UNIVERSITÄT LEIPZIG
Fakultät für Mathematik und Informatik
Institut für Informatik

RAP: Ribozeq Analysis Pipeline

Masterarbeit

Leipzig, den 31. s.September 2020

Betreuender Hochschullehrer:
Dr. rer. nat. Jörg Fallmann
Reviewer:
Prof. Dr. rer. nat. Peter F. Stadler
Professur für Bioinformatik
Institut für Informatik
Universität Leipzig

vorgelegt von:

Dr. med. Christiane Gärtner
Studiengang Bioinformatik

Contents

Acronyms

ann. Annotation.

C. Clostridium.

CFG context-free grammar.

CLIP Cross-linked immunoprecipitation.

CM covariance model.

CYK Cocke-Younger-Kasami.

D. Desulfitobacterium.

DAG directed acyclic graph.

Des. Desulfovibrio.

DNA Desoxyribonucleic acid.

DSM *Deutsche Sammlung von Mikroorganismen* - German Collection of Microorganisms.

F. Fervidicella.

HBV Hepatitis B virus.

HDV Hepatitis delta virus.

HH Hammerhead.

HITS High-throughput sequencing of RNA.

HMM Hidden-Markov-Model.

miRNA micro ribonuleic acid.

MMP Maximal Mappable Prefix.

mRNA messenger ribonuleic acid.

NCBI National Center for Biotechnology Information.

ncRNA non-coding ribonucleic acid.

nt nucleotide.

P. Paenibacillus.

PCR Polymerase Chain Reaction.

pf_JF peakfinder from J. Fallmann.

RAGATH RNAs Associated with Genes Associated with twister and hammerhead ribozymes.

RAP Ribo-seq Analysis Pipeline.

RF cm-models from Rfam database.

RNA Ribonucleic acid.

rRNA ribosomal ribonucleic acid.

S. Schistosoma.

SAM Sequence Alignment/Map.

SCFG stochastic context-free grammar.

SSV Single ungapped Segment Viterbi algorithm.

STAR Spliced Transcripts Alignment to a Reference.

tRNA transfer ribonuleic acid.

UCSC University of California Santa Cruz.

UMI Unique Molecular Identifier.

VS Varkud satellite.

Z cm-models built by Z. Weinberg.

Chapter 1

Introduction

1.1 RNA and the origin of life on earth

Nucleic Acids Ribonucleic acid (RNA) and Desoxyribonucleic acid (DNA) are nucleic acids composed of alternating sugar and phosphate residues of individual nucleotides linked by phosphodiester bonds. DNA and RNA are long, linear chains (polymers) of nucleotides. The chemical composition of nucleic acids and their structure of repetitive nucleotide units allow them to function as both information carrier and mediator. RNA is composed of a sugar ribose and the nucleotides uracil, guanine, cytosine, and adenine, whereas DNA consists of a sugar deoxyribose and has thymin as a base instead of uracil (Figure 1.1, Figure 1.2). The bases that are part of DNA and RNA are either purin bases (adenine and guanine) or pyrimidines (uracil, thymin, and cytosin). In the following the nucleotide bases are abbreviated with their initial letter: adenine-A, cytosin-C, thymin-T, uracil-U, guanine-G. Purines consist of a heterocyclic aromatic ring skeleton of six atoms with an additional imidazole ring, pyrimidines are one-carbon nitrogen ring bases (Figure 1.2). Chemical modifications of the bases are widespread, especially in ribosomal ribonucleic acid (rRNA) and transfer ribonuleic acid (tRNA), e.g., inosine, pseudouridine, and dihydrouracil. The length of RNA molecules ranges from $\sim$ 21nucleotide (nt) long micro ribonuleic acid (miRNA)s to $\sim$19000nt long (Manolio et al., 2009).

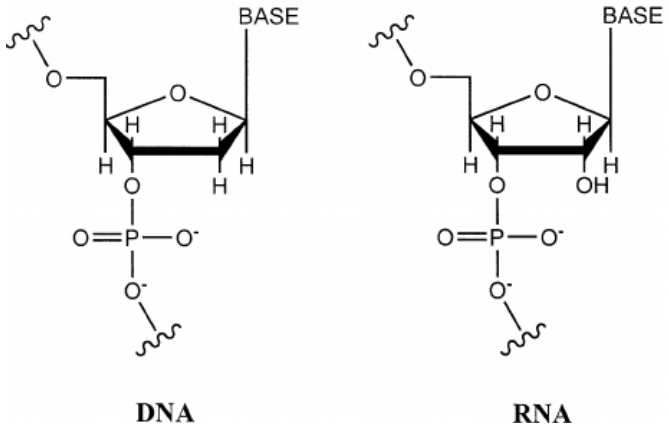


Figure 1.1: **Structure of single DNA and RNA molecules.** DNA is composed of Desoxyribose, whereas RNA contains Ribose. Adapted from [57]

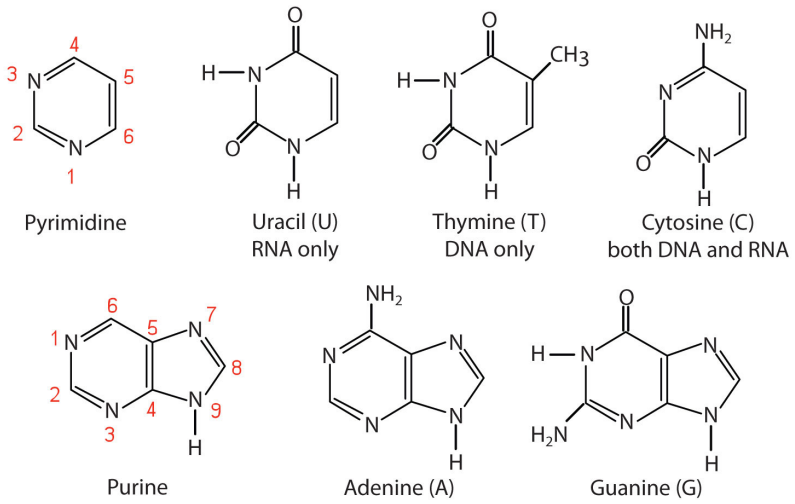


Figure 1.2: **Purin and Pyrimidine bases of RNA (guanine, cytosine, adenine, and uracil) and DNA (thymin instead of uracil).**

Secondary structure and their motifs in RNA DNA is a double-stranded molecule and holds together due the Watson-Crick base pairings between A-T and C-G. DNA folds amongst others into the widely known double helix structure, discovered by Watson and Crick and represents the secondary structure of DNA. This form is known as B-DNA, it is right-handed with about 10–10.5 base pairs per turn [26]. On the contrary, the secondary structure of RNA is more complex and is described below.

While the primary structure describes the RNA sequence,the order of the nucleobases from its 5' to 3' end, the secondary structure of RNA molecules is formed by base pairing within an RNA strand. Even though RNA is a single-stranded molecule, it forms complex base-pairing interactions due to its ability to form hydrogen bonds stemming from the hydroxyl group in the ribose sugar. In secondary structure diagrams, paired residues are indicated by connecting lines. Each RNA molecule builts a characteristic secondary structure, an example for the secondary structure of the Hammerhead (HH) type 3 ribozyme can be seen in Figure 1.7. The secondary structure of RNA molecules can be visualized with the so called dot-bracket notation. Base pairs are represented by matching pairs of parenthesis () and unpaired nucleotides by dots. The dot-bracket notation for the HH-3 molecule is shown in Figure 1.7. The bases A-U and G-C built the Watson and Crick base pairs, but in RNA molecules, the so-called Wobble pair G-U is also typical and approximately as stable as the base pair A-U. Beside these pairings of the Watson-Crick edge, RNA molecules can also interact on the so-called Hoogsteen edge (defined by the purine positions 6, 7, and 8 or the pyrimidine positions 4 and 5) or the sugar edge (formed by the 2' hydroxyl group of the ribose with purine positions 2 and 3 or with the pyrimidine oxygen atom at position 2). Due to the multiple possibilities of intra- and inter-molecular interactions RNA can act in multiple functions [24].

While the secondary structure motif of DNA is mostly a double-helix, RNA builts more diverse motifs in its secondary structure. Figure 1.3 shows the different motifs in RNA. A stem is built by two adjacent base pairs that do not include another base. An interior loop is closed by a base pair on each side and contains at least one unpaired nucleotide on each strand between the base pairs. In contrast to that, a bulge consists of unpaired nucleotides on one side of the strand with two closing base pairs around it.

A hairpin is also a sequence of unpaired bases, but it is closed by only one base pair. A loop closed by more than two basepairs is called a multibranch loop [24].

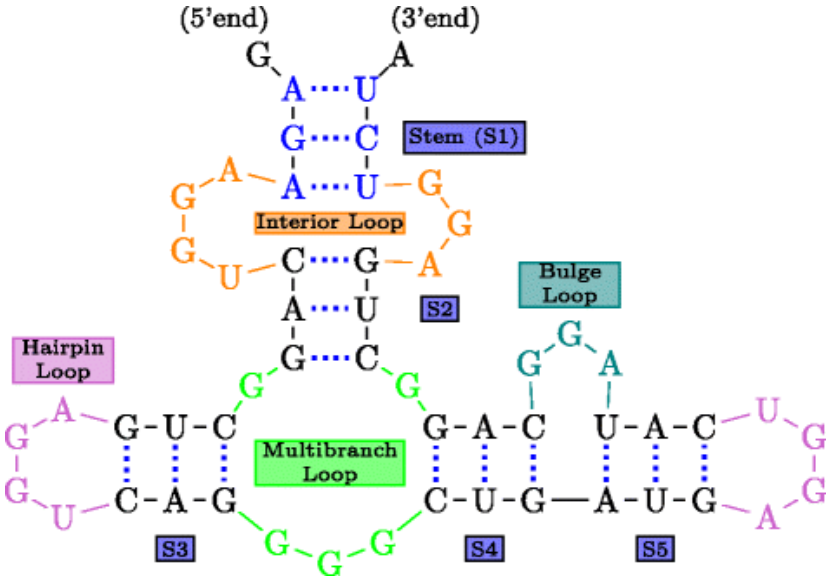


Figure 1.3: **Different RNA motifs in the RNA secondary structure.** The single structures are explained in detail in the paragraph 'Secondary structure and their motifs in RNA'. From [3].

The transcription process The transcription describes the process of copying information from a strand of the DNA into a new RNA molecule, thus enabling the information to leave a cells nucleus. RNA contains the same information as the transcribed DNA section, but it is not an exact copy. The reverse complement of the template DNA sequence is trabscribed into the RNA molecule. Transcription is a complex process and is only briefly presented below and shown in Figure 1.4.

The transcription process starts with the initiation step. The DNA molecule unwinds, and the strands separate to form an open complex. The enzyme RNA polymerase binds to the template strand's so-called promoter with the help of transcription factors. The promoter is a region 'upstream' (towards the 5' end) that controls the transcription of the genes but is not transcribed itself. Transcription factors enhance or repress the transcription of genes by binding to the promoter region and either make it better or less accessible to the RNA polymerase. The template strand is the strand from which the sequence is read.

In the elongation step, the RNA polymerase moves along the template strand, while nucleotides are being added to the RNA molecule. During this process the RNA polymerase moves from 3' to the 5' end of the DNA while the RNA is produced from 5' to 3' end.

The transcription is terminated at the termination side, and the RNA template is then released from the DNA molecule. The RNA molecule at this stage is called pre-messenger ribonuleic acid (mRNA)-molecule. The pre-mRNA-molecules go through some processing steps. For mRNA, these steps are adding a 5' cap, and a 3' polyadenylated tail, and RNA splicing. Splicing is when introns are removed from the mRNA molecule, and the exons are connected [26]. Modifications of RNA bases are also widespread. These are described in the following.

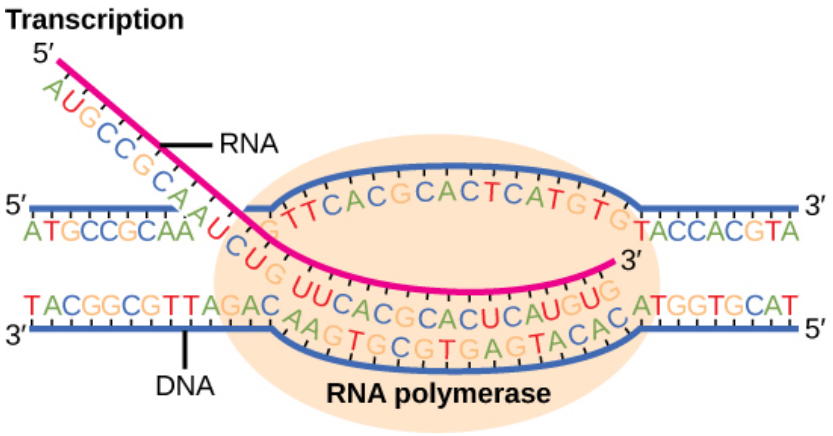


Figure 1.4: **The transcription process.** The RNA-polymerase binds to the template strand of the DNA and transcribes the DNA sequence into a RNA molecule that is identical to the sequence of the coding strand of the DNA. Thymin is replaced by Uracil in RNA molecules. From [11]

RNA modifications Nucleotide modifications appear in all three kingdoms of life. Especially nucleobases of tRNAs (more details below) are widely modified. A relatively common modification is adenosine-to-inosine editing. A double-stranded RNA specific adenosine deaminase catalyzes the deamination of adenosine after transcription. Also, methylations of RNA bases are of research interests. Methylation is a reversible process by which methyl groups are added to the bases or the ribose by various RNA methyltransferases. Methylation impacts numerous biological processes, e.g., translation of mRNA. In tRNA and rRNA molecules N^1-methyladenosine, methylated guanosines, 5-methylcytosine, and a 2'-O-methylations occur [62]. In tRNA and mRNA molecules also pseudouridine molecules (ψ uridines) are found. Pseudouridines are isomers of the nucleoside uridine in which the uracil is attached via a carbon-carbon instead of a nitrogen-carbon glycosidic bond. Pseudouridine is biosynthesized from uridine with the help of ψ synthases. The Carbon-carbon bound gives the molecule more rotational freedom and conformational flexibility [47]. There are many more RNA modifications, and new ones are still being discovered, and their biological significance studied.

Coding RNA molecules and the translation process mRNA molecules are translated into a sequence of amino acids in a process called translation. Although this work deals with RNA molecules that are not transplanted into an amino acid sequence, the process of translation is briefly explained below. As shown in Figure 1.5, the translation takes part in the ribosomes. These are cell particles that are composed of rRNA and associated proteins. They occur both as free particles and bound to the endoplasmatic reticulum (only in eucaryotic cells), which is then called rough endoplasmatic reticulum. They are composed of two different subunits (large and small) in bacteria and eucaryotes. Ribosomes translate the information coded in mRNA into an amino acid sequence that then builds a peptide or protein. Important terms for translation are tRNA, codons, and amino acids. Codons are triplets of nucleotides. Each codon refers to an amino acid and is recognized by a specific tRNA or is a stop codon. tRNA molecules are non-coding RNA molecules that carry an amino acid. The secondary structure contains three hairpin loops that form the shape of a three-leafed clover. The anticodon loop has three nucleotides that bind specifically to the codon

of the mRNA and so decode it. The amino acid is bound to the "CCA" tail at the 3' end of the tRNA sequence. The amino acid loaded onto the tRNA to form aminoacyl-tRNA, is covalently bonded to the 3'-hydroxyl group on the "CCA" tail. Each tRNA has attached its corresponding specific amino acid. Amino acids are organic molecules that contain both an amine- ($-NH_2$) and a carboxyl- ($-COOH$) group. A side chain that is specific for each amino acid completes the molecule. In humans, 20 amino acids are encoded by triplet codons and incorporated in peptides or proteins. Furthermore, two so-called "non-canonical" amino acids are also used to built amino acid chains: selenocysteine (found in all domains of life) and pyrrolysine (found only in some archaea and one bacterium). They are encoded via variant codons (surprisingly, selenocysteine is encoded by a stop codon).

The translation process is divided into three steps: initiation, elongation, and termination. In the initiation step, the ribosome assembles around the mRNA and the tRNA carrying the belonging amino acid, e.g., for the tRNA binding to the sequence "AUG" it is methionine., is attached to the start codon, i.e., "AUG". The present of the start-codon is not sufficient to begin the translation process. It starts due to initiation factors and nearby sequences (e.g., Shine-Dalgarno sequence in bacteria). In the elongation step, the tRNA transfers amino acids to the tRNA corresponding to the next codon. Molecules of rRNA catalyze the peptidyl transferase reaction, forming peptide bonds between the amino acids, linking them together. Then the ribosome moves to the next codon. A chain of amino acids is created that way. Once the ribosome reaches a stop codon, the ribosome releases the polypeptide. A stop codon is a codon that does not code for an amino acid, so no tRNA molecule can bind to it, and the mRNA can then be bound by release factors. After that, the amino acid chain goes through some post-translational modification steps, e.g., covalent and other enzymatic modifications of the proteins [26].

Non-coding RNA non-coding ribonucleic acid (ncRNA) molecules are not translated into a protein, e.g. tRNA or miRNA in contrast to coding RNA molecules described before. One must say, that the exact number of non coding RNAs as well as the exact biological role of some of them remain still unclear. The smallest non-coding RNA is the so-called miRNA with a size of 20nt, which are involved in the regulation of mRNA translation. In contrast long non-protein coding RNA is defined by a length $>$ 200nts. The human genome encodes several thousand of this RNAs. Their crucial roles in a variety of biological processes range from epigenetic control of chromatin, promoter-specific gene regulation, mRNA stability, and imprinting. E.g., Xist RNA, which is 19,000nt long, is involved in the inactivation of the X chromosome [52].

More than 80% of human disease-associated loci are associated with non-protein-coding regions [51]. ncRNAs represent a highly divers group of regulatory RNAs with respect to characteristics, localization and modes of action [37]. Ribozymes, that are investigated in this study, are also non-coding RNAs and are described in detail in section 1.2.

Properties of RNA RNA carries the DNA-encoded information from the nucleus into the cytosol. In this process DNA is working as a template for transcription of RNA. RNA can also catalyse chemical reactions in the form of ribozymes (section 1.2). A third attribute of RNA is that it acts as regulator of gene expression. It was shown that RNA molecules can interact with the general transcription machinery. This leads to an

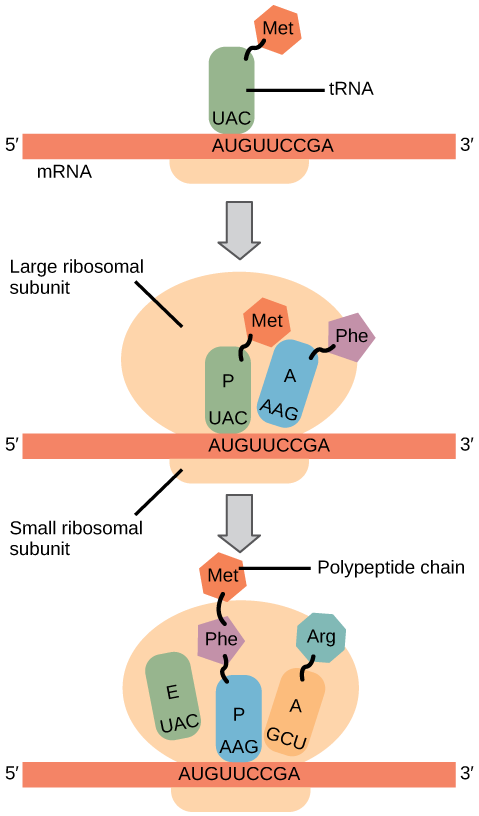


Figure 1.5: **The translation process.** From top to bottom: The translationstarts with an initiation step, is followed by the elongation and is then terminated, once a so called stop-codon is reached. From [11].

inhibition of transcription [70]. In general, transcription can be altered, disrupted, or truncated by RNA molecules. Other regulators are RNA thermometers that regulate gene expression in response to temperature changes [37].

RNA world hypothesis The various functions that RNA molecules can perform led to the hypothesis of the RNA-world. This theory describes a stage in the evolution of life that uses RNA and not DNA for information storage in organisms. The self-replicating RNA molecules proliferated before DNA or proteins. The properties of different RNA molecules that support this theory are catalytic activity, storage and replication of genetic information, and regulatory functions. The self-cleaving ribozymes investigated in this thesis are an example for RNA molecules that can catalyze chemical reactions. miRNA molecules are examples for regulatory RNA molecules. RNA molecules are less stable than DNA molecules. It could be possible that the fragility of RNA molecules could be counteracted by methylation. These functions and properties taken together show the importance of RNA in living organisms [27].

1.2 A short introduction to ribozymes

Catalytic RNA or RNA-protein complexes with catalytic active RNA are called ribozymes (**ribo**nucleic acid en**zymes**). Catalytic RNA molecules were discovered in the 1980s by Thomas Cech and Sidney Altman, which earned the Nobel Price in chemistry for their discovery of catalytic properties of RNA in 1989. There are 14 ribozyme

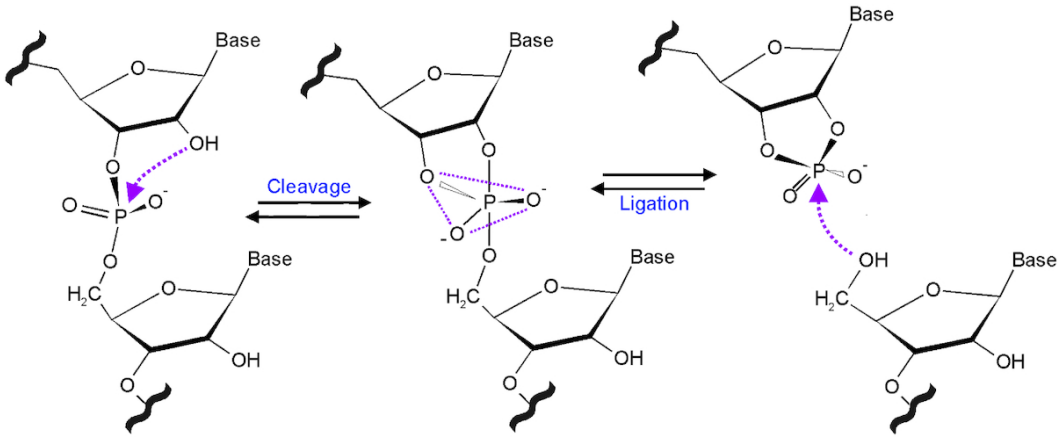
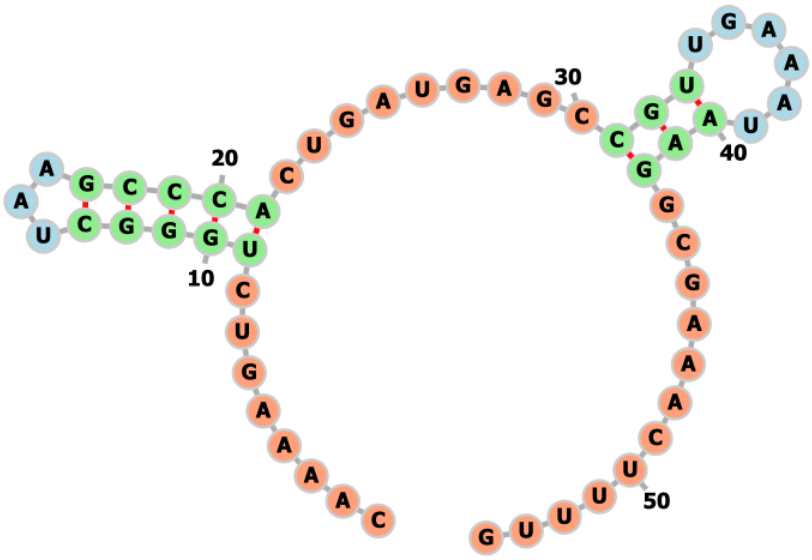


Figure 1.6: **Cleavage mechanism of small ribozymes.** Small ribozymes cut their phosphate backbone at a specific position to enable their biological function. The resulting two RNA fragments then have either a 2',3'-cyclic phosphate(2',3'-cP) or a 5'-hydroxyl group (5'-OH). Small ribozymes also catalyze the ligation of RNA fragments. From [75]

classes known in nature, nine of which are self-cleaving ribozymes. Ribozymes appear in all kingdoms of life. The general functions of ribozymes can be cleavage or ligation of RNA and DNA, peptide bond formation, linking amino acids to peptide chains as part of the ribosome, RNA splicing, virus replication, or tRNA synthesis. Ribozyme can work due to folding in specific tertiary structure. This process requires K^+ and Mg^{2+} ions. The tertiary structure forms the catalytic center. Summed up, RNA is both genetic material and a catalyst [75].

There is a rough separation into large and small ribozymes. Ribozymes that catalyze their intramolecular cleavage are called small ribozymes; examples are HH, Hairpin, and Hepatitis delta virus (HDV). They cut their phosphate backbone at a specific position to enable their biological function. The resulting two RNA fragments have then either a 2',3'-cyclic phosphate(2',3'-cP) or a 5'-hydroxyl group (5'-OH) ([75], Figure 1.6). Small ribozymes also catalyze the ligation of RNA fragments. Large ribozymes are, for example, group I and II introns and the ribosome, the spliceosome, and RNase P. RNase P works like protein RNases but has a catalytic core comprised solely of RNA [4]. Group I and II introns are self-splicing ribozymes that excise themselves from precursor RNAs and ligate the flanking exons to yield mature RNA [64].

The Hammerhead ribozyme (HH) Best studied ribozyme type is the HH ribozyme. The self-cleavage is performed by forming a conserved tertiary structure. HH ribozymes are found in all kingdoms of life. It is built from a highly conserved catalytic center of 15 nucleotides surrounded by three double helices (I to III). The secondary structure of the ribozyme reshapes a HH shark and was the namesake for this ribozyme type (Figure 1.7). Depending on the open-ended helix, there are three possible circularly permuted forms of HH-ribozymes, named type I, II or III. Their biological role extends beyond processing of satellite RNA and viroid replication products, and into the dominion of cellular functions. The genomic localisation of the HH-ribozymes suggest their important biological role. They are found in 3'-untranslated regions of several mammalian C-type lectin type II (CLEC2) genes and located between the stop-codon and poly-A signal sequence [66].



.....((((...))))).....((((.....))))..........

Figure 1.7: **Hammerhead type3, secondary structure.** Top: The secondary structure was calculated with RNAalifold [6] from the Vienna rna package with the cm-models from Rfam database (RF) alignment from HH-type3 (RF0008 from [61]). Visualization was done with forna from the Vienna rna package [30]. Gap columns were removed from the visualization. Bottom: Secondary structure of the HH type3 molecule in dot-bracket notation.

The twister ribozyme The twister ribozyme was discovered in 2014 via bioinformatics analysis [63]. The RNA motif occurs in both bacteria and eukarya species. The secondary structure of the consensus sequence is small but complex. It is composed of three stems conjoined by internal and terminal loops. Two pseudoknots provide tertiary structure contacts that are critical for catalytic activity. The secondary structure is similar to the Egyptian hieroglyph "twister flax," so this hieroglyph was name-giving for this ribozyme type. Comparable to the HH ribozymes, the twister motif could be circularly permutated, leading to three different ribozyme types: Twister-P1, Twister-P3, and Twiter-P5 [63]. Phylogenetic analyses revealed at least ten strongly conserved nucleotides within loop L1 and loop L4, the terminal loop. These two loops are also brought together via a double-pseudoknot arrangement. There are highly conserved guanine and adenine at the cleavage site. One could think that they are important contributors to cleavage chemistry of the twister ribozyme. Some localisations of the twister ribozyme in the genome suggest that the Twister ribozyme could play a role in innate immune defense [23].

The hepatitis delta virus The HDV (hepatitis delta virus) ribozyme is another type of ribozymes investigated in the present study. The hepatitis delta virus is a small single-stranded RNA virus that can infect humans. It is called a satellite virus, because it can only replicate in the presence of the Hepatitis B virus (HBV) (hepatitis B virus) infection in the same person. The infection can be simultaneous with HBV or as a superinfection in patients that are HBV carriers or have a chronic HBV infection. The circular RNA genome is in a complex with proteins that are encoded in its genome. Cellular polymerases replicate the genome. It has an envelope containing host phospholipids and three kinds of HBV envelope protein - large, medium, and small hepatitis B surface antigens [38].

The HDV ribozyme The HDV RNA is divided into distinct domains: the coding sequence for a protein, the delta-antigen, and a ~350nt viroid-like sequence that is for protection against cellular RNases and contains self-cleaving ribozymes necessary for replication. The HDV ribozyme is built of five paired regions that form two stacks, which are linked by single-stranded joining strands J1/2 and J4/2. A single nucleotide upstream and about seventy nucleotides downstream of the cleavage site are sufficient for self-scission [74]. The HDV ribozyme uses a nucleotide for general acid-base catalysis: a cytosine acts as the general acid and a metal ion bound water molecule as the general base [75].

Further ribozyme classes Other ribozyme types investigated in this study are ribozymes with the acronym RAGATH in their names. This acronym refers to a bioinformatic strategy to find new self-cleaving ribozymes: RNAs Associated with Genes Associated with twister and HH ribozymes (RNAs Associated with Genes Associated with twister and hammerhead ribozymes (RAGATH)). With this method, twister-sister, pistol, and hatchet ribozyme classes were found. Moreover, unusual examples of HH and HDV ribozymes were found, e.g. RAGATH-2-HDV that is examined in this study. The also found pistol ribozyme performs its cleavage with an internal phosphoester transfer mechanism [76].

1.3 Investigated species

Six different species were investigated in this study, namely *Schistosoma (S.) mansoni*, a water-born parasite of humans, that is known to express a high amount of HH-ribozymes [48], and the following bacteria species: Clostridium (C.) sporosphaeroides, Fervidicella (F.). metallireducens, Paenibacillus (P.) polymyxa, Desulfovibrio (Des.) vulgaris, and Desulfitobacterium (D.) dehalogenans.

Schistosoma mansoni *S. mansoni* causes Schistosomiasis, a disease with acute symptoms like fever, headache, and allergic reactions. After that, the urogenital tract, the intestine tract, the liver, and the spleen are the most affected organ systems. During its life cycle (Figure 1.8), *S. mansoni* has two hosts, humans and snails. *S. mansoni* differentiates in the adult form into two different sexes, but the male and female worms live permanently together in humans' blood vessels. Adult worms have a size of 6-22mm. The pairs lay up to a thousand eggs daily, mainly in the mesenteric vein. With the help of a particular enzyme, the eggs can go through the vessel wall to the adjacent organs, e.g., urinal or intestinal tract. The eggs are then excreted and pass

into freshwater, where the larvae, called miracidia, hatch from the eggs. Special snails are the intermediate hosts for the miracidia. In the snails, they evolve into cercariae. After that, they leave the snails throughout their breath hole back to the freshwater. The cercariae can now infect their final host humans through their skin. Within the veins, the cercariae arrive at the portal vein system, where the development to the adult form and the differentiation to the female or male sex occur [7].

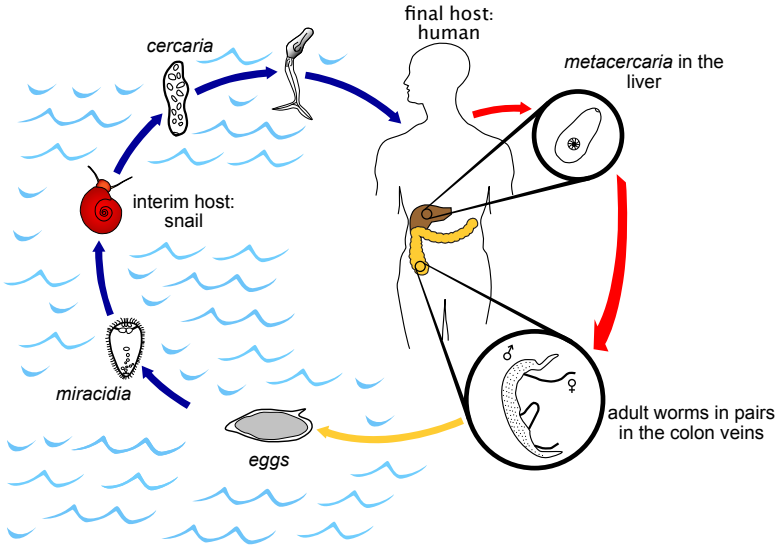


Figure 1.8: **Life cycle of *S. mansoni*** The eggs of the *S. mansoni* are excreted and pass into freshwater. There the larvae (miracidia) hatch from the eggs. The miracidia enter snails, which are an intermediate host for *S. mansoni*. There they evolve to cercariae and come again to freshwater. Through the skin, they come into humans' blood vessels and through the liver to the mesenteric veins. The adult worms live in pairs in the mesenteric veins and lay their eggs. With a special enzyme, the eggs come into the colon. The eggs are then excreted with the faeces. Adapted from [9].

Investigated bacteria species Clostridium is a genus of Gram-positive bacteria. Some Clostridium species are pathogen to humans, e.g., C. difficile or C. botulinum. They are obligately anaerobic, which means that they only grow without the presence of oxygen. There is less known about C. sporosphaeroides (*Deutsche Sammlung von Mikroorganismen* - German Collection of Microorganisms (DSM)1294), which was analyzed in this study [69]. Fervidicella metallireducens (DSM25808) is a thermophilic, anaerobic bacterium from geothermal waters with no known pathology for humans [46]. P.polymyxa (DSM36) is a facultative anaerobic, Gram-positive bacterium found in soil, plants, and marine sediments. It can fix nitrogen so that it may have future applications in agriculture [50]. Additionally, the antibiotic compounds polymyxin were extracted from P.polymyxa strands [56]. The species Des. vulgaris (DSM644) is Gram-negative, anaerobic, sulfate-reducing, and ubiquitous[13]. D. dehalogenens (DSM9161) is an anaerobic, Gram-positive bacteria species. They are facultative organohalide respiring bacteria capable of reductively dechlorinating chlorophenolic compounds and tetrachloroethene [73].

bacteria species	expected ribozyme	reference
C. sporoshaeroides (DSM1294)	Twister-P5	[63]
F. metallireducens (DSM25808)	HHtype2 & Twister-P1	[59]
P.polymyxa (DSM36)	Pistol	[25]
Des. vulgaris (DSM644)	HHtype2	[59]
D. dehalogenans (DSM9161)	Twister-P5	[60]

Table 1.1: **Expected ribozymes in bacteria species that were investigated in this study.**

Expected ribozyme classes Up to date, one would expect to find expressed HH-type1, Twister-P1, and RAGATH-2-HDV ribozymes in *S. mansoni* ([48], [23], [77]). Different ribozyme types are expected in the bacterial specimens; Table 1.1 gives an overview about them.

1.4 Sequence alignments

In bioinformatics, alignments are used, among other things, to investigate sequence homologies. This is based on the state that new sequences are adapted from pre-existing sequences rather than invented *de novo*. So one assumes that similar sequences, e.g., for a gene, belong to related species. There are multiple ways to compare sequences of two or more species; some are described in detail below [19].

The models used for covariance model (CM) search were created from alignments. So in the following the principles of building alignments in bioinformatics should be described.

Pairwise alignments The simplest case is comparing two sequences of a gene from two different species. Let assume that the sequence from the species *rat* is `ATAAC`, and from the species *dog* is `ATAAG`. The alignment of both sequences could be done base per base as seen in Figure 1.9, indicating either a "match" if the bases are equal, or "mismatch" if not. "Matches" are marked as green lines, whereas "mismatches" are marked with red lines in the figure. One way to calculate a score could be to calculate the Hamming distance between the two sequences. The Hamming distance sums up the "mismatching" positions of two strings and would be 1 in the example because only the 5th position's characters do not match.

To compare different pairwise alignments, the "matches" and "mismatches" could be scored, e.g., $+1$ for a "match" on a position and -1 for a "mismatch". The score in the example would be 4.

Multiple sequence alignments Sequence alignments can be performed for more than two sequences, too. An example is given in Figure 1.10. Additional to the two sequences presented in the paragraph before, we have a third sequence `CTAGG` from the species *owl*. One strategy for multiple alignments is called progressive. First the most similar sequences (in our example *rat* and *dog*) are aligned and the third sequence *owl* is added to this alignment. Sometimes an alignment score is better if a so-called "gap" is added to a sequence, e.g., in the *owl* sequence at the fourth position. It is indicated by a "-" in Figure 1.10. A 6th alignment column is inserted to the total

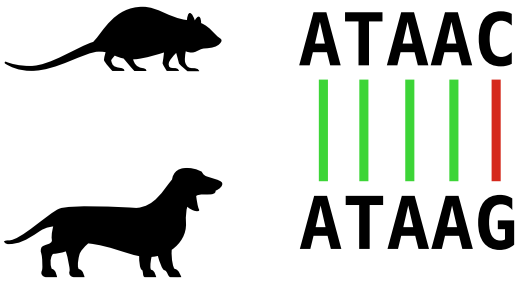


Figure 1.9: **Example for a pairwise sequence alignment.** Matches are connected by a green line, mismatches with a red one.

alignment containing "gaps" in the *dog* and *rat* sequence and *G* in the *owl* sequence. Multiple sequence alignments are a useful tool to analyse the evolutionary relationship of organisms. In this thesis, multiple sequence alignments were the basis of the used ribozyme CM models [19].

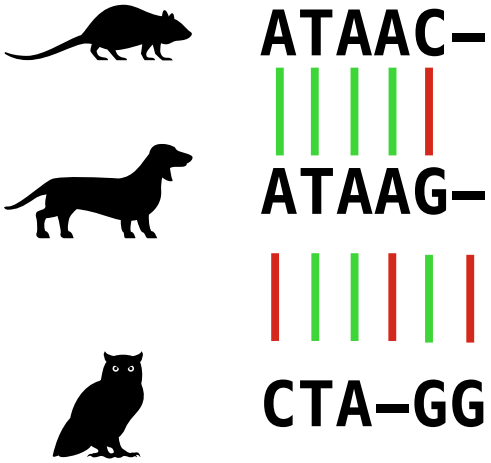


Figure 1.10: **Example for a progressive multiple sequence alignment.** First, the mouse and dog sequences are aligned, wherby green lines indicate matches and red lines mismachtes. Thereafter the thirs sequence from owl is aligned against the first two sequences. On the fourth position, a gap (-) is inserted in the sequence for a better matching score. Additional, a gap is integrated at the end of the two sequences, that were aligned first, due to each sequence in the alignment has to be of the same length.

Global vs. local alignments There are different alignment strategies in either pairwise or multiple sequence alignments. Sequences can be globally aligned so that two sequences were aligned from beginning to end, aligning each letter in each sequence only once. This is useful if one compares sequences of a similar length. In contrast, a local alignment ignores gaps in the shorter sequence's beginning or end while aligning it to the longer sequence. This approach is used if only similarities of subsequences are expected or if the sequences are very different in length [19]. A third approach is

the so-called glocal ("**glo**cal-lo**cal**) alignment, which is a hybrid of the first two forms, where either start or end of the sequences are stated to be aligned. The local and glocal approaches are used, e.g. in the tool *cmsearch* in this study [39].

Mapping Mapping is the alignment of sequenced reads to a genome. It is a crucial step in the analysis of sequencing data. If a reference genome or transcriptome is available, the reads can be mapped directly to this genome or transcriptome. This is called genome or transcriptome mapping. If there is no reference available, the reads have first to be assembled into longer contigs (called de novo assembly). Now one can map the reads to this de novo assembly to quantify the expression of particular reads. In this pipeline, genome mapping is performed as reference genomes for all investigated species are available. There are some challenges in mapping relatively short reads to a whole genome. First, find the correct location of each read in a potentially large quantity of reference data while distinguishing between technical sequencing errors and valid genetic variation within the sample. Second, one read can map to multiple locations equally well. These reads are called multi-mapping reads. Paired-end reads during sequencing can reduce this problem because the two paired reads have to map in a certain distance and specific order [65].

The mapping approach of the Spliced Transcripts Alignment to a Reference (STAR) mapper used in this thesis is to find the Maximal Mappable Prefix (MMP) first. The MMP in STAR is implemented as suffix arrays. A suffix array is the set of suffixes of the genome sorted lexicographically. After the MMP is found, it is extended in the second step [17]. The algorithm of STAR is described in detail in section 2.8.

1.5 Covariance and Hidden Markov Models

The software Infernal used in this study utilises methods based on CMs and Hidden-Markov-Model (HMM)s. Infernal includes programs to build a CM from an alignment (cmbuild), search a target sequence database with a CM (cmsearch), and create multiple sequence alignments of putative homologs with a CM (cmalign; [39]). Therefore bothHMMs, CMs and some important related algorithms will be explained in this section.

Markov Models Covariance models are a generalisation of HMMs. So in this paragraph Markov models are introduced. A Markov model is a stochastic model used to model randomly changing systems and a Markov chain is a simple case of a Markov model. A Markov model can be modeled by a directed graph, whereby the vertices indicate different states and the edges connect the states with a certain transition probability as an attribute. Assuming a sequence is displayed as a path. This path itself follows a simple Markov chain, so the probability of a state depends only on the previous state. The changes in the states are called transitions and their probability transition probability. Figure 1.11 gives an example for a Markov model with the states s_1 to s_x and the transition probabilities p between the individual states. If one walks along with the vertices and edges of this model, one gets a Markov path [19].

Hidden Markov Models HMMs describe a statistical model that augments a Markov model. It is a Markov model for which the state is only partially observ-

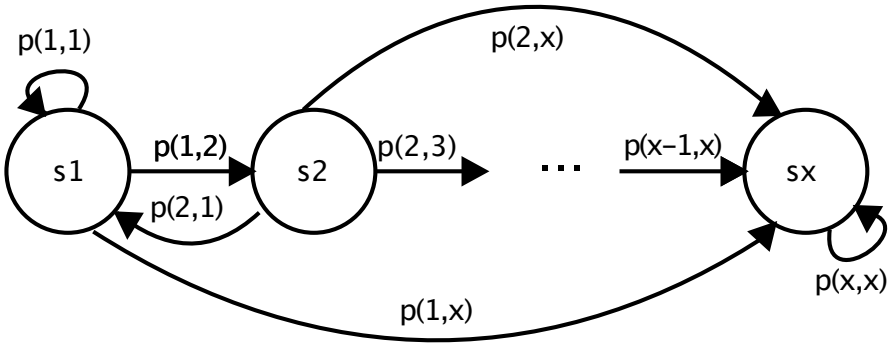


Figure 1.11: **Example of a Markov model.** s_1 to s_x are different states, p indicate the different transition probabilities from state to state.

able. In other words, observations are related to the state of the system, but they are typically insufficient to precisely determine the state. An example for a simple HMM is shown in Figure 1.12. Suppose a prisoner in a dungeon can not look outside. He wants to know what the weather is like. So the hidden states of the HMM in this example would be sunny and rainy (shown as circles). The prisoner judges the weather by the condition of the guards' shoes (dirty or clean, squares). He knows that there is an equal probability that the shoes are dirty or clean when it rains, but that there is a higher probability of dirty shoes when it rains. Furthermore, the prisoner knows that a sunny day with 30% is followed by a rainy day (transmission probabilities, black arrows). In this example, the guards' shoes are the visible emissions of the HMM states (probabilities to different emissions from a certain state are written beside the red arrows). With these pieces of information, the prisoner can calculate the probability for sunny or rainy weather [19].

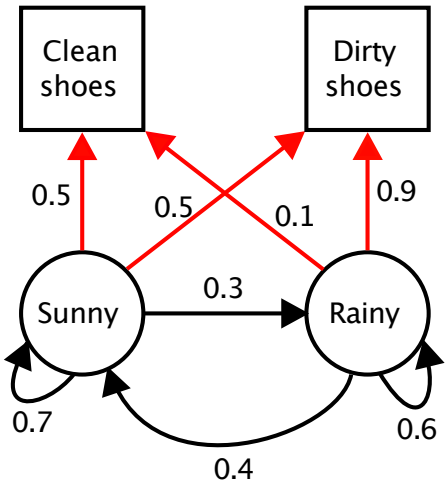


Figure 1.12: **Example of a simple Hidden Markov Model (HMM).** A prisoner can calculate the probability of rainy or sunny weather by the condition of the guards' shoes using a HMM. Details are given in the paragraph 'Hidden Markov Models'.

The Viterbi algorithm The Viterbi algorithm is used to find the most probable sequence of hidden states of an HMM based on a sequence of observations. The algorithm is implemented using dynamic programming. This programming principle breaks complex problems down to simpler sub-problems in a recursive manner. The algorithm's pseudocode is represented as follows (adapted from [29]):

Input: observations of len T, graph of states s of len N
Returns: best-path, best-path-probability

Let viterbi[N,T] be a path probability matrix
Let backpointers[N,T] be a matrix that stores the most probable path
Let a be transition probabilities
Let b be observation probabilities
Let P_s be the start probability distribution over the states
Let $b_i(o_t)$ be emission probabilities, expressing the probability of an observation o_t being generated from a state i

for $s = 1$ to N **do**
  viterbi[s, 1] $\leftarrow P_s \cdot b_s(o_1)$
  backpointers[s, 1] $\leftarrow 0$
end for
for $t = 2$ to T **do**
  for $s = 1$ to N **do**
    viterbi[s,t] $\leftarrow \max_{s'=1}^{N}$ viterbi$[s', t-1] \cdot a_{k,s'} \cdot b_s(o_t)$
    backpointer[s,t] $\leftarrow \arg\max_{s'=1}^{N}$ viterbi$[s', t-1] \cdot a_{s',s} \cdot b_s(o_t)$
  end for
end for
best-path-probability $\leftarrow \max_{s=1}^{N}$ viterbi[s,T]
best-path-pointer $\leftarrow \arg\max_{s=1}^{N}$ viterbi[s,T]
best-path $\leftarrow$ the path starting at state best-path-pointer, that follows backpointer[] to states back in time

The Viterbi algorithm calculates both the probability of the best path and the best paths probabilty. The time complexity of the Viterbi algorithm is $O(N \times |T|^2)$ [29]. This algorithm mirrors the strategy from the HMM stages of *cmsearch* to find the regions that match best to the consensus sequence of the CM model.

Formal grammars HMMs are stochastic regular grammars, and CMs could be described as stochastic context-free grammar (SCFG). Therefore regular grammars and SCFG are described below.

A formal grammar describes how to form strings from a language's alphabet valid according to its syntax. A grammar is defined as a 4-tuple $G = \{N, T, R, S\}$, where S is the start symbol (with $S \in N$), R is a set of production rules, and N and T are nonterminal and terminal alphabets with $N \cap T = \emptyset$.

An element (W, α) in R is called a production rule and is written $W \rightarrow \alpha$. W refers to the left-hand side of the production rule and α to their right-hand side. The

left-hand side must have at least one nonterminal symbol. Grammars are generators and generate a language in a recursive way [29].

Regular grammar Every regular grammar describes a regular language. In the Chomsky hierarchy of grammars, regular grammars are named `type 3` grammars with the most restrictions [10]. The left-hand side must be a single nonterminal, and the right-hand side can be either empty, a single terminal by itself, or with a single nonterminal. Sequences are produced from left to right with regular grammars [1]. All production rules in regular grammars must be of the form:

$W \to \alpha W$, where α is a terminal in T, and W is a nonterminal in N
$W \to \alpha$, where W is in N, and α is in T
$W \to \epsilon$, where W is in N, and ϵ denotes the empty string.

Context-free grammars Context-free grammars (CFGs) are `type 2` grammars in the Chomsky hierarchy of grammars [10]. They are less restrictive than regular grammars so that the production of palindromes or text copies is possible. The productions are of the form

$W \to \alpha$, where α is a string of symbols in $T \cup N$, and W is a nonterminal in N.

One can replace W with an α, wherever one finds a W, regardless of the context. For example, the RNA secondary structure is a kind of palindrome language and can be represented with a context-free grammar. Additional rules allow the grammar to create long-distance pairwise correlations between terminal symbols. The left side's production rule must always be a single nonterminal symbol as in the regular grammars, whereas the right side can be a combination of terminal and nonterminal symbols [24].

An example for a context-free grammar (CFG) that models RNA stem loops with two basepairs and a "gcaa" or "gaaa" loop is given below:

$S \to aXu|cXg|gXc|uXa$, where S stands for the start symbol, $\{a, c, g, u\} \in T$, and $X \in N$,
$X \to aYu|cYg|gYc|uYa$, with $\{X, Y\} \in N$ and $\{a, c, g, u\} \in T$,
$Y \to gaaa|gcaa$, with $\{Y\} \in N$ and $\{a, c, g, u\} \in T$.

For example, in Figure 1.13, a short RNA molecule with its secondary structure and its parse tree for its sequence *acgaaagu* with the secondary structure ((....)) is shown (adapted from [18]).

Stochastic context-free grammars (SCFG) SCFGs can be described as HMMs where the linear path of state transitions is replaced by a tree of states (RNA book) or as CFGs with additional probability distribution on productions, leading to: $G = (N, T, R, S, P_p)$. P_p represents the probabilities $P_p : R \to (0, 1)$. Each production from the example of the paragraph above would be given a probability [24].

$P_p(S \to aXu) = 0.3$
$P_p(S \to uXa) = 0.3$

$P_p(S \to cXg) = 0.2$
$P_p(S \to gXc) = 0.2$
$P_p(X \to aYu) = 0.2$
$P_p(X \to uYa) = 0.2$
$P_p(X \to cYg) = 0.3$
$P_p(X \to gYc) = 0.3$
$P_p(Y \to gaaa) = 0.4$
$P_p(Y \to gcaa) = 0.6$

For the example in Figure 1.13 the sequence *acgaaagu* with the secundary structure ((....)) would have a probability of 0.036.

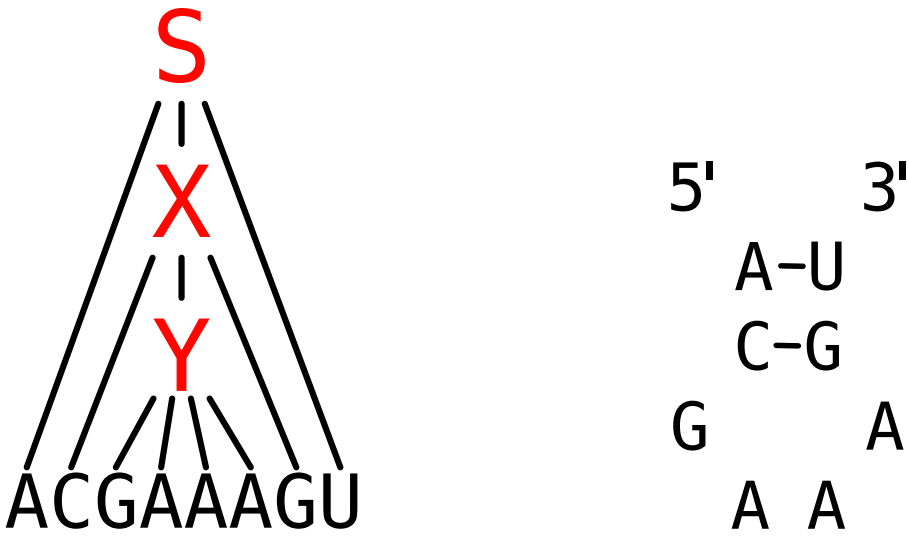


Figure 1.13: **Example of a parse tree for the CFG explained in the paragraph 'Context free grammars' and the resulting secundary structure of the RNA.** Adapted from [18].

The CYK and the inside algorithms The Cocke-Younger-Kasami (CYK) algorithm was developed to parse a CFG. The inventors John Cocke, Daniel Younger, and Tadao Kasami were name-giving for it. The CYK algorithm answers the question, whether a string can be derived under the given CFG. The algorithm is also implemented using dynamic programming. For the CYK algorithm, this means that one starts with small substrings and expands them. The time complexity of the CYK is O(n^3 |R|), the space complexity O(n^2 |N|), where n is a the length of the word in the grammar, |R| is the number of rules and |N| is the number of nonterminals in the grammar [19]. A pseudo-code of the CYK is shown below (adapted from [8]):

Input: string S consisting of n words: $a_1...a_n$

Let the grammar contain r non-terminal symbols $R_1...R_r$
Let R_s be the start symbol of the grammar
Let P[n,n] be a matrix array of sets of non-terminals
Initialize all elements of P to an empty set
for $i = 1$ to n **do**

```
  for each terminal production R_j → a_i do
    set R_j as a member of P[1,i]
  end for
end for

for i = 2 to n do
  for for each j = 1 to n-i+1 do
    for k = 1 to i-1 do
      for production R_A → R_B R_C do
        if R_B is a member of P[k,j] AND R_C a member of P[i-k,j+k] then
          set R_A as a member of P[i,j]
        end if
      end for
    end for
  end for
end for
if R_s is a member of P[n,1] then
  S is member of language
else
  S is not member of language
end if
```

If one wants to find the most probable structure using a SCFG given a specific sequence, the CYK algorithm can be modified. Instead of storing True in the fields of the array, they are replaced with products of probabilities. The adapted algorithm is shown below (adapted from [24]):

Input: string S consisting of n words: $a_1 ... a_n$

```
Let the grammar contain r non-terminal symbols R_1...R_r
Let R_s be the start symbol of the grammar
Let P[n,n, Pr] be a matrix array of sets of non-terminals
Initialize all elements of P to 0
for i = 1 to n do
  for each terminal production R_j → a_i do
    set P[i,i] = P(R_j → a_i)
  end for
end for

for i = 2 to n do
  for for each j = 1 to n − i + 1 do
    for k = 1 to i − 1 do
      for production R_A → R_B R_C do
        if P[k, j,R_B] · P[i-k, j+k,R_C] > P[i, j,R_A] then
          set P[i, j, R_A] = P[i, k, R_B] · P[k + 1, j, R_C] · P(R_A → R_B R_C)
        end if
      end for
    end for
```

end for
end for
P[1,n,S] is the probability of the best parse.

The Inside algorithm In contrast to the CYK algorithm the Inside algorithm computes the total probability of all derivations consistent with a given sequence, based on a SCFG and not only the probability of the best parse. The Inside algorithm is identical to the CYK algorithm for finding the highest probability parse, except that the maxima for the highest probable parse are replaced with sums. Consequently, its time complexity and space complexity are identical to those of the CYK algorithm [24].

Covariance Models CMs are probabilistic models of the multiple sequence alignment and secondary structure of an RNA family (Eddy and Durban, 1994). So the CM is useful for searching databases for homologs of an RNA family. It is of interest to search for such homologies because it implies a similar function or evolutionary homology. To achieve this, the secondary structure of the RNA molecules must also be considered due to the better conservation of the secondary structure compared to the primary structure (sequence). CMs are profile stochastic context-free grammars, analogous to HMMs. A profile model's key feature is its position-specificity: each position of the input alignment is modeled independently. This considers the level of conservation at each position when scoring / aligning candidate family members ([24], [19].

The algorithm for the CM presented by Eddy and Durban (1994) is implemented using dynamic programming [20]. A RNA CM is based on an ordered binary tree that can capture all pairwise interactions of the secondary structure of the RNA molecule. Tertiary structure interactions, pseudoknots, and interactions involving more than two bases can not be integrated into the CM. Each vertex in the representing tree belongs to one of eight possible states corresponding to a different type of structural element: base-pair, single-stranded residue (bulge at the left or right side), beginning of the complete structure, bifurcation, beginning of a substructure (left or right side), and end of a substructure. Only consensus columns (columns in that fewer than half of the residues are marked as gaps) are modeled with vertices. Each consensus column corresponds to a vertex with either the attribute base-pair or single-stranded residue. Base-pair vertices emit two residues at once. The other vertex types are necessary only to model the possible branching patterns of RNA structures. A parse tree and its corresponding sequence are generated from a CM just like in a context-free grammatic with the main difference that each production rule (emission and transition) has an associated probability. Sequences are generated from the outside-in instead of from left to right as in regular grammars like HMMs, by starting at the root vertex (REF Dissertation Eric Nawrocki). For each subsequence, all possible states are calculated, subtrees evolve from bifurcations. For each calculation, one must consider both the transition to the next state and the emission probability in the state as determined by training data ([19], [20]).

With the CMs several RNA analysis problems could be (partially) solved: consensus secondary structure prediction, multiple sequence alignment, and database similarity

searching [19].

Filter steps in Infernal's cmsearch Infernal's *cmsearch* combines the previously described principles (CM and HMM) and their belonging algorithms (Viterbi and CYK) to find hits of cm models in a genome with high sensitivity. Therefore different filter steps are implemented in *cmsearch* to recognize subsequences with CM hits better from filter step to filter step, but also they need more computational requirement from filter to filter (Figure 1.14). Each filter has certain cut-offs to exclude windows from the search [39].

The first *cmsearch* stage contains of HMM Single ungapped Segment Viterbi algorithm (SSV), Viterbi, local and glocal Forward filters. The HMM SSV filter detects hits in long sequences. It depends on ungapped alignment segments and scans each sequence in the database for high-scoring ungapped alignment segments with the Viterbi algorithm. Together with a certain window around them, these segments are now aligned to a profile HMM with a fast Viterbi algorithm. In this filter stage gapped alignments are allowed. Then they go through a local Forward filter. Optional a quick "bias filter" is then applied to the segments. These Segments have false-positive good scores but do not mirror suitable matches to the profile HMM. The next filter is a local Forward filter that runs the segments aligned to a profile HMM. The local mode allows the alignment to begin and end at any position of the model. The likelihood of the target sequence given the profile HMM is summed up over all possible alignments. The next step is a glocal Forward filter. The windows are aligned to the profile HMM again. This time only alignments that begin at the first position and end at the final position of the model are allowed. Here the full Forward algorithm is performed. The stage is called glocal because the alignment to the model is global in principle, but it is potentially local with respect to the sequence window (alignments can start and end anywhere in the window). The second stage contains two filter stages that use CM algorithms. They score both the conserved structure and the sequence of the potential hit. The first of them is the CYK algorithm. It is slow compared to the previous filters, but only a few potential hits are left in this stage. Subsequences that are not filtered out in this stage are brought to a CM Inside algorithm, and, if they also survive this stage, displayed ([19], [39]).

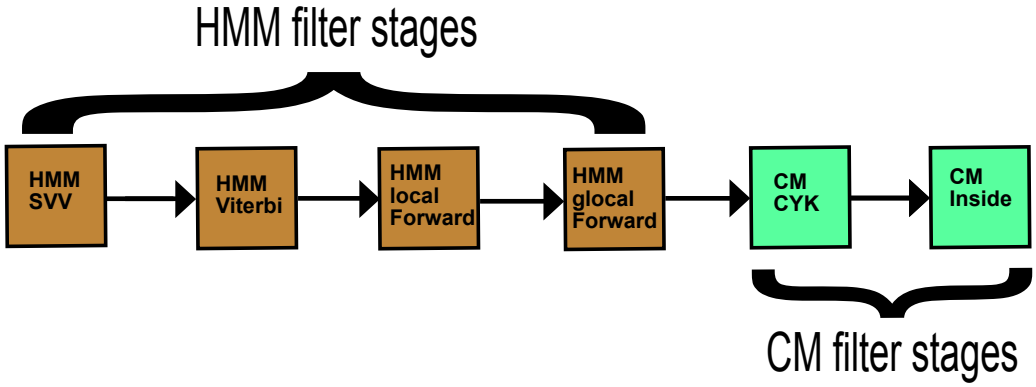


Figure 1.14: **Filter stages of Internals' *cmsearch*.**

1.6 Aims

- To date, in most species, ribozymes are not or insufficiently annotated. Therefore, the present work's first goal is to establish a method for a sufficient annotation of ribozymes in the species' genomes.
- A novel experimental method established by C. Weinberg and J. Olzog (Leipzig University) enables the direct sequencing of ribozymes. The sequencing data then should be analysed for diffenrential expression of ribozymes in vivo. In this thesis a pipeline is to be developed, which allows the automated analysis of the sequence data.
- This pipeline should integrate the deduplication of Unique Molecular Identifier (UMI)s and quantification of transcription with a peak finding step.
- The pipeline should be composed in a way that it can be used by users without a strong bioinformatics background.

Chapter 2

Methods

2.1 Workflow Overview

Figure 2.1 outlines the general workflow established in this thesis. Laboratory experiments (green box, section 2.5) lead to sequencing data subsequently analyzed by myself. Different alignments of ribozyme types (section 2.3) were analyzed with covariance models with the software *cm-search* from Infernal ([39], section 1.5, section 2.4). After that, the results of cmsearch were analyzed and quantified. Ribozymes were then annotated in each genome separately. The mapping of reads from sequencing experiments was performed with the STAR mapper ([15], section 2.8). Before the mapping, UMIs were extracted with the software UMI-tools section 2.7, and after the mapping step, the marked reads were also deduplicated with UMI-tools. Occurences of read-UMI combination were quantified. To find peaks in the sequencing data two peak finding tools, were compared: Piranha, that is typically used to analyze Cross-linked immunoprecipitation (CLIP)-Seq experiments, and a peak finder implemented by J. Fallmann [67], which is in the further work named pf_JF (section 2.9). An intersection of the peaks with the new ribozyme annotation then shows the transcribed ribozymes found in the different species. These results now can be transferred to genome browsers: University of California Santa Cruz (UCSC) or for *S. mansoni* Wormbase (section 2.11). The orange boxes in Figure 2.1 show steps that were integrated into the analyzation pipeline within this thesis. The steps shown in the white boxes were implemented by J. Fallmann in the pilot phase of this study.

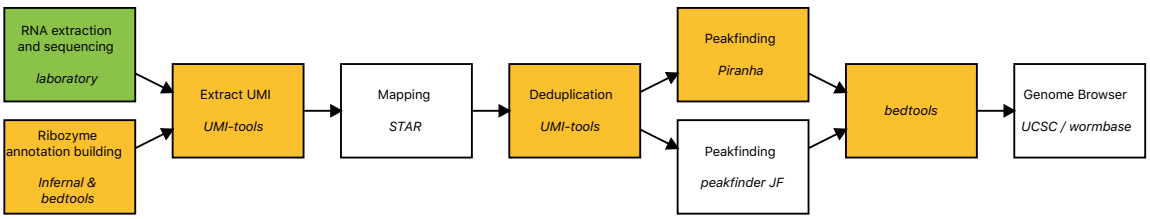


Figure 2.1: **Flowchart of steps in the pipeline.** The green box marks the laboratory step, the white boxes show steps that were implemented by J. Fallmann, and the orange boxes show steps that were included into the pipeline during the creation of the thesis.

2.2 Official genome annotation files

The genome files and annotations for the bacterial species have been obtained from National Center for Biotechnology Information (NCBI) (DSM1294: [40], DSM25808: [41], DSM36: [42], DSM644: [43], DSM9161: [44]).

For *S. mansoni* [45] the current annotation, PRJEA36577, is taken from the Wormbase database [79].

2.3 Missing annotation of ribozymes

Different CMs build out of sequence alignments were compared to find a ribozyme annotation in the genomes of the previous described bacteria species (section 1.3) and *S. mansoni*. Doing so is necessary because only a few ribozyme types are annotated in the current annotations, especially in the bacteria species. Also in the used annotation of *S. mansoni* only ribozymes of the type HH are annotated. This step is important to later assign the transcribed reads from the Ribozeq experiments to see whether ribozymes or other RNA molecules were sequenced.

Table 2.1 lists the investigated cm-files, whereby the files indicated with 'RF' are from the rfam database, Zasha Weinberg created the cm-models indicated by 'cm-models built by Z. Weinberg (Z)' with Infernals' *cmbuild* software [34]. Different sources of the CMs were used to cover all ribozyme types that are known to date. Rfam does not provide CMs for all ribozyme types, so in this case, the models from Z are used.

On some cm-models, there is not much difference between 'RF' and 'Z' cm-models for the same ribozyme, but for HH-type1, HH-type2 and HH-type3 alignments of cm-files from 'RF' and 'Z' are independent of each other. Multiple HDV-Ribozyme models could find different (partially overlapping) hits. On this occasion, the rfam cm-file HDV_ribozyme is the classic alignment for that ribozyme, whereas RAGATH-2-HDV is built as a variant version of it [76]. The RAGATH-1-HH ribozyme is a variant of HH ribozymes. One would not expect any hit for the VS ribozyme in *S. mansoni* or the investigated bacteria species. Up to date, it could only be detected in the mitochondria of a particular fungus species. This 'alignment' consists of only one sequence.

2.4 Finding hits in cm-files with Infernals' *cm-search*

The aim of Infernals' *cm-search* is to compare covariance models against a sequence database. This software was used to find optimal annotations of ribozymes for each investigated species, in the following modes: mid, default (in this work referred to as std), and hmmonly (in this work referred to as hmm). *cm-search* consists of different filter steps to find significant hits, whereby the first five filter stages are build by different HMM filters. In the introduction, one can read more about the HMM algorithms mentioned below (section 1.5).

The filter stages recognize subsequences containing CM hits better with an increasing number of filter stages, but runtime is longer from stage to stage. The single filters are described in more detail in section 1.5. The thresholds in the *std* filter option depend on the size of the database, so they are less sensitive for bigger datasets to prevent a sharp deceleration in running time. Higher sensitivity can be achieved with the *mid*

Table 2.1: **Informations about cm-files used in this study** *clen*: the number of columns from the alignment defined as consensus (match) columns; *nseq*: The number of sequences that the profile was estimated from. *eff-nseq*: The effective number of sequences that the profile was estimated from, after Infernal applied an effective sequence number calculation such as the default entropy weighting. RF files are from Rfam database, other files are build by Zasha Weinberg as described in (section 2.3).

ribozyme name	**clen**	**nseq**	**eff-nseq**
HH-type1_Z	93	10300	36.261032
HH_1_RF	46	29	29.000000
HH_type2_Z	95	2009	10.227243
HH_II_RF	67	24	6.688477
HH_3_RF	58	82	8.615906
HH_3_Z	77	504	25.497620
HDV-F-prausnitzii_Z	72	870	9.541512
HDV-F-prausnitzii_RF	74	48	5.000977
pistol_Z	70	450	13.887405
Pistol_RF	70	45	8.396301
RAGATH-3-twister-sister_Z	83	173	4.543045
Twister-sister_RF	84	4	1.449219
RAGATH-2-HDV_Z	91	1618	9.526760
HDV_ribozyme_RF	91	33	1.575073
Hatchet_RF	162	8	3.261719
hatchet_Z	141	196	9.178528
Hairpin_RF	52	5	3.071289
HH_9_RF	77	33	1.567017
HH_10_RF	125	18	0.962402
RAGATH-1-HH_Z	78	22	2.290771
Twister-P1_Z	65	1872	25.011749
Twister-P3_Z	73	10	4.809570
Twister-P5_Z	58	264	31.384644
VS_Z	167	1	0.566406

option. Here filter stages 1 and 2 are turned off. At *hmmonly* option, the software runs profile HMM searches instead of CM searches in the different filter stages, which reduces the runtime. This option is recommended initially for sequences without predicted base pairings in the RNA. So at the *hmmonly* option, only filter stages 1 to 3 are passed through with strict p-value thresholds compared to mid or std search options [39].

The previously described different options of *cm-search* were tested to find as many annotations of ribozymes as possible in the genomes of the different species.

2.5 Detection of ribozyme transcription from total RNA with RNASeq

The data preparation experiments were performed by V.J. Olzog as part of her PhD-Thesis at the Christina Weinberg Lab at the Institute for biochemistry, University of Leipzig.

The total RNA-isolation from Bacteria cultures or the Schistosomes was followed by the depletion of rRNA. Specific adapter-ligation methods were used to enrich ribozyme fragments to investigate self-cleaving ribozyme. To capture the fragment with a 2',3'-cP, a procedure consisting of ligation using the A. thaliana tRNA ligase was established. This enzyme is responsible for the repair of RNA breaks with 2',3'-cyclic phosphate and 5'-OH ends in the cell. In contrast, other RNA ligases connect the 3'-hydroxyl and 5'-phospho ends. The ligation was followed by dephosphorylation of the 2'-phosphate, reverse transcription, and 16 PCR amplification cycles. After that, Illumina high-throughput sequencing was performed. In future experiments, also the capturing of the fragment with the 5'-OH group using the RtcB ligase for the adapter-ligation should be established.

2.6 Pre-processsing

Pre-processing includes quality control of the sequencing data and adapter trimming. The quality control was run on the FastQ files with the software *FastQC*, thus getting an overview over the read length distribution, sequencing quality per base, and sequencing quality per read [5]. With *FastQC*, one gets an impression where possible pitfalls in the analysis of the data could be, e.g., short reads or a bad sequencing quality. After that, *Trim Galore* was used for both quality and adapter trimming [33]. Before adapter removal, low-quality bases at the 3' end of the reads are removed to enhance the reads' overall quality. Then the adapter is removed from the 3' end of the reads. In the next step, reads with a length under a user-specific cut-off are sorted out, thus analysing only reads of the right length and quality.

2.7 Find and deduplicate reads amplified with PCR by using UMIs

UMIs were integrated into the adapter sequence to exclude errors due to PCR amplification. UMIs allow one to distinguish whether a particular sequence was transcribed more frequently or multiplied only during the PCR cycles. 3'-adapters were structured as follows:

$$5' - NNNNNNNNTGGAATTCTCGGGTGCCAAGG - 3'.$$

"N" here stands for a random nucleotide, and the sequence of 8 random nucleotides forms the UMI, whereas the following sequence is fixed. Using UMI-tools ([49]) the UMIs with the reads belonging to it are first extracted at the level of the FASTQ files. Then the reads are mapped as usual. After mapping, duplicates are removed with the UMI-tools dedup option. The resulting SAM files can now be used for peak finding [68].

2.8 Mapping of the reads

The mapping of the reads to the reference genomes was performed with the STAR mapper [15]. STAR was developed to map reads from RNASeq experiments to the corresponding genome. STAR is designed to align the non-contiguous sequences directly to the reference genome. The STAR algorithm consists of two major steps: 1: seed searching and 2: clustering, stitching, and scoring. First, the longest exact matching subsequence for each read is searched and referred to as 'seed1'. Within the unmapped subsequence, the STAR mapper searches different mapping coordinates for the longest subsequence out of this remaining subsequence, which is then referred to as 'seed2'. If STAR could not find an exact matching sequence for each part of the read, the previously found seeds are extended. If this extension is not of the desired quality, the part of the read will be clipped. Second, after the seeds are found, the reads are stitched together to a complete read again. This step is performed with the help of an anchoring algorithm. The best alignment was determined by a score based on the scoring of mismatches, indels, gaps, matches [17].

STAR runs with the default setting options except for 'genomeSAindexNbases', representing the length (bases) of the SA pre-indexing string, which was set to 13 [16]. It has to be said that the STAR mapper enables reads to map multiple, i.e. at several locations, onto the genome. Due to the ribozyme sequences, some of which frequently occur in the genome, the multi mapping reads cannot be ignored but must be considered separately. This can be done by a so-called fraction count when counting the number of reads at a specific position [16].

2.9 Peak finding

Two peak finding strategies were compared to find peaks in the mapped reads in this study: a peak finder established by J. Fallmann (after this referred to as pf_JF) and the Piranha peak finder. In analogy to the evaluation of CLIP experiments[1], where one searches for the exact positions at the beginning of the transcription process, mapped files were processes so that for each read, the nucleotide at its 3'-end is kept and saved in the correspondending bedgraph files. This strategy should be useful because the ribozyme sequencing experiments described in section 2.5 grip the ribozyme's fragment at the 3' end. So one could state that the 3' end of the ribozyme fragments should be enriched. That is the ribozyme splice side one wants to detect in this workflow. In the experiments, the 5'-end fragment of the ribozyme after cutting is analysed up to date. For the other method using the RtcB ligase to capture the 3' fragment of the ribozyme, the bedgraph files must be created so that the nucleotide on the 5' end of the reads is retained. This procedure was used because the experiment setup leads to reverse transcription starting at the ribozyme-specific intersection site, and therefore a

[1] E.g., High-throughput sequencing of RNA (HITS)-CLIP experiments are performed to analyze protein-RNA-interactions. CLIP starts with an in-vivo cross-linking step using ultraviolet light, that builds covalent bonds between RNA and RNA-binding proteins. With immunoprecipitation, the investigated protein can be isolated. Then RNA-adapters are brought to the 3'-and 5'-ends of the RNA strands. During reverse transcription from 3'-end to 5'-end of the RNA, the cDNA synthesis is truncated at the nucleotides that bind the protein. On the other hand, in the cDNA, synthesized from 5'- to 3'-end of the RNA-strand, mutated nucleotides could be integrated due to UV light treatment. The cDNA is then amplificated via PCR and sequenced. The nucleotides before the truncated reads are of interest and investigated by peak callers for these experiments [12].

strong signal at this position should be determined as precisely as possible. The files processed in this way were then evaluated using the different peak finder tools.

Piranha is a peak finder based on a zero-truncated negative binomial regression model. The mapped reads are binned based on the start nucleotide. After that, Piranha counts the number of reads starting in each bin. In CLIP experiments, the cross-linking event is considered to take place at the nucleotides before the start of the read. In the investigation of the ribozymes, the start of the read is considered as the cleavage-site. Bins with a count >0 are taken and fitted to a probability distribution. In principle, one has the choice between four different distributions. However, the zero-truncated negative binomial distribution is set as default and recommended for CLIP-seq datasets. Piranha reports per default peaks with a p-value of 0.05 [72], [71]. In this study, Piranha was run with the default setting, which means the zero-truncated negative binomial regression model is used.

The pf_ JF takes as input the bedgraph files, and based on the profile out of these files, the peaks are called. Within a sliding window, the most mapped reads' position is taken if the frequency of mapped hits on this position is higher than a user-defined cut-off. Then the area is enlarged in 3'- and 5'-direction. If the rate of mapped reads on a position drops under a specific value (default 30% of the maximum value), the region is no longer expanded. Then a p-value is added to the peak [67].

2.10 Analyses with *bedtools intersect*

Bedtools was used for different purposes within this thesis. First, with *bedtools intersect*, one can find overlapping hits from the *cm-searches* with the various search options and also the number of hits that are not covered by other search modes. Second, peaks found with Piranha or the peak finder from JF were compared. Third, the detected peaks were intersected with my ribozyme annotations and the official annotations to distinguish the peaks [54]. Figure 2.2 shows the functional principle of bedtools intersect.

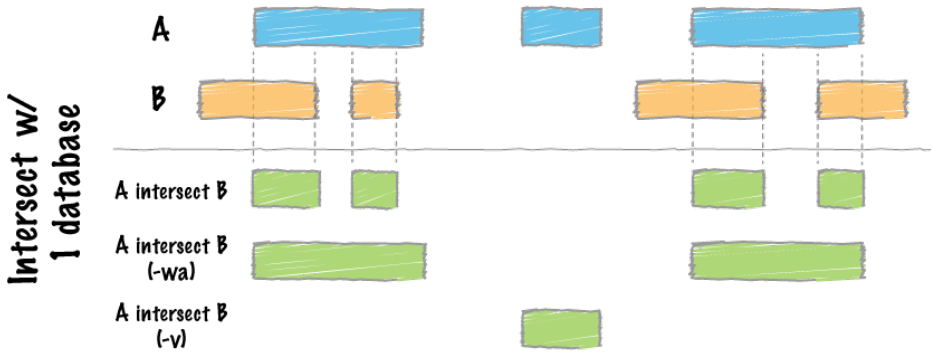


Figure 2.2: **Bedtools intersect options.** A intersect B without another option outputs all sections from B that overlap with A. Using the option -wa, any section out of A that is (partially) overlapped by a section from B is completely written into the output. The option -v outputs the sections of A that are not (even partially) overlapped by any section of B. From [55].

2.11 Genome browsers and visualisation of the results

WormBase ParaSite WormBase ParaSite is a biological database; one can find information about genomes of nematode model organisms like *Caenorhabditis elegans*. Further, genomes of other related phyla like Platyhelminthes, to whom *S. mansoni* belongs, are stored. WormBase can be used as an information resource and as a place to publish and distribute results. Tracks can be added to the genome browser to visualise RNASeq results and different analysis approaches, e.g., different peak finding tools. In this visualisations, one could see whether a peak represents a ribozyme by analysing the belonging sequence of a peak in the browser [53].

USCS genome browser The UCSC genome browser is an interactive website offering access to genome sequence data from vertebrate and invertebrate species. It is built on a MySQL database and allows a fast visualisation and querying of data. It is possible to publish one's data on custom tracks that can be added alongside the original data. In this thesis, mapping and sequencing tracks were added to the data [36].

2.12 Implementation of the pipeline with Snakemake

The software Snakemake was introduced in 2012 by Köster and Rahmann [32]. It enables the automatisation of workflows for recurrent analysis tasks and, thus, reproducibility of analyses. The workflows are defined in terms of rules that describe how to create output files from input files. The workflow is written down in a so-called *Snakefile* that defines rules in a Snakemake specific definition language, which is an extension of the Python programming language. These rules can be generalized with named wildcards. Dependencies between different rules are determined by Snakemake and created as a directed acyclic graph (DAG) of jobs to enable a parallelisation and thus more efficient processing of all tasks. The package manager Conda ([14]) is integrated in Snakemake, thus the software *bedtools*, *UMI-tools*, *STAR*, *Infernal*, and *Piranha*, that were used in this thesis, could be executed with the Snakemake workflow. Additional external scripts for statistics written in R and Python and plotting could be integrated into the workflow. The task is to find a composition of rules for a set of targets in this thesis as the first step find a good ribozyme annotation for the investigated species (as described in section 2.4). The workflows are executed in three phases: Initialisation phase (parsing), DAG phase (building the DAG), scheduling phase (execute DAG).

2.13 Statistical analysis

Output files from *bedtools* and *cmsearch* were statistically analyzed using Python's statistic package and R. For distributions, medians, and 95% confidence interval were calculated. Additionally, for evaluation of *UMI-tools* output files, Pearson and Spearman correlation coefficients were calculated. It measures the linear relationship between two variables. The Pearson correlation can take values from -1 to 1, whereas -1 indicates a negative correlation, 1 a positive correlation, and 0 no correlation between two

parameters. It is calculated by the covariances (cov) of two variables divided by the product of their standard deviations (σ):

$$P_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

In contrast to that, the Spearman correlation coefficient is a nonparametric measure of rank correlation. It mirrors a statistical dependence between the ranking of two variables and how well a monotonic function can describe a relationship between two variables. The coefficient can take values from -1 to 1, as described before for the Pearsons correlation coefficient:

$$r_s = P_{rg_X, rg_Y} = \frac{cov(rg_X, rg_Y)}{\sigma_{rg_X} \sigma_{rg_Y}}$$

P denotes the usual Pearson correlation coefficient but applied to the rank variables, cov stands for the covariance of the rank variables, and σ are the standard deviations of the rank variables [28].

Additionally, local polynomial regression fitting (loess) is done in R for scatterplots of *UMI-tools* results. For that fitting, the fit at point x is made using points in a neighborhood of x, weighted by their distance from x [58].

2.13.1 Tools and programs

The Pipeline was created with Snakemake 5.24.2 [31]. The pipeline uses the following tools and programs in the RAP workflow: TrimGalore [33], bedtools 2.29.2 [54], samtools 1.3 [2], STAR 2.5.1b [15], UMI-tools 1.0.1 [49], FastQC 0.11.8 [5], Infernal 1.1.2 [34], Piranha 1.2.1 [35].

Evaluation and statistical analysis were done using Python 3.7.5 with Python's statistic module ([22]) and R 3.6.2 [21]. Plotting was performed with R's `ggplot2` [78].

The implementation of the pipeline RAP (version 0.1) is avaiable at `https://github.com/Tschi`

Chapter 3

Results

3.1 Semi-automatic ribozyme annotation

Generel observations In order to annotate all ribozyme types that may occur in all investigated species, *cm-search* was run with three different search options for all previously described ribozyme cm-models (section 2.3). Hits, regardless of the e-value threshold and with an e-value cut-off of 0.05, were quantified and compared.

The number of hits and ribozyme types was significantly lower in the investigated bacteria species than in *S. mansoni*. Since it is known from previous studies that many ribozyme sequences are present in the genome of S. mansoni, this was to be expected. Further explanations for detection of ribozyme types in relation to the composition of the alignments of the CM models and the investigated species can be found at the end of the section in subsection 3.1.3

In general, the mid search option showed most hits, but for some ribozyme types, e.g., Twister-P1 and Twister-P3, most hits were discovered by hmm search option in most species. An explanation for that fact could be that sequences in Twister ribozymes are stronger conserved compared to the other ribozymes, so the hmm search, which is based on sequence conservation, can detect more hits for ribozymes of this type.

For nearly all species, the hits found by *cm-search* were shorter in length in the hmm search than the hits from mid and std search. Ribozymes that were known before in the different species could be found in the *cm-search* in all species except DSM36. One can note that the hits from the mid and std search options were mostly as long as the corresponding CM-model. The filter stages in *cm-search* are based on HMMs and CMs. The three different search options compared in this thesis were hmm, which contains only the HMM-based filter stages, std, and mid search. The mid and std searches also include the CM filter stages, but with different filter settings. The hmm only setting should find hits with conserved sequences, while the additional CM filter stages also take the secondary structure conservation into account. Therefore one could assume that the hits found using mid or std search are more likely functional ribozymes than the hits from hmm search. The hmm search finds subsequences in ribozymes conserved, but it is not certain that these are functional or expressed. An indication that the hmm hits do not show complete ribozymes is the observation that the hmm search hits are usually shorter than the cm model. However, for its function, the ribozyme needs its entire length. Therefore the shorter hmm hits are worse than the hits from the std and mid search.

Detailed plots and tables of quantification of *cm-search* results with and without

an e-value cut-off are shown in Appendix A.

Details of interesting results for the individual species are described below.

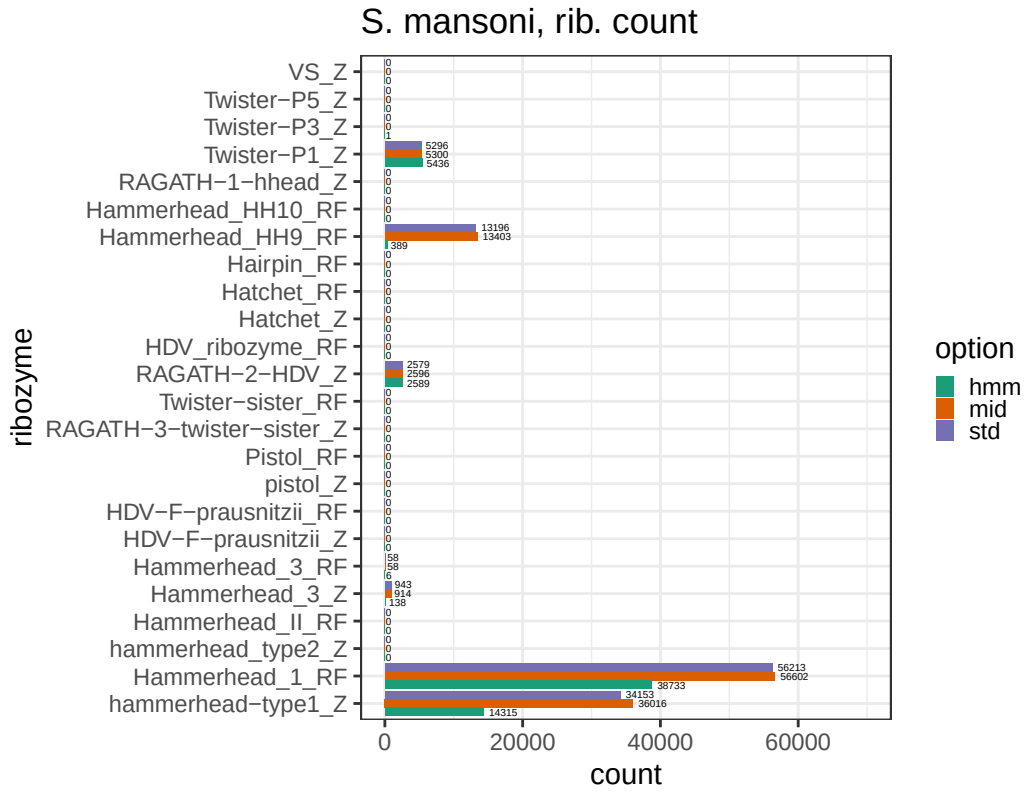
3.1.1 Schistosoma mansoni

cm-search in Schistosoma mansoni revealed most hits for HH-type1, HH-type3, and HH-9 ribozymes. Most hits are found with the mid-search option for most ribozyme types, but for some ribozyme types, hmm finds more, even significant hits. Examples are Twister-P1, and RAGATH-1-HH (Figure 3.1, Table A.1). But for this hits it must be taken into account that the median length even of the significant hits was lower compared to the hits from mid or std search option or the length of the CM models (Figure 3.2a, Table 2.1). Hits with an e-value below 0.05 are found for the following ribozymes (mid search option): Twister-P1 5300, RAGATH-2-HDV 2596, and HH_9 13403. For HH_3 and HH_type1 one could compare number of hits for the cm-models from Zand RF. It is noticeable that for HH-type 1, more hits were found with the rfam model for all search options and independent of using an e-value threshold. E.g., 56641 hits were found with the mid-search option in the model from RF, 36016 in the model from Z. For HH-type3, on the other hand, with the cm-model created by Z. Weinberg, more hits could be found: 914 vs. 58 with the mid-search option (Table A.1). No significant hits were found for the following ribozyme types: pistol, HH-type2, Twister-P5, RAGATH-1-HH, pistol, Hatchet, HH-10, Hairpin, and the HDV-ribozyme in *S. mansoni* (Figure A.1). It is also of interest to look at the percentage composition of hits found with different *cm-search* strategies. With a cut-off of 0.05 in e-value, it is markable, that the hmm search option found much fewer hits for the ribozyme types Hammherhead-9 and HH-type 3 than with the other two search options (Figure 3.1b).

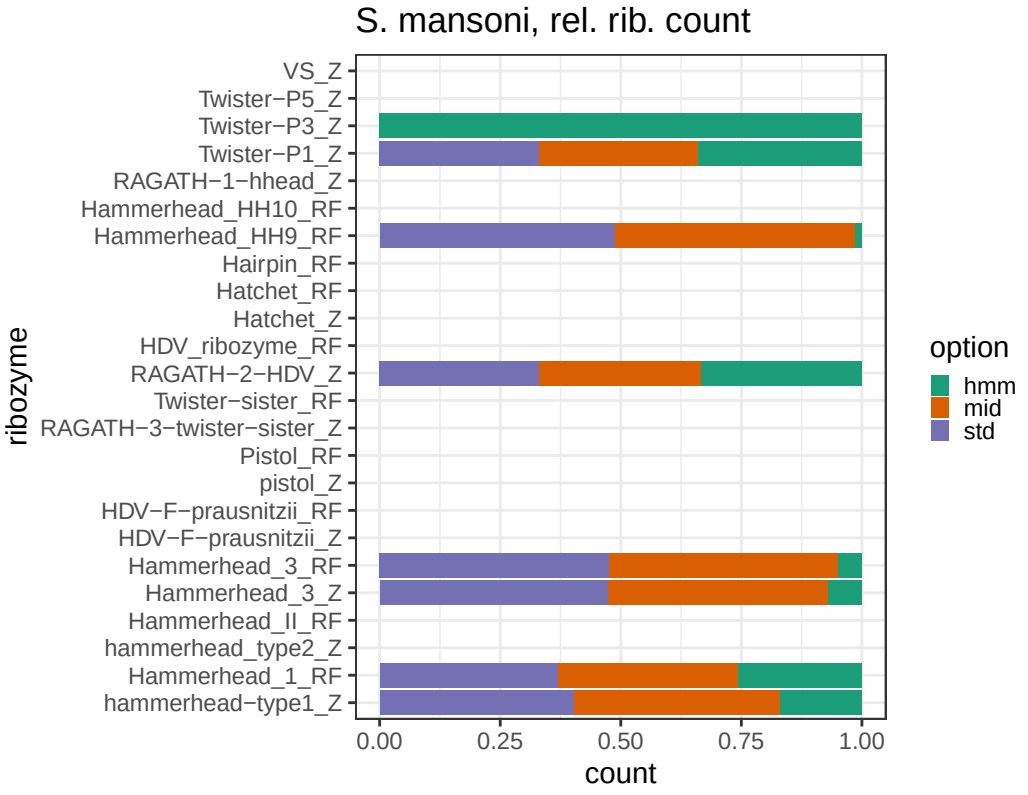
As mentioned before, the length distribution shows a longer median length for all ribozyme types with std or mid search option compared to the hmm option (Figure 3.2a). Together with the observation that the distribution of e-values shows the opposite picture (the median e-value in hits from the hmm search option is higher compared to the other two options, Figure 3.2b), assuming that hits from hmm search are only partial matches of ribozymes. Corresponding to this observation is the distribution of hits comparing e-value vs. length for the single ribozyme types (Figure 3.4). It shows no dependence on the length of the hit's e-value but on the used search mode. These observations will be discussed in more detail in chapter 4.

One could think that the number of hits found by a special *cm-search* option could be dependent on the alignment depth of the cm-model, where alignment depth means the number of sequences in a cm-model. So additionally, the relative distribution of found hits to alignment depth was calculated. One could expect that hmm search can find more hits in cm-models with fewer sequences because hmm hits were, on average, shorter for all ribozymes. The percentage of hits found with hmm search mode was higher for alignments with fewer sequences without a threshold, but this trend is not evident with an e-value limit of 0.05 (Figure 3.3). Thus, in Schistosoma mansoni analysis, this hypothesis could not be confirmed for significant hits. Hits from hmm search seem to be worse because the hmm search takes only the conserved sequence of the ribozymes into account, not the structure, but this is essential for the function of a ribozyme.

To date, HH-type1, RAGATH-2-HDV, and twister-P1 ribozyme types have been annotated in *S. mansoni*. The occurences of sequences of these ribozyme types could be confirmed in this thesis, and additional, significant hits for HH_3 and HH_9 could be found in the genome of *S. mansoni*. This was expected due to sequence and structure similarities of HH_1, HH_3, and HH_9.

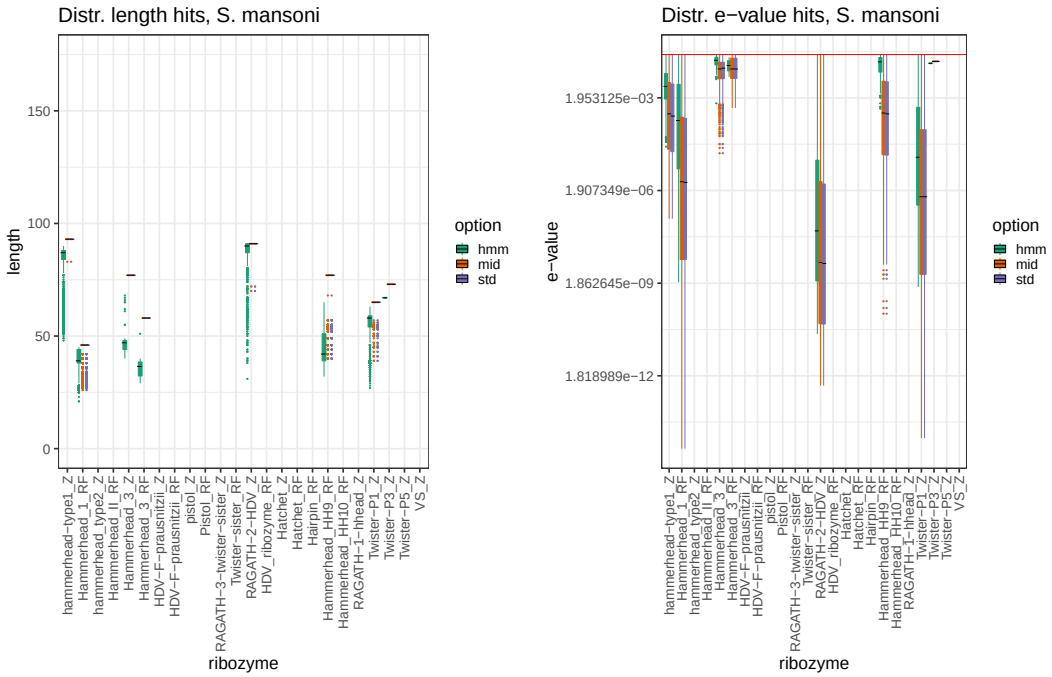


(a)



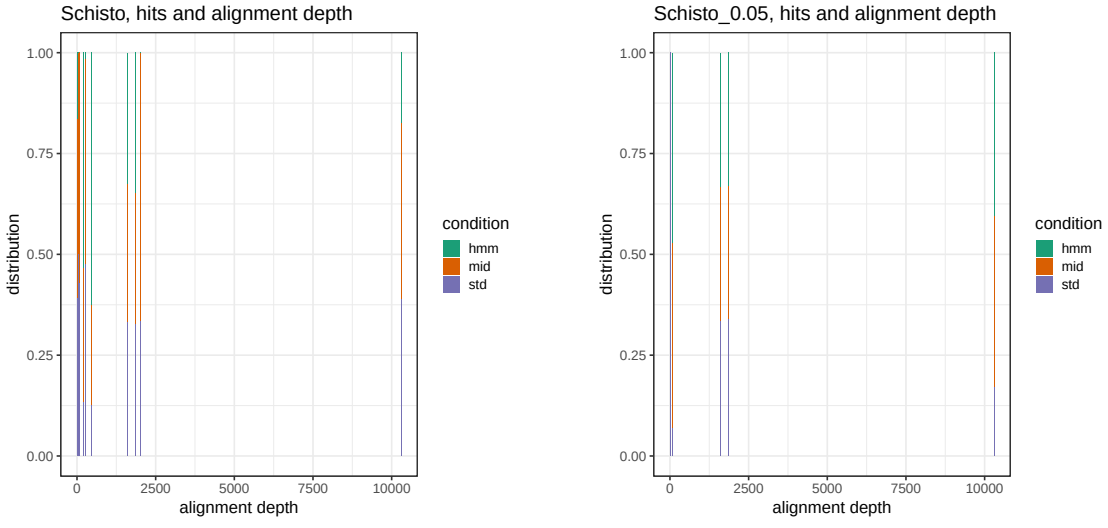
(b)

Figure 3.1: **Schistosoma mansoni: Ribozyme count, hits with e-value threshold 0.05.** Absolute count (a), relative count (b)



(a) (b)

Figure 3.2: **S. mansoni *cm-search* hits with e-value** < 0.05**.** Length distribution of the hits (a), e-value distribution of the hits (b).



(a) (b)

Figure 3.3: **S. mansoni: relative counts of hmm, mid and std for each ribozyme depending on the alignment depth.** (a) all hits, (b) significant hits.

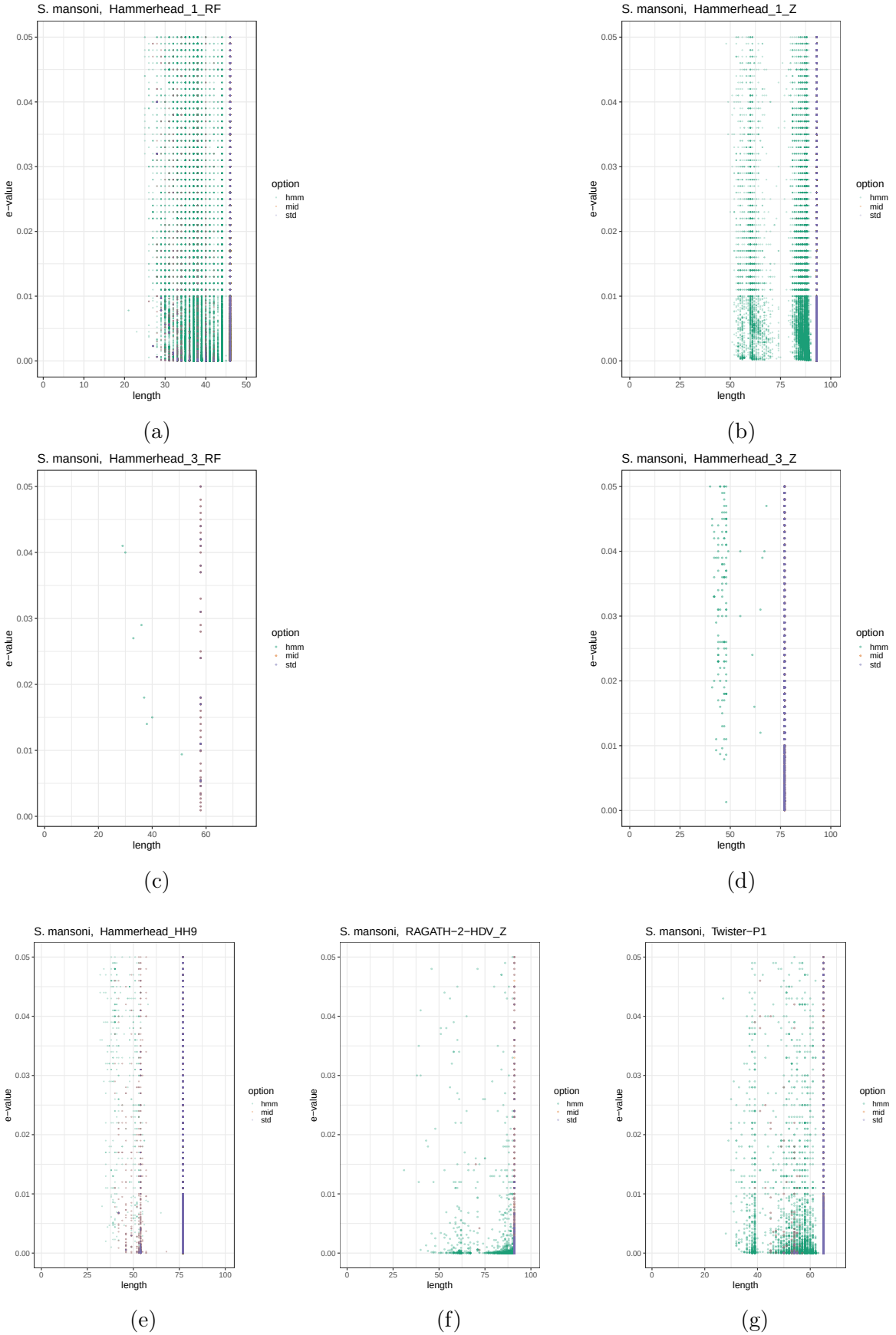


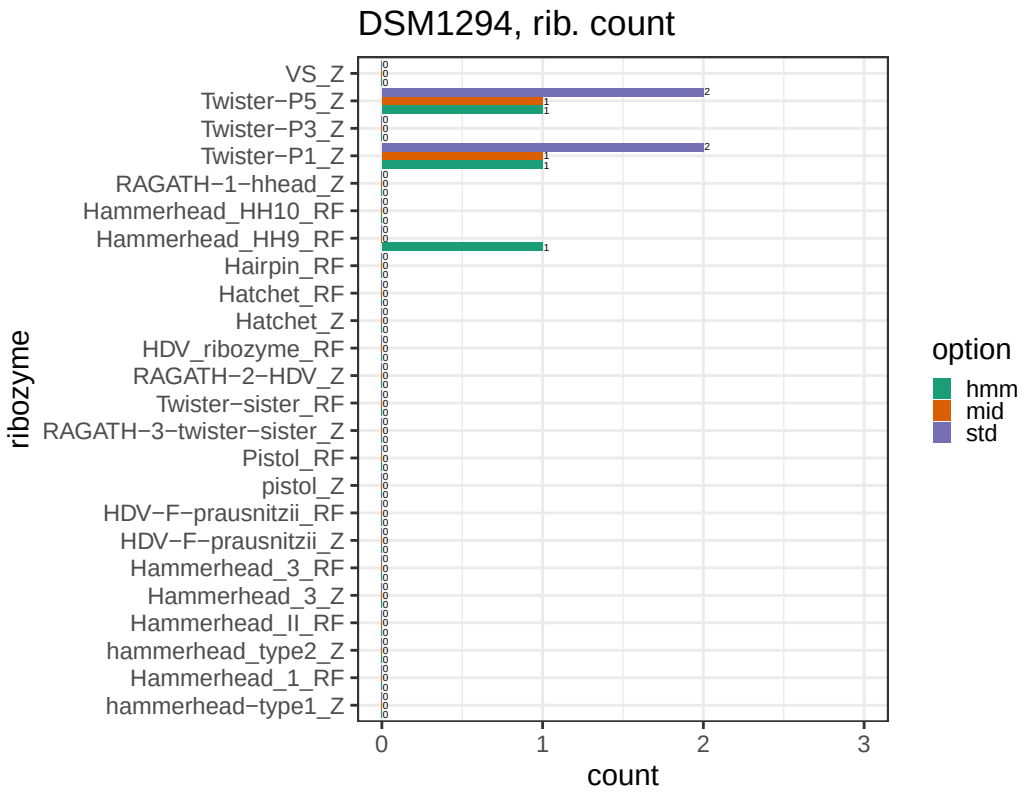
Figure 3.4: **S. mansoni: E-value vs. length for *cm-search* significant hits (E-value<0.05) in Schistosoma mansoni.** (a) HH_type1_Z (b) HH_type1_RF (c) HH_type3_Z (d) HH_type3_RF (e) HH_type9 (f) RAGATH2-HDV (g) Twister-P1

3.1.2 Bacteria species

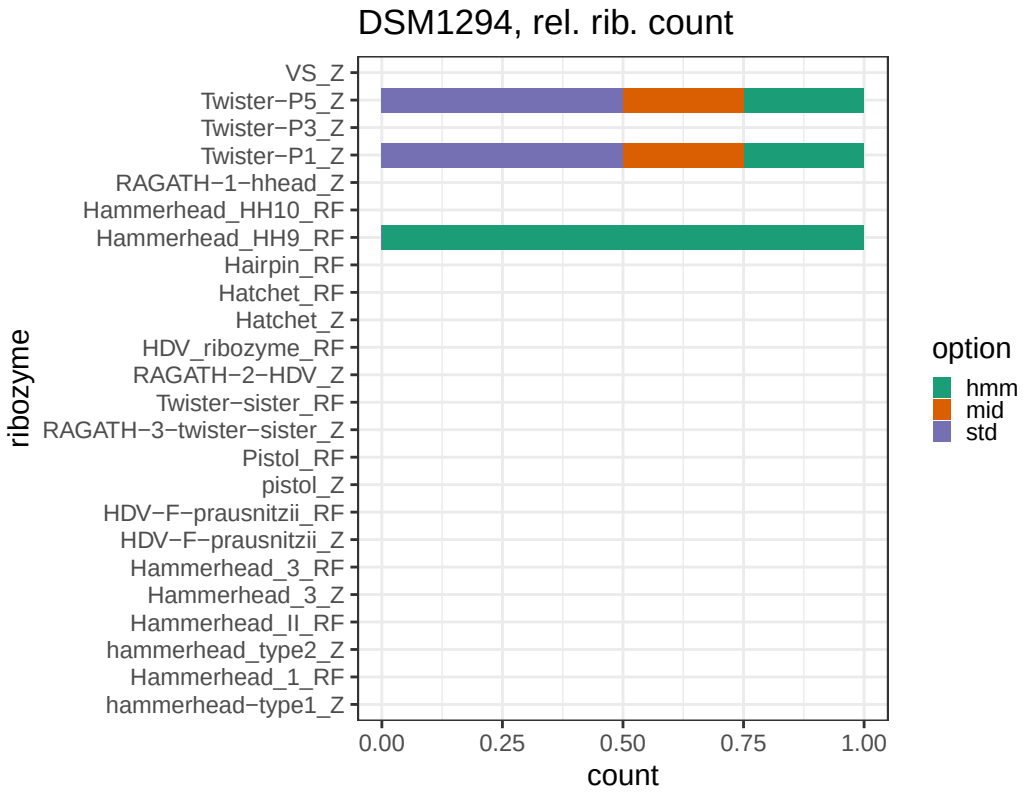
The ribozyme sequencing method was also performed in some bacteria species to answer whether the ribozyme sequencing method also works in organisms with a low known number of annotated and probably lowly expressed ribozymes. In the bacteria genomes from NCBI no ribozymes were annotated. In order to prove later whether a certain peak in the sequencing results belongs to an annotated ribozyme, the annotations of the *cm-search* with the ribozyme type cm-models were appended to the genomes.

In the following interesting aspects are presented, which were noticed during the search for a ribozyme annotation in the bacteria. The other results, as well as non-significant hits, are shown in Appendix A.

Clostridium sporosphaeroides - DSM1294 A low number of ribozyme hits was found in DSM1294 independent of the used CM-search mode (Table A.2). Significant hits could be detected for Twister-P5, Twister-P1 (two hits with std search options each), and HH-9. However, the hit for HH-9 was detected by the hmm search option only (Figure 3.5). In contrast to the other investigated species, the hits found with hmm search were not much shorter than the other search options' hits. Surprisingly for Twister-P1, the hits found with mid and std search options were shorter than the cm-model length. In contrast, the hmm hits were as long as the cm-model length (Figure 3.6a). The hits for the Twister-P5 ribozyme were as long as the cm-model. Without a threshold in e-value, there is generally a low percentage of hits with the hmm *cm-search* option, but it changes while integrating an e-value threshold of 0.05. However, it should be noted that there were very few hits for this option overall (Figure 3.7, Figure A.3a). As mentioned before, one could assume that sequences in Twister ribozymes are stronger conserved compared to the other ribozymes. In DSM1294, mainly Twister ribozymes were detected, so in this species, hmm search option is even better compared to mid or std search. Whether these hits are biological relevant is to be proven by further sequencing experiments.
It was former known that in DSM1294 Twister-P5 sequences can be found in the genome, but the occurence of Twister-P1 is a new discovery in the study.

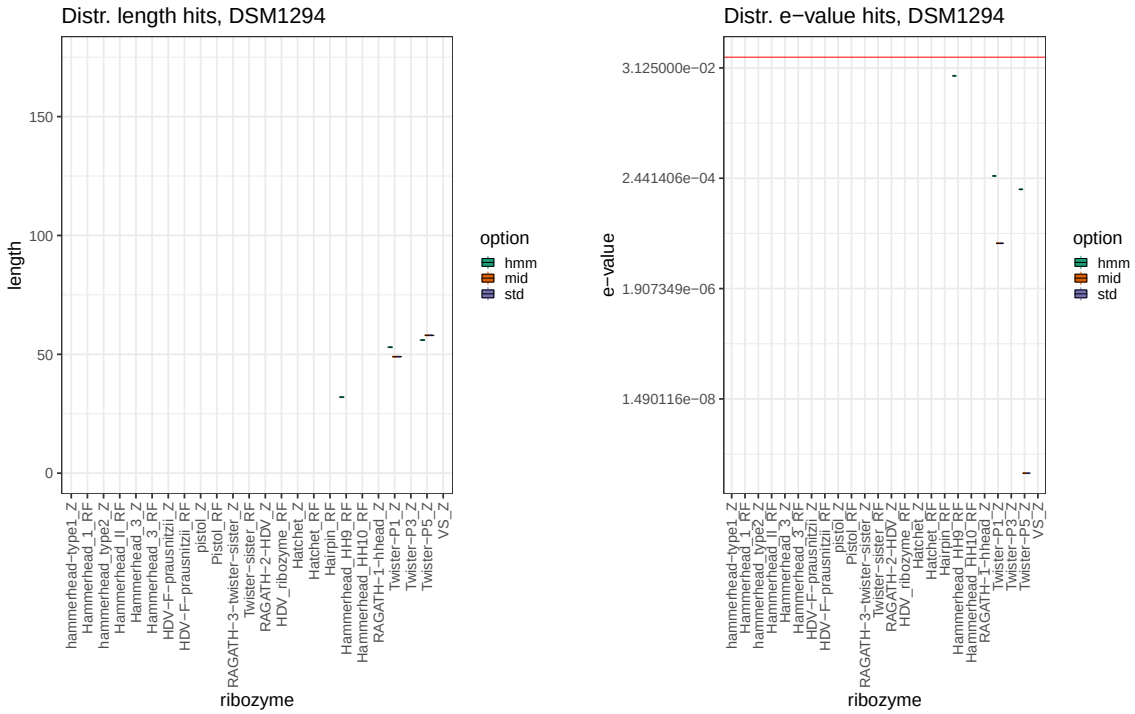


(a)



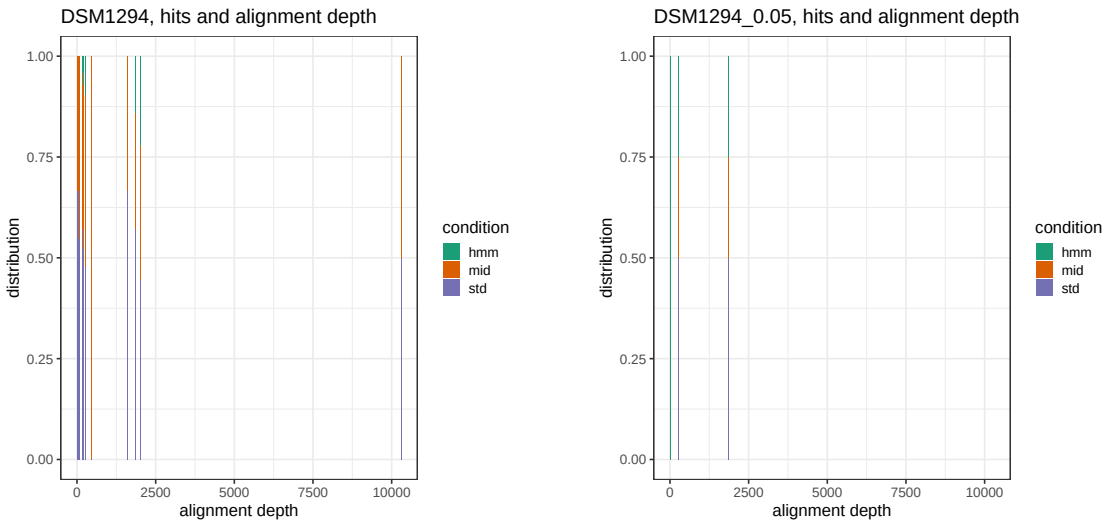
(b)

Figure 3.5: **DSM1294: Absolute and relative count for *cm-search* hits with different search options with e-value ≤ 0.05.** Absolute count (a), relative count (b).



(a) (b)

Figure 3.6: **DSM1294: Distribution of length and e-value of *cm-search* hits with e-value** ≤ 0.05**.** (a) length (b) e-values.



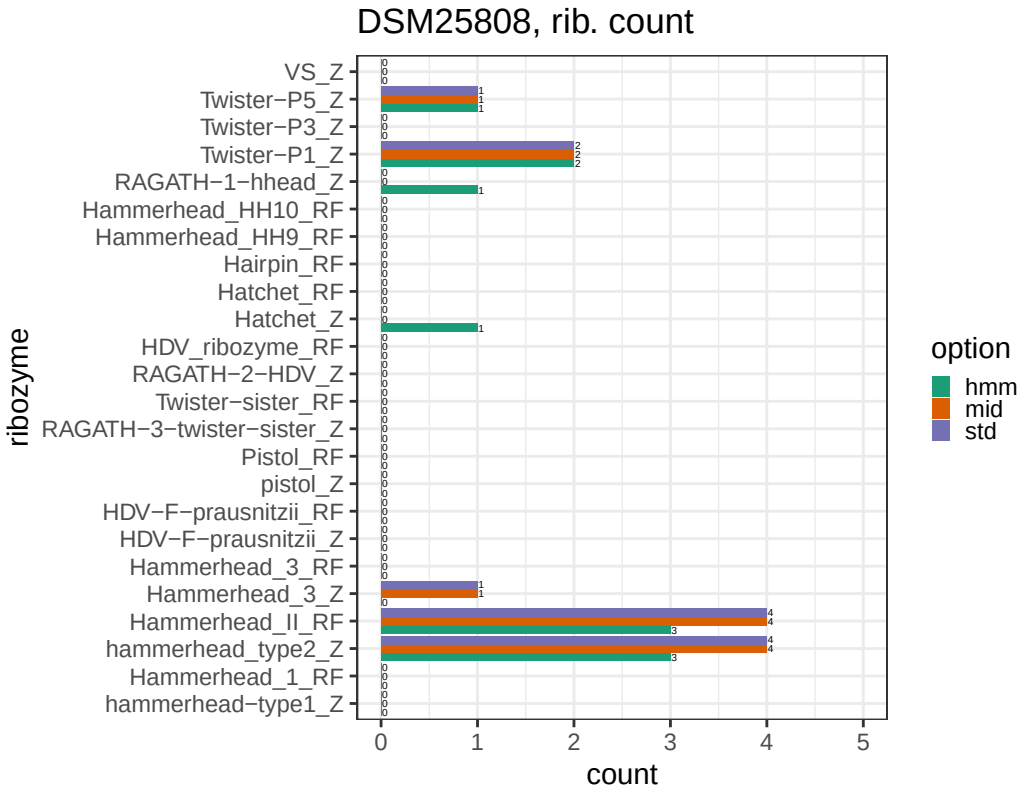
(a) (b)

Figure 3.7: **DSM1294: Alignment depth to procentual hits found by hmm, mid and std *cm-search*** Without a threshold in e-value (a) vs. e-value ≤ 0.05 (b).

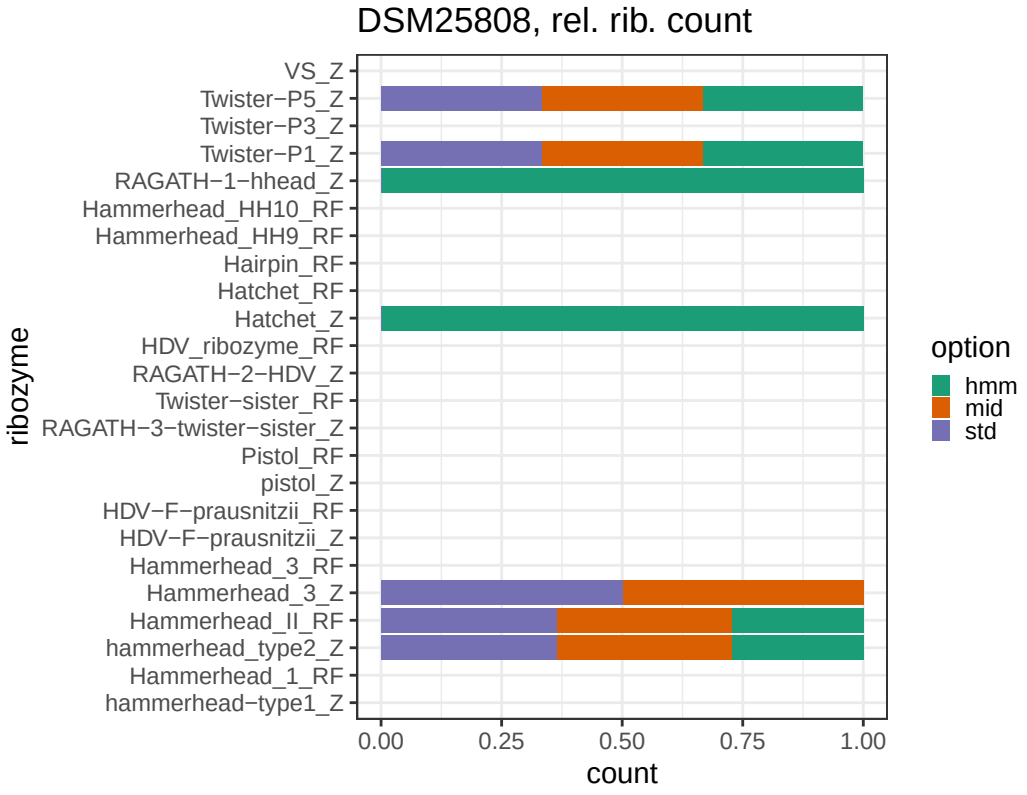
Fervidicella metallireducens - DSM25808 In F. metallireducens (DSM25808), hits in seven cm-models were found with significant e-values, wherby two of them (RAGATH-1-HH, Hatchet) were only detected with one hit by hmm search. The other detected ribozyme types were: Twister-P5 and HH-3, Twister-P1 and HH-type2 (Z and RF; Figure 3.8, Table A.3).

Evaluation of the length distribution of found hits showed in median shorter hits for hmm search, whereas mid and std search had no differences in median length (Figure 3.9a). Also, hmm hits have higher e-values than hits found with mid and std search (Figure 3.9b). The scatter plots in Figure 3.10 show the ratio of length to the e-value of hits in the textitcm-search. For significant hits, e-values of hmm search hits were higher than those from mid or std search, but there is no difference in length of the hits. The length of most hits is as long as the length of the underlying cm-models, but the hmm hit in Twister-P5 was shorter. This trend could be seen for most of the hmm search hits in all species.

With *cm-search* it was possible to build annotations of occurrences from Twister-P1 and HH-type2 in the genome of DSM25808. Moreover, in this study, significant hits were found for Twister-P5 and HH_3 with all search options.

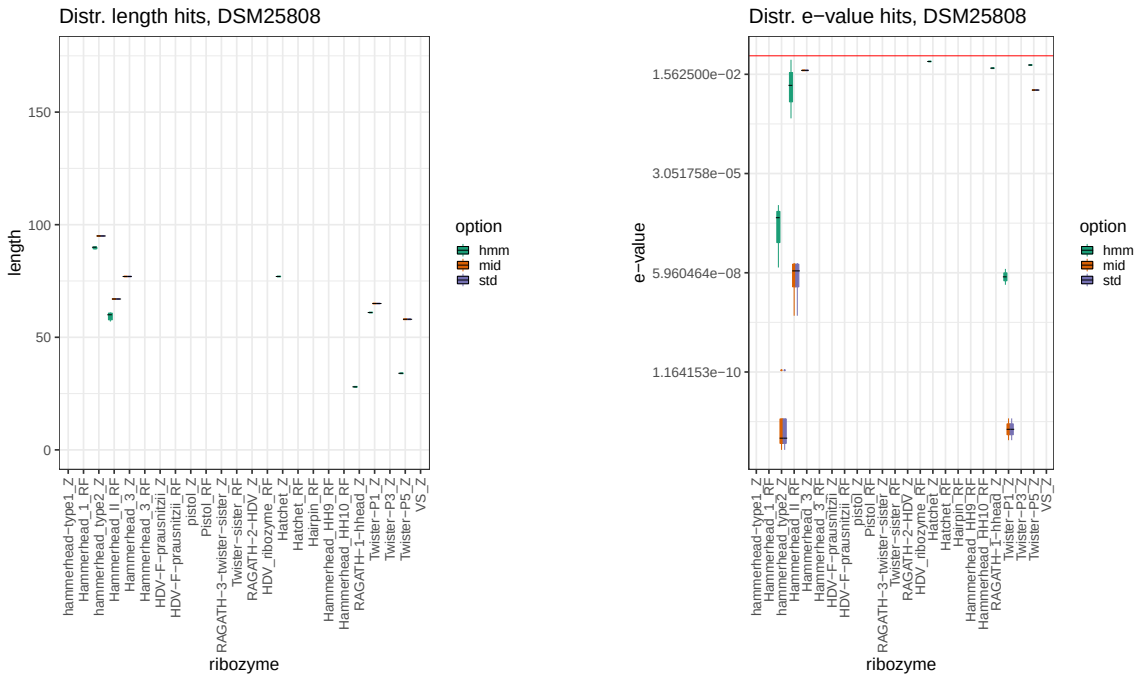


(a)



(b)

Figure 3.8: **Hits in *cm-search* for F. metallireducens (DSM25808).** Hits with different search options, e-value ≤ 0.05. (a) absolute (b) relative.



(a) (b)

Figure 3.9: **Distribution of length (a) and e-value (b) of *cm-search* hits with e-value ≤ 0.05 in DSM25808.**

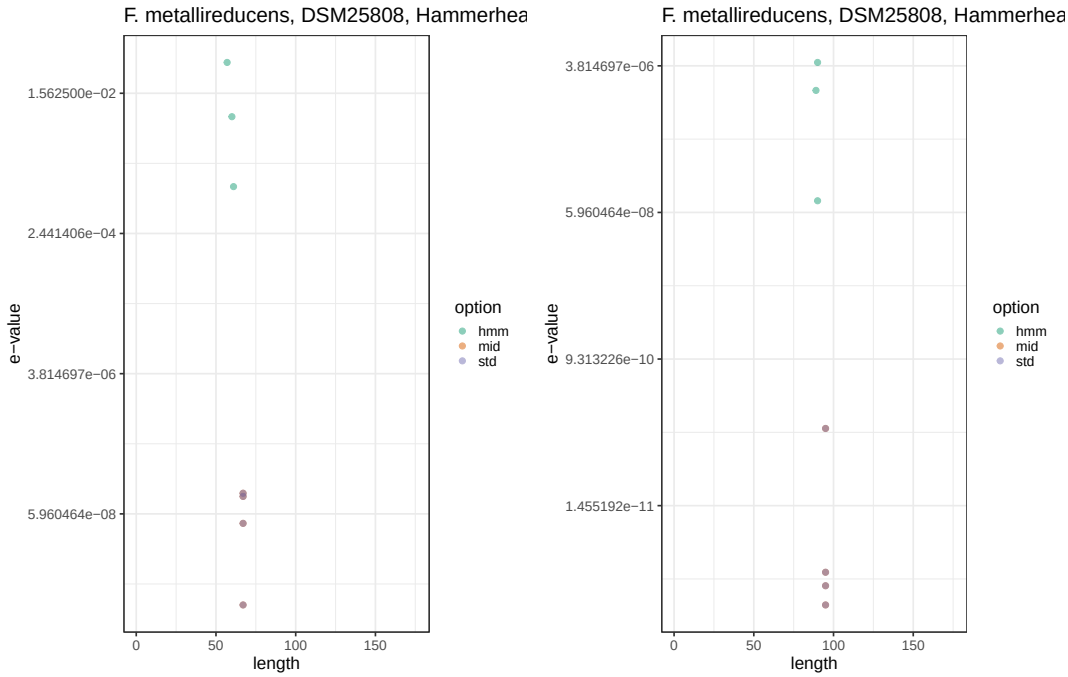
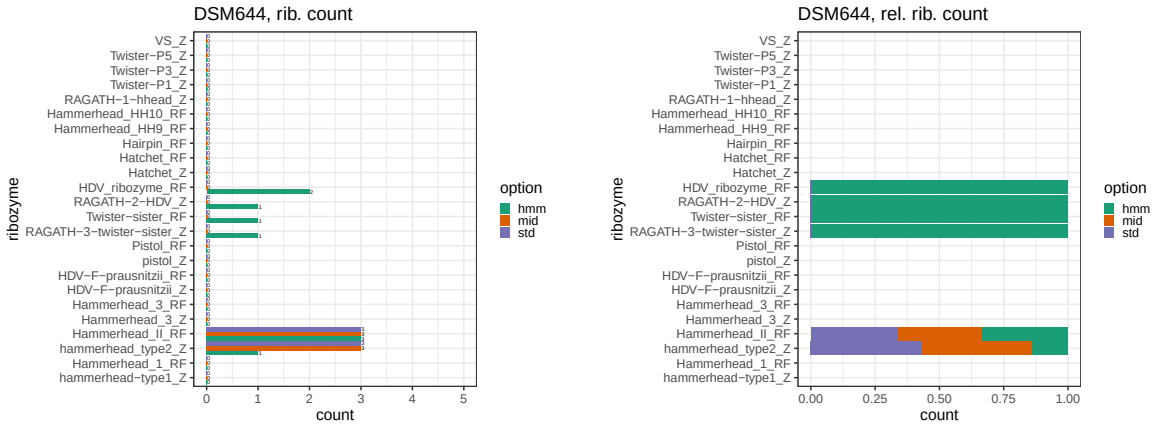


Figure 3.10: **E-value vs. length for ribozyme hits (e-value ≤ 0.05) in *cm-search* in F. metallireducens, DSM25808.** Left: HH-type2_RF, right: HH-type2_Z

Paenibacillus polymyxa - DSM36 It is described that in P.polymyxa, the pistol ribozyme sequence can be found in the genome. However, in this study, no significant hit for this species could be found for any ribozyme type with any search option. The "best hit" for the pistol ribozome with mid or std search had an e-value >2.0 and can therefore not be considered a significant hit. Therefore, this section will not discuss in detail the results of the search without e-value cut-off. The visualization of these results can be found in Appendix A.

Desulfovibrio vulgaris - DSM644 As espected, in D. vulgaris (DSM644) hits for HH-type2 could be fund with all search options. Additionally, with the hmm option there were hits for HDV-ribozyme, RAGATH-2-HDV, twister-sister, and RAGATH-3-twister-sister found (Table A.5). All these hits are shorter compared to the length of the cm-models (Table 2.1, Figure A.13). In both cm-models (Zand RF) for HH-type2 *cm-search* found 3 hits with std and mid search (Figure 3.11). Length of the hits did not differ between the different search options for HH-type2 ribozymes, but in e-value (Figure 3.12).



(a) (b)

Figure 3.11: **DSM644: Absolute and relative hit count for found ribozymes with *cm-search* and an e-value threshold of ≤ 0.05.** Hits with different search options. Absolute count (a), relative count (b).

Desulfitobacterium dehalogenans - DSM9161 Although it is already known that hits from the Twister-P5 ribozyme can be found in the genome of DSM9161, this study also found significant hits for Twister-P1, RAGATH-2-HDV, and HH type1-3 (Figure 3.13, Table A.6). The different HH ribozyme types were found with up to two hits, respectively, depending on the search option. Twister-P5 could be detected up to eight times.

Similar to the results from the other investigated species, hits found with the hmm option were shorter and had a higher dispersion in length, the e-value distribution was comparable to those in the other species (Figure 3.14). For significant hits, the ribozymes' length was as long as the length of the belonging cm-model. For Twister-P1, the length was shorter than the length of the cm-model, but one should keep in mind that this hit was done with the hmm search option.

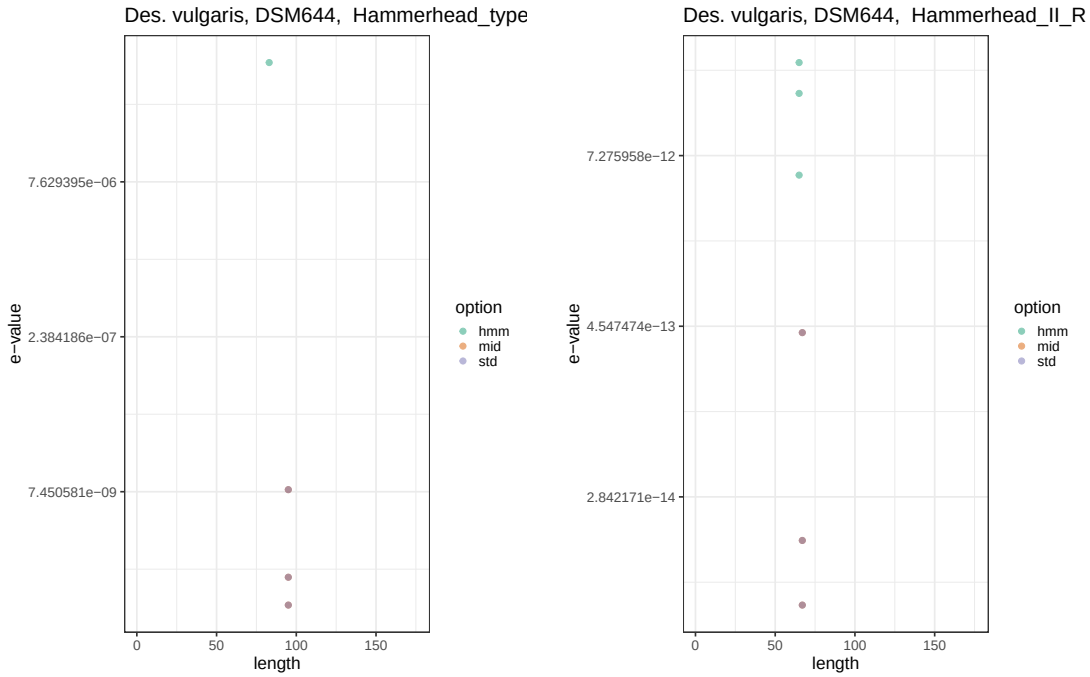


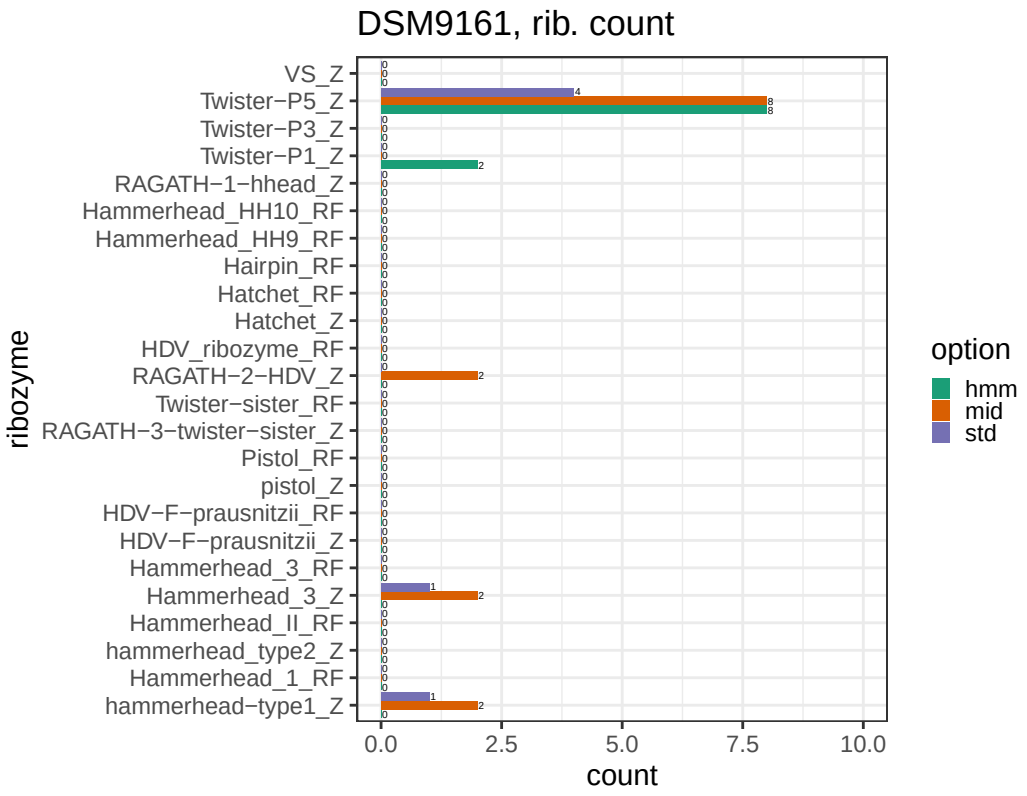
Figure 3.12: **E-value vs. length for significant (E-value $<$ 0.05) hits at *cm-search* in D. vulgaris (DSM644)** Left: HH-type2_Z, right: HH-type2_RF

3.1.3 Summary *cm-search*

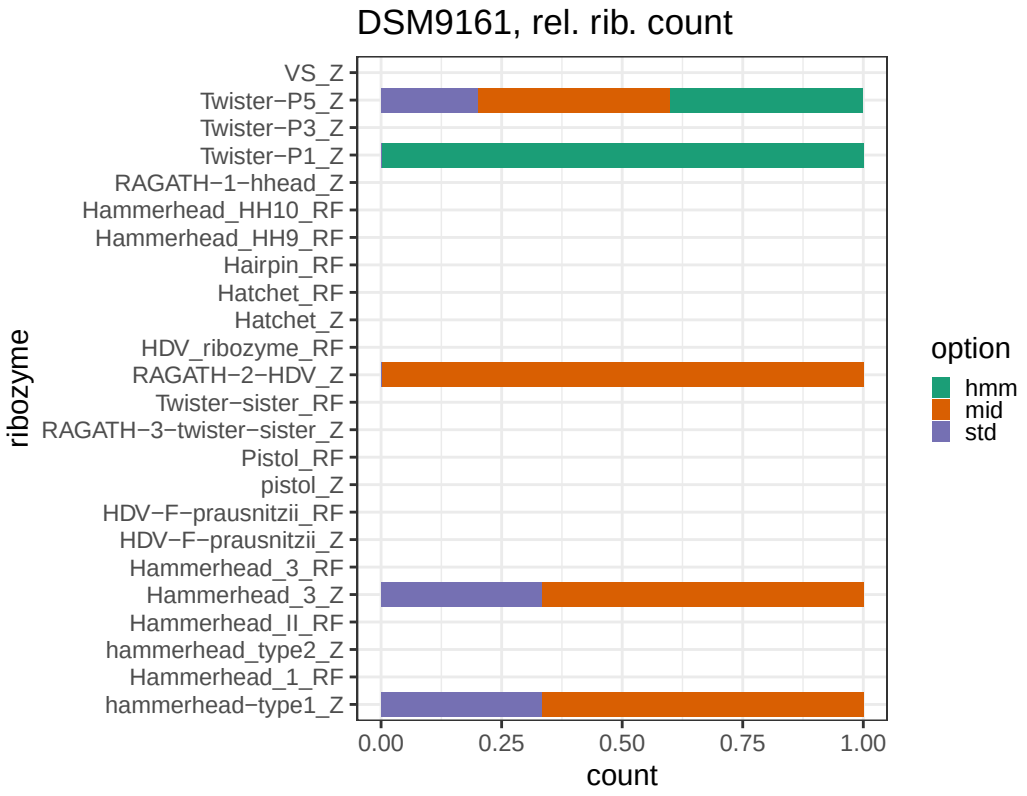
Taken the results of *cm-search* together for all bacteria species except P.polymyxa (DSM36), the expected ribozyme species (Table 1.1) could be found with mid and std search options. Moreover, other ribozyme types could be detected with significant e-values and expected lengths. This is also true for *S. mansoni*.

It can be assumed that the composition of the sequences in the alignment affects the finding of hits in the investigated bacterial species. The alignments of Twister-P1, HH_9, HH_3, and HH_I consist mainly of sequences from eukaryotes, those from Twister-P5, Pistol, and HH_II of sequences from bacteria. Thus, it is not surprising, that Twister-P1, HH_9, HH_3, and HH_I could be found with significant hits in *S. mansoni*. These hits were also about as long as the used cm-model. There were also hits for Pistol and Twister-P5 in *S. mansoni*. They were from the same length as the corresponding CM model, but without a significant p-value. This could be due to the composition of the CM models with mainly bacteria sequences, but it is also possible that these ribozymes are not part of the genome in eucaryotes or only in few species (Figure A.2).

Surprisingly, in DSM1294 Twister-P1 could be detected with significant hits with a length corresponding to the length of the CM model, although the corresponding cm-model is mainly composed of sequences from eukaryotes. Other ribozymes could also be detected. The hits from mid and std search were as long as the CM model length, but they do not have a significant p-value (Figure A.4a). It seems that the composition of CM models do not influence the length of the hits in *cm-search* found with std or mid search options, but the significance of the hits. This is not particularly surprising since *cm-search* primarily examines the conserved structure of an RNA and not primarily conserved sequences. Ribozymes need the correctly conserved secondary structure more than one conserved sequence.

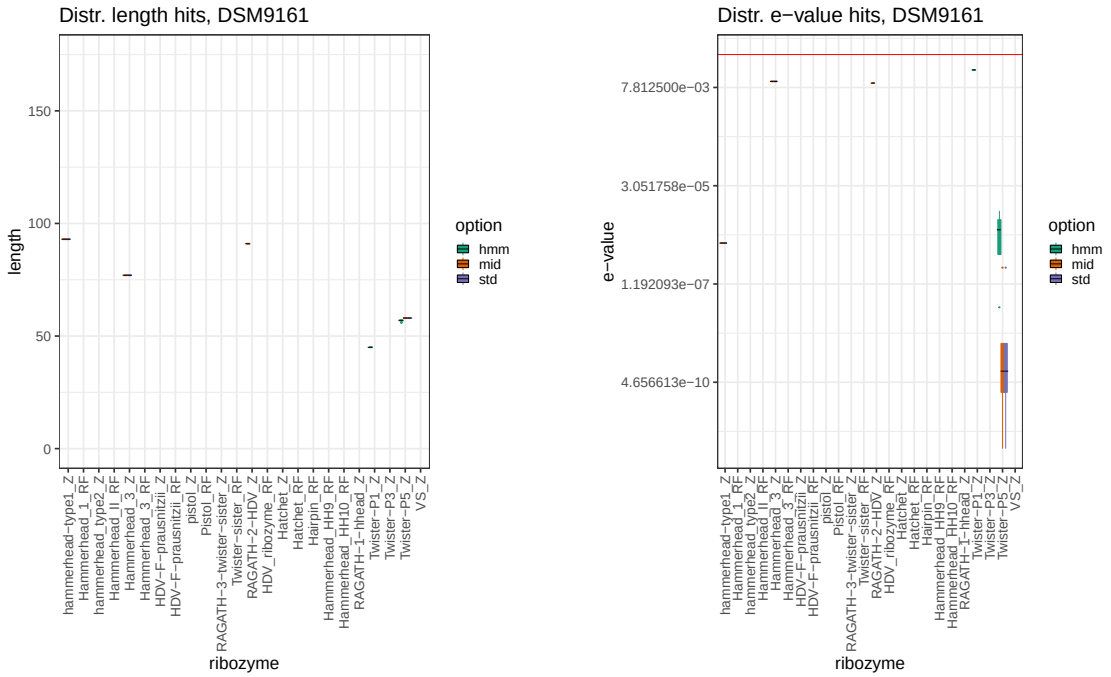


(a)



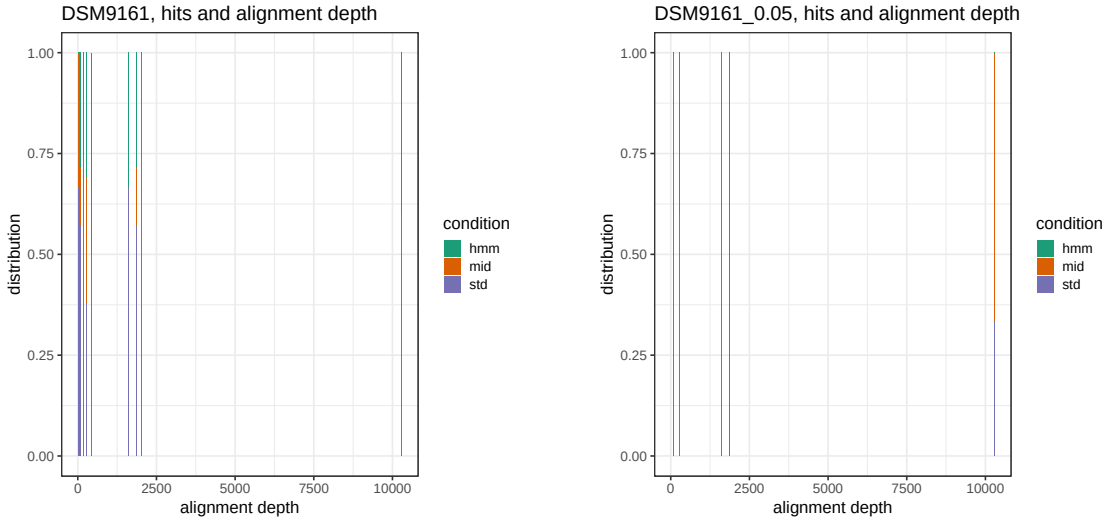
(b)

Figure 3.13: **DSM9161: Absolute and relative count of ribozyme hits with *cmsearch*.** Hits with different search option and e-value ≤ 0.05. (a) absolute, (b) relative count.



(a) (b)

Figure 3.14: **DSM9161: length and e-value distribution of ribozyme hits with *cm-search* with an e-value ≤ 0.05.** Length (a), e-value (b).



(a) (b)

Figure 3.15: **DSM9161: Alignment-depth of cm-models to relative hit count.** All (a) vs. hits with an e-value ≤ 0.05 (b).

3.2 Intersection analysis

The hits in the output of the different *cm-search* runs were compared with the tool *bedtools intersect*. On the one hand, one could see which number of cm-search hits is not covered by hits from another search option. On the other hand, out of the statistics from the cm-search, one could see that hits from hmm search are shorter than hits from mid and std search, so that one wants to exclude the hypothesis that two hmm hits cover one hit in mid or std search.

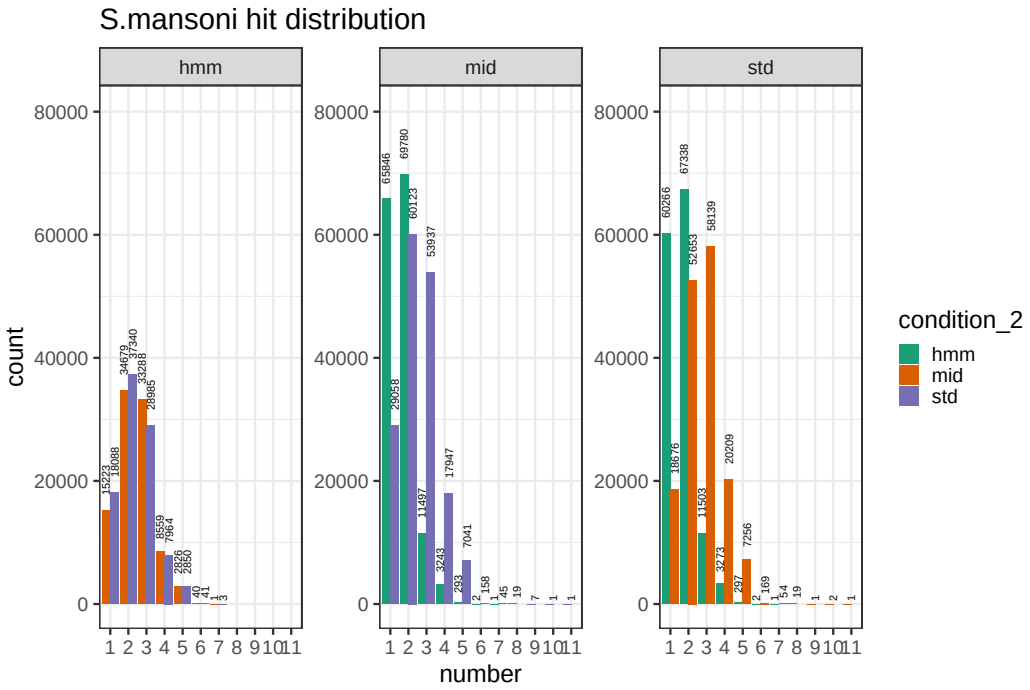
For all investigated species, there was not a high number of hits in std search, that were not covered by mid search option.

Schistosoma mansoni In *S. mansoni*, most hits are covered with one to three hits from other search options. It should be noted that hits from mid and std searches are most frequently covered by one , some by two hits out of the hmm search (Figure 3.16). This fact is surprising because of the smaller mean length of hits in the hmm search (Figure 3.2a). Around 4000 hits from hmm search are not covered by mid and std hits, whereas around 23000, respectively 18000 hits from mid and std search are not covered by hmm search. Mid search hits only missed 88 hits from std search (Figure 3.16b). The distribution to single ribozyme types shows, that hmm hits for Hairpin, HDV-ribozyme, Pistol, Hatchet, HH_II, HH_10, RAGATH_I_HH are not covered with hits from mid or std search. Again it should be noted, that the mean length for hmm hits is shorter than for std or mid search option, but mid or std hits do not cover more hits of hmm for this ribozyme types (Figure B.1). One exception here is HH_3, for which two hits from hmm search cover most hits of mid and std search (Figure B.1 top row right, purple and green bar).

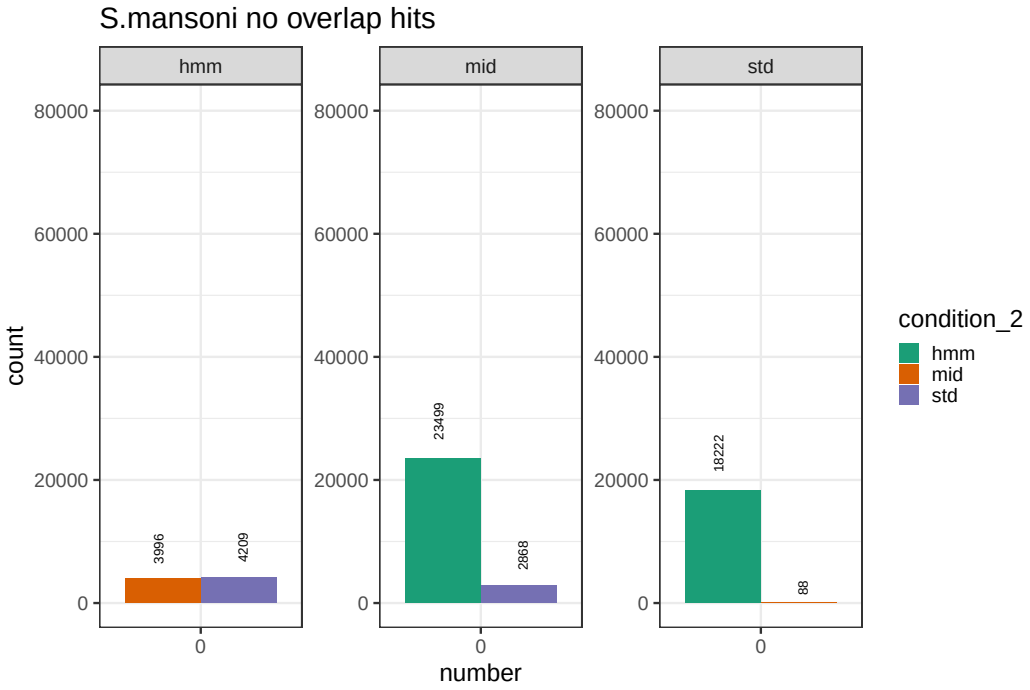
Taken together the results of the intersection analysis exclude the hypothesis that hits from mid or std search are mostly covered by two hmm hits.

Bacteria species Bedtool intersect analysis of the cm-search hits showed similar results in all five Bacteria species. There are small relative proportions of hits in hmm search that are not covered by hits from mid and std *cm-search*. On the contrary, the hits from hmm search do not cover most of the mid and std search hits (Figure 3.17). Most hits that are covered by hits from another search option are covered by one hit out of this. So certain search modes of cm-search do not find several parts of the same annotated ribozyme (Figure B.2).

Summary intersection analysis Mid and std options have cm-search steps included, while hmm search option consists only of hmm steps, which do not take a conserved secondary structure into account. These results are not surprising since the structure is essential rather than a conserved sequence for ribozymes' function. Taken together, one can conclude that for building the analysis pipeline, std or mid search option seem to be sufficient for ribozyme annotation in species genomes.



(a)



(b)

Figure 3.16: **S. mansoni: number of hits dependend on the *cm-search* mode.** A *bedtools intersect* results: Intersection of search option 1 und 2. option 1 is written on the plot, option 2 in the legend. B Number of hits not covered by the other search option.

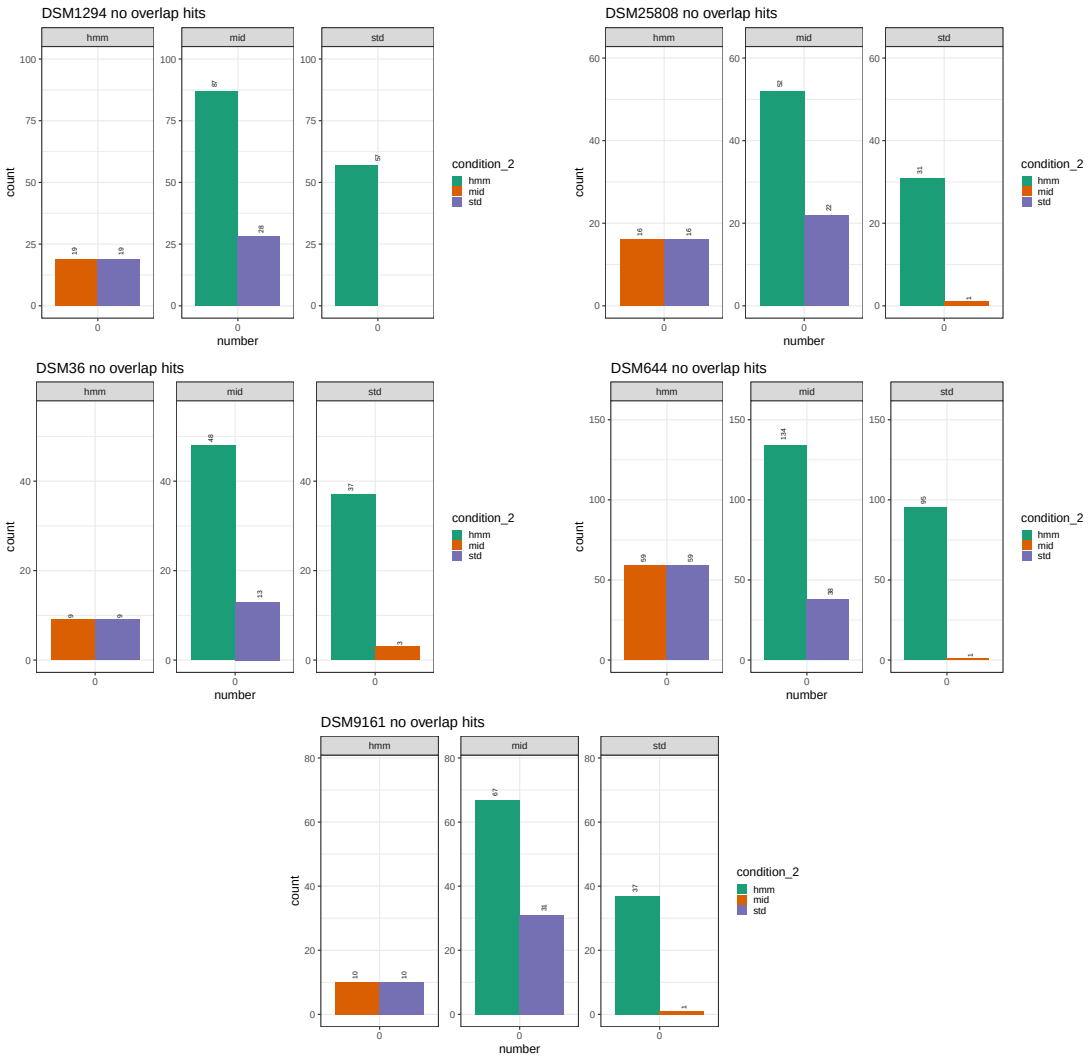


Figure 3.17: **Bacteria: Count of not overlapping hits depending on the *cmsearch* condition.** From upper left to down right: DSM1294, DSM25808, DSM36, DSM644, DSM9161

Intersection of the official annotations with the ribozyme annotations Since no other ribozyme types apart from HH ribozymes have been added to the genome of *S. mansoni* so far (Figure 3.18), one should get a deeper insights into the *S. mansoni's* genome in the wormbase browser. For that example sites for each not annotated ribozyme types (twister-P1 ,RAGATH-2-HDV, HH-3) were visualized to prove, whether it could be a valide ribozyme hit. For twister-P1 and RAGATH-2-HDV the specific cleavage sites could be shown in the example sites (Figure 3.20, Figure 3.21). The annotation site of the HH-type3 ribozyme overlays the annotation of HH-type1 from the WormBase ParaSite (Figure 3.19). This may be due to the sequence similarity of the different HH ribozymes.

In Figure 3.18 one can see that for HH_I around 1/3 of the hits from *cm-search* the ribozyme was annotated in the genome from WormBase ParaSite yet, whereas nearly all annotation sites of HH_9 overlap. The missing annotation sites of ribozymes in the WormBase ParaSite genome were distributed to protein-coding, low complexity, repeat, exon regions, and moreover to miRNA, rRNA, tRNA, and mRNA. For each of these cases, example sites were proven. In all these cases, cleavage sites of the ribozymes could be found within this annotation sites. It might be a hint that the new annotation sites should be integrated into the genome's annotation.

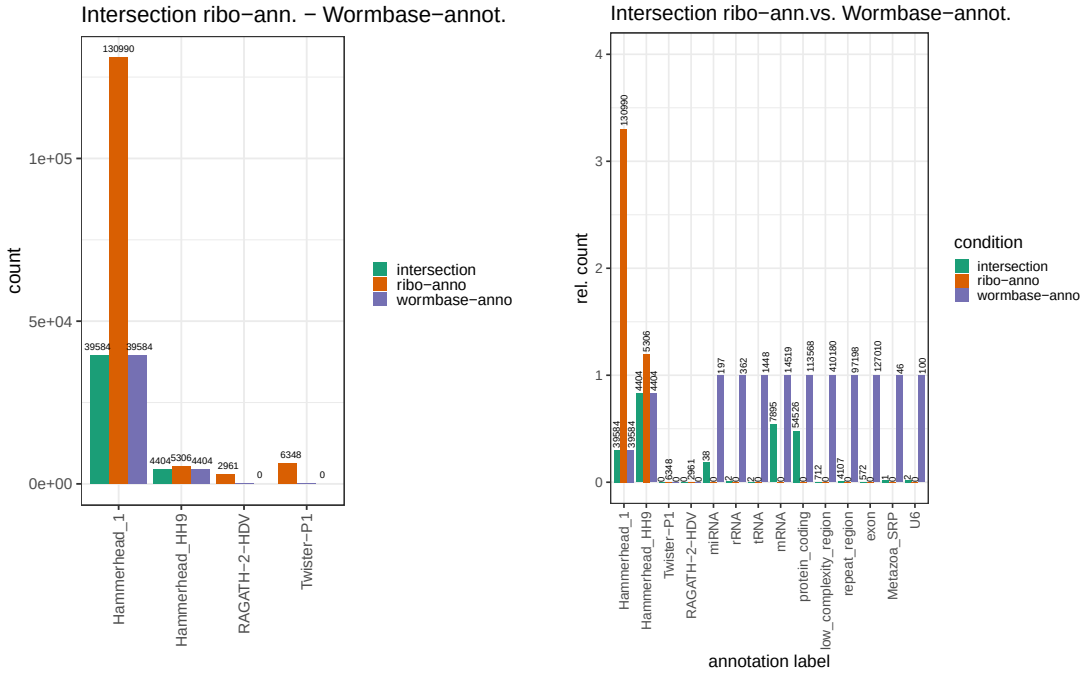


Figure 3.18: **Intersection of the ribozyme annotation with the official Wormbase annotation from *S. mansoni*** Left: Absolute count of hits from intersection of ribozyme-annotation with the official wormbase annotation. Right: Ratio of hits from Intersection ribo-anno vs. Wormbase-anno to either ribozyme annotation (for ribozymes) or official wormbase annotation (labels other than ribozymes).

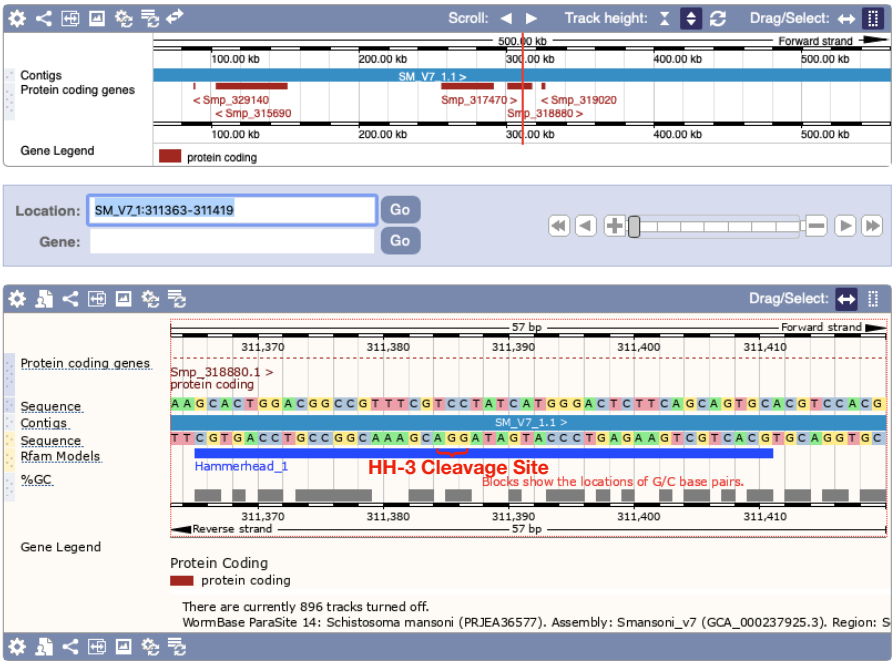


Figure 3.19: **Detailed visualisation of a Hammerhead-type3 annotation site in*S. mansoni* genome.** Hammerhead-type3 annotation out of *cm-search*, but Hammerhead_1 annotation in the WormBase annotation. The Hammerhead specific cleavage site could be found in the middle of the annotation. From [53].

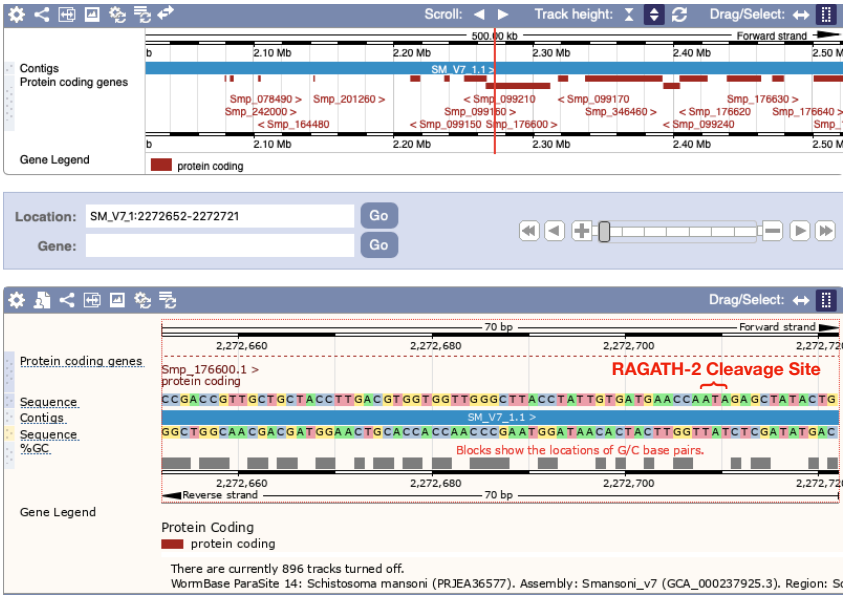


Figure 3.20: **Detailed visualisation of a RAGATH-2-HDV annotation site in *S. mansoni* genome.** RAGATH-2-HDV annotation out of *cm-search*, that is not annotated in the WormBase annotation. The RAGATH-2-HDV specific cleavage site could be found within the annotation site. FROM [53].

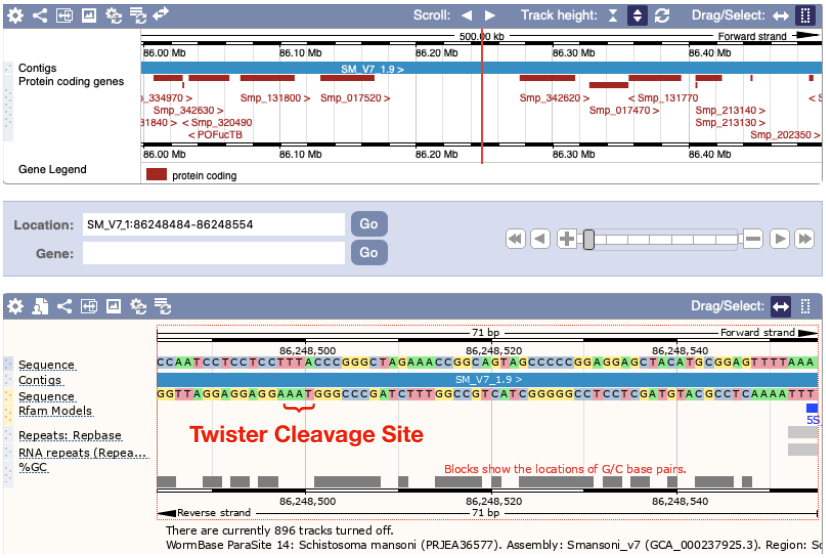
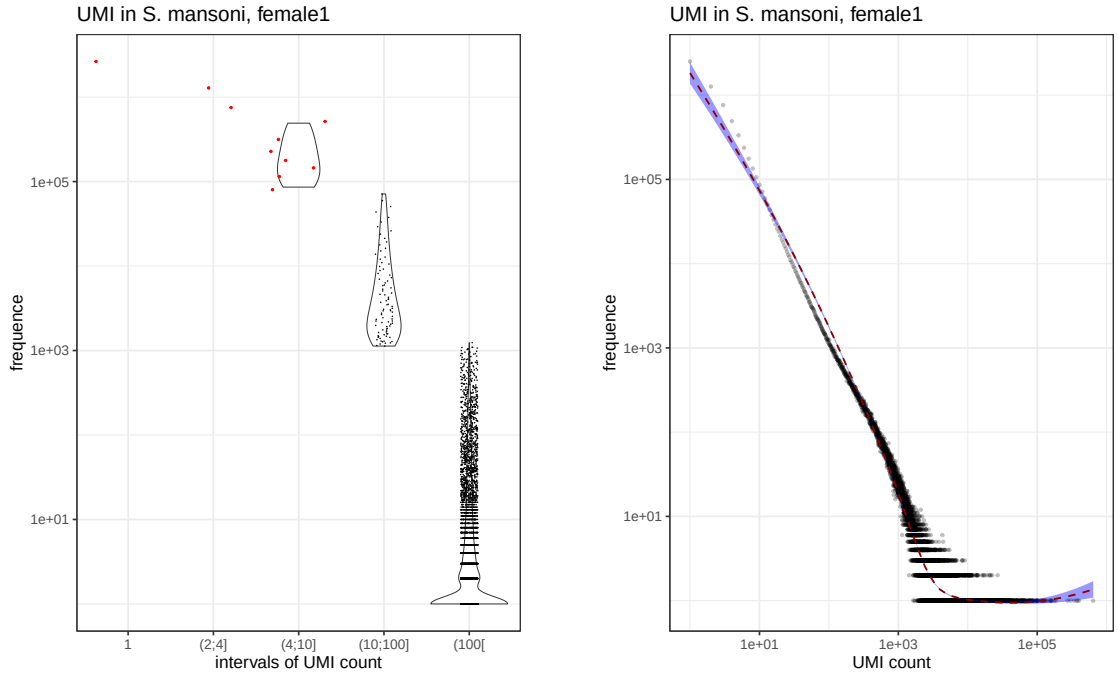


Figure 3.21: **Detailed visualisation of a Twister-P1 annotation site in *S. mansoni* genome.** Twister-P1 annotation out of *cm-search*, that is not annotated in the WormBase annotation. The Twister-P1 specific cleavage site could be found at the end of the annotation site. From [53].

3.3 Dedup and count UMIs

To exclude high replication of reads during PCR steps, UMIs were integrated into the adapter sequence for RibozymeSeq experiments, as described previously (section 2.7). UMIs were extracted with UMI-tools after sequencing and before mapping of the Illumina reads. Due to the marking of duplicated reads with a specific character combination during the UMI-tools extraction step, frequency analysis was possible. The results were similar for all samples, so for example, *S. mansoni* female1 results are shown below. In Figure 3.22a one can see the distribution of UMI frequences. Unique read-UMI combinations are most frequent, whereas read-UMI combinations, that occur 1000times and more are rare (Figure 3.22b). The results show the importance of using deduplication after mapping to avoid false high peaks in the further RibozymeSeq analysis. Testing with Spearman rank correlation test was done for all samples. There was a negative correlation between occurence of read-UMI correlation and their corresponding frequency. Results are summarized in Table 3.1.



(a) (b)

Figure 3.22: **Frequence of UMI in sequencing data of *S. mansoni*, female 1.** A Violin plot of frequence of binned occurence of read-UMI combinations. B Scatter plot of frequence of occurence of read-UMI combination together. In purple local smoothing of the values.

Table 3.1: **Spearman rank correlation coefficient for number of UMI-read combination vs. their frequences**

sample	ρ-value	p-value
S. mansoni, female1	-0.8662494	$<$ 2.2e-16
S. mansoni, female2	-0.9509563	$<$ 2.2e-16
S. mansoni, female3	-0.8616479	$<$ 2.2e-16
S. mansoni, male1	-0.8078201	$<$ 2.2e-16
S. mansoni, male2	-0.9648563	$<$ 2.2e-16
S. mansoni, male3	-0.821012	$<$ 2.2e-16
DSM1294	-0.867152	$<$ 2.2e-16
DSM25808	-0.8410847	$<$ 2.2e-16
DSM36	-0.8606152	$<$ 2.2e-16
DSM644	-0.8491182	$<$ 2.2e-16
DSM9161	-0.8793625	$<$ 2.2e-16

3.4 Peak finding

Compare peaks found with Piranha and pf_JF To get an idea, whether Piranha or the pf _JF founds the same peaks in the Ribozeq experiments, an intersection of peak finding files was performed. In Figure 3.23, one can see that using the unique mapped reads for peak-finding, Piranha can not find a peak, whereas the pf_JF finds only four peaks, which in addition, however, are under the threshold of a peak size of 30. The diagrams indicated with 'sorted' show all mapped reads, including the multi mapped ones. At a threshold of peak height 30, there are some unmatched reads from pf_JF ("own" in the diagram), but most unmapped hits from pf_JF are under the threshold. With a threshold of 100, Piranha finds all hits over the peak size of 100 that were found by pf_JF. So one can assume that the tool Piranha finds all relevant peaks in the mapped data and can be integrated in the snakemake analysis pipeline RAP.

Finding peaks that represent ribozymes The sequencing method should detect self-cleaving ribozymes. For that a annotation was built that especially detects ribozymes. In this step, it should be checked if ribozymes can be found among the peaks. One aim of this work is to find new annotation sites for ribozymes and confirm them with the sequencing data.

In *S. mansoni* peakfinding out of the data without deduplication after mapping of the reads, several ribozyme peaks could be detected (Figure 3.24, upper row). Intersection with the ribozyme annotation shows 1036 peaks from the pf_JF in the female samples and 208 in the male ones. The Piranha peakfinder finds 72 peaks for HH ribozymes. Additionally, there are peaks in the mapping data, that intersect with RAGATH-2-HDV and Twister-P1 ribozymes in the ribozyme annotation. RAGATH-2-HDV and Twister-P1 ribozymes are not annotated in the wormbase annotation so that no intersection hits can be found there. The amount of peaks found with pf_JF is higher for all conditions. In the samples from male *S. mansoni*, peaks could only be seen with the pf_JF, but in a smaller quantity than in the female samples. After deduplication using UMI-tools, peakfinding with Piranha can not reveal peaks that intersect with ribozymes in the ribozyme and the wormbase annotation (Figure 3.24, bottom row). There is a small number of peaks in female *S. mansoni* samples for HH ribozymes. In male *S. mansoni* samples, additionally, two peaks of RAGATH-2-HDV ribozyme can be detected in deduplicated mapping files.

Experiments in the Bacteria species revealed one hit in DSM644, and also DSM9161 (Figure 3.25). Surprisingly, without deduplication, HH ribozyme is found in DSM644, with deduplication HDV-ribozyme. In DSM9161, the peakfinder Piranha finds one hit each for HH and Twister-P5 ribozyme.

Summing up, it is the first time that a direct sequencing method for ribozymes was used and that peaks representing ribozymes could be detected, even in bacteria species.

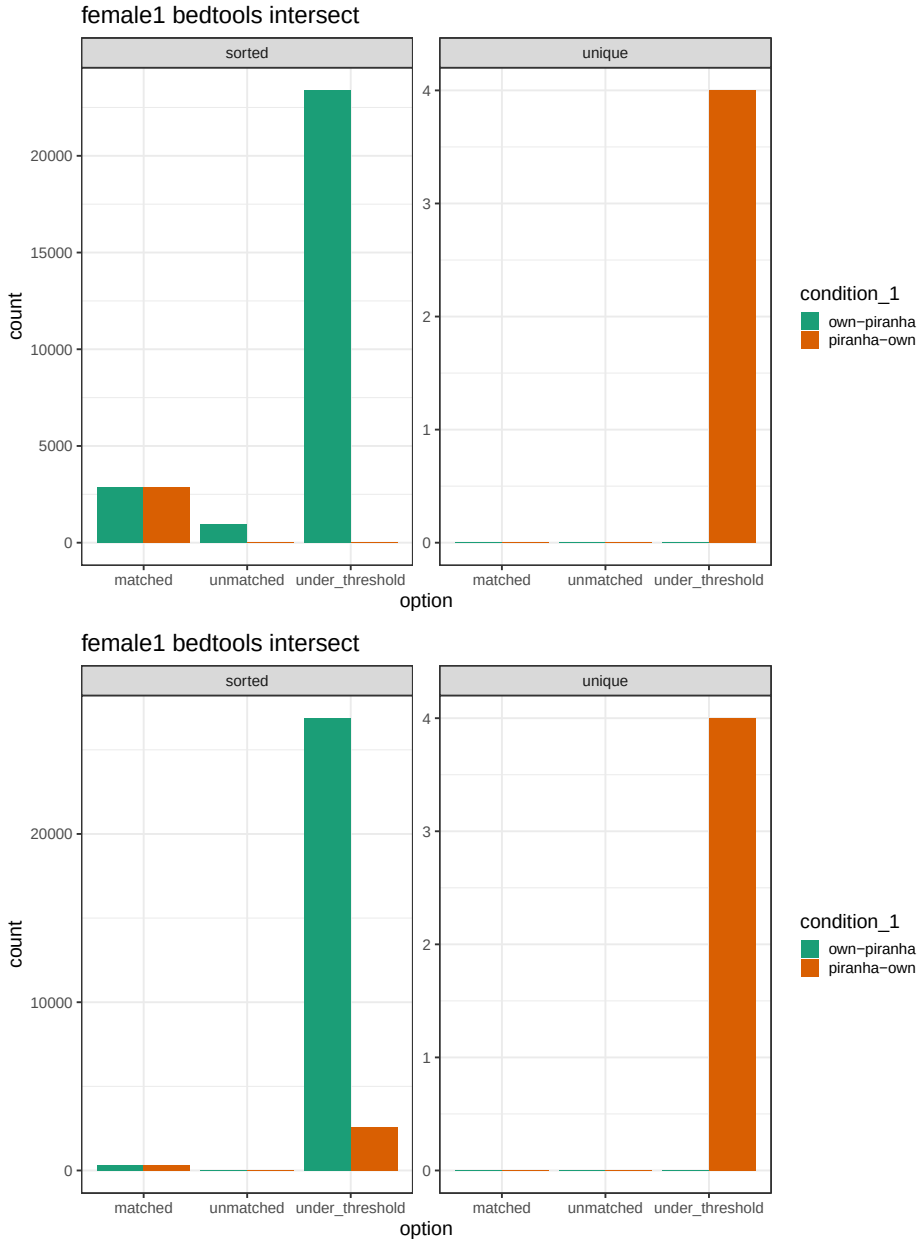


Figure 3.23: **Intersection results from peakfinding with Piranha intersected with pf_PF and vice versa for the *S. mansoni* female1 sample** A Threshold for peak-height 30. B Threshold for peak-height 100. sorted = all mapped reads, unique = unique mapped reads, own = pf_JF, pir = Piranha

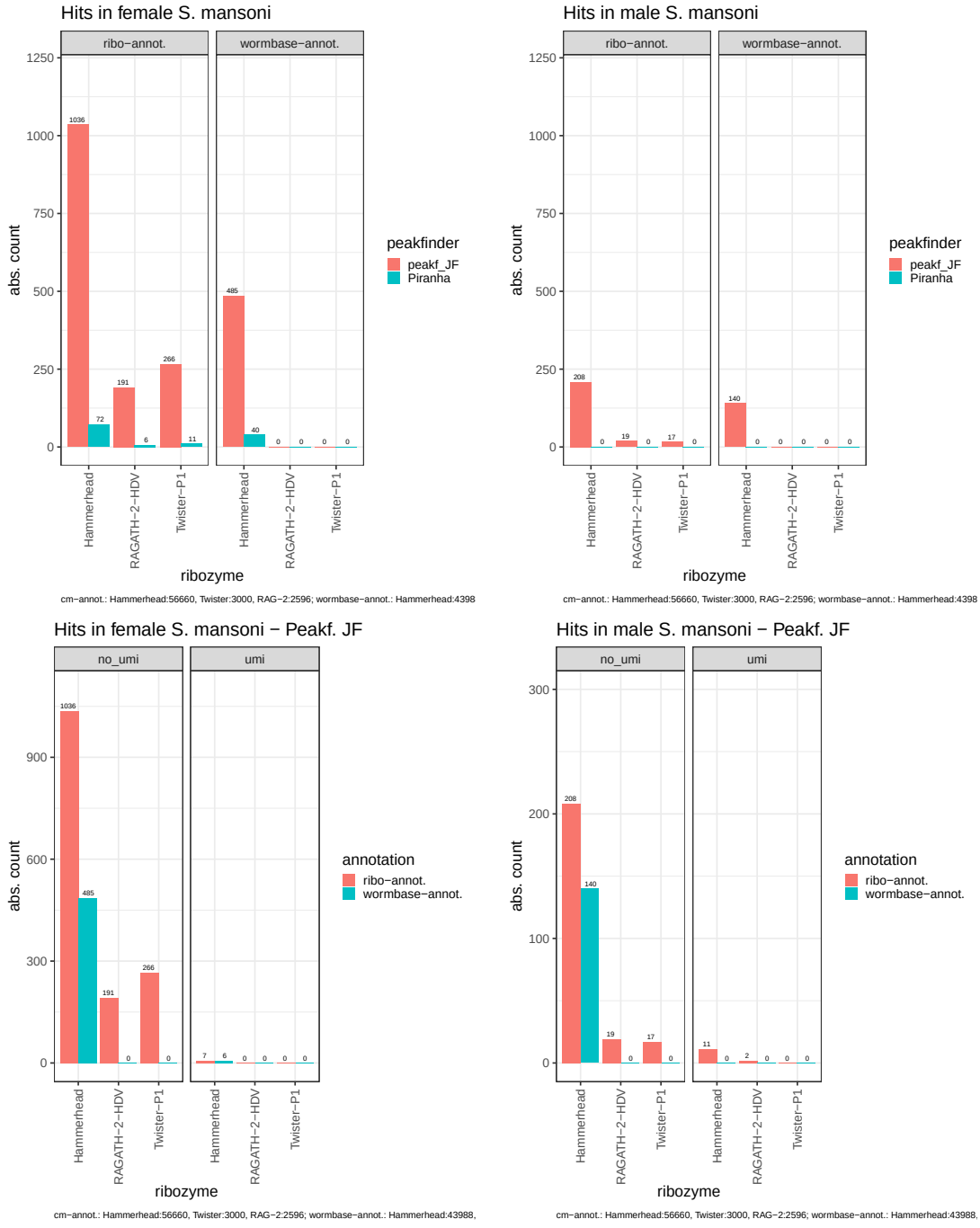


Figure 3.24: **Ribozyme hits found in *S. mansoni* samples.** From upper left to down right: no_UMI female and male *S. mansoni*, pf_JF in female and male *S. mansoni*

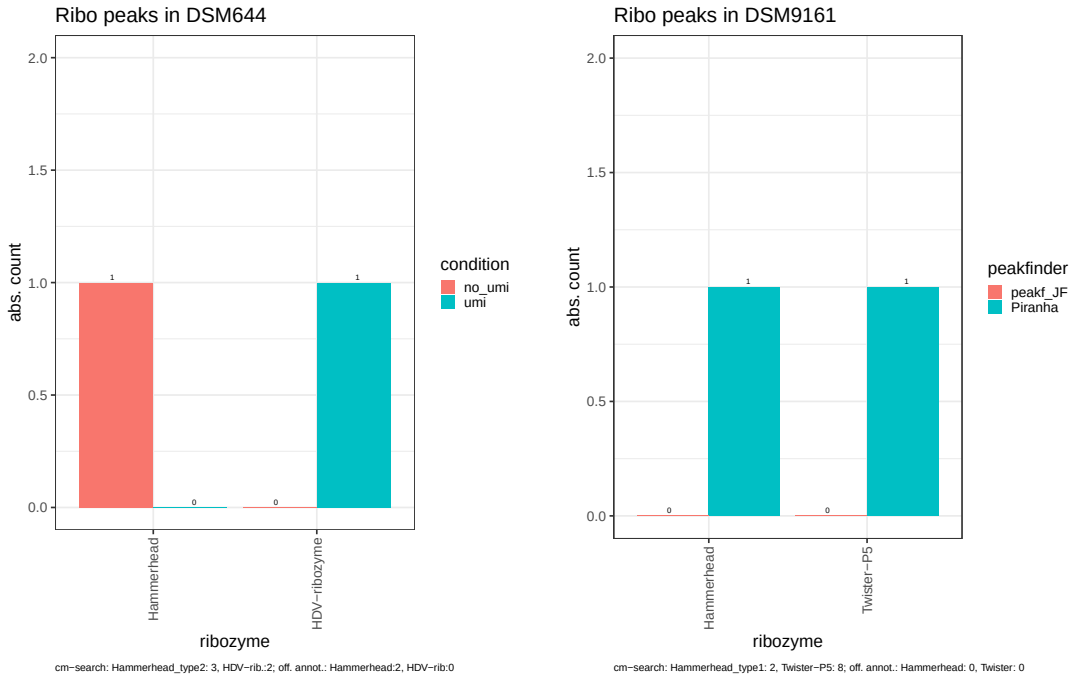


Figure 3.25: **Ribozyme hits found in bacteria species.** However, there were no peaks found for ribozymes in the other investigated bacteria species.

Analysis of the other peaks After finding the ribozyme peaks and excluding them from the mapping data, it is of interest to see to what elements the other reads of the sequencing experiments mapped. For that, the results of the two peakfinders were separated from one another. To see the effect of deduplication,the peak distribution was analyzed for each sample and peakfinder with and without deduped mapping files.

Since ribozymes were detected in the samples of *S. mansoni*, DSM644 and DSM9161, the distribution of oeaks to locations other than ribozymes are shown below. The results of the other species can be found in the appendix (Figure D.1 to Figure D.4).

In the female samples from *S. mansoni* analyzed with Piranha (Figure 3.26, upper row), many hits mapped to protein-coding-, repeat-, and exon regions. After deduplication, the frequency of other peaks is smaller, but now the relative amount that is assigned to tRNA and rRNA is higher. In contrast, no peaks in exon regions were found in the deduplicated runs. Peakfinding with the pf_JF surprisingly showed two peaks for HH-9 ribozyme that were not covered by the ribozyme annotation (Figure 3.26, lower row). Most hits in the run without deduplication were found in protein-coding-, and exon-regions, whereas after deduplication, relatively more peaks were localized to rRNA and repeat-regions.

Peakfinding with Piranha in DSM644 surprisingly has more peaks with deduplication compared to the run without deduplication. relatively most hits were found in protein-coding-, and exon-regions, for regions annotated as tRNA and rRNA (Figure 3.27, upper row). The pf_JF found more hits without deduplication. The relative amount of peaks that localize to rRNA-, and tRNA-coding regions is smaller after deduplication (Figure 3.27, lower row).

In DSM9161, one can see that the absolute number of peaks is lower after deduplication independent of the peakfinder used. Peakfinding with Piranha shows a high amount for exon-, tRNA-, and rRNA-regions after deduplication. Without deduplication, nearly three-quarters of the hits located to protein-coding-regions (Figure 3.28, upper row). With the pf_JF, one could see a lower relative part of hits mapping to rRNA- and tRNA-regions. Most peaks mapped to exon-regions and pseudogenes (Figure 3.28, lower row).

Although the Ribozeq experiments contained peaks that could be mapped to ribozymes, many peaks were found that were mapped to other locations. This was also the case after deduplication with UMI-tools, and it could be a potential pitfall in the Ribozeq experiment analysis. Nevertheless, a deduplication should be performed if a quantitative transcriptional analysis of ribozymes is to be performed or if different experimental conditions should be compared to each other.

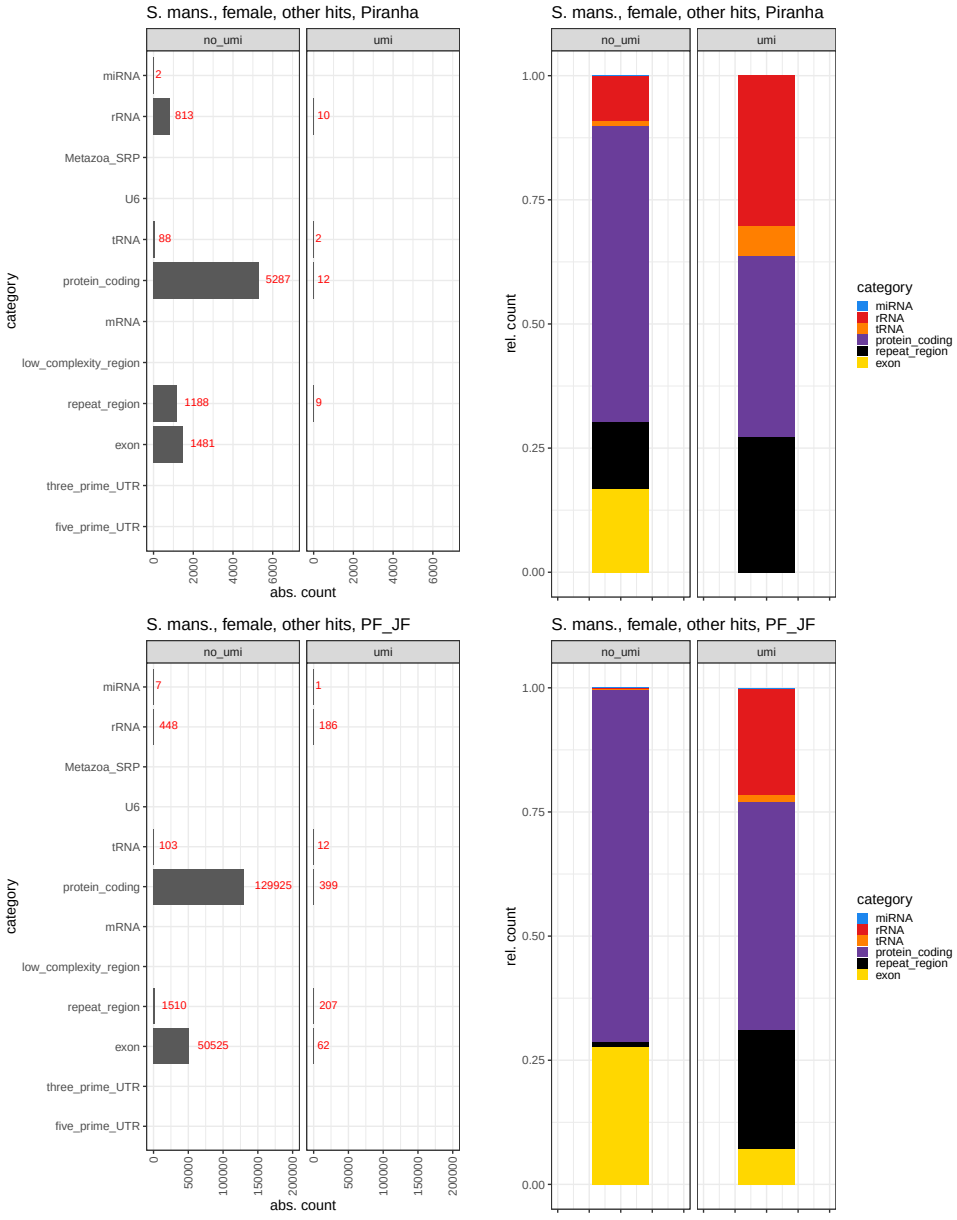


Figure 3.26: **Intersection of peaks other than ribozymes with the "official" annotations in female *S. mansoni*.**

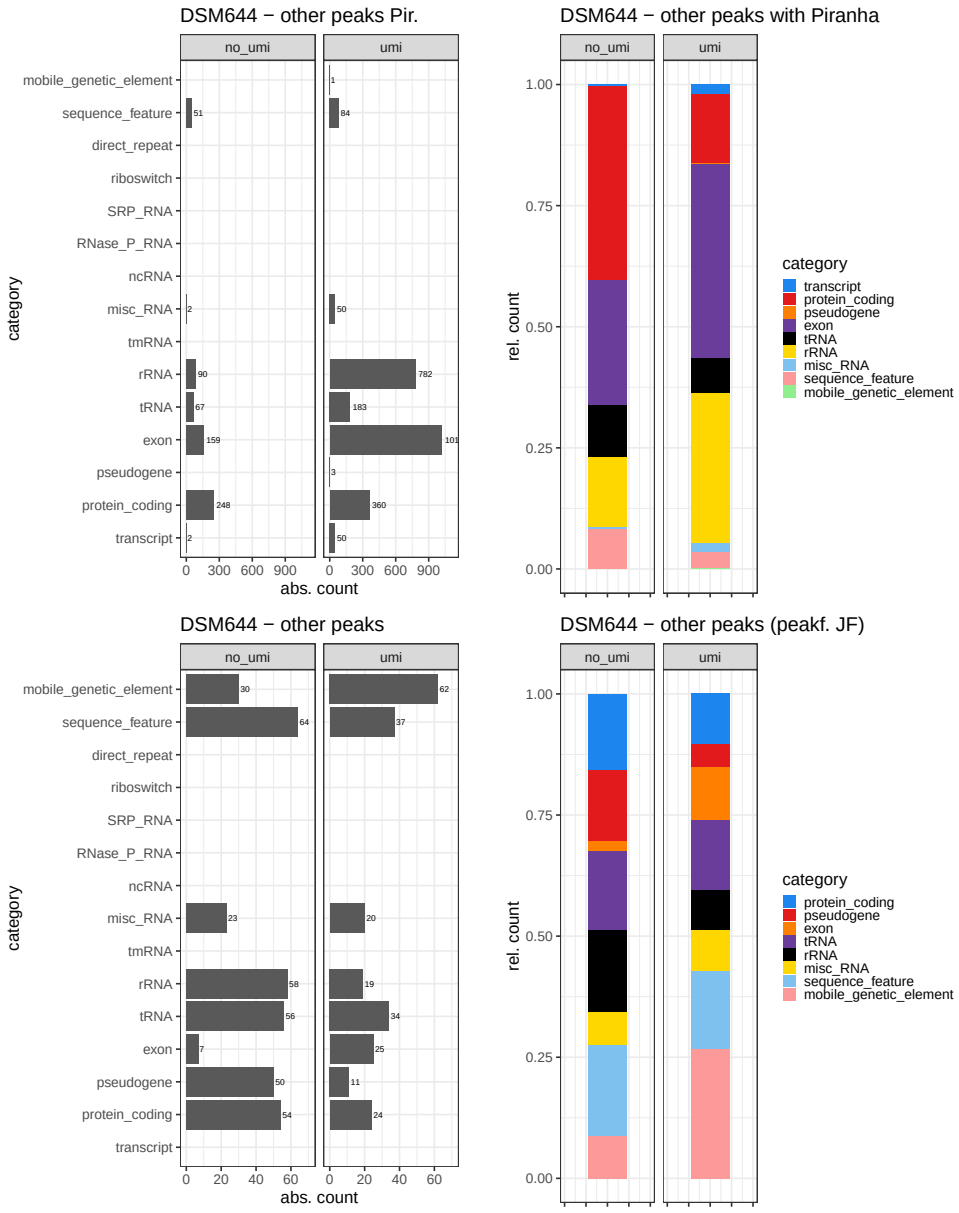


Figure 3.27: **Intersection of peaks other than ribozymes with the "official" annotations in DSM644**

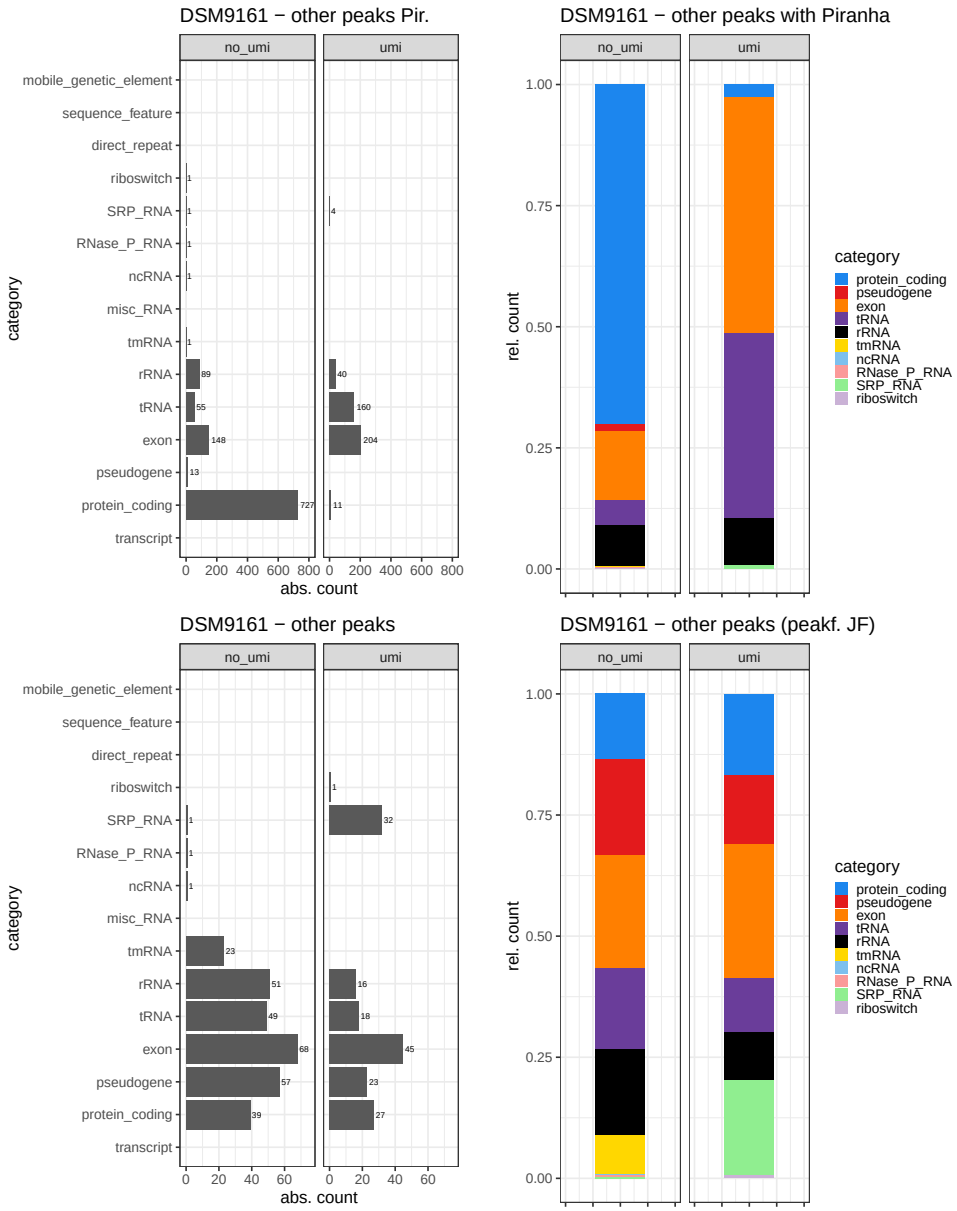


Figure 3.28: **Intersection of peaks other than ribozymes with the "official" annotations in DSM9161**

Chapter 4

Discussion

This study came up with a pipeline to analyze output from Ribozeq experiments. Since this sequencing method is new, no already established pipeline could be used for the analysis. Additionally, the occurrences of the individual ribozyme types in the genomes of the investigated species had to be annotated because this annotation was insufficient in the available genome data. This chapter summarizes and evaluates the findings and discusses the limitations of this study.

First, a ribozyme annotation was built for each investigated species using *cm-search*. To find all possible annotation sites, different search modes of *cm-search* were run. The mid and hmm search modes exclude some filter steps of cm-search, in the hope to find more potential annotation sides, even with reduced specificity. For all species, *cm-search* reveals most, even significant, hits with the mid search option, but taken together, the standard search option found most of these hits, too. Length and e-value distribution of mid and std search hits were equivalent. In contrast, hits from hmm search option were in median shorter and with higher e-values. The shorter length of these hits could be caused by the fact that only parts of the ribozymes are recognized. It may be in the nature of HMMs that do not include the secondary structure of RNA. One could assume that one hit out of mid or std search could be covered by two hits from hmm search option, but intersect analysis could disprove this assumption. Most likely, the hits from the hmm search are parts of the ribozyme sequence, or they could be pseudogenes.

The amount of both different ribozyme types and annotation sites per species showed that most of both could be found in *S. mansoni*. Since *S. mansoni* is known to express a high amount of ribozyme, this observation is not surprising at all. Except for DSM36 (P.polymyxa) in all investigated species could be found at least the expected ribozyme types with significant hits in *cm-search*. Additionally, new ribozyme types could be annotated in *S. mansoni*, DSM1294, DSM25808, and DSM9161 (Table 4.1). Although the hits that were integrated as annotation sites for ribozymes had a significant e-value, expression studies are necessary to prove that it is an actual transcription side for ribozymes.

Although it was already known that Twister and RAGATH-2-HDV could be found in *glss mansonis'* genome, it is unclear why these ribozymes were not annotated in the recent genome files from WormBase ParaSite. It is also the case for the bacteria species and their recent genome files. So these missing annotations could be included in the genomes using the pipeline.

species	previous known ribozyme types	new annotated ribozymes
S. mansoni	HH-type1, Tw-P1,RAG-2	HH_9, HH-type3
DSM1294	Twister-P5	Twister-P1
DSM25808	HHtype2 & Twister-P1	Twister-P5, HH-type3
DSM36	Pistol, but not found	no other
DSM644	HHtype2	no other
DSM9161	Twister-P5	RAG-2, HH-type1, HH-type3

Table 4.1: **Studied species, expected ribozyme classes and new annotated ribozyme classes per species.** RAG-2 = RAGATH-2-HDV, Tw = Twister

Intersection analysis showed that hits cover most hits from mid search from std search, so one can conclude that for further ribozyme annotation in genomes without ribozyme annotations, std option from *cmsearch* should be sufficient for further investigation of ribozymes in other species.

Intersection studies and verification of hits within the WormBase genome browser could show that HH_9 and HH-type3 are often annotated together with annotation sites of HH-type1. The reason for that could be a sequence similarity, and no ribozyme expression studies up to date that could show which HH type is exactly coded in this site.

Although the Ribozeq experiments contained peaks that could be mapped to ribozymes, many peaks were found that were mapped to other locations. This was also the case after deduplication with UMI-tools, and it could be a potential pitfall in the Ribozeq experiment analysis. Thus, the sequencing method itself should be adapted to be more specific for ribozymes. It would also be conceivable to develop bioinformatic strategies to filter the reads, such as the mapper on tRNA sites in advance. As far as these strategies are not developed, one could think about also using multi-mapped reads for analysis, while only taking unique mapped reads into account. Currently, no ribozyme peaks could be detected while analyzing unique mapped reads.

The results from extracting and deduplication of UMI-read-combinations show that this step is crucial to prevent false high expression results; thus, single UMI-read-combinations were found in a very high amount within the sequencing data. So this step should be a part of future experiments, too.

A comparison of the two peak finders revealed that *Piranha* could be used for the pipeline without losing significant peaks. Due to the pre-processing of input files for this tool, it will also be possible to analyze the results with the second sequencing method that catches the 5' end of the ribozyme cleavage site, once it is established.

The creation of an analysis pipeline with Snakemake seems to be the right choice, so future users only have to adapt the configuration file that specifies the desired steps that should be performed and the input files' names and locations. So the pipeline should be easy to use even for people without a strong bioinformatics background.

In this study, the goal of creating a functional pipeline to analyze the output of Ri-

bozeq experiments could be reached. Moreover, integrating *cm-search* into the pipeline, it is possible to annotate ribozymes into the genome of the species of interest. This works even there are no ribozymes annotated before.

Summing up, it is the first time that a direct sequencing method for ribozymes was used and evaluated with a bioinformatics pipeline that semi-automated finds annotation sites of ribozymes and detects peaks within the sequencing data, even in species with a few amount of ribozyme sites (in this study bacteria species).

Further analyses could include the analysis of different experimental conditions, e.g., comparing the transcription in male or female individuals of *S. mansoni* or changes of ribozyme expression in different stages of development. It could also be important to perform more laboratory experiments to prove the different annotation sites in the investigated species and find ribozymes that cleave itself near the 5' end (e.g., Twister ribozymes). The last experiment could be realized with the further described capturing of the fragment with the 5'-OH group using the RtcB ligase for the adapter-ligation.

Abstract

To date, there is no analysis pipeline for Ribozyme Sequencing (Ribozeq) experiments available. Such a pipeline is necessary to evaluate the results of a newly developed method for catching the 2',3'-cP or 5'-OH group ends of self-cleaving ribozymes with special ligases.

Different steps were integrated into the pipeline: annotating missing ribozyme sites, trimming and mapping the reads, deduplication, peak finding, transferring the results to a Genome Browser, and statistical analysis of the transcription of ribozymes under different experimental conditions. In order to improve the data evlauation from the Ribozeq experiments this thesis comes along with comparing different options or tools, especially for annotation site detection and peak finding.

The Ribozeq Analysis Pipeline (RAP) was implemented in snakemake and is available at github.com/Tschichen/RAP. The pipeline is designed so that even users with little bioinformatics background can evaluate their experiments with it.

Bibliography

[1] *Debugging by Thinking - A Multidisciplinary Approach.* Hewlett-Packard Development Company, Amsterdam, Booston, Heidelberg, 2004.

[2] Samtools. `http://www.htslib.org`, 2019. Online; accessed 15 September 2020.

[3] A. Achar and P. Saetrom. Rna motif discovery: a computational overview. *Biology Direct*, 10:10:61, 2015.

[4] S. Altman. Ribonuclease p. *Philos Trans R Soc Lond B Biol Sci*, 366:2936–2941, 2011.

[5] S. Andrews. Fastqc. `https://github.com/s-andrews/FastQC`, 2019. Online; accessed 05 April 2020.

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

[7] J. Boissier, S. Morand, and H. Mone. A review of performance and pathogenicity of male and female schistosoma mansoni during the life-cycle. *Parasitology*, 119(5):447–454, 1999.

[8] D. Boji and M. Bojovi. A streaming dataflow implementation of parallel cocke–younger–kasami parser. *Advances in Computers*, 104:159–199, 2017.

[9] CDC. Parasites - schistosomiasis- biology. `https://www.cdc.gov/parasites/schistosomiasis/biology.html`, 2020. Online; accessed 18 August 2020.

[10] N. Chomsky. Three models for the description of language. *IRE Transactions on Information Theory*, 2:113–124, 1956.

[11] M. A. Clark, M. Douglas, and J. Choi. *Biology 2e.* OpenStax, Houston, Texas, USA, 2018.

[12] R. B. Darnell. Umi-tools: Modelling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Wiley Interdiscip Rev RNA*, 1(2):266–86, 2010.

[13] R. Devereux, S. H. He, C. L. Doyle, S. Orkland, D. A. Stahl, J. LeGall, and W. B. Whitman. Diversity and origin of desulfovibrio species: phylogenetic definition of a family. *Journal of Bacteriology*, 172(7):3609–19, 1990.

[14] A. S. Distribution. Anaconda, version 4.8.4. https://anaconda.com, 2020. Online; accessed 20 September 2020.

[15] A. Dobin. github - star 2.7. https://github.com/alexdobin/STAR, 2009-2019. Online, accessed 30 June 2020.

[16] A. Dobin. Star manual 2.7.5a. github.com/alexdobin/STAR/blob/master/doc/STARmanual.pdf, 2020. Online, accessed 30 June 2020.

[17] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. Star: ultrafast universal rna-seq aligner. *BIOINFORMATICS*, 29:15–21, 2015.

[18] R. D. Dowell and S. R. Eddy. Evaluation of several lightweight stochastic context-free grammars for rna secondary structure prediction. *BMC Bioinformatics*, 5:71 (2004), 2004.

[19] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK, 1998.

[20] S. R. Eddy and R. Durban. Rna sequence analysis using covariance models. *Nucleic Acids Research*, 2(11):2079–2088, 1994.

[21] T. R. for statistical computing. R for mac os x. https://cran.r-project.org/bin/macosx/, 2020. Online; accessed 04 January 2020.

[22] P. S. Foundation. Python, version 3.7.5. https://www.python.org/, 2020. Online; accessed 01 November 2019.

[23] J. Gebetsberger and R. Micura. Unwinding the twister ribozyme: from structure to mechanism. *WIREs RNA*, 8:e1402, 2017.

[24] J. Gorodkin and W. L. Ruzzo, editors. *RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods*. Springer, Heidelberg, Germany, 2014.

[25] K. A. Harris, C. E. Lünse, S. Li, K. I. Brewer, and R. R. Breaker. Biochemical analysis of pistol self-cleaving ribozymes. *RNA*, 21(11):1852–1858, 2015.

[26] P. C. Heinrich, M. Müller, and L. Graeve, editors. *Löffler/Petrides Biochemie und Pathobiochemie*. Springer, Heidelberg, Germany, 2014.

[27] P. G. Higgs and N. Lehman. The rna world: molecular cooperation at the origins of life. *Rev Genet*, 16:7–17, 2015.

[28] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer, Heidelberg, Germany, 2013.

[29] D. Jurafsky and J. H. Martin, editors. *Speech and Language Processing*. Prentice Hall, Upper Saddle River, New Jersey, 2009.

[30] P. Kerpedjiev, S. Hammer, and I. L. Hofacker. Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. *Bioinformatics*, 31(20):3377–3379, 06 2015.

[31] J. Koester. Snakemake. `https://github.com/snakemake/snakemakes`, 2020. Online; accessed 15 September 2020.

[32] J. Koester and S. Rahmann. Snakemake—a scalable bioinformatics workflow engine. *Bioinformarics*, 28(19):2520–2522, 2012.

[33] F. Krueger. Trim galore. `https://github.com/FelixKrueger/TrimGalore/blob/master/Docs/Trim_Galore_User_Guide.md`, 2019. Online; accessed 09 September 2020.

[34] E. lab. Infernal. `http://eddylab.org/infernal/`, 2019. Online; accessed 12 December 2019.

[35] T. S. Lab. Piranha. `http://smithlabresearch.org/software/piranha/`, 2012. Online; accessed 15 March 2020.

[36] C. M. Lee, G. P. Barber, and J. C. et al. Ucsc genome browser enters 20th year. *Nucleic Acids Research*, 48(D1):D756–D761, 2020.

[37] T. A. M. Manolio, F. S. Collins, N. J. Cox, and et al. Finding the missing heritability of complex diseases. *Nature*, 461:747–753, 2009.

[38] S. Mauss, T. Berg, J. Rockstroh, C. Sarrazin, and H. Wedemeyer, editors.

[39] E. Nawrocki and S. Eddy. Infernal user's guide. `http://gensoft.pasteur.fr/docs/infernal/1.1.2/Userguide.pdf`, 2016. Online; accessed 12 December 2019.

[40] NCBI. Dsm1294 annotation, asm38329v1. `ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/383/295/GCF_000383295.1_ASM38329v1/GCF_000383295.1_ASM38329v1_genomic.fna.gz`, 2019. Online, 10 October 2019.

[41] NCBI. Dsm25808, fervidicella_assembly. `ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/601/455/GCF_000601455.1_Fervidicella_assembly/GCF_000601455.1_Fervidicella_assembly_genomic.fna.gz`, 2019. Online, accessed 10 October 2019.

[42] NCBI. Dsm36 annotation, asm16498v2. `ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/164/985/GCF_000164985.3_ASM16498v2/GCF_000164985.3_ASM16498v2_genomic.fna.gz`, 2019. Online, accessed 10 October 2019.

[43] NCBI. Dsm644 annotation, asm19575v1. `ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/195/755/GCF_000195755.1_ASM19575v1/`, 2019. Online, accessed 10 October 2019.

[44] NCBI. Dsm9161 annotation, asm24315v3. `ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/243/155/GCF_000243155.2_ASM24315v3/GCF_000243155.2_ASM24315v3_genomic.fna.gz`, 2019. Online, accessed 10 October 2019.

[45] NCBI. S. mansoni genome, accession: Prjea36577 id: 36577 schistosoma mansoni strain:puerto rico. https://www.ncbi.nlm.nih.gov/bioproject/PRJEA36577, 2020. Online, accessed 17 February 2020.

[46] C. Ogg and B. Patel. Fervidicella metallireducens gen. nov., sp. nov., a thermophilic, anaerobic bacterium from geothermal waters. *International journal of systematic and evolutionary microbiology*, 60:1394–400, 09 2009.

[47] R. J. Ontiveros, J. Stoute, and K. F. Liu. The chemical diversity of rna modifications. *Biochem J*, 476(8):1227–1245, 2019.

[48] E. M. Osborne, J. E. Schaak, and V. J. Derose. Characterization of a native hammerhead ribozyme derived from schistosomes. *RNA*, 11(2):187–196, 2005.

[49] C. Oxford. Umi-tools. https://github.com/CGATOxford/UMI-tools, 2020. Online; accessed 20 September 2020.

[50] K. P. Padda, A. Puri, and C. P. Chanway. Plant growth promotion and nitrogen fixation in canola (brassica napus) by an endophytic strain of paenibacillus polymyxa and its gfp-tagged derivative in a long-term study. *Botany*, 94(12):1209–1217, 2016.

[51] A. F. Palazzo and E. S. Lee. Non-coding rna: what is functional and what is junk? *Frontiers in Genetics*, 6:2, 2015.

[52] S. Panni, R. C. Lovering, P. Porras, and S. Orchard. Non-coding rna regulatory networks. *Biochim Biophys Acta Gene Regul Mech*, 1863(6):194417, 2020.

[53] W. ParaSite. Wormbase parasite,v ersion: Wbps14 (ws271). https://parasite.wormbase.org/index.html, 2019. Online, accessed 20 September 2020.

[54] A. R. Quinlan and I. M. Hall. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[55] A. R. Quinlan and N. Kindlon. bedtools intersect. https://bedtools.readthedocs.io/en/latest/content/tools/intersect.html?highlight=intersect, 2020. Online; accessed 20 September 2020.

[56] G. A. Quinn, A. P. Maloy, S. McClean, B. Carney, and J. W. Slater. Lipopeptide biosurfactants from paenibacillus polymyxa inhibit single and mixed species biofilms. *Biofouling*, 28(10):1151–66, 2012.

[57] A. Randazzo, V. Esposito, O. Ohlenschläger, R. Ramachandran, and L. Mayol. Nmr solution structure of a parallel lna quadruplex - scientific figure on researchgate. www.researchgate.net/figure/Chemical-structure-of-DNA-RNA-and-LNA_fig1_8526091, 2004. Online, access 23 Aug, 2020.

[58] RDocumentation. Rdocumentation, loess. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/loess, 2020. Online; accessed 15 September 2020.

[59] Rfam. Family: Hammerhead-type2 (rf02276), species distribution. `http://rfam.xfam.org/accession/AZQP01000024.1?seq_start=34132&seq_end=34066`, 2020. Online; accessed 18 August 2020.

[60] Rfam. Family: Twister-p5, species distribution. `http://rfam.xfam.org/family/RF02684#tabview=tab4`, 2020. Online; accessed 18 August 2020.

[61] Rfam. Hammerhead type 3, alignments. `https://rfam.xfam.org/family/RF00008#tabview=tab2`, 2020. Online; accessed 01 September 2020.

[62] G. Romano, D. Veneziano, G. Nigita, and S. P. Nana-Sinkam. Rna methylation in ncrna: Classes, detection, and molecular associations. *Frontiers in Genetics*, 9:243, 2018.

[63] A. Roth, Z. Weinberg, A. G. Y. Chen, P. B. Kim, T. D. Ames, and R. R. Breaker. A widespread self-cleaving ribozyme class is revealed by bioinformatics. *Nat Chem Biol*, 10(1):56–60, 2014.

[64] R. Saldanha, G. Mohr, M. Belfort, and A. M. Lambowitz. Group i and group ii introns. *FASEB*, 7:15–24, 1993.

[65] S. Schbath, V. M. andMatthias Zytnicki, J. Fayolle, V. Loux, and J.-F. Gibrat. Mapping reads on a genomic sequence: An algorithmic overview and a practical comparative analysis. *J Comput Biol*, 19(6):796–813, 2012.

[66] W. G. Scott, L. H. Horan, and M. Martick. The hammerhead ribozyme: Structure, catalysis and gene regulation. *Prog Mol Biol Transl Sci*, 120:1–23, 2013.

[67] V. Sedlyarov, J. Fallmann, F. Ebner, J. Huemer, L. Sneezum, M. Ivin, K. Kreiner, A. Tanzer, C. Vogl, I. Hofacker, and P. Kovarik. Tristetraprolin binding site atlas in the macrophage transcriptome reveals a switch for inflammation resolution. *Mol Syst Biol*, 12:868, 2016.

[68] T. S. Smith, A. Heger, and I. Sudbery. Umi-tools: Modelling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Res*, 27(3):491–499, 2017.

[69] S. Suerbaum, G.-D. Burchard, S. H. E. Kaufmann, and T. F. Schulz, editors. *Medizinische Mikrobiologie und Infektiologie*. Springer, Heidelberg, Germany, 2020.

[70] A. E. Trotochaud and K. M. Wassarman. 6s rna function enhances long-term cell survival. *J Bacteriol*, 186(15):4978–4985, 2004.

[71] M. Uhl, T. Houwaart, G. Corrado, P. R. Wright, and R. Backofen. Computational analysis of clip-seq data. *Methods*, 118-119:60–72, 2017.

[72] P. J. Uren, E. Bahrami-Samani, S. C. Burns, M. Qiao, F. V. Karginov, E. Hodges, G. J. Hannon, J. R. Sanford, L. O. F. Penalva, and A. D. Smith. Site identification in high-throughput rna–protein interaction data. *Bioinformatics*, 28 no. 23:3013–3020, 2012.

[73] R. Villemur, M. Lanthier, and R. B. F. Lepine. The desulfitobacterium genus. *FEMS Microbiology Reviews*, 30(5):706–33, 2006.

[74] C.-H. T. Webb and A. Luptak. Hdv-like self-cleaving ribozymes. *RNA Biology*, 8:5:719–727, 2011.

[75] C. E. Weinberg, Z. Weinberg, and C. Hammann. Novel ribozymes: discovery, catalytic mechanisms, and the quest to understand biological function. *Nucleic Acids Research*, 47:9480–9494, 2019.

[76] Z. Weinberg, P. B. Kim, T. H. Chen, S. Li, K. A. Harris, C. E. Luense, and R. R. Breaker. New classes of self-cleaving ribozymes revealed by comparative genomics analysis. *Chem Biol*, 11(8):606–10, 2015.

[77] Z. Weinberg, C. E. Lünse, K. A. Corbino, T. D. Ames, J. W. Nelson, A. Roth, K. Perkins, M. E. Sherlock, and R. R. Breaker. Detection of 224 candidate structured rnas by comparative analysis of specific subsets of intergenic regions. *Nucleic Acids Res*, 45(18):10811–10823, 2017.

[78] H. Wickham, W. Chang, L. Henry, and et al. ggplot2. `https://cloud.r-project.org/web/packages/ggplot2/index.html`, 2020. Online; accessed 15 September 2020.

[79] WORMBASE. Wormbase, s. mansoni annotation, version: Ws276. `ftp://ftp.wormbase.org/pub/wormbase/parasite/releases/WBPS11/species/schistosoma_mansoni/PRJEA36577/schistosoma_mansoni.PRJEA36577.WBPS11.annotations.gff3.gz`, 2020. Online; accessed 17 February 2020.

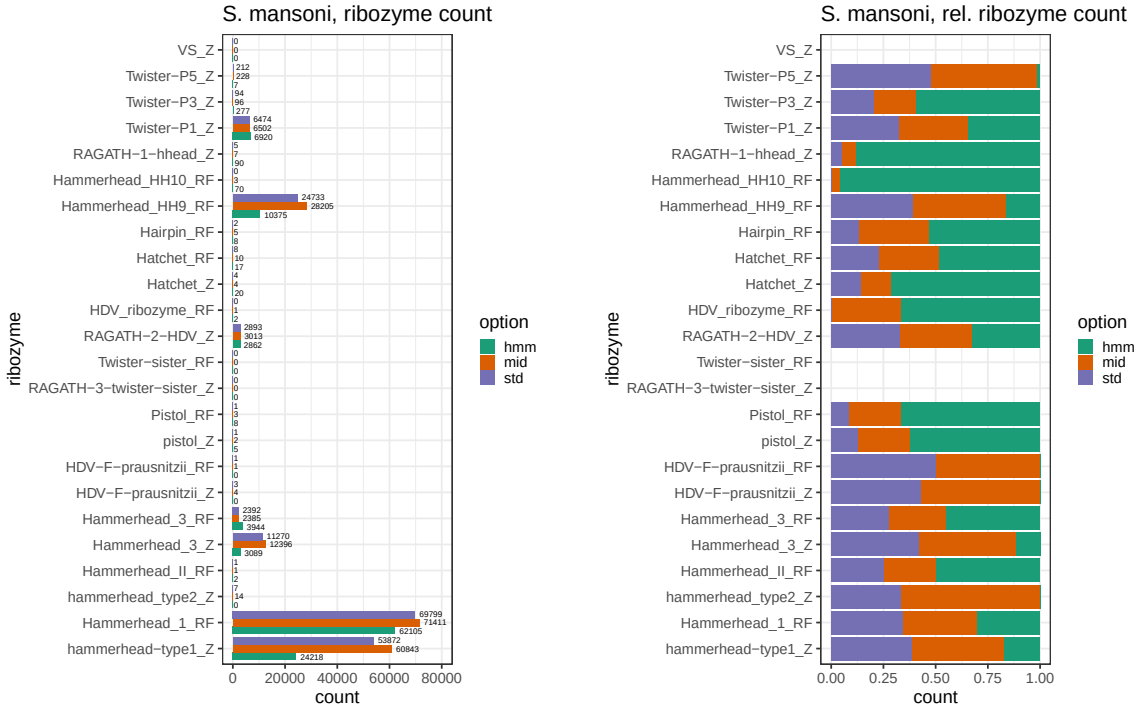
Appendices

Appendix A

Ribozyme hits from *cm-search*

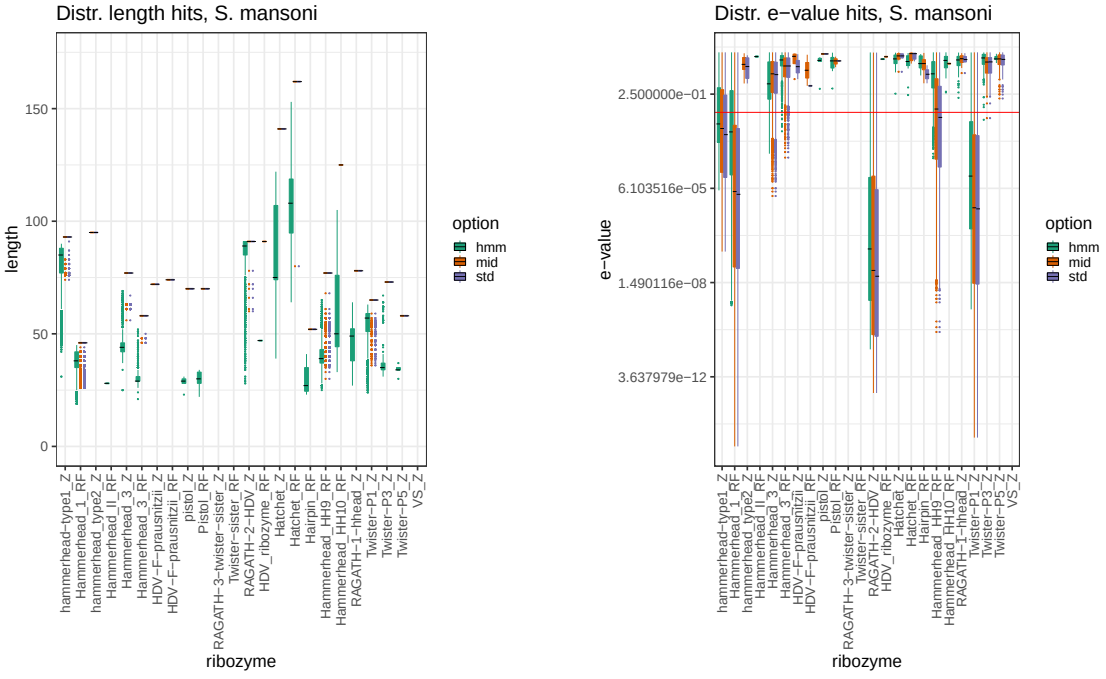
Table A.1: **Counts of found ribozymes in Schistosoma mansoni.** *cm-search* options hmm, mid option and std. In brackets is the number without e-value limit.

Ribozyme	hmm	mid	std
Twister-P5_Z	0 (7)	0 (212)	0 (228)
Twister-P3_Z	1 (277)	0 (69)	0 (94)
Twister-P1_Z	5436 (6920)	5300 (6502)	5296 (6474)
RAGATH-1-HH_Z	0 (90)	0 (7)	0 (5)
HH_10_RF	0 (70)	0 (3)	0 (0)
HH_9_RF	389 (10375)	13403 (28205)	13196 (24733)
Hairpin_RF	0 (8)	0 (5)	0 (2)
Hatchet_RF	0 (17)	0 (10)	0 (8)
Hatchet_Z	0 (20)	0 (4)	0 (4)
RAGATH-2-HDV_Z	2589 (2862)	2596 (3013)	2579 (2893)
pistol_Z	0 (5)	0 (2)	0 (1)
Pistol_RF	0 (8)	0 (3)	0 (1)
HDV-F-prausnitzii_RF	0 (0)	0 (4)	0 (3)
HDV-F-prausnitzii_Z	0 (0)	0 (1)	0 (1)
HH_3_RF	6 (3944)	58 (2385)	58 (2392)
HH_3_Z	138 (3089)	914 (12396)	943 (11270)
HH_II_RF	0 (2)	0 (1)	0 (0)
HH_type2_Z	0 (0)	0 (14)	0 (17)
HH_1_RF	38733 (62105)	56641 (71411)	56211 (69799)
HH-type1_Z	14315 (53872)	36016 (60843)	34153(60843)



(a) (b)

Figure A.1: **Schistosoma mansoni: Ribozyme count with no threshold in e-value** absolute count (a), relative count (b).

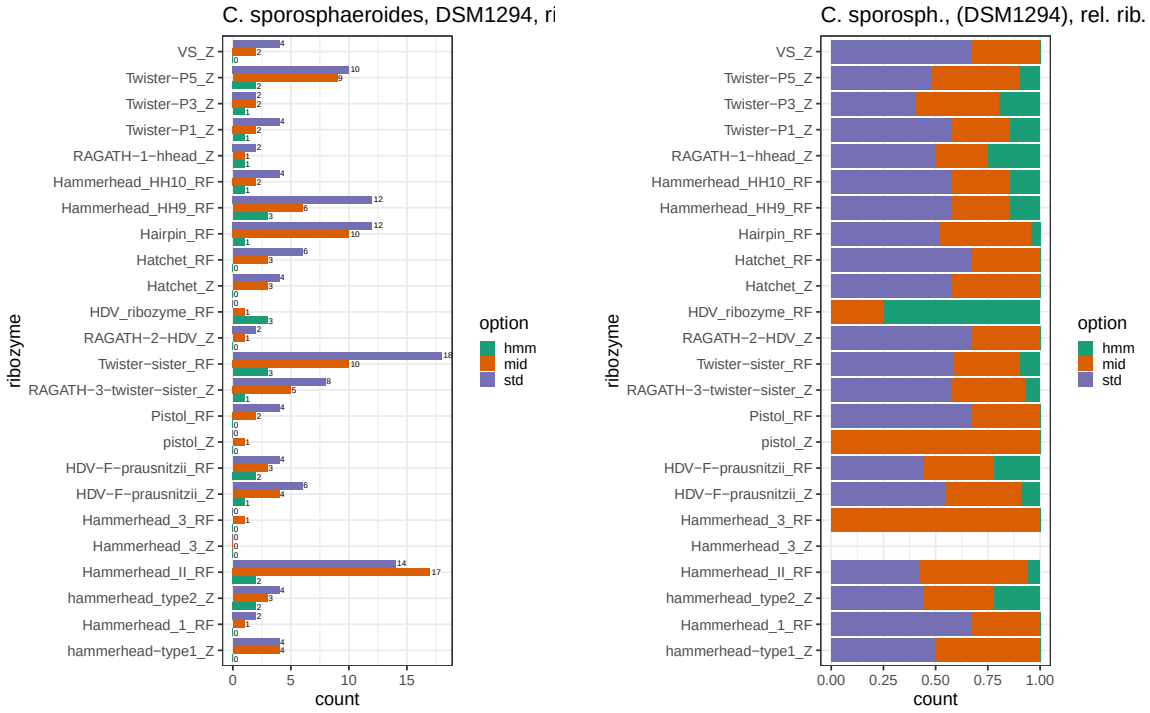


(a) (b)

Figure A.2: **S. mansoni: eval vs. length for Schisto all.**

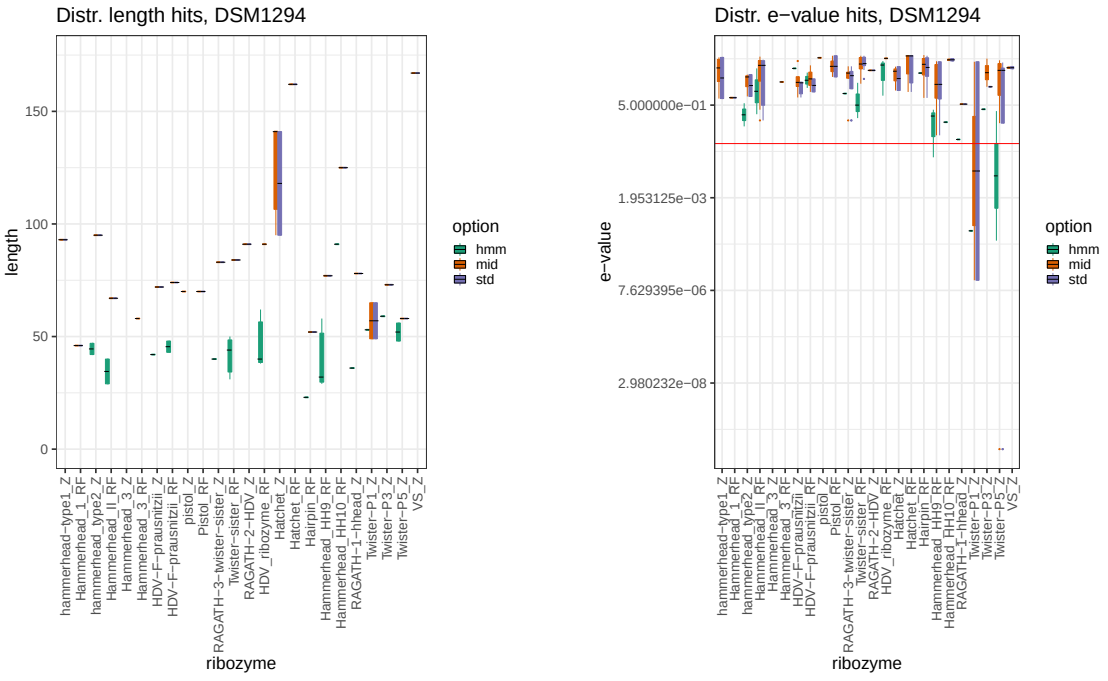
Table A.2: **Counts of found ribozymes in Clostridium sporoshaeroides (DSM1294).** *cm-search* options: hmm, mid, std. In brackets is the number without e-value limit

Ribozyme	hmm	mid	std
VS_Z	0 (0)	0 (2)	0 (4)
Twister-P5_Z	1 (2)	1 (9)	2 (10)
Twister-P3_Z	0 (1)	0 (2)	0 (2)
Twister-P1_Z	1 (1)	1 (2)	2 (4)
RAGATH-1-HH_Z	0 (1)	0 (1)	0 (2)
HH_10_RF	0 (1)	0 (2)	0 (4)
HH_9_RF	1 (3)	0 (6)	0 (12)
Hairpin_RF	0 (1)	0 (10)	0 (12)
Hatchet_RF	0 (0)	0 (3)	0 (6)
Hatchet_Z	0 (0)	0 (3)	0 (4)
RAGATH-2-HDV_Z	0 (0)	0 (1)	0 (2)
Twister-sister_RF	0 (3)	0 (10)	0 (18)
RAGATH-3-twister-sister_Z	0 (1)	0 (5)	0 (8)
pistol_Z	0 (0)	0 (1)	0 (0)
Pistol_RF	0 (0)	0 (2)	0 (4)
HDV-F-prausnitzii_RF	0 (2)	0 (3)	0 (4)
HDV-F-prausnitzii_Z	0 (1)	0 (4)	0 (6)
HH_3_RF	0 (0)	0 (1)	0 (0)
HH_II_RF	0 (2)	0 (17)	0 (14)
HH_type2_Z	0 (2)	0 (3)	0 (4)
HH_1_RF	0 (0)	0 (1)	0 (2)
HH-type1_Z	0 (0)	0 (4)	0 (4)



(a) (b)

Figure A.3: **DSM1294.:Absolute and relative count for *cm-search* Hits with different search options.** Without a threshold in e-value. absolute count (a), relative count (b).



(a) (b)

Figure A.4: **DSM1294: length and e-value distribution of hits without a threshold in e-value.** (a) length, (b) e-value.

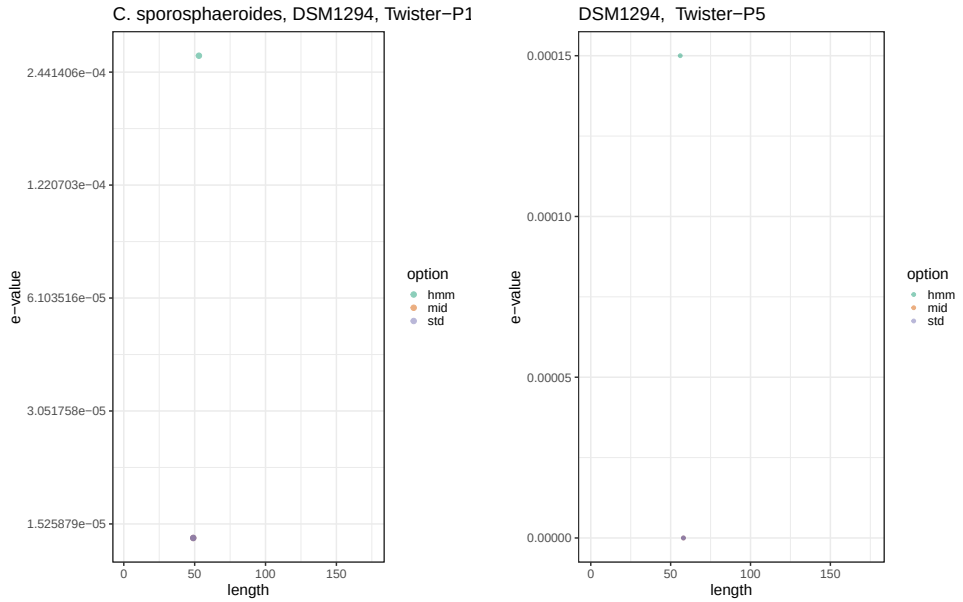
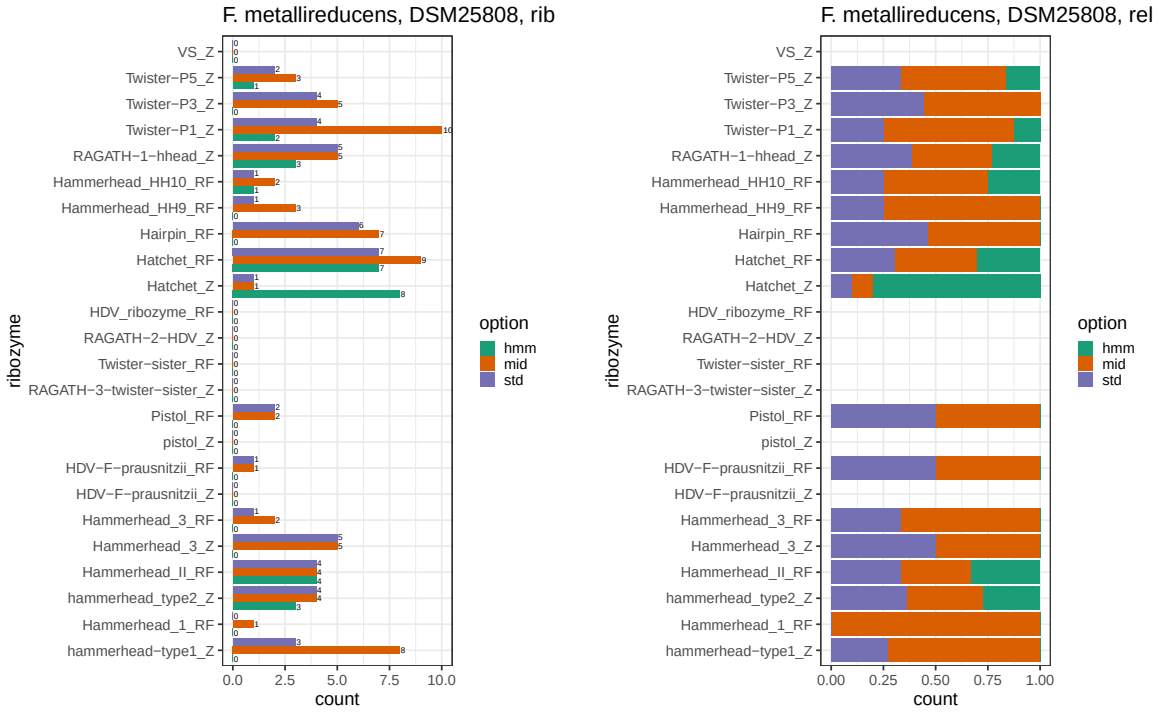


Figure A.5: **DSM1294: E-value vs. length for *cm-search* significant hits (e-value ≤ 0.05) in DSM1294.** Left: Twister-P1, right: Twister-P5

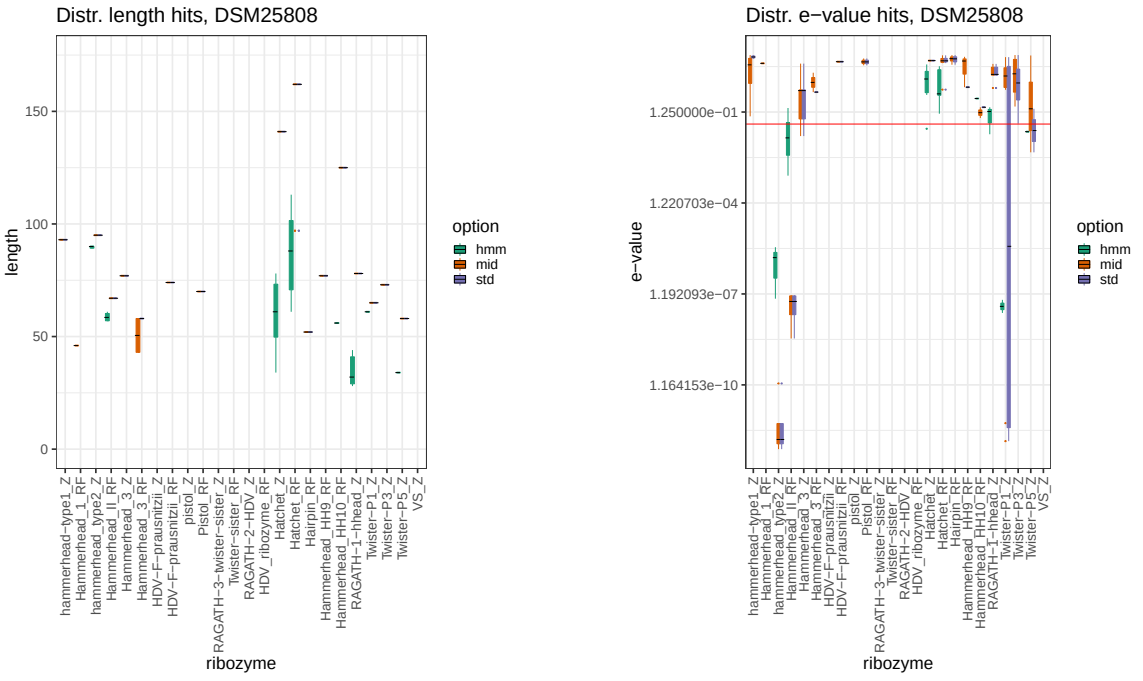
Table A.3: **Counts of found ribozymes in Fervidicella metallireducens (DSM25808).** *cm-search* options: hmm, mid, std. In brackets is the number without e-value limit.

Ribozyme	hmm	mid	std
Twister-P5_Z	1 (1)	1 (3)	1 (2)
Twister-P3_Z	0 (0)	0 (5)	0 (4)
Twister-P1_Z	2 (2)	2 (10)	2 (4)
RAGATH-1-HH_Z	1 (3)	0 (5)	0 (5)
HH_10_RF	0 (1)	0 (2)	0 (1)
HH_9_RF	0 (0)	0 (3)	0 (1)
Hairpin_RF	0 (0)	0 (7)	0 (6)
Hatchet_RF	0 (7)	0 (9)	0 (7)
Hatchet_Z	1 (8)	0 (1)	0 (1)
pistol_Z	0 (0)	0 (2)	0 (2)
HDV-F-prausnitzii_RF	0 (0)	0 (1)	0 (1)
HH_3_RF	0 (0)	1 (2)	1 (1)
HH_3_Z	0 (0)	0 (5)	0 (5)
HH_II_RF	3 (4)	4 (4)	4 (4)
HH_type2_Z	3 (3)	4 (4)	4 (4)
HH_1_RF	0 (0)	0 (1)	0 (0)
HH-type1_Z	0 (0)	0 (8)	0 (3)



(a) (b)

Figure A.6: **Hits in *cm-search* for F. metallireducens (DSM25808). Hits with different search options without a threshold in e-value.** Absolute count (a) , relative count (b).



(a) (b)

Figure A.7: **Distribution of length (a) and e-value (b) of *cm-search* hits in DSM25808.** No e-value threshold

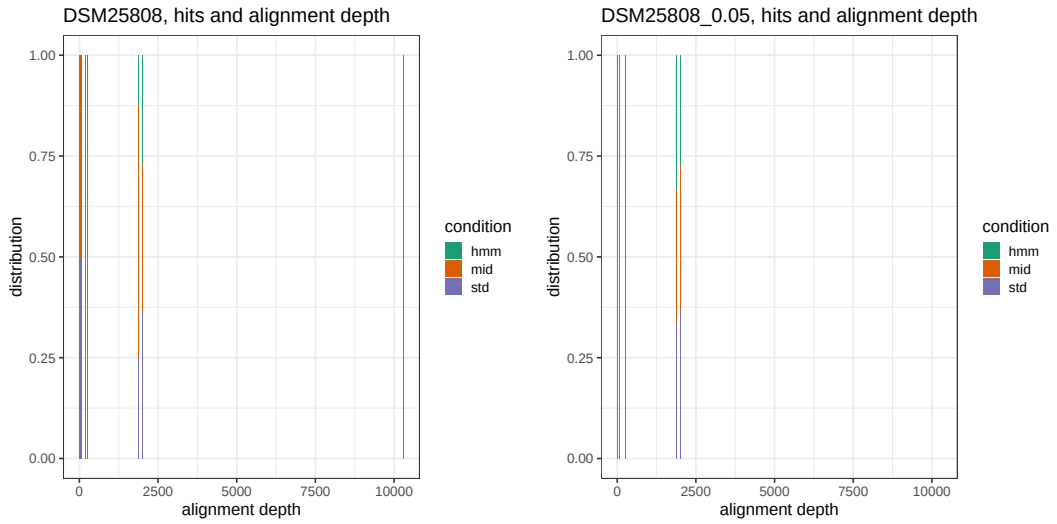
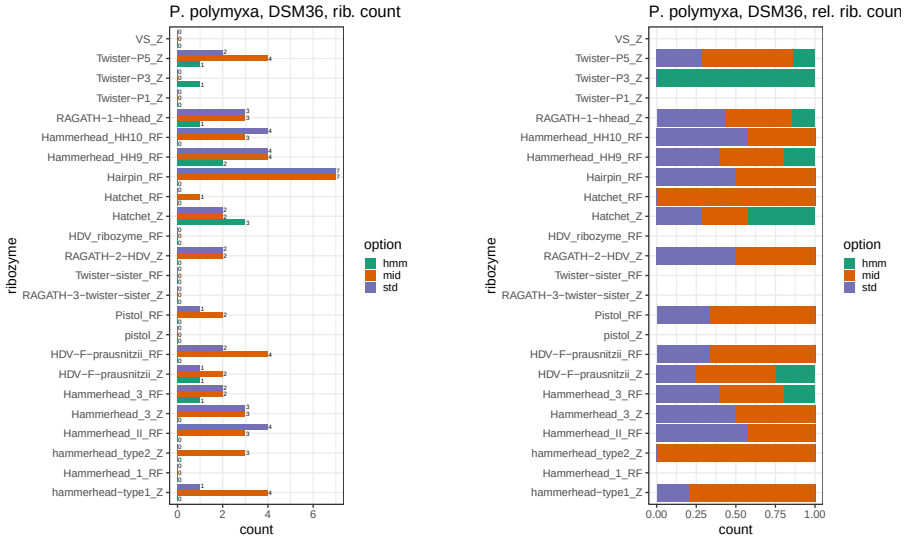


Figure A.8: **F. metallireducens(DSM25808): Alignment depth to procentual hits found by hmm, mid and std *cm-search*** Left: without threshold in e-value. Right: e-value ≤ 0.05

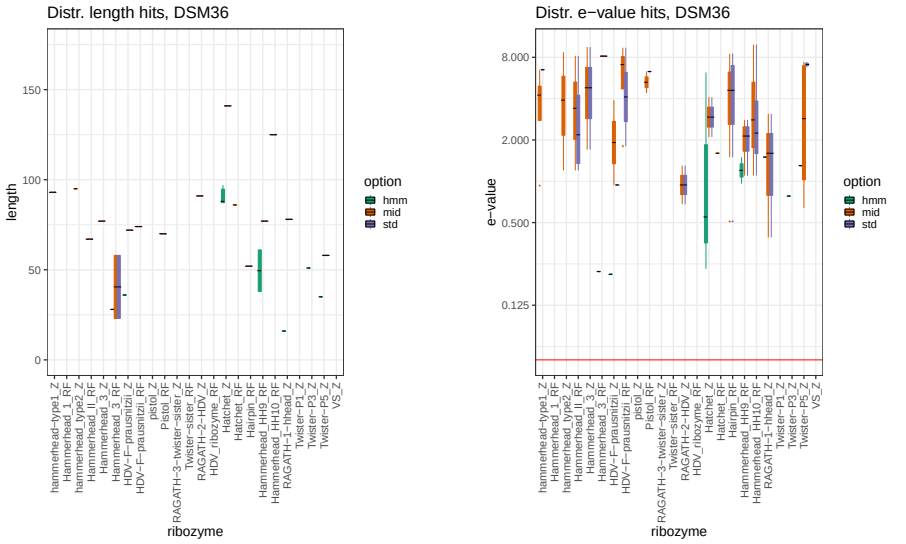
Table A.4: **Paenibacillus polymyxa (DSM36): count of found ribozymes.** In brackets the count of hits without limit in e-value is shown.

Ribozyme	hmm	mid	std
Twister-P5_Z	0 (1)	0 (4)	0 (2)
Twister-P3_Z	0 (1)	0 (0)	0 (0)
RAGATH-1-HH_Z	0 (1)	0 (3)	0 (3)
HH_10_RF	0 (0)	0 (3)	0 (4)
HH_9_RF	0 (2)	0 (4)	0 (4)
Hairpin_RF	0 (0)	0 (7)	0 (7)
Hatchet_RF	0 (0)	0 (1)	0 (0)
Hatchet_Z	0 (3)	0 (2)	0 (2)
Pistol_RF	0 (0)	0 (2)	0 (1)
HDV-F-prausnitzii_RF	0 (0)	0 (4)	0 (2)
HDV-F-prausnitzii_Z	0 (1)	0 (2)	0 (1)
HH_3_RF	0 (1)	0 (2)	0 (2)
HH_3_Z	0 (0)	0 (3)	0 (3)
HH_II_RF	0 (0)	0 (3)	0 (4)
HH_type2_Z	0 (3)	0 (3)	0 (0)
HH-type1_Z	0 (0)	0 (4)	0 (1)



(a) (b)

Figure A.9: **DSM36: count of found ribozymes.** Absolute counts (a) vs. relative counts (b). However, there where no hits found for e-value ≤ 0.05.



(a) **Length distribution of found ribozymes in P.polymyxa** (b) **E-value distribution of found ribozymes in DSM36.**

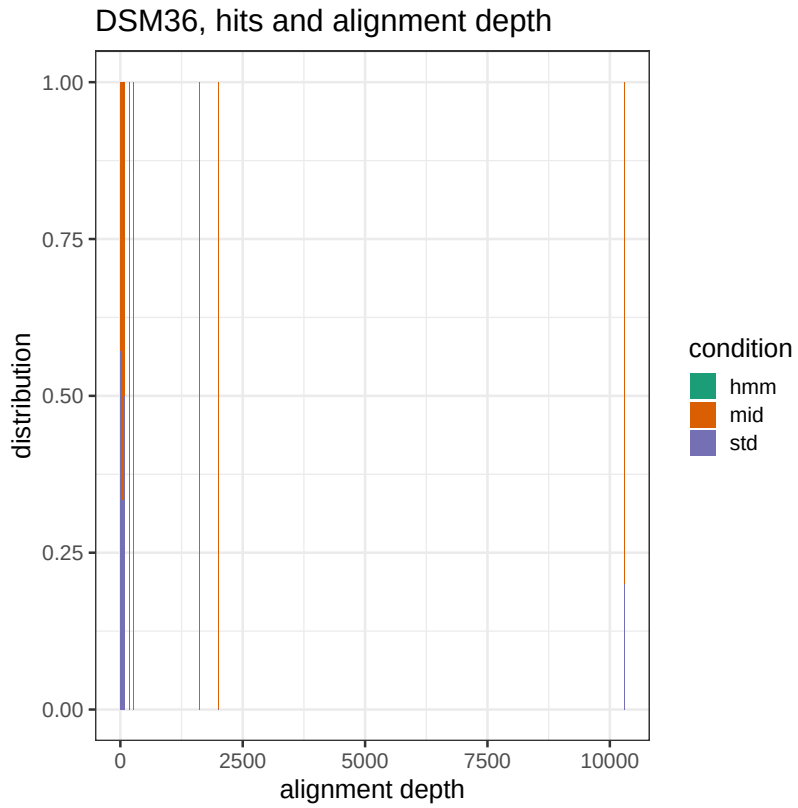
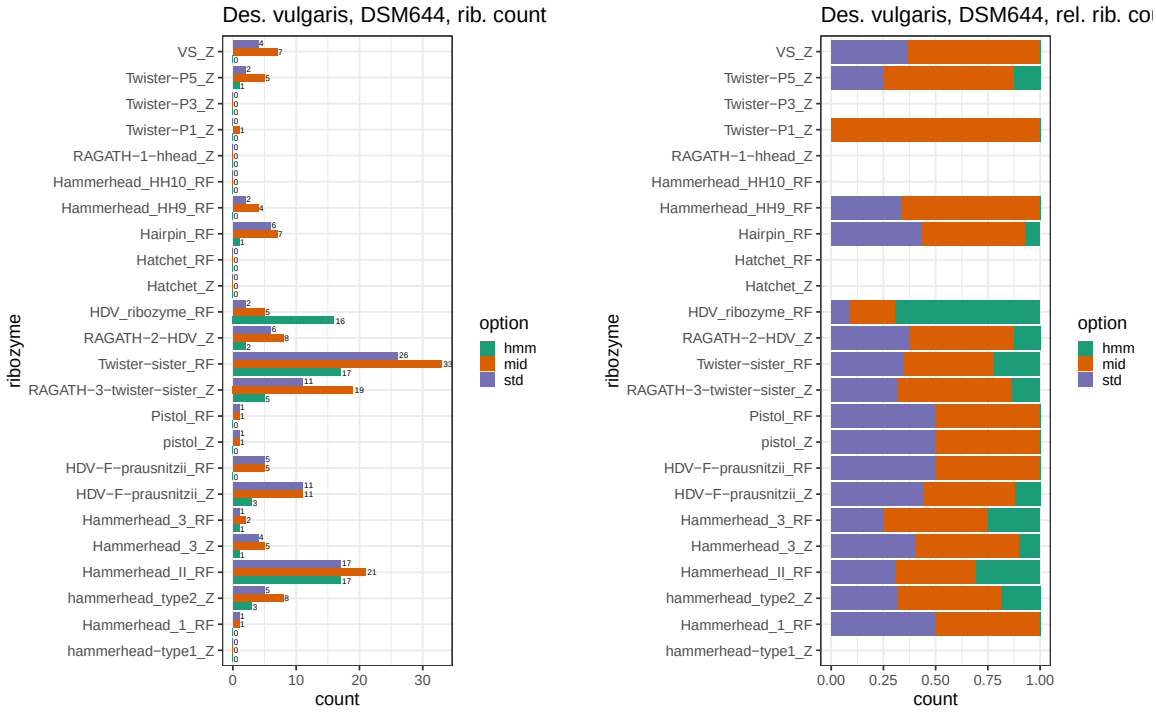


Figure A.11: **DSM36: Alignment depth to procentual hits found by hmm, mid and std *cm-search***

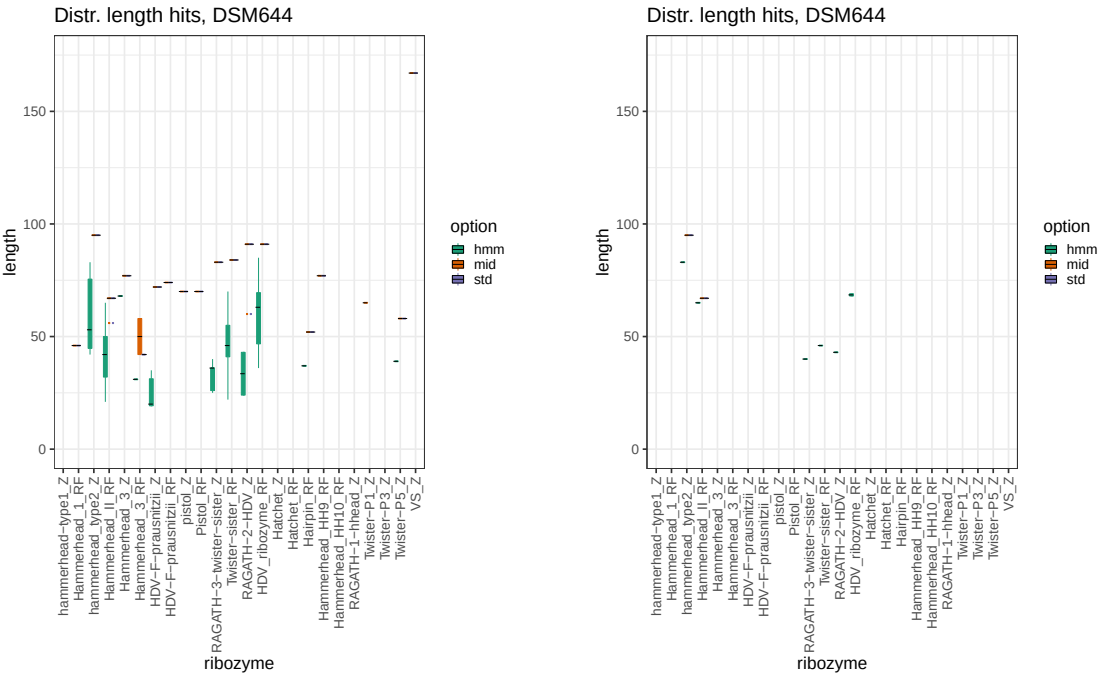
Table A.5: **Found Ribozymes in DSM644 (Desulfovibrio vulgaris).** *cm-search* options: hmm, mid, std. In brackets is the number without e-value limit.

Ribozyme	hmm	mid	std
VS_Z	0 (0)	0 (7)	0 (4)
Twister-P5_Z	2 (2)	0 (5)	0 (2)
Twister-P1_Z	0 (0)	0 (1)	0 (0)
HH_9_RF	0 (0)	0 (4)	0 (2)
Hairpin_RF	0 (1)	0 (7)	0 (6)
HDV_ribozyme_RF	2 (16)	0 (5)	0 (2)
RAGATH-2-HDV_Z	1 (2)	0 (8)	0 (6)
Twister-sister_RF	1 (17)	0 (33)	0 (26)
Twister-sister_Z	1 (5)	0 (19)	0 (11)
Pistol_RF	0 (0)	0 (1)	0 (1)
pistol_Z	0 (0)	0 (1)	0 (1)
HDV-F-prausnitzii_RF	0 (0)	0 (5)	0 (5)
HDV-F-prausnitzii_Z	0 (3)	0 (11)	0 (11)
HH_3_RF	0 (1)	0 (2)	0 (1)
HH_3_Z	0 (1)	0 (5)	0 (4)
HH_II_RF	3 (17)	3 (21)	3 (17)
HH_type2_Z	1 (3)	3 (8)	3 (5)
HH_1_RF	0 (0)	0 (1)	0 (1)



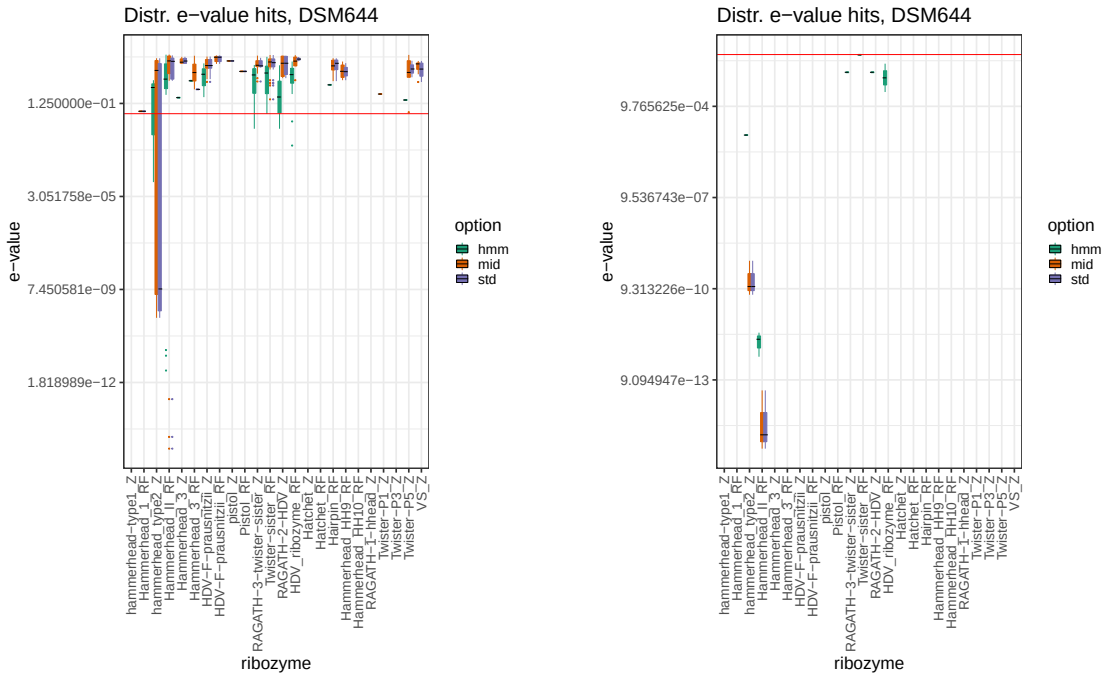
(a) (b)

Figure A.12: **DSM644: Absolute and relative hit count for found ribozymes with *cm-search*. Hits with different search options and without a threshold in e-value** Absolute count (a), relative count (b).



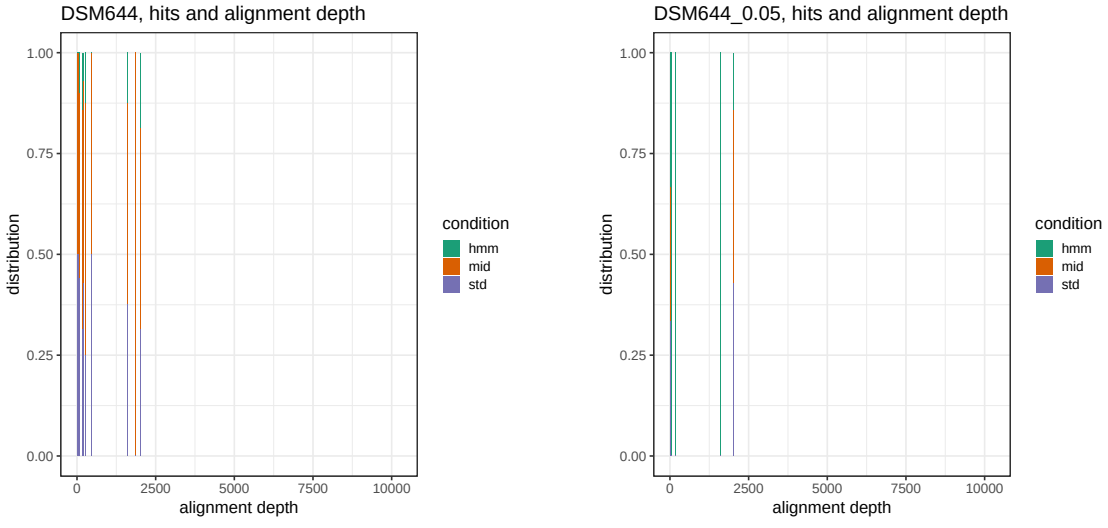
(a) (b)

Figure A.13: **Length for ribozyme hits with *cm-search* in DSM644.** All hits (a) vs. hits with an e-value ≤ 0.05 (b).



(a) (b)

Figure A.14: **E-value distribution for ribozyme hits with *cm-search* in DSM644.** All hits (a) vs. hits with an e-value ≤ 0.05 (b).

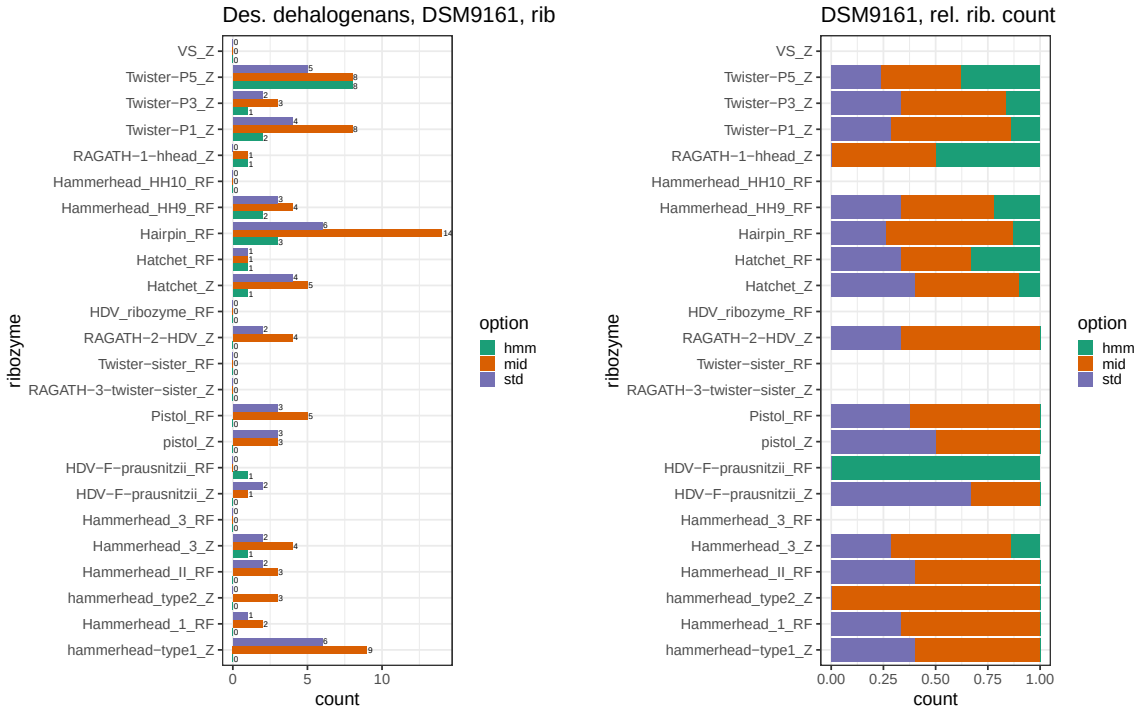


(a) (b)

Figure A.15: **DSM644: Alignment depth to procentual hits found by hmm, mid and std *cm-search*.** Without threshold in e-value (a). E-value limit ≤ 0.05 (b).

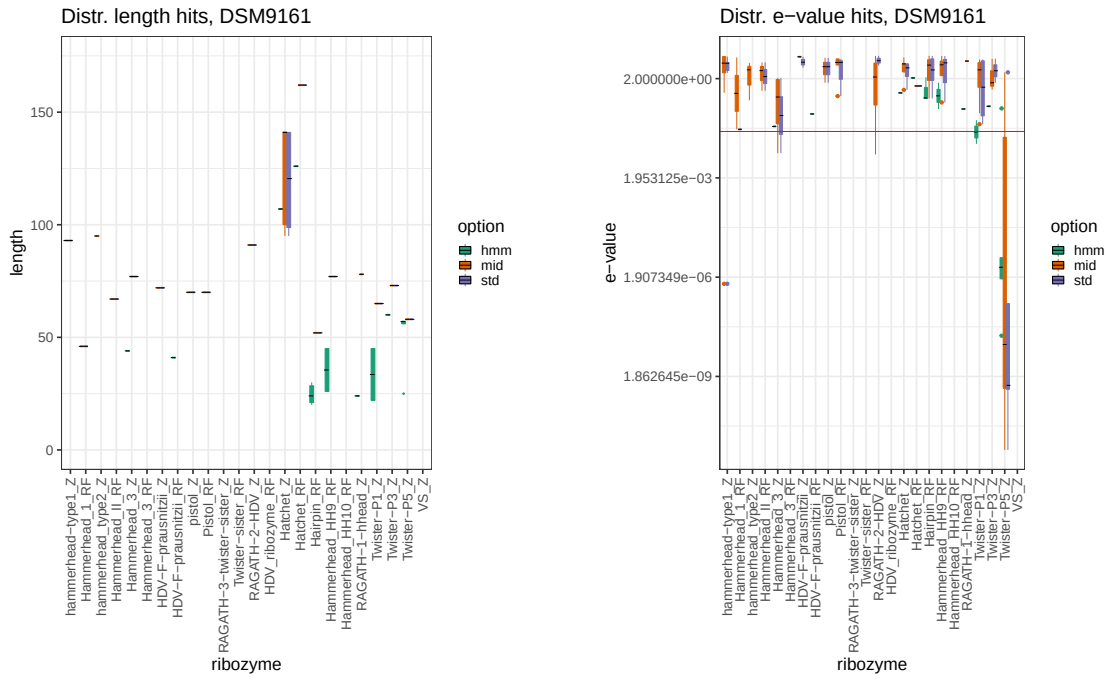
Table A.6: **Absolute counts of found ribozymes in DSM9161.** Significant (e-value ≤ 0.05) hit count (all hits).

Ribozyme	hmm	mid	std
Twister-P5_Z	8 (8)	8 (8)	4 (5)
Twister-P3_Z	0 (1)	0 (3)	0 (2)
Twister-P1_Z	1 (2)	0 (8)	0 (4)
RAGATH-1-HH_Z	0 (0)	0 (1)	0 (1)
HH_9_RF	0 (2)	0 (4)	0 (3)
Hairpin_RF	0 (3)	0 (14)	0 (6)
Hatchet_RF	0 (1)	0 (1)	0 (1)
Hatchet_Z	0 (1)	0 (5)	0 (4)
RAGATH-2-HDV_Z	(0)	2 (4)	0 (2)
Pistol_RF	0 (0)	0 (5)	0 (3)
pistol_Z	0 (0)	0 (3)	0 (3)
HDV-F-prausnitzii_RF	0 (1)	0 (0)	0 (0)
HDV-F-prausnitzii_Z	0 (0)	0 (1)	0 (2)
HH_3_Z	0 (1)	2 (4)	1 (2)
HH_II_RF	0 (0)	3 (3)	3 (0)
HH_type2_Z	0 (0)	3 (2)	3 (1)
HH_1_RF	0 (0)	2 (9)	1 (6)



(a) (b)

Figure A.16: **DSM9161: Absolute (a) and relative (b) count of ribozyme hits with*cm-search*.** Hits with different search options without a threshold in e-value.



(a) (b)

Figure A.17: **Length and e-value of ribozyme hits in DSM9161.** Length (a), e-value (b).

Appendix B

Evaluation of *cm-search* hits - Intersection analyses

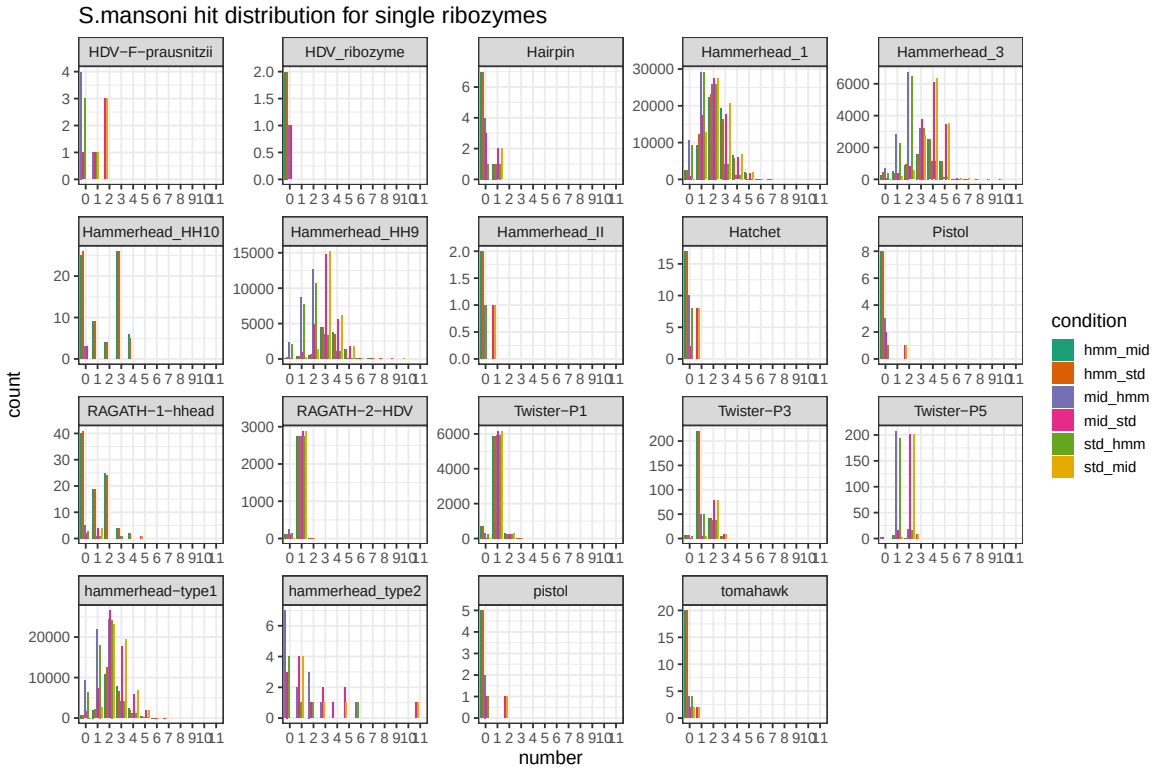


Figure B.1: **distribution of overlap hits on the single ribozyme types**

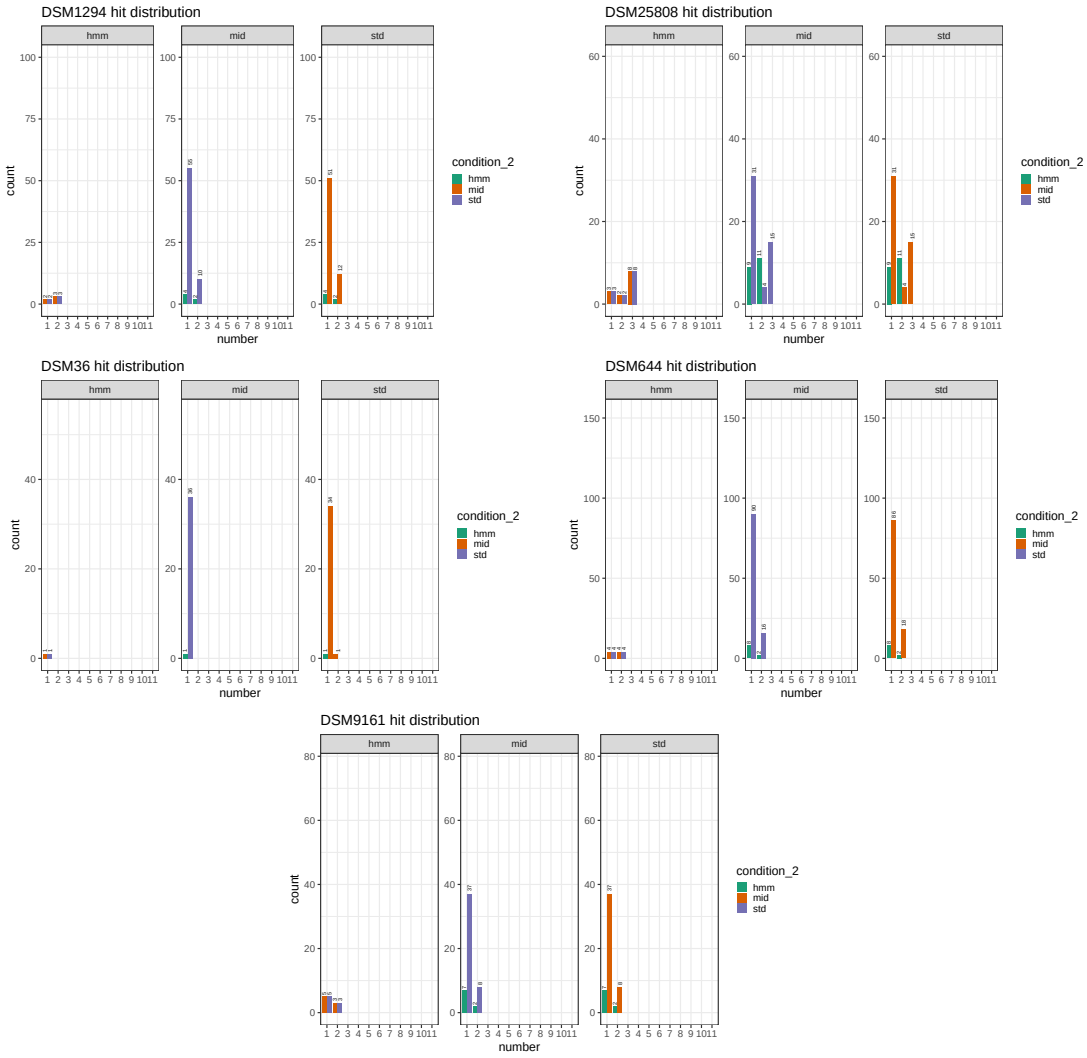


Figure B.2: **Bacteria: Number of hits dependend on the *cm-search* mode.** From upper left to down right: DSM1294, DSM25808, DSM36, DSM644, DSM9161

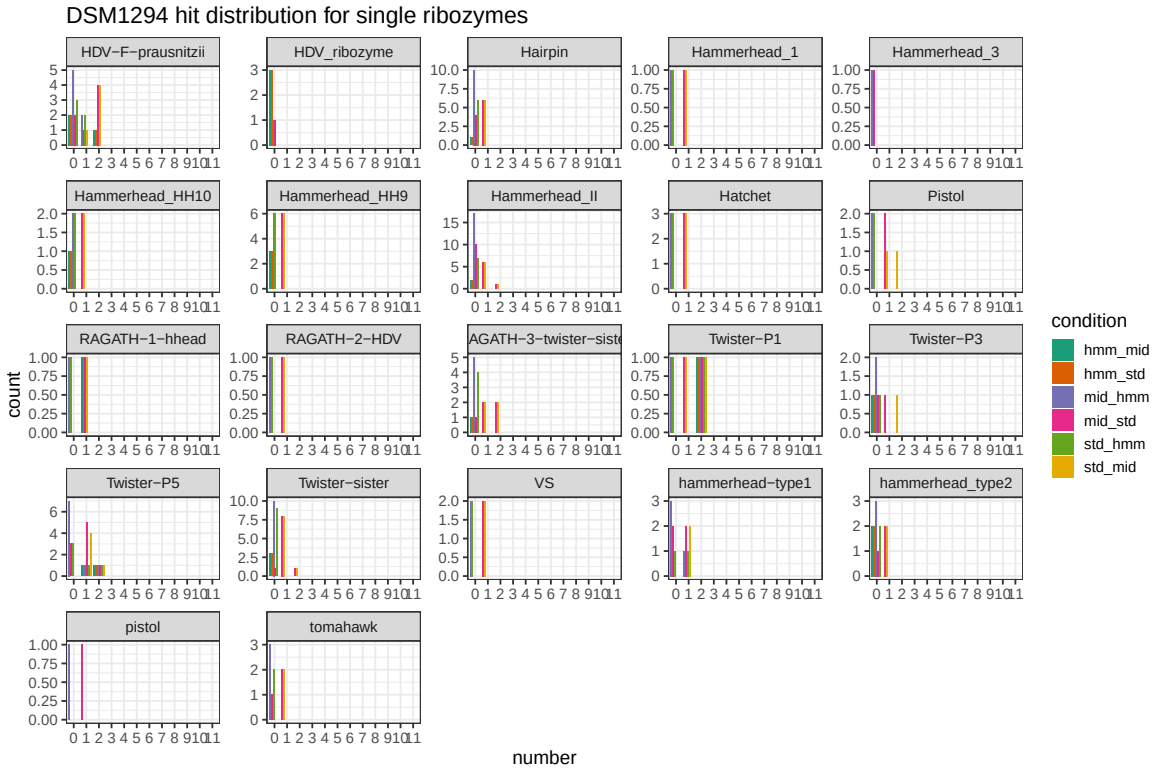


Figure B.3: **distribution of overlap hits on the single ribozyme types**

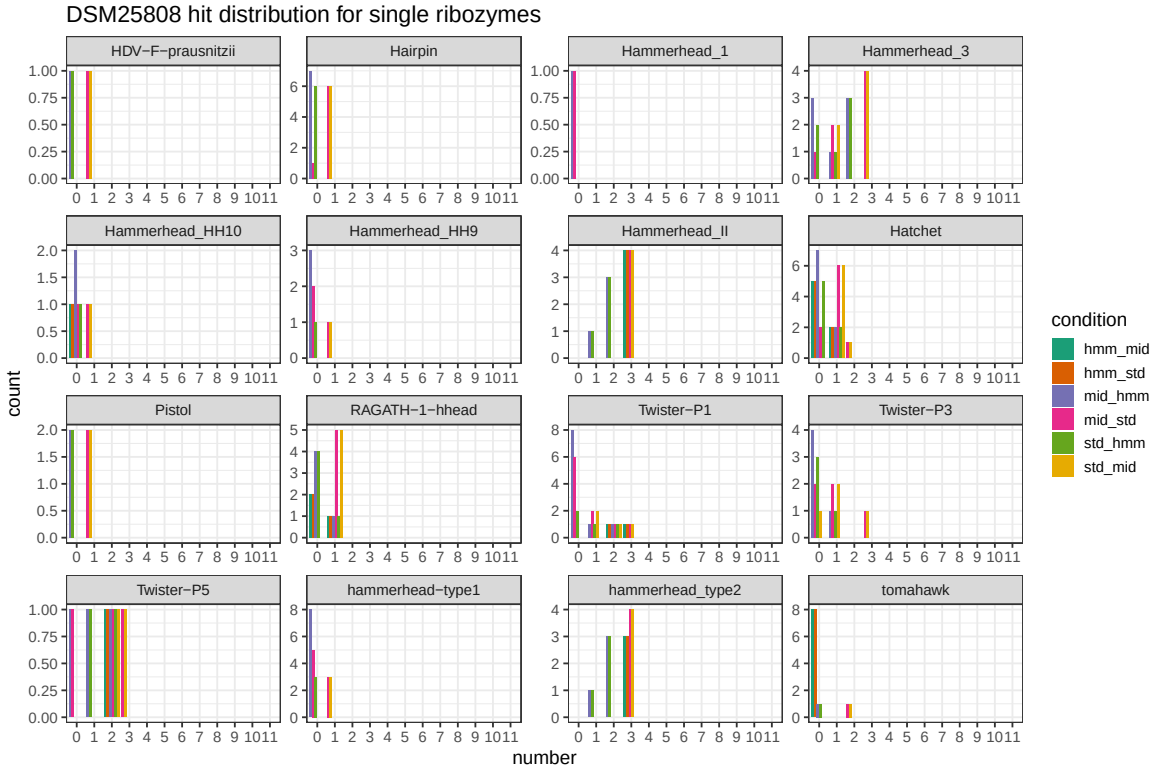


Figure B.4: **distribution of overlap hits on the single ribozyme types**

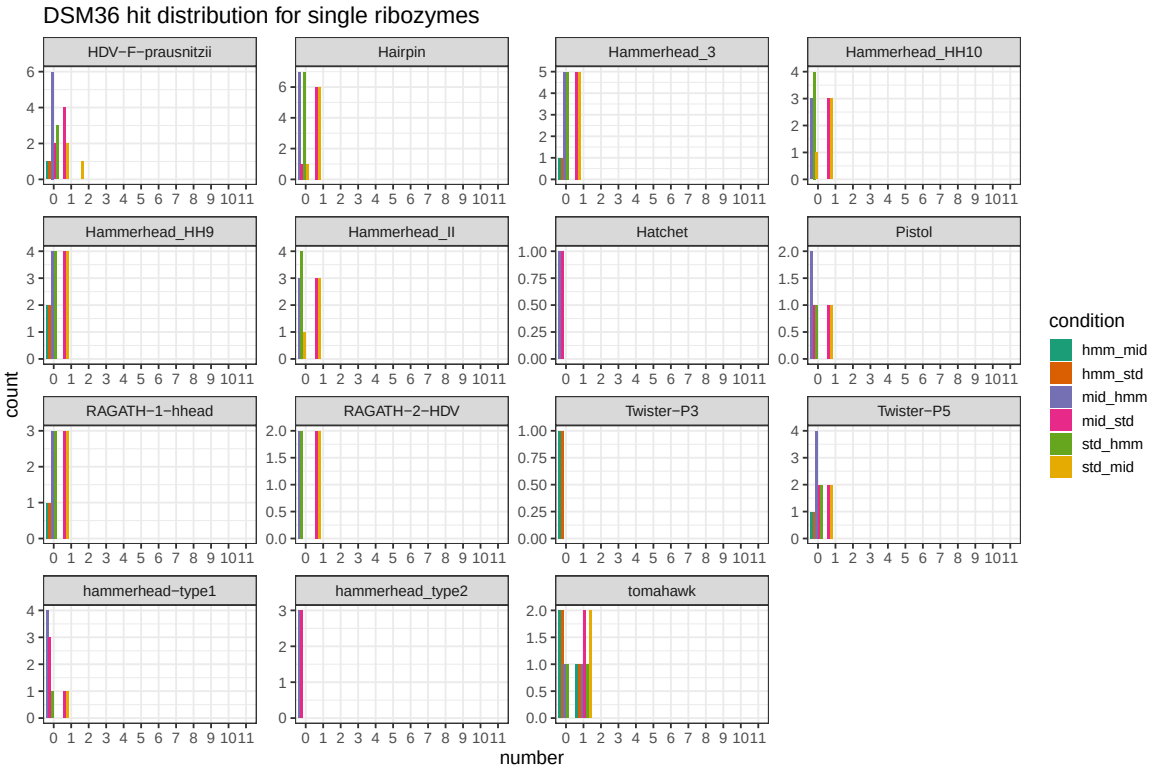


Figure B.5: **distribution of overlap hits on the single ribozyme types**

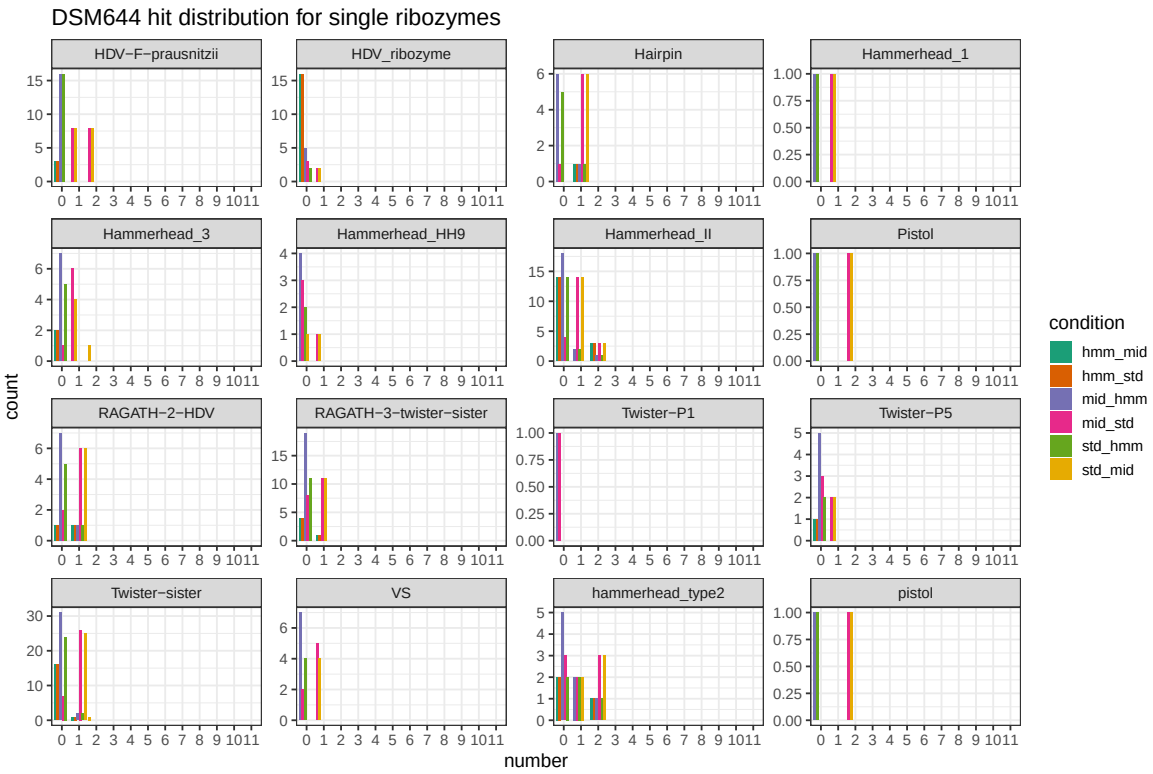


Figure B.6: **distribution of overlap hits on the single ribozyme types**

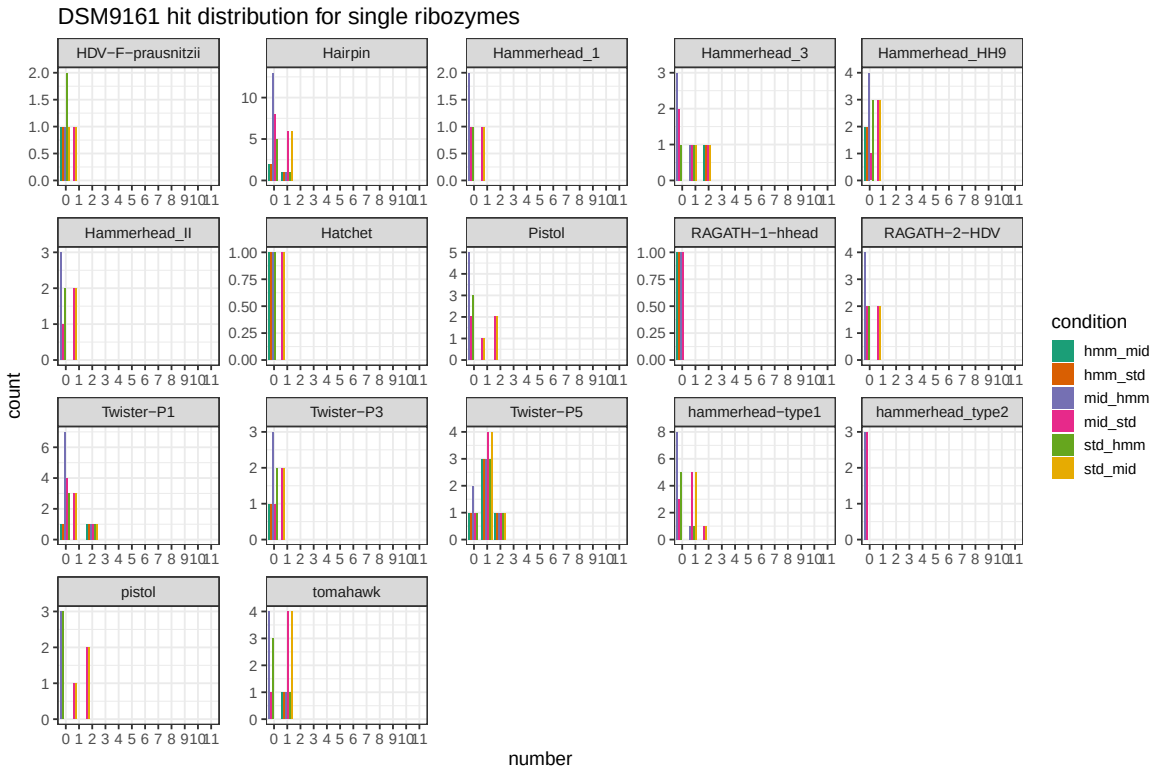
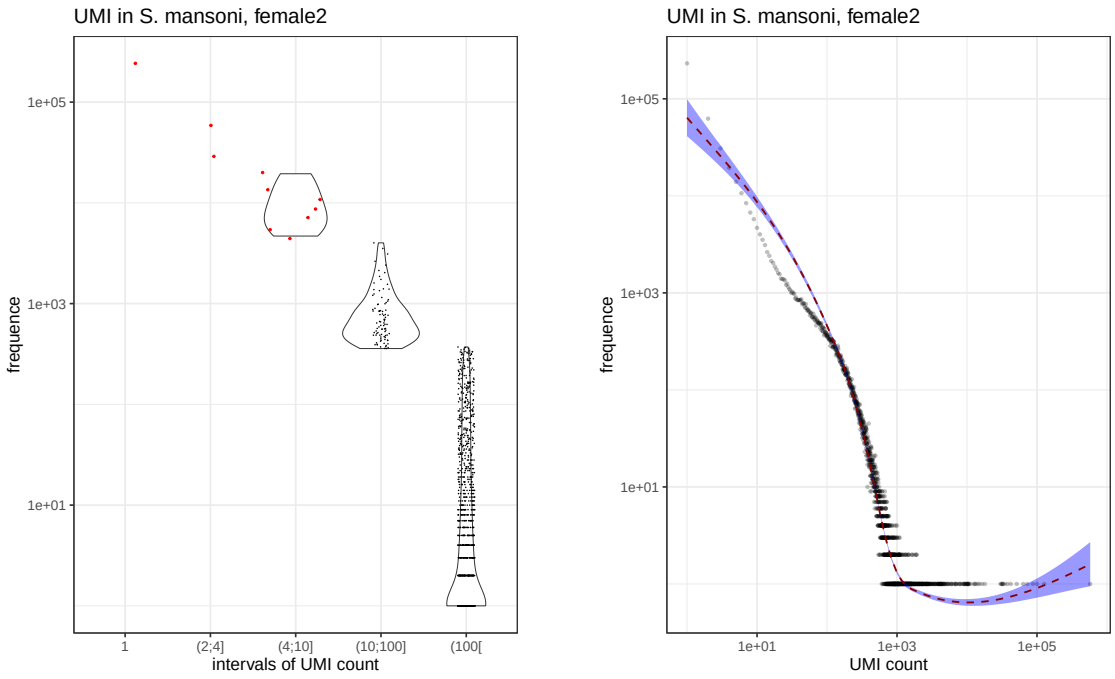


Figure B.7: **distribution of overlap hits on the single ribozyme types**

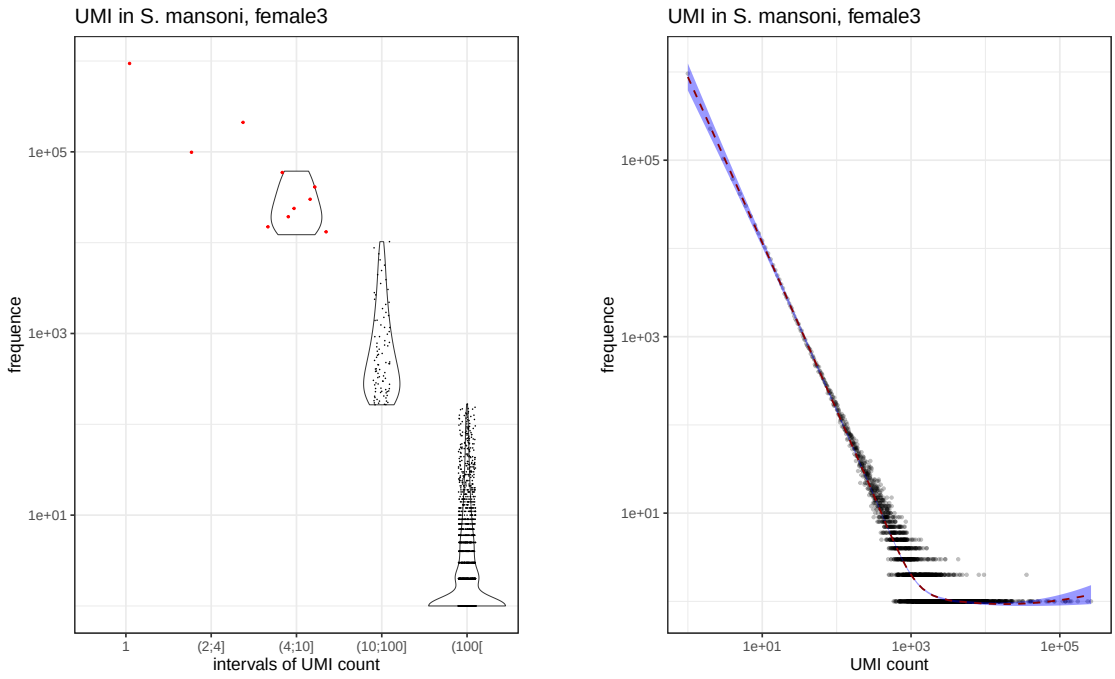
Appendix C

Analysis of UMI-tools results



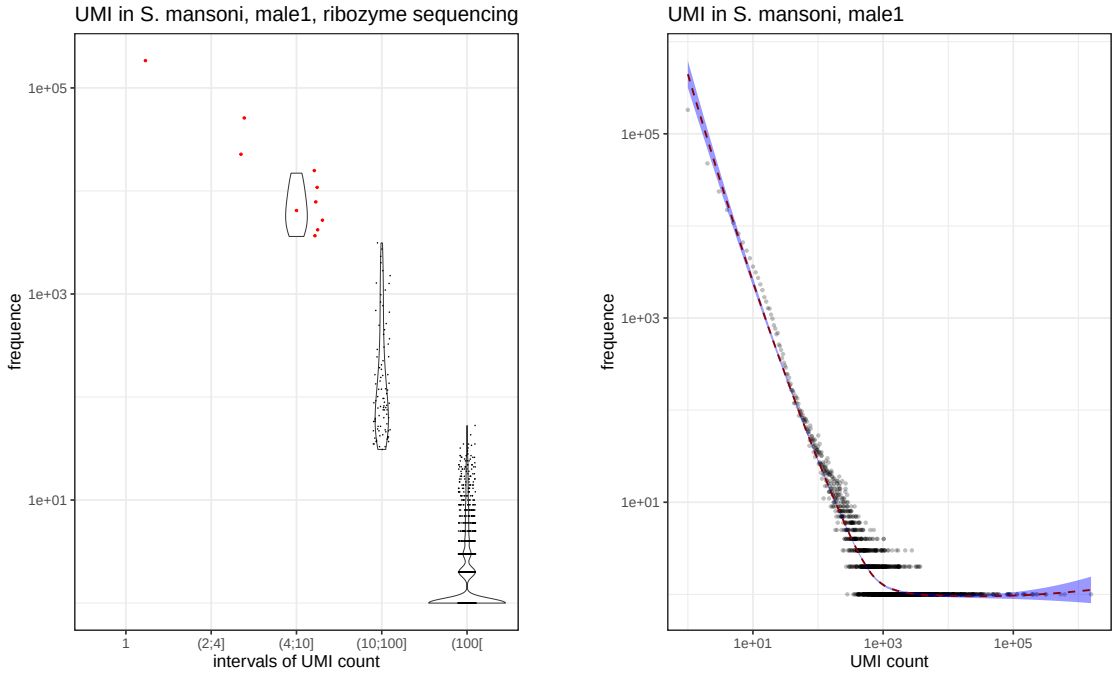
(a) (b)

Figure C.1: **Frequence of UMI in sequencing data of *S. mansoni*, female 2.**



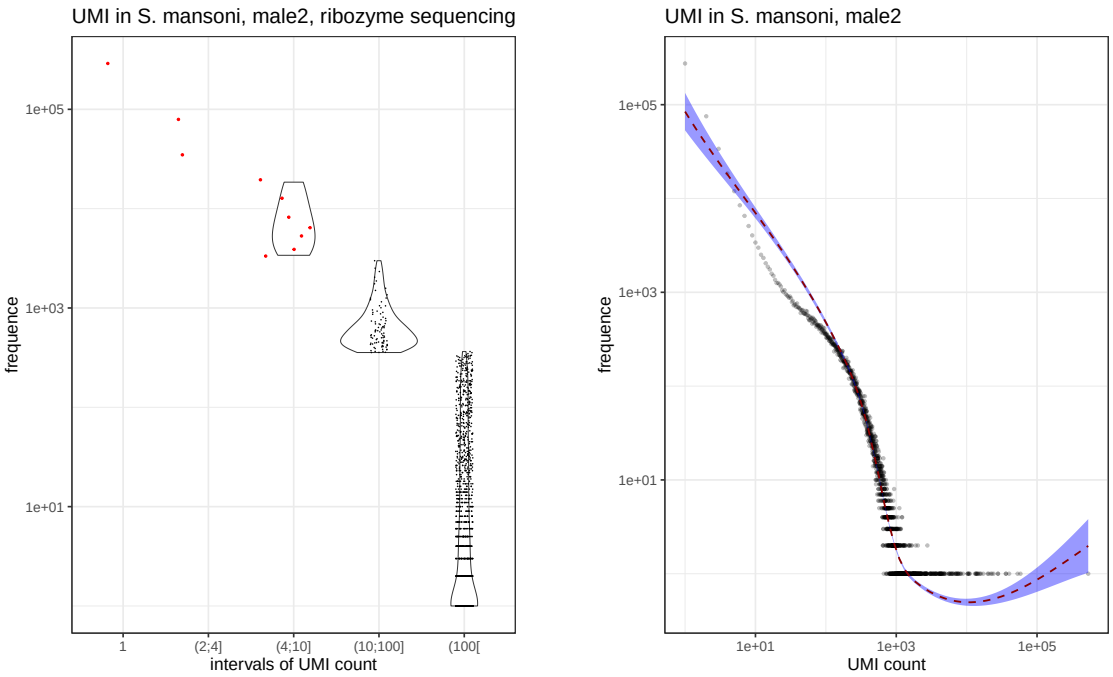
(a) (b)

Figure C.2: **Frequence of UMI in sequencing data of *S. mansoni*, female 3.**



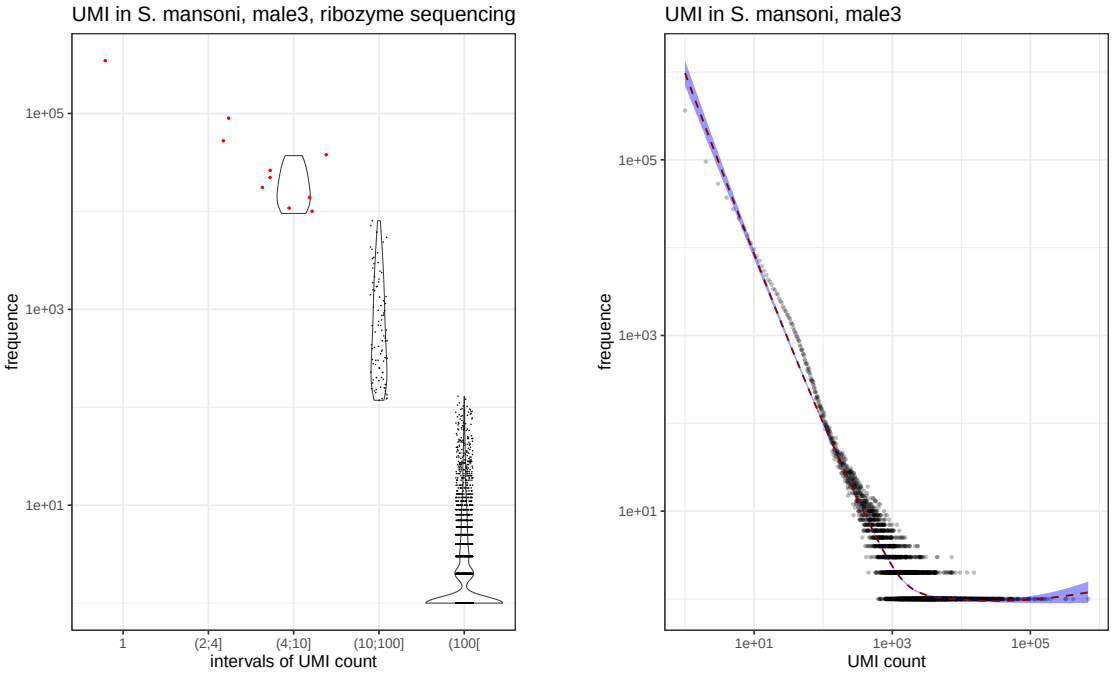
(a) (b)

Figure C.3: **Frequence of UMI in sequencing data of *S. mansoni*, male 1.**



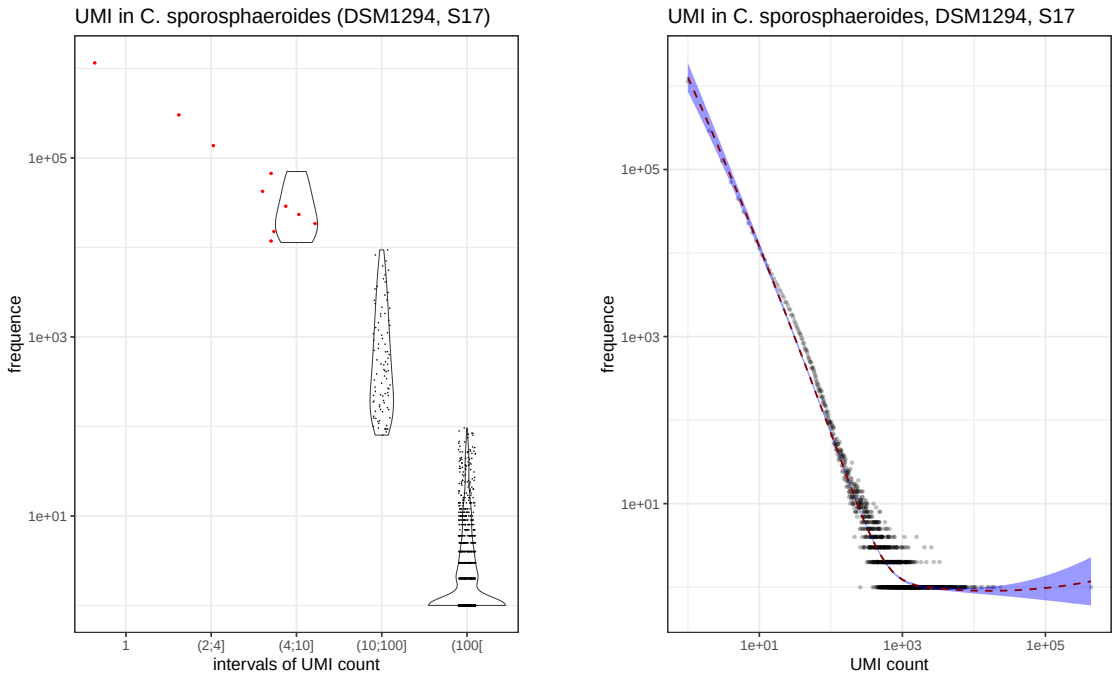
(a) (b)

Figure C.4: **Frequence of UMI in sequencing data of *S. mansoni*, male 2.**



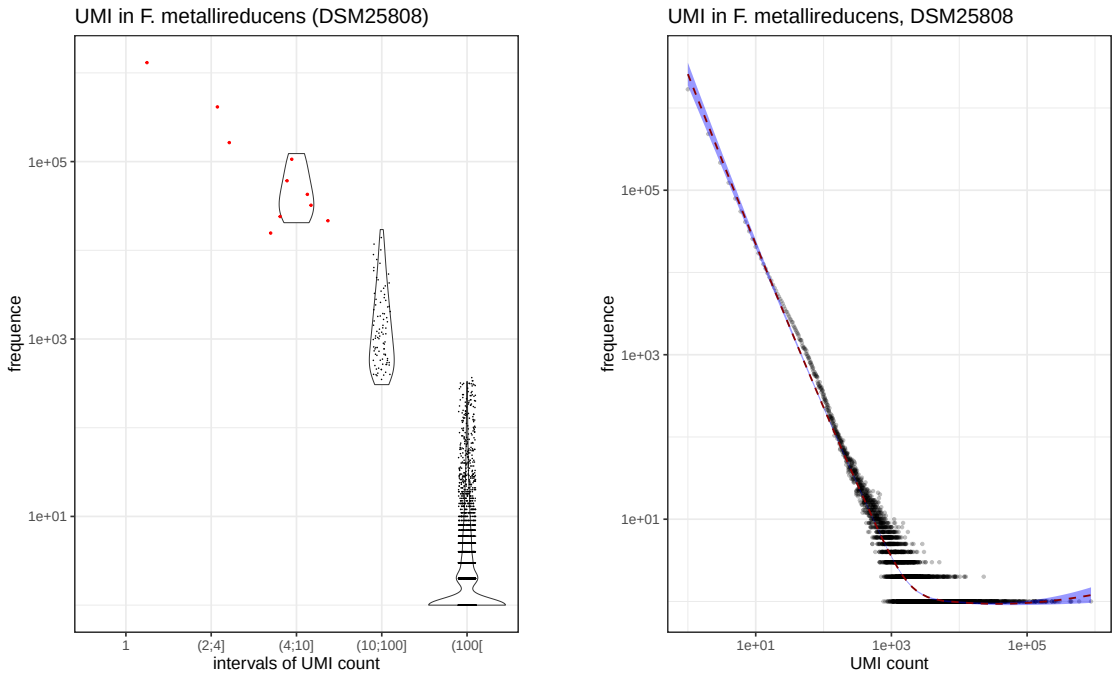
(a) (b)

Figure C.5: **Frequence of UMI in sequencing data of *S. mansoni*, male 3.**



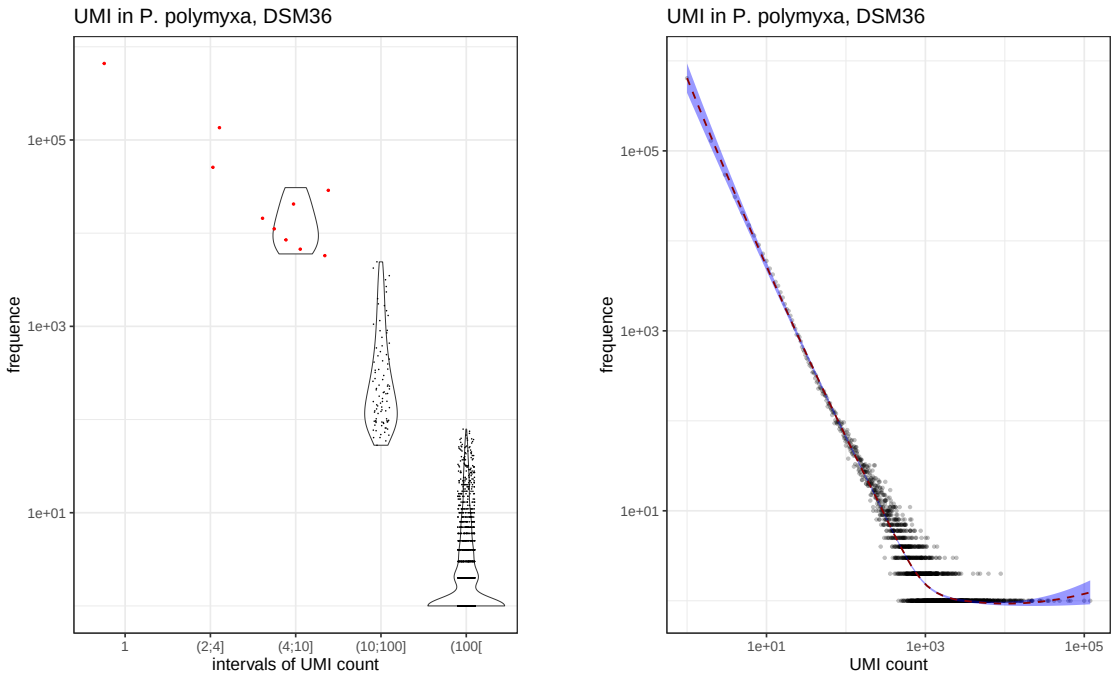
(a) (b)

Figure C.6: **Frequence of UMI in sequencing data of C. sporosphaeroides, DSM1294-1.** A Violin plot of frequence of binned occurence of read-UMI combinations. B Scatter plot of frequence of occurence of read-UMI combination together. In purple local smoothing of the values.



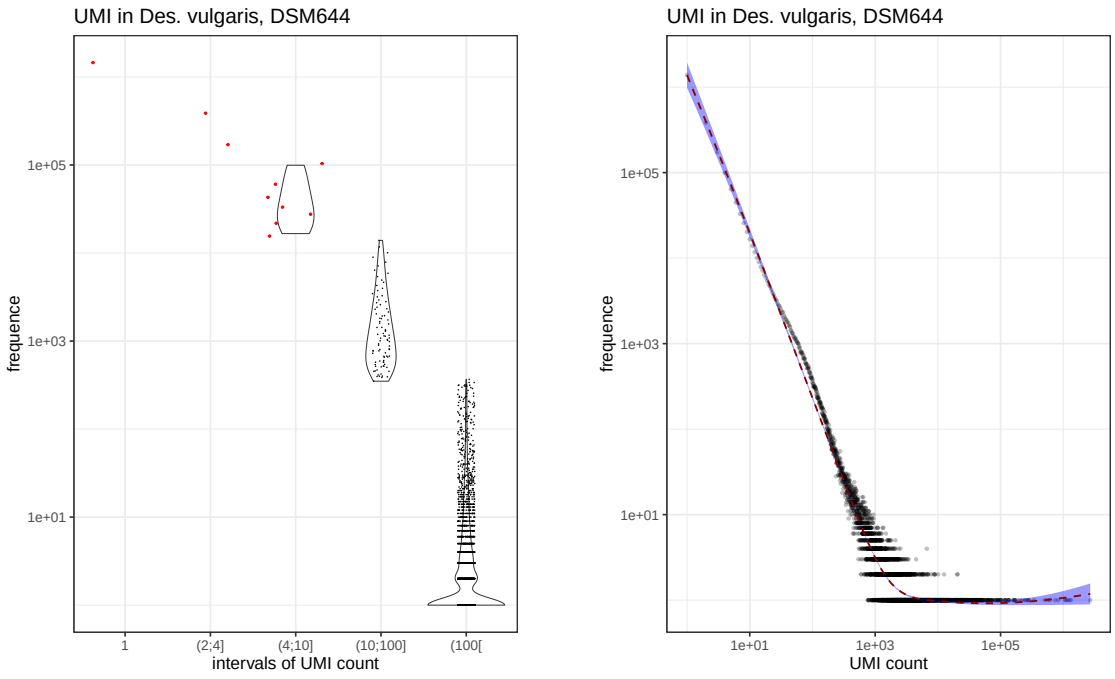
(a) (b)

Figure C.7: **Frequence of UMI in sequencing data of F. metallireducens, DSM25808**



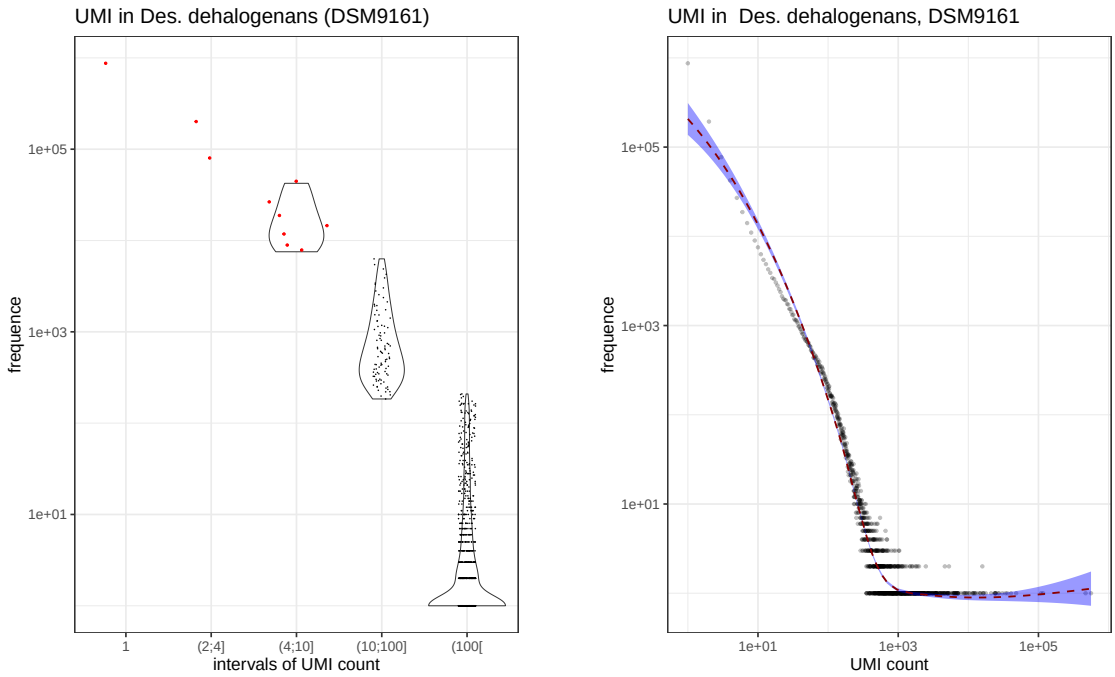
(a) (b)

Figure C.8: **Frequence of UMI in sequencing data of P.polymyxa, DSM36**



(a) (b)

Figure C.9: **Frequence of UMI in sequencing data of D. vulgaris, DSM644**



(a) (b)

Figure C.10: **Frequence of UMI in sequencing data of D.dehalogenans, DSM9161**

Appendix D

Peaks other than ribozymes

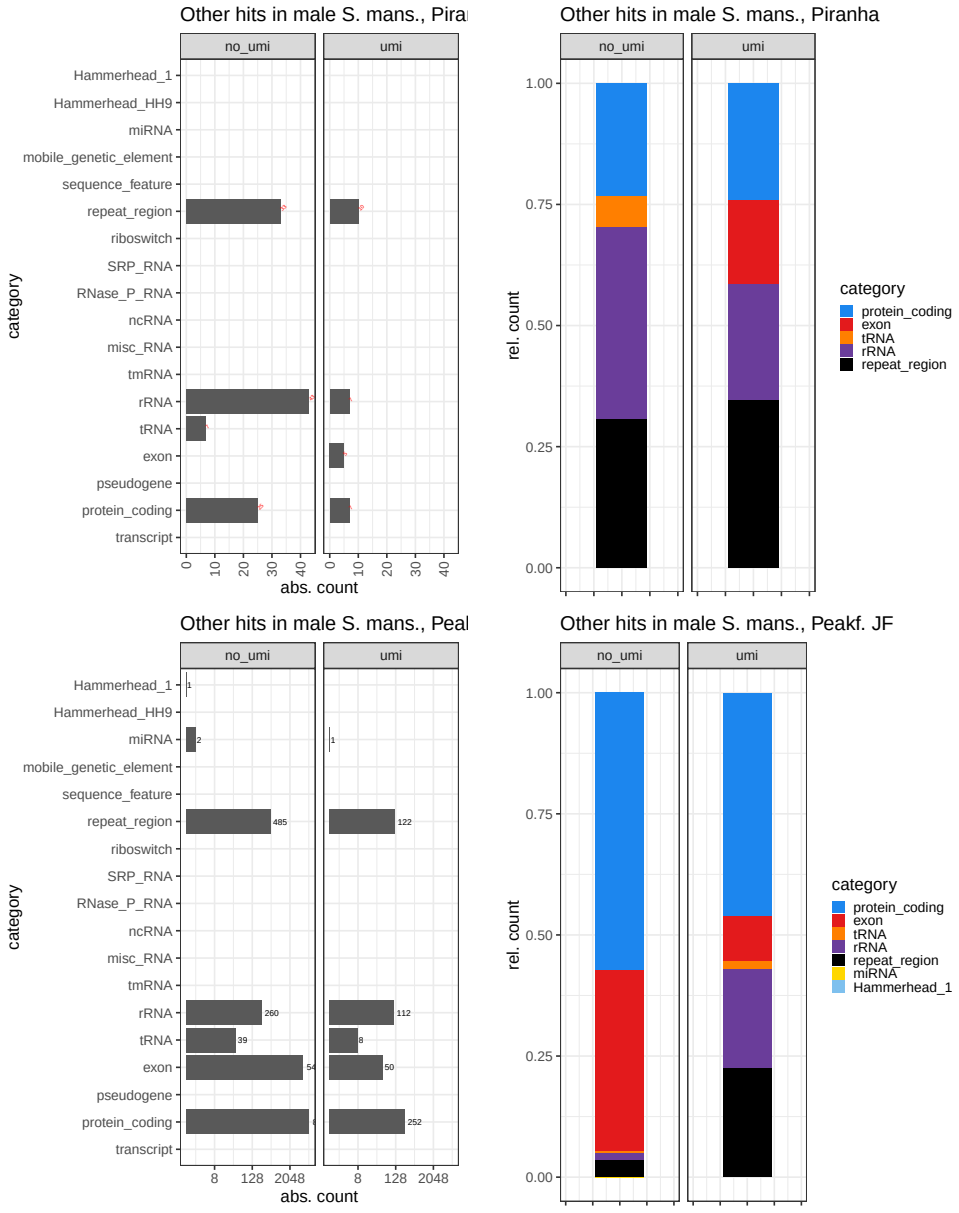


Figure D.1: **Intersection of peaks other than ribozymes with the "official" annotations in male *S. mansoni*.**

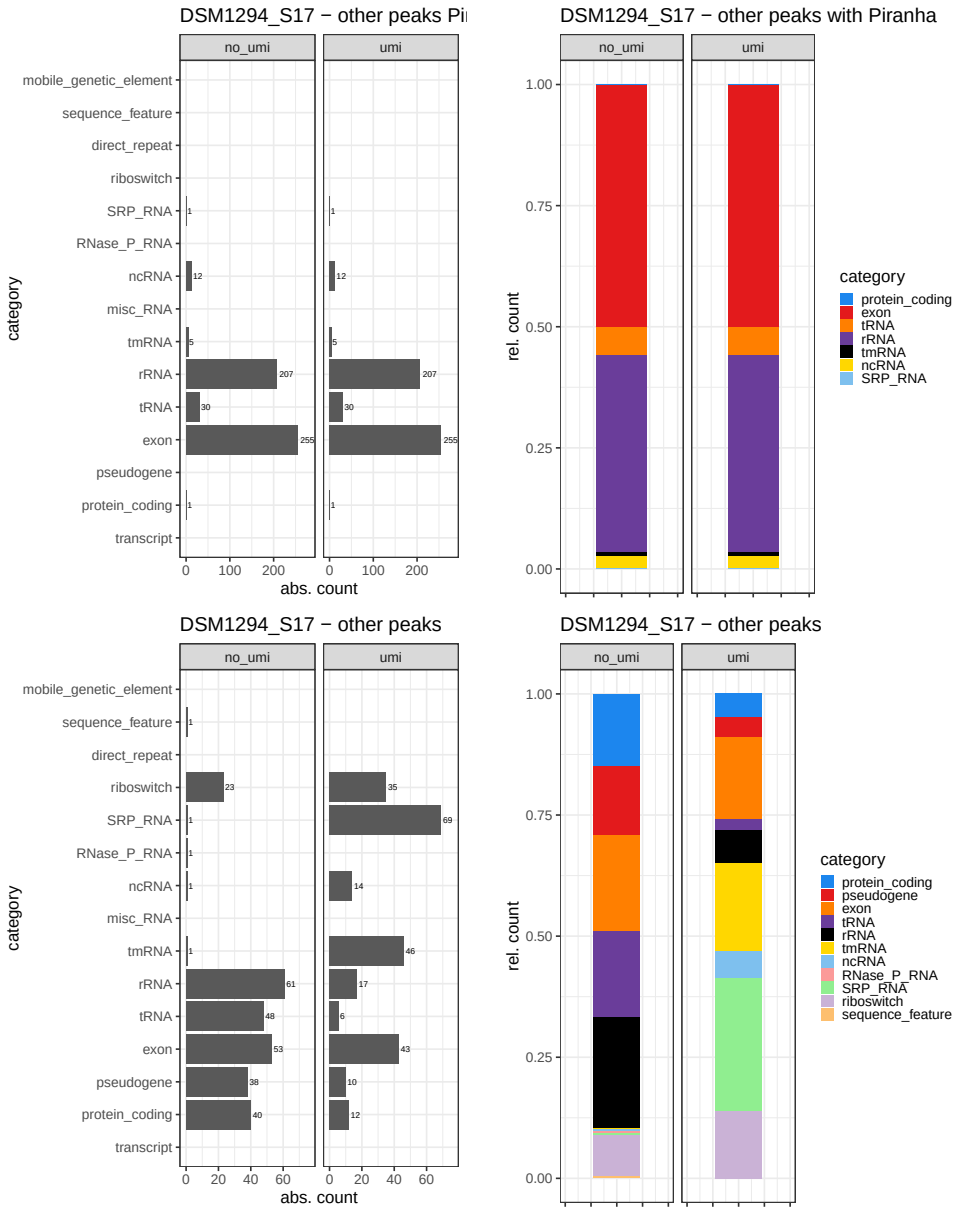


Figure D.2: **Intersection of peaks other than ribozymes with the "official" annotations in DSM1294_S17**

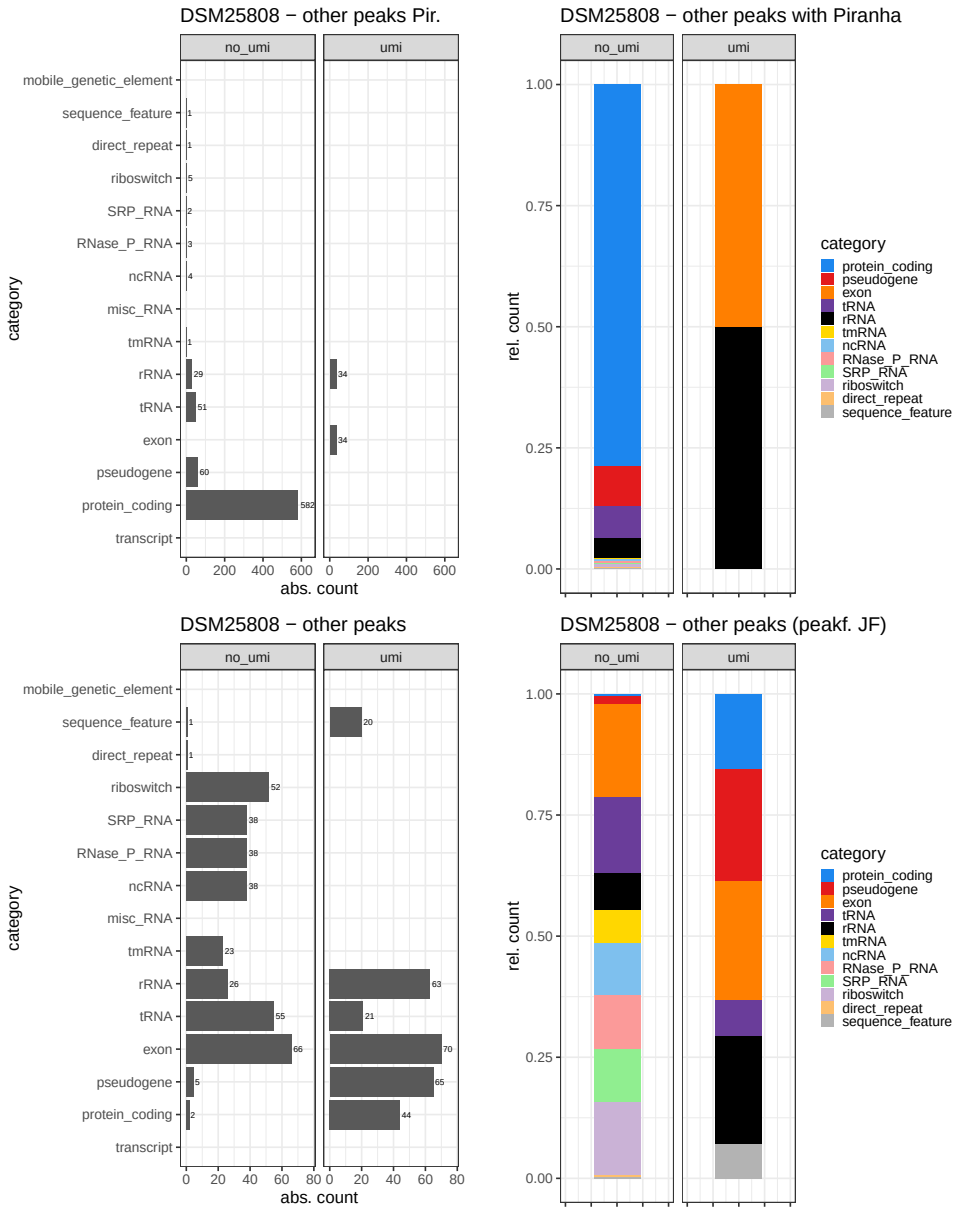


Figure D.3: **Intersection of peaks other than ribozymes with the "official" annotations in DSM25808**

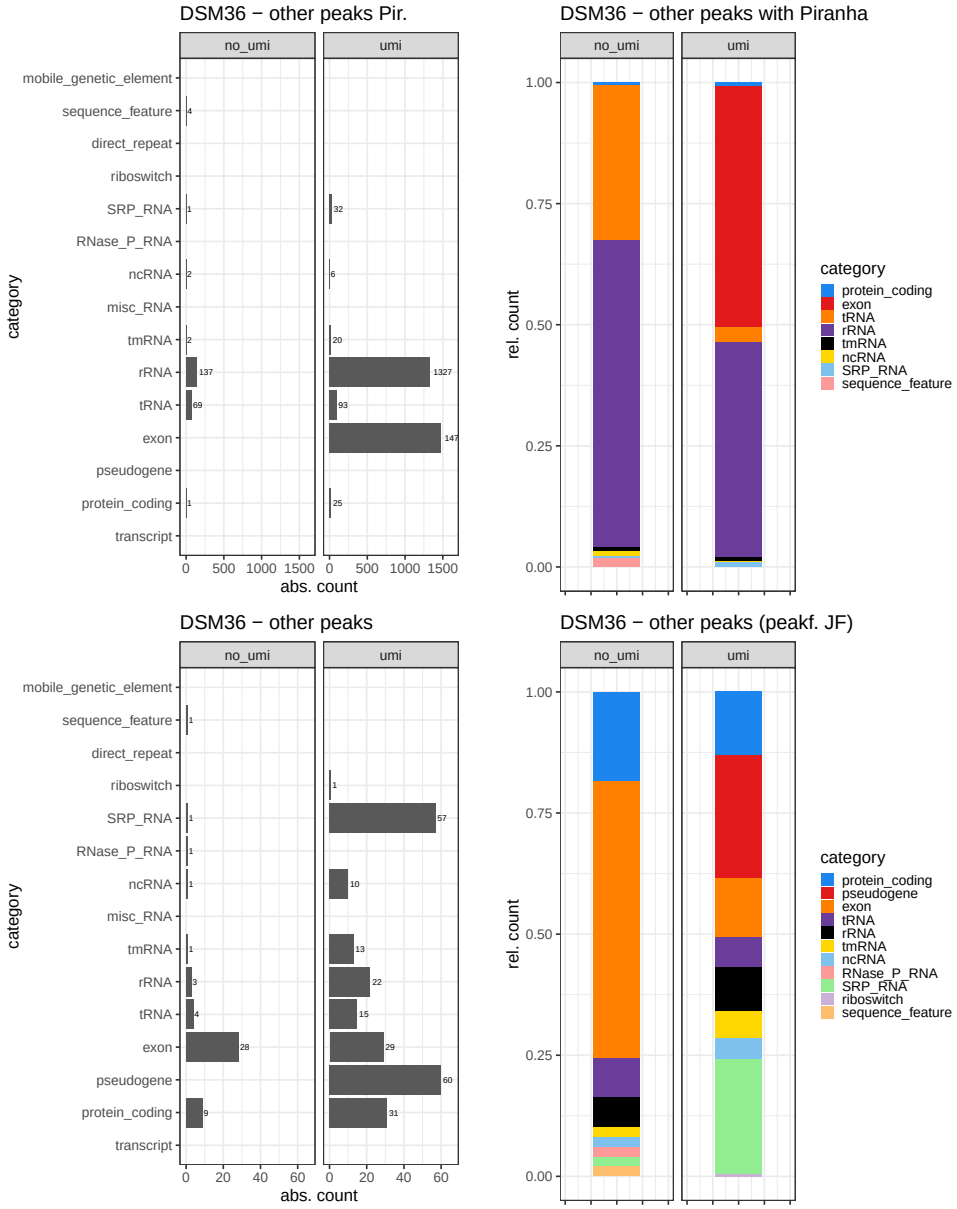


Figure D.4: **Intersection of peaks other than ribozymes with the "official" annotations in DSM36**

Selbstständigkeitserklärung

Ich versichere, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe, insbesondere sind wörtliche oder sinngemäße Zitate als solche gekennzeichnet. Mir ist bekannt, dass Zuwiderhandlung auch nachträglich zur Aberkennung des Abschlusses führen kann.

Ich versichere, dass das elektronische Exemplar mit den gedruckten Exemplaren übereinstimmt.

Leipzig, xx.xx.2020
Christiane Gärtner