VaulteR - A pipeline for vault associated RNA detection from RNA-sequencing

Yue Gao

September 2017

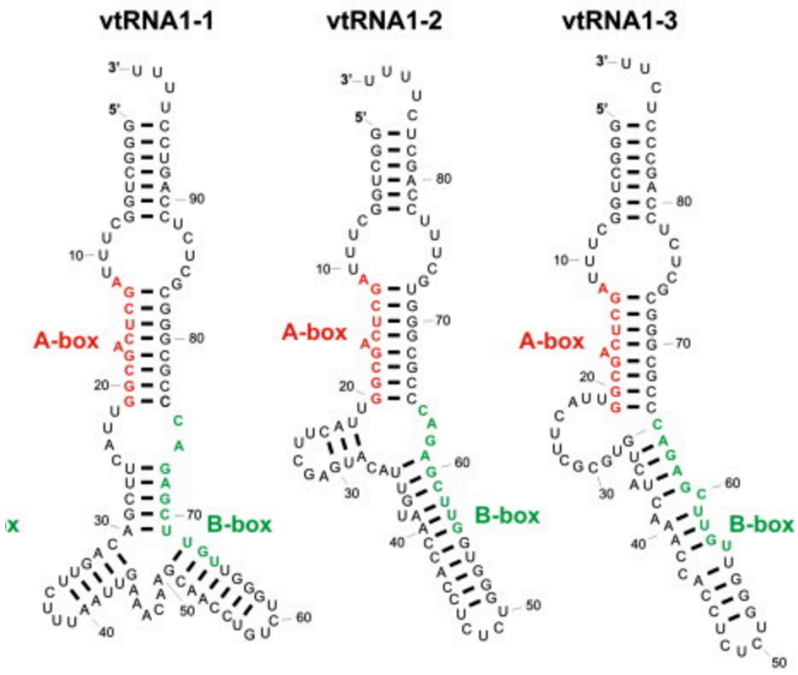
COMPUTATIONAL BIOLOGY UNIT(CBU) & Sea Lice Research Centre
Department of Informatics
University of Bergen

Master Degree Thesis

VaulteR - A pipeline for vault associated RNA detection from RNA-sequencing

by Yue Gao

September 2017



supervised by Michael Dondrup
COMPUTATIONAL BIOLOGY UNIT(CBU) & Sea Lice Research Centre
Department of Informatics
University of Bergen

Abstract

Vaults are highly conserved ribonucleoprotein complexes of unknown function. They have so far been found to be present in high numbers among higher eukaryotes including mammals, amphibians, and avians, as well as lower eukaryotes including deuterostomes and the slime mold (*Dictyostelium discoideum*). The aim of this thesis is to design a pipeline for vault associated RNA detection from RNA-sequences. And especially try to detect vtRNA in the Salmon Louse. The genome of the atlantic salmon louse, a major parasite of salmonids, affecting the global aquaculture industry.

The thesis presents three methods of detecting vtRNA, one way is to find the peaks in the alignment of reads and search for the high coverage sequences in Rfam to check the existence of vtRNA. Another way is by predicting the secondary structures of the high coverage sequences, drawing a dendrogram with hierarchical clusters according to the dissimilarity matrix of RNA secondary structures, and then analysing key features of secondary structures of the known vtRNA in order to filter the candidates. At last, the third method is by detecting motifs, such as A-Box and B-Box, in candidate sequences with the MEME Suite.

The result of this thesis is a pipeline that can effectively detect vtRNA, and a set of novel candidate sequences which can probably act as vtRNA in the salmon louse genome.

Acknowledgements

I would like to thank my supervisor, Michael Dondrup, from at the Department of Informatics, University of Bergen. Without his invaluable input, continued support and guidance this thesis would not have been possible. I would like to thank Christiane Eichner for providing the biological sample and extracting RNA for this thesis.

Table of contents

1 Introduction

1.1. Vaults and their structure

Vaults, first described in 1986, are large cytoplasmic ribonucleoprotein (RNP) particles found in nearly all eukaryotic cells. The vault complex is mainly comprised of four major components in multiple copies: major vault protein (MVP), two minor vault proteins (VPARP and TEP1), and a small untranslated RNA ranging between 80 and 150 nucleotides[1]. The particle is abundant in all cells of many higher eukaryotes and highly conserved throughout evolution; the high conservation of Vault protein sequences implies some kind of functional importance. Vaults may be able to open and close and Vault ribonucleoprotein particles open into flower-like structures, with octagonal symmetry[2]. vtRNA comprises less than 5% of the total mass of a vault particle and stoichiometric calculations on data from rat liver vaults suggest that each vtRNA is present in approximately 16 copies per particle[3]. This would therefore suggest that one RNA is associated with each petal. Vaults have been implicated in a broad range of cellular functions including nuclear-cytoplasmic transport, mRNA localization, drug resistance, cell signaling, nuclear pore assembly, and innate immunity[34]. It is also found that vaults (especially the MVP) were over-expressed in cancer patients who were diagnosed with multidrug resistance, that is the resistance against many chemotherapy treatments[8]. Although this does not prove that increased number of vaults led to drug resistance, it does hint at some sort of involvement. This has potential in discovering the mechanisms behind drug-resistance in tumor cells and improving anticancer drugs[9]

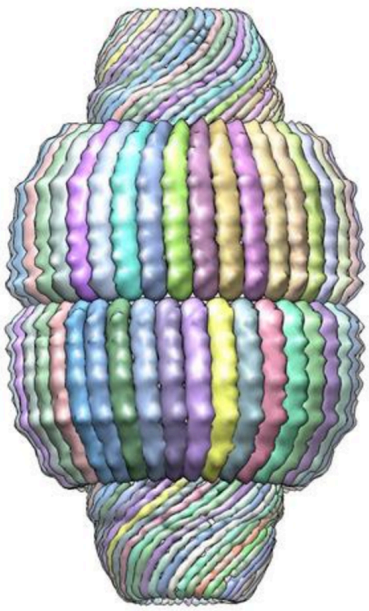


Fig.1 Structure of the Vault complex from rat liver.

1.2. vtRNA, secondary structure and functions

VtRNA, close to the end caps of Vaults, has a species-specific length, ranging between 86 and 141 bases[4]. vtRNA has been found in human, rat, mouse and bullfrog. In rats there is a single gene that encodes the rat vtRNA, whereas in humans there are four separate genes (hvg1–4) that encode highly related vtRNAs[4]. Hvg1 encodes a 98 nt RNA while hvg2 and hvg3 encode similar 88 nt RNAs. All these three are found on Chromosome 5 and show little sequence conservation between species except for their A and B boxes, which are internal polymerase III elements. However hvg4, which is found on the X chromosome, does not appear to be expressed[5]. Even though it varies in length, the vtRNA can be folded into similar secondary stem-loop and unusual symmetries structure. The current belief is that the vtRNA do not have a structural role in the vault protein, but rather play some kind of functional role[6].

Since the function of vtRNA remains unknown, so does the mechanism of action. It is hypothesized that at least in species with multiple vtRNAs such as humans, the ratio of what vtRNA species are associated with vaults may have a functional implication on drug resistance[7]. vtRNAs from different species are all predicted to form a stem-loop structure. The role of the stem-loop in vtRNA is still unknown; however, it is possible that the loop regions may be involved in mechanism via interaction with other RNAs or proteins. Regulation of vtRNA, is hypothesized to be controlled by the two closely spaced B boxes along with the 5’ flanking sequence[2].

According to recent studies, vtRNA is thought to have some implications in stress response, drug resistance and cancer. A study, using cryo-electron microscopy, has determined that vtRNAs are found close to the end caps of vaults. This positioning of the RNA indicates that they could interact with both the interior and exterior of the vault particle[32]. Overall, the current belief is that the vtRNAs do not have a structural role in the vault protein, but rather play some kind of functional role.[33]

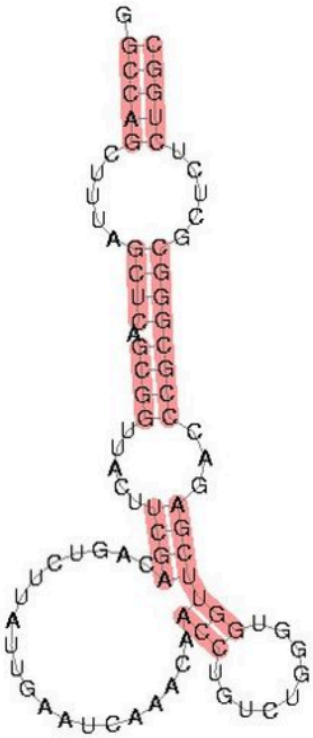


Fig.2 vtRNA with the stem-loop structure

1.3. Relevance of research on vtRNA in the Atlantic salmon louse

The notion that vaults might play a role in drug resistance was suggested by the molecular identification of the lung resistance-related (LRP) protein as the human MVP[10]. MVP/LRP was found to be overexpressed in many chemoresistant cancer cell lines and primary tumor samples of different histogenetic origin. Several, but not all, clinico-pathological studies showed that MVP expression at diagnosis was an independent adverse prognostic factor for response to chemotherapy[10]. The hollow barrel-shaped structure of the vault complex and its subcellular localization indicate a function in intracellular transport. It was therefore postulated that vaults contributed to drug resistance by transporting drugs away from their intracellular targets and/or the sequestration of drugs. However, even though there has been an expanding body of research on vtRNA, there has yet to be a solid conclusion on the exact function. To take a closer look into vtRNA at the genomic level could unravel more secrets.

The Atlantic salmon louse (*Lepeophtheirus salmonis*) is a and a serious threat to global and in particular Norwegian aquaculture. It is an ectoparasitic copepod *(Arthropoda;Crustacea)* primarily found on salmonid fishes where it feeds on the hosts skin, blood and mucus and can cause lesions that result in osmotic imbalance and stress. Salmon lice affect host physiology, suppress host immune responses and are suspected as vectors for other pathogen. If not kept under control, it represents a potentially severe burden for farmed and wild salmons[27]. The costs for salmon louse treatment are estimated to exceed 5 billion per year in Norway alone (Frank Nilsen, personal communication). In this study, we use the Atlantic salmon louse genome as a reference and aim to design a pipeline to detect vtRNA in salmon lice.

Consequently research has been conducted towards better understanding of the molecular fundament facilitating the success of the salmon louse. By 2012 the salmon louse 600 Mbp genome has been sequenced to significant coverage (<300X) using Sanger, 454 and Illumina sequencing (both shotgun and PE libraries)[28]. Several assembly strategies have been pursued and a pipeline for comparing assemblies has been established. It is sequenced by Illumina to generate the whole genes of Leopeophtheirus salmonis in order to the following data analyse and the implentation of the pipeline.

1.4. Problem description

The arrival of high-throughput sequencing technology has provided researchers with an opportunity to systematically identify most, if not all, of the vtRNA. Thus, determining expression of known and novel vtRNA from small RNA sequencing data is an important issue in the era of next generation sequencing[48].

While the function of vtRNAs is still unknown, due to their unique semi-conserved variable structure, these molecules have become useful in developing new research methods. One example of this is seen in the fact that vtRNAs are now used to benchmark the performance of the recently created research query tool, fragrep2[23].

Query tools are used to find regions of similar biological sequences amongst species. However, one problem that these tools (e.g. most famously, “Blast”) have is that they struggle to identify sequences that contain insertions and deletions. These highly variable structural changes cause problem in detecting homology in weakly conserved sequences, such as vtRNA and other non-coding RNA (ncRNA).

Fragrep2 seeks to solve this problem by using a pattern-based algorithm that can match or approximately match exact sequences of motifs within the desired molecule[23]. In order to help build fragrep2, the scientists needed a test molecule, and found vtRNAs to be perfect since vtRNAs generally have two very well-conserved sequences, surrounded by regions of high variability.

While quite successful in detecting novel ncRNA, tools such as fragrep2 do not take secondary structures of RNA other specific features of ncRNA such as U tails or other signal sequences into account.

1.5 Goals and Research Questions

In the following, I describe the construction of a novel pipeline, called vaulteR which can de-novo detect from RNA-seq data, using the known characteristics of vtRNA described before. I will first give an overview of tools in the pipeline, and then introduce the three main methods of detecting vtRNA. Finally, I will summarise the results from running the pipeline on simulated data and real data from the Atlantic salmon louse. I will attempt to answer the following research questions.

Question 1: Is Blast, as the traditional way of detecting RNA, suitable for detecting vtRNA?

Question 2: How can we make use of the secondary structure and other key feathers in detecting vtRNA?

Question 3: How can sequence motifs be used in detecting vtRNA ?

Goals: The pipeline aims to integrate the most reliable tools and make them applicable for RNA de-novo detection, especially with semi-conserved and highly variable structures.

2. Construction of the vtRNA detecting pipeline

2.1. Introduction on tools used by vaultR

A pipeline consists of a chain of data-processing elements, arranged so that the output of each element is the input of the next. For the construction of the vtRNA detecting pipeline, effective bioinformatics software tools are necessary for analysing and processing the results. The pipeline integrates the tools together in order to get the final output, which is to detect if there is vtRNA in the dataset.

A. Simulation of Illumina single-end reads: ART

ART is a set of simulation tools to generate synthetic next-generation sequencing reads[12]. ART simulates sequencing reads by mimicking real sequencing process with empirical error models or quality profiles summarized from large recalibrated sequencing data. ART can also simulate reads using user own read error model or quality profiles. ART supports simulation of single-end, paired-end/mate-pair reads of three major commercial next-generation sequencing platforms: Illumina's Solexa, Roche's 454 and Applied Biosystems' SOLiD. Here the pipeline uses ART illumine to generate single-end reads. ART can also be used to test or benchmark a variety of method or tools for next-generation sequencing data analysis, including read alignment, de novo assembly, SNP and structure variation discovery. ART outputs reads in the FASTQ format, and alignments in the ALN format. ART can also generate alignments in the SAM alignment or UCSC BED file format.

B. Mapping of short reads to the reference genome: BWA

BWA (Burrows-Wheeler Alignment Tool) is a software package for mapping low-divergent sequences against a large reference genome[13], such as the human genome. It consists of three algorithms: BWA-backtrack, BWA-SW and BWA-MEM. The first algorithm is designed for Illumina sequence reads up to 100bp, while the rest two for longer sequences ranged from 70bp to 1Mbp. BWA-MEM and BWA-SW share similar features such as long-read support and split alignment, but

BWA-MEM, which is the latest, is generally recommended for high-quality queries as it is faster and more accurate.
BWA-MEM also has better performance than BWA-backtrack for 70-100bp Illumina reads. Since the length of vtRNA is between 80bp and 150bp, the pipeline uses BWA-MEM for mapping. The BWA-MEM algorithm performs local alignment and output SAM file.

C. Samtools

SAM (Sequence Alignment/Map) format is a generic format for storing large nucleotide sequence alignments[35]. Samtools provide various utilities for manipulating alignments in the SAM format, BAM (Binary Alignment/Map) and CRAM formats, including sorting, merging, indexing and generating alignments in a per-position format[14].

Samtools is a suite of programs for interacting with high-throughput sequencing data .It consists of three separate repositories: Samtools, BCFtools and HTSlib[15]. In the pipeline, Samtools is used for reading, writing, editing, indexing and viewing SAM/BAM/CRAM format files, which are the result of read mapping by BWA.

D. IGV

The Integrative Genomics Viewer (IGV) is a lightweight visualization tool that enables intuitive real-time exploration of diverse, large-scale genomic datasets on standard desktop computers. It supports flexible integration of a wide range of genomic data types including aligned sequence reads, mutations, copy number, RNAi screens, gene expression, methylation, and genomic annotations[16]. IGV makes use of efficient, multi-resolution file formats to enable real-time exploration of arbitrarily large datasets over all resolution scales, while consuming minimal resources on the client computer[17].
With the help of IGV, the indexed and sorted BAM file on IGV can be visualised in order to see the aligned regions.

E. Quality control: FastQC

FastQC aims at providing a simple way to perform quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis and statistic summery[18].

The main functions of FastQC are:

- Import of data from BAM, SAM or FastQ files (any variant)
- Providing a quick overview to tell you in which areas there may be problems
- Summary graphs and tables to quickly assess your data
- Export of results to an HTML based permanent report
- Offline operation to allow automated generation of reports without running the interactive application

FastQC aims to analyse and assess the quality of raw read data and remove low quality reads for more accurate alignment.

F. Cmscan

Infernal ("INFERence of RNA ALignment") is for searching DNA sequence databases for RNA structure and sequence similarities. It is an implementation of a special case of profile stochastic context-free grammars called covariance models (CMs). A CM is like a sequence profile, but it scores a combination of sequence consensus and RNA secondary structure consensus, so in many cases, it is more capable of identifying RNA homologs that conserve their secondary structure more than their primary sequence which means Infernal cmscan is used to search the CM-format Rfam database[19].

The Rfam database is a collection of multiple sequence alignments and covariance models representing non-coding RNA families, each represented by multiple sequence alignments, consensus secondary structures and covariance models (CMs)[20].
The candidate sequences are written in FASTA format and are searched with Cmscan to find the similarities between the candidate sequences and RNA database.

G. VienneRNA Package/RNA-fold

The ViennaRNA Package consists of a C code library and several stand-alone programs for the prediction and comparison of RNA secondary structures[21].
RNA secondary structure prediction through energy minimization is the most used function in the package. There are three kinds of dynamic programming algorithms for structure prediction: the minimum free energy algorithm which yields a single optimal structure, the partition function algorithm which calculates base pair probabilities in the thermodynamic ensemble, and the suboptimal folding algorithm which generates all suboptimal structures within a given energy range of the optimal energy[21].
Here, the pipeline uses RNAfold, which is one of the core programs of the Vienna RNA package. It can be used to predict the minimum free energy (MFE) secondary structure of single sequences using the dynamic programming algorithm originally proposed by Zuker and Stiegler.
The input is a single RNA or DNA sequence in plain text or FASTA format, and the output contains the predicted MFE secondary structure in the usual dot-bracket notation, together with a detailed thermodynamic description according to the loop-based energy model and 2D graph[21].

H. MEME

MEME (Multiple EM for Motif Elicitation) is a tool for discovering novel, ungapped motifs (recurring, fixed-length patterns) motifs in a group of related DNA or protein sequences. MEME takes as input a group of DNA or protein sequences and outputs as many motifs as requested up to a user-specified statistical confidence threshold. MEME uses statistical modeling techniques to automatically choose the best width, number of occurrences, and description for each motif[22].

A motif is a sequence pattern that occurs repeatedly in a group of related sequences[36]. MEME represents motifs as position-dependent letter-probability matrices which describe the probability of each possible letter at each position in the pattern. Individual MEME motifs do not contain gaps. Patterns with variable-length gaps are split by MEME into two or more separate motifs.

2.2. Simulation of short-read and Real data from atlantic salmon louse

2.2.1. Generating simulated data for testing

As a first step, the sequences of the known vtRNAs are retrieved from Ensembl Biomart[37], which is an easy-to-use web-based tool that allows extraction of data. The present known vtRNA in human genome 38 are vtRNA 1-1, vtRNA 1-2, vtRNA 1-3 and vtRNA 2-1 in Chromosome 5, vtRNA 2-2 in Chromosome 2 and vtRNA 3-1in Chromosome X. The vtRNAs retrieved from Ensembl Biomart are given in Table 1.

Approved Symbol	Approved Name	Previous Symbols	Synonyms	Chromosome
VTRNA1-1	vault RNA 1-1	VAULTRC1	vtRNA1-1, hvg-1, HVG1, vRNA, VR1	5q31.3
VTRNA1-2	vault RNA 1-2	VAULTRC2	vtRNA1-2, hvg-2, HVG2, VR2	5q31.3
VTRNA1-3	vault RNA 1-3	VAULTRC3	vtRNA1-3, hvg-3, HVG3, VR3	5q31.3
VTRNA2-1	vault RNA 2-1	MIR886, MIRN886, VTRNA2	vtRNA2, hvg-5, CBL-3, hsa-mir-886, nc886	5q31.1
VTRNA2-2P	vault RNA 2-2, pseudogene			2p14
VTRNA3-1P	vault RNA 3-1, pseudogene	VAULTRC4, VTRNA3P	vtRNA3P, hvg-4, HVG4	Xp11.22

Table 1. Known vtRNAs in human genome 38 from Ensembl Biomart

The simulated data has to contain all these six known vtRNA in order to make sure it can generate enough vtRNA reads. Then, these vtRNA are mixed with other non-coding RNA (ncRNA) families, since vtRNA is a non-coding RNA and is likely to occur in a mixture with other nvRNA's in real data as well.

As the next step, the sequences are stored in FASTA file format, filtered by length, and only those sequences with a length ranging from 80 to 150 bases are kept, to imitate the size-selection that will be performed in real data since all known vtRNA fall into this range.

2.2.2. Real data from the Atlantic salmon louse

Next-generation sequencing (NGS)[38], also known as high-throughput sequencing, is the catch-all term used to describe a number of different modern sequencing technologies including: Illumina (Solexa) sequencing, Roche 454 sequencing, Ion torrent: Proton / PGM sequencing and SOLiD sequencing.

These technologies allow us to sequence DNA and RNA much more quickly and cheaply than the previously used Sanger sequencing, and as such have revolutionised the study of genomics and molecular biology. And currently, there are ten high-throughput sequencing platforms and the Illumina platforms is the leading platform for high-throughput sequencing[24].

RNA-sequencing, uses next-generation sequencing (NGS) to reveal the presence and quantity of RNA in a biological sample at a given moment in time, RNA-Seq can look at different populations of RNA to include total RNA, small RNA, such as miRNA, tRNA, vtRNA and ribosomal profiling[25].

For small-RNA and non-coding RNA sequencing, library preparation is modified. The cellular RNA is selected based on the desired size range. For small RNA targets, such as vtRNA, the RNA is isolated through size selection. This can be performed with a size exclusion gel, through size

selection magnetic beads, or with a commercially developed kit. Once isolated, linkers are added to the 3' and 5' end then purified. The final step is cDNA generation through reverse transcription[26].

Single-read sequencing involves sequencing DNA from only one end, and is the simplest way to utilize Illumina sequencing. By leveraging proprietary reversible terminator chemistry and a novel polymerase, this solution delivers large volumes of high-quality data, rapidly and economically.

While Paired-end sequencing allows users to sequence both ends of a fragment and generate high-quality, alignable sequence data. Paired-end sequencing facilitates detection of genomic rearrangements and repetitive sequence elements, as well as gene fusions and novel transcripts. Since paired-end reads are more likely to align to a reference, the quality of the entire data set improves. All Illumina next-generation sequencing (NGS) systems are capable of paired-end sequencing. Paired-end reads, which means that for each DNA fragment, we have sequence data from both ends. The sequences are therefore stored in two separate files (one for the data from each end).

In the study, total RNA from a mixture of all lif-cycle stages of the Atlantic salmon louse was generated by Christiane Eichner at the Sea Lice Research Centre, Bergen, Norway. Library preparation and sequencing was done by the Norwegian sequencing centre, Oslo, using the Illumina Sequencing Protocol and the NextSeq 500 platform. Sequencing was done for a size selected fraction (80-150 bases) of the total RNA, and resulted in approximate 50 million single-end reads of 76bp in length.

2.3. The vtRNA detecting pipeline

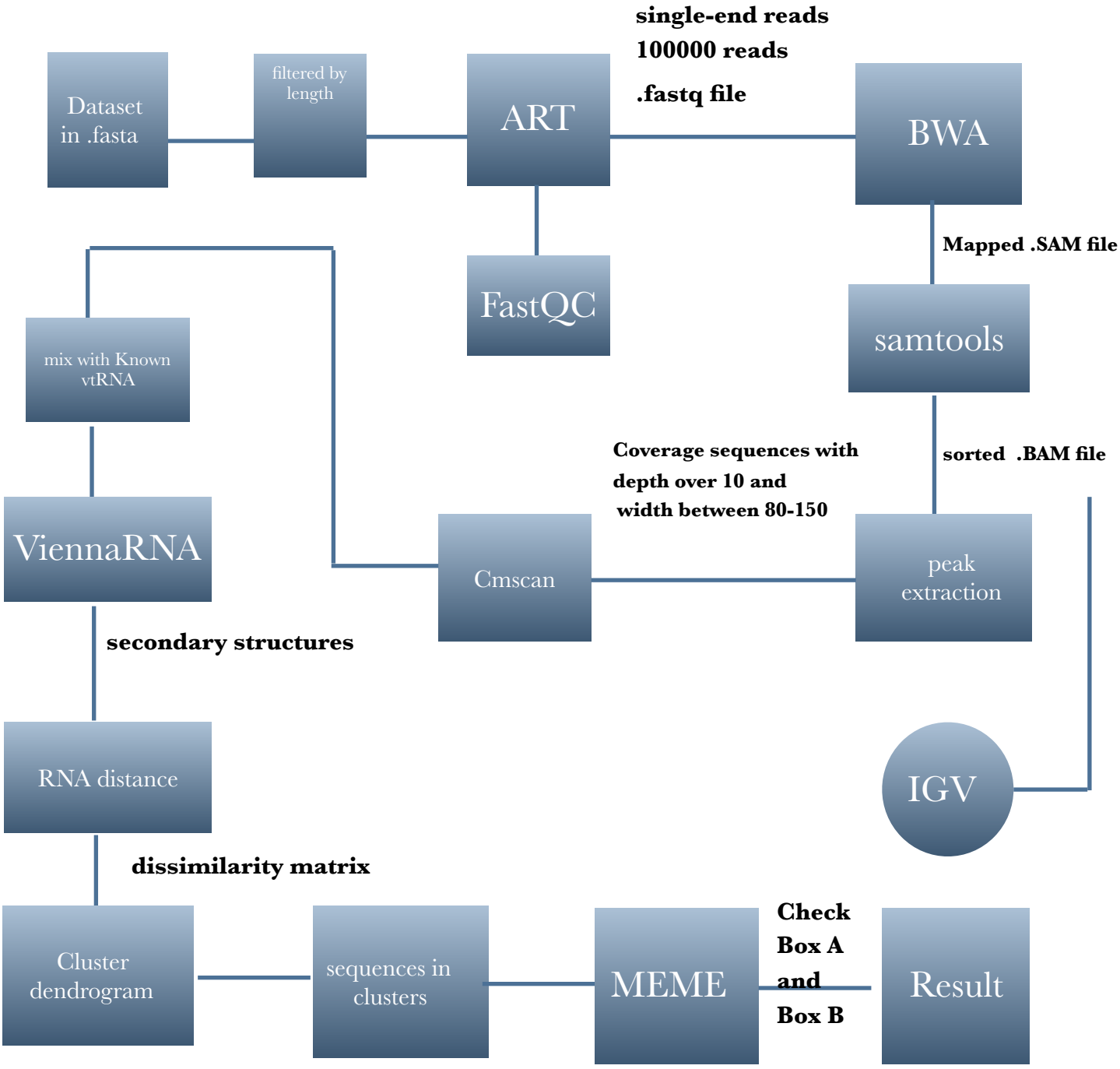


Fig.4 Overview of vaultR - the vtRNA detecting pipeline workflow

The workflow of the pipeline for detecting vtRNA (vaultR) is depicted in Fig.5. In the beginning, it generates a dataset in FASTA format and filters by length between 80 to 150 since known vtRNA has such length range. Then vaultR uses ART to generate sufficiently large single-end reads from the filtered dataset and checks the quality of the reads with FastQC. Then, with the help of BWA, vaultR maps the reads back to the reference genome and generates a SAM file. samtools is used to transfer the format from SAM to BAM and sort the BAM file. The high coverage sequences on reference genome by IGV. Then it comes to the peak extraction which is to extract the high coverage sequences with depth over 10 and width between 80 and 150. The sequences of high coverage are annotated with Cmscan using the Rfam database and a report about whether there is vtRNA in the dataset is generated. If not, go to the next step: vaultR mixes the high-coverage sequences with known vtRNAs and generates the secondary structures by ViennaRNA. From that the dissimilarity matrix between the sequences by their secondary structures is computed and a cluster dendrogram is generated. By looking at the sequences in each cluster, especially those cluster together with the known vtRNAs, novel candidates are found. Then by MEME, motifs are generated for each cluster, the vaultR checks the existence of Box A and Box B motifs and attempts to finally rank the candidate vtRNA sequences.

3. Methodologies to detect vtRNA

3.1. Peak extraction and high coverage sequences searching

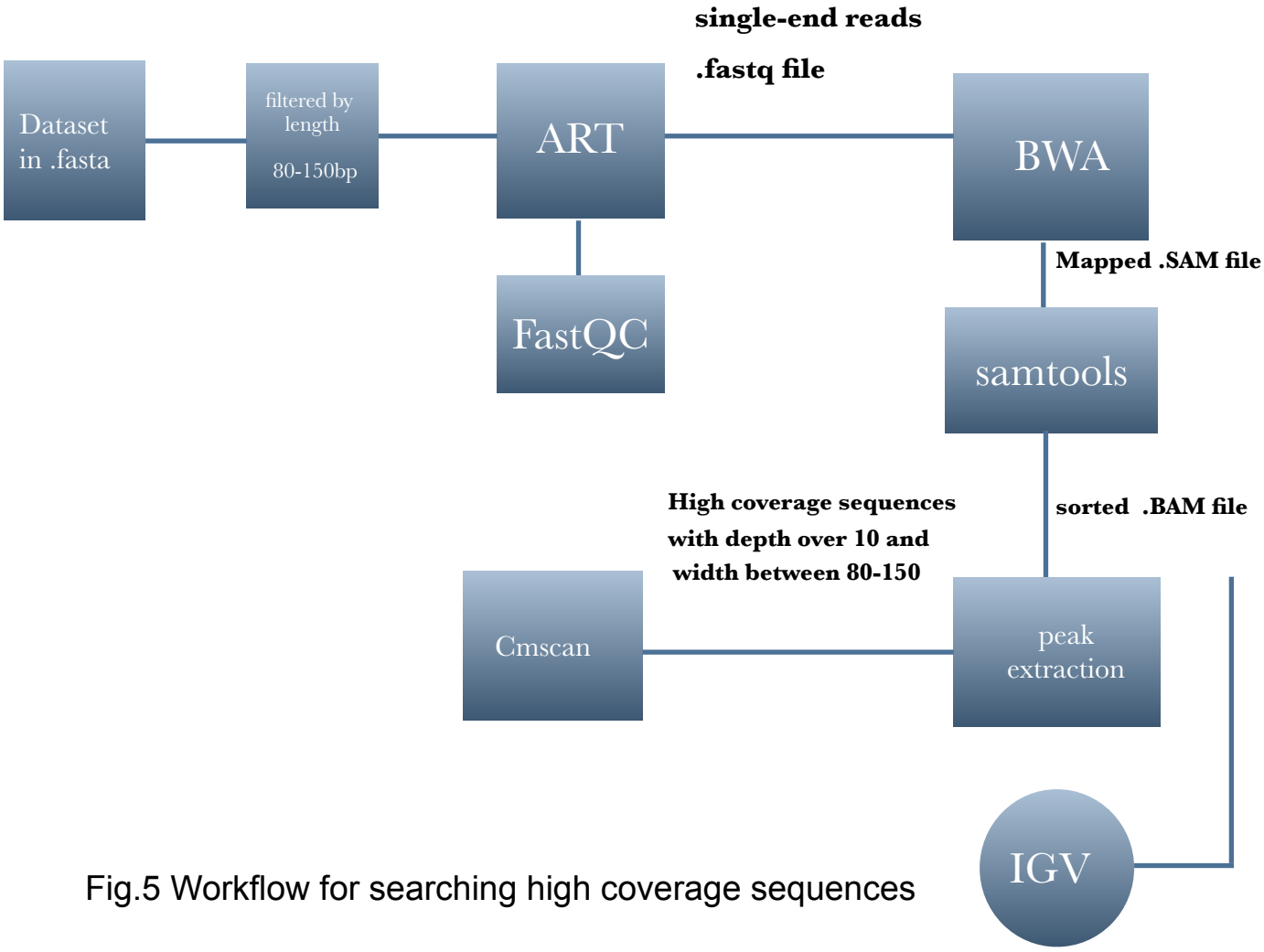


Fig.5 Workflow for searching high coverage sequences

The input data is prepared in FASTA format, and filtered to contain sequences of length 80-150bp. Generally, vtRNA has a length that ranges between 86 and 141 bases, depending on the species. Then, the pipeline generates single-end reads by ART. ART simulates both single-end and paired-end sequencing reads of the three main commercial next-generation sequencing platforms: 454, Illumina and SOLiD. The built-in read length and read error profiles were derived from large sets of actual real sequencing data. ART supports all three types of common sequencing errors: base substitutions, insertions and deletions.

After preparation of data, the pipeline generates FASTQ, SAM and ALN files by ART and maps the reads in the FASTQ file back to human genome 38, or other reference genome.

Then, as the third step, quality control of the generated reads is performed. Quality control and filtering of sequencing reads is one of the most important steps in the pre-processing of sequencing reads. However, it is not always trivial to figure out which reads needs adjustment and which can be left untouched. And here to assess the quality of the source data. The most convenient tool for this task is FastQC.

Sequencing reads can be assembled de-novo into a full genome or mapped to an already-assembled reference genome of a related organism. However, no sequencing technology is perfect and raw reads inevitably contain mistakes: sequencing errors. The probability of an error for each nucleotide of each read is always written in a FASTQ file. Therefore, the very first step of fragment analysis is quality control and filtering on the FASTQ file. This step aims to remove low quality reads.

Mapping by BWA:

The BWA tool uses the Ferragina and Manzini matching algorithm to find exact matches, similar to Bowtie[29]. For all the algorithms, BWA first needs to construct the FM-index for the reference genome (the index command). And alignment algorithms are invoked with different sub-commands: aln/samse/sampe for BWA-backtrack, bwasw for BWA-SW and mem for the BWA-MEM algorithm.

For longer sequences ranged from 70bp to 1Mbp, BWA-MEM performs better. BWA-MEM is a new alignment algorithm for aligning sequence reads or long query sequences against a large reference genome such as human. It automatically chooses between local and end-to-end alignments. The algorithm is robust to sequencing errors and applicable to a wide range of sequence lengths from 70bp to a few megabases. For mapping 100bp sequences, BWA-MEM shows better performance than several state-of-art read aligners to date[29].

There are two steps, Indexing and mapping:
The first step of using BWA is to make an index of the reference genome in FASTA format. Then using bwa-mem for mapping, it generates a SAM file, which is technically human-readable.

When configuring to the BWA application, one of the most important parameters is how many mismatches you will allow between a read and a potential mapping location for that location to be considered a match. It sets as the default (4% of the read length)[29]. And for the single-end reads, use “bwa samse” as command.

Processing the output with Samtools:
Like BWA, Samtools also go through several steps before data are in usable form. First, it generates its own index of the reference genome with Samtools, and the reference genome should always be the same. Next, a SAM file is converted into a BAM file. (A BAM file is just a binary version of a SAM file.) Then sort and index the BAM file.

Then, aligned reads can be viewed by using the Integrative Genomics Viewer (IGV), BAM form is preferred than SAM form, which is the recommended format for IGV. IGV requires that both SAM and BAM files be sorted by position and indexed, and that the index files follow a specific naming convention. Specifically, a BAM index file should be named by appending .BAI to the bam file name. A SAM index filename is created by appending .SAI.

Peak Extraction：

Peak calling is a computational method used to identify areas in a genome that have been enriched with aligned reads as a consequence of performing sequencing[31]. A peak is called where either the number of reads exceeds a pre-determined threshold value or where there is a minimum enrichment compared to background signal, often in a sliding window across the genome. The parameters for identifying peaks can be adjusted, sometimes leading to very different numbers of peaks being called.

For extracting the peak regions, first read genomic alignments from the BAM file into a GappedReads object in R[39]. A GappedReads object contains all the information contained in a GAlignments object plus the sequences of the queries. Then vaultR counts the number of reads at each position on the reference genome, which is represented in a set of ranges. After that, it extracts and keeps those regions with a coverage depth over 10, and a width between 80 and 150 and notes the positions. Then it finds the positions back in the reference genome and get the coverage sequences, writes these coverage sequences into a FASTA file and runs cmscan on all regions with high coverage sequences, finally it searches Rfam using the FASTA file and check if vtRNA exists.

3.2. Secondary structures, prediction and analysis

Biomolecules exhibit a close interplay between structure and function. While prediction of tertiary structure is usually infeasible, the area of RNA secondary structures is an example where computational methods have been highly successful. The prediction of RNA structure has received increasing attention over the last decade as the number of known functional RNA sequences, called non-coding RNA (ncRNA), has increased. And the conserved structures are of particular interest, since conservation of structure in spite of sequence variation implies that the structure must be functionally important. VtRNA, as the highly conserved noncoding RNA, can be known better through the secondary structures[49].

To understand the mechanism of action of a RNA, the structure must be known. RNA secondary structure prediction, using thermodynamics, can be used to develop hypotheses about the structure of an RNA sequence. Secondary structure prediction is a set of techniques in bioinformatics that aim to predict the secondary structures of proteins and nucleic acidsequences based only on knowledge of their primary structure. For nucleic acids it means predicting the formation of nucleic acid structures like helixes and stem-loopstructures through base pairing and base stacking interactions.

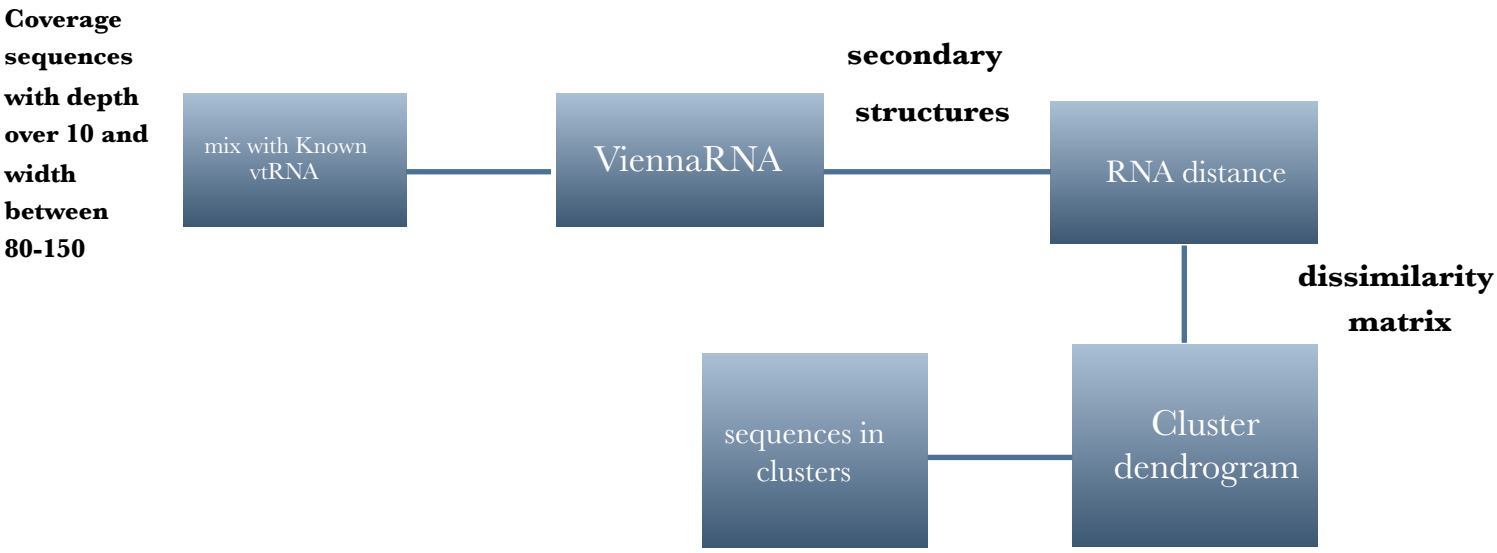


Fig.6 Workflow for vtRNA detection by secondary structure prediction and analyse

1. First is to mix the high coverage candidate sequences with the known vtRNAs. Predict the secondary structure of all the candidate sequences and the known vtRNA with ViennaRNA[21] secondary structure prediction. It generates two kinds of secondary structures: MFE secondary structure by minimum free energy and centroid secondary structure by thermodynamic ensemble prediction. Here use Minimum Free Energy Structure for further analyse.

Minimum Free Energy Structure(MFE)[40]: The minimum free energy structure of a sequence is the secondary structure that is calculated to have the lowest value of free energy. It is synonymous with natural-mode structure, but it is not necessarily the structure that forms in nature. The MFE structure of an RNA sequence is the secondary structure that contributes a minimum of free energy. This structure is predicted using a loop-based energy model and the dynamic

programming algorithm introduced by Zuker et al. As an RNA secondary structure can be uniquely decomposed into loops and external bases the loop-based energy model treats the free energy F(s) of an RNA secondary structures as the sum of the contributing free energies F_L of the loops Lcontained in s. According to the chosen energy parameter set and a given temperature (defaults to 37 °C) the secondary structure s that minimizes F(s) is computed[40].

The lower the free energy, the more likely the structure will form, which means that the lower the thermodynamic energy of the structure, the more stable it generally is. However, this is calculated using Zuker's algorithm[40] which is accurate for secondary structure predictions. If working with specific family or group of RNAs then attempt to correlate the secondary structural motifs such as stem loops - bulges or junctions in the RNA structure with the free energy value.

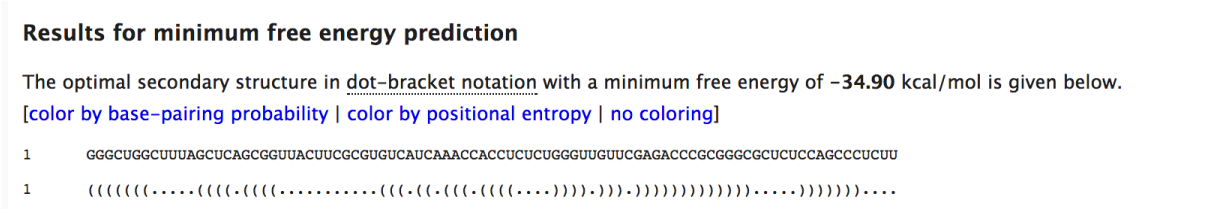


Fig.7 shows the MFE prediction of secondary structure for vtRNA 1-3 by ViennaRNA, which is expressed by dot-bracket notation.

2. According to the MFE secondary structures of the candidate sequences. Calculate dissimilarities between RNA secondary structures with RNAdistance[41][42][43][44][45].

Read RNA secondary structures and calculates one or more measures for their dissimilarity, based on tree or string editing (alignment). In addition it calculates a "base pair distance" given by the number of base pairs present in one structure, but not the other. For structures of different length base pair distance is not recommended. RNAdistance accepts structures in bracket format, where matching brackets symbolize base pairs and unpaired bases are represented by a dot “.”, which is the dot bracket form of the secondary structure.

Then take the structures of the known vtRNA as references and compare the distances between all the candidate sequences and then make a dissimilarity matrix. The lower result is, the more similar between the two structures are.

3. The next step is hierarchical cluster analysis on a set of dissimilarities in order to generate cluster dendrogram. A dendrogram[46] is a tree diagram frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering. Dendrograms are often used in computational biology to illustrate the clustering of genes or samples, sometimes on top of heatmaps. The dendrogram is a visual representation of the compound correlation data. The individual compounds are arranged along the bottom of the dendrogram and referred to as leaf nodes. Compound clusters are formed by joining individual compounds or existing compound clusters with the join point referred to as a node. At each dendrogram node there is a right and left sub-branch of clustered compounds.

4. Analyse the candidate sequences which cluster the same with known vtRNAs. Check the secondary structures and structural features. And the pipeline goes to the next step, de novo detection of motifs.

3.3. De novo detection of motifs in vtRNA candidates

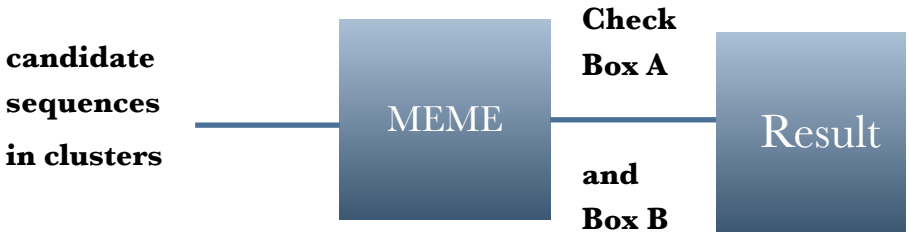


Fig.8 Workflow for detecting motifs in candidate sequences by MEME

Motifs[36]: Sequence motifs are short, recurring patterns in DNA that are presumed to have a biological function. Often they indicate sequence-specific binding sites for proteins such as nucleases and transcription factors (TF). Others are involved in important processes at the RNA level, including ribosome binding, mRNA processing (splicing, editing, polyadenylation) and transcription termination. Nowadays, computational methods are generating a flood of putative regulatory sequence motifs by searching for overrepresented (and/or conserved) DNA patterns upstream of functionally related genes (for example, genes with similar expression patterns or similar functional annotation)[36].

VtRNA genes have been cloned from several vertebrates including rat, mouse, and humans. Their copy numbers vary, as does the length of the encoded RNA. By comparing the upstream regions of the vertebrate vRNA genes, a 25 bp conserved sequence and a TATA box can be identified. Furthermore, the unique arrangement of the internal promoter boxes is conserved in the expressed human vRNA genes even though a new RNA polymerase III termination sequence has evolved between the two B boxes[47].

The vRNA contains two B-box elements and one A-box element (type-2 promoter elements). A and B boxes are binding sites for TFIIIC which positions TFIIIB immediately upstream of the gene. Subsequently, TFIIIB directs binding of RNA polymerase III, which initiates transcription. The vRNA contains a TATA box sequence at position -25 and an assumed proximal sequence element at position -70 (with respect to transcription initiation site). Additionally, viable 5' flanking sequence is required for transcription. Also, at high transcription factor concentrations, the presence of the two B boxes inhibits vRNA transcription. It is postulated that the two closely spaced B boxes along with the 5' flanking sequence provide a mechanism for the regulation of vRNA gene activity.

According to the conserved structure of the vtRNA with two B-box elements and one A-box element, we aim to detect the motifs with MEME Suite and check if there is sequences with such A-B-Box structures.

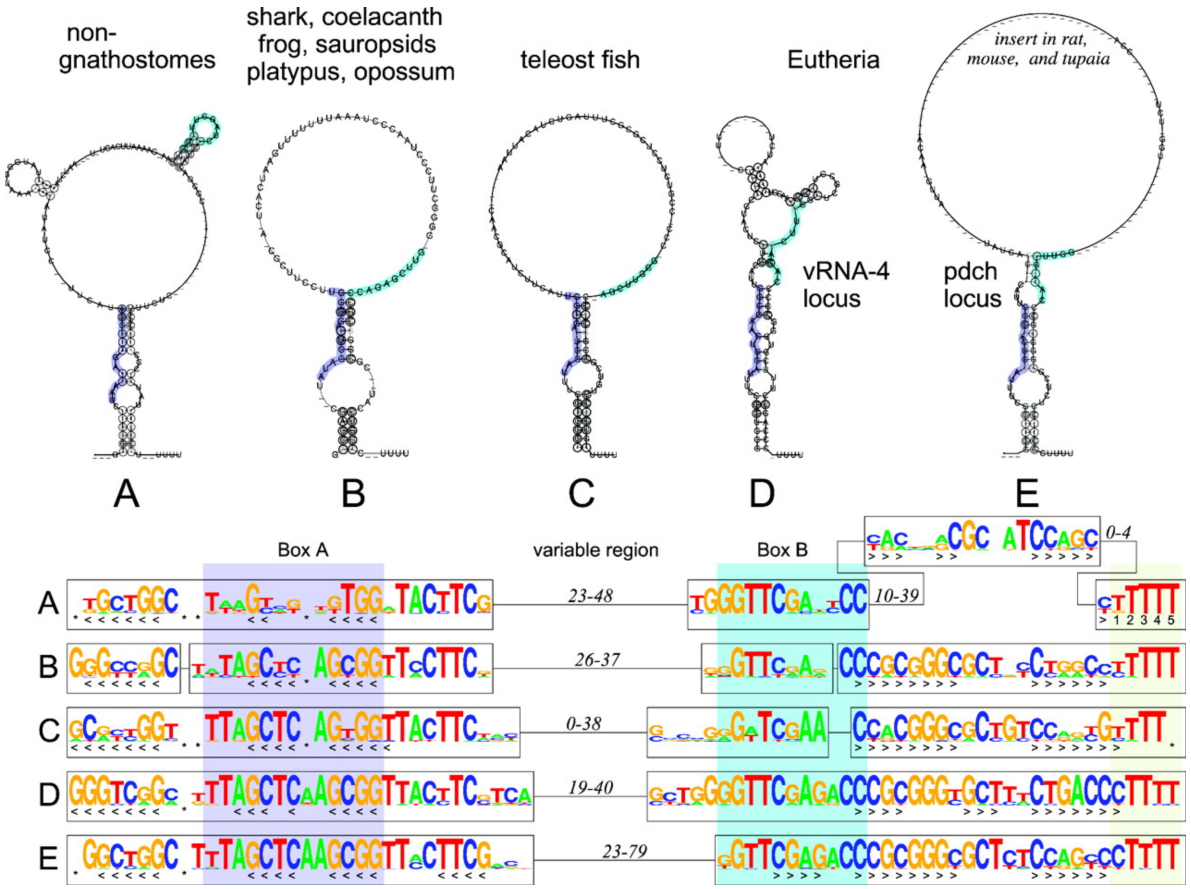


Fig.9

Fig.9 Comparison of consensus secondary structures and sequence logos derived from separate alignments of the deuterostome vtRNAs. vtRNAs form a conserved panhandle-like secondary structure with a well-conserved extended stem-loop structure connecting 5′ end and 3′end of the molecule. This structure also involves the box A sequence. The box B, on the other hand, does not take part in conserved structural features, albeit in vertebrates, the stem-loop structure overlaps the last 1 or two nt of the box B. In the basal lineages, box B and the 3′ side of the stem-loop structure are separated by at least 10 nt of intervening sequence. The base pairing of box A likely contributes to the sequence conservation in the 3′ region of the vtRNAs.

4. Results and summary

4.1. Results from simulated data

For simulated data:

Statistic results from FASTQC for quantity control.

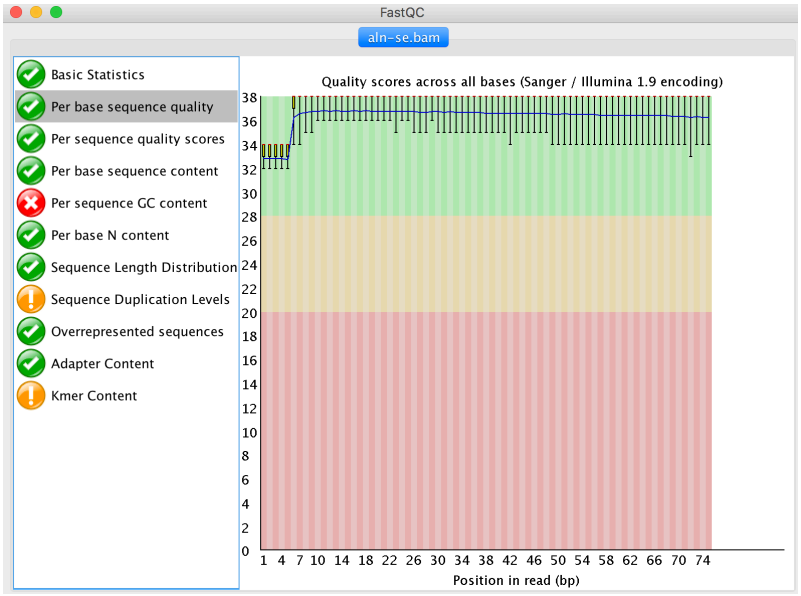


Fig.10 Basic statistics of the reads sequences from the simulated data

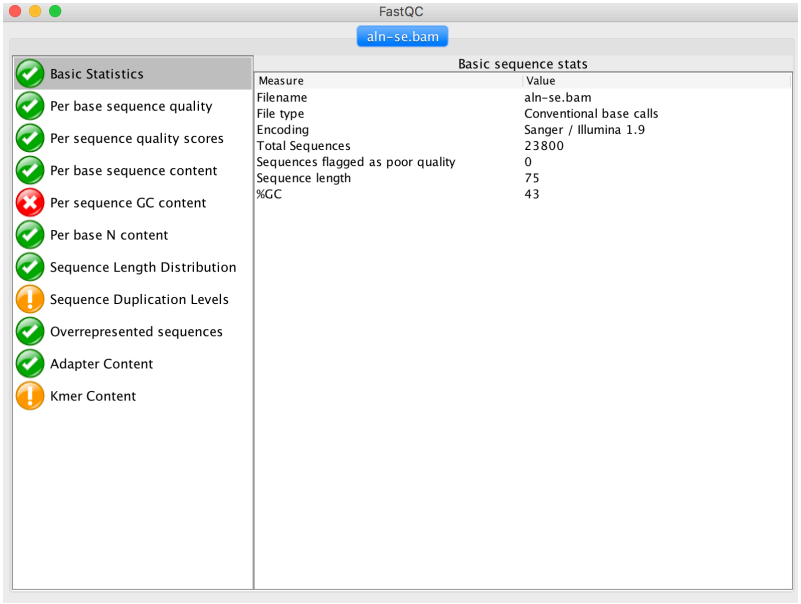


Fig.11 Per base sequence quality of reads sequences from the simulated data

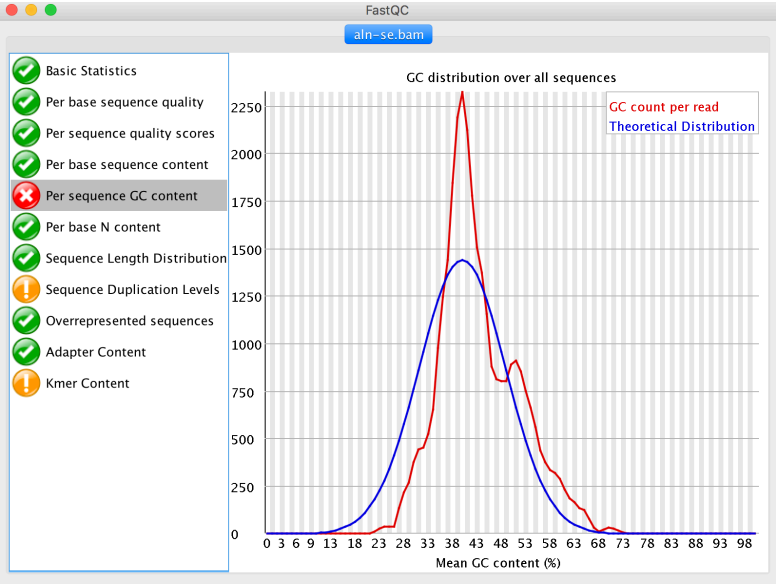


Fig.12 Per sequence GC content of reads sequences from the simulated data

As quality control, for the “per sequence quality scores”, the quality of the simulated reads is quite good and the curve is very smooth, so there is no need to cut any reads in the dataset. The only one problem is “Per sequence GC content”, In a normal random library it is expected to see a roughly normal distribution of GC content where the central peak corresponds to the overall GC content of the underlying genome. Since it is not known that the GC content of the genome the modal GC content is calculated from the observed data and used to build a reference distribution. An unusually shaped distribution could indicate a contaminated library or some other kinds of biased subset. A normal distribution which is shifted indicates some systematic bias which is independent of base position. If there is a systematic bias which creates a shifted normal distribution then this won't be flagged as an error by the module since it does not know what your genome's GC content should be. If the secondary peak is very sharp it's probably a specific contaminant - often something which is found by the overrepresented sequences module.while the “Per sequence GC content” does not affect the results so much, so the process can go further for reads mapping after quality control.

The quality of simulation reads are fine. As the simulated data is mixed with known vtRNAs, put back the mapped SAM file to the Hg 38 as reference genome, check if there are high coverage sequences at where the positions vtRNAs locate.

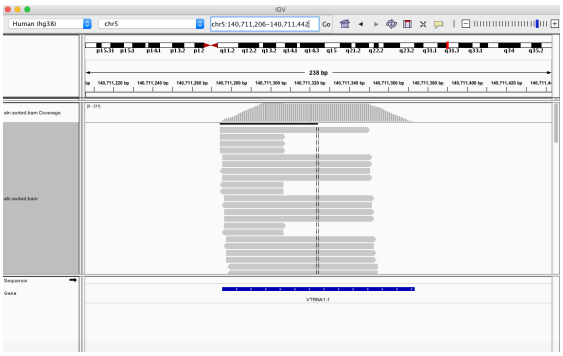


Fig.13 high coverage sequence at the position of vtRNA 1-1 at Hg 38

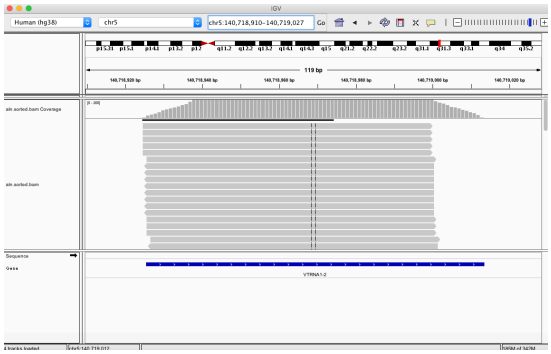


Fig.14 high coverage sequences at the position of vtRNA 1-2 at Hg 38

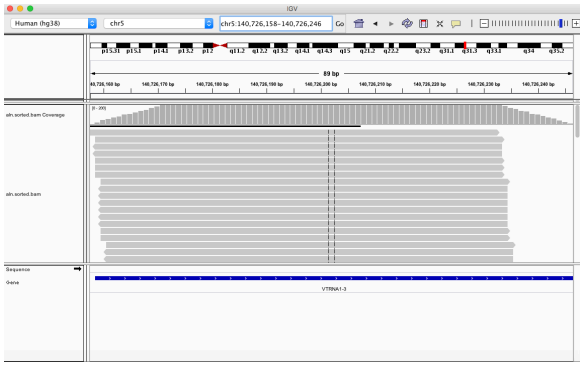


Fig.15 high coverage sequences at the position of vtRNA 1-3 at Hg 38

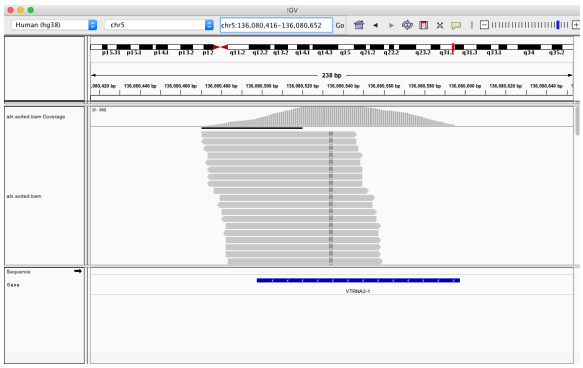


Fig.16 high coverage sequences at the position of vtRNA 2-1 at Hg 38

High coverage sequences at the positions where Vaults locate are found in the reference genome. So the method is feasible for simulated data, which is obviously seen at IGV. Then extract the high coverage sequences and write them in FASTA file, search the sequences with Cmscan, here is the result:

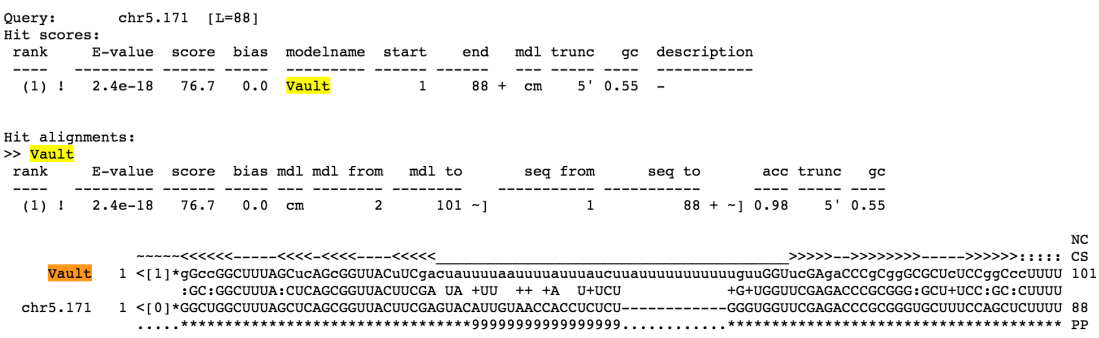


Fig.17 searching results from simulated data by Cmscan

As the graph shows, vtRNAs are found after searching in the Rfam. In the thesis, we generate a whole pipeline with three different methods to detect vtRNA. For the first method, we take use of the high coverage sequences and search for them in the Rfam with cmscan and it works for the simulated data.The advantage is that result is quite intuitive for the users, while it is ineffective for identifying vtRNA sequences which are very highly conserved at the nucleotide level. And even though it does find vtRNAs in the simulated data, let us see the results from salmon louse.

4.2. Results from atlantic salmon louse data

For the real data from salmon louse, first check the reads quality by FASTQC

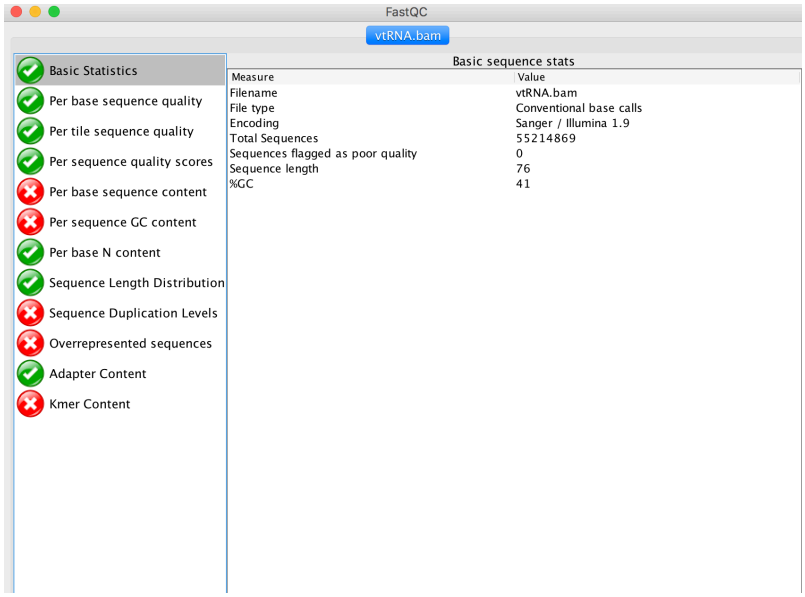


Fig.18 Basic statistics of the reads sequences from the salmon louse

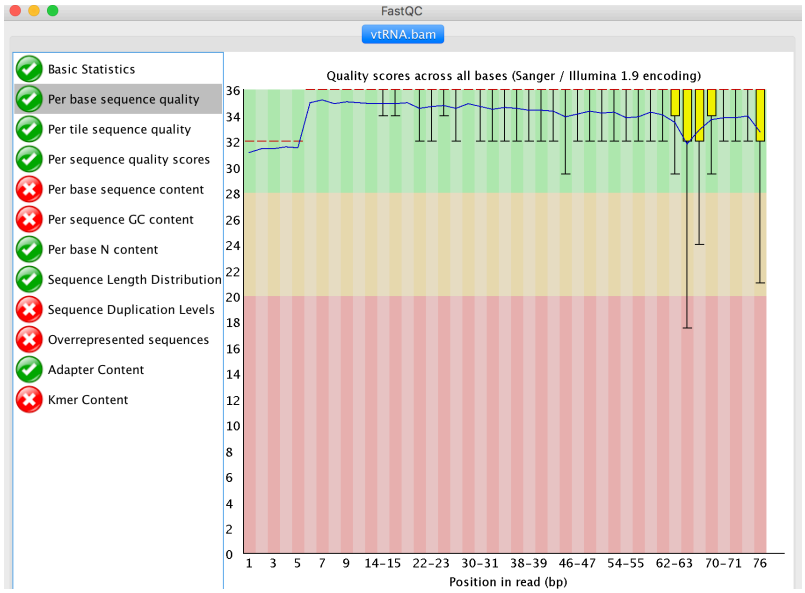


Fig.19 Per base sequence quality of reads sequences from the salmon louse

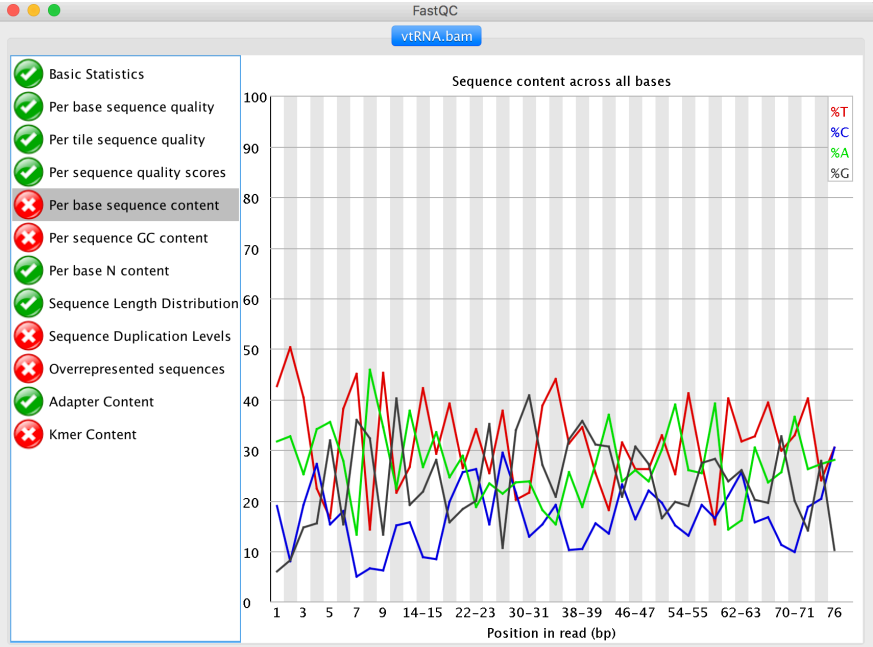


Fig.20 Per base sequence content of reads sequences from the salmon louse

For the real data from salmon louse, it generates 55 million reads with length 76 bases, the per base sequence quality on Fig.18 are relative smooth. The next problem is unusual per-base sequence content on Fig. 20. We expect to see flat lines that represent the percentages of A, C, T, and G in the genome. However, there are often biases (particularly at the start of reads). And it is clearly seen that the biased sequence along the run. While this does not affect the following results. Just keep going to the next step for peek extraction and searching for high coverage sequences.

1. Read genomic alignments from the BAM file into a GappedReads object.

```
GAlignments object with 49412244 alignments and 0 metadata columns:
                   seqnames strand       cigar    qwidth     start       end     width     njunc
                      <Rle>  <Rle> <character> <integer> <integer> <integer> <integer> <integer>
          [1]  LSalAtl2s126      -         76M        76    350633    350708        76         0
          [2] LSalAtl2s1318      -         76M        76     98466     98541        76         0
          [3] LSalAtl2s1699      +         76M        76      5197      5272        76         0
          [4]  LSalAtl2s378      -         76M        76    615051    615126        76         0
          [5]  LSalAtl2s126      -         76M        76    350633    350708        76         0
          ...           ...    ...         ...       ...       ...       ...       ...       ...
  [49412240] LSalAtl2s1699      +         76M        76      5197      5272        76         0
  [49412241] LSalAtl2s1699      +         76M        76      5196      5271        76         0
  [49412242] LSalAtl2s7121      -         76M        76        46       121        76         0
  [49412243] LSalAtl2s1699      +         76M        76      5196      5271        76         0
```

Fig.21 high coverage sequences with start position, end position and width

2. Count the number of coverages at each position, which is represented in a set of ranges.

```
RleList of length 36095
$LSalAtl2s1
integer-Rle of length 3702309 with 878 runs
  Lengths: 226830     76   1016     76   1604     76   3230     76    926     76    335     76    121 ...
18     58   2165     76  18784     76    973     76   1255     76   3135     76   4966
  Values :      0      1      0      1      0      2      0      1      0      1      0      1      0 ...
3      1      0      1      0      1      0      1      0      1      0      1      0
```

Fig.22 Sequences with number of coverages at each position on reference genome.

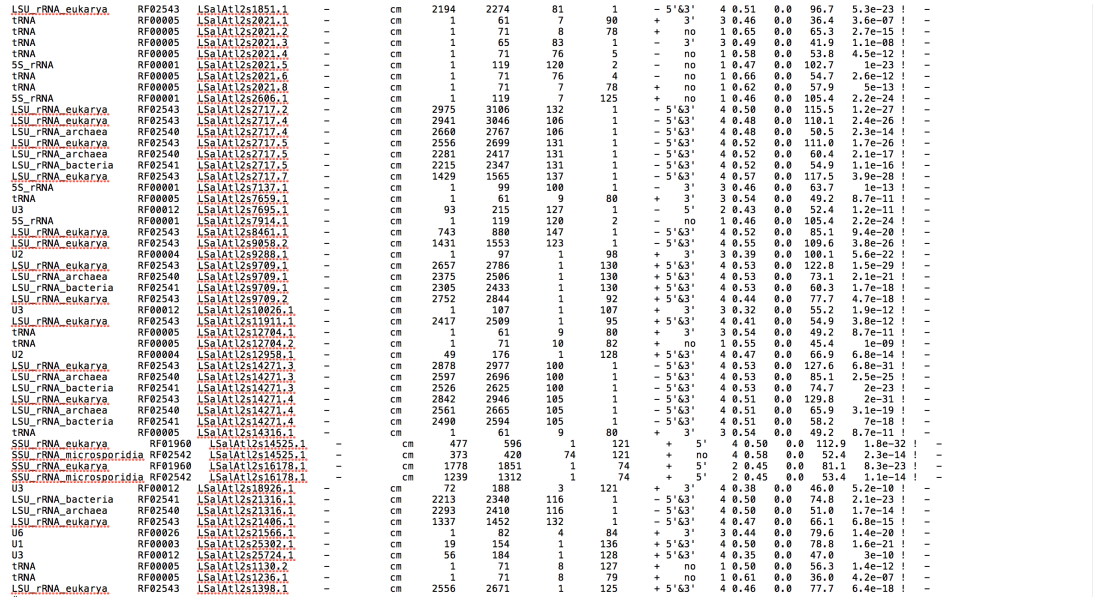
3. Extract and keep those coverages with depth over 10, and width between 80 and 150 and note the positions.

```
views:
      start     end width
[1] 3596953 3597041    89 [10 10 10 10 10 10 11 14 14 14 16 19 19 21 22 23 24 24 24 24 24 25 25 25 25 25
25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 24 24 24 25 26 26 26 27 27 27 ...]
```

Fig.23 Sequences after Peak extraction with number of coverages at each position, followed with start position, end position and width.

5. Find the positions back in the reference genome and get the coverage sequences and write these coverage sequences in FASTA file and run Cmscan to check if vtRNA exists.

```
#target name         accession query name           accession mdl mdl from   mdl to seq from   seq to strand trunc pass   gc  bias  score   E-value inc description of target
#-------------------- --------- -------------------- --------- --- -------- -------- -------- -------- ------ ----- ---- ---- ----- ------ --------- --- ---------------------
LSU_rRNA_eukarya      RF02543   LSalAtl2s229.2       -         cm     2265     2373        1      109      + 5'&3'    4 0.46   0.0  122.0   2.1e-29 !   -
LSU_rRNA_archaea      RF02540   LSalAtl2s229.2       -         cm     2021     2129        1      109      + 5'&3'    4 0.46   0.0   57.6   1.4e-16 !   -
LSU_rRNA_eukarya      RF02543   LSalAtl2s229.3       -         cm     2843     2975        1      133      + 5'&3'    4 0.53   0.0  175.6   3.8e-43 !   -
LSU_rRNA_archaea      RF02540   LSalAtl2s229.3       -         cm     2562     2694        1      133      + 5'&3'    4 0.53   0.0  105.0   1.7e-31 !   -
LSU_rRNA_bacteria     RF02541   LSalAtl2s229.3       -         cm     2491     2623        1      133      + 5'&3'    4 0.53   0.0   96.4   1.4e-30 !   -
SSU_rRNA_eukarya      RF01960   LSalAtl2s256.1       -         cm      955     1069      115        1      - 5'&3'    4 0.44   0.0  105.2   3.6e-30 !   -
SSU_rRNA_microsporidia RF02542  LSalAtl2s256.1       -         cm      597      711      115        1      - 5'&3'    4 0.44   0.0   68.0   3.2e-19 !   -
tRNA-Sec              RF01852   LSalAtl2s257.1       -         cm        1       89       84        1      -    3'    3 0.55   0.0   62.4   2.1e-13 !   -
LSU_rRNA_eukarya      RF02543   LSalAtl2s258.1       -         cm     1842     1970        1      129      + 5'&3'    4 0.49   0.0  134.2   1.8e-32 !   -
LSU_rRNA_archaea      RF02540   LSalAtl2s258.1       -         cm     1736     1867        1      129      + 5'&3'    4 0.49   0.0   54.2   1.9e-15 !   -
U6                    RF00026   LSalAtl2s322.1       -         cm        1       82       81        1      -    3'    3 0.44   0.0   79.6   1.4e-20 !   -
U6                    RF00026   LSalAtl2s322.2       -         cm        1       82       85        1      -    3'    3 0.45   0.0   86.8   1.2e-22 !   -
snosnR60_Z15          RF00309   LSalAtl2s383.2       -         cm        1       98        5       80      +    no    1 0.41   0.0   53.5   3.7e-09 !   -
tRNA                  RF00005   LSalAtl2s2.1         -         cm        6       71        1       79      +    5'    2 0.57   0.0   48.4   1.5e-10 !   -
U5                    RF00020   LSalAtl2s50.1        -         cm        1      116        9      129      +    no    1 0.35   0.0   79.5   2.7e-14 !   -
snosnR60_Z15          RF00309   LSalAtl2s52.3        -         cm        1       98        1       93      +    no    1 0.32   0.0   53.4   4.3e-09 !   -
SSU_rRNA_eukarya      RF01960   LSalAtl2s54.1        -         cm     1071     1154        1       85      + 5'&3'    4 0.46   0.0   52.9   2.3e-14 !   -
U5                    RF00020   LSalAtl2s61.2        -         cm        1      104      105        1      -    3'    3 0.38   0.0   77.2   6.8e-14 !   -
LSU_rRNA_eukarya      RF02543   LSalAtl2s67.1        -         cm     1881     1971        1       91      + 5'&3'    4 0.46   0.0   89.1   5.4e-21 !   -
U6                    RF00026   LSalAtl2s75.1        -         cm        1       81        3       82      +    3'    3 0.45   0.0   79.0   2.1e-20 !   -
tRNA                  RF00005   LSalAtl2s447.3       -         cm        1       71        8       91      +    no    1 0.57   0.0   63.7     9e-15 !   -
LSU_rRNA_eukarya      RF02543   LSalAtl2s465.1       -         cm     2344     2433        1       90      + 5'&3'    4 0.50   0.0   94.1   2.8e-22 !   -
5_8S_rRNA             RF00002   LSalAtl2s529.1       -         cm        1      117        7      120      +    3'    3 0.47   0.0   65.2   9.7e-17 !   -
5_8S_rRNA             RF00002   LSalAtl2s618.1       -         cm        1       80       77        1      -    3'    3 0.56   0.0   51.1   1.2e-12 !   -
U6                    RF00026   LSalAtl2s625.1       -         cm        1       82       81        1      -    3'    3 0.44   0.0   79.6   1.4e-20 !   -
U6                    RF00026   LSalAtl2s625.2       -         cm        1       81       80        1      -    3'    3 0.45   0.0   79.0   2.1e-20 !   -
U6                    RF00026   LSalAtl2s137.2       -         cm        1       82        4       84      +    3'    3 0.44   0.0   79.6   1.4e-20 !   -
U6                    RF00026   LSalAtl2s137.3       -         cm        1       82        3       83      +    3'    3 0.44   0.0   79.6   1.4e-20 !   -
LSU_rRNA_eukarya      RF02543   LSalAtl2s146.1       -         cm      201      333        1      131      + 5'&3'    4 0.45   0.0   79.9   1.9e-18 !   -
U5                    RF00020   LSalAtl2s163.1       -         cm        1      107        2      112      +    3'    3 0.39   0.0   74.4     3e-13 !   -
5_8S_rRNA             RF00002   LSalAtl2s167.3       -         cm        5      108        1      102      + 5'&3'    4 0.51   0.0   63.9     2e-16 !   -
LSU_rRNA_eukarya      RF02543   LSalAtl2s173.1       -         cm     1401     1494        1       95      + 5'&3'    4 0.56   0.0   59.2   2.9e-13 !   -
LSU_rRNA_eukarya      RF02543   LSalAtl2s183.1       -         cm     2315     2409        1       95      + 5'&3'    4 0.48   0.0  103.8   9.3e-25 !   -
U4                    RF00015   LSalAtl2s227.1       -         cm        1      140      142        2      -    no    1 0.48   0.0   74.5   1.7e-16 !   -
tRNA                  RF00005   LSalAtl2s684.1       -         cm        1       62       18       80      +    3'    3 0.57   0.0   34.6     1e-06 !   -
LSU_rRNA_eukarya      RF02543   LSalAtl2s709.1       -         cm     3079     3151       80        8      -    5'    4 0.51   0.0   52.1   1.7e-11 !   -
LSU_rRNA_eukarya      RF02543   LSalAtl2s719.1       -         cm     2356     2464        1      109      + 5'&3'    4 0.52   0.0   89.7   4.7e-21 !   -
tRNA-Sec              RF01852   LSalAtl2s724.1       -         cm        1       87       82        1      -    3'    3 0.56   0.0   60.5   6.5e-13 !   -
U1                    RF00003   LSalAtl2s25084.1     -         cm       18      154      138        1      - 5'&3'    4 0.50   0.0   80.0   7.2e-22 !   -
SSU_rRNA_eukarya      RF01960   LSalAtl2s26581.1     -         cm     1249     1355      107        1      - 5'&3'    4 0.48   0.0  123.2   1.1e-35 !   -
LSU_rRNA_eukarya      RF02543   LSalAtl2s26842.1     -         cm     1822     1963      142        1      - 5'&3'    4 0.51   0.0  131.4     1e-31 !   -
LSU_rRNA_archaea      RF02540   LSalAtl2s26842.1     -         cm     1716     1860      142        1      - 5'&3'    4 0.51   0.0   56.4   4.2e-16 !   -
SSU_rRNA_eukarya      RF01960   LSalAtl2s27909.1     -         cm      906      994        1       89      + 5'&3'    4 0.38   0.0   69.5     2e-19 !   -
tRNA                  RF00005   LSalAtl2s28940.1     -         cm        1       71       76        5      -    no    1 0.58   0.0   53.8   4.7e-12 !   -
tRNA                  RF00005   LSalAtl2s29146.1     -         cm        1       71       73        2      -    no    1 0.58   0.0   53.8   4.5e-12 !   -
tRNA                  RF00005   LSalAtl2s30337.1     -         cm        1       71       76        5      -    no    1 0.64   0.0   58.1   2.7e-13 !   -
LSU_rRNA_eukarya      RF02543   LSalAtl2s30460.1     -         cm      864     1001      139        1      - 5'&3'    4 0.50   0.0  139.9   6.4e-34 !   -
LSU_rRNA_archaea      RF02540   LSalAtl2s30460.1     -         cm      839      977      139        1      - 5'&3'    4 0.50   0.0   64.7     1e-18 !   -
tRNA                  RF00005   LSalAtl2s30615.1     -         cm        1       65       76        1      -    3'    3 0.55   0.0   56.2   9.8e-13 !   -
tRNA                  RF00005   LSalAtl2s31030.1     -         cm        1       71       73        1      -    no    1 0.57   0.0   49.6   6.8e-11 !   -
SSU_rRNA_eukarya      RF01960   LSalAtl2s32224.2     -         cm      401      485        1       85      + 5'&3'    4 0.49   0.0   82.5   2.2e-23 !   -
SSU_rRNA_microsporidia RF02542  LSalAtl2s32224.2     -         cm      258      342        1       85      + 5'&3'    4 0.49   0.0   55.3     2e-15 !   -
LSU_rRNA_eukarya      RF02543   LSalAtl2s33285.1     -         cm        1      107      107        1      -    3'    3 0.42   0.0  112.4   6.7e-27 !   -
LSU_rRNA_archaea      RF02540   LSalAtl2s33285.1     -         cm      130      248      116        1      - 5'&3'    4 0.40   0.0   67.8   8.5e-20 !   -
LSU_rRNA_bacteria     RF02541   LSalAtl2s33285.1     -         cm      150      267      116        1      - 5'&3'    4 0.40   0.0   59.3   3.4e-18 !   -
tRNA                  RF00005   LSalAtl2s35394.1     -         cm        1       71        8       79      +    no    1 0.58   0.0   53.8   4.5e-12 !   -
U1                    RF00003   LSalAtl2s92.1        -         cm        1      148        1      147      +    3'    3 0.49   0.0  110.6   5.7e-31 !   -
U1                    RF00003   LSalAtl2s92.2        -         cm       25      166        1      140      +    5'    2 0.51   0.0   96.4   9.4e-27 !   -
U1                    RF00003   LSalAtl2s92.3        -         cm       25      158        1      132      + 5'&3'    4 0.52   0.0   84.9   2.4e-23 !   -
U1                    RF00003   LSalAtl2s92.4        -         cm       54      158        1      103      + 5'&3'    4 0.51   0.0   62.7   7.6e-17 !   -
U1                    RF00003   LSalAtl2s92.5        -         cm       22      158        1      135      + 5'&3'    4 0.51   0.0   88.1   2.8e-24 !   -
U1                    RF00003   LSalAtl2s92.6        -         cm       55      166        1      110      +    5'    2 0.50   0.0   72.9   8.2e-20 !   -
U1                    RF00003   LSalAtl2s92.7        -         cm       48      158        1      109      + 5'&3'    4 0.51   0.0   64.9   1.9e-17 !   -
LSU_rRNA_eukarya      RF02543   LSalAtl2s98.2        -         cm      307      387        1       81      + 5'&3'    4 0.41   0.0   77.2   5.7e-18 !   -
LSU_rRNA_bacteria     RF02541   LSalAtl2s98.2        -         cm      457      538        1       81      + 5'&3'    4 0.41   0.0   50.5   2.1e-15 !   -
LSU_rRNA_eukarya      RF02543   LSalAtl2s98.5        -         cm      799      914        1      116      + 5'&3'    4 0.56   0.0  118.0   2.4e-28 !   -
LSU_rRNA_archaea      RF02540   LSalAtl2s98.5        -         cm      774      890        1      116      + 5'&3'    4 0.56   0.0   67.2   1.4e-19 !   -
LSU_rRNA_eukarya      RF02543   LSalAtl2s98.6        -         cm     1104     1213        1      109      + 5'&3'    4 0.47   0.0   84.1   1.2e-19 !   -
LSU_rRNA_eukarya      RF02543   LSalAtl2s98.7        -         cm     1241     1374        1      137      + 5'&3'    4 0.48   0.0  120.6   6.2e-29 !   -
LSU_rRNA_archaea      RF02540   LSalAtl2s98.7        -         cm     1179     1287        1      109      +    5'    4 0.54   0.0   62.3   5.7e-18 !   -
LSU_rRNA_eukarya      RF02543   LSalAtl2s98.9        -         cm     1879     1998        1      120      + 5'&3'    4 0.46   0.0   95.5   1.6e-22 !   -
LSU_rRNA_eukarya      RF02543   LSalAtl2s102.1       -         cm     2192     2272       81        1      - 5'&3'    4 0.51   0.0   95.9   8.7e-23 !   -
U4                    RF00015   LSalAtl2s107.3       -         cm        1      140      148        9      -    no    1 0.47   0.0   78.8   1.4e-17 !   -
LSU_rRNA_eukarya      RF02543   LSalAtl2s116.1       -         cm      914      996        1       83      + 5'&3'    4 0.48   0.0   79.6   1.4e-18 !   -
U1                    RF00003   LSalAtl2s136.1       -         cm       12      154        1      143      + 5'&3'    4 0.52   0.0   85.9   1.3e-23 !   -
LSU_rRNA_eukarya      RF02543   LSalAtl2s2717.8      -         cm     1115     1258      143        1      - 5'&3'    4 0.50   0.0  137.6   2.6e-33 !   -
LSU_rRNA_eukarya      RF02543   LSalAtl2s2717.9      -         cm     1058     1165      111        1      - 5'&3'    4 0.44   0.0   77.6     6e-18 !   -
tRNA                  RF00005   LSalAtl2s2894.1      -         cm        1       71       76        5      -    no    1 0.58   0.0   53.8   4.7e-12 !   -
5S_rRNA               RF00001   LSalAtl2s2894.2      -         cm        1       75       75        1      -    3'    3 0.41   0.0   43.6   1.2e-08 !   -
5S_rRNA               RF00001   LSalAtl2s2894.3      -         cm        1      110      110        1      -    3'    3 0.45   0.0   84.6   4.4e-19 !   -
tRNA                  RF00005   LSalAtl2s2894.4      -         cm        1       71       73        1      -    no    1 0.66   0.0   54.7   2.5e-12 !   -
tRNA                  RF00005   LSalAtl2s2894.5      -         cm        1       71        9       80      +    no    1 0.53   0.0   60.9     6e-14 !   -
tRNA                  RF00005   LSalAtl2s2894.6      -         cm        1       62       96        1      -    3'    3 0.56   0.0   34.0   1.9e-06 !   -
tRNA                  RF00005   LSalAtl2s2894.7      -         cm        1       71       82        1      -    no    1 0.59   0.0   72.5   2.6e-17 !   -
tRNA                  RF00005   LSalAtl2s2894.8      -         cm        1       69        5       97      +    3'    3 0.47   0.0   50.6   4.1e-11 !   -
tRNA                  RF00005   LSalAtl2s2894.9      -         cm        1       71        7       78      +    no    1 0.64   0.0   58.1   2.8e-13 !   -
tRNA                  RF00005   LSalAtl2s2894.10     -         cm        1       71       17       87      +    no    1 0.66   0.0   55.5   1.6e-12 !   -
tRNA                  RF00005   LSalAtl2s2894.11     -         cm        1       71        7       78      +    no    1 0.64   0.0   58.1   2.8e-13 !   -
tRNA                  RF00005   LSalAtl2s2894.12     -         cm        1       71      137       65      -    no    1 0.58   0.0   60.8   8.5e-14 !   -
tRNA                  RF00005   LSalAtl2s2894.13     -         cm        1       71        6       99      +    no    1 0.49   0.0   59.7   1.6e-13 !   -
tRNA                  RF00005   LSalAtl2s2894.14     -         cm        1       71        6       77      +    no    1 0.64   0.0   58.1   2.7e-13 !   -
tRNA                  RF00005   LSalAtl2s2894.15     -         cm        1       71        8       78      +    no    1 0.65   0.0   65.3   2.7e-15 !   -
tRNA                  RF00005   LSalAtl2s2894.16     -         cm        1       71        7       78      +    no    1 0.60   0.0   58.6   2.1e-13 !   -
tRNA                  RF00005   LSalAtl2s2894.17     -         cm        1       66      115        1      -    3'    3 0.50   0.0   41.8   1.5e-08 !   -
tRNA                  RF00005   LSalAtl2s2894.18     -         cm        1       69       87        1      -    3'    3 0.49   0.0   49.3   9.2e-11 !   -
tRNA                  RF00005   LSalAtl2s3165.1      -         cm        1       61       74        1      -    3'    3 0.58   0.0   40.7   2.1e-08 !   -
tRNA                  RF00005   LSalAtl2s4604.1      -         cm        1       63       97        1      -    3'    3 0.57   0.0   35.7   6.6e-07 !   -
tRNA                  RF00005   LSalAtl2s4604.2      -         cm        1       71       19       90      +    no    1 0.62   0.0   57.9   3.6e-13 !   -
tRNA                  RF00005   LSalAtl2s4916.1      -         cm        1       71       75        3      -    no    1 0.66   0.0   54.7   2.6e-12 !   -
LSU_rRNA_eukarya      RF02543   LSalAtl2s5200.1      -         cm     2968     3110        1      143      + 5'&3'    4 0.51   0.0  131.4     1e-31 !   -
tRNA                  RF00005   LSalAtl2s6836.1      -         cm        1       71       73        2      -    no    1 0.58   0.0   49.7   6.4e-11 !   -
5S_rRNA               RF00001   LSalAtl2s7121.1      -         cm        1      119      121        3      -    no    1 0.46   0.0  105.4   2.2e-24 !   -
tRNA                  RF00005   LSalAtl2s837.1       -         cm        1       63        7       80      +    3'    3 0.51   0.0   41.8     1e-08 !   -
SSU_rRNA_eukarya      RF01960   LSalAtl2s874.1       -         cm      785      921        1      139      + 5'&3'    4 0.37   0.0   64.2   1.4e-17 !   -
U6                    RF00026   LSalAtl2s887.1       -         cm        1       82        4       84      +    3'    3 0.44   0.0   79.6   1.4e-20 !   -
SSU_rRNA_eukarya      RF01960   LSalAtl2s943.1       -         cm     1319     1485        1      140      + 5'&3'    4 0.41   0.0   83.9   1.3e-23 !   -
tRNA                  RF00005   LSalAtl2s1130.1      -         cm        1       67        8       92      +    3'    3 0.49   0.0   45.6     1e-09 !   -
LSU_rRNA_eukarya      RF02543   LSalAtl2s1442.1      -         cm     1400     1506        1      108      + 5'&3'    4 0.58   0.0   80.8     9e-19 !   -
SSU_rRNA_eukarya      RF01960   LSalAtl2s1561.1      -         cm     1584     1698      115        1      - 5'&3'    4 0.52   0.0  123.8   7.7e-36 !   -
SSU_rRNA_microsporidia RF02542  LSalAtl2s1561.1      -         cm     1090     1202      115        1      - 5'&3'    4 0.52   0.0   76.5   7.7e-22 !   -
SSU_rRNA_archaea      RF01959   LSalAtl2s1561.1      -         cm     1240     1354      115        1      - 5'&3'    4 0.52   0.0   72.0   8.5e-20 !   -
SSU_rRNA_bacteria     RF00177   LSalAtl2s1561.1      -         cm     1285     1399      115        1      - 5'&3'    4 0.52   0.0   62.2   2.4e-18 !   -
snoR639               RF00291   LSalAtl2s1699.10     -         cm        1      140        2      136      +    no    1 0.35   0.0   54.7   1.7e-11 !   -
```



```
LSU_rRNA_eukarya     RF02543   LSalAtl2s1851.1      -          cm     2194     2274       81        1      - 5'&3'    4 0.51   0.0   96.7   5.3e-23 !   -
tRNA                 RF00005   LSalAtl2s2021.1      -          cm        1       61        7       90      +    3'    3 0.46   0.0   36.4   3.6e-07 !   -
tRNA                 RF00005   LSalAtl2s2021.2      -          cm        1       71        8       78      +    no    1 0.65   0.0   65.3   2.7e-15 !   -
tRNA                 RF00005   LSalAtl2s2021.3      -          cm        1       65       83        1      -    3'    3 0.49   0.0   41.9   1.1e-08 !   -
tRNA                 RF00005   LSalAtl2s2021.4      -          cm        1       71       76        5      -    no    1 0.58   0.0   53.8   4.5e-12 !   -
5S_rRNA              RF00001   LSalAtl2s2021.5      -          cm        1      119      120        2      -    no    1 0.47   0.0  102.7     1e-23 !   -
tRNA                 RF00005   LSalAtl2s2021.6      -          cm        1       71       76        4      -    no    1 0.66   0.0   54.7   2.6e-12 !   -
tRNA                 RF00005   LSalAtl2s2021.8      -          cm        1       71        7       78      +    no    1 0.62   0.0   57.9     5e-13 !   -
5S_rRNA              RF00001   LSalAtl2s2606.1      -          cm        1      119        7      125      +    no    1 0.46   0.0  105.4   2.2e-24 !   -
LSU_rRNA_eukarya     RF02543   LSalAtl2s2717.2      -          cm     2975     3106      132        1      - 5'&3'    4 0.50   0.0  115.5   1.2e-27 !   -
LSU_rRNA_eukarya     RF02543   LSalAtl2s2717.4      -          cm     2941     3046      106        1      - 5'&3'    4 0.48   0.0  110.1   2.4e-26 !   -
LSU_rRNA_archaea     RF02540   LSalAtl2s2717.4      -          cm     2660     2767      106        1      - 5'&3'    4 0.48   0.0   50.5   2.3e-14 !   -
LSU_rRNA_eukarya     RF02543   LSalAtl2s2717.5      -          cm     2556     2699      131        1      - 5'&3'    4 0.52   0.0  111.0   1.7e-26 !   -
LSU_rRNA_archaea     RF02540   LSalAtl2s2717.5      -          cm     2281     2417      131        1      - 5'&3'    4 0.52   0.0   60.4   2.1e-17 !   -
LSU_rRNA_bacteria    RF02541   LSalAtl2s2717.5      -          cm     2215     2347      131        1      - 5'&3'    4 0.52   0.0   54.9   1.1e-16 !   -
LSU_rRNA_eukarya     RF02543   LSalAtl2s2717.7      -          cm     1429     1565      137        1      - 5'&3'    4 0.57   0.0  117.5   3.9e-28 !   -
5S_rRNA              RF00001   LSalAtl2s7137.1      -          cm        1       99      100        1      -    3'    3 0.46   0.0   63.7     1e-13 !   -
tRNA                 RF00005   LSalAtl2s7659.1      -          cm        1       61        9       80      +    3'    3 0.54   0.0   49.2   8.7e-11 !   -
U3                   RF00012   LSalAtl2s7695.1      -          cm       93      215      127        1      -    5'    2 0.43   0.0   52.4   1.2e-11 !   -
5S_rRNA              RF00001   LSalAtl2s7914.1      -          cm        1      119      120        2      -    no    1 0.46   0.0  105.4   2.2e-24 !   -
LSU_rRNA_eukarya     RF02543   LSalAtl2s8461.1      -          cm      743      880      147        1      - 5'&3'    4 0.52   0.0   85.1   9.4e-20 !   -
LSU_rRNA_eukarya     RF02543   LSalAtl2s9058.2      -          cm     1431     1553      123        1      - 5'&3'    4 0.55   0.0  109.6   3.8e-26 !   -
U2                   RF00004   LSalAtl2s9288.1      -          cm        1       97        1       98      +    3'    3 0.39   0.0  100.1   5.6e-22 !   -
LSU_rRNA_eukarya     RF02543   LSalAtl2s9709.1      -          cm     2657     2786        1      130      + 5'&3'    4 0.53   0.0  122.8   1.5e-29 !   -
LSU_rRNA_archaea     RF02540   LSalAtl2s9709.1      -          cm     2375     2506        1      130      + 5'&3'    4 0.53   0.0   73.1   2.1e-21 !   -
LSU_rRNA_bacteria    RF02541   LSalAtl2s9709.1      -          cm     2305     2433        1      130      + 5'&3'    4 0.53   0.0   60.3   1.7e-18 !   -
LSU_rRNA_eukarya     RF02543   LSalAtl2s9709.2      -          cm     2752     2844        1       92      + 5'&3'    4 0.44   0.0   77.7   4.7e-18 !   -
U3                   RF00012   LSalAtl2s10026.1     -          cm        1      107        1      107      +    3'    3 0.32   0.0   55.2   1.9e-12 !   -
LSU_rRNA_eukarya     RF02543   LSalAtl2s11911.1     -          cm     2417     2509        1       95      + 5'&3'    4 0.41   0.0   54.9   3.8e-12 !   -
tRNA                 RF00005   LSalAtl2s12704.1     -          cm        1       61        9       80      +    3'    3 0.54   0.0   49.2   8.7e-11 !   -
tRNA                 RF00005   LSalAtl2s12704.2     -          cm        1       71       10       82      +    no    1 0.55   0.0   45.4     1e-09 !   -
U2                   RF00004   LSalAtl2s12958.1     -          cm       49      176        1      128      + 5'&3'    4 0.47   0.0   66.9   6.8e-14 !   -
LSU_rRNA_eukarya     RF02543   LSalAtl2s14271.3     -          cm     2878     2977      100        1      - 5'&3'    4 0.53   0.0  127.6   6.8e-31 !   -
LSU_rRNA_archaea     RF02540   LSalAtl2s14271.3     -          cm     2597     2696      100        1      - 5'&3'    4 0.53   0.0   85.1   2.5e-25 !   -
LSU_rRNA_bacteria    RF02541   LSalAtl2s14271.3     -          cm     2526     2625      100        1      - 5'&3'    4 0.53   0.0   74.7     2e-23 !   -
LSU_rRNA_eukarya     RF02543   LSalAtl2s14271.4     -          cm     2842     2946      105        1      - 5'&3'    4 0.51   0.0  129.8     2e-31 !   -
LSU_rRNA_archaea     RF02540   LSalAtl2s14271.4     -          cm     2561     2665      105        1      - 5'&3'    4 0.51   0.0   65.9   3.1e-19 !   -
LSU_rRNA_bacteria    RF02541   LSalAtl2s14271.4     -          cm     2490     2594      105        1      - 5'&3'    4 0.51   0.0   58.2     7e-18 !   -
tRNA                 RF00005   LSalAtl2s14316.1     -          cm        1       61        9       80      +    3'    3 0.54   0.0   49.2   8.7e-11 !   -
SSU_rRNA_eukarya     RF01960   LSalAtl2s14525.1     -          cm      477      596        1      121      +    5'    4 0.50   0.0  112.9   1.8e-32 !   -
SSU_rRNA_microsporidia RF02542 LSalAtl2s14525.1     -          cm      373      420       74      121      +    no    4 0.58   0.0   52.4   2.3e-14 !   -
SSU_rRNA_eukarya     RF01960   LSalAtl2s16178.1     -          cm     1778     1851        1       74      +    5'    2 0.45   0.0   81.1   8.3e-23 !   -
SSU_rRNA_microsporidia RF02542 LSalAtl2s16178.1     -          cm     1239     1312        1       74      +    5'    2 0.45   0.0   53.4   1.1e-14 !   -
U3                   RF00012   LSalAtl2s18926.1     -          cm       72      188        3      121      +    3'    4 0.38   0.0   46.0   5.2e-10 !   -
LSU_rRNA_bacteria    RF02541   LSalAtl2s21316.1     -          cm     2213     2340      116        1      - 5'&3'    4 0.50   0.0   74.8   2.1e-23 !   -
LSU_rRNA_archaea     RF02540   LSalAtl2s21316.1     -          cm     2293     2410      116        1      - 5'&3'    4 0.50   0.0   51.0   1.7e-14 !   -
LSU_rRNA_eukarya     RF02543   LSalAtl2s21406.1     -          cm     1337     1452      132        1      - 5'&3'    4 0.47   0.0   66.1   6.8e-15 !   -
U6                   RF00026   LSalAtl2s21566.1     -          cm        1       82        4       84      +    3'    3 0.44   0.0   79.6   1.4e-20 !   -
U1                   RF00003   LSalAtl2s25302.1     -          cm       19      154        1      136      + 5'&3'    4 0.50   0.0   78.8   1.6e-21 !   -
U3                   RF00012   LSalAtl2s25724.1     -          cm       56      184        1      128      + 5'&3'    4 0.35   0.0   47.0     3e-10 !   -
tRNA                 RF00005   LSalAtl2s1130.2      -          cm        1       71        8      127      +    no    1 0.50   0.0   56.3   1.4e-12 !   -
tRNA                 RF00005   LSalAtl2s1236.1      -          cm        1       71        8       79      +    no    1 0.61   0.0   36.0   4.2e-07 !   -
LSU_rRNA_eukarya     RF02543   LSalAtl2s1398.1      -          cm     2556     2671        1      125      + 5'&3'    4 0.46   0.0   77.7   6.4e-18 !   -
```

Fig. 24 Results for the salmon louse data after searching the high coverage sequences with Cmscan

Results:

Blast is commonly used for sequence similarity searches which finds regions of similarity between biological sequences, however, blast-based searches beyond mammals have not been successful. And it is ineffective for identifying vtRNA sequences which are very highly conserved at the nucleotide level. Through searching the high coverage candidate sequences in Rfam by Cmscan, the pipeline unfortunately cannot find the vtRNA in the salmon louse gene.

So we need to consider the secondary structures of vtRNA and make the cluster dendrogram by the dissimilarity matrix. Take the structures of the known vtRNA as references and compare the distances between all the candidate sequences and then make a dissimilarity matrix . The lower result is, the more similar between the two structures are.

The next step is hierarchical cluster analysis on a set of dissimilarities and generate cluster dendrogram. A dendrogram is a tree diagram frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering. Dendrograms are often used in computational biology to illustrate the clustering of genes or samples, sometimes on top of heatmaps. The dendrogram is a visual representation of the compound correlation data. The individual compounds are arranged along the bottom of the dendrogram and referred to as leaf nodes. Compound clusters are formed by joining individual compounds or existing compound clusters with the join point referred to as a node. At each dendrogram node we have a right and left sub-branch of clustered compounds.

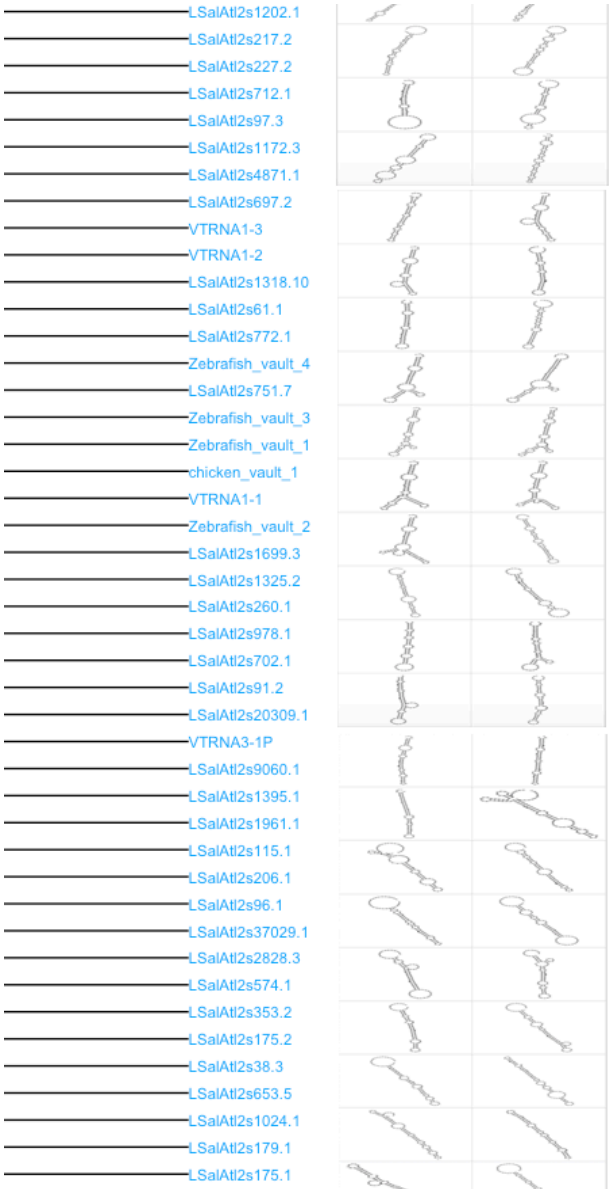


Fig.25 part of the cluster dengrogram of high coverage sequences from salmon louse data

According to cluster dendrogram, most of the known vtRNA are assigned into the same cluster, which means that they have low dissimilarities. It is noticed that the secondary structures in the same clusters, especially those which are closed to the known vtRNA, have the similar structures. This implies that these structures can probably be the vtRNA and the pipeline can take those as candidate sequences for further research. But it can only reduce the number of candidate sequences by this method since the specific secondary structures of vtRNA are not certain.

Motif GCYURAVWGANAGCDRAWCCM MEME-2 sites sorted by position p-value

Sequence name	Start	P-value		Site	
LSalAtl2s1172.3	27	6.15e-10	AAGACCCAUA	GCCUCAGAGAGAGCGAAUCCA	AGAAUGGCGU
LSalAtl2s175.2	60	2.21e-09	CGGGUGAGGA	GCUUGAAAUACAGCUGAACCC	ACCUUACCUG
LSalAtl2s751.7	51	4.63e-09	UCCUUAAAUG	AGCUAACUGCAAGCGGAUCCC	UAUUUAUUUA
LSalAtl2s175.1	60	6.50e-09	AGGGAAAGGA	GCUUGAAACAUAGCUGAACCC	WCCUUUUCUU
LSalAtl2s96.1	65	1.60e-07	AUCUGUCAAG	ACUGAAAUGGGAGCUGAACGC	AUGAUCAAGA
LSalAtl2s653.5	41	2.36e-07	GUUUGCAUUG	GCCAAAGAUCCAGAGUAUCCC	UGAAGUUCAG
LSalAtl2s1699.3	75	2.66e-07	CCACCAUAAU	GCUUAACUGAUUUCAAAUCCA	AGAAAUUGCU
LSalAtl2s20309.1	58	2.83e-07	UAAACCUAAA	GGCGUAGUGAAAGCAAAUGCA	GGGAUGAUCC

De novo motifs detection: The vtRNA contains two B-box elements and one A-box element, if there are sequences with such structures, they can probably be vtRNA. Use the MEME for finding the motifs in the clusters and highlight the motifs at the secondary structures.

Having generated the cluster dendrogram according to the dissimilarity matrix from the last step with the sequence names and corresponding secondary structures. In total there are 8 clusters and most of the known vtRNA clustered in the “BLUE” cluster (Fig.25) with kind of similar

Motif CYAARGASAWYVSAGUGKDUUCCMWGGHCNAUMNAKUNGRDUMCMMSG MEME-1 sites sorted by position p-value

Sequence name	Start	P-value		Site	
LSalAtl2s9060.1	8	2.30e-23	GUGGAUU	CCAAGGACAAUCCAGUGGAUUCCAAGGUCAAUCCAGUGGAUUCCAAGG	WCAAUCCAGU
LSalAtl2s4871.1	16	1.32e-22	AUGGUGGAUU	CCAAGGACAAUCCAGUGGAUUCCAAGGACAAUCCAGUGGAUUCCAAGG	ACAAUCCAAU
LSalAtl2s697.2	9	1.11e-16	GCGCUACA	CUGAAGGGAUCAGCGUGUUUUCCCUGGCCGAGAGGUUCGGGUAACCCG	UUGAACCCCC
LSalAtl2s702.1	35	9.03e-16	GGUGACUCCU	ACAAGGACUUUGGAGAGGUUUUCCUUCCCCUUAUCUGUGAGUACCACG	GMC
LSalAtl2s179.1	21	4.64e-15	AGGGCUGAGU	CUCAAUAGAUCGCAGUGUGGUGGCUGSUCUACCAAGUACGACMCCCCG	GCCGGUACAU

Motif ACYUCYWCHUBRAWANRUCMA MEME-3 sites sorted by position p-value

Sequence name	Start	P-value		Site	
LSalAtl2s574.1	16	5.77e-09	UUUACAGAUU	ACUUCUUCUUGGAUACACCAA	CAGUUCUACC
LSalAtl2s217.2	10	4.39e-08	CUUGUCGAU	ACCUCCACAUUGGUAGAUCAA	AUGACCGGUG
LSalAtl2s712.1	30	1.11e-07	GAUAUCUGGU	ACCUCUUACAUGAAAGCUCAA	AACCCAUGAU
LSalAtl2s653.5	77	1.81e-07	UUCAGCAUUA	ACUUCUCCUUCAAUUUGUCCU	UCUACUUGUG
LSalAtl2s260.1	59	9.06e-07	ACCCAAAAUA	ACUUCCUCAAGGCCUAAUCCA	UCCUUUGCUU
LSalAtl2s1202.1	90	1.74e-06	UCCAACCCGU	ACUUCUUCUUGAUAAGAGCCU	UG
LSalAtl2s175.1	81	2.91e-06	AGCUGAACCC	WCCUUUUCUUGAAUGCGUCAU	GUGCCUU
LSalAtl2s206.1	14	3.16e-06	CGUUCAAAUU	UCCUUCCCUUGGAUCGGACAA	AAGGUUUGGA
LSalAtl2s1395.1	44	3.16e-06	CACGGACUUG	ACUUCUUCUUUGAUGUCUGCG	AAUGCAUUUU
LSalAtl2s38.3	34	4.01e-06	GAUUCCCUUU	AACUCUACCAACAAAUAUCAA	GUAUCUAUUC
LSalAtl2s978.1	43	6.79e-06	GUAUUAAUGA	ACUUCUAACUUCACAUGACGA	UGAAUCCUUG
LSalAtl2s1325.2	55	7.29e-06	GACUUGUCUA	ACUUUCAAGACAAAUCAUCCA	GUCAAAAGGG
LSalAtl2s115.1	23	8.99e-06	CRUAUAAUGG	UACUCCAAAUCGAGAAAUCCA	UACGUAUAUA

secondary structures, which means that these candidate sequences are in the same cluster with the known vtRNAs and can probably be the vtRNA. Detect the motifs of the sequences in the cluster with MEME Suite.

In the “BLUE” cluster and other clusters, first remove the known vtRNA since the known vtRNA here act as a reference, while they shouldn't be counted into detecting.

Fig.26 Motifs generated from candidate sequences from the “BLUE” cluster (Fig.25)

By default MEME, it finds 3 motifs. It tries to find the best motifs first but due to the enormous search space it is impossible to guarantee that they will always be listed best to worst. Always check the P-value of the motifs found by MEME as sometimes the motifs found will not be statistically significant. Generally if a motif has an P-value larger than 0.05 it is not significant.

Map the motifs back to the sequences and pick those sequences with at least 2 motifs in the same sequence.

1. Highlight the sequences with motifs in the secondary structures in all of the clusters. While not all of the sequences in different clusters have motifs.
2. Find the sequences with 3 motifs in different cluster since we know vtRNA has one box A and two box B elements. Those sequences can with motifs can be the potential vtRNA for further research.

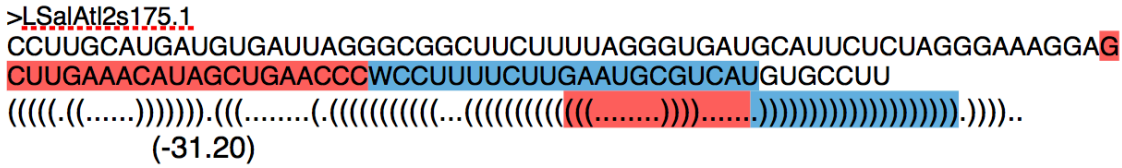


Fig.26 candidate sequence with two motifs

Final candidate:

The pipeline finds only one sequence with two motifs which could be box A and Box B, the secondary structure of LSalAt2s175.1 is depicted in Fig. 27

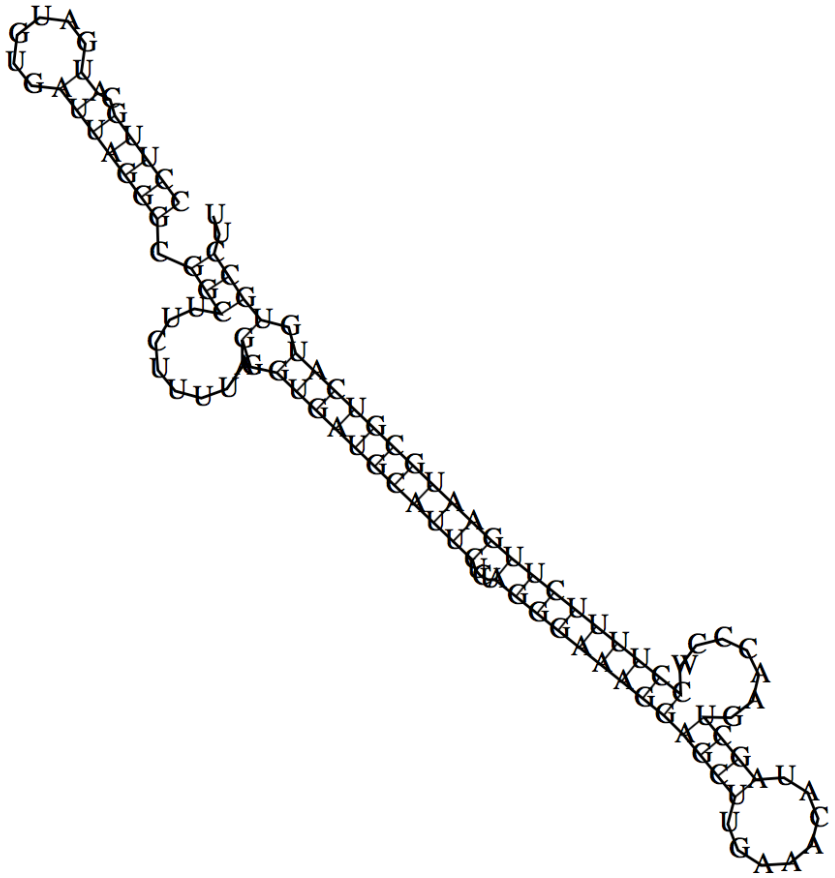


Fig.27 MFE secondary structure of LSalAt2s175.1

This sequence is a good candidate for a vtRNA, but for further validating the result, it should be verified by a laboratory experiment, such as purification of the whole vault and sequencing of all bound RNA. The result can only be checked by the help of further laboratory work.

5. Discussion and Further work

In my thesis, I have described the development and application of a vaultR, a pipeline for de-novo detection of vault-RNA from RNA-sequencing data. I have tested vaultR on two data sets, simulated data and real data from Atlantic salmon louse, an important fish parasite. The pipeline successfully detects vtRNA in the simulated data using Rfam and cmscan, while it fails to find vtRNA in real data from the Atlantic salmon louse that way. Thus, the pipeline goes further to secondary structure prediction and cluster analysis, and then to de novo detection of motifs in vtRNA candidates for real data. As a final result, there are some good candidate sequences which match the structural features of vtRNA.

However, the main obstacle is that we do not know the true specific structure of vtRNA, and there is little relevant research on that topic. The available structures are conserved only in the small stem portion of the vtRNA, and amount of validated vtRNA is relatively small. By now there are only 6 from human, 2 from chicken, 1 from mouse, 1 from rat and 4 from zebrafish. Machine learning algorithms like decision trees could be a good way to predict the result if there are large enough training data in the future.

For the candidate sequences, all the bioinformatic work is finished here. The pipeline is able to find the potential vtRNA and reduce the number of candidate sequences. With these in hand, we need to come back to laboratory to get a validated vtRNA sequence and compare the potential vtRNA candidates which are generated by the pipeline with the results of the experiment.

There are also some more structural features which can be used for identifying vtRNA, for example, termination signal poly U-tail at 3’ end, and at least 2 U are unpaired, and the poly U-tail are not far away from initial pair. Most of the vtRNA has also opening stems/bulges. This can also be as the factor to identify vtRNA. More research needs to be done on these sequence features and other ncRNA motifs to develop better algorithms for vtRNA detection in the future.

Current Rfam structure

5 out of **19** basepairs are significant at E-value=0.05

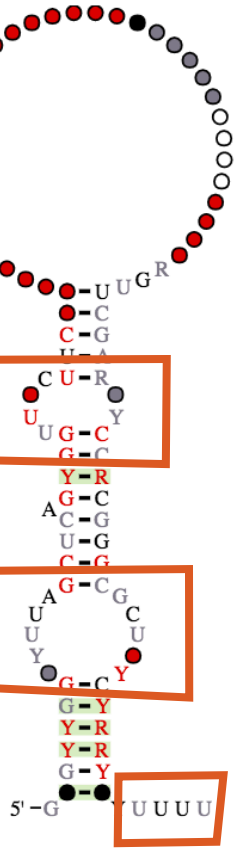


Fig.28 vtRNA generated from Rfam

6. References / Bibliography:

[1] Kedersha, N.L., et al., Vaults. II. Ribonucleoprotein structures are highly conserved among higher and lower eukaryotes. J Cell Biol, 1990. 110(4): p. 895-901.

[2] V.A. Kickhoefer, R.P. Searles, N.L. Kedersha, M.E. Garber, D.L. Johnson, L.H. Rome. Vault ribonucleoprotein particles from rat and bullfrog contain a related small RNA that is transcribed by RNA polymerase III. J. Biol. Chem., 268 (1993), pp. 7868–7873.

[3] Kedersha, N. L., Heuser, J. E., Chugani, D. C. and Rome, L. H. (1991) Vaults. III. Vault Ribonucleoprotein Particles Open into Flower-like Structures with Octagonal Symmetry. J Cell Biol 112, 225-35.

[4] van Zon A, Mossink MH, Scheper RJ, Sonneveld P, Wiemer EA (September 2003). "The vault complex". Cellular and Molecular Life Sciences. 60 (9): 1828–37. PMID 14523546. doi:10.1007/s00018-003-3030-y.

[5] V.A. Kickhoefer, K.S. Rajavel, G.L. Scheffer, W.S. Dalton, R.J. Scheper, L.H. Rome. Vaults are up-regulated in multidrug-resistant cancer cell lines. J. Biol. Chem., 273 (1998), pp. 8971–8974.

[6] Rome, Leonard. "Vaults. Novel nano particles". http://www.vaults.arc.ucla.edu. Computing Technologies Research Lab. External link in |website= (help)

[7] van Zon, A., et al., Multiple human vtRNAs. Expression and association with the vault complex. J Biol Chem, 2001. 276(40): p. 37715-21.195-202.

[8] Mossink MH, van Zon A, Scheper RJ, Sonneveld P, Wiemer EA (October 2003). "Vaults: a ribonucleoprotein particle involved in drug resistance?". Oncogene. 22(47): 7458–67. PMID 14576851. doi:10.1038/sj.onc.1206947.

[9] Kickhoefer VA, Vasu SK, Rome LH (May 1996). "Vaults are the answer, what is the question?". Trends in Cell Biology. 6 (5): 174–8. PMID 15157468. doi:10.1016/0962-8924(96)10014-3.

[10] Marieke H Mossink[1], Arend van Zon[1], Rik J Scheper[2], Pieter Sonneveld[1] and Erik AC Wiemer[1]. Vaults: a ribonucleoprotein particle involved in drug resistance? Oncogene (2003) 22, 7458–7467. doi:10.1038/sj.onc.1206947

[11] Standler, Peter F.; Chen, Julian J.-L.; Hackermuller, Jorg (June 2, 2009). "Evolution of Vault RNAs". Molecular Biology and Evolution. 26 (9): 1975–1991. doi:10.1093/molbev/msp112.

[12] Weichun Huang,1,* Leping Li,1 Jason R. Myers,1,† and Gabor T. Marth2,* ART: a next-generation sequencing read simulator. Bioinformatics. 2012 Feb 15; 28(4): 593–594.

[13] Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60. [PMID: 19451168]

[14] Li H.*, Handsaker B.*, Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9. [PMID: 19505943]

[15] Li H A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011 Nov 1;27(21):2987-93. Epub 2011 Sep 8. [PMID: 21903627]

[16] James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. Integrative Genomics Viewer. Nature Biotechnology 29, 24–26 (2011)

[17] Helga Thorvaldsdóttir, James T. Robinson, Jill P. Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics 14, 178-192 (2013).

[18] Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

[19] The EMBL-EBI bioinformatics web and programmatic tools framework. (2015 July 01) Nucleic acids research 43 (W1) :W580-4 PMID: 25845596

[20] Rfam 12.0: updates to the RNA families database. Eric P. Nawrocki, Sarah W. Burge, Alex Bateman, Jennifer Daub, Ruth Y. Eberhardt, Sean R. Eddy, Evan W. Floden, Paul P. Gardner, Thomas A. Jones, John Tate and Robert D. Finn
[21] Lorenz, Ronny and Bernhart, Stephan H. and Höner zu Siederdissen, Christian and Tafer, Hakim and Flamm, Christoph and Stadler, Peter F. and Hofacker, Ivo L. ViennaRNA Package 2.0 Algorithms for Molecular Biology, 6:1 26, 2011, doi: 10.1186/1748-7188-6-26

[22] Timothy L. Bailey and Charles Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers", Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pp. 28-36, AAAI Press, Menlo Park, California, 1994.

[23] Axel Mosig, Julian Chen, Peter F. Stadler, Homology Search with Fragmented Nucleic Acid Sequence Patterns, Proc. Worksh. Alg. Bioinf. (WABI), 2007.

[24] Sam Behjati1,2 and Patrick S Tarpey1 "What is next generation sequencing?" Arch Dis Child Educ Pract Ed. 2013 Dec; 98(6): 236–238. Published online 2013 Aug 28. doi: 10.1136/archdischild-2013-304340
[25] Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS (August 2012). "The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments". Nat Protoc. 7 (8): 1534–50. PMC 3535016 . PMID 22836135. doi:10.1038/nprot.2012.086
[26] https://en.wikipedia.org/wiki/RNA-Seq
[27] Rolf Brudvik Edvardsen, , Sussie Dalvin, Tomasz Furmanek, Ketil Malde, Stig Mæhle, Bjørn Olav Kvamme, Rasmus Skern-Mauritzen "Gene expression in five salmon louse (Lepeophtheirus salmonis, Krøyer 1837) tissues" https://doi.org/10.1016/j.margen.2014.06.008
[28] http://sealouse.imr.no/ "The Salmon Louse Genome Project"

[29] Benchmarking short sequence mapping tools Ayat Hatem, Doruk Bozdağ, Amanda E Toland, Ümit V Çatalyürek BMC Bioinformatics. 2013; 14: 184. Published online 2013 Jun 7. doi: 10.1186/1471-2105-14-184 PMCID: PMC3694458

[30] Bronwen L. Aken, Sarah Ayling, Daniel Barrell1, Laura Clarke, Valery Curwen, Susan Fairley, Julio Fernandez Banet, Konstantinos Billis, Carlos Garcín Girón, Thibaut Hourlier, Kevin Howe, Andreas Kähäri, Felix Kokocinski, Fergal J. Martin, Daniel N. Murphy, Rishi Nag, Magali Ruffier, Michael Schuster, Y. Amy Tang, Jan-Hinnerk Vogel, Simon White, Amonida Zadissa, Paul Flicek and Stephen M. J. Searle The Ensembl gene annotation system Database 2016, baw093 doi: 10.1093/database/baw093

[31] Valouev A, et al. (September 2008). "Genome-wide analysis of transcription factor binding sites based on ChIP-seq data". Nature Methods. 5: 829–834. PMC 2917543 . PMID 19160518. doi:10.1038/nmeth.1246.

[32] Kong, Lawrence B; Siva, Amara C; Kickhoefer, Valerie A (March 20, 2000). "RNA location and modeling of a WD40 repeat domain within the vault". RNA. 6 (6): 890–900. doi:10.1017/s1355838200000157.

[33] Rome, Leonard. "Vaults. Novel nano particles". http://www.vaults.arc.ucla.edu. Computing Technologies Research Lab. External link in |website= (help).

[34] Berger W, Steiner E, Grusch M, Elbling L, Micksche M (January 2009). "Vaults and the major vault protein: novel roles in signal pathway regulation and immunity". Cellular and Molecular Life Sciences. 66 (1): 43–61. PMID 18759128. doi:10.1007/s00018-008-8364-z.

[35] Heng Li,1,† Bob Handsaker,2,† Alec Wysoker,2 Tim Fennell,2 Jue Ruan,3 Nils Homer,4 Gabor Marth,5 Goncalo Abecasis,6 Richard Durbin,1,* and 1000 Genome Project Data Processing Subgroup7 “The Sequence Alignment/Map format and SAMtools” Bioinformatics. 2009 Aug 15; 25(16): 2078–2079. Published online 2009 Jun 8. doi: 10.1093/bioinformatics/btp352

[36] Patrik D'haeseleer "What are DNA sequence motifs?" Nature Biotechnology 24, 423 - 425 (2006) doi:10.1038/nbt0406-423

[37] Bronwen L. Aken, Sarah Ayling, Daniel Barrell1, Laura Clarke, Valery Curwen, Susan Fairley, Julio Fernandez Banet, Konstantinos Billis, Carlos Garcín Girón, Thibaut Hourlier, Kevin Howe, Andreas Kähäri, Felix Kokocinski, Fergal J. Martin, Daniel N. Murphy, Rishi Nag, Magali Ruffier, Michael Schuster, Y. Amy Tang, Jan-Hinnerk Vogel, Simon White, Amonida Zadissa, Paul Flicek and Stephen M. J. Searle The Ensembl gene annotation system Database 2016, baw093 doi: 10.1093/database/baw093

[38] Jerzy K. Kulski "Next-Generation Sequencing — An Overview of the History, Tools, and "Omic" Applications" DOI: 10.5772/61964

[39] Herve Pages, Valerie Obenchain, Martin Morgan GenomicAlignments. R package version 3.4.1.

[40] Zuker, M. and Stiegler, P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information Nucleic Acid Res. 9(1): 133-148, 1981

[41] R. Lorenz, S.H. Bernhart, C. Hoener zu Siederdissen, H. Tafer, C. Flamm, P.F. Stadler and I.L. Hofacker (2011), "ViennaRNA Package 2.0", Algorithms for Molecular Biology: 6:26

[42] I.L. Hofacker, W. Fontana, P.F. Stadler, S. Bonhoeffer, M. Tacker, P. Schuster (1994), "Fast Folding and Comparison of RNA Secondary Structures", Monatshefte f. Chemie: 125, pp 167-188

[43] B.A. Shapiro (1988), "An algorithm for comparing multiple RNA secondary structures" CABIOS: 4, pp 381-393

[44] B.A. Shapiro, K. Zhang (1990), "Comparing multiple RNA secondary structures using tree comparison", CABIOS: 6, pp 309-318

[45] W. Fontana, D.A.M. Konings, P.F. Stadler and P. Schuster P (1993), "Statistics of RNA secondary structures", Biopolymers: 33, pp 1389-1404

[46] Everitt, Brian (1998). Dictionary of Statistics. Cambridge, UK: Cambridge University Press. p. 96. ISBN 0-521-59346-8.Galili, T (2015).“dendextend: an R package for visualizing,adjusting and comparing trees of hierarchical clustering”. Bioinformatics. 31: 3718 3720. PMC 4817050 .PMID 26209431. doi 10.1093/bioinformatics btv428
[47] Stadler PF1, Chen JJ, Hackermüller J, Hoffmann S, Horn F, Khaitovich P, Kretzschmar AK, Mosig A, Prohaska SJ, Qi X, Schutt K, Ullmann K. “Evolution of vault RNAs” Mol Biol Evol. 2009 Sep;26(9) 1975-91. doi: 10.1093/molbev/msp112. Epub 2009 Jun 2.
[48] Jason A. Reuter, Damek Spacek, and Michael P. Snyder ,”High Throughput Sequencing Technologies” Mol Cell.2015 May 21;58(4): 586-597. doi: 10.1016/j.molcel.2015.05.004
[49] Ivo L. Hofacker “Vienna RNA secondary structure server” Nucleic Acids Res. 2003 Jul 1; 31(13): 3429–3431.

7. Figures:

Fig.1 Tanaka H, Kato K, Yamashita E, Sumizawa T, Zhou Y, Yao M, Iwasaki K, Yoshimura M, Tsukihara T (January 2009). "The structure of rat liver vault at 3.5 angstrom resolution". Science. 323 (5912): 384–8. PMID 19150846. doi:10.1126/science.1164975

Fig.2 Rfam 12.0: updates to the RNA families database. Eric P. Nawrocki, Sarah W. Burge, Alex Bateman, Jennifer Daub, Ruth Y. Eberhardt, Sean R. Eddy, Evan W. Floden, Paul P. Gardner, Thomas A. Jones, John Tate and Robert D. Finn

Fig.7 Lorenz, Ronny and Bernhart, Stephan H. and Höner zu Siederdissen, Christian and Tafer, Hakim and Flamm, Christoph and Stadler, Peter F. and Hofacker, Ivo L. ViennaRNA Package 2.0 Algorithms for Molecular Biology, 6:1 26, 2011, doi: 10.1186/1748-7188-6-26

Fig.9 Stadler PF1, Chen JJ, Hackermüller J, Hoffmann S, Horn F, Khaitovich P, Kretzschmar AK, Mosig A, Prohaska SJ, Qi X, Schutt K, Ullmann K. “Evolution of vault RNAs” Mol Biol Evol. 2009 Sep;26(9): 1975-91. doi: 10.1093/molbev/msp112. Epub 2009 Jun 2.

Fig.10.11.12.18.19.20 Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Fig.13.14.15.16 Helga Thorvaldsdóttir, James T. Robinson, Jill P. Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics 14, 178-192 (2013).

Fig.17. The EMBL-EBI bioinformatics web and programmatic tools framework. (2015 July 01) Nucleic acids research 43 (W1) :W580-4 PMID: 25845596

Fig.21.22.23 R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Fig.24 Rfam 12.0: updates to the RNA families database. Eric P. Nawrocki, Sarah W. Burge, Alex Bateman, Jennifer Daub, Ruth Y. Eberhardt, Sean R. Eddy, Evan W. Floden, Paul P. Gardner, Thomas A. Jones, John Tate and Robert D. Finn

Table 1.Bronwen L. Aken, Sarah Ayling, Daniel Barrell1, Laura Clarke, Valery Curwen, Susan Fairley, Julio Fernandez Banet, Konstantinos Billis, Carlos Garcín Girón, Thibaut Hourlier, Kevin Howe, Andreas Kähäri, Felix Kokocinski, Fergal J. Martin, Daniel N. Murphy, Rishi Nag, Magali Ruffier, Michael Schuster, Y. Amy Tang, Jan-Hinnerk Vogel, Simon White, Amonida Zadissa, Paul Flicek and Stephen M. J. Searle The Ensembl gene annotation system Database 2016, baw093 doi: 10.1093/database/baw093