). Looking at the predictions shared only between two sets, the merged set shared eight predictions with each other set. The paired and paired_unpaired set shared 117 predictions with only each other. Between all three sets 184 tRNA predictions were shared.



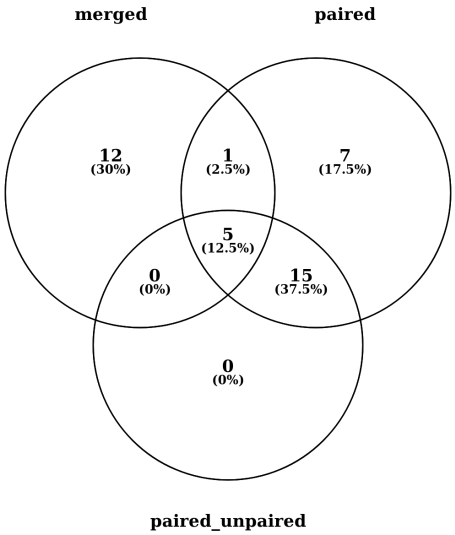
(a) tRNAs



(b) miRNAs



(c) H/ACA snoRNAs



(d) C/D snoRNAs

Figure 3.5.: Visualisation of the overlapping ncRNAs predicted by the DARIO pipeline in *Orussus abietinus* with the three different datasets sorted by how the read sets were constructed. ncRNAs were classified as being the same one if at least 70 % of the sequence overlapped.

3.3.4. *de novo* Prediction of miRNAs in *Athalia rosae*

The DARIO pipeline predicted 400 miRNAs in *A. rosae* with the merged dataset, 440 with the paired, and 401 with the paired_unpaired. Between 172 and 202 miRNAs were removed because they either overlapped with exons from the official gene set or overlapped with another Dario prediction. Our final list of *de novo* predicted miRNAs contains 228 genes for the merged set, 238 for the paired, and 218 for the paired_unpaired (table 3.6). Shared between all three sets 99 miRNAs were predicted (figure 3.4b). The most overlap between only two sets was found between paired and paired_unpaired with 105 miRNA, whereas the merged set only shared further 17 (paired) or 7 (paired_unpaired). Only 7 miRNAs were only predicted in the paired_unpaired datasets which makes this the set with the lowest number of unique predictions. The merged set had the most unique predictions with 105 unique miRNAs.

3.3.5. *de novo* Prediction of miRNAs in *Orussus abietinus*

The DARIO pipeline predicted between 1,291 (merged) and 1,441 miRNAs (paired_-unpaired) in *O. abietinus*. After sorting out those that overlap with an exon of a protein-coding gene, any other already predicted ncRNA, or another DARIO prediction between 974 (merged) and 1,105 miRNAs (paired_unpaired) remained (table 3.6). Most miRNAs were predicted with the paired_unpaired set (1,105), followed by the paired set (1,061), and the merged set (974). Of these predictions 645 were found in all three sets (figure 3.5b). The paired and the paired_unpaired sets shared additional 410 predictions, whereas the merged set had 304 that were only predicted in this set (figure 3.5b).

3.3.6. *de novo* Prediction of snoRNAs in *Athalia rosae*

The Dario pipeline predicts two types of snoRNAs, H/ACA snoRNAs and C/D snoRNAs. Dario predicted between 63 (paired) and 110 H/ACA snoRNAs (paired-_unpaired) and

between 162 (merged) and 191 C/D snoRNAs (paired_unpaired) (table 3.6). After sorting out between 35 (paired) and 55 H/ACA snoRNAs (paired_unpaired) remained, as well as between 96 (merged)and 122 C/D snoRNAs (paired_-unpaired). The most H/ACA snoRNAs were predicted with the paired_unpaired dataset (55), followed by the merged set (39), and the least were predicted in the paired set (35). Regarding the C/D snoRNAs, the highest number was predicted with the paired_unpaired set (122), followed by the paired set (115), and the lowest number with the merged set (96). Of the H/ACA snoRNAs, 9 were predicted as present in all three datasets, whereas 46 C/D snoRNAs were present in all three sets (figure 3.4c, 3.4d). In both cases the most overlap between only two sets was found between paired and paired_unpaired (22 H/ACA snoRNAs (figure 3.4c), 64 C/D snoRNAs (figure 3.4d)).

3.3.7. *de novo* Prediction of snoRNAs in *Orussus abietinus*

Using the DARIO pipeline we predicted snoRNAs of the two types H/ACA and C/D in *O. abietinus*. Before curation of the candidates DARIO predicted between 84 (merged) and 314 (paired) H/ACA and 24 (merged) and 48 (paired) C/D snoRNAs. After filtering between 55 (merged) and 194 (paired) H/ACA snoRNAs remained (figure 3.5c), and between 18 (merged) and 27 (paired) C/D snoRNAs (figure 3.5d). The least H/ACA snoRNAs were predicted using the merged set (55), followed by the paired_unpaired set (160) and the paired set (194). Between the paired and the paired_unpaired sets we predicted 115 H/ACA snoRNAs only found in these two sets, two between merged and paired_unpaired and none between merged and paired (figure 3.5c). Only four H/ACA snoRNAs were found only with the paired_unpaired sets, 14 only in the merged set, and 36 in the paired set. Regarding the C/D snoRNAs we found none that were only found with the paired_unpaired set, seven only with the paired set and 12 using the merged set (figure 3.5d). No C/D snoRNAs were shared only between the merged and the paired_unpaired sets, one was shared between the merged and the paired sets, and 15 between the paired and the paired_unpaired sets.

We predicted 39 H/ACA and 5 C/D snoRNAs as present in all the sets. Overall we were able to predict more H/ACA snoRNAs than C/D.

3.3.8. RNAz

Due to a problem with the script 'rnazAnnotate.pl' we discarded the RNAz analysis. The script takes the RNAz results that were generated so far and compares them with other annotations, such as the OGS or our DARIO results. The results were grouped into loci which contained at least one location. The 'rnazAnnotate.pl' script misgrouped the results in some cases by combining loci regardless on their position in the genome, i.e. two different scaffolds being combined.

3.3.9. *de novo* Prediction of lncRNAs in *Athalia rosae*

Additionally to the two lncRNAs (Sphinx 1, Sphinx 2) identified through homology analysis a *de novo* analysis of the genome produced 3,613 more lncRNA candidates (table 3.7). All candidates contained between one and 11 exons. The majority of the predictions were lncRNAs with only one exon (3,014 = ∼83.4 %), and the less lncRNA are predicted the more exons they have.
FEELnc also predicts a protein-coding gene for a potential interaction with an lncRNA. These interactions are categorised into different types and subtypes and are differently ranked (see figure 2.2). For *A. rosae* 8,804 lncRNA-gene interactions were predicted. Of these interactions 3,573 were classified as the best ones following the interaction criteria. For 40 lncRNA (∼1.11 %) it was not possible to predict a gene interaction. The types are genic and intergenic. For genic, the subtypes are overlapping, containing, and nested with the additional locations of exonic and intronic. Intergenic subtypes are divergent, convergent, and same strand with the locations upstream and downstream. For intergenic interactions the best gene partner is the one closest to the lncRNA and for genic ones exonic gene partners (see figure 2.2). Due to our settings it is possible to have more than one interaction partner for an lncRNA. The majority of the interactions belonged to the intergenic type. However, most intergenic type interaction (1,832)

could not be added to one of the subtypes and are therefore classified as 'unknown strand' (table 3.8). The second most interactions (470) are classified as 'genic'-'nested'-'intronic'. No interactions were predicted as 'divergent'-'downstream' or 'convergent'-'upstream'.

Table 3.7.: Number of lncRNAs predicted in four Hymenoptera species. The numbers for *A. mellifera* and *N. vitripennis* were taken from the official gene sets and the ones for *A. rosae* and *O. abietinus* were predicted using the FEELnc pipeline.

	A. rosae	*O. abietinus*	*A. mellifera*	*N. vitripennis*
Number of lncRNAs	3,613	5,121	4,749	605

3.3.10. *de novo* Prediction of lncRNAs in *Orussus abietinus*

In *O. abietinus* we also identified two lncRNAs through homology prediction (Sphinx 1, Sphinx 2) and identified 5,121 lncRNA candidates through *de novo* prediction. The number of exons per lncRNA varied between one and nine. We predicted 4,338 (=∼84.7 %) lncRNAs with only one exon. Looking at the predicted lncRNA-gene interactions we got 9,786 possible interactions. Of these 4,797 were classified as the best interaction. Also 324 (=∼6.32 %) lncRNAs had no interaction partner identified. The most interactions belonged to the intergenic type with 'unknown strand' subtype (2,998). If those are removed however the most interactions would be classified as genic (table 3.8). Excluding the 'unknown strand' subtype most interactions were predicted as 'genic'-'nested'-'intronic' (544). We found no interactions of the types 'intergenic'-'divergent'-'downstream' and 'intergenic'-'convergent'-'upstream'.

Table 3.8.: FEELnc lncRNA-gene interaction results for *A. rosae* (Aros), *O. abietinus* (Oabi), *A. mellifera* (Amel), and *N. vitripennis* (Nvit). Only the best interaction for each lncRNA was added to this table. loc = location, up = upstream, down = downstream, ex = exonic, int = intronic

type	subtype	loc	Aros	Oabi	Amel	Nvit
intergenic	divergent	up	71 (2 %)	81 (1.7 %)	596 (13.3 %)	128 (22.1 %)
		down	0	0	0	0
	convergent	up	0	0	0	0
		down	48 (1,3 %)	69 (1.4 %)	316 (7 %)	49 (8.4 %)
	same strand	up	77 (2.2 %)	87 (1.8 %)	513 (11.4 %)	65 (11.2 %)
		down	88 (2.5 %)	82 (1.7 %)	413 (9.2 %)	64 (11 %)
	unkown strand	up	940 (26.3 %)	1380 (28.8 %)	0	0
		down	892 (25 %)	1618 (33.7 %)	0	0
genic	overlapping	ex	427 (12 %)	468 (9.8 %)	479 (10.6 %)	88 (15.2 %)
		int	8 (0.2 %)	5 (0.1 %)	137 (3 %)	18 (3.1 %)
	containing	ex	128 (3.6 %)	151 (3.1 %)	23 (0.5 %)	2 (0.3 %)
		int	10 (0.3 %)	17 (0.4 %)	67 (1.5 %)	13 (2.2 %)
	nested	ex	414 (11.6 %)	295 (6.1 %)	328 (7.3 %)	36 (6.2 %)
		int	470 (13.2 %)	544 (11.3 %)	1626 (36.1 %)	117 (20.2 %)
Total			3573	4797	4498	580

3.3.11. lncRNA-protein-coding gene interaction in *Apis mellifera* and *Nasonia vitripennis*

Extracted from the respective OGS we provided 4,749 lncRNAs for *A. mellifera* and 605 for *N. vitripennis* to the FEELnc_classifier. For 251 (=∼5.29 %) in *A. mellifera* and 25 (=∼4.13 %) in *N. vitripennis* no interaction was found. In total 25,696 (*A. mellifera*) and 3,702 (*N. vitripennis*) lncRNA-gene interactions were predicted. Of these 4,498 (*A. mellifera*) and 580 (*N. vitripennis*) were classified as 'isBest'. For both species the majority of interactions belong to the 'genic' type in contrast to *A. rosae* and *O. abietinus* (table 3.8). Most interactions for *A. mellifera* were classified as 'genic'-'nested'-'intronic' (1,626), for *N. vitripennis* as 'intergenic'-'divergent'-'upstream' (128). None of the 'intergenic' interactions were classified with an 'unknown strand'.

4. Discussion non-coding RNAs

4.1. Database curation

Specialised databases are a useful for the identification of ncRNAs in as yet not annotated organisms. However, their usefulness depends a lot on the curation, data availability, and completeness. Both miRBase and Rfam rely on user interaction. Both databases are curated, but while manually checking the families we found entries assigned to the wrong organism or the wrong family. In quite a few cases it was a bacterial sequence that was found in an organism and was identified as belonging to Metazoa (Ludwig et al., 2017).

Another problem is that not all available data are included in these specialised databases. This is due to the fact that researchers have to send in their data to be included and there is no automated process that includes newly published data fitting into these databases.

A lot of ncRNAs have their own specialised databases, which in some cases only contain those of one organism. This decentralisation makes it harder to get a conclusive overview over the available data. Also, a large number of different databases increases the chance for some of them not having long time support. This creates the possibility of data getting lost as the databases vanish or don't get updated. It also creates the possibility of different sets for the same organism existing, which can create a problem in reproducability. Depending on the database, thoroughness on the documentation of how the data were generated varies, which can make it harder to create new data that is fitting or compare different datasets.

Curation is another matter in all these smaller, specialised databases. For both Rfam and miRBase the process is documented, but for all the smaller databases additional effort is needed to guarantee they have the same or a very similar standard as other databases.

Of course one could argue that the NCBI database provides a lot of this data. However, what is not present in this database is information about the families that both Rfam and miRBase provide. For most ncRNAs a seed region or another conserved part is important for identifying the family relationship. This information is not provided by the NCBI, and neither are ncRNA family models that can be used for further analysis. The way the NCBI database is organised makes it difficult to find all relevant data. As stated above the family information is missing, which is problematic in cases where especially miRNA families were combined without updating the naming scheme. The mir-2 family is a good example for this, where some members are named mir-13 for historical reasons, but new additions to this family still follow the naming scheme as it shows which single members are closest related.

Using a centralised database increases the data available for analyses all in the same format without having to search through several different databases. The current non-standardised format of different databases makes it harder to combine data. This makes it less likely for researchers to combine as many datasets as possible for a comprehensive analysis.

For our analysis we only used the data on ncRNAs available in the two databases Rfam and miRBase. The databases contain miRNAs, tRNAs, rRNAs, snRNAs, and snoRNAs, but no piRNAs and only a very limited number of lncRNAs. The selected databases reduced our species set as well as the number of annotated ncRNAs for those species that we used in our analysis. We accepted these restrictions for our analyses because the curation and ncRNA family information in the Rfam and miRBase were deemed more important than a more complete dataset.

The other ncRNAs missing from these databases should be found through our *de novo*

analyses, however, we did not check for an overlap between our predictions and ncRNA predictions from other databases.

4.2. Homology prediction of non-coding RNAs

The pattern of the miRNAs identified in *A. rosae* and *O. abietinus* fits with the known patterns for Hymenoptera miRNAs. From the results of the other Hymenoptera present in the miRBase we expected to identify miRNAs in 65 different families. With miRNAs found belonging to 60 (*O.abietinus*) and 62 (*A. rosae*) different families we stayed slightly below this expectation. However, except for two families (mir-3718, mir-3747), we found all expected families in at least one of *A. rosae* and *O. abietinus*. The two known members of the mir-3747 family and the two of mir-3718 listed in miRBase are found in *A. mellifera*. We did not find it in our other Hymenoptera, making it likely that this miRNA evolved in the lineage leading to the honeybee, most likely after the split of Aculeata and the remaining Apocrita.

In *A. rosae* we identified miRNAs belonging to 60 of those families present in Hymenoptera and one other (mir-1923) which is not present in the other Hymenoptera species or in *T. castaneum* (table 3.1). mir-1923 has been so far only identified in *Bombyx mori* and *Acyrthosiphon pisum*, making it an insect specific miRNA family that is not shared between many species. The function of this family is not known and it is therefore impossible to create a hypothesis on the actual distribution of this family in insects.

In *O. abietinus* we identified miRNAs of 62 of the Hymenoptera miRNA families and no unexpected ones (table 3.1). Overall the pattern of the ncRNAs predicted through homology are very similar to other Hymenoptera.

In both *A. rosae* and *O. abietinus* we did not find some miRNA families known from other Hymenoptera. These families showed a mixed present-absent pattern in the different Hymenoptera, making it difficult to extrapolate any lineage specific losses and gains. They could just be missing from the genome assemblies or be really absent. Further research is needed to answer this question.

Some of those miRNA families can be found as multiple copies in the Hymenoptera genomes, such as mir-2, which has five copies in most species. In those cases we expected to identify similar copy numbers. As found in other species, we identified in both *A. rosae* and *O. abietinus* a cluster of the mir-2 family. A cluster of the mir-2 family is also present in *A. mellifera* and *N. vitripennis* (figure 3.2). In *A. rosae*, *O. abietinus*, and *N. vitripennis* the mir-2 cluster consists of five copies and has the same miRNAs at the ends (mir-2b and mir-2c). In *A. mellifera*, mir-2c is not present, but mir-2b also marks the start of the cluster. In this species we have a total of six mir-2 genes creating the cluster. However, the mir-2b is orientated into a different direction than the cluster and is completely nested inside mir-2-1. The middle part of the cluster varies slightly. In three species mir-13a follows the first mir-2 copy of the cluster (*O. abietinus*, *A. mellifera*, *N. vitripennis*) and three have mir-2a as second to last (*A. rosae*, *A. mellifera*, *N. vitripennis*). Other miRNA cluster we found split over different scaffolds. Better assemblies can shed light onto these cases if the spatial orientation is conserved or not. Especially methods that produce long reads, such as PacBio or the Oxford Nanopore Technology sequencing systems.

The biggest problem comparison-wise is that in *A. mellifera* a different naming scheme was used. Even if the total composition of the cluster varies, it seems that one end of the cluster is conserved in Hymenoptera. Furthermore, in all four species mir-71 can be found next to the cluster end where mir-2c is if present or would be located if missing outside the cluster.

Only 5 miRNA families are present in the miRBase that are only present in Hymenoptera (mir-928, mir-3477, mir-3478, mir-3718, mir-3747). If one compares this number with miRNAs lineage specific to Diptera (around 50 families present in miRBase are only found in Diptera) the number of known families is smaller.

4.3. *de novo* prediction of non-coding RNAs

In general it is important to realise that the full ncRNA repertoire of a species can never be identified through homology prediction only. It is expected that all species have some species specific ncRNAs that will not be present in a database. Also, if one does not work with model organisms or species closely related to these, lineage specific ncRNAs will not be found.

We expected to identify at least one tRNA gene for each amino acid present in multiple copies. This was true for both *A. rosae* and *O. abietinus*. Our numbers of 177 tRNA genes (*A. rosae*) and 158 (*O. abietinus*) are lower than the ones reported from Behura et al. (2010) for *Nasonia* and *Apis* (221 and 199). However, they show a similar number of tRNA genes and the overall number of tRNA genes can vary a lot between species (e.g. 85 in *Drosophila melanogaster or 496 in Bombyx mori* (Behura et al., 2010)), as it is dependent on the codon usage of a species. tRNAs containing introns are known from several species (Behura et al., 2010). However, which tRNA contain introns varies. In Hymenoptera and other insects they have been found in tRNA-Tyr, -Ile, and -Leu genes (Behura et al., 2010). We also only identified introns in these tRNAs and as is known from *A. mellifera* and *N. vitripennis* we found no tRNA-Tyr without introns. tRNAs containing introns have been shown to be involved in base modification of the anticodon triplet (Behura et al., 2010), but which tRNAs contain introns varies between species.

We identified two different types of snoRNAs, H/ACA and C/D. Our homology analysis resulted in only 14 snoRNA families in *A. rosae* and 11 in *O. abietinus*. This small set likely caused the high number of false-positives DARIO predicted (table 3.6). Another reason could be that our RNAseq-reads were of a quality that allowed DARIO to correctly predict their stack pattern. Comparing the numbers of snoRNAs known from other insects, we find that the ensembl Metazoa database (Zerbino et al., 2017) lists 7 snoRNAs for *A. mellifera* and 8 for *N. vitripennis*, but 292 for *D. melanogaster*. We assume that we did not identify all snoRNAs in *A. rosae* and *O. abietinus* as each

snoRNA can only direct one or two rRNA modification and this alone would point to an expected number of snoRNAs over 200 (Bachellerie et al., 2002).

Influence of different short-read preparation

Using the DARIO pipeline we compared three different types of trimmed and mapped reads. Our first set was merging the paired-end reads and mapping only those, the second was using only those reads that were still a complete pair after trimming, and the third mapped all reads that remained after trimming. Strictly speaking the second set is a subset of the third. Our results show that it makes a difference how the mapped reads are treated beforehand. The biggest difference was between the read set using merged reads and the two others with unmerged reads. There seems to be a core set of reads that can be mapped regardless of their treatment before mapping. However, looking at the predictions shared between all sets, we found one set of the three that had an ncRNA set unique to this one that was similar in size. Our results do not lead to a recommendation of the best way to treat reads before mapping but shows that it is important to look at all three sets and put further work into it. Additional lab work should be done to look at the validation of our results.

We used a conservative method to create our final set of ncRNAs by removing all those predicted ncRNAs that were either overlapping exons (strand independent), ncRNA predictions or overlapping another DARIO prediction. ncRNAs and exons of protein-coding genes can be found at the same region of a genome, but then they exist on different strands. Also, ncRNAs can be present in UTRs which are not distinguished from exons in *A. rosae* and *O. abietinus* in the official gene set. Our sequenced reads do not contain strand information, so it is not possible to check the strandedness of our prediction even though DARIO itself does predict a strand. This might exclude true positive predictions from our final list but probably also lessens the false positive results we would get.

In mammals, it has been shown that snRNAs and snoRNAs underwent massive expansions over time which coincided with the diversification of said group (Hoeppner et al.,

2018). Even though it is still up for debate whether all those expansions led to more functional ncRNAs, it would be interesting to take a look at insects in this regard. For this, an analysis of additional insect lineages needs to be done. Our analysis can therefore only be seen as a first, but important, step in this direction.

Our *de novo* prediction of lncRNAs in *A. rosae* and *O. abietinus* showed that the trends of lncRNA-gene interaction are similar between the species even though the absolute numbers are different. The majority of the predicted interactions are of the intergenic type for both species, even though this includes still a large number were the subtypes could not be determined.

In total numbers our study predicted more lncRNAs in *A. rosae* and *O. abietinus* than are present in the official gene sets of *N. vitripennis* and *A. mellifera*. Especially *Nasonia* stands out with a current number of 784, which is way lower than all others. The most likely explanation is the pretty recent focus of lncRNA research and not a lot of work being done on these organisms so far. The total number of lncRNAs that are supposed to be present in a genome can not be identified through our analysis and further work on this, as well as the conservation of lncRNAs between insects, has to be done.

4.4. Non-coding RNA repertoire of *Athalia rosae* and *Orussus abietinus*

Ideally we would have been able to identify the complete ncRNA repertoire of the two Hymenoptera species. However, our exclusion of certain ncRNA types (e.g. piRNAs) from our analysis made this impossible. For those ncRNA types that we looked at we significantly increased the number of identified ncRNAs. This shows that the usage of only homology prediction in as yet not annotated species is never enough to build a conclusive picture of the gene repertoire. Of course this is still not a comprehensive set of species from all Hymenoptera lineages, but a far broader set than was available beforehand.

Our *de novo* prediction of miRNAs and snoRNAs relied on short RNAseq libraries. These were whole-body and only from adults. This makes it hard to identify tissue or stage specific ncRNAs because they are lowly expressed if at all in these transcriptomes.
The best way to get a good idea of the repertoire is a combination of homology and *de novo* prediction with well sequenced genomes, extensive short RNAseq reads, and additional lab work to validate the predictions.
The basis for a good homology prediction is a big evidence base from various closely related species, ideally from the same lineage. For this, the research focus needs to shift from a couple of well studied model organisms to a broad variety of non-model organisms.

5. Methods conserved non-coding elements

5.1. CNEr

We identified CNE candidates in the four Hymenoptera species *Apis mellifera*, *Athalia rosae*, *Nasonia vitripennis*, and *Orussus abietinus* using the R Bioconductor package 'CNEr' (Tan, 2015). This package uses pairwise whole genome alignment (WGA) and genome annotation to identify CNEs. The pairwise alignments were created using the program last (version 744) (Kiełbasa et al., 2011) with the MAM8 seed (Frith and Noé, 2014). A total of six runs with CNEr were done to get results from all possible pairwise genome alignments. For the analysis we followed the CNE identification guideline by Ge Tan (`http://rpusb.com/yang2/CNEr3`) with some changes: Our definition of CNEs included only sequences of ≥ 100 bp and a minimal conservation of 70 %. Due to that we changed the window sizes for CNEr to 100 and used the identity thresholds 70, 90, and 99 %.

CNEr only identifies CNE candidates in regions that do not already contain a known gene. For this it uses the genome annotation of the species. We provided CNEr with all annotated exons of this species, which we extracted from the OGSs for each species (Aros v1.0, Oabi v1.0, Amel v4.5 (GCF_000002195.4_-Amel_4.5) (Elsik et al., 2014), Nvit v.2.1 (GCA_000002325.2) (Werren et al., 2010)). The list of CNE candidates were further checked for repeats using the CNEr internal blat function. All remaining candidates were then further analysed using custom Perl scripts.

Our sampling included four species, but CNEr only supports pairwise analysis. We did six pairwise analyses and we combined the three different CNEr output files per species.

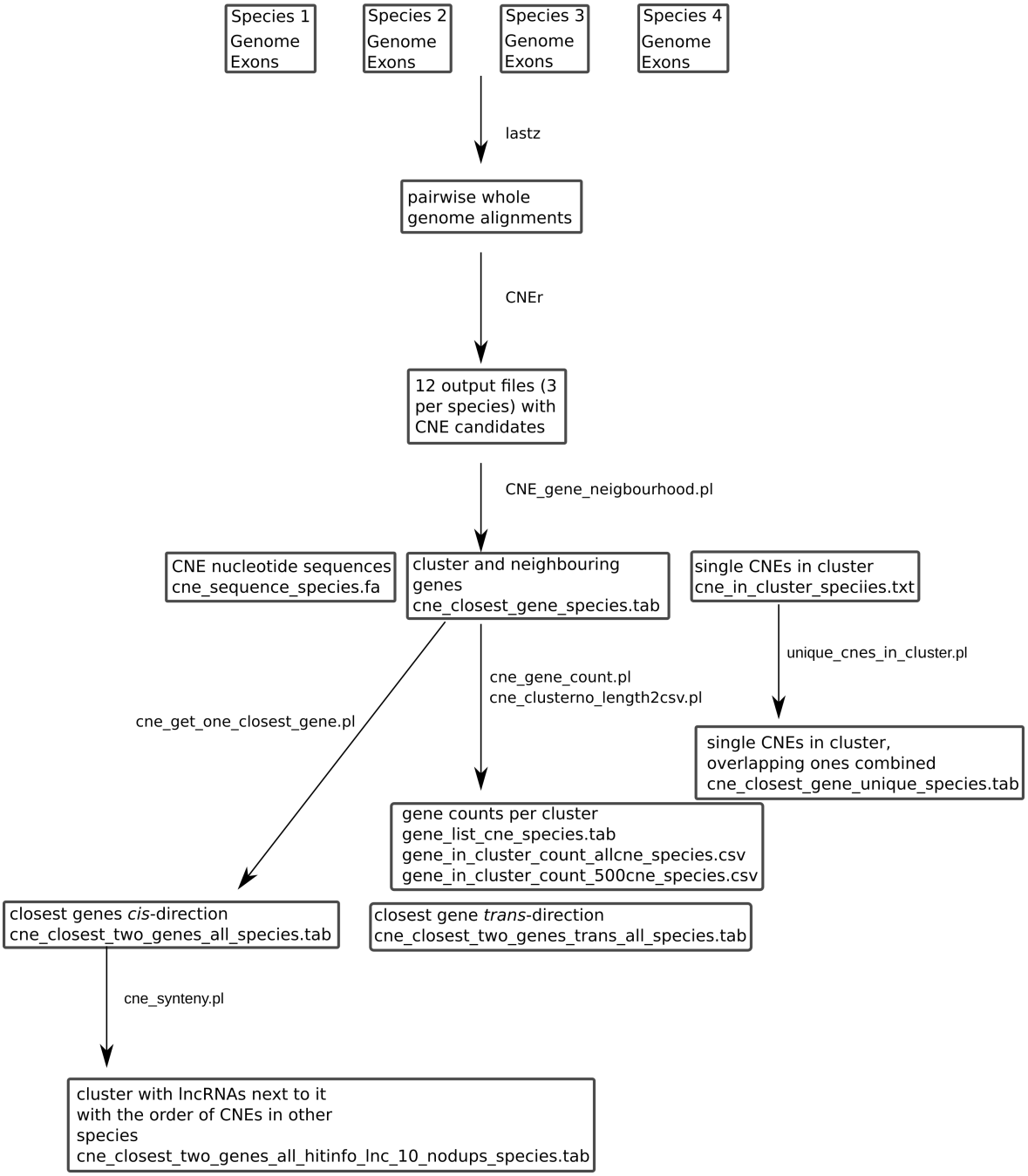


Figure 5.1.: Graphical overview of the steps in the pipeline used for CNE prediction.

5.2. CNE_gene_neighbourhood.pl

To this end, we developed a custom perl script CNE_gene_neighbourhood.pl to combine the output files. The script needs one reference species with which the three different analyses were conducted. All CNE candidates were sorted and combined if they overlap by at least one nucleotide using the reference species. Also, for combined CNEs the borders of the CNE were adapted to include the longest sequence possible. Each CNE of the reference species ended up with at least one CNE candidate in another species. The CNEs in the other species were not checked for overlap in this script.
The CNEs of the reference species were not only checked for overlap but were also sorted into cluster. Two CNEs belonged to the same cluster if they are $\leq$20,000 bp apart from each other. Woolfe et al. (2004) showed that still 85 % of CNEs cluster within 370 kb distance, however, we chose this conservative distance to take into account our two assemblies that are not at chromosome level.
This script produces three output files. One contains the nucleotide sequence of each CNE (cne_sequence_species.fa), the second information to each cluster (scaffold, start, stop, count of CNEs, distance to scaffold end) and all genes that were found within $\leq$500 kb distance to this cluster or within it (cne_closest_gene-_species.tab), and the third the position information for each CNE in a cluster (cne_in_cluster-_species.txt). We used 500 kb as the distance because Woolfe et al. (2004) found that 93 % of the CNE cluster they identified had a *trans-dev* gene within this distance. The species part of the file name is a placeholder for the reference species. These files still listed overlapping CNEs separately but were combined into one in the next step.

5.3. unique_cnes_in_cluster.pl

The next script unique_cnes_in_cluster.pl takes the file cne_in_cluster_species.-txt as input and merges all the overlapping CNEs. It also checks if two clusters should be

merged. This case happened because the borders of each cluster, meaning the most outwards placed CNE, are expanded with each CNE that gets added. In some cases the distance between two clusters was $\leq$ 20,000 bp after the finished analyses and that classifies the CNEs as belonging to the same cluster. If two clusters were 20,030 bp apart and one was extended by 31 bp they would now count as one cluster. This reevaluation was not done in the previous steps.

5.4. cne_gene_count.pl

The script cne_gene_count.pl takes the file with all genes neighbouring a cluster and counts how often each gene was present and saves this to a file (gene_list_cne_-species.tab). It also created two files that contained for each cluster the numbers of genes found upstream, downstream or within a cluster. The first one contained all clusters (gene_in_-cluster_count_allcne_species.csv) and the second only those clusters that had a minimal distance of 500 kb to each scaffold end (gene_in_cluster_count-_500cne_species.csv). We included the second file because the number of genes neighbouring clusters with $\leq$ 500 kb distance to the scaffold end might be artificially lower as the search for genes stops at the end of a scaffold even though the actual chromosome might be longer. The length of each individual cluster was added using the script cne_clusterno_length2csv.pl.

5.5. cne_get_one_closest_gene.pl

Further analysis focused on just the closest genes in *cis* on each cluster side. This gene was identified using the script cne_get_one_closest_gene.pl. It took the list of genes per cluster (cne_closest_gene_species.tab) created by the CNE_gene-_neighbourhood.pl script and the annotation of the species to find it. The gene that was closest to the cluster is selected. Also the direction of the gene towards the cluster orientation was checked. If the closest gene was in *cis* to the cluster it was added with

the cluster info to the file cne_closest_two_genes_all_species.tab. In case this gene was in *trans* to this file 'na' was added and the cluster with the gene information was stored in the file cne_closest_two_genes_trans_all_species-.tab.

5.6. cne_synteny.pl

The next step was to get the information for each gene that was identified as closest to a cluster. We checked what type it is, meaning protein-coding or lncRNA, focusing further on those clusters that had an lncRNA identified as a possible interaction partner. For those clusters with at least one lncRNA in *cis* as interaction partner we checked the synteny of the single CNEs in the cluster.

The script cne_synteny.pl was used for this. It takes a file with the cluster information (scaffold, start/stop position, closest gene (upstream/downstream), distance to the gene (upstream/downstream) number of CNEs in the cluster, distance to scaffold end) of interest of the reference species, in this case those clusters with an lncRNA next to them consisting of ≥ 10 CNEs, the CNEr output files, and the file for each species containing the final CNE coordinates as provided by the unique_cne_in_cluster.pl script. This script provides information on the partner for each CNE in a cluster together with the position of the CNE in the other species. The file this script produces (cne_closest_two_genes_all_hit-info_lnc_10_no-dups_species.tab) contains for each cluster provided a list of CNE matches in the other species. As a link between the CNEs it provided, this file allows us the check for synteny between species.

5.7. cne_diff_species_ident.pl

To visualise and compare the CNEr results, we created venn diagrams for each species. For each species we collected the CNEs predicted in all three runs of CNEr and compared them. Predictions were counted as the same if at least 1 nt was overlapping. Using the script cne_diff_species_ident.pl we created an ID for each CNE that made them comparable between the three output files. The list created for each output was then

passed on to Venny (Oliveros, 2015) which created the venn diagrams.

6. Results conserved non-coding elements

6.1. CNE prediction

The way we set up the CNE identification pipeline, we always used one species as a reference and got CNE predictions in pairwise comparisons with the three other species. The total amount of CNEs predicted per species varied between 5,740 and 12,462 CNEs (table 6.1). Most CNEs were predicted in interaction with *O. abietinus* (12,462), followed by *A. mellifera* (9,887), *A. rosae* (7,263), and *N. vitripennis* (5,740). In all analyses we identified CNEs that were present in all four species, with varying numbers. *N. vitripennis* was the species with the lowest number of CNEs found in all three other species (316) (figure 6.1c), followed by *A. rosae* with 321 (figure 6.1b), *A. mellifera* with 410 (figure 6.1a), and in *O. abietinus* with 490 (figure 6.1d) most CNEs found in all species were identified.

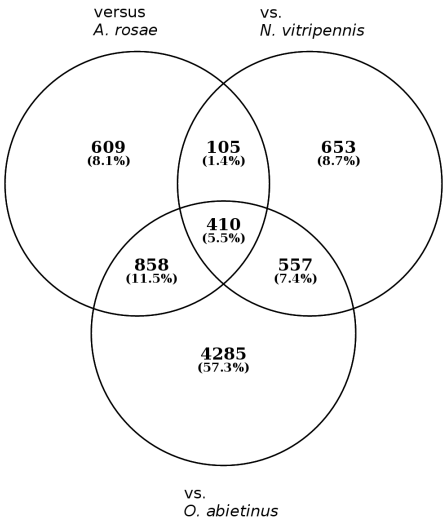
The most CNEs were always predicted in the comparison with *O. abietinus* and only a fraction of the CNE candidates were identified in all four species (between 5.2 % in *O. abietinus* and 7.7 % in *N. vitripennis*). The most CNEs identified in a pairwise comparison were found between *A. mellifera* and *O. abietinus* (4,285) and the least between *A. mellifera* and *A. rosae* (609). The majority of predicted CNEs was identified in three out of the four species. The total number of CNEs was reduced during further analysis through combining the overlapping CNEs. In *A. mellifera* we found 410 CNEs

that were present in all three pairwise analyses. The most CNEs found in only one analysis was found with *O. abietinus* (4,285) (figure 6.1a).

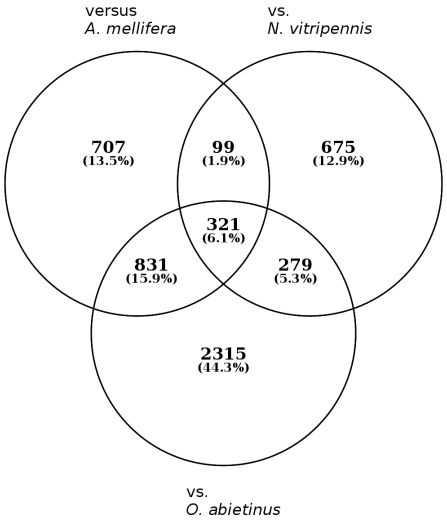
Through combining of overlapping CNEs the total number dropped in all four species by over 1,000 CNEs (table 6.1).

Table 6.1.: Number of CNEs identified by CNEr sorted by species, number of CNEs left after overlapping ones were combined, size of the assembly (Mb), and N50 (kb) of the assembly.

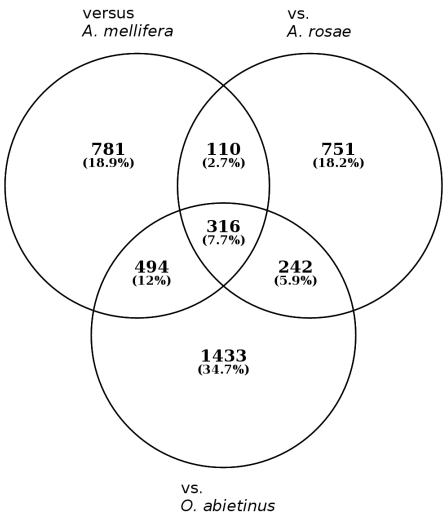
Species	CNEr results	Combined	Assembly size	N50
Athalia rosae	7,263	5,224	164	1370
Orussus abietinus	12,462	9,449	201	2370
Apis mellifera	9,887	7,474	250	997
Nasonia vitripennis	5,740	4,127	295	708



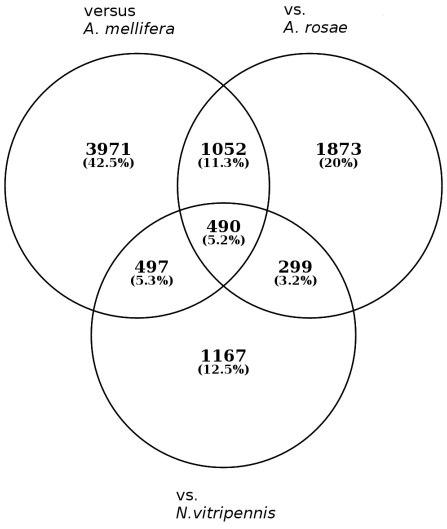
(a) *Apis mellifera*



(b) *Athalia rosae*



(c) *Nasonia vitripennis*



(d) *Orussus abietinus*

Figure 6.1.: Total number of CNE candidates identified with CNEr with each species as reference species. Overlapping CNEs are not combined.

6.2. CNE cluster analysis

In further analysis these CNE predictions were sorted into clusters if they were ≤ 20 kb apart and only the CNEs after combining overlapping ones were used. Combining the

CNEs into clusters we found that the lowest number of clusters was found in *A. rosae* (1,599) and the highest in *O. abietinus* (2,088) (table 6.4). The number of CNEs per cluster varied. In all species the majority of CNEs were not found in clusters but as single CNEs (table 6.2, figure 6.2). The largest cluster group consisted of 2-4 CNEs per cluster. The cluster groups in table 6.2 were chosen arbitrary to visualise it better. The number of clusters decreases the more CNEs are included. All four species had at least one cluster that contained over 100 CNEs (table 6.2). Overall the largest cluster CNE count wise was found in *O. abietinus* with 342 CNEs, followed by *A. mellifera* (228), *A. rosae* (217), and *N. vitripennis* (175) (table 6.2).

Note that the maximal amount of CNEs per cluster is also dependent on the assembly. The larger the assembled scaffolds are, the bigger a cluster can get.

Table 6.2.: Number of CNE clusters, grouped by CNE numbers, for each species. Grouping was chosen arbitrary. Last row shows the number of CNEs making up the largest cluster of a species.

Species	1	2-4	5-9	10-49	50-99	≥ 100	largest cluster
Athalia rosae	839	524	156	72	8	1	217
Orussus abietinus	1,004	673	240	143	28	7	342
Apis mellifera	1,046	585	174	122	21	3	228
Nasonia vitripennis	989	441	132	61	2	1	175

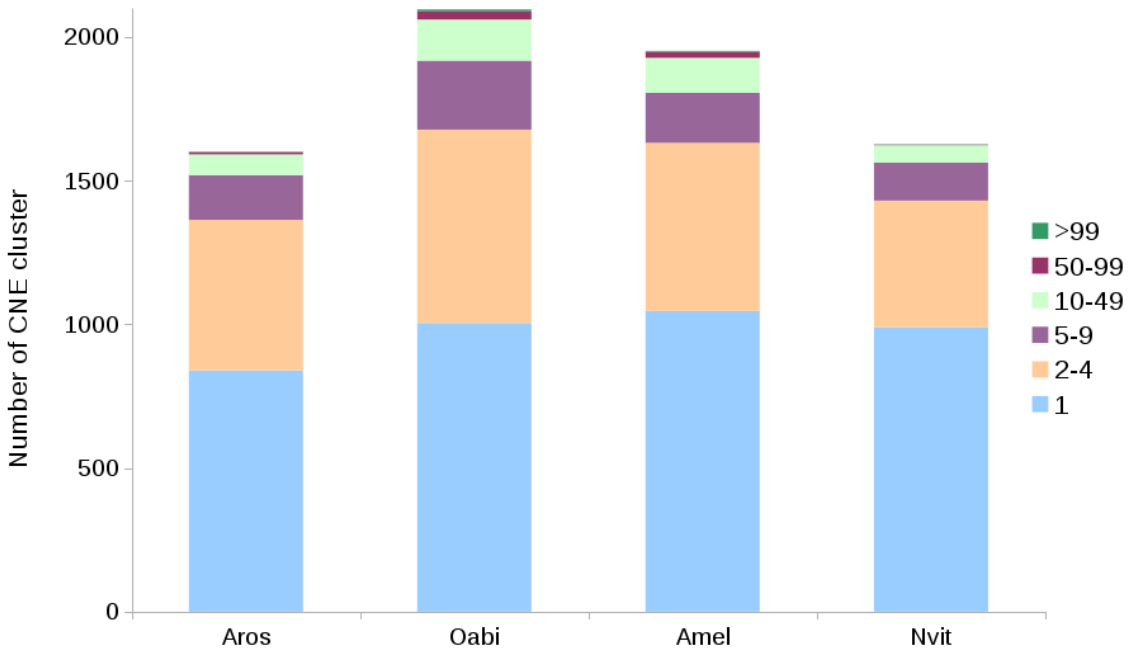


Figure 6.2.: Visualisation of the numbers of CNEs per cluster seen in table 6.2 in the four species.

6.3. CNE gene interaction

To each cluster we assigned the two closest protein-coding genes or lncRNAs (one for each side of the cluster) if they were in *cis* orientation to the cluster (table 6.4), which will be called associated gene. Except in *A. rosae* the majority of clusters had no gene in *cis* direction assigned (table 6.4). Following this protein-coding genes were the next biggest type assigned to cluster (table 6.4). The number of different lncRNAs identified next to clusters varied between 36 (*N. vitripennis*) and 548 (*A. mellifera*). In some cases one lncRNA was assigned to two clusters when no other gene was found between the two clusters. The number of unique lncRNAs was lower. In *N. vitripennis* we found 27 lncRNAs, 56 in *O. abietinus*, 292 in *A. rosae*, and 435 in *A. mellifera*. In both *A. mellifera* and *O. abietinus* the number of lncRNAs found upstream and downstream of a cluster was comparable (38 in both directions in *O. abietinus* and 288 to 260 in *A. mellifera*), whereas in *N. vitripennis* all lncRNAs were found upstream. In *A. rosae*, the majority of lncRNAs was found downstream of CNE clusters (table 6.5).

Table 6.4.: Total amount of CNE clusters per species as well as the count of clusters with at least one lncRNA in *cis* as the closest gene, number of clusters with protein-coding genes (gene) in *cis*. N/A shows the number of clusters where no gene was found next to it or were the closest gene was in *trans*.

Species	Cluster	Gene	lncRNA	N/A
Athalia rosae	1,599	1,464	322	1,142
Orussus abietinus	2,088	2,040	76	2,060
Apis mellifera	1,948	1,248	548	2,100
Nasonia vitripennis	1,625	325	36	2,555

As lncRNAs are not described as interaction partners with CNEs, we looked at the protein-coding gene/lncRNA ratios (table 6.3). In *A. mellifera* lncRNAs do not occur more often next to a cluster than would be expected. In two species they occur less often than expected, 1.3 times lesser in *A. rosae* and 12 times lesser in *O. abietinus*. In *N. vitripennis* the were found twice as often as expected neighbouring a cluster in *cis*.

Table 6.3.: Ratios of lncRNA/protein-coding genes in each species. First number is calculated from all lncRNAs and protein-coding genes present in the OGS, second is calculated from the lncRNAs and protein-coding genes that were found in *cis* next to a CNE cluster.

Species	Ratio whole annotation	Ratio CNE cluster neighbours
Athalia rosae	0.30	0.22
Orussus abietinus	0.46	0.037
Apis mellifera	0.44	0.43
Nasonia vitripennis	0.045	0.11

Between 325 (*N. vitripennis*) and 2,040 (*O. abietinus*) different protein-coding genes were identified as neighbouring a cluster in *cis* direction. Except for *N. vitripennis*, over 1,000 genes were identified as neighbouring a cluster in *cis*: 1,248 (*A. mellifera*), 1,464 (*A. rosae*), 2,088 (*O. abietinus*).

Table 6.5.: CNE clusters with an lncRNA in *cis* direction next to it. Total includes every occurrence of an lncRNA in the right direction next to a CNE cluster, upstream is the total count of those found upstream, downstream the total count found downstream of a cluster, and unique lncRNAs is the count of different lncRNAs identified.

Species	total	upstream	downstream	unique lncRNAs
Athalia rosae	322	22	300	292
Orussus abietinus	76	38	38	56
Apis mellifera	548	288	260	435
Nasonia vitripennis	36	36	-	27

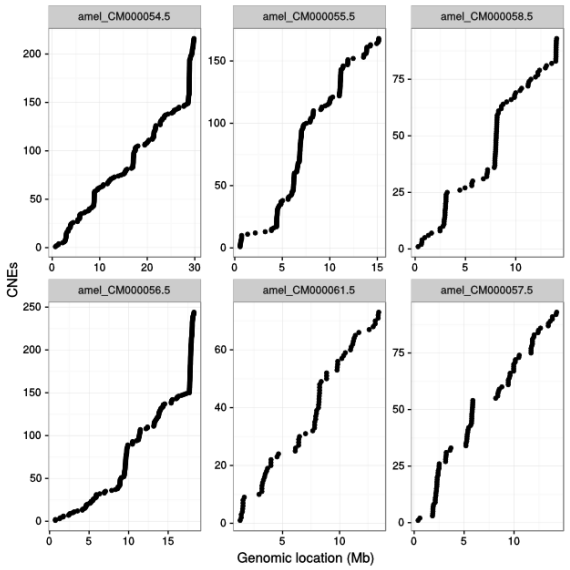
For each CNE cluster we set a maximum distance of 500 kb in which a gene had to be located. This distance was chosen as other studies showed that genes of interest tend to be located inside this region. The distance between the cluster and the closest gene varied between 1 bp (found in all four species) and 483,349 bp (*O. abietinus*).

As stated above we found between 36 and 548 cases of lncRNAs next to a CNE cluster in *cis* direction in one species (table 6.5). The highest number of 548 was found in *A. mellifera* with 288 lncRNAs found upstream and 260 found downstream of a cluster. Reducing this number to unique lncRNAs 435 genes remain.

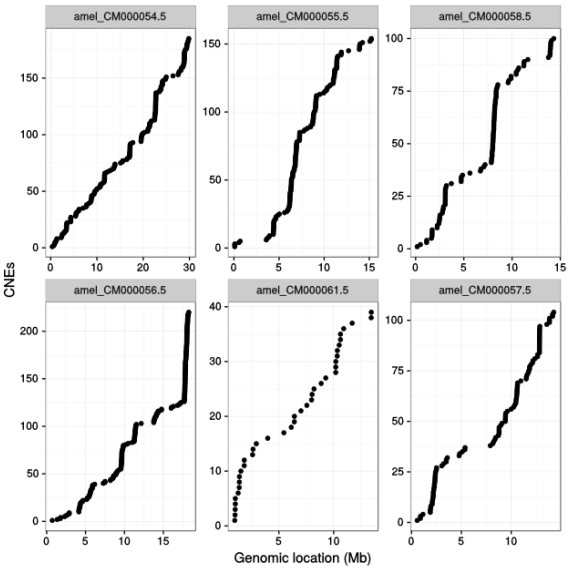
In *N. vitripennis* all lncRNAs identified as the associated gene of a cluster were found upstream, whereas in all three other species lncRNAs were found both upstream and downstream of CNE clusters. In *A. mellifera* and *O. abietinus* the number between upstream and downstream were similar, whereas in *A. rosae* the majority of lncRNAs was found downstream (300 genes downstream, 22 upstream) (table 6.5).

We selected the six largest scaffolds to get a look at the CNE distribution. The predicted CNEs were not equally distributed along the scaffolds of each species. This distribution is visualised in figures 6.3, 6.4, 6.5, and 6.6. For each species three graphs are shown, to show the results of the pairwise comparisons. Looking at figure 6.3 a) we see the distribution of the CNEs identified in the comparison of *A. mellifera* and *A. rosae*. Each of the six subplots shows the distribution on one scaffold. On the x-axis we see the genomic location of a CNE in Mb and on the y-axis the accumulative number of CNEs. In *A. mellifera* we see on some of those scaffolds only small gaps, meaning locations on

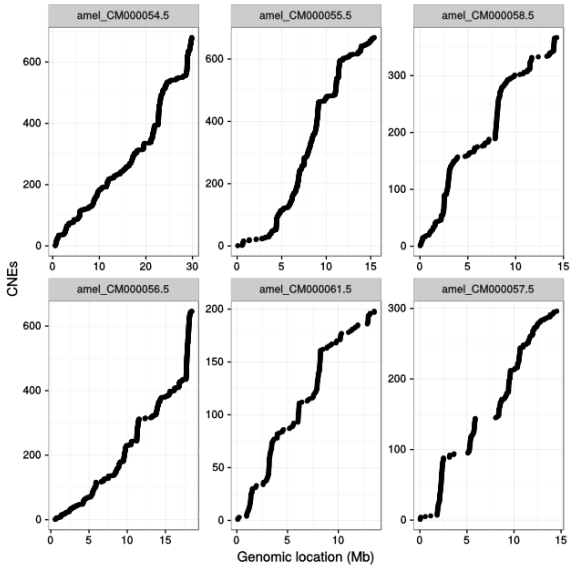
the scaffold where no CNEs were identified, and long stretches with no gaps in the CNE distribution, whereas on others we find a lot of single CNEs (figure 6.3). This uneven distribution is especially noticeable in *A. rosae* (figure 6.4). *A. rosae* has less CNEs identified on the six largest scaffold compared to the other three species (up to 60 in *A. rosae* compared to up to 600 in the other species). In all four species the distribution of CNEs between all three pairwise comparisons is similar (figures 6.3, 6.4, 6.5, 6.6). Scaffold 1 of *O. abietinus* is an example where the majority of CNEs were identified in the middle of the scaffold. This leads to a high increase of the total number over a small amount of basepairs (figure 6.6). Note that the scale of the y-axis is not unified.



(a) *Athalia rosae*

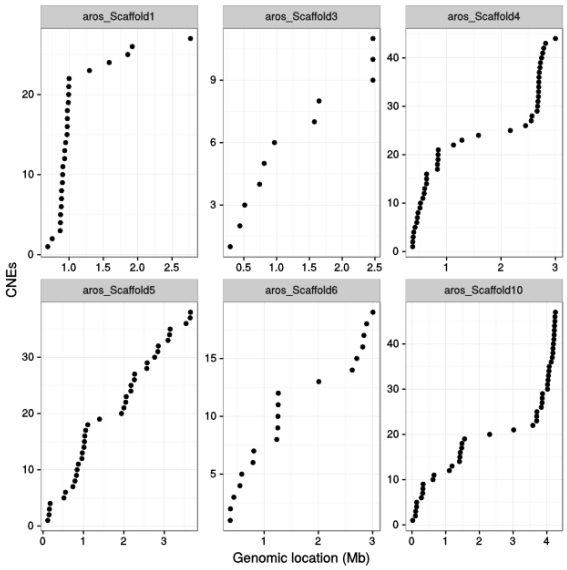


(b) *Nasonia vitripennis*

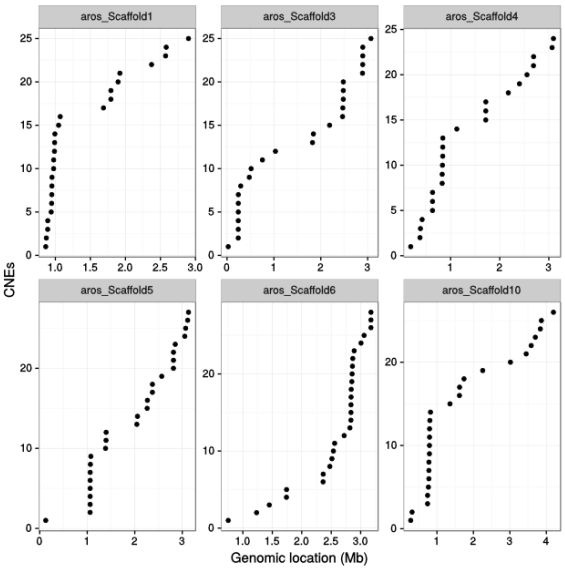


(c) *Orussus abietinus*

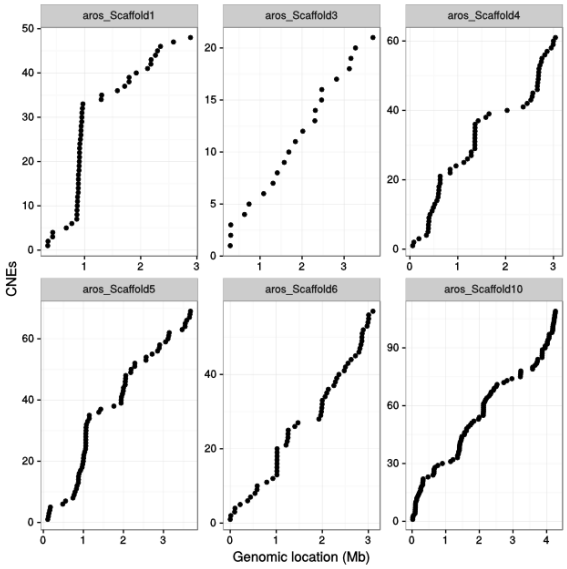
Figure 6.3.: Distribution of CNE prediction in *Apis mellifera* differentiated by species. Only the results for the six longest scaffolds are shown. The number of CNEs is the accumulative total amount found on this scaffold. x-axis shows the genomic location on the scaffold, y-axis the number of CNEs. The results are for pairwise comparisons between species.



(a) *Apis mellifera*

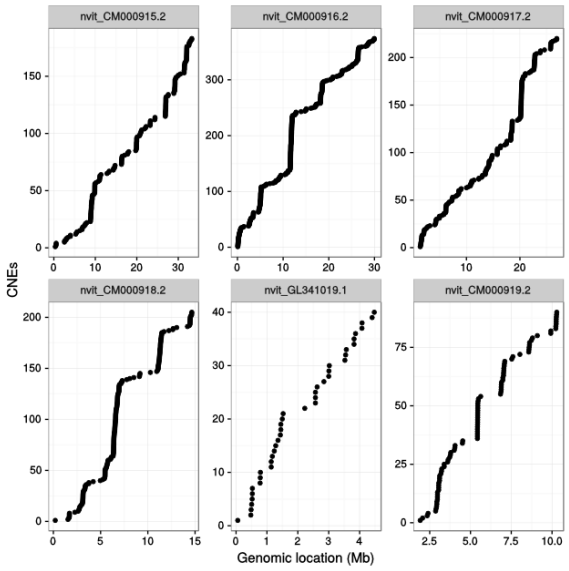


(b) *Nasonia vitripennis*

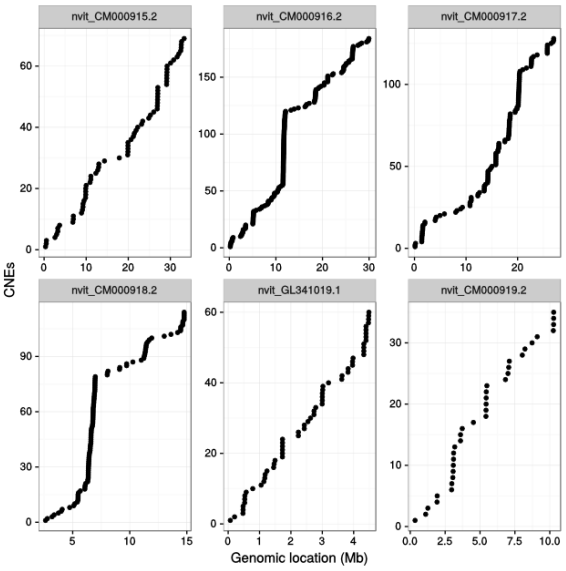


(c) *Orussus abietinus*

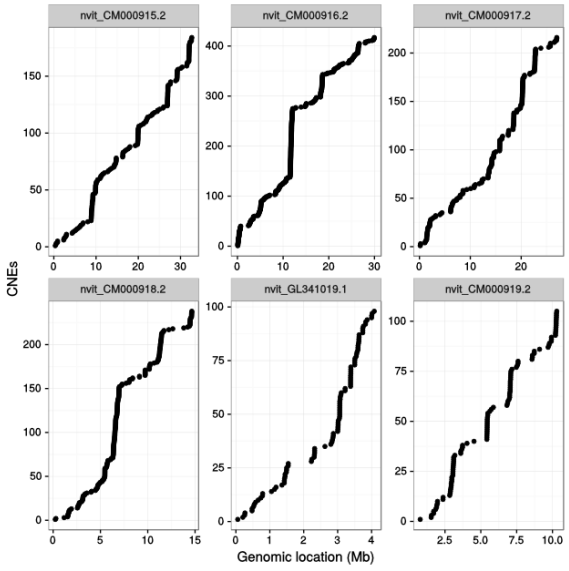
Figure 6.4.: Distribution of CNE prediction in *Athalia rosae* differentiated by species. Only the results for the six longest scaffolds are shown. The number of CNEs is the accumulative total amount found on this scaffold.



(a) *Apis mellifera*

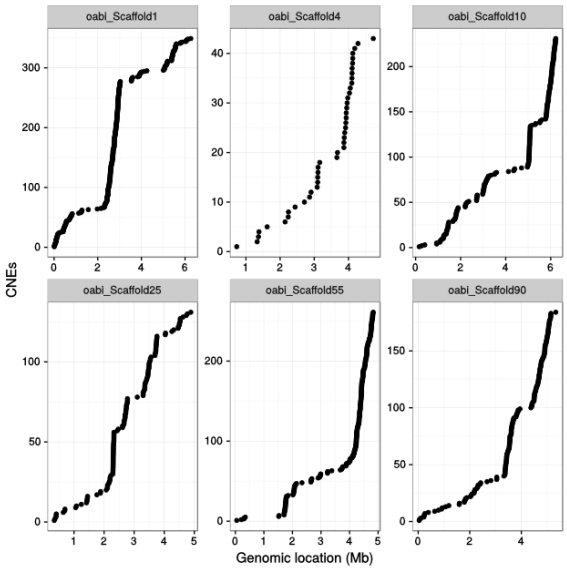


(b) *Athalia rosae*

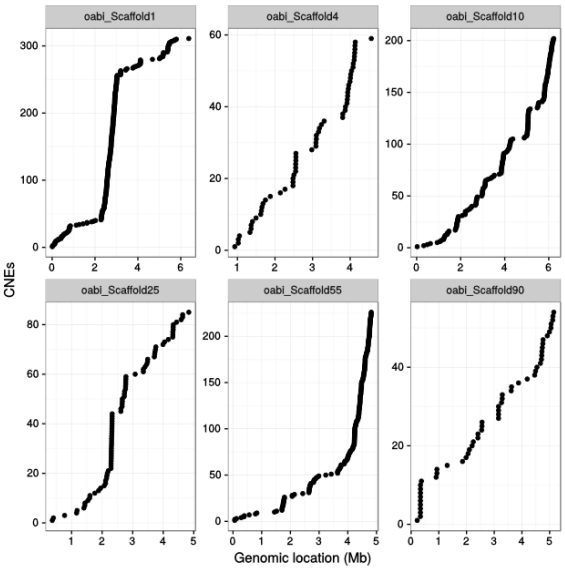


(c) *Orussus abietinus*

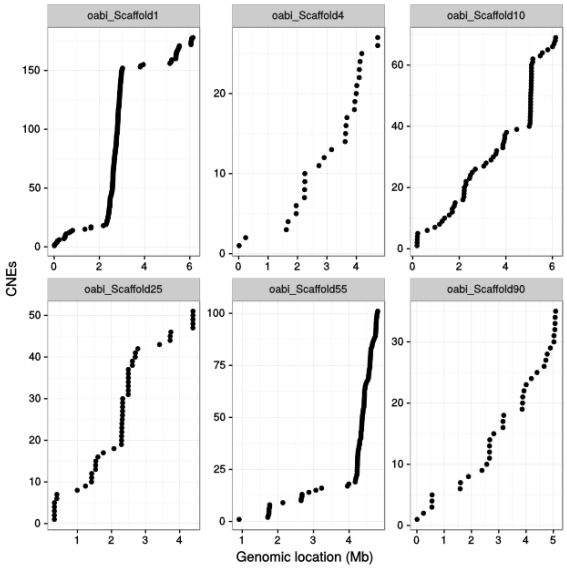
Figure 6.5.: Distribution of CNE prediction in *Nasonia vitripennis* differentiated by species. Only the results for the six longest scaffolds are shown.The number of CNEs is the accumulative total amount found on this scaffold.



(a) *Apis mellifera*



(b) *Athalia rosae*



(c) *Nasonia vitripennis*

Figure 6.6.: Distribution of CNE prediction in *Orussus abietinus* differentiated by species. Only the results for the six longest scaffolds are shown.The number of CNEs is the accumulative total amount found on this scaffold.

6.4. CNE cluster synteny

For the CNE cluster synteny analysis we focused on clusters containing at least 10 CNEs that had at least one lncRNA identified next to it. For *A. mellifera* this was the case for 47 cluster, two in *N. vitripennis*, 12 in *A. rosae*, and three in *O. abietinus* (table 6.6). Except for two clusters in *A. mellifera* all of these had only one lncRNA neighbouring the cluster. The closest a lncRNA was found to a CNE cluster were 1,417 bp (*A. rosae*).

Table 6.6.: Number of clusters consisting of ≥ 10 CNEs with an lncRNA as the associated gene with the information of the shortest distance between the cluster and the lncRNA found.

Species	Number of clusters	Shortest distance lncRNA-cluster (bp)
Athalia rosae	12	1,417
Orussus abietinus	3	3,092
Apis mellifera	47	2,046
Nasonia vitripennis	2	19,101

Each CNE in a cluster of a given reference species had at least one CNE hit in one of the other three species. Comparing the order of the CNEs between the different species we found that this order is at least partially conserved.

In all four species the cluster arrangement with lncRNAs next to it was at least partially conserved in one other species. We did not look into those clusters that had only protein-coding genes next to them.

One example using *A. mellifera* as the reference species showed that the cluster on scaffold CM000062.5 ranging from 8282101 to 8403340 was partially found in all three other species. This cluster consists of 33 CNEs. Of these CNEs 32 were also found in *O.abietinus*, however six CNEs were found twice. These CNEs were found in two groups on the same scaffold in the same order just with a different orientation. In *A. rosae*, nine CNEs of this cluster were found without duplications and seven in *N. vitripennis*, also without duplications. In *A. rosae* all CNEs corresponding to this cluster were found on the same scaffold, in the same orientation, and with less than 17 kb distance to each other. In *N. vitripennis* also all CNEs were present on the same cluster, but

in two cases the distance was more than 20 kb to the next CNE. Also the last CNE identified was found in a different orientation than the rest of the cluster.

In *O. abietinus*, the CNEs were spread over three different scaffolds. The distribution over the three scaffolds was not random but followed the order of the CNEs found in the cluster in *A. mellifera*. The first 17 CNEs were found on Scaffold 40, all in the same order as in *A. mellifera* with less than 13 kb distance. The next 12 CNEs were found on Scaffold 667, however this included the six duplicated CNEs. The order was also the same as in *A. mellifera*, with less than 19 kb distance between the CNEs. The final nine CNEs were found on Scaffold 70. The first six of these were found in order, with less than 13 kb distance, the seventh had 64 kb distance to the other CNEs. The final two CNEs were found in the 64 kb space stated before with an additional change in orientation, making the distance between CNEs at this position only 47 kb long.

7. Discussion conserved non-coding elements

The focus of CNE research so far has been on vertebrate genomes (Polychronopoulos et al., 2017). There are quite a few species and lineages were vertebrate CNEs were identified, also between quite distantly related species such as human and puffer fish, where the last common ancestor occurred 430 years ago (Aparicio et al., 1995). Even after quite a long divergence time vertebrates still tend to have a high sequence similarity between species. This shows for example in the high alignment rate between human and puffer fish (12 % of the puffer fish genome can be aligned to the human genome). In insects, the focus lies on UCEs and is mostly centred on Drosophilids. Insect genomes are more divergent after the same time span than vertebrate genomes. Between different Hymenoptera genomes with a divergence time of 240 my (Misof et al., 2014), we were able to align 2-10 %. This alignment rate was enough to identify CNEs, as these are highly conserved regions that are of interest of us.

The biggest hurdle for the identification of CNEs is the availability of well sequenced and annotated genomes of species that are closely related. To be able to identify CNEs at least one well sequenced and annotated genome is necessary, depending on the method used to identify CNEs. This one species is used as a reference to identify conserved regions in other genomes, regardless whether whole genome alignments or a sliding window approach are used.

Using whole genome alignments for CNE identification requires more good quality genomes with good annotation as well as specialised seeding schemes for the species that are aligned. In our work we used the MAM8 seeding scheme, which is based on the substitution patterns in mammals (Frith and Noé, 2014). The WGAs we used might be improved by using an insect or arthropod specific seeding scheme, which does not currently exist.

So far, the focus on the gene that the CNE or the whole cluster is interacting with, lay on protein-coding genes. We could show that lncRNAs are also in distances and orientation to CNE clusters that could point towards an interaction between these two and an additional protein-coding gene. We calculated the lncRNA-protein-coding gene ratio for each species both for the whole annotation and the identified cluster partner. For *N. vitripennis* we noticed that an lncRNA was twice as highly likely to be neighbouring a cluster in *cis*-direction than would be expected if this was a random occurrence. Of course this number highly depends on the annotation of the genome and the assembly quality, as some studies assume that the number of lncRNAs vastly outnumbers the protein-coding genes (Quinn and Chang, 2016). Also, the total number of genes (including N/A) might be lower than expected due to the cluster number, because in some cases a cluster was found between a gene and another cluster. These two clusters then had the same gene identified as neighbouring.

Still we found that the majority of our gene-CNE cluster neighbours were not real interactions due to either no gene being found next to the cluster in a 500 kb distance or the gene having an orientation towards the cluster that is not *cis*. Because very little is known about the interaction between CNEs and their genes we made the assumption that their orientation to each other is important. If future research into this topic shows that their orientation is not important, our results regarding how many genes are identified as potential interaction partner for a CNE cluster could change considerably.

It has been shown that the protein-coding genes associated with CNEs are mostly involved in developmental regulation. This is also an area where lncRNAs have one of their functions. The problem with lncRNAs is that their general functions are known, as in what the whole class does, but only for a small number the function of a specific lncRNA is known, such as sphinx that regulates the male courtship behaviour (Legeai and Derrien, 2015). The combination of their high abundance and their presumed function makes them a point of interest regarding CNEs. So far a possible interaction between CNEs, a developmental protein-coding gene, and an lncRNA has not been studied. Because we only did computational analysis of CNEs we cannot say that the lncRNA neighbouring a CNE is really involved in a CNE-gene-interaction. But it presents an idea that should be further looked into, i.e., looking into genomes with better studied lncRNAs to see if this relationship also exists there and using experimental set ups.
We looked at the synteny between those CNE clusters that have an lncRNA as their possible interaction partner. We found out that there does seem to exist synteny of the clusters between different species, as in all cases we looked at this synteny was at least partially conserved. However, the clusters might not be identified in one species because the single CNEs inside have a larger distance to another than we defined as a cluster. Recombination maps would be an interesting further study to see how much recombination actually happens inside a CNE cluster. We did not look into what created the different distances between single CNEs. Another point is that the definition of a CNE cluster is somewhat arbitrary set to a maximum distance of 20 kb between to CNEs. An expansion of our cluster definition could probably show that our syntenic CNE hits are arranged in clusters in more than one genome, but only if no rearrangements of the genome happens. As we did not look into those clusters that had only protein-coding genes next to them, we cannot conclude an association between the lncRNAs and the synteny.
Some studies have shown that these conserved regions harbour transposable elements, although it is not clear yet whether the insertion of TEs is enhanced in these regions

(Manee et al., 2018). Inserted TEs could be responsible for the different distances. To show if TEs are indeed found between the different CNEs a comparison with a TE annotation of the genome is needed.

Our study showed that CNEs are still identifiable over an evolutionary distance of 240 my in insect lineages with a low similarity between their genomes. In vertebrates it has been shown that the CNEs that are conserved between distantly related lineages are different from those found in closer related groups (i.e. mammals), meaning that there is only a partial overlap between these CNE groups (Woolfe et al., 2004). It would be interesting to see if this also holds true for Hymenoptera. Also, it would be interesting to see how large the divergence time between species has to be before no CNEs are identifiable anymore. This also raises the question if there are CNEs that are conserved in all Metazoa. The first step would be to look how much of these highly differed genomes can still be aligned.

8. Conclusions

We were able to get a first look at the ncRNA repertoire in the two Hymenoptera species *A. rosae* and *O. abietinus*. Taking a closer look at their repertoire and comparing it with that of other Hymenoptera, we found that there seems to be a conserved set of ncRNAs for Hymenoptera. Most of the already known ncRNA families are not Hymenoptera specific. Through our additional *de novo* analyses we showed that the ncRNA repertoire is more extensive than the current state of knowledge suggests. Also we showed that using homology methods is not sufficient to identify the full repertoire of a species and is especially not qualified to find potentially lineage specific ncRNAs.
We showed that it is possible to identify CNEs between insect species that have a divergence time of 240 my. So far the only research on CNEs, or specifically UCEs, was done in Drosophilids, which have less variation in their sequences due to their shorter divergence time. Most of the CNE research so far has been focused on vertebrates, which possess more conserved areas of the genome even after a long divergence time. Our results showed that it is possible and necessary to broaden this research to other lineages. Only through additional research the origin of CNEs can be identified, as well as their level of conservation. It has been shown in vertebrates that the CNEs of mammals are only a subgroup of the CNEs identified between vertebrates (Woolfe et al., 2004). This shows that CNEs are still evolving and the variance between different groups should be the subject of future research.
Furthermore we found an interesting spatial relationship between CNEs and lncRNAs. lncRNAs were more often present in *cis*-direction next to a CNE cluster than would be

expected by chance. This conserved orientation and both being involved in regulatory processes could be a sign that they are a regulatory unit. As little is known how exactly CNEs regulate genes this needs to be addressed in further research.
Overall our research showed that both ncRNAs as well as CNEs are important parts of genomes that should not be neglected in genomic analyses.

Bibliography

Aparicio, S., Morrison, A., Gould, A., Gilthorpe, J., Chaudhuri, C., Rigby, P., Krumlauf, R., Brenner, S. Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, Fugu rubripes. *Proceedings of the National Academy of Sciences*, 92(5):1684–1688, 1995.

Aravin, A. A., Hannon, G. J., Brennecke, J. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science*, 318(5851):761–764, 2007.

Aravin, A. A., Lagos-Quintana, M., Yalcin, A., Zavolan, M., Marks, D., Snyder, B., Gaasterland, T., Meyer, J., Tuschl, T. The small RNA profile during Drosophila melanogaster development. *Developmental Cell*, 5(2):337–350, 2003.

Aravin, A. A., Naumova, N. M., Tulin, A. V., Vagin, V. V., Rozovsky, Y. M., Gvozdev, V. A. Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the D. melanogaster germline. *Current Biology*, 11(13):1017–1027, 2001.

Bachellerie, J.-P., Cavaillé, J., Hüttenhofer, A. The expanding snoRNA world. *Biochimie*, 84(8):775–790, 2002.

Baker, M. Long noncoding RNAs: the search for function. *Nature Methods*, 8:379, 2011.

Bantounas, I., Phylactou, L., Uney, J. RNA interference and the use of small interfering RNA to study gene function in mammalian systems. *Journal of Molecular Endocrinology*, 33(3):545–557, 2004.

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

Behura, S. K., Stanke, M., Desjardins, C. A., Werren, J. H., Severson, D. W. Comparative analysis of nuclear tRNA genes of Nasonia vitripennis and other arthropods, and relationships to codon usage bias. *Insect Molecular Biology*, 19(s1):49–58, 2010.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S., Haussler, D. Ultraconserved Elements in the Human Genome. *Science*, 304(5675):1321–1325, 2004.

Bolger, A. M., Lohse, M., Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

Brannan, C. I., Dees, E. C., Ingram, R. S., Tilghman, S. M. The product of the H19 gene may function as an RNA. *Molecular and Cellular Biology*, 10(1):28–36, 1990.

Bregliano, J., Picard, G., Bucheton, A., Pelisson, A., Lavige, J., l'Héritier, P. Hybrid dysgenesis in Drosophila melanogaster. *Science*, 207(4431):606–611, 1980.

Brennecke, J., Aravin, A. A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., Hannon, G. J. Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell*, 128(6):1089–1103, 2007.

Burke, W. D., Malik, H. S., Jones, J. P., Eickbush, T. H. The domain structure and retrotransposition mechanism of R2 elements are conserved throughout arthropods. *Molecular Biology and Evolution*, 16(4):502–511, 1999.

Burton, N. O., Burkhart, K. B., Kennedy, S. Nuclear RNAi maintains heritable gene silencing in Caenorhabditis elegans. *Proceedings of the National Academy of Sciences*, 108(49):19683–19688, 2011.

Bushnell, B. BBMap: a fast, accurate, splice-aware aligner. 2014.

Calin, G. A., Liu, C.-g., Ferracin, M., Hyslop, T., Spizzo, R., Sevignani, C., Fabbri, M., Cimmino, A., Lee, E. J., Wojcik, S. E., et al. Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. *Cancer Cell*, 12(3):215–229, 2007.

Cam, H. P., Sugiyama, T., Chen, E. S., Chen, X., FitzGerald, P. C., Grewal, S. I. Comprehensive analysis of heterochromatin-and RNAi-mediated epigenetic control of the fission yeast genome. *Nature Genetics*, 37(8):809, 2005.

Chambeyron, S., Seitz, H. Insect small non-coding RNA involved in epigenetic regulations. *Current Opinion in Insect Science*, 1:1–9, 2014.

Chen, J., Shishkin, A. A., Zhu, X., Kadri, S., Maza, I., Guttman, M., Hanna, J. H., Regev, A., Garber, M. Evolutionary analysis across mammals reveals distinct classes of long non-coding RNAs. *Genome Biology*, 17(1):19, 2016.

Ciganda, M., Williams, N. Eukaryotic 5S rRNA biogenesis. *Wiley Interdisciplinary Reviews: RNA*, 2(4):523–533, 2011.

Denli, A. M., Tops, B. B., Plasterk, R. H., Ketting, R. F., Hannon, G. J. Processing of primary microRNAs by the Microprocessor complex. *Nature*, 432(7014):231, 2004.

Desset, S., Meignin, C., Dastugue, B., Vaury, C. COM, a heterochromatic locus governing the control of independent endogenous retroviruses from Drosophila melanogaster. *Genetics*, 164(2):501–509, 2003.

Doench, J. G., Petersen, C. P., Sharp, P. A. siRNAs can function as miRNAs. *Genes & Development*, 17(4):438–442, 2003.

Dritsou, V., Deligianni, E., Dialynas, E., Allen, J., Poulakakis, N., Louis, C., Lawson, D., Topalis, P. Non-coding RNA gene families in the genomes of anopheline mosquitoes. *BMC Genomics*, 15(1):1038, 2014.

Dykxhoorn, D. M., Novina, C. D., Sharp, P. A. Killing the messenger: short RNAs that silence gene expression. *Nature Reviews Molecular Cell Biology*, 4(6):457, 2003.

Eliceiri, G. Small nucleolar RNAs. *Cellular and Molecular Life Sciences*, 56(1):22–31, 1999.

Elsik, C. G., Worley, K. C., Bennett, A. K., Beye, M., Camara, F., Childers, C. P., de Graaf, D. C., Debyser, G., Deng, J., Devreese, B., et al. Finding the missing honey bee genes: lessons learned from a genome upgrade. *BMC Genomics*, 15(1):86, 2014.

Faircloth, B. C. Identifying conserved genomic elements and designing universal bait sets to enrich them. *Methods in Ecology and Evolution*, 8(9):1103–1112, 2017.

Faircloth, B. C., Branstetter, M. G., White, N. D., Brady, S. G. Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera. *Molecular Ecology Resources*, 15(3):489–501, 2015.

Fang, S., Zhang, L., Guo, J., Niu, Y., Wu, Y., Li, H., Zhao, L., Li, X., Teng, X., Sun, X., et al. NONCODEV5: a comprehensive annotation database for long non-coding RNAs. *Nucleic Acids Research*, 46(D1):D308–D314, 2017.

Fasold, M., Langenberger, D., Binder, H., Stadler, P. F., Hoffmann, S. DARIO: a ncRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Research*, 39(suppl_2):W112–W117, 2011.

Fatica, A., Tollervey, D. Making ribosomes. *Current Opinion in Cell Biology*, 14(3):313–318, 2002.

Fire, A., Albertson, D., Harrison, S. W., Moerman, D. Production of antisense RNA leads to effective and specific inhibition of gene expression in C. elegans muscle. *Development*, 113(2):503–514, 1991.

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., Mello, C. C. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, 391(6669):806, 1998.

Frith, M. C., Noé, L. Improved search heuristics find 20 000 new alignments between human and mouse genomes. *Nucleic Acids Research*, 42(7):e59–e59, 2014.

Ghildiyal, M., Zamore, P. D. Small silencing RNAs: an expanding universe. *Nature Reviews Genetics*, 10(2):94, 2009.

Girard, A., Sachidanandam, R., Hannon, G. J., Carmell, M. A. A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature*, 442(7099):199, 2006.

Glazov, E. A., Pheasant, M., McGraw, E. A., Bejerano, G., Mattick, J. S. Ultraconserved elements in insect genomes: a highly conserved intronic sequence implicated in the control of homothorax mRNA splicing. *Genome Research*, 15(6):800–808, 2005.

Griffiths-Jones, S. RALEE—RNA ALignment editor in Emacs. *Bioinformatics*, 21(2):257–259, 2004.

Grimson, A., Srivastava, M., Fahey, B., Woodcroft, B. J., Chiang, H. R., King, N., Degnan, B. M., Rokhsar, D. S., Bartel, D. P. Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature*, 455(7217):1193–1197, 2008.

Gruber, A. R., Findeiß, S., Washietl, S., Hofacker, I. L., Stadler, P. F. RNAz 2.0: improved noncoding RNA detection. In *Biocomputing 2010*, pages 69–79. 2010.

Haller, F., von Heydebreck, A., Zhang, J. D., Gunawan, B., Langer, C., Ramadori, G., Wiemann, S., Sahin, Ö. Localization-and mutation-dependent microRNA (miRNA) expression signatures in gastrointestinal stromal tumours (GISTs), with a cluster of co-expressed miRNAs located at 14q32. 31. *The Journal of Pathology*, 220(1):71–86, 2010.

Hammond, S. M., Bernstein, E., Beach, D., Hannon, G. J. An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells. *Nature*, 404(6775):293, 2000.

Han, J., Lee, Y., Yeom, K.-H., Kim, Y.-K., Jin, H., Kim, V. N. The Drosha-DGCR8 complex in primary microRNA processing. *Genes & Development*, 18(24):3016–3027, 2004.

Hannon, G. J. RNA interference. *Nature*, 418(6894):244, 2002.

Hardison, R. C. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends in Genetics*, 16(9):369–372, 2000.

Henras, A. K., Plisson-Chastang, C., O'Donohue, M.-F., Chakraborty, A., Gleizes, P.-E. An overview of pre-ribosomal RNA processing in eukaryotes. *Wiley Interdisciplinary Reviews: RNA*, 6(2):225–242, 2015.

Hoeppner, M. P., Denisenko, E., Gardner, P. P., Schmeier, S., Poole, A. M. An Evaluation of Function of Multicopy Noncoding RNAs in Mammals Using ENCODE/FANTOM Data and Comparative Genomics. *Molecular Biology and Evolution*, 35(6):1451–1462, 2018.

Hoffmann, S., Otto, C., Kurtz, S., Sharma, C. M., Khaitovich, P., Vogel, J., Stadler, P. F., Hackermüller, J. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLOS Computational Biology*, 5(9):e1000502, 2009.

Hua-Van, A., Le Rouzic, A., Boutin, T. S., Filée, J., Capy, P. The struggle for life of the genome's selfish architects. *Biology direct*, 6(1):19, 2011.

Huang, X., Tóth, K. F., Aravin, A. A. piRNA Biogenesis in Drosophila melanogaster. *Trends in Genetics*, 33(11):882–894, 2017. ISSN 0168-9525. Transposable Elements.

Jakubczak, J. L., Burke, W. D., Eickbush, T. H. Retrotransposable elements R1 and R2 interrupt the rRNA genes of most insects. *Proceedings of the National Academy of Sciences*, 88(8):3295–3299, 1991.

Jandura, A., Krause, H. M. The New RNA World: Growing Evidence for Long Noncoding RNA Functionality. *Trends in Genetics*, 33(10):665–676, 2017.

Jones-Rhoades, M. W., Bartel, D. P., Bartel, B. MicroRNAs and their regulatory roles in plants. *Annual Review of Plant Biology* , 57:19–53, 2006.

Jühling, F., Mörl, M., Hartmann, R. K., Sprinzl, M., Stadler, P. F., Pütz, J. tRNAdb 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Research*, 37(suppl_1):D159–D162, 2008.

Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E. P., Rivas, E., Eddy, S. R., Bateman, A., Finn, R. D., Petrov, A. I. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Research*, 46(D1):D335–D342, 2018.

Kiełbasa, S. M., Wan, R., Sato, K., Horton, P., Frith, M. C. Adaptive seeds tame genomic sequence comparison. *Genome Research*, 21(3):487–493, 2011.

Kim, H. S., Murphy, T., Xia, J., Caragea, D., Park, Y., Beeman, R. W., Lorenzen, M. D., Butcher, S., Manak, J. R., Brown, S. J. BeetleBase in 2010: revisions to provide comprehensive genomic information for Tribolium castaneum. *Nucleic Acids Research*, 38(suppl_1):D437–D442, 2009a.

Kim, S. Y., Pritchard, J. K. Adaptive evolution of conserved noncoding elements in mammals. *PLOS Genetic*, 3(9):e147, 2007.

Kim, V. N., Han, J., Siomi, M. C. Biogenesis of small RNAs in animals. *Nature Reviews Molecular Cell Biology*, 10(2):126, 2009b.

Kim, V. N., Nam, J.-W. Genomics of microRNA. *TRENDS in Genetics*, 22(3):165–173, 2006.

Kozomara, A., Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research*, 42(D1):D68–D73, 2013.

Ku, H.-Y., Lin, H. PIWI proteins and their interactors in piRNA biogenesis, germline development and gene expression. *National Science Review*, 1(2):205–218, 2014.

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

Lau, N. C., Robine, N., Martin, R., Chung, W.-J., Niki, Y., Berezikov, E., Lai, E. C. Abundant primary piRNAs, endo-siRNAs, and microRNAs in a Drosophila ovary cell line. *Genome Research*, 19(10):1776–1785, 2009.

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

Lee, Y., Jeon, K., Lee, J.-T., Kim, S., Kim, V. N. MicroRNA maturation: stepwise processing and subcellular localization. *The EMBO journal*, 21(17):4663–4670, 2002.

Lee, Y., Kim, M., Han, J., Yeom, K.-H., Lee, S., Baek, S. H., Kim, V. N. MicroRNA genes are transcribed by RNA polymerase II. *The EMBO journal*, 23(20):4051–4060, 2004.

Legeai, F., Derrien, T. Identification of long non-coding RNAs in insects genomes. *Current Opinion in Insect Science*, 7:37–44, 2015.

Leis, J. P., Keller, E. B. Protein chain-initiating methionine tRNAs in chloroplasts and cytoplasm of wheat leaves. *Proceedings of the National Academy of Sciences*, 67(3):1593–1599, 1970.

Lestrade, L., Weber, M. J. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Research*, 34(suppl_1):D158–D162, 2006.

Lewis, M. A., Quint, E., Glazier, A. M., Fuchs, H., De Angelis, M. H., Langford, C., Van Dongen, S., Abreu-Goodger, C., Piipari, M., Redshaw, N., et al. An ENU-induced mutation of miR-96 associated with progressive hearing loss in mice. *Nature Genetics*, 41(5):614, 2009.

Lorenz, R., Bernhart, S. H., Zu Siederdissen, C. H., Tafer, H., Flamm, C., Stadler, P. F., Hofacker, I. L. ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(1):26, 2011.

Lowe, T. M., Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, 25(5):955–964, 1997.

Ludwig, N., Becker, M., Schumann, T., Speer, T., Fehlmann, T., Keller, A., Meese, E. Bias in recent miRBase annotations potentially associated with RNA quality issues. *Scientific Reports*, 7(1):5162, 2017.

Manee, M. M., Jackson, J., Bergman, C. M. Conserved noncoding elements influence the transposable element landscape in Drosophila. *Genome Biology and Evolution*, page evy104, 2018.

Marco, A., Hooks, K., Griffiths-Jones, S. Evolution and function of the extended miR-2 microRNA family. *RNA Biology*, 9(3):242–248, 2012.

Marcovitz, A., Jia, R., Bejerano, G. "reverse genomics" predicts function of human conserved noncoding elements. *Molecular Biology and Evolution*, 33(5):1358–1369, 2016.

Maxwell, E., Fournier, M. The small nucleolar RNAs. *Annual Review of Biochemistry*, 64(1):897–934, 1995.

Misof, B., Liu, S., Meusemann, K., Peters, R. S., Donath, A., Mayer, C., Frandsen, P. B., Ware, J., Flouri, T., Beutel, R. G., et al. Phylogenomics resolves the timing and pattern of insect evolution. *Science*, 346(6210):763–767, 2014.

Muerdter, F., Olovnikov, I., Molaro, A., Rozhkov, N. V., Czech, B., Gordon, A., Hannon, G. J., Aravin, A. A. Production of artificial piRNAs in flies and mice. *RNA*, 18(1):42–52, 2012.

Nawrocki, E. P., Burge, S. W., Bateman, A., Daub, J., Eberhardt, R. Y., Eddy, S. R., Floden, E. W., Gardner, P. P., Jones, T. A., Tate, J., et al. Rfam 12.0: updates to the RNA families database. *Nucleic Acids Research*, 43(D1):D130–D137, 2014.

Nawrocki, E. P., Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29(22):2933–2935, 2013.

Nobrega, M. A., Ovcharenko, I., Afzal, V., Rubin, E. M. Scanning human gene deserts for long-range enhancers. *Science*, 302(5644):413–413, 2003.

Ohno, S. An argument for the genetic simplicity of man and other mammals. *Journal of Human Evolution*, 1(6):651–662, 1972.

Ojala, D., Montoya, J., Attardi, G. tRNA punctuation model of RNA processing in human mitochondria. *Nature*, 290(5806):470, 1981.

Oliveros, J. C. VENNY. An interactive tool for comparing lists with Venn Diagrams. 2007. 2015.

Olovnikov, I., Ryazansky, S., Shpiz, S., Lavrov, S., Abramov, Y., Vaury, C., Jensen, S., Kalmykova, A. De novo piRNA cluster formation in the Drosophila germ line triggered by transgenes containing a transcribed transposon fragment. *Nucleic Acids Research*, 41(11):5757–5768, 2013.

Olsen, P. H., Ambros, V. The lin-4 regulatory RNA controls developmental timing in Caenorhabditis elegans by blocking LIN-14 protein synthesis after the initiation of translation. *Developmental Biology*, 216(2):671–680, 1999.

Paten, B., Earl, D., Nguyen, N., Diekhans, M., Zerbino, D., Haussler, D. Cactus: Algorithms for genome multiple sequence alignment. *Genome Research*, 21(9):1512–1528, 2011.

Phizicky, E. M., Hopper, A. K. tRNA biology charges to the front. *Genes & Development*, 24(17):1832–1860, 2010.

Polychronopoulos, D., King, J. W., Nash, A. J., Tan, G., Lenhard, B. Conserved non-coding elements: developmental gene regulation meets genome organization. *Nucleic Acids Research*, 45(22):12611–12624, 2017.

Prud'Homme, N., Gans, M., Masson, M., Terzian, C., Bucheton, A. Flamenco, a gene controlling the gypsy retrovirus of Drosophila melanogaster. *Genetics*, 139(2):697–711, 1995.

Quek, X. C., Thomson, D. W., Maag, J. L., Bartonicek, N., Signal, B., Clark, M. B., Gloss, B. S., Dinger, M. E. lncRNAdb v2. 0: expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Research*, 43(D1):D168–D173, 2014.

Quinn, J. J., Chang, H. Y. Unique features of long non-coding RNA biogenesis and function. *Nature Reviews Genetics*, 17(1):47, 2016.

Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., Brugmann, S. A., Goodnough, L. H., Helms, J. A., Farnham, P. J., Segal, E., et al. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell*, 129(7):1311–1323, 2007.

Ruby, J. G., Stark, A., Johnston, W. K., Kellis, M., Bartel, D. P., Lai, E. C. Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. *Genome Research*, 17(12):1850–1864, 2007.

Senti, K.-A., Brennecke, J. The piRNA pathway: a fly's perspective on the guardian of the genome. *Trends in Genetics*, 26(12):499–509, 2010.

Slotkin, R. K., Martienssen, R. Transposable elements and the epigenetic regulation of the genome. *Nature Reviews Genetics*, 8(4):272, 2007.

Smith, C. M., Steitz, J. A. Sno storm in the nucleolus: new roles for myriad small RNPs. *Cell*, 89(5):669–672, 1997.

Srivastava, A., Schlessinger, D. Structure and organization of ribosomal DNA. *Biochimie*, 73(6):631–638, 1991.

Stallman, R. EMACS: The Extensible, Customizable Display Editor. 519a. 1981.

Tan, G. *CNEr: CNE Detection and Visualization*, 2015. R package version 1.6.2.

Tarver, J. E., Sperling, E. A., Nailor, A., Heimberg, A. M., Robinson, J. M., King, B. L., Pisani, D., Donoghue, P. C. J., Peterson, K. J. miRNAs: small genes with big potential in metazoan phylogenetics. *Molecular Biology and Evolution*, 30(11):2369–82, 2013.

Tarver, J. E., Taylor, R. S., Puttick, M. N., Lloyd, G. T., Pett, W., Fromm, B., Schirrmeister, B. E., Pisani, D., Peterson, K. J., Donoghue, P. C. Well-annotated microRNAomes do not evidence pervasive miRNA loss. *Genome Biology and Evolution*, 10(6):1457–1470, 2018.

Thibaud-Nissen, F., Souvorov, A., Murphy, T., DiCuccio, M., Kitts, P. *The NCBI Handbook [Internet]*, chapter Eukaryotic Genome Annotation Pipeline. Bethesda (MD): National Center for Biotechnology Information (US), 2013. Available from: https://www.ncbi.nlm.nih.gov/books/NBK169439/.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., Salzberg, S. L., Wold, B. J., Pachter, L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511, 2010.

Vilardo, E., Nachbagauer, C., Buzet, A., Taschner, A., Holzmann, J., Rossmanith, W. A subcomplex of human mitochondrial RNase P is a bifunctional methyltransferase—extensive moonlighting in mitochondrial tRNA biogenesis. *Nucleic Acids Research*, 40(22):11583–11593, 2012.

Volders, P.-J., Verheggen, K., Menschaert, G., Vandepoele, K., Martens, L., Vandesompele, J., Mestdagh, P. An update on LNCipedia: a database for annotated human lncRNA sequences. *Nucleic Acids Research*, 43(D1):D174–D180, 2014.

Wang, Y., Jiang, Feng, Wang, Huimin, Song, Tianqi, Wei, Yuanyuan, Yang, Meiling, Zhang, Jianzhen, Kang, Le. Evidence for the expression of abundant microRNAs in the locust genome. *Scientific Reports*, 5:13608, 2015.

Wapinski, O., Chang, H. Y. Long noncoding RNAs and human disease. *Trends in Cell Biology*, 21(6):354–361, 2011.

Warnefors, M., Hartmann, B., Thomsen, S., Alonso, C. R. Combinatorial gene regulatory functions underlie ultraconserved elements (UCEs) in Drosophila. *Molecular Biology and Evolution*, page msw101, 2016.

Warren, W. C., Hillier, L. W., Graves, J. A. M., Birney, E., Ponting, C. P., Grützner, F., Belov, K., Miller, W., Clarke, L., Chinwalla, A. T., et al. Genome analysis of the platypus reveals unique signatures of evolution. *Nature*, 453(7192):175, 2008.

Wei, Y., Chen, S., Yang, P., Ma, Z., Kang, L. Characterization and comparative profiling of the small RNA transcriptomes in two phases of locust. *Genome Biology*, 10(1):R6, 2009.

Werren, J. H., Richards, S., Desjardins, C. A., Niehuis, O., Gadau, J., Colbourne, J. K., Group, N. G. W., et al. Functional and evolutionary insights from the genomes of three parasitoid Nasonia species. *Science*, 327(5963):343–348, 2010.

Woolfe, A., Goodson, M., Goode, D. K., Snell, P., McEwen, G. K., Vavouri, T., Smith, S. F., North, P., Callaway, H., Kelly, K., et al. Highly conserved non-coding sequences are associated with vertebrate development. *PLOS Biology*, 3(1):e7, 2004.

Wu, M., Jolicoeur, N., Li, Z., Zhang, L., Fortin, Y., L'abbe, D., Yu, Z., Shen, S.-H. Genetic variations of microRNAs in human cancer and their effects on the expression of miRNAs. *Carcinogenesis*, 29(9):1710–1716, 2008.

Wucher, V., Legeai, F., Hédan, B., Rizk, G., Lagoutte, L., Leeb, T., Jagannathan, V., Cadieu, E., David, A., Lohi, H., Cirera, S., Fredholm, M., Botherel, N., Leegwater, P. A., Le Béguec, C., Fieten, H., Johnson, J., Alföldi, J., André, C., Lindblad-Toh, K., Hitte, C., Derrien, T. FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Research*, 45(8):e57, 2017.

Yekta, S., Tabin, C. J., Bartel, D. P. MicroRNAs in the Hox network: an apparent link to posterior prevalence. *Nature Reviews Genetics*, 9(10):789, 2008.

Ylla, G., Fromm, B., Piulachs, M.-D., Belles, X. The microRNA toolkit of insects. *Scientific Reports*, 6:37736, 2016.

Yue, J.-X., Kozmikova, I., Ono, H., Nossa, C. W., Kozmik, Z., Putnam, N. H., Yu, J.-K., Holland, L. Z. Conserved noncoding elements in the most distant genera of cephalochordates: the Goldilocks principle. *Genome Biology and Evolution*, 8(8):2387–2405, 2016.

Zamore, P. D., Haley, B. Ribo-gnome: the big world of small RNAs. *Science*, 309(5740):1519–1524, 2005.

Zanni, V., Eymery, A., Coiffet, M., Zytnicki, M., Luyten, I., Quesneville, H., Vaury, C., Jensen, S. Distribution, evolution, and diversity of retrotransposons at the flamenco locus reflect the regulatory properties of piRNA clusters. *Proceedings of the National Academy of Sciences*, 110(49):19842–19847, 2013.

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C. G., et al. Ensembl 2018. *Nucleic Acids Research*, 46(D1):D754–D761, 2017.

A. Appendix

A.1. Prediction of non-coding RNAs

The tables A.1 and A.2 show the coordinates of the ncRNAs that were predicted through the homology analysis (both miRBase and Rfam), as well as the tRNAscan-SE predictions.

Table A.1.: Coordinates of all regulatory elements and ncRNAs that were predicted in *Athalia rosae*, after manual curation.

Name	Scaffold	Start	End	Strand
tRNA_Pro-1	1	1433526	1433597	+
U12	1	1511977	1512125	-
K_chan_RES-1	1	2833066	2833179	+
5S_rRNA-1	2	170446	170564	+
5S_rRNA-2	2	170716	170854	+
K_chan_RES-2	2	1497229	1497347	+
tRNA_Val-1	2	1642903	1642975	+
tRNA_Leu-1	2	1668913	1668996	+
tRNA_Thr-1	3	1264092	1264163	+
tRNA_Thr-2	3	1452256	1452329	-
tRNA_Ser-1	3	1453561	1453642	+
tRNA_Phe-1	3	1454541	1454613	+
tRNA_Ser-2	3	1455794	1455875	+
Aro-mir-375	3	2588571	2588662	-
tRNA_Cys-1	3	2658276	2658347	+
tRNA_Pro-2	3	2712853	2712924	-
tRNA_Ser-3	4	1002965	1003046	+
tRNA_Lys-1	4	1003374	1003446	+
tRNA_Asn-1	4	1003743	1003816	-
U5-1	4	1004293	1004414	-
U5-2	4	1005341	1005462	+

Continued on next page

Table A.1.: Continued from previous page.

Name	Scaffold	Start	End	Strand
U5-3	4	1018005	1018128	-
U5-4	4	1029647	1029771	+
tRNA_Pro-3	4	1169861	1169932	+
Aro-mir-219	4	1751471	1751573	-
tRNA_Gln-1	4	2449617	2449688	+
tRNA_Pro-4	5	1046515	1046586	-
Aro-mir-929	5	1202068	1202165	-
U6atac	5	1429993	1430097	-
tRNA_Ser-4	5	1472246	1472327	+
U1-1	5	2486577	2486738	+
tRNA_Ala-1	5	2800605	2800677	-
tRNA_Ala-2	5	2800767	2800839	-
tRNA_Ala-3	5	2800920	2800992	-
tRNA_Ala-4	5	2801081	2801153	-
tRNA_Arg-1	5	3445625	3445697	-
tRNA_Met-1	6	115917	115988	-
Aro-mir-263b	6	351357	351447	-
U2-1	6	1028869	1029062	-
U2-2	6	1040818	1041011	-
tRNA_His-1	6	1286444	1286515	+
U1-2	6	1306590	1306752	-
Aro-mir-279d	6	1386159	1386253	+
Aro-mir-11	6	1559210	1559291	-
Aro-mir-10a	6	2023801	2023896	-
Aro-let-7	6	2024110	2024209	-
Aro-mir-10b	6	2024618	2024691	-
tRNA_Phe-2	7	685484	685556	-
tRNA_Phe-3	7	685645	685717	-
tRNA_Tyr-1	7	686538	686634	-
tRNA_Tyr-1_Intron	7	686574	686597	-
tRNA_Asp-1	7	687397	687468	+
tRNA_Asp-2	7	687545	687616	+
tRNA_Asp-3	7	687694	687765	+
tRNA_Asp-4	7	687843	687914	+
tRNA_Met-2	7	792101	792172	-
tRNA_Gly-1	7	812607	812678	-
tRNA_Leu-2	7	835273	835352	+
tRNA_Gln-2	7	835519	835590	-
tRNA_Ala-5	7	835799	835870	+
tRNA_Leu-3	7	839988	840071	+
tRNA_Gln-3	7	1096854	1096925	+
tRNA_Leu-4	7	1229422	1229503	+

Continued on next page

Table A.1.: Continued from previous page.

Name	Scaffold	Start	End	Strand
tRNA_Arg-2	7	1323188	1323260	-
Aro-mir-9b	8	911691	911780	-
Aro-mir-9d	8	911824	911916	-
tRNA_Met-3	8	1112111	1112182	+
tRNA_Lys-2	8	1112404	1112476	+
tRNA_Ile-1	8	1112655	1112728	+
tRNA_Cys-2	8	1162633	1162704	+
tRNA_Met-4	8	1276723	1276795	+
U6-1	8	1287237	1287343	+
U6-2	8	1289329	1289435	-
tRNA_Pro-5	8	1356938	1357009	-
U6-3	8	1357349	1357455	-
tRNA_Glu-1	9	332701	332772	+
K_chan_RES-3	9	1630457	1630570	-
K_chan_RES-4	9	1641017	1641130	-
tRNA_Glu-2	9	1853331	1853402	+
tRNA_Gly-2	9	1853623	1853694	+
tRNA_Gly-3	9	1859166	1859236	-
tRNA_Gly-4	9	1859560	1859631	-
tRNA_Arg-3	10	1070988	1071060	+
SNORD31-1	10	1319758	1319829	-
SNORD31-2	10	1320060	1320128	-
SNORD31-3	10	1320322	1320389	-
SNORD31-4	10	1320581	1320646	-
Aro-mir-980	10	1903681	1903768	-
Rnase_MRP	10	3312516	3312779	-
Aro-mir-927	11	870330	870416	+
Aro-mir-iab-8	12	417229	417321	-
Aro-mir-iab-4	12	417234	417315	+
tRNA_Tyr-2	12	993592	993693	+
tRNA_Tyr-2_Intron	12	993629	993657	+
tRNA_Leu-5	13	343636	343718	+
tRNY_Lys-3	13	1211483	1211555	+
SNORD18-1	13	1264137	1264205	+
SNORD18-2	13	1264667	1264736	+
Aro-mir-928	13	1862522	1862621	-
Aro-mir-25a	13	1914138	1914241	+
Aro-mir-25c	13	1914415	1914511	+
tRNA_Met-5	13	2516187	2516259	-
tRNA_Arg-4	13	2516769	2516841	+
tRNA_His-2	13	2518416	2518487	-
U2-3	15	52209	52392	-

Continued on next page

Table A.1.: Continued from previous page.

Name	Scaffold	Start	End	Strand
tRNA_Arg-5	15	238215	238287	-
tRNA_Arg-6	15	331485	331557	+
tRNA_Tyr-3	15	338807	338907	-
tRNA_Tyr-3_Intron	15	338807	338907	-
5_8S_rRNA-1	16	90960	91114	-
18S_rRNA-1	16	99366	101200	+
Aro-mir-263a	16	230855	230947	-
SNORD33	17	891328	891407	+
tRNA_Trp-1	17	1263752	1263823	+
tRNA_Val-2	18	591985	592057	-
tRNA_Lys-4	18	853539	853611	-
tRNA_Trp-2	18	858202	858273	+
Aro-mir-2765	18	938207	938312	+
tRNA_Tyr-4	18	1608505	1608597	+
tRNA_Tyr-4_Intron	18	1608542	1608561	+
Aro-mir-282	19	430551	430641	-
tRNA_Arg-7	19	1214761	1214833	-
tRNA_Leu-13	20	841054	841173	+
tRNA_Leu-13_Intron	20	841092	841128	+
tRNA_Glu-3	20	841336	841407	+
tRNA_Glu-4	20	841612	841683	+
tRNA_Leu-14	20	957645	957763	-
tRNA_Leu-14_Intron	20	957690	957725	-
tRNA_Glu-5	20	958495	958566	+
tRNA_Glu-6	20	958771	958842	+
U11	21	488637	488770	+
Histone3-1	21	2279129	2279173	-
Histone3-2	21	2282281	2282326	+
U1-3	22	169986	170149	-
tRNA_His-3	22	550439	550510	+
tRNA_Thr-3	22	689649	689722	-
tRNA_Asn-2	22	690035	690108	-
tRNA_Leu-6	22	2178946	2179028	-
tRNA_Pro-6	22	2539025	2539096	+
U1-4	22	2713663	2713813	+
Aro-mir-71	23	253314	253411	+
Aro-mir-2c	23	253775	253855	+
Aro-mir-2d	23	253985	254062	+
Aro-mir-2e	23	254331	254416	+
Aro-mir-2a	23	254603	254688	+
Aro-mir-2b	23	254818	254901	+
tRNA_Cys-3	26	159479	159550	+

Continued on next page

Table A.1.: Continued from previous page.

Name	Scaffold	Start	End	Strand
Aro-mir-34	26	306065	306158	-
Aro-mir-277	26	307949	308033	-
Aro-mir-317	26	316144	316236	-
Aro-mir-965	26	1420301	1420392	+
tRNA_Gln-4	28	848964	849035	+
tRNA_Arg-8	28	865411	865483	+
tRNA_Asp-5	28	876157	876228	+
Snopsi18S-1377	28	877006	877134	+
tRNA_Lys-5	29	150731	150803	+
tRNA_Ala-6	29	2462845	2462916	-
tRNY_Lys-6	29	2465497	2465569	+
tRNA_Gly-5	29	2466564	2466634	+
tRNA_His-4	29	2466927	2466998	+
Aro-mir-279a	30	1474727	1474818	+
Aro-mir-279b	30	1475040	1475124	+
Aro-mir-124	30	1715134	1715231	-
tRNA_Ala-7	30	1818767	1818838	-
tRNA_Gln-5	30	2247032	2247103	+
Aro-mir-971	32	893457	893559	+
tRNA_Gly-6	32	1760401	1760471	-
tRNA_Asp-6	32	1760664	1760735	-
Histone3-3	32	1815208	1815253	-
Histone3-4	32	1816266	1816311	+
Aro-mir-2796	32	2502286	2502385	+
tRNA_Leu-7	32	2724703	2724782	+
Aro-mir-7	34	318810	318903	-
Aro-mir-25b	34	376919	377015	-
tRNA_Pro-7	34	625361	625432	-
tRNA_Ser-5	34	836463	836544	-
tRNA_Gln-6	36	148833	148904	+
tRNA_Met-6	36	155596	155667	-
tRNA_Ser-6	36	230624	230705	+
tRNA_Val-3	36	309911	309983	-
tRNA_Val-4	36	310290	310362	+
tRNA_Val-5	36	310455	310527	+
Aro-mir-193	36	403355	403444	+
Aro-mir-2788	36	407954	408051	+
tRNA_Val-6	36	951495	951567	+
SNORA57	37	183510	183647	-
tRNA_Thr-4	37	905020	905092	+
tRNA_Leu-8	37	973628	973709	+
K_chan_RES-5	37	1155836	1155946	+

Continued on next page

Table A.1.: Continued from previous page.

Name	Scaffold	Start	End	Strand
Aro-mir-10c	38	818030	818101	+
Aro-mir-10d	38	868885	868975	-
5S_rRNA-3	39	122799	122917	-
5S_rRNA-4	39	132597	132712	-
5S_rRNA-5	39	142963	143086	-
5S_rRNA-6	39	148828	148942	-
Metazoa_SRP-1	40	355109	355407	+
tRNA_Ala-8	40	896827	896898	-
tRNA_Ala-9	40	1208833	1208904	+
Aro-mir-252	40	1209240	1209344	-
tRNA_Ile-2	40	1281043	1281116	+
ACEA_U3-1	41	280994	281209	-
tRNA_Ala-10	41	282552	282624	-
tRNA_Ala-11	41	297290	297362	-
tRNA_Arg-9	41	400121	400193	-
tRNA_Ile-3	41	418591	418664	-
tRNA_Asp-7	41	422237	422308	-
tRNA_Thr-5	41	426535	426608	-
tRNA_Arg-10	41	428924	428996	+
tRNA_Ile-4	41	429270	429343	-
tRNA_Ser-7	41	434848	434929	+
RnaseP_nuc-1	41	1455640	1455982	-
Aro-mir-216	44	387971	388051	+
Aro-mir-3477	44	389196	389294	+
Aro-mir-12	44	389626	389696	+
tRNA_Gly-7	44	459497	459568	+
Aro-mir-279c	44	481830	481921	-
Aro-mir-210	49	280120	280214	-
tRNA_Pro-8	51	177557	177628	+
tRNA_Pro-9	51	177909	177980	+
tRNA_Ala-12	51	181173	181245	+
tRNA_Thr-6	51	181414	181487	+
tRNA_Thr-7	51	255484	255556	-
Aro-bantam	52	488311	488391	+
Aro-mir-932	53	918485	918593	+
tRNA_Arg-11	54	100659	100731	-
U1-5	54	316054	316215	-
U1-6	54	316986	317147	-
U1-7	54	322228	322389	-
U1-8	54	325637	325798	-
Aro-mir-1	56	27718	27810	-
tRNA_Thr-8	56	359070	359141	+

Continued on next page

Table A.1.: Continued from previous page.

Name	Scaffold	Start	End	Strand
Aro-mir-31	58	321391	321478	-
ACEA_U3-2	59	780929	781142	-
tRNA_Ser-8	59	781333	781414	+
Metazoa_SRP-2	59	922320	922616	-
tRNA_Thr-9	61	339996	340067	-
Aro-mir-3478	61	530879	530958	-
Aro-mir-279	61	531015	531110	+
Aro-mir-9c	61	531156	531246	+
Aro-mir-2944	61	531427	531529	+
Aro-mir-1000	61	786936	787028	-
SNORA79	62	319487	319622	+
SnosnR60_Z15-1	63	826108	826197	-
SnosnR60_Z15-2	63	826417	826506	-
SnosnR60_Z15-3	63	826729	826817	-
SnosnR60_Z15-4	63	827120	827209	-
SnosnR60_Z15-5	63	827491	827579	-
5_8S_rRNA-2	66	7307	7461	-
U2-4	68	86044	86217	+
tRNA_Ile-5	71	277794	277867	+
tRNA_Thr-10	72	508757	508828	-
SNORD49	72	525256	525330	+
5S_rRNA-7	74	116702	116820	-
Histone3-5	74	1120520	1120566	-
Histone3-6	74	1121662	1121706	+
Histone3-7	74	1128458	1128503	-
Histone3-8	74	1138828	1138873	+
tRNA_Asn-3	74	1370688	1370761	-
Aro-mir-305	75	320599	320683	-
Aro-mir-275	75	320791	320879	-
Histone3-9	75	889397	889442	-
Histone3-10	75	892588	892633	-
Histone3-11	75	893721	893767	+
Histone3-12	75	894603	894648	-
Histone3-13	75	895869	895913	+
Histone3-14	75	1153597	1153643	-
Histone3-15	75	1154859	1154905	+
Histone3-16	75	1156903	1156947	+
U1-9	75	1319180	1319322	-
tRNA_Gln-7	77	484106	484177	+
tRNA_Pro-10	77	484344	484415	-
tRNA_Gln-8	77	484638	484709	+
tRNA_Leu-9	77	687759	687840	-

Continued on next page

Table A.1.: Continued from previous page.

Name	Scaffold	Start	End	Strand
Aro-mir-67a	79	1699	1800	-
Aro-mir-278	79	103718	103814	+
Aro-mir-1923	79	695478	695565	+
SCARNA8	79	872022	872152	+
tRNA_Lys-7	79	1069400	1069472	-
U2-5	80	317345	317537	-
Aro-mir-6012	80	733636	733773	-
Aro-mir-133	81	45091	45189	+
U4-1	84	427265	427405	+
tRNA_Leu-10	87	722654	722735	-
U4-2	87	733168	733308	+
snoU43	87	825174	825249	-
Aro-mir-184	88	116244	116340	-
tRNA_Asn-4	88	310198	310271	+
Aro-mir-276	91	186540	186634	+
tRNA_Tyr-5	92	980509	980602	+
tRNA_Tyr-5_Intron	92	980546	980566	+
Aro-mir-750	100	475828	475927	+
Aro-mir-1175	100	476053	476152	+
Aro-mir-137	102	321092	321190	-
tRNA_Glu-7	108	125008	125079	-
tRNA_Leu-11	108	125293	125375	+
tRNA_His-5	109	669378	669449	+
tRNA_Ser-9	111	691367	691448	+
U6-4	112	69576	69681	+
Aro-mir-981	112	487634	487724	-
18S_rRNA-2	115	102146	103660	+
Aro-mir-315	117	119236	119322	-
Trna_Trp-3	119	95686	95757	+
SNORD57-1	119	529202	529271	+
SNORD57-2	119	529479	529550	+
Sphinx_1	119	573152	573253	+
Sphinx_2	119	573668	573831	+
Aro-mir-33	121	164366	164452	+
tRNA_Glu-8	123	348736	348807	-
Aro-mir-46	123	400982	401074	-
tRNA_Arg-12	123	408431	408503	+
28S_rRNA-1_partial_3prime	125	83375	84067	-
R2	125	84068	91353	-
18S_rRNA-3	125	87025	88957	-
Aro-mir-6497	125	91556	91730	-
28S_rRNA-2_partial_5prime	125	91354	94132	-

Continued on next page

Table A.1.: Continued from previous page.

Name	Scaffold	Start	End	Strand
5_8S_rRNA-3	125	94719	94873	-
18S_rRNA-3	125	96001	97914	-
tRNA_Gly-8	133	149604	149674	+
tRNA_Gly-9	133	149790	149860	+
tRNA_Gly-10	133	149979	150049	+
SNORA16	141	15821	15954	+
tRNA_Gly-11	143	144508	144578	+
tRNA_Val-7	145	176543	176615	-
tRNA_Gly-12	145	190361	190432	-
tRNA_Asn-5	145	319523	319596	-
tRNA_Ile-7	145	431981	432072	+
tRNA_Ile-7_Intron	145	432019	432036	+
tRNA_Phe-4	145	533079	533151	+
5_8S_rRNA-4	149	11898	12052	+
tRNA_Met-7	150	164643	164714	-
tRNA_Met-8	150	164816	164887	-
5S_rRNA-8	152	108889	109007	-
5S_rRNA-9	159	1	85	-
tRNA_Pro-11	161	17575	17646	+
Aro-mir-9a	161	134454	134546	+
tRNA_Pro-12	161	250689	250760	+
tRNA_Ile-6	161	390009	390082	+
tRNA_Ala-13	166	232945	233017	-
tRNA_Ser-10	166	240175	240256	-
tRNA_Ile-8	167	59899	59990	-
tRNA_Ile-8_Intron	167	59935	59952	-
tRNA_Val-8	172	312695	312767	+
tRNA_Val-9	172	317269	317341	-
tRNA_Lys-8	173	349230	349302	-
tRNA_Lys-9	173	353283	353355	+
tRNA_Ser-11	173	359445	359526	+
tRNA_Arg-13	176	67431	67503	+
tRNA_Met-9	176	81150	81222	-
tRNA_Glu-9	176	100004	100075	+
tRNA_Glu-10	176	100156	100227	+
tRNA_Leu-12	176	106577	106660	-
tRNA_Ser-12	176	172584	172665	-
Aro-mir-307b	176	179939	180040	-
Aro-mir-190	178	123855	123950	-
Aro-mir-8	179	177532	177623	-
U4atac	186	147167	147302	-
Arthropod_7SK	190	86563	86838	+

Continued on next page

Table A.1.: Continued from previous page.

Name	Scaffold	Start	End	Strand
RnaseP_nuc-2	192	37270	37576	-
SNORD36-1	194	120466	120535	+
SNORD36-2	194	121505	121575	+
tRNA_Ala-14	199	196160	196231	-
tRNA_Arg-14	199	227392	227464	+
Aro-mir-316	212	58268	58352	+
tRNA_Lys-10	212	123676	123748	-
tRNA_Ala-15	236	487211	487283	+
tRNA_Gly-13	238	6383	6455	+
tRNA_Thr-11	238	6472	6543	+
5S_rRNA-10	363	5783	5901	-
5S_rRNA-11	363	5988	6106	-
Aro-mir-14	474	2323	2417	+

Table A.2.: Coordinates of all regulatory elements and ncRNAs that were predicted in *Orussus abietinus*, after manual curation.

Name	Scaffold	Start	End	Strand
U12	1	930553	930702	-
Oab-mir-210	1	1283063	1283157	+
tRNA_Asn-1	1	2959736	2959809	+
tRNA_Thr-1	1	2960113	2960186	+
tRNA_Tyr-1	1	3121509	3121595	+
tRNA_Tyr-1_Intron	1	3121546	3121559	+
tRNA_Arg-1	1	3123523	3123595	-
U1-1	1	4455016	4455177	+
U1-2	1	4456139	4456300	-
U1-3	1	4457071	4457232	-
tRNA_Ile-1	1	4507292	4507365	+
tRNA_Ile-2	1	4507955	4508028	+
tRNA_Ile-3	1	4508345	4508418	+
tRNA_Ile-4	1	4508538	4508611	+
tRNA_Ser-1	1	5161536	5161617	+
U1-4	2	36681	36844	-
tRNA_Thr-2	2	576079	576150	-
tRNA_Ala-1	2	1056957	1057029	+
tRNA_Arg-2	2	2442274	2442346	+
Rnase_MRP	2	2805931	2806208	+
U2-1	2	3521344	3521517	-
Oab-mir-282	3	206061	206151	+
Oab-mir-932	3	333874	333982	-

Continued on next page

Table A.2.: Continued from previous page.

Name	Scaffold	Start	End	Strand
U2-2	3	445664	445856	-
Oab-mir-190	3	1257791	1257885	-
Oab-mir-87	3	1344741	1344825	-
5S_rRNA-1	3	1859953	1860069	+
U2-3	3	3162892	3163084	-
tRNA_Asp-1	3	3381573	3381644	-
5S_rRNA-2	4	1023077	1023195	-
tRNA_Leu-1	4	1776042	1776125	-
Oab-mir-965	4	1966553	1966644	-
SCARNA8	4	2468726	2468854	-
tRNA_Arg-3	4	2761017	2761089	+
tRNA_Pro-1	4	2802085	2802156	+
5S_rRNA-3	4	4054289	4054408	+
tRNA_Ile-5	5	537640	537713	-
K_chan_RES-1	6	271644	271755	+
5S_rRNA-4	6	699486	699597	+
5S_rRNA-5	6	702106	702217	-
Oab-mir-219	6	835051	835152	-
Oab-mir-2944	7	1344549	1344650	-
Oab-mir-9c	7	1344854	1344944	-
Oab-mir-996	7	1344998	1345093	-
Oab-mir-3478	7	1345150	1345228	+
tRNA_His-1	7	1388018	1388089	+
tRNA_Arg-4	7	1389891	1389963	-
tRNA_Met-1	7	1390519	1390591	+
Oab-mir-1000	7	2260319	2260412	+
Oab-mir-8	8	835867	835959	-
tRNA_Arg-5	9	1213090	1213162	-
tRNA_Ala-2	9	1518666	1518737	-
Oab-mir-981	9	2229972	2230066	+
tRNA_Gln-1	9	2408190	2408261	+
Oab-mir-133	9	2555006	2555104	-
Oab-mir-1	9	2695303	2695380	-
5S_rRNA-6	10	340838	340960	+
5S_rRNA-7	10	647821	647939	+
Oab-mir-125	10	703187	703282	-
Oab-let-7	10	705647	705746	-
Oab-mir-100	10	716391	716488	-
5S_rRNA-8	10	811116	811238	-
Oab-mir-980	10	2852896	2852986	-
tRNA_Asn-2	10	2934926	2934999	-
SNORD31	10	4305874	4305938	+

Continued on next page

Table A.2.: Continued from previous page.

Name	Scaffold	Start	End	Strand
Oab-mir-79	10	5758988	5759078	-
Oab-mir-9b	10	5759137	5759229	-
Oab-mir-283	12	128994	129084	+
Oab-mir-3477	12	130305	130403	+
U4atac	12	845973	846101	-
Oab-mir-2788	12	1152310	1152406	+
Oab-mir-12	13	130802	130881	+
Oab-mir-9a	13	1220697	1220789	+
K_chan_RES-2	13	1527118	1527231	-
5S_rRNA-9	13	1735085	1735201	-
tRNA_Tyr-2	13	2061137	2061224	-
tRNA_Tyr-2_Intron	13	2061453	2061467	-
tRNA_Tyr-3	13	2061417	2061504	-
tRNA_Tyr-3_Intron	13	2061173	2061187	-
tRNA_Arg-6	13	3748331	3748403	-
Oab-mir-11	15	136848	136929	+
Oab-mir-263a	15	297665	297757	-
Snopsi28S-1192	15	1066447	1066585	+
5S_rRNA-10	16	1043233	1043350	+
tRNA_Asp-2	17	314009	314080	-
tRNA_Val-1	17	2171027	2171099	-
tRNA_Arg-7	17	2877027	2877099	+
SNORA3	19	1342612	1342745	-
Oab-mir-275	20	1394993	1395084	+
Oab-mir-305	20	1395156	1395240	+
SNORA79	20	2042237	2042369	-
Oab-bantam	21	67032	67112	-
5S_rRNA-11	21	533839	533952	+
tRNA_Ile-7	21	1363942	1364033	+
tRNA_Ile-7_Intron	21	1363980	1363997	+
tRNA_Thr-3	21	2373213	2373285	-
tRNA_Val-2	22	753276	753348	-
tRNA_Val-3	22	754583	754655	-
tRNA_Val-4	22	755645	755717	-
U1-5	23	1172700	1172851	-
Oab-mir-34	23	2114914	2115005	-
Oab-mir-277	23	2122734	2122818	-
Oab-mir-317	23	2138851	2138943	-
U6atac	23	2463193	2463297	-
tRNA_Ser-2	23	2517504	2517585	+
Oab-mir-375	25	1479610	1479701	+
tRNA_Gly-1	25	2337673	2337743	-

Continued on next page

Table A.2.: Continued from previous page.

Name	Scaffold	Start	End	Strand
tRNA_Gly-2	25	2338008	2338078	-
5S_rRNA-12	25	3815022	3815143	+
Oab-mir-278	25	4268603	4268695	-
Oab-mir-67	25	4299820	4299923	+
K_chan_RES-3	25	4362117	4362230	+
Oab-mir-7	25	4682301	4682386	-
Oab-mir-25	25	4789747	4789845	-
tRNA_Glu-1	25	4871592	4871663	-
tRNA_Leu-11	25	4871940	4872061	-
tRNA_Leu-11_Intron	25	4871985	4872023	-
tRNA_Val-5	25	4872353	4872425	-
tRNA_Lys-1	26	350425	350497	-
5S_rRNA-13	28	158005	158220	-
5S_rRNA-14	28	161689	161807	-
5S_rRNA-15	28	165825	165943	+
5S_rRNA-16	28	166500	166618	+
snoU43	29	9280	9356	-
5S_rRNA-17	29	353413	353531	-
SNORA57	29	824418	824560	+
5S_rRNA-18	29	837533	837654	+
5S_rRNA-19	33	642981	643100	+
tRNA_Gly-3	37	59913	59984	+
tRNA_Gly-4	37	67257	67327	-
tRNA_Gly-5	37	67603	67673	-
tRNA_Gly-6	37	67935	68006	-
tRNA_Glu-2	37	72557	72628	+
Oab-mir-137	37	742647	742745	-
Oab-mir-46	37	850395	850487	-
Oab-mir-184	37	2088531	2088626	+
tRNA_Ile-6	38	228990	229062	+
tRNA_Leu-12	38	639374	639491	+
tRNA_Leu-12_Intron	38	639412	639446	+
tRNA_Glu-3	38	639814	639885	+
tRNA_Pro-2	38	1683453	1683524	+
SnoMe28S-Am982	39	219821	219893	-
5S_rRNA-20	39	237753	237869	+
5S_rRNA-21	39	241196	241312	-
tRNA_Pro-3	40	115252	115323	-
tRNA_Trp-1	42	660175	660246	+
tRNA_Asn-3	42	660516	660589	-
tRNA_Trp-2	42	1356424	1356495	-
tRNA_Met-2	42	2747578	2747650	-

Continued on next page

Table A.2.: Continued from previous page.

Name	Scaffold	Start	End	Strand
Oab-mir-2765	42	2808617	2808718	-
Oab-mir-971	45	478800	478896	-
tRNA_Cys-1	53	1938407	1938478	-
tRNA_Gln-2	54	286892	286963	-
tRNA_Ile-8	54	310866	310959	+
tRNA_Ile-8_Intron	54	310904	310923	+
5S_rRNA-22	54	497806	497923	+
Oab-mir-2796	55	678696	678799	+
tRNA_Glu-4	55	907293	907364	+
tRNA_Glu-5	55	920108	920179	-
tRNA_Glu-6	55	923290	923361	-
tRNA_Glu-7	55	956765	956836	-
tRNA_Leu-2	55	957004	957087	+
ACEA_U3	55	2867421	2867635	-
tRNA_Ala-3	55	2869039	2869111	-
tRNA_Glu-8	55	2977306	2977377	+
Oab-mir-6012	55	3170710	3170847	-
tRNA_Lys-2	55	3397734	3397806	-
tRNA_Lys-3	55	3476167	3476239	+
Oab-mir-iab-8	55	4144454	4144546	+
Oab-mir-iab-4	55	4144460	4144541	-
Oab-mir-10a	55	4681645	4681735	+
Oab-mir-10b	55	4721032	4721106	-
R2_retro_el-1	60	58984	59177	+
R2_retro_el-2	60	155806	155995	+
tRNA_Glu-9	61	675806	675877	-
tRNA_Arg-8	62	90404	90476	+
tRNA_Leu-3	62	1904003	1904084	+
tRNA_Glu-10	62	2223704	2223775	+
tRNA_Leu-4	62	2252372	2252451	+
tRNA_Gln-3	62	2252643	2252714	-
tRNA_Ala-4	62	2253169	2253240	+
tRNA_Leu-5	62	2258970	2259053	-
tRNA_Gln-4	62	2493305	2493376	-
tRNA_Gln-5	62	2493445	2493516	-
tRNA_Gln-6	62	2493585	2493656	-
tRNA_Gln-7	62	2493725	2493796	-
tRNA_Gln-8	62	2559318	2559389	+
tRNA_Gln-9	62	2559603	2559674	+
tRNA_Tyr-4	62	2559996	2560091	-
tRNA_Tyr-4_Intron	62	2560229	2560250	-
tRNA_Tyr-5	62	2560193	2560287	-

Continued on next page

Table A.2.: Continued from previous page.

Name	Scaffold	Start	End	Strand
tRNA_Tyr-5_Intron	62	2560032	2560054	-
K_chan_RES-4	62	3865471	3865586	+
tRNA_Ser-3	64	25722	25803	+
Oab-mir-927	64	801800	801887	-
tRNA_Val-6	64	1067585	1067657	-
tRNA_Leu-6	64	1259974	1260057	-
tRNA_Asp-3	64	1287193	1287264	-
tRNA_Gly-7	65	676686	676757	-
tRNA_Gly-8	65	678095	678166	-
tRNA_Gln-10	65	692980	693051	-
tRNA_Asp-4	65	699236	699307	+
tRNA_Tyr-6	65	706370	706461	-
tRNA_Tyr-6_Intron	65	706406	706424	-
tRNA_Met-3	65	948660	948731	-
tRNA_Lys-4	65	1022097	1022169	+
5S_rRNA-23	65	1241026	1241144	+
Oab-mir-279a	66	312044	312133	+
tRNA_Arg-9	67	441040	441112	+
U6-1	67	638273	638379	+
R2_retro_el-3	68	264390	264579	+
snosnR60_Z15-1	69	388480	388564	-
snosnR60_Z15-2	69	388842	388921	-
5S_rRNA-24	69	863051	863169	-
tRNA_Thr-4	69	922716	922787	+
tRNA_Asn-4	70	2055303	2055376	+
tRNA_Lys-5	70	2055699	2055771	-
tRNA_Asp-5	70	2056165	2056236	+
tRNA_Ser-4	70	2056827	2056908	-
tRNA_Asp-6	70	2059793	2059864	+
tRNA_Thr-5	70	2189330	2189401	-
tRNA_Phe-1	70	2195890	2195962	-
tRNA_Phe-2	70	2196030	2196102	-
5S_rRNA-25	78	152782	152893	-
5S_rRNA-26	79	455097	455217	+
tRNA_Ala-5	86	916813	916885	-
tRNA_Ala-6	86	917151	917223	-
tRNA_Ala-7	86	917305	917377	-
tRNA_Ala-8	86	918124	918196	-
tRNA_Ala-9	86	918278	918350	-
Oab-mir-2b	86	1378646	1378728	-
Oab-mir-2a	86	1378974	1379057	-
Oab-mir-2d	86	1379139	1379224	-

Continued on next page

Table A.2.: Continued from previous page.

Name	Scaffold	Start	End	Strand
Oab-mir-13a	86	1379451	1379523	-
Oab-mir-2c	86	1379849	1379926	-
Oab-mir-71	86	1380202	1380298	-
Oab-mir-14	86	1471709	1471803	-
5S_rRNA-27	87	76326	76509	-
tRNA_Asp-7	87	90068	90139	-
Oab-mir-33	88	121724	121811	-
tRNA_Ser-5	89	231886	231967	-
tRNA_Leu-7	90	1325696	1325775	-
Oab-mir-315	90	3183103	3183188	-
5S_rRNA-28	90	5129033	5129150	-
5S_rRNA-29	94	516836	516950	-
U2-4	95	83308	83499	-
5S_rRNA-30	103	78	196	+
5S_rRNA-31	103	29473	29590	+
Oab-mir-929	113	238631	238728	+
Oab-mir-252	113	365047	365158	+
tRNA_Ala-10	113	365413	365484	-
tRNA_Arg-10	116	3360610	3360682	+
tRNA_Gln-11	117	46744	46815	+
Oab-mir-1175	120	945403	945502	-
5S_rRNA-32	121	18959	19080	+
5S_rRNA-33	125	126967	127092	+
tRNA_Val-7	126	2254486	2254558	-
tRNA_Leu-8	126	2941564	2941647	+
tRNA_Pro-4	131	937181	937252	-
Oab-mir-25b	131	1735163	1735241	+
Oab-mir-25c	131	1735374	1735458	+
tRNA_Gln-12	131	2567871	2567942	-
tRNA_Ser-6	131	3095025	3095106	+
tRNA_Thr-6	131	3097672	3097745	-
tRNA_Ser-7	131	3098056	3098137	-
tRNA_Trp-3	131	3098412	3098483	-
tRNA_Leu-9	131	3102431	3102512	-
U4	131	3108303	3108443	+
tRNA_Lys-6	131	3211003	3211075	-
tRNA_Lys-7	131	3225144	3225216	-
tRNA_Met-4	131	3681057	3681128	-
tRNA_Met-5	131	3681325	3681396	+
tRNA_Met-6	131	3681650	3681721	+
5S_rRNA-34	136	383037	383155	-
18S_rRNA-1	139	81747	83639	+

Continued on next page

Table A.2.: Continued from previous page.

Name	Scaffold	Start	End	Strand
Arthrophod_7SK	140	441017	441314	+
Histone3-1	145	22704	22749	-
Histone3-2	145	23840	23884	+
Histone3-3	145	35009	35053	+
Histone3-4	145	36276	36322	+
Sphinx_1	145	73975	74072	+
Sphinx_2	145	74382	74532	+
tRNA_Phe-3	145	122843	122915	-
tRNA_Ser-8	145	402493	402574	-
tRNA_Ser-9	145	402911	402992	-
tRNA_Leu-10	145	524944	525025	+
Histone3-5	145	672817	672863	+
Histone3-6	145	673309	673355	-
tRNA_Arg-11	145	677580	677652	-
Oab-mir-276	145	824248	824342	+
tRNA_Gly-9	145	1281726	1281796	+
tRNA_His-2	145	1282099	1282170	+
tRNA_Lys-8	145	1282890	1282962	-
tRNA_Gly-10	145	1283880	1283950	+
tRNA_Gly-11	145	1284891	1284961	-
tRNA_Ala-11	145	1285979	1286050	+
tRNA_Pro-5	145	1673893	1673964	-
tRNA_Lys-9	145	1773376	1773448	-
5S_rRNA-35	150	125902	126022	-
5S_rRNA-36	150	150134	150253	-
Oab-mir-316	151	575503	575589	+
U1-6	151	900646	900808	+
Oab-mir-928	151	1062109	1062208	+
Oab-mir-31	151	1376209	1376296	+
tRNA_Ser-10	151	1521263	1521344	-
Oab-mir-124	156	61239	61336	+
tRNA_Tyr-7	160	52191	52286	-
tRNA_Tyr-7_Intron	160	52227	52249	-
RnaseP_nuc	160	101783	102068	-
Oab-mir-279d	170	1289079	1289172	-
18S_rRNA-2	177	14619	16312	+
5S_rRNA-37	178	389220	389336	-
tRNA_Met-7	182	176587	176658	-
tRNA_His-3	217	184861	184932	-
Metazoa_SRP	220	56463	56760	-
5S_rRNA-38	228	32389	32508	-
Oab-mir-29	238	55287	55376	-

Continued on next page

Table A.2.: Continued from previous page.

Name	Scaffold	Start	End	Strand
tRNA_Thr-7	254	37639	37711	-
U11	266	6405	6537	-
tRNA_Pro-6	269	43041	43112	-
Oab-mir-263b	304	71335	71425	+
Oab-mir-279	309	559628	559723	+
U5-1	315	6762	6883	+
tRNA_Cys-2	315	11041	11112	+
tRNA_Thr-8	315	17807	17880	-
tRNA_Thr-9	423	445	518	-
tRNA_Ala-12	423	915	987	-
tRNA_Pro-7	423	2616	2687	-
tRNA_Pro-8	423	10193	10264	-
tRNA_His-4	461	94810	94881	-
Histone3-7	464	148020	148066	+
Histone3-8	464	152065	152111	-
tRNA_Val-8	482	858	930	+
Histone3-9	508	8388	8434	+
Histone3-10	508	11263	11308	+
tRNA_Met-8	633	763	835	+
U6-2	633	15671	15777	+
U5-2	770	1	87	+
U5-3	770	527	645	-
U5-4	770	1200	1321	+

A.2. Electronic supplement

The electronic supplement is available on the enclosed CD.

A.2.1. Scripts

This folder contains all self written scripts that were mentioned in this thesis. See chapters 2 and 5 for further details how the scripts were used, which input they need and the output they produce.

A.2.2. ncrna_results

This folder includes direct output files of used programs, as well as results of further analysis of the ncRNA analysis with subfolders for *A. rosae*, *O. abietinus*, and the FEELnc results of *A. mellifera* and *N. vitripennis*.

Also, the lists of the families we removed from the Rfam/miRBase analysis are included.

The folders of *A. rosae* and *O. abietinus* have the same structure. They contain one folder with the FEElnc results (predicted lncRNAs, predicted gene interactions), the results of DARIO (direct output for the three read sets, filtered predictions), the results of the Infernal prediction (direct output miRBase (cmsearch_g_species), automatically filtered miRBase alignments (species_cmsearch_aln, stockholm format), manually filtered alignments (species_cmsearch_aln_sortout), direct output Rfam as a table (species_cmsearch_12.tbl), filtered output Rfam (species_cmsearch_rfam_aln, stockholm format)), as well as the direct output of tRNAscan-SE (species_trnascan.out).

A.2.3. CNE_results

This folder contains the results of the CNE analysis. It contains one folder with the direct CNEr outputs for all pairwise comparisons. Also, it contains outputs produced by the perl-scripts. See chapter 5 and figure 5.1 for futher information.

Acronyms

Aub	Aubergine
bp	base pair
CNE	conserved non-coding element
CNS	conserved non-coding sequence
DNA	deoxyribonucleic acid
dsRNA	double-stranded RNA
Hox gene	homeobox gene
ITS	internal transcribed spacer
lincRNA	long intergenic non-coding RNA
lncRNA	long non-coding RNA
LSU	large transcriptional subunit
MFE	minimal free energy
miRISC	microRNA-induced silencing complex
miRNA	micro RNA
mRNA	messenger RNA
MRP RNA	mitochondrial RNA processing
MSA	multiple sequence alignment
mya	million years ago
my	million years

ncRNA non-coding RNA

OGS official gene set

piRNA PIWI-interacting RNA

Pol II RNA polymerase II

rasiRNA repeat-associated RNA

rDNA ribosomal DNA

RISC RNA-induced silencing complex

RNA ribonucleic acid

RNAi RNA interference

rRNA ribosomal RNA

scaRNA small Cajal body-specific RNA

siRNA small interfering RNA

SSU small transcriptional subunit

snoRNA small nucleolar RNA

snRNA small nuclear RNA

SRP RNA signal recognition particle RNA

ssRNA single-stranded RNA

TE transposable element

TFBS transcription factor binding sites

tRNA transfer RNA

TSS transcription start site

UCE ultraconserved element

UTR untranslated region

WGA whole genome alignment

Zuc Zucchini

List of Figures

1.1 Schematic overview of the secondary structure of four different ncRNAs: miRNA, tRNA, H/ACA snoRNA, C/D snoRNA. 3
1.2 Biogenesis pathway of a miRNA. 7
1.3 Visualisation of the piRNA ping-pong loop. 15
1.4 Graphical overview of the rDNA cluster. 20

2.1 Graphical overview of the steps in the pipeline used for homology prediction of miRNAs. 37
2.2 FEELncclassifier description. Sub classification of intergenic and genic lncRNA/transcripts interactions by the FEELncclassifier module. Taken from Wucher et al. (2017). 42

3.1 Absolute numbers of different ncRNAs found through homology analysis. 47
3.2 Graphical overview of the mir-2 cluster. 49
3.3 Visualisation of tRNA clusters found in *A. rosae* and *O. abietinus*. . . . 55
3.4 Visualisation of overlapping ncRNAs predicted using the DARIO pipeline in *A. rosae*. 57
3.5 Visualisation of the overlapping ncRNAs predicted by the DARIO pipeline in *Orussus abietinus*. 61

5.1 Graphical overview of the steps in the pipeline used for CNE prediction. 76

6.1 Visualised number of CNE candidates identified with CNEr. 83
6.2 Visualisation of the numbers of CNEs per cluster seen in table 6.2 in the four species. 85
6.3 Distribution of CNE prediction in *Apis mellifera* differentiated by species. Only the results for the six longest scaffolds are shown. The number of CNEs is the accumulative total amount found on this scaffold. x-axis shows the genomic location on the scaffold, y-axis the number of CNEs. The results are for pairwise comparisons between species. 89
6.4 Distribution of CNE predictions in *Athalia rosae*. 90
6.5 Distribution of CNE prediction in *Nasonia vitripennis*. 91
6.6 Distribution of CNE prediction in *Orussus abietinus*. 92

List of Tables

1.1 A selection of different definitions of conserved non-coding elements and ultraconserved elements. 28

3.1 List of all miRNAs present in seven insect species. 49
3.1 Continued from previous page. 50
3.1 Continued from previous page. 51
3.2 List of all regulatory elements and ncRNAs, excluding miRNAs and tRNAs, present in seven insect species. 52
3.2 Continued from previous page. 53
3.3 Read counts of the different datasets that were prepared for DARIO. . 53
3.4 Results of the tRNA *de novo* prediction. 58
3.5 List of tRNA families containing introns. 58
3.6 Results of *de novo* ncRNA prediction using the DARIO pipeline. 59
3.7 Number of lncRNAs predicted in four Hymenoptera species. 65
3.8 FEELnc lncRNA-gene interaction results for *A. rosae*, *O. abietinus*, *A. mellifera*, and *N. vitripennis*. 66

6.1 Number of CNEs identified by CNEr sorted by species, number of CNEs left after overlapping ones were combined, size of the assembly (Mb), and N50 (kb) of the assembly. 82
6.2 Number of CNE clusters, grouped by CNE numbers. 84
6.4 Total amount of CNE clusters and number of cluster with lncRNA/protein-coding genes. 86
6.3 Ratios of lncRNA/protein-coding genes. 86
6.5 CNE clusters with lncRNA in *cis* direction next to it. 87
6.6 Number of clusters consisting of ≥ 10 CNEs with an lncRNA as the associated gene with the information of the shortest distance between the cluster and the lncRNA found. 93

A.1 Coordinates of all regulatory elements and ncRNAs that were predicted in *Athalia rosae*, after manual curation. 115
A.2 Coordinates of all regulatory elements and ncRNAs that were predicted in *Orussus abietinus*, after manual curation. 124

Danksagung

Auch wenn es schwierig ist allen Leuten, die mich in den letzten Jahren unterstützt haben zu danken, versuche ich es. Zuerst natürlich Prof. Dr. Bernhard Misof, der mir in seiner Arbeitsgruppe die Möglichkeit gegeben hat diese Arbeit zu verwirklichen. Am meisten bedanken möchte ich mich bei Dr. Alexander Donath, der diese Arbeit und mich betreut hat und mir immer mit seinem Rat und Wissen beiseite stand, auch wenn ich an den Skripten verzweifeln wollte.
Desweiteren danke ich Dr. Lars Podsiadlowski, der diese Arbeit begutachtet, und Prof. Dr. Dietmar Quandt und Prof. Dr. Thomas Döring dafür, dass sie sich zur Teilnahme an der Prüfungskommision bereit erklärt haben.
Weiterer Dank gilt allen Mitgliedern der Graduate School on Genomic Biodiversity Research für interessante Gespräche und fürs Brainstormen.
Nicht zu vergessen sind die Menschen außerhalb der Arbeit, die mich mit Motivation, Aufmunterung und ihrer Freundschaft versorgt haben. Ganz besonders Anna-Lisa Hahnen, die sich die Mühe gemacht hat diese Arbeit gegenzulesen und meinem Englisch etwas zu helfen.
Zu guter Letzt danke ich meiner Familie, für einfach alles.

Electronic Supplement