Understanding snoRNA interaction in healthy tissues through network expansion

Laboratory of Biological Data Mining - Project proposal

Artusi Alessia
Cretti Stefano
Ossanna Vittoria
Piazzi Andrea

December 14, 2022

1 Introduction

In the universe of RNA, small nucleolar RNAs (*snoRNAs*) consist of a family of short non-coding RNAs that are typically transcribed from non-coding regions such as introns or intra-gene sequences. There are two main categories of snoRNA, namely *H/ACA box* and *C/D box*. These classes can be distinguished due to their 3D-structure and their main functions; the former is known to perform pseudouridylation on specific RNA target sequences, while the latter regulates rRNA methylation. In the recent years many new functions of snoRNA, beyond rRNA modification, have been found; these functions take part in several biological processes and could be crucial in many cancers and neuro-degenerative diseases. These new functions are therefore a relevant topic for us to focus on, since they could be exploited for therapeutic purposes or as disease bio-markers.

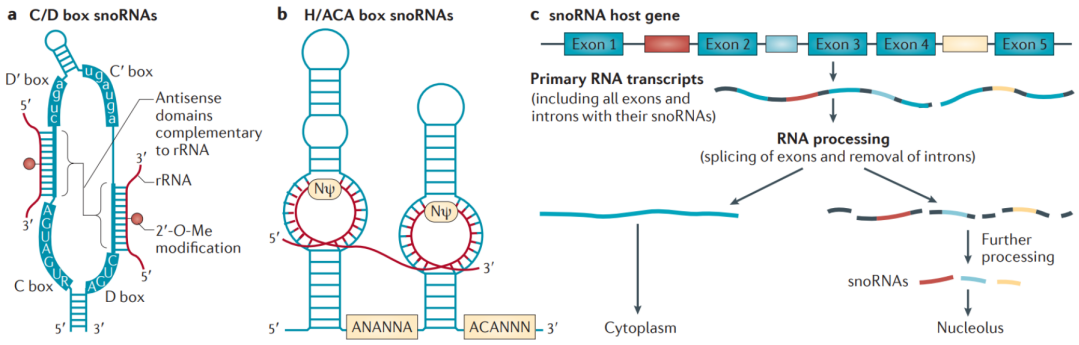

Figure 1: General overview of snoRNA's structure and behaviour

This project sets its main goal in studying snoRNAs that have been reported to be dis-regulated in cancer, and then defining which genes are causally related to them in physiological conditions. This is because, through a better understanding of their regulatory gene network, we could further advance the knowledge regarding their interaction mechanisms and consequently their clinical applications. Of particular interest is the correlation of snoRNAs with their host genes; literature shows that, unlike what would be expected, the expression of several snoRNAs does not correlate with that of their host genes, but a clear reason is yet to be defined [1].

For this analysis, gene-network expansion of all expressed snoRNAs will be performed, starting from a dataset consisting of TGIRT-seq data coming from seven healthy tissues. In order to perform the expansions, our project will exploit the *gene@home* project which hosts the NES^2RA algorithm. Moreover, a manually curated list of cancer-related snoRNAs will be defined starting from literature. This list plus information coming from differential expression analysis and snoDB2 will be used to enrich the expanded gene network.

After the enrichment, the main analyses that we intend to perform on the gene network are:

1. identification of putative novel interactions
2. gene set enrichment analysis on host-related snoRNAs and non-host-related snoRNAs
3. clustering analyses on snoRNAs according to several features

2 Data

2.1 Expression matrix

According to literature, conventional RNA-seq techniques fail to capture most of the snoRNA expression due to their 3D structure. The most promising sequencing method to get a reliable estimate of snoRNA expression is TGIRT-seq; this is a procedure that uses a fragment of a thermostable polymerase that allows to work at higher temperatures therefore denaturating even snoRNAs. Given this premise, we looked for a TGIRT-seq dataset and we found one from a study of Fafard-Couture et al. [2]. This dataset contains RNA expression levels in TPM (*transcript per million*) for seven healthy human tissues, namely breast, ovary, prostate, testis, skeletal muscle, liver and brain. For each tissue, three samples from unrelated individuals were used. The expression levels of roughly 55 thousands transcripts are present in the dataset.

2.2 snoDB2

snoDB2 [3] is an online interactive database that has been developed with the goal of consolidating information on snoRNAs. This wide database uses as its sources relevant available databases (those being *snoRNAbase*, *snOPY* and *snoRNA Atlas*), annotations from *RedSeq*, *Ensembl* and eventually manual curation. For each snoRNA, this source provides name and aliases, box type (H/ACA or C/D), host gene symbol, targets and their types, expression, Ensembl id, genomic coordinates and the nucleotide sequence.

2.3 Tumor-related snoRNAs

We manually curated a list of snoRNAs with known or presumed connection to cancer, involved both as tumor suppressor and/or oncogene. The information was extracted from literature available on PubMed; for each snoRNA in the list we reported some relevant features that could be used in the analyses downstream. We retrieved the connections from papers published from 2019 and onward, since among those there are several reviews trying to harmonize previous findings. Since snoRNA nomenclature is not unified and rather unclear, using snoDB2 as a medium, we converted the gene names into Ensembl ids; still, despite our efforts, some connections might have been missed due to this lack of cohesion.

3 Pipeline

In the following section we describe the pipeline we are going to use to carry out this project, which is shown graphically in Figure 2.
In the flowchart we represented:

- in blue the resources from which we obtained the data
- in grey the intermediate steps and data elaboration processes
- in orange the expected results

We downloaded the expression matrix from *GEO* [4]; this matrix was already pre-processed and normalized in TPM. We therefore performed feature selection keeping transcripts with at least 1 TPM in at least 20% of the samples. The filtered matrix contains around 13,500 genes.

Our analysis on the filtered expression matrix starts with differential expression analysis among the different types of tissues; from this list we then extract the differentially expressed snoRNAs. We would then like to perform gene set enrichment analysis (*GSEA*) using GO ontology; given the

absence of ontology for most snoRNAs (since they are non-coding genes), we aim to focus on the GSEA of the host genes of those snoRNAs.

From the filtered dataset we extracted the list of expressed snoRNAs, consisting of roughly 390 snoRNAs. Ideally, we would like to expand all of these snoRNAs starting from the filtered dataset using NES^2RA; however, given the time constraints, if the computation of the expansions becomes unfeasible, we will expand only a manually curated list of snoRNAs that have some sort of relation to tumor initiation and/or progression (in the worst case scenario only snoRNAs related to a specific type of tumor).

The expanded network will be further enriched using information stored in snoDB2 [3] and PubMed publications.

The analyses we intend to perform on the network will result in the following information:

- list of novel interactions through a comparative analysis with known interactions from snoDB2 and literature
- causal relations of snoRNAs with their host genes; moreover gene set enrichment analysis using GO ontology will be performed on the lists of host-related snoRNAs and non-host-related snoRNAs
- several clustered networks, according to different features, namely pathology, box type, causal relations with host gene

It is important to notice that these analyses might not be feasible, since we do not know a priori the connectedness of the graph. We expect to have at least some subnetworks having riboproteins as central nodes; moreover we know from literature that some snoRNAs interact with each other. Still, there is a chance that the resulting graph is too disconnected to perform clustering.

Moreover, further analyses could be performed on grouped snoRNAs, such as by pathology, in order to identify common mechanisms of action or molecular targets; again, this is only tentative, since it is subordinate to having enough snoRNAs per group in order to have statistically significant results.

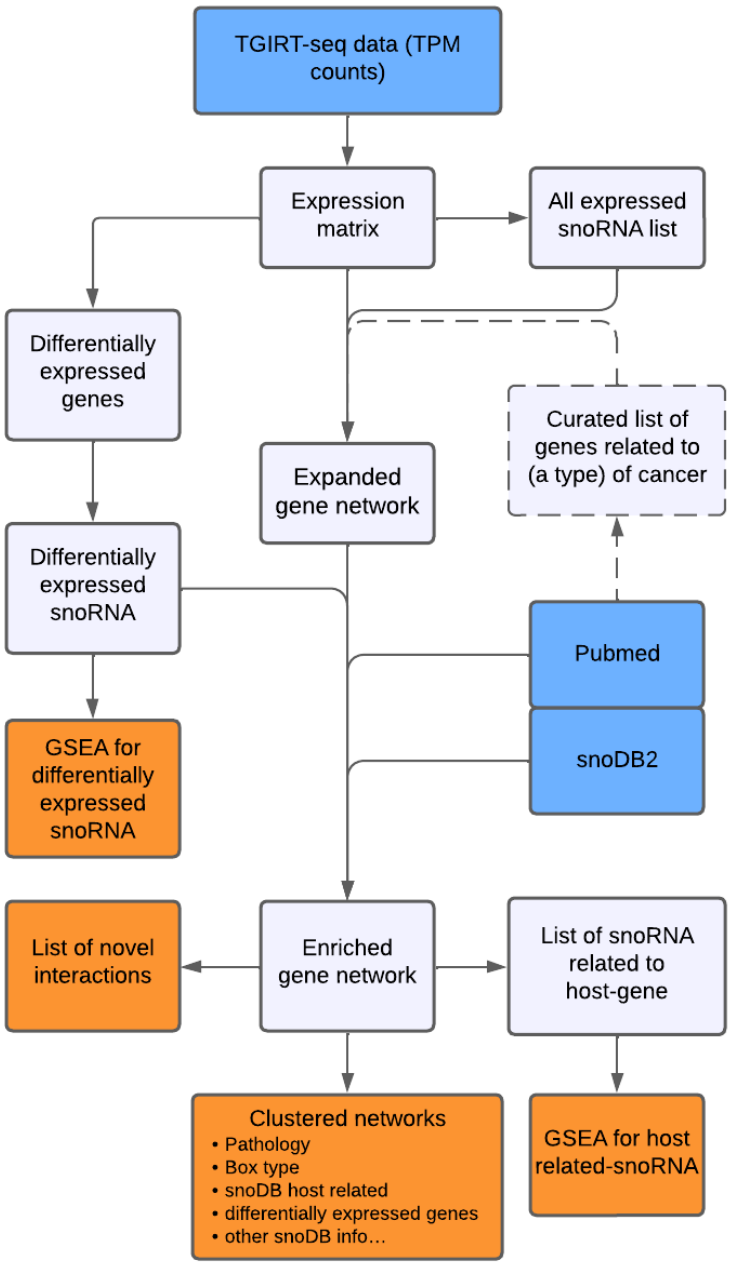

Figure 2: Project pipeline

References

[1] Philia Bouchard-Bourelle ; Clément Desjardins-Henri; Darren Mathurin-St-Pierre; Gabrielle Deschamps-Francoeur; Étienne Fafard-Couture; Jean-Michel Garant; Sherif Abou Elela; Michelle S Scott. snodb: an interactive database of human snorna sequences, abundance and interactions. *National Library of Medicine*, 2020.

[2] Étienne Fafard-Couture, Danny Bergeron, Sonia Couture, Sherif Abou-Elela, Michelle S. Scott. Annotation of snoRNA abundance across human tissues reveals complex snoRNA-host gene relationships. *Genome Biology*, 2021.

[3] Bouchard-Bourelle P;Desjardins-Henri C;Mathurin-St-Pierre D;Deschamps-Francoeur G;Fafard-Couture É;Garant JM;Elela SA;Scott MS;. Snodb: An interactive database of human snorna sequences, abundance and interactions. *Nucleic acids research*, 2020.

[4] Eric W Sayers; Evan E Bolton; J Rodney Brister; Kathi Canese; Jessica Chan;, Donald C Comeau; Ryan Connor . Database resources of the national center for biotechnology information. *NCBI*, 2022.