Development of a computational workflow to design a circular microRNA sponge in the context of RNA therapy, with a focus on melanoma cancer

Under the direction of :

Dr. Shailendra Gupta,

Pr. Olaf Wolkenhauer,

SBI Rostock

Pr. Yannick Andéol,

UPMC

Author,

Etienne Rolland

Abstract

This document is a concise presentation of the work accomplished during my master's degree internship at the Rostock SBI department. The project was to develop a computational workflow which design a circular RNA, in order to sponge microRNAs of interest.

The core material was the triplexRNA database, a database of cooperative microRNAs.

After a quick introduction, we will see first how the design script proceed to build the circular RNA. After this, we will detail of the first workflow and its results, as well as the second workflow. Finally, the conclusion will discuss the limits of the first and second workflow.

Chapter 1

Introduction

MicroRNAs

MicroRNAs are 20-24 nt RNA molecules, which can form a complex with AGO family proteins members and other ribonucleoprotein , to exert a post-transcriptional repression [1,2,3,4]. This complex is design as microRNA-induced silencing complex (miRISC) [4].

This repression is exert by a inhibition of the translation's initiation, and eventually, by a messenger RNA destabilization [2,4].

The repression is mediated by the binding of the complex miRISC to the target messenger RNA, the complementarity between the microRNA sequence and the messenger RNA being the vector of recognition of the target and of the specificity of the interaction

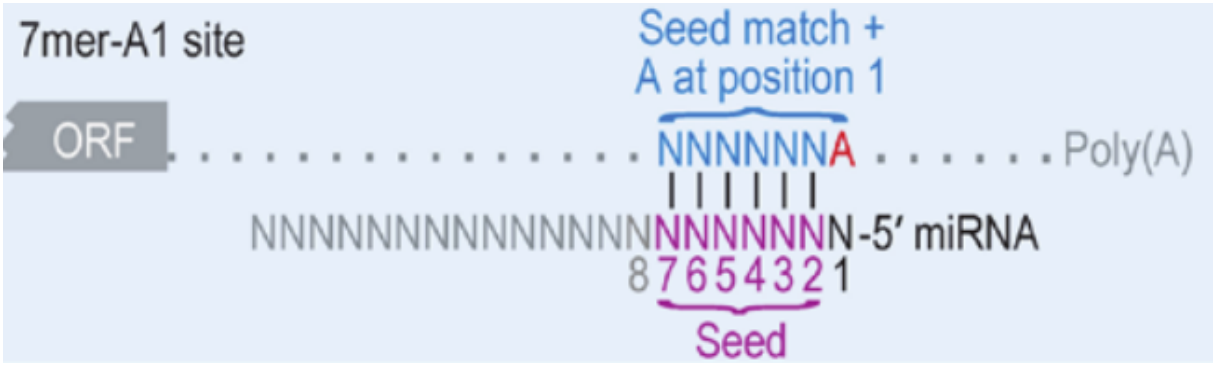


Figure 1.1: Schematic representation of the binding of a microRNA on a messenger RNA, for a 7mer site. Illustration taken from [3]

In that sequence, the most important is the "seed region" (see the figure above for a schematic representation), a complementarity in this region beeing enought to for the complex to bind the messenger RNA, and most of the time, mandatory. This seed region is located in the 5' region of the microRNA, and could be a 6, 7 or 8mer [2, 3]. By extension, the seed region is also in the 3' region of the microRNA binding site.

Nonetheless, despite this evidences, the microRNAs' network is still poorly understood [1, 4]. The reason includes : the functional redundancy of the microRNAs, the lack of clear phenotype when one microRNA is knock-down, the combinational of the repression, etc [1, 4]. Their levels of expression, like many non coding RNA, is altered in disease, including cancer [4].

Inhibition and level of mRNA

While the description of the mechanism of the translational inhibition above suggest that the levels of microRNA-inhibited mRNAs remain unchanged,

more recent work has demonstrated that the repression of many miRNA targets is frequently associated with their destabilization [4].

Micro-array studies of transcript levels in cells and tissues in which microRNA levels were experimentally altered, revealed marked changes in the abundance of validated or predicted miRNA targets, consistent with an role for miRNAs in mRNA destabilization [4].

MicroRNAs : a paradigm based on cooperativity

There is experimental evidences for a microRNA cooperativity in the post-transcriptional repression they exert, when their seeds region on the messenger RNA are in close proximity $13-35$ nt [2, 5].

Since this could be an effective way to understand how the microRNAs' network exert its repression, a model based on this cooperativity has been tested in sillico. The model predicted the expression level of p21 in 9 of the 12 tissues considered for the simulation [6].

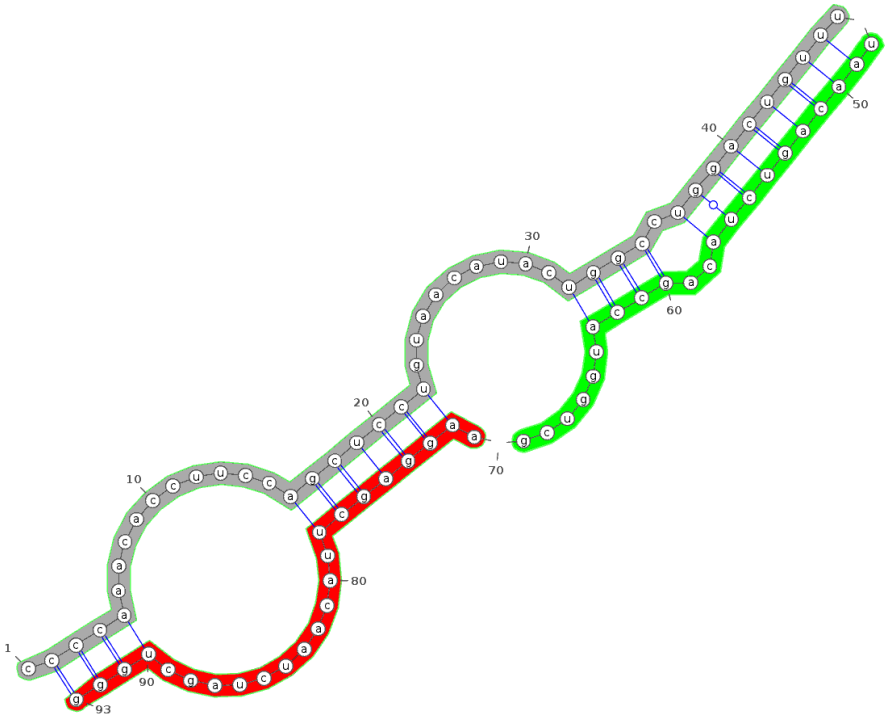


Figure 1.2: An example of canonical triplex, from the triplexRNA database

The TriplexRNA database is based on this work. Based on a computational workflow to predict them [7], the database provide putative triplexes, triplexes composed by two microRNAs and their mutual messanger RNA target (three RNA molecules for the triplex, nonobsting the AGO proteins). This report will sometimes refer to this database as the "triplex", for convenience purposes.

Circular RNAs

Mostly formed by backsplicing [8, 9], the circular RNA are a class of long non coding RNA. While the exact nature of their function remains elusive [8, 9, 10], it has been proved that one of

them is to ”sponge” the microRNAs [9, 10].
By binding the microRNAs, a circular sponge decoy the microRNAs from their original targets. Since this targets are messenger RNAs, and that the microRNAs exert a repression, the sponge influence the level of expression of the genes regulated by this microRNA, and induce an up-regulation of their traduction [9, 10, 11].

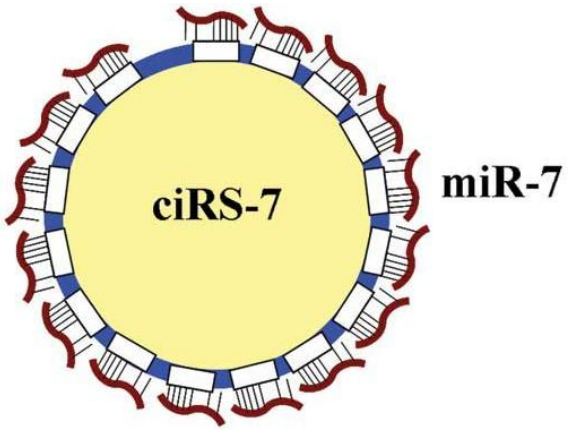


Figure 1.3: Illustration of the ”sponging” capability of the circular RNAs. Illustration taken from [9]

This ability to up-regulate the expression of genes, coupled with a long half-life due to their circular shape [9], make them a good candidate as a therapy tools.

That could be the case for disease involving a deregulation of microRNAs' level of expression, like in cancer as stated before [4], and their phenotypic consequences in various disease [12], or in case the enhancing of traduction of one gene would be suitable.

Some circular RNAs and their absence have also been report to be directly involved in diseases. For example, a circular RNA which sponge miR-9, circMTO1, has been found to be highly correlated with the prognosis of patient suffering from human hepatocellular carcinoma (HCC) [11]. Moreover, the circular RNA could suppress HCC progression in vivo [11], leading us to think that circular RNAs could be a good therapy tool, once bypassed the usual problems of addressing and delivering of this molecules, common to every RNA therapy [12].

Context of the work

The first objective was to develop an executable which design a circular RNA sponge, to up-regulate the level of expression of a gene(s) of interest, using the information contained inside the triplexRNA database. This executable could be then integrated into further release of the

triplexRNA. This part was done with success, and this concise report will first show how the executable which design the circular RNA works. The executable which can be found here. The procedure to chooses which microRNAs should be sponge is detailed later in the report.

A secondary focus was made on the melanoma cancer. The main motivation behind this was to extend the previous workflow to upregulate genes of interest, to a workflow to upregulate a pathways of interest. Indeed, the triplexRNA can be query for a list of genes and for pathways (the triplex integrate most of the KEGG pathways), including the KEGG melanoma pathways.

Why differential expression analysis

The other problem is that KEGGs pathways, the way that they are incoded inside the database, are just a list of genes, not telling if ... The idea was to find genes differentially expressed between early and late stages of the melanoma cancer, using the Xena database [13] to obtain the level of expression of the genes composing the KEGG Melanoma pathway.

The choice of the Melanoma cancer have been done because of the potential experimental outcome. For example, the circular RNA sponge could have been tested in cells culture : the experimental design would have been to test if the artificial introduction of the circular RNA would have induced a phenotype reversion of the late stage phenotype toward an early stage phenotype.

The whole workflow and controls are detailed below, and the limits of such a demarch explored inside the conclusion.

Why differential expression analysis

The other problem is that KEGGs pathways, the way that they are incoded inside the database, are just a list of genes, not telling if ...

Chapter 2

Results

Circular RNA design script

The circular RNA design script is write in Python, is executed in command line, and take as argument, not genes names, but the microRNAs' names that the sponge should decoy from theirs original target. The procedure to choose which microRNAs to sponge is detailed later.

More precisly, the circular RNA design script have to mandatory input :

- a file contening the names of the microRNAs that the user want the sponge to decoy from their original target. The names have to be the same than the one in the triplexRNA database. This file is a txt file with just the name of the microRNAs, one per line.
- a file containing a priorities for each microRNAs. Basically, the script build the sponge in a recursive way, and add as many binding site as the priority number allocated to each microRNA on each recursion. For example, for a sponge with twice the number of binding site for one microRNAs compare to the other, the file will contain the integer 2 and 1. This file is a txt file with just one integer per line, and have to be in the same order than the microRNAs they refer to.

The script will then query the triplexRNA database, using the provided names to get the mimat IDs which identify this microRNA in the database. This ID is then used to query Mirbase, to get the sequence of the microRNA.

There is several reasons to this. First of all, while the sequences of the microRNAs are encode inside the triplexRNA database, the orientation of the strand is not indicate (5' and 3' extremity), and this step is then a security control, assuring the sequence of the microRNA is in the right orientation.

Also, the name of some microRNAs can be different between Mirbase and the triplexRNA, depending on their respective update. For example, most of the microRNAs have the mention 5p or 3p in their name, indicating from which part of the hairpin they come from, a mention sometimes absent from the name inside the triplexRNA database. This step is then a disambiguation phase.

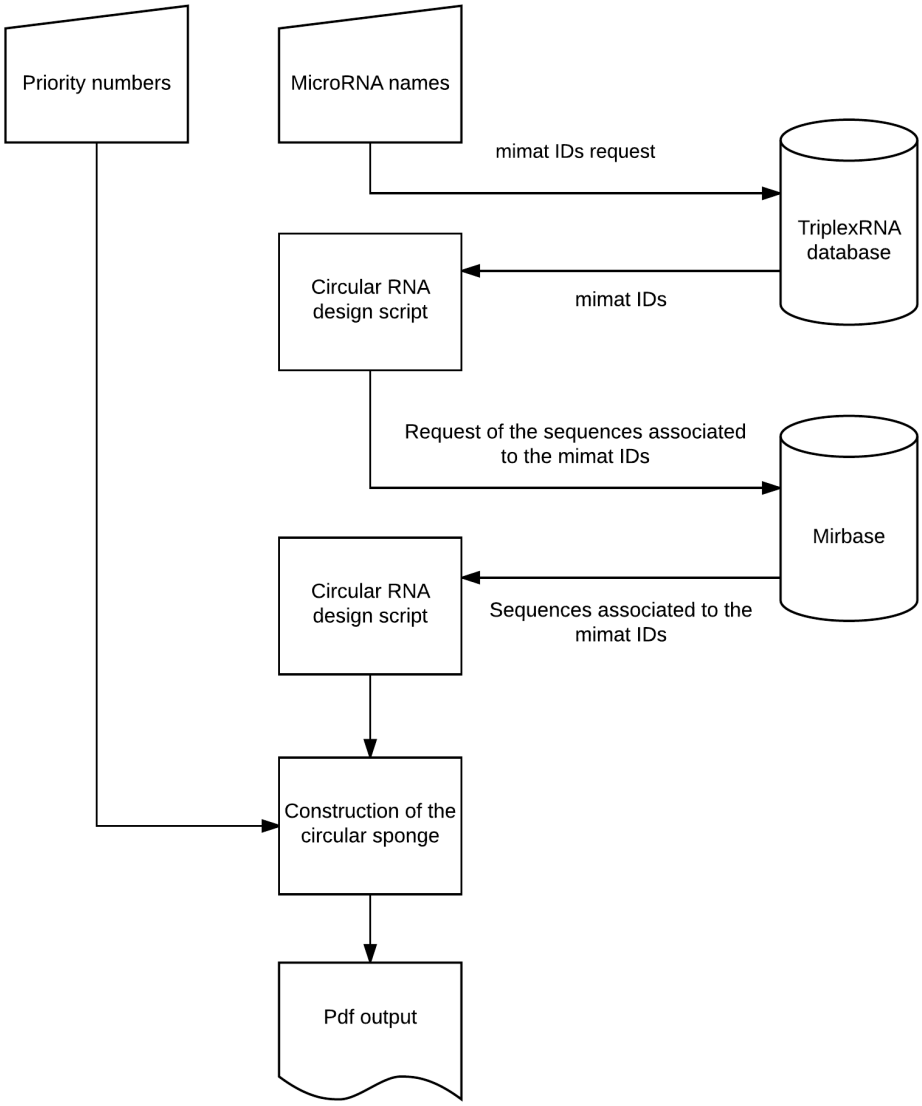


Figure 2.1: Details of the request done by the circular RNA design script.

Once the sequences of the microRNAs are obtain by the circular design script, it will use them to build the most efficient binding site possible for each of the microRNAs. Those building sites, one per microRNAs, will be then repeat on the circular sponge, separated by nucleotides according to another parameter, the distance of cooperativity [5].

In order to do so, the design script is implemented with rules extract from the scientific literature :

- a A is systematically put in first position (in 5') of the microRNA's binding site. There is an overrepresentation of conserved adenosines flanking the seed complementary sites [14], among other data suggesting that a site with a A in first position outperform the

others one, including the one with a complementary nucleotide to the microRNAs [3].

- the seed region (from the 2nd to the 8th nt, cf introduction) is build by inserting complementary nucleotides to the microRNAs' ones, forming a "8-mer" site. See [2, 3] for further details and representation.

- a A is put in the ninth position of the binding site, because there is an overrepresentation of conserved adenosines flanking the seed complementary sites [14], and because the microRNAs binding sites which perform the best are the ones with a local hight content of Adenine (A) and Uracile (U), over any other variable or predicator [3].

- the 12th from 10th nucleotides are not complementary to the microRNAs', and U and A if possible. The aim is to avoid any Ago2 mediated cleavage of the circular RNA, cleavage which occur in case of near perfect complementarity between the microRNA sequence and the binding site [4]. This nucleotide are U, or A if not possible.

- the nucleotides from 13 to 16 (included, so four in total) are complementary to the - microRNAs' one. The reason is that Watson-Crick pairing to four contiguous nucleotides produce an additionnal pairing in 3' [2] and that most downregulation is associated to this additionnal pairing if it start at the 13th position [2], additionnal downregulation which can be translate as a stronger binding of the microRNA on the binding site [2].

As we just mentionned, the microRNAs binding sites which perform the best are the ones with a local hight content of A and U, over any other variable or predicator [3].
This is probably correlated to binding site accessibility, and a low potential of forming a self-interacting secondary structure [3].
To take advantage of this two features, in order to improve the efficiency of the circular RNA sponge, the script design the sponge's sequence with as many U and A as possible.
That means that, outside the seed region, where a strict complementarity is required, or in the additional 3' pairing region (cf above), the sequence is write with U or A.
A cognate problem is that some proteins target and bind some AU rich sequences, known as AREs [15]. However, since this AREs sequences contain some A [15], the safest option is to specify a U prior to an A when writing the sponge's sequence, when no other specific nucleotides is required for complementarity between the sponge and the microRNA.

Each binding site is terminated by an additional number of nucleotides, in order to separate its seed from the seed of the following binding site by the distance of cooperativity, as define

in [5]. The distance of cooperativity is the distance between the two seed, distance which insure that the two microRNAs will coopere to bind and exert their translational repression on a messenger RNA [5].
The distance of cooperativity is set by default to 17, but can be specified by an additional parameters when running the script.

The binding sites for one microRNA are organized in cluster by the design script, considering the observation that Ago2 shuttles between adjacent target and then that neighboring sites could cooperate to retain the Ago2-miRNA complex. [16].

During the recursive call, the design script build the clusters by adding as many binding site to the cluster as the priority number allocated to this microRNA. The recursive call is interrupt when the limit size (set by default to 300, but can be specified), is about to be exceeded by the addition of a new binding site.

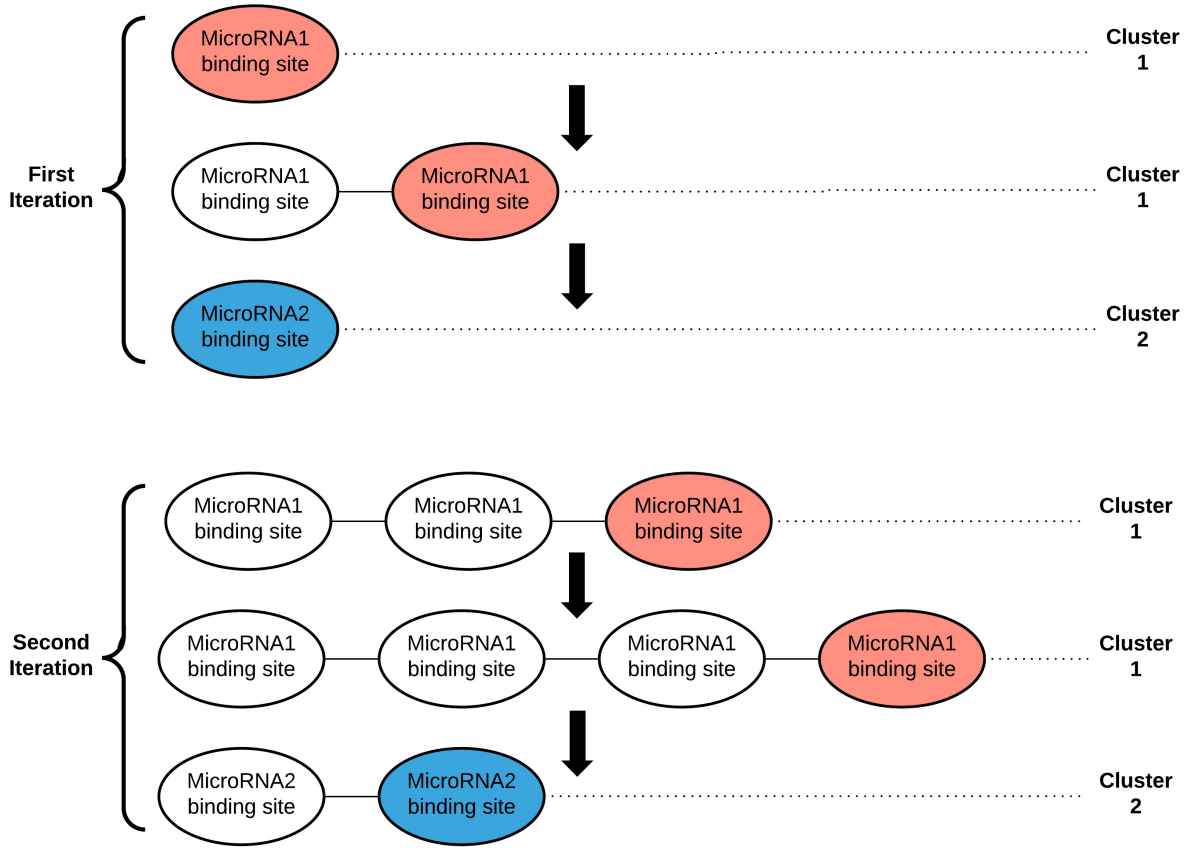


Figure 2.2: Schema of the circular RNA's construction. During the recursive call, the function add on each cluster of microRNA' binding site, as many binding site as the priority number for this microRNA, until the limit size is about to be exceeded. Please take note that the addition of the binding sites is done in a iterative way.

After this, the design script compute the minimum free energy of the circular RNA for every arrangement of clusters of binding sites in the sequence, and select the sequence of the circular RNA with the lowest minimum free energy for the output.

Finaly, a pdf report with the sequence of the circular sponge is output by the design script. This pdf report details and justifies every step of the design to the user (organization of the binding site in cluster, the design rich in U, etc).
This pdf report also provide quality controls in the form of miranda alignments. In this section are exposed the 3 best alignments (if they exists) for each cluster against every human microRNAs. The script is provide with the file mature.fa, a fasta format sequences of all mature miRNA sequences, download from miRbase in may 2017.

Each alignments have for title the name of the microRNA performing the alignment, and the name of the microRNA for which the cluster has been design.

2.2.5 Mir_name : hsa-miR-21-5p, cluster for : hsa-miR-21

Score : of the alignement :152.000000

```
miRNA:     3′ aguuguAGUCAGACUAUUCGAu 5′
                      |||| :  |||||||
CircRNA:   5′ auuuuuUCAGAUAAAUAAGCUa 3′
```

Figure 2.3: Example of one of the alignment found in the pdf report.

This allow to check easily if the best alignment is performed by the microRNAs for which the binding site was created, and make sure that the binding site will not be bound by another microRNAs. Or even worse, to ensure that no microRNA is susceptible to trigger a Ago2 mediated cleavage of the sponge.

In such case, an override procedure is implement. Providing a txt file with a sequence, the user can redo the whole design process with a binding sequence specified for one or more microRNAs. The construction of the circular sponge is then redo in the same recursive way, except for the creation of the specified binding site, and then the quality control are reperformed and a pdf report is recreated.

Since all the quality control are performed again, this allow the user to refine the construction of the sponge, especially in the case of one miss design, as describe above.

Most of the case of competition by another microRNA are due to a U:G wobble, an allowance between a G in the microRNA's sequence and a U in the sequence of the circular RNA (there is a good example in the figure 2.3), the probability of such an event beeing incredibly improved by the design rich in U of the sponge.

In that case the pdf output advise to first replace this problematic U by a A, if possible, to keep the low probability of the binding site to form a self secondary structure.

First workflow

Af we stated before, our goal was to investigate the design of a workflow to upregulate a biological pathways of interest with a circular RNA sponge.

Our concern is that KEGG pathways are only implemented as a list of genes involved in disease, inside the triplexRNA database [7], without any further indication on their involvement. We don't know from the base if this genes are down-regulated, up-regulated, or inhibited in the biological conditions of interest.

A solution is to perform a differential expression analysis on this subset of genes provide by the triplexRNA, to find which genes are lowly expressed in the disease, and potentially responsible of its phenotype.

As we stated in the introduction, the repression exert by microRNA can result in a destabilization of the mRNA target and a change in the level of expression, but this is does not now necessary occur, and microRNA-inhibited mRNAs can have their level unchanged.

Here, we do not presume that the change in level of expression of mRNA is caused by the change in the expression of microRNAs, as it occur in cancer [4], even if it's a possibility.

We are looking for genes which are potentially responsible for the phenotype. This genes would be then good candidatures to be upregulated or having their expression improved using a circular sponge.

A transversal and complementary approach to the differential expression analysis is a feature selection, operated on the genes.

The feature selection is a field of the machine learning domain [17], dedicated to select feature among a set of data, in order to improve the generalization of models, or get a better understanding of the features and their relationship to the response variables [17]. It is this last aspect which we interest us.

To check the accuracy of such selection, a random forest classifier [17] has been trained with this set of genes, and the error of classification has been plot.

This approach is still a bit naive, especially regarding the complexity and heterogeneity of cancers [18, 19], and the uncertainty behind the causes of the differential expression. For this reason, two control will be perform, one for this two approach.

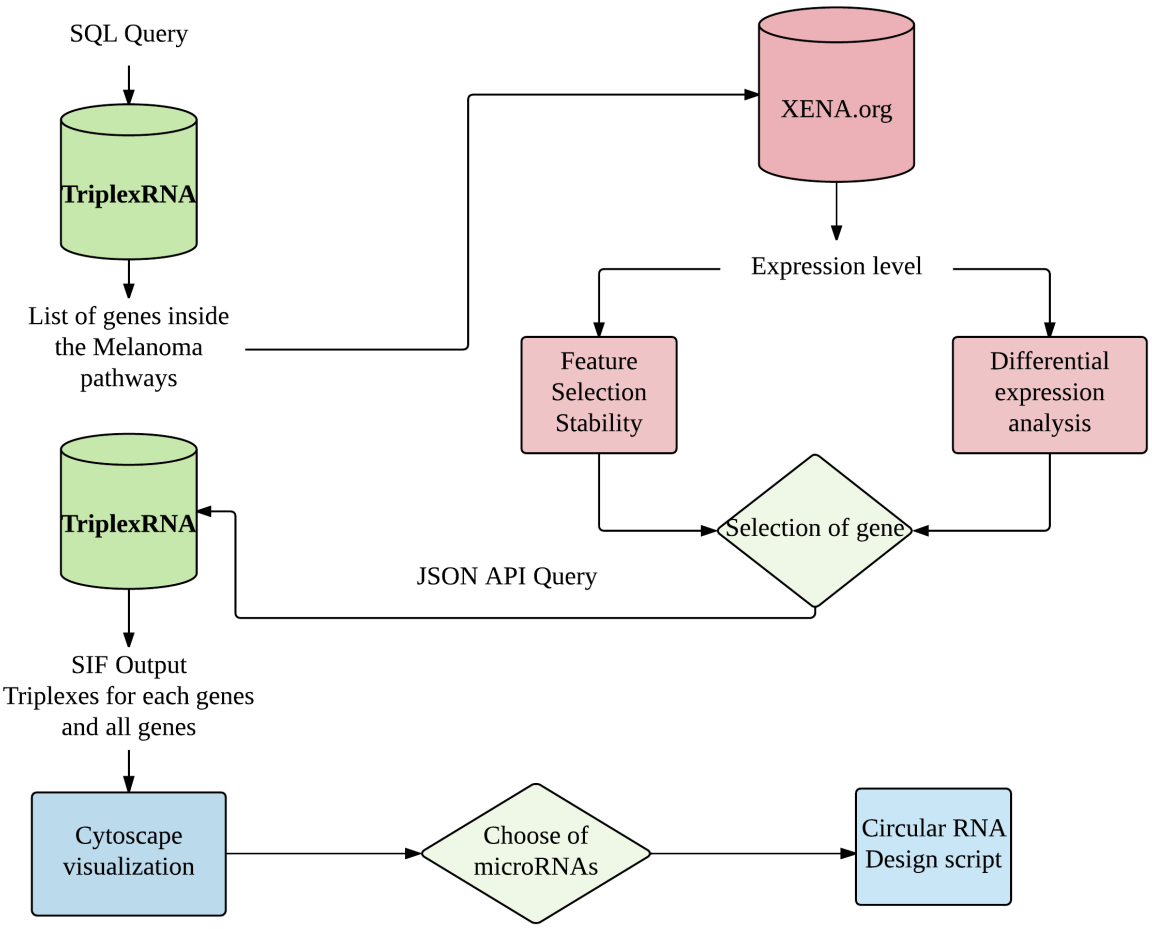
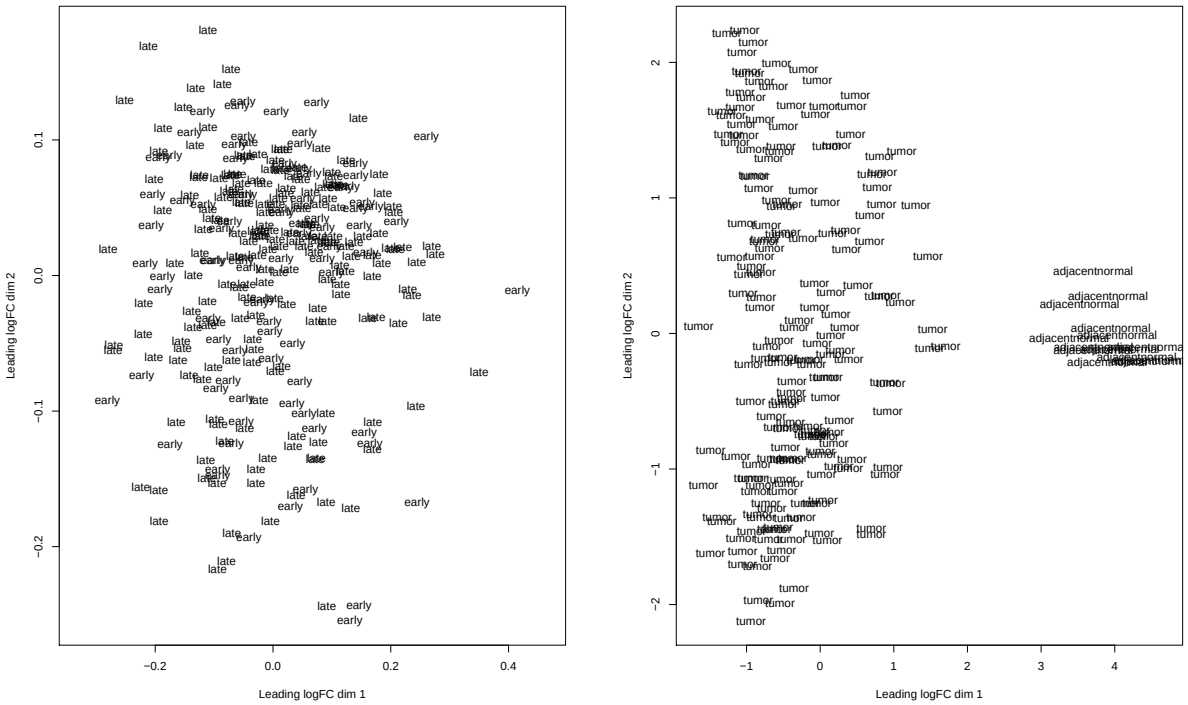


Figure 2.4: The structure of the first workflow. The retrieving of the genes composing the Melanoma KEGG pathways using a SQL query is the first step.

Differential expression analysis



(a) Scatterplot where the distance between sample is a projection of the log2 fold change distance, for Melanoma cancer data, between late melanoma and early melanoma.

(b) Scatterplot where the distance between sample is a projection of the log2 fold change distance, for Colorectal cancer data, between normal adjacent and cancer cells.

Figure 2.5: Comparison of the projection fold change distance between melanoma sample and colorectal cancer after a differential expression analysis. The scatterplot of the colorectal cancer show the expected output for this analysis, with a good separation of the two groups.

Feature selection using randomized lasso

The feature selection method employed was the Stability Selection [20], a method based on numerous subsampling of the data set - here, different samples - with different subsets of feature - here, the genes -, in combination with a method of feature selection.

This different subsampling are then aggregated, and a ranking is produced using the features selection method retain. This Stability approach perform well in dataset with a lot of correlation in the data [20], assuming that the model is sparse - meaning that only a few genes are responsible for the difference between the two conditions-.

This framework of stability selection has already been used in biology [21], and is imple-

mented for the lasso method inside the scikit learn library [22] under the name of Randomized-Lasso, a method which has already been deployed to identify cancer driver genes [23].

To test this selection of gene, a random forest classifier [17] has been trained using only this selection of genes as variable, and the error of classification has been plot (see below).

The reason for choosing a random forest classifier is its reliability and the fact that they require less tuning in comparison to other classification method [17]. This make them ideal to test a selection of genes in a repetitive manner inside a workflow.

Nonetheless, the plotting of the classification error, for the whole melanoma data, but also for subtypes [18] (NRAS, BRAF), did not allow to think that the genes selection was pertinent (cf figure).

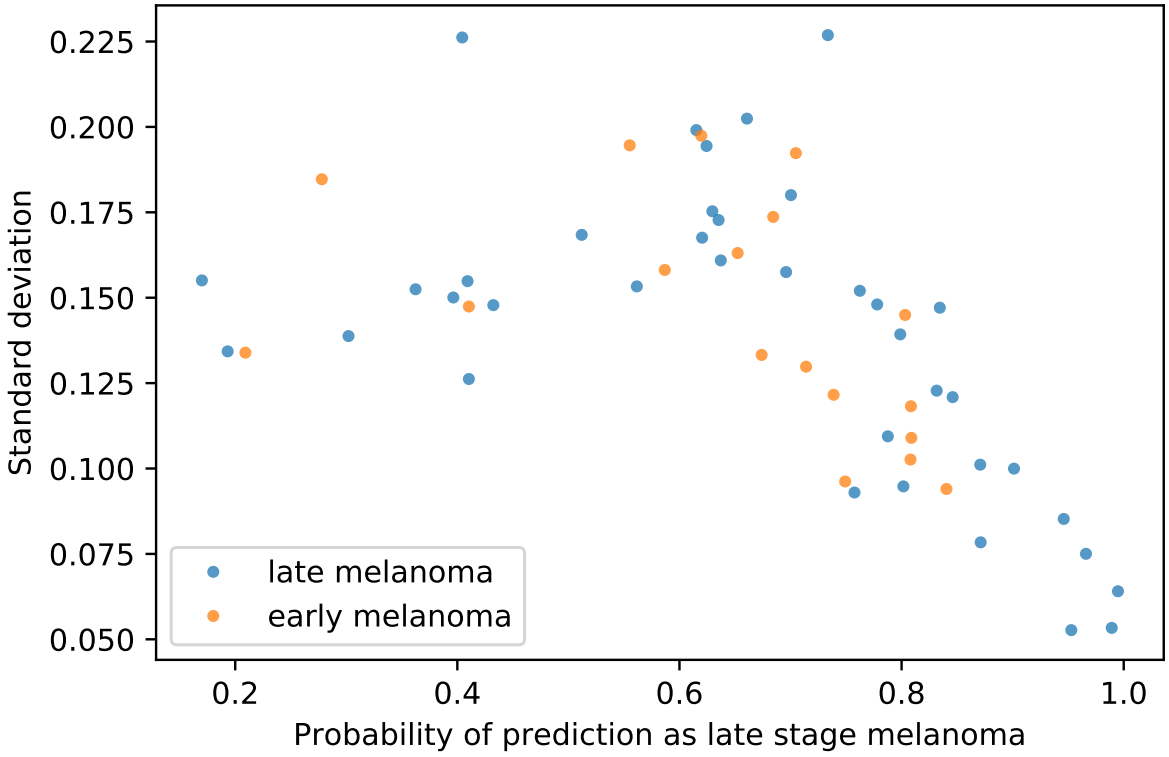


Figure 2.6: Error of classification of the random forest classifier, using a selection of genes inside the KEGG melanoma pathways, selection output by the feature selection approach

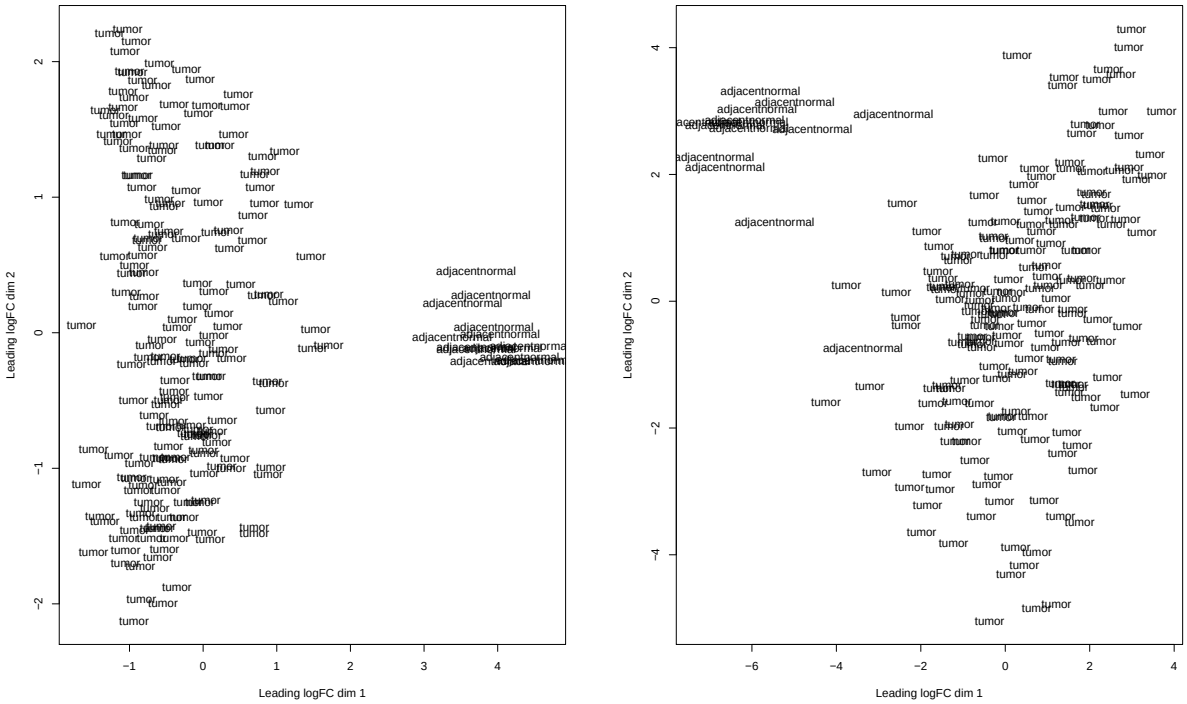
Current workflow

To illustrate the current workflow, we will create a sponge in the context of a RNA therapy for colorectal cancer, using some of the genes differentially expressed between the normal condition and the tumor condition. Of course, while this reduced set of genes is sufficient to distinguish between the two condition, they are not necessary responsible of the difference in the

phenotype. Such a diagnostic would require more investigation, using ontologies and literature. Here we take select them for illustration purpose.

A normal use of the workflow rely on the expertise of the user and that the user provide a set of downregulated genes he is interested in.

Diffential expression analysis, Colorectal Cancer



(a) Projection of the fold change between normal cells and tumor cells for colorectal cancer.

(b) Projection of the fold change between normal cells and tumor cells for colorectal cancer, using only the top 12 of the differentially expressed genes.

Figure 2.7: Multidimensional scaling plot of distances, using the plotMDS() function from edgeR. In the context of using a reduced number of genes to distinguish between the two conditions (normal adjacent cells and cancerous cells), we observe a good separation with the top 12 of the differentially expressed genes.

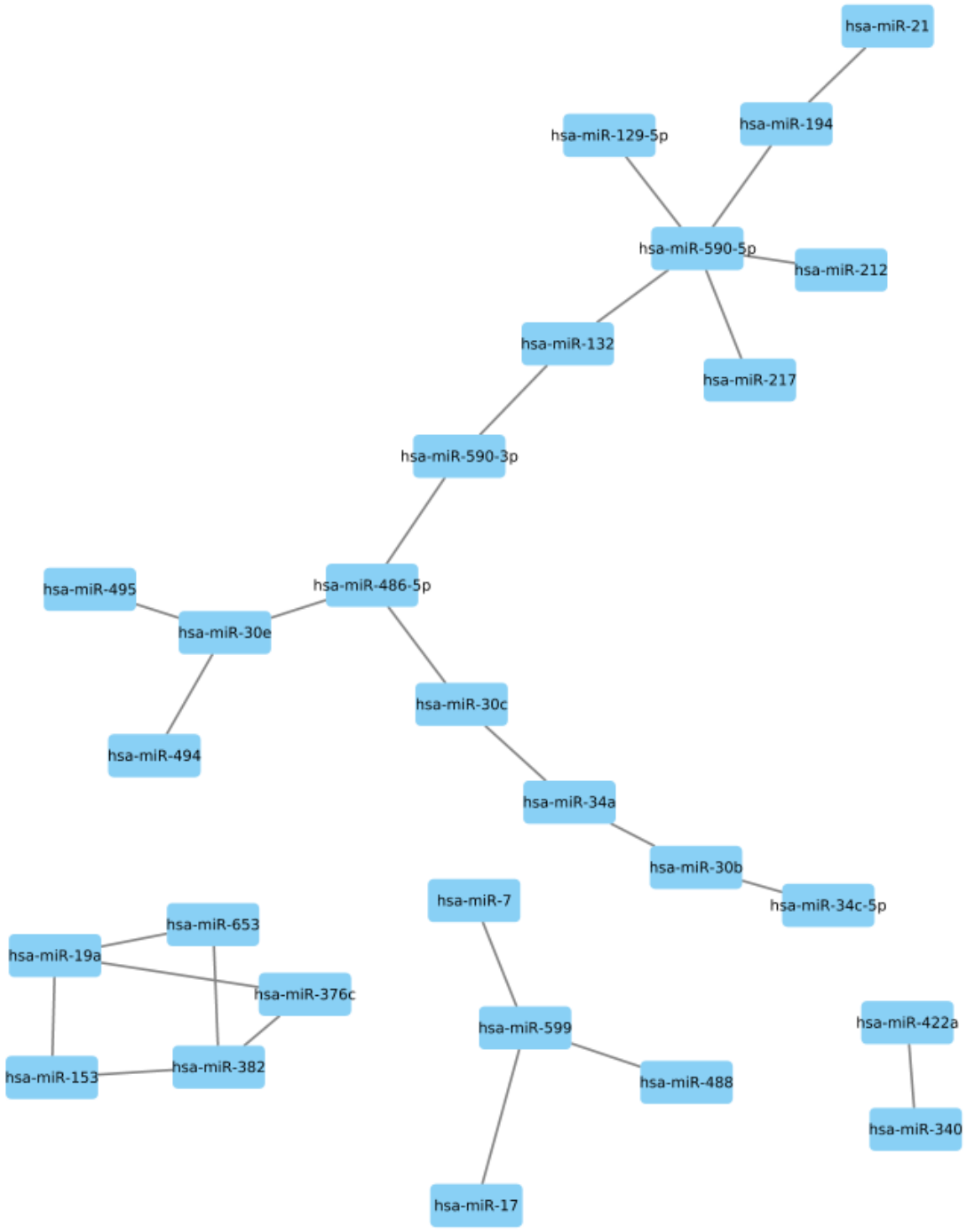


Figure 2.8: Here the cytoscape visualization for 3 downregulated gene in colorectal cancer, compare to normal adjacent cells BEST4 (4 triplexes), LGI1 (20 triplexes), KRT24 (14 triplexes). Here hsa-miR-590-5p is a potent target for the circular RNA design script

Chapter 3

Conclusion and Perspectives

Outcomes

Circular design script

The major outcome to this are the two executables which can be found here.

While one is only a local API to query a local instance of the triplexRNA, the other one can be used by anyone and independently from the workflow.

The circular design script can be used to design a circular microRNA sponges, get the quality control and refine the design with the ”override” procedure, whether or not the previous workflow has been used to retrieve the microRNAs to be targeted.

While this script can be used separately from the triplexRNA, it’s require that the user use the same names as the ones inside the triplexRNA database.

Workflow

The workflow retrieve all the triplexes involved in the repression of a list of genes passed as input, using the triplexRNA database [7]

This workflow provide a output file for a graphical visualization of the cooperation between

microRNAs which exert a repression in a cooperative manner on the level of expression of the genes passed as input.

That way the workflow provide a tool to visualize which microRNAs are involved in the most pairs of cooperative microRNAs and whose that the sponging by a circular RNA will affect the most pairs and remove the most repression on the list of genes passed as input.

The workflow finally provide a design script to automatize the construction of a circular RNA to decoy particular microRNAs, outputting a pdf report justifying the key points of the design and providing qualities control.

The design script allow the user to specify parameters of his own interest : the list of microRNAs to be decoy, a list of priority which is an intuitive way to prioritize the sponging of some microRNAs, and a maximum size according to the user synthesis facilities. The size distance between the seed can be specified to adapt to new scientific evidence, or particular wishes of the user.

Finally, the script allow to refine the design of the circular RNA sponge, through a repetition of execution of the design script, with some of or all binding sequence for microRNAs specified, instead of leaving the construction to the the automatic design, coupled with the repetition of the qualities controls. This process is ease by having files as inputs for the design script.

The workflow rely on the triplexRNA database, allow the user to retrieve all the canonical triplexes for a list of genes, to visualize them in a way that allow to find the most important microRNAs using cytoscape, and finally to build a circular RNA to sponge them, and eventually to refine the process.

While there is some limits to this workflow (see perspectives below), ...

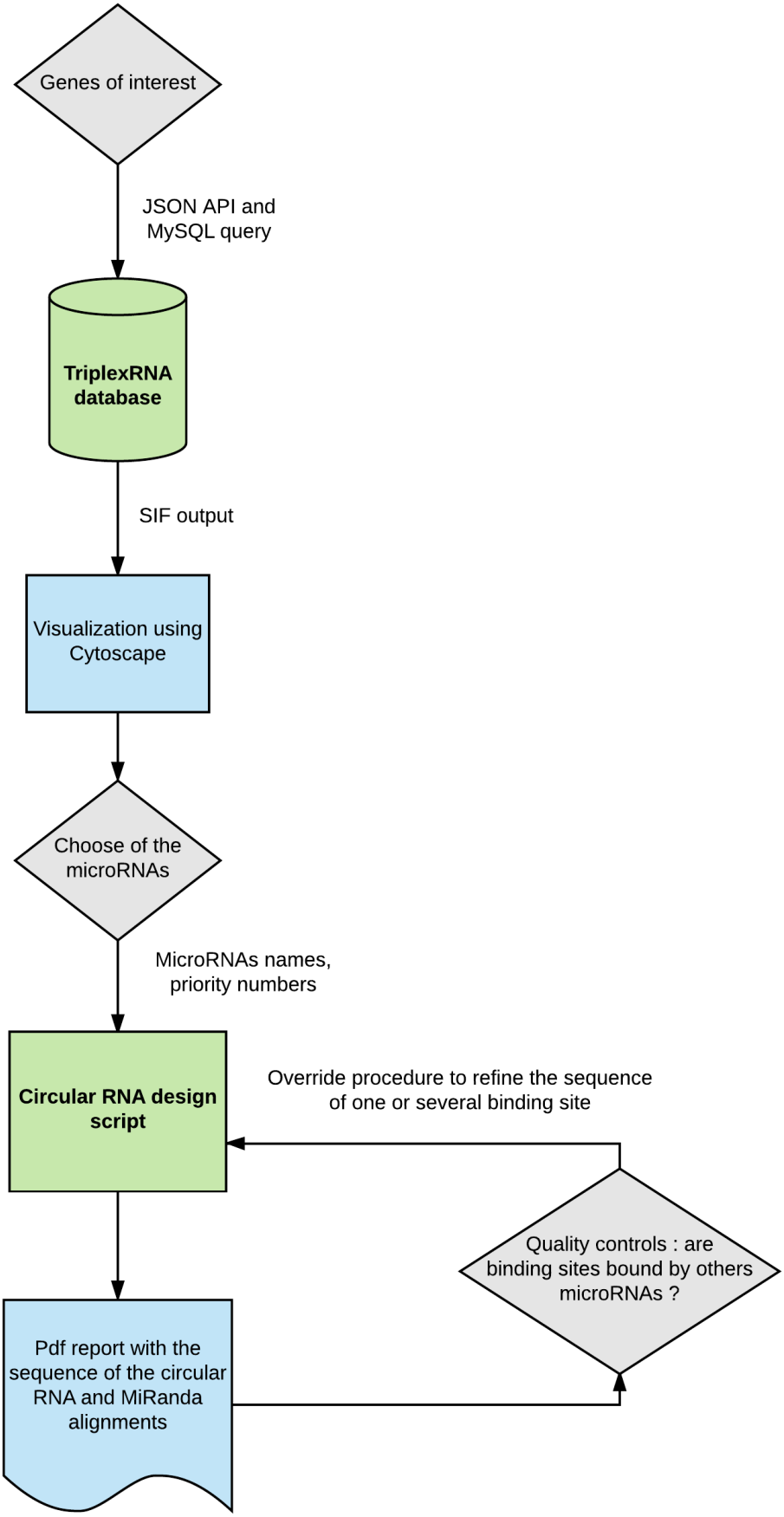


Figure 3.1: Schematic representation of the final workflow

Down-regulation and translationnal inhibition

A really important point is that a genes negatively regulated by the microRNAs, will not appear as differentially expressed in a RNA-seq : the mRNA will not be degraded, but the transduction of the mRNA will be inhibited.

A more pertinent approach would be to investigate directly the microRNAs level of expression, and cross the data with the triplexes referenced inside the tripleRNA database or, more directly, rely on the expertise of the user. Indeed the user can have - through western blot experiments for example -, genes of interest, not down-regulated(?) but for which the translation of the messenger RNA is inhibited.

KEGGs map

The problem of finding which genes are down-regulated/

Indeed, the KEGG pathways are only a list of genes inside the triplex RNA, which doesn't tell us which gene should be up-regulated/de-repressed by the circular RNA sponge. A similar information, however, is on the KEGGs map, where the effect of each genes/proteins are indicated.

Considering this, an other approach will be to used the KEGGs map to select genes of interest and look at the microRNAs.

This is a variation of a feature implemented in the triplexRNA database, which allow a visualization of the mutual target genes for selected microRNA pair, within the KEGG disease pathway.

But this feature is for two microRNAs only, and by then ignore the relatives importances of microRNAs among all the pairs, nor it can be done in the direction we would need (a selection of genes downregulated on the map towards all the microRNAs pairs).

Anyway, a collection of manually drawn map is not the best material to carry such a task, especially in the optics of long maintenance.

cirMOT1 in hepatocellular carcinoma

One thing we can wonder is : does the workflow retrieve Mir-9 as a microRNAs to be target with a circular RNA in the case of the carcinoma cancer ?

While there is little to criticize in the demarche exposed in the article (Mir-9 and circMOT1 are clearly identified as responsible for the down expression of p21 and for the inhibition of the repressive effect of Mir-9, respectively), MiR-9 is actually no part of the triplexes involving the CDKN1A messenger RNA. There is several reason which could explain this :

- Among the 99 microRNAs identified using miRanda as susceptible to bind cirMOT1, only 20 of them have been screened for specific enrichment after RNA precipitation in vivo, considering that they were previously describe as involved in the hepatocellular carcinoma.

  While this analytic procedure is entirely relevant in the case of a cancer study, right now I would deplore the lack of information about the 80 other ones. Also, while the case of the sponge ciRS-7 [10] is abundantly cited along the article [11], there is no reason to think that most of the natural sponge decoy only one microRNA. The supplementals data of Memczak et al, [10] are rich in predicted circular RNAs with a seed match for more than one microRNA family.

- MiR-9 is over expressed in the hepatocellular carcinoma, and this is why the researcher selected this microRNA, among other, for screening after the RNA precipitation. And in this situation of over expression, low affinity sites for the microRNA (6-nt sites) become occupied and functionnal [24].

  Nonetheless, the triplexRNA database has been construct based on binding site prediction extracted from microRNA.org, in which only binding sites with good conservation and prediction scores were retained [7], binding sites which are assumed to be the functionnal ones in normal circomstances.

  That make very likely that the workflow constructing the triplexRNA database does not take into account some microRNAsbinding sites which are functional uniquely in some diseases, and by that, it make possible that some triplexes, specifics to particular disease and associated with specific over expression of some microRNAs, are missing in the triplexRNA database.

Visualization of the triplexes

While Cytoscape provide a comfortable way to visualize the pairs of microRNAs which repress the expression of a set of genes, it can't be run in command line and then can't be part of a pipeline. The choice of cytoscape was done mainly for convenience purpose, and because the author of this document didn't find a Python library which would have allow him to fulfill the function of triplexes' visualization.

A prospective work for a core installation inside the triplexRNA of the executable which design a circular sponge, to enable the creation of the sponge from a web interface; would be the development of a visualization tools which can be run by a program, and which does not require for the user to run a graphical interface by itself.

Chapter 4

Material and Methods

TriplexRNA database

The TriplexRNA database provide putatives triplexes according to the workflow published here [7], and is available at the following url : https://triplexrna.org/

A triplex is a conformation defined as a messenger RNA and a pair of microRNAs bound on it, in a cooperative manner.

Two microRNAs are a putative pair of cooperating microRNAs if the seed region of their binding site are separated by a distance of 13-35 nt [5]

A triplex is supposed functional - meaning : the pair of cooperative microRNAs exert a corepression - if the triplex adopt a conformation called ”canonical triplex”, discribing a conformation where the two microRNAs bound the messenger RNA without overlap or mutual exclusion, with a preservation of the seed binding for both of the microRNAs.

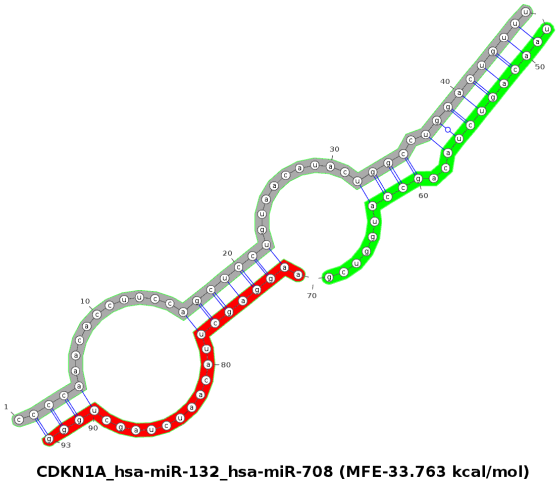


Figure 4.1: An example of canonical triplex, from the triplexRNA database

The triplex database can be interrogate through the web instance, using a JSON query. The database is also composed by three SQL table :

- A table listing the triplexes for their cognate genes in the mouse.
- A table listing the triplexes for their cognate genes in the human.
- A table associating genes and their cognate triplexes with a Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways identifier (ID) [25].

Each of them can be query using MySQL. The triplexRNA database which is interrogate with a MySQL request in the script Query Triplex is a local instance, different from the web instance called by the JSON queries (but both contains the same data).

Request of triplexes and gene inside the Melanoma Pathways

The list of the genes inside the KEGG Melanoma pathway [25] was retrieved by querying the corresponding SQL table for all the genes associated with the KEGG ID of the pathways (hsa05218).

The triplexes for this list of genes was then retrieved using the JSON query for a list of genes, as describe below.

SIF format, output of the JSON query

The sif format is the input format used for visualization using cytoscape [26].

This format specifies only nodes and interactions and it is written as follow :

nodeA relationship_type nodeB
nodeC relationship_type nodeA
nodeD relationship_type nodeE nodeF
Which, for the purpose of visualizating the microRNAs involved in the most triplexes, became, as outputted by the executable which retrieve all the triplexes for a list of genes :

microRNAname ID_of_the_triplexes microRNA2name

JSON query of a list of genes

The script which retrieve the functionnal triplexes take as input the name of a txt file containing the list of genes.

This list is then used to query the triplexRNA database with a JSON query and keep only the functionnal triplexes for each gene. The JSON query retrieve data under the form of JSON dictionnary that the script can then parse. Here an example of JSON query to retrieve all the triplexes associated with CDKN1A :

```
import requests
import json
response = requests.get("""http://www.sbi.uni-rostock.de/triplexrna/
JSON/Human/gene/CDKN1A""")
if response.status_code == 200:
  results = json.loads(response.content)
```

```
# results can be now accessed as a Python dictionary
```

The script also filter the triplexes considered as redundant. The output is wrotte in a sif format, in a separate folder, with a global report.

Filtering of the redondant triplexes

The redundant triplexes are filtered out considering the following : if a triplex in the list of triplexes associated to a gene has :

- the same position of beginning or ending for the microRNA's binding position of one already retained microRNA, for the first microRNA composing the triplex
- or the same position of beginning or ending for the microRNA's binding for the second microRNA composing the triplex,
- and a name of one microRNA in common with a triplex already retain for the sif output of this gene;

Then this triplex is considered as redundant and ignored for the sif output. The triplexRNA database which is interrogated with the MySQL request for the filtering is a local instance, different from the web instance which is the one queried with the JSON request.

Genes level of expression, Melanoma cancer

The level of gene expression for the Melanoma cancer used for the differential expression are RNA-seq data, published by the The Cancer Genome Atlas Network (TGCA) [18], from the project TGCA - Melanoma (SKCM).

The gene expression profiles was measured experimentally using the Illumina HiSeq 2000 RNA Sequencing platform by the University of North Carolina TCGA genome characterization center. This data are level 3 data, meaning that the gene-level transcription estimates are in log2(transformed RSEM normalized count + 1), and have been produced from a cohort of 331 patient [18].

The level of expression for the genes composing the KEGG melanoma pathways have been download through the Xena browser (http://xena.ucsc.edu/) [13]. UCSC's Xena is a data server-based platform that contain functional genomics data and output them in response to request done through a web graphical interface.

The Xena browser was used to retrieve the level of gene expression for each sample, associated with a phenotype [pathologic_stage], the categories of phenotypes going from the stage 0 to IV. This indices describe the extent of the cancer. See the pdf from the american joint committee on cancer [27]

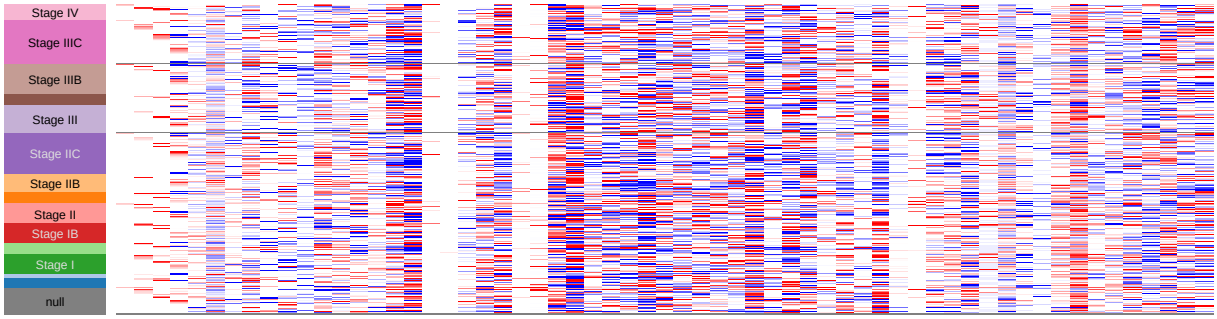


Figure 4.2: Web visualization of the level of gene expression, associated with the phenotype [pathalogic_stage]. This levels of expression can be then download.

Colorectal cancer data

The level of genes expression for the colorectal cancer was extracted from the R package curatedCRCData [28], a package providing manually curated data of microarray and RNA-seq data from the TCGA-COAD project.

The level of gene expression are RNA-seq level 3 data, in log2 (transformed RSEM normalized count + 1), the same format than the RNA - seq data of the TGCA - SKCM project.

Graphical visualization of the cooperation between microRNAs

The visualization of the cooperation between microRNAs in a form of graph has been done using Cytoscape [26], an open source software for integrating biomolecular interaction networks.

The format used as input to construct the network is the .sif format, cf the SIF format section above.

Differential expression analysis for melanoma and colorectal cancer data

The differential genes expression analysis between early and late melamoma stage has been done using gene-level transcription estimates in log2(x+1) transformed RSEM normalized count, provide by the TGCA-SKCM project, using the Limma package in R [29].

After Quantile normalization to equalize the library sizes [29], Limma has been used to fit a linear model to the already log 2 - transformed data using an empirical Bayes method [29].

The same procedure has been employed for the colorectal cancer data.

Randomized Lasso

The randomized lasso has been used on the gene-level transcription estimates in log2(x+1) transformed RSEM normalized count, provide by the TGCA in the project SKCM, using the python scikit-learn library [22].

Random Forest

The Random Forests has been trained using 2500 tree, 0 random state, directly on the gene-level transcription estimates in log2(x+1) transformed RSEM normalized count provide by the TGCA in the project SKCM, using the python scikit-learn library [22]. The testing set of data has been generate by selecting at random 20 percent of the total data set.

Random Forest Classification Error Plotting

The plotting of the error of the Random Forest has been done using the python library forestci.

Minimum Free Energy Calculation and Vizualization

The Minimum Free Energy Calculation is performed using the tool RNAfold of the package ViennaRNA 2.3.5. [30]

MicroRNAs alignments

The circular design script is provided with a copy of the mature.fa file from the last release of mirbase, downloaded in may2017, containing all the mature microRNAs referenced in the database.

An alignement using miRanda [31] is performed for each cluster against every human microRNAs, with the options "-noenergy -strict -sc 150". This options restrict the alignements to the ones with a strict alignment in the seed region (positions 2-8) region and eliminate results with little to no biological sense.

Pdf report

The pdf report is created using the web tool TEXPILE, developed by Martin Scharm. TEXPILE is a web service design to be an online compiler for LaTeX projects. The report is first write in the Tex language by the circular RNA design script. This Tex file is then convert into a pdf file by the TEXPILE tool.

TEXPILE's code is hosted at GitHub: github.com/binfalse/TEXPILE

Bibliography

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

[2] A. Grimson, K. K. Farh, W. K. Johnston, P. Garrett-Engele, L. P. Lim, and D. P. Bartel. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell*, 27(1):91–105, Jul 2007.

[3] David P Bartel. Micrornas: target recognition and regulatory functions. *cell*, 136(2):215–233, 2009.

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

[5] P. Saetrom, B. S. Heale, O. Sn?ve, L. Aagaard, J. Alluin, and J. J. Rossi. Distance constraints between microRNA target sites dictate efficacy and cooperativity. *Nucleic Acids Res.*, 35(7):2333–2342, 2007.

[6] X. Lai, U. Schmitz, S. K. Gupta, A. Bhattacharya, M. Kunz, O. Wolkenhauer, and J. Vera. Computational analysis of target hub gene repression regulated by multiple and cooperative miRNAs. *Nucleic Acids Res.*, 40(18):8818–8834, Oct 2012.

[7] U. Schmitz, X. Lai, F. Winter, O. Wolkenhauer, J. Vera, and S. K. Gupta. Cooperative gene regulation by microRNA pairs and their identification using a computational workflow. *Nucleic Acids Res.*, 42(12):7539–7552, Jul 2014.

[8] Erika Lasda and Roy Parker. Circular rnas: diversity of form and function. *Rna*, 20(12):1829–1842, 2014.

[9] L. Peng, X. Q. Yuan, and G. C. Li. The emerging landscape of circular RNA ciRS-7 in cancer (Review). *Oncol. Rep.*, 33(6):2669–2674, Jun 2015.

[10] S. Memczak, M. Jens, A. Elefsinioti, F. Torti, J. Krueger, A. Rybak, L. Maier, S. D. Mackowiak, L. H. Gregersen, M. Munschauer, A. Loewer, U. Ziebold, M. Landthaler,

C. Kocks, F. le Noble, and N. Rajewsky. Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature*, 495(7441):333–338, Mar 2013.

[11] Dan Han, Jiangxue Li, Huamin Wang, Xiaoping Su, Jin Hou, Yan Gu, Cheng Qian, Yun Lin, Xiang Liu, Mingyan Huang, Nan Li, Weiping Zhou, Yizhi Yu, and Xuetao Cao. Circular rna mto1 acts as the sponge of mir-9 to suppress hepatocellular carcinoma progression. *Hepatology*, pages n/a–n/a.

[12] J. C. Burnett and J. J. Rossi. RNA-based therapeutics: current progress and future prospects. *Chem. Biol.*, 19(1):60–71, Jan 2012.

[13] Jingchun Zhu, Brian Craft, Mary Goldman, Melissa Cline, Mark Diekhans, and David Haussler. Abstract b1-07: Using the ucsc xena platform to integrate, visualize, and analyze your own data in the context of large external genomic datasets. *Cancer Research*, 75(22 Supplement 2):B1–07–B1–07, 2015.

[14] B. P. Lewis, C. B. Burge, and D. P. Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1):15–20, Jan 2005.

[15] M. Spasic, C. C. Friedel, J. Schott, J. Kreth, K. Leppek, S. Hofmann, S. Ozgur, and G. Stoecklin. Genome-wide assessment of AU-rich elements by the AREScore algorithm. *PLoS Genet.*, 8(1):e1002433, Jan 2012.

[16] S. D. Chandradoss, N. T. Schirle, M. Szczepaniak, I. J. MacRae, and C. Joo. A Dynamic Search Process Underlies MicroRNA Targeting. *Cell*, 162(1):96–107, Jul 2015.

[17] Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The elements of statistical learning: data mining, inference, and prediction, 2nd Edition.* Springer series in statistics. Springer, 2009.

[18] Akbani et al. Genomic Classification of Cutaneous Melanoma. *Cell*, 161(7):1681–1696, Jun 2015.

[19] Evangelia Koutsogiannouli, Athanasios G Papavassiliou, and Nikolaos A Papanikolaou. Complexity in cancer biology: is systems biology the answer? *Cancer medicine*, 2(2):164–177, 2013.

[20] Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

[21] A. C. Haury, F. Mordelet, P. Vera-Licona, and J. P. Vert. TIGRESS: Trustful Inference of Gene REgulation using Stability Selection. *BMC Syst Biol*, 6:145, Nov 2012.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[23] H. Park, S. Imoto, and S. Miyano. Recursive Random Lasso (RRLasso) for Identifying Anti-Cancer Drug Targets. *PLoS ONE*, 10(11):e0141869, 2015.

[24] R. Denzler, S. E. McGeary, A. C. Title, V. Agarwal, D. P. Bartel, and M. Stoffel. Impact of MicroRNA Levels, Target-Site Complementarity, and Cooperativity on Competing Endogenous RNA-Regulated Gene Expression. *Mol. Cell*, 64(3):565–579, Nov 2016.

[25] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, 45(D1):D353–D361, Jan 2017.

[26] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13(11):2498–2504, Nov 2003.

[27] American Joint Committee on Cancer. Melanoma of the skin staging, 7th edition. 2009.

[28] Riester M Parsana P and Waldron L. "curatedcrcdata: Clinically annotated data for the colorectal cancer transcriptome.".

[29] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, 43(7):e47, Apr 2015.

[30] R. Lorenz, S. H. Bernhart, C. Honer Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker. ViennaRNA Package 2.0. *Algorithms Mol Biol*, 6:26, Nov 2011.

[31] A. J. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, and D. S. Marks. MicroRNA targets in Drosophila. *Genome Biol.*, 5(1):R1, 2003.