Analysis of snoRNA expression and interactions in several healthy tissues

Laboratory of Biological Data Mining - Final report

Artusi Alessia
Cretti Stefano
Ossanna Vittoria

December 14, 2022

1 Abstract

Small nucleolar RNAs (snoRNAs) are mid-size non-coding RNAs required for ribosomal RNA modification, implying a ubiquitous tissue distribution linked to ribosome synthesis. The study of the human snoRNome is getting increasingly interesting as new snoRNAs functions and roles are discovered. Our project focuses on the analysis of snoRNAs expression and interaction in seven healthy human tissues, such that this information can be exploited in studies on snoRNA-disease association.

We started from TGIRT-RNA sequencing data of 21 samples from seven healthy human tissues (brain, breast, liver, ovary, prostate, skeletal muscle, and testis) to investigate the distribution and tissue-specificity of snoRNAs. We noticed a predominant and specific expression of snoRNAs in brain, as reported in literature, and in reproductive tissues (ovary and testis);an over-representation analysis showed enrichment in known generic pathways - rna modification, nucleolar - but also some not explored pathways - protein binding and protein translation -.

From the expression matrix and two lists of genes of interest (consisting in snoRNAs and riboproteins), we carried on gene network expansions through the gene@home project. We performed a validation through snoRNA's canonical targets: only few of these genes have been found with high relative frequency. From the same networks we looked for correlation between snoRNAs and their hosts: our findings agree with the current state of the art, as we did not find sufficient proof of correlation.

Data and scripts are publicly available in the GitHub repository of the project.

2 Introduction

In the universe of RNA, small nucleolar RNAs (*snoRNAs*) consist of a family of non-coding RNAs that are typically transcribed from non-coding regions such as introns or intra-genic sequences. The biogenesis of most intronic snoRNAs includes cotraascription with the host gene, splicing, debranching of lariat intron, and exonucleolytic digestion. SnoRNAs consist of around 30-600 nucleotides and are divided into two main categories, 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 particular in its 2'-O-ribose methylation. The box C/D family is characterized by a kink-turn (k-turn) structure and contains two conserved sequence elements: box C (RUGAUGA) in the 5' region and box D (CUGA) in the 3' region. H/ACA box snoRNAs are composed of conserved box H (ANANNA, where N represents any nucleotide) and box ACA (ACA) motifs; these snoRNAs are marked with a hairpin-hinge-hairpin-tail secondary structure.

In 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.

This project sets its main goal in studying snoRNAs abundance and in 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 and riboproteins 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

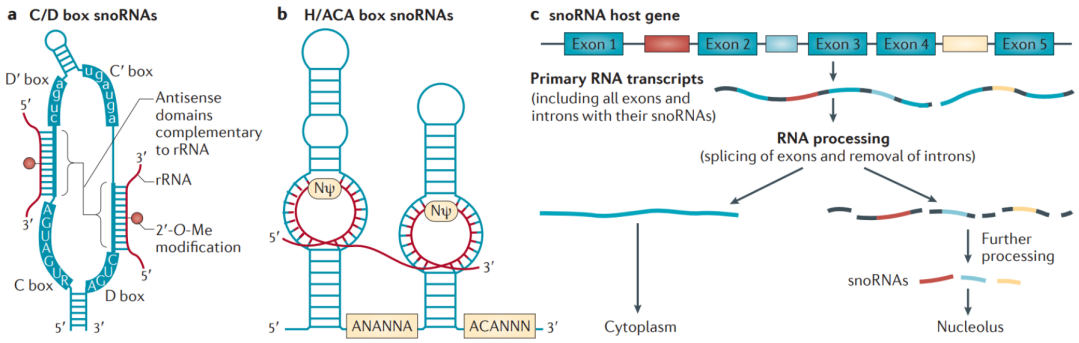


Figure 1: General overview of snoRNA's structure and behaviour

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. validation of the networks obtained from the expansions with the gene@home project through a comparison with current literature and **some statistical test that Stefano is running**
2. searching for links between snoRNAs and their host through the networks generated from the expansions
3. differential expression analysed on the seven tissues in a *1 vs all* approach
4. over-expression analysis on host of tumor-related snoRNAs and among the overall set of snoRNA that are currently available from snoDB2
5. over expression analysis on host-related snoRNAs differentially expressed in healthy tissue and every host-related snoRNAs expressed in the dataset

This project has been realized using a combination of python and R scripts published in the GitHub repository, along with all software specifications of packages and versions.

3 Materials and Method

Firstly, we will go over the expression datasets analysed throughout the project. Then we will describe the content of the databases used to retrieve the various annotations; our usage of said databases will be briefly introduced and further explained during the processing steps that make use of them. Finally we will discuss the processing and analysis pipeline of the project breaking it into chunks.

3.1 Expression datasets

3.1.1 TPM-counts matrix

According to literature, conventional RNA-Seq techniques fail to capture most of the snoRNA expression due to their 3D structure [2]. The most promising sequencing methods to get a reliable estimate of snoRNA expression are low structure bias RNA-Seq approaches; among those, RNA-Seq using a thermostable group II intron reverse transcriptases (TGIRT-Seq) is of particular interest. TGIRT-Seq is a procedure that uses a fragment of a thermostable polymerase which allows to work at higher temperatures therefore denaturating all types of nucleic acids, even snoRNAs; this fact

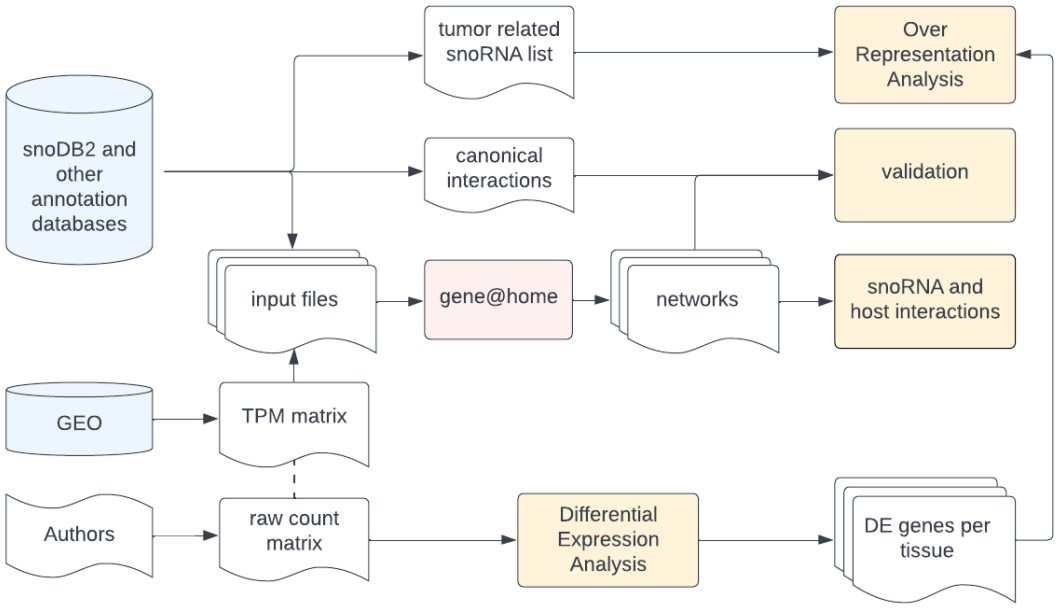


Figure 2: This pipeline describes the overall process of our work. Starting from two separate files (the TPM and raw count matrix) we performed several analyses (orange boxes): validation of the networks expanded through the gene@home project, searched for snoRNA-host expression correlations, differential expression analysis and over representation analysis.

makes TGIRT-Seq optimal for studying not only snoRNAs but all kind of structured RNAs (tRNAs for instance).

Given this premise, we selected a TGIRT-Seq dataset from a study of Fafard-Couture et al. [2]. This dataset contains RNA expression levels in transcripts per million (TPMs) 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. All prostate, testis, skeletal muscle and brain samples, plus two of the liver samples come from male donors; all ovary and breast samples plus one of the liver samples come from female donors. Regarding the age of the donors (available for all samples but breast and ovary ones), the minimum age is 24 years, the maximum is 76 years, the average is 47 years, and the median is 43. For each sample, the expression levels of roughly 55 thousands transcripts are present in the dataset.

It is important to notice that the samples, before sequencing, have been ribodepleted, meaning that the ribosomes have been removed; this was done because ribosomal RNAs are very abundant and they would therefore reduce the dynamic range of detection for less abundant transcripts. Moreover the objective of the authors of the paper was to study snoRNAs interactions other than the ones with rRNAs.

3.1.2 Raw-counts matrix

In order to perform differential expression analysis using the most common analysis packages, the expression levels must be expressed in raw-counts, since the packages internally perform appropriate filtering and normalization steps [3][4]. Since TPM-counts are normalized in such a way that information on the starting library size is lost, and it is therefore impossible to compute raw-counts from

TPM-counts, we contacted the authors of the publication we selected ([2]) asking for the raw-counts matrix of the study. The raw-counts table we were given is actually not the one used for the study, but rather a slightly different version obtained from a re-analysis of the raw sequencing data from the same samples of the study, using a more detailed annotation file (especially regarding snoRNA sequences).

We considered switching to the normalized version of this matrix also for the expansions (discussed later). Still, considering that the computation of the expansions had already started using the TPM matrix, the fact that we had to necessarily use the raw-counts matrix for differential expression analysis, and the fact that the raw-counts matrix is just a re-analysis of the same data, we were willing to accept the supposedly small discrepancy. Moreover we though that, if differences were to arise from the usage of one matrix with respect to the other, given how closely related the two are, those differences would most likely be small and statistically debatable.

3.2 Databases

3.2.1 Gene Expression Omnibus

Gene Expression Omnibus (GEO) is a public functional genomics data repository. Its archive freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomics data submitted by the research community. We downloaded the TPM-counts matrix (split in 3 separate files) from GEO Series archive.

3.2.2 snoDB2

snoDB2 [5] 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 (derived from text mining) and their types, expression, Ensembl id, genomic coordinates and the nucleotide sequence (and others). We used this database for id conversion, to define the host-gene (if present) and the targets.

3.2.3 PubMed

PubMed is a free search engine accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics. The United States National Library of Medicine (NLM) at the National Institutes of Health maintain the database as part of the Entrez system of information retrieval. We used this database to define a list of snoRNAs related to any type of tumor.

3.2.4 HUGO Gene Nomenclature Committee

The *HUGO Gene Nomenclature Committee* (*HGNC*) [6] is responsible for approving unique symbols and names for human loci, including protein coding genes, ncRNA genes and pseudogenes, to allow unambiguous scientific communication. All approved symbols are stored in the HGNC database, which keeps records of gene nomenclature, gene groups and associated resources including links to genomic, proteomic and phenotypic information. We used this database to retrieve the list of ribosomal proteins (according to GO ontology) and for some id conversions.

3.2.5 Biomart

Biomart is a web based tool that allows extraction of data without any programming knowledge or understanding of the underlying database structure. We mostly used the R package version of

Biomart for id conversion and to download GO ontologies. In fact, this R package enables the retrival of large amounts of data in a uniform way, the biggest resource offered is the Ensembl database. [7]

3.3 Pipeline

3.3.1 Generating input files for the expansions

This step is represented in Figure 3. Firstly we manually downloaded from GEO the three files containing the TPM-counts, those having identifiers GSE157846, GSE126797 and GSM4838073 respectively. We merged and formatted those files into a single TPM-counts matrix. After testing some different sets of filtering parameters (for the grid of parameters and the evaluation metrics used, refer to Table), we opted to perform feature selection keeping transcripts with at least 1 TPM in at least 20% of the samples. The filtered TPM-counts matrix (from here onward referred to as filtered matrix) contains the expression levels of around 13,500 transcripts per sample.

Filter	Genes	snoRNA	Hosts	Hostless	Paired
(0, 0.0)	58884	951	1045	371	571
(1, 0.1)	19341	420	913	37	363
(1, 0.2)	13509	396	800	29	339
(1, 0.3)	11690	385	738	25	329
(2, 0.2)	9492	378	622	20	320
(2, 0.5)	5145	345	343	11	252

Table 1: This table reports the filtering results. In the first column we find the number of TPM per sample accepted. Next we find the number of genes that correspond to the criteria, number of snoRNAs, number of hosts of snoRNAs, hostless and paired snoRNA.

Then, we manually downloaded the entirety of the information present in snoDB2; from it we extracted the list of all annotated snoRNAs. Consequently, from HUGO we downloaded the full list of genes annotated as coding for non-mitochondrial ribosomal proteins according to the GO ontology. We were interested in the list of ribosomal proteins since by including them in the expansions, we hoped to improve the chances of obtaining a connected network, rather than a set of disjoint elements, given that we expected them to act as hub-nodes. We chose ribosomal proteins rather than ribosomal RNAs (which would be the most sensible choice given their primary function in rRNA modification), since those are not present in our filtered matrix (due to ribodepletion as previously mentioned); ribosomal proteins then become the second best choice since many snoRNAs are known to interact with ribosomal proteins (as per snoDB2 information). Moreover, we speculated that it is likelier for ribosomal proteins to be found causally related to snoRNAs in the absence of rRNAs, rather than in their presence, given that the skeleton procedure of the PC-algorithm cannot condition with respect to the rRNAs. We excluded mitochondrial ribosomal proteins since, as far as our knowledge, there is no documentation of snoRNAs acting inside the mitochondria.

By intersecting the genes present in the filtered matrix and these two lists (all snoRNAs and non-mitochondrial ribosomal proteins), we obtained two lists on which the expansions are based on:

- a list of 396 snoRNAs present in the filtered matrix
- a list of 86 non-mitochondrial ribosomal proteins present in the filtered matrix

Each transcript of these two lists would then be expanded individually.

Using these last two lists we also created three mock-up local gene networks, one containing all ribosomal proteins, one containing all snoRNAs and finally one containing all snoRNAs and all ribosomal proteins. These local gene networks were also given as inputs for expansion.

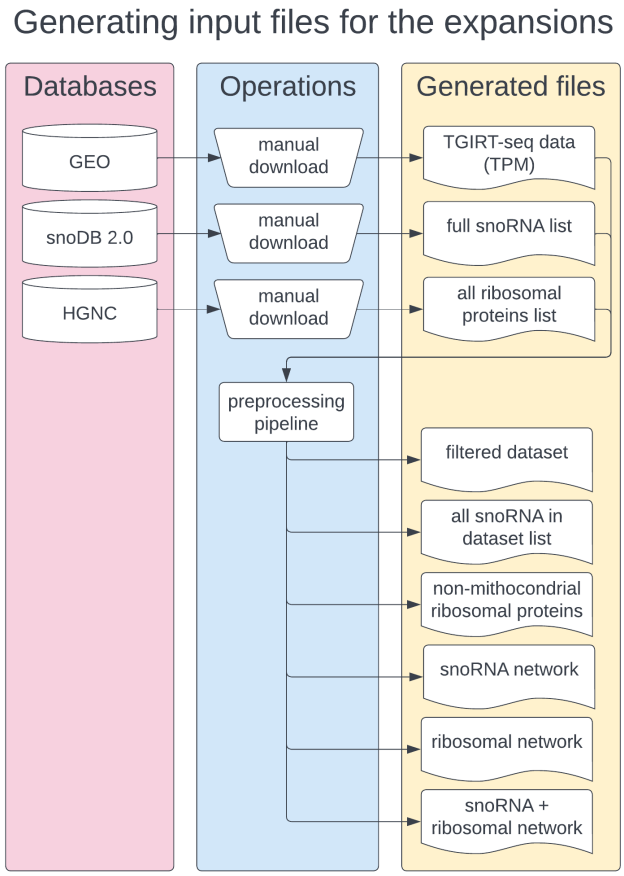


Figure 3: The pipeline above represent the process for generating the filtered matrix and for defining the list of genes of interest and local gene networks to expand.

3.3.2 Gene and network expansion

This step is described in Figure 6b. The next milestone of our project was to expand, starting from the filtered matrix, the lists of genes and the local gene networks described in the previous section, that is, to find new causal relations among the genes of the network (where the individual gene is considered as a sub-case in which the starting network contains only that gene) and with transcripts not in the starting network. In order to do this, we exploited the project *gene@home* [8].
The gene@home project provides a tool to expand a pre-existing local gene network (LGN, namely a subset of the entire gene network of an organism) using transcriptomic data; this means identifying new genes connected to the ones already present in the LGN via causal relations. To perform this task, the gene@home project uses expanded and specialized versions of the PC-algorithm [9], namely NESRA, NES^{2}RA [10] and OneGenE [11].
Since these algorithms are rather computationally expensive, the gene@home project takes advantage of the TN-Grid, which is a way to perform distributed and voluntary computing. The TN-Grid is based on the BOINC (Berkeley Open Infrastructure for Network Computing) system, and as such it has two main components: work generator and validator.

As stated previously, the expansion of the ribosomal proteins was driven by the idea that, by

including them, they would act as hub-nodes given the demonstrated interactions with multiple snoRNAs, therefore making easier to create a connected graph.

For all the expansions that we carried on, we kept the same parameters in order to obtains results as homogeneous and comparable as possible. After testing for execution time using the provided bench-marking script, we decided to set `alpha_value = 0.05`, `iterations = 1000` and `t-size = 4000`. The `alpha_value` corresponds to a set of significance values, `iterations` is the number of NES^{2}RA iterations, while `t-size` is the number of transcript in each tile.

3.4 Expansion aggregation

This step is described in Figure 4. For the sake of simplicity and ease of access, we aggregated the results from each lists of individual gene expansions in a single file. This was done iterating over the single interaction files (automatically downloaded from gene@home) and storing the name of both the interactors, along with the relative frequency of the interaction (defined as number of times the genes have been found causally related over the number of times they have been included in the same expansion tile). This led to files uniformly formatted as `gene_x, gene_y, relative_frequency`. Every network expansion was also parsed into the same format in order to facilitate comparison.

Since some snoRNAs interact with each other, while aggregating the individual gene expansions we would come across the situation in which we had the causal relation between snoRNA A and snoRNA B both in the expansion of snoRNA **A** and in the expansion of snoRNA **B**. In order to properly merge the single interaction files, we therefore needed to define a good metric to aggregate the value of the relative frequencies of the causal relations. For this reason, we decided to keep the average of the relative frequencies found in the two files. It is reasonable to state that, if in the expansion of snoRNA A, snoRNA A and snoRNA B are causally related, then in the expansion of snoRNA B, snoRNA A and snoRNA B should be causally related too; for this reason, even if one of the expansion files of the two snoRNAs in question does not contain the causal relation among the two snoRNAs while the other does, the average is computed between the relative frequency of the file in which is present and zero for the other file.

Overall, this process led to the creation of six files:

1. expansion of the network composed of the 396 snoRNAs (SN).
2. expansion of the network composed of the 86 ribosomal proteins (RN). .
3. expansion of the network composed of the 396 snoRNAs and 86 ribosomal proteins (SRN)
4. aggregated individual expansions of the 396 snoRNAs (ASN)
5. aggregated individual expansions of the 86 ribosomal proteins (ARN)
6. aggregated individual expansions of the 396 snoRNAs and 86 ribosomal proteins (ASRN)

3.4.1 Expansion validation

This step is described in Figure 5. In order to validate the biological relevance of the expansions, we compared the known interactions of the snoRNAs (from here onward referred to as canonical interactions) with those found using gene@home; the higher the overlap of the two, the higher the confidence in the biological pertinence of the results. The canonical interactions were retrieved starting from snoDB2. In this database, for each snoRNA of interest, we considered the following fields: snRNA targets, lncRNA targets, protein coding targets, snoRNA targets, miRNA targets, tRNA targets, ncRNA targets, pseudogene targets and other targets. Target data in snoDB2 includes known targets from rRNA annotated in snoRNAbase and rRNA targets confirmed by RiboMethSeq. Non-canonical interactors (all non-rRNA targets) included in the database were reported in literature

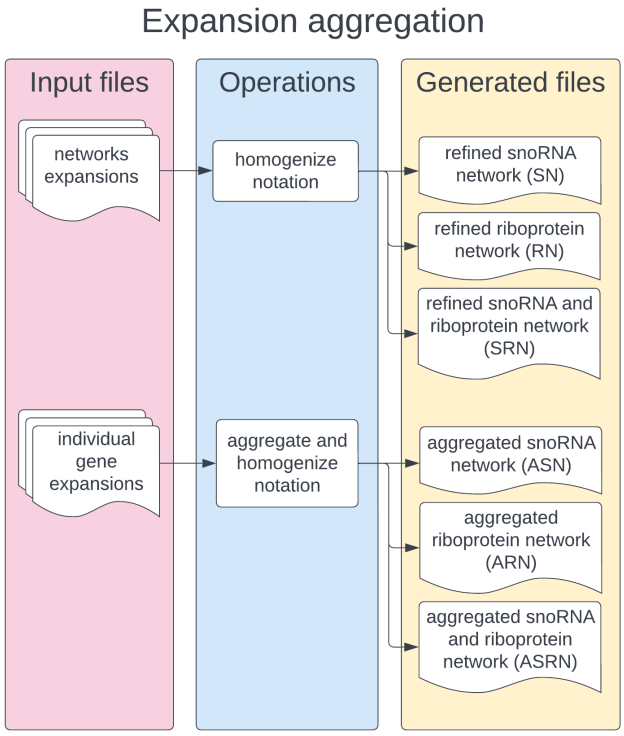


Figure 4: This pipeline describe the process of getting a uniform notation for every network and expansion we considered in this study.

to be experimentally validated [5] [12] [13]. Please note that the field *rrna targets* was omitted since the dataset that we used for the expansions was ribodepleted.

Given that the nomenclature conventions for the genes are highly heterogeneous, especially among different classes of genes (protein-coding, ncRNA-coding, tRNA-coding), to the point where direct comparison of gene lists becomes almost impossible, we decide to translate all gene names to Ensembl ids. This was crucial since the canonical interactors in snoDB2 are identified through their gene name, while almost all the genes in the filtered matrix are identified through their Ensembl id. For each canonical interactor, we tried to associate one (or more) Ensembl ids corresponding to its gene name (or aliases). The association of the gene name (or its aliases) to an Ensembl id was carried on through the annotation of some datasets, namely:

- *Biomart* (from Ensembl) for most of the genes
- *snoDB2* for missing snoRNA ids
- *HGNC* datasets for other missing genes (especially ribosomal proteins)

Although we considered several sources, some genes have no Ensembl id or we could not find one: among all canonical interactors considered from snoDB2, for the 396 snoRNAs of interest, 57 genes have no Ensembl id associated; we are aware that this entails that we might have missed some interactions but they do constitute a very small fraction of the overall number of canonical interactors. A complete list of genes with no Ensembl id associated is available in the GitHub repository.

After gene name conversion to Ensembl id, we used these canonical interactions to validate the results of the expansions. For each of the six files mentioned in the previous section, we checked - with different thresholds for the relative frequency of causal relation - how many of the canonical interactions were found by the algorithm used for the expansion.

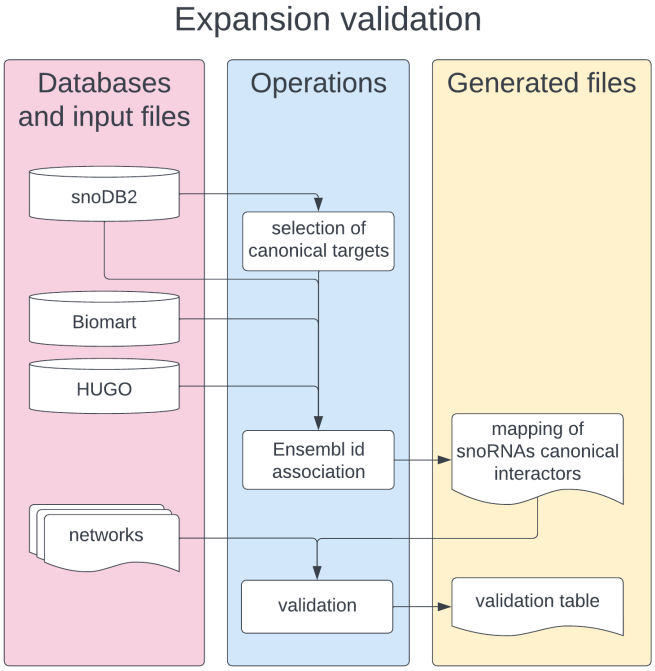


Figure 5: This pipeline describe the overall process we performed in order to obtain a validation of the networks.

3.4.2 Tumor-related snoRNA list curation

This step is described in Figure 6a. While the computationally intensive and time consuming expansion step was running, we manually curated a list of snoRNA documented to be dysregulated in some type of human cancer. The idea was to define if some snoRNAs or classes of snoRNAs are particularly related to a specific type of cancer, and in a second moment to use this information to annotate the network obtained from the expansions to analyze potential clustering; both these ideas, as we explain in the results sections, had to be scrapped in favor of a more generic over-representation analysis of GO-terms in tumor related snoRNAs (explained in the following paragraphs).

We started by searching for publications on snoRNAs dysregulation in tumors on pubmed.gov (query using "snorna tumor" in any field). We restricted our research to papers published in the last four years (from 2019 to 2022, both included), assuming that all good quality results present in publications from previous years would be included in the copious number of reviews on the topic published in the last four years.

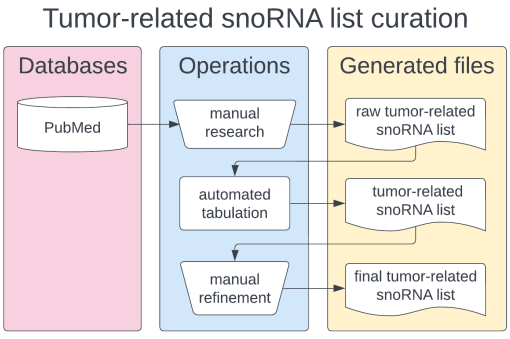
We included all studies performed on human or human-derived cell cultures (immortalized ones too), regardless of tissue and tumor type, which correlate in any way a snoRNA to the pathology. Since the genesis of a tumor induces general transcriptional instability, causing dysregulation of many genes and low transcription levels for most genes that should be inactive, it is in general very easy

to fit a classifier to distinguish tumoral from healthy tissues due to these patterns. Also due to this transcriptional instability, most snoRNAs display higher expression levels in tumoral tissues; for this reason we decided to exclude papers presenting large numbers of differentially expressed snoRNAs since deemed of low statistical relevance, while keeping those analyzing individual snoRNAs or small snoRNA signatures (less then 10 snoRNAs).

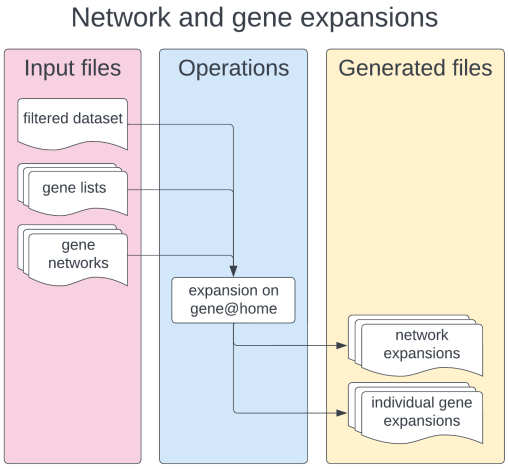
In order to produce this list, we started from a set of snoRNAs that have been published by a review paper [14], in which the authors reported a table of 48 snoRNAs involved in different types of human cancers and we progressively expanded upon it. We collected a total of 228 snoRNAs that are supposed to have a connection with cancer.

We encountered several difficulties into defining this list, since snoRNA nomenclature is unclear, redundant and inconsistent: we made an endeavor in order to find all references that connect snoRNAs with cancer, still - despite our efforts - some connections might have been missed due to this lack of coherence. We also need to report that, just like for the expansion validation, since our filtered dataset is defined on genes with a unique Ensembl id, we mapped the standard names (or aliases or deprecated nomenclature) into Ensembl ids. Also in this case, the association is not unique nor complete. In fact, some snoRNAs lack an Ensembl id, some other have more than one: for this latter case, we kept every unique label, since it is possible - from a biological point of view - that a snoRNA gene is present in more than one copy along the genome.

We divided the snoRNA according to the type of tumor they were associated to. We defined 16 major groups of tumors: bladder, brain, cervical, colon, gallbladder, gastric, hepatocellular, lung, ovarian, prostate, renal, leukemia, myeloma, osteosarcoma, breast aggressive and non-aggressive. Some snoRNAs did not fall into any of these classes or belonged to multiple of them. This division did not end up mattering since, aside from a couple groups, the number of snoRNAs belonging to each group was too low for any statistically meaningful analysis. Since any grouping according to biological features would be rather arbitrary and heterogeneous, we decided to just consider them jointly, as a list of snoRNA dysregulated in some kind of tumor.



(a) This pipeline shows the process used to obtain a refined list of tumor related snoRNAs.



(b) This pipeline shows the process used to obtain the expansions, both as single genes and as network expansions.

Figure 6

3.5 Differential Expression Analysis

This step is described in Figure 7a. The differential expression analyses (DEA) were performed using the R package `DESeq2`. As described above, the least biased methods to perform such analyses are the one relying on raw counts data (and not TPM); DESeq2 requires as input a table of raw-counts, which we could neither retrieve from our TPM-counts matrix nor it was available on GEO or recount3. Luckily the raw-counts table was kindly provided by the authors of the paper in an updated form (described in section 3.1.2). DESeq2 came in handy also because it does not need the records of the gene lengths, which would have been hard to retrieve (especially for ncRNAs and tRNAs), as it performs an internal normalization which computes the geometric mean for each gene across all samples, to finally divide the counts of the gene by this mean. This computation corrects the library size and RNA composition bias, while the intra-sample normalization is not performed.

Our focus is to understand whether there are snoRNAs specific for each type of tissue, and whether they are associated with any particular pathway or function. We performed differential expression analysis in a one vs all manner (since the pairwise approach was not sufficiently informative). In this way we obtained the lists of the genes differentially expressed in a tissue with respect to the average of the others for all types of genes, not only for the snoRNAs.
The DESeqDataSet object was built from the raw-counts matrix using `DESeqDataSetFromMatrix()` and it was filtered using a threshold of at least 5 counts per gene in the 20% of the samples. The filtered dataset was passed to `DESeq()` which estimates the size factor for the normalization, the gene-wise dispersion, and fits a model to the dataset in input through a negative-binomial distribution. The output files were saved in the form of tables where each row is a gene with five columns of information each, in particular:

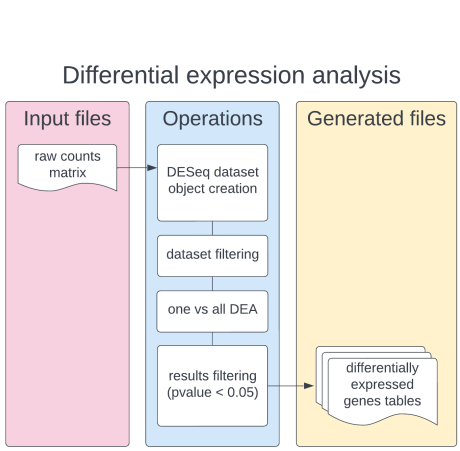
- `baseMean`: average of normalized count values divided by size factor
- `log2FoldChange`: effect size estimate
- `lfcSE`: standard error estimate for LFC
- `stat`: Wald test
- `pvalue`: Wald test p-value
- `padj`: Benjamini-Hochberg adjusted p-value

Each output gene list was filtered to retain only significant differentially expressed genes setting a cutoff on the p-value of 0.05. The analyses were carried out in `R` and the script is available at Github repository.

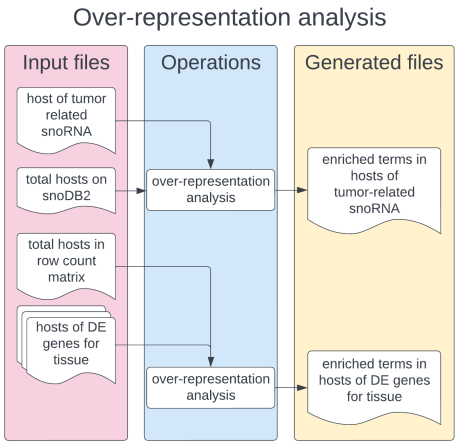
3.6 Over-representation Analysis

This step is described in Figure 7b. Over-representation analysis (ORA) is a test used to define whether a list of labelled items of interest contains significantly more items with a specific label with respect to a random sample of the same size from a reference list. Translated to biological applications, it can be rephrased as, given a list of genes with some function (usually a GO or KEGG annotation), test if a function is more represented (enriched) in the list with respect to a list of randomly sampled genes from a reference list of genes. For each function an enrichment p-value is computed (and adjusted for multiple testing), which basically states the confidence of the enrichment (the lower it is the higher the confidence). The difficult part is usually defining a proper reference list.

ORA is significantly easier to perform and analyze with respect to Gene Set Enrichment Analysis (GSEA), but it requires the assumption of independence of the items in the list, requires and arbitrary



(a) This pipeline describe the process of performing Differential Expression Analysis



(b) This pipeline describe the process of performing Over-representation analysis

Figure 7

p-value cutoff and is in general less powerful. On the other hand GSEA is more powerful and makes almost no assumptions, but computing the permutations required for the analysis can be slow and it requires that there are absolutely no duplicates in the list; moreover it requires some generic statistic (p-value, correlation, entropy reduction or others) associated to each item. These last two points made us opt out from using GSEA.

snoRNAs are non-coding genes and usually are not annotated with regards to their function; on the other hand most snoRNAs are placed in the non-coding regions of some other gene, the so-called host gene. We therefore decided to analyze whether a list of snoRNAs linked to a particular condition or tissue tends to have hosts with a specific function. It is important to notice that one gene might be the host of multiple snoRNAs, and this is especially true for long-non-coding genes; moreover it has been shown that the expression levels of snoRNAs and protein-coding hosts tend to not correlate, while they tend to for non-coding hosts. For this reason, it is not possible to consider multiple snoRNAs with the same host as independent. It is though a fair assumption that snoRNA with no host (placed in intra-genic regions) are independent. In short, computing the lists of interest and the background lists, we consider duplicate hosts just once, while we introduce a "host-less snoRNA" placeholder for each host-less snoRNA. This makes for a more fair comparison than only considering the snoRNAs with hosts in both the list of interest and the reference list.

We performed ORA for two different conditions:

- **ORA for hosts of snoRNAs tumor related**:

  As stated in the previous section, we gathered a list of snoRNAs from manual curation. We therefore compiled the list of Ensembl ids of the host genes of these snoRNAs (when available), resulting in a list of 304 hosts of interest (before de-duplication and padding). The background list of this analysis was the list of all host of all snoRNAs in snoDB2, composed of 995 genes (before de-duplication and padding); this is because the list of interest, the tumor-related snoRNAs list, comes from the literature and could therefore include any snoRNA.

- **ORA for hosts of snoRNAs differentially expressed in tissues**:

  We also performed ORA for the hosts of snoRNAs that have been detected as differentially expressed in some tissue. In order to do this, we considered the differential expression analyses

in format "one vs all". This means that we kept the snoRNAs defined as differential expressed in a specific tissue compared to the average of the other six tissues. For each of the seven tissue, we collected a list of hosts related to the differential espressed snoRNAs. In particular we gathered:

- 33 hosts for breast vs average of all other tissues
- 44 hosts for liver vs average of all other tissues
- 14 hosts for prostate vs average of all other tissues
- 331 hosts for brain vs average of all other tissues
- 196 hosts for testis vs average of all other tissues
- 48 hosts for skeletal muscles vs average of all other tissues
- 155 hosts for ovary vs average of all other tissues

(all the values refer to before de-duplication and padding). The list of interest is therefore the list of hosts of DE snoRNAs for each tissue in turns, while the reference list is always the set of all snoRNAs present in our raw-counts matrix (which is the one on which we performed DE). The reference list is composed of 976 hosts (before de-duplication and padding).

The analysis was performed mainly using the R package `clusterProfiler`.

3.7 Code availability and software specifics

Every step of this project has been realized using a combination of python and R scripts published in the GitHub repository, along with all software specifications of packages and versions.

4 Results

4.1 Network analyses

In this section, we describe the results obtained from the expansion through the NES^{2}RA algorithm and the gene@home project.

4.1.1 Network validation

In order to further proceed with downstream analyses on the networks, we performed a validation with data extracted from literature. In particular, we tried to validate the networks through the canonical correlations of snoRNAs and other genes. As specified in Section 3.4.1, we acquired associations between Ensembl labels of snoRNAs of interest with their canonical interactors (always annotated with Ensembl identification). Since we are taking into account canonical interactors coming from literature and experimental verification, we expect these interaction as as resut of the expansions.

Therefore, we went through each of the six networks and we kept track of every canonical interaction found from the expansions within their relative frequency. This analysis has been carried on with different thresholds. Along with the cutoff of the relative frequency value, we further analyzed cases of stricter filtering values on the initial TPM matrix. With this further step we decrease the number of genes present in the dataset to higher values (gene with 5 TPM in at least 40% and 10 TPM in at least 40% of samples in the dataset). This reduces the number of canonical interactors present in the dataset: when this happens, it means that the supposed canonical interactors was not enough expressed, therefore it is plausible that the algorithm did not find the connection for this very same reason. Still, the fraction of canonical interactors is really low in every condition analyzed, thus not reliable for more in depth analyses. In Table 4 in the Attachment 1 we report the results of the validation analysis.

4.1.2 snoRNAs and host correlation

Given the networks we obtained from the NES^{2}RA expansions, we looked for links between the snoRNAs - object of this research - and their hosts. We extracted a list of host Ensembl identification labels of their host from snoDB2 and looked for their interactions in the networks. In Table 2 we report the total number of snoRNA-host interactions that we found in each dataset, for a list of thresholds for the relative frequency.

From literature, it is known that snoRNAs do not often correlate with their host, the case in which is most common to see a correlation is whenever the latter belongs to the lncRNA family [1]. In line with this consideration, each network we obtained agrees with this fact: among the 396 snoRNAs expanded through the project gene@home, only few genes demonstrated to be correlated to their host. Among them, two snoRNAs - hosts correlation have been hound with a higher relative frequency (details in Table 3). Still, the relative frequency for one of them is still relatively low.

In the particular case reported in Table 3, we highlight that both MEG8 and SNHG4 (hosts of the snoRNAs) belong to the lncRNA family. Keeping in mind the considerations of the very last paragraph 4.1.1, we find that these results are coherent with previous literature.

	Thresholds (on frel)			
Networks	0	0.05	0.5	0.9
RN	0	0	0	0
SN	4	2	1	0
SRN	5	2	1	1
ASN	0	0	0	0
ARN	11	2	1	0
ASRN	11	2	1	0

Table 2: This table report the number of interaction snoRNA to its host found in each network analyzed along with the a threshold considered. The first item in each row corresponds to the specific network: RN - riboproteins network expansion, SN - snoRNAs network expansion, SRN - snoRNA and riboproteins network expansion, ASN - aggregation of snoRNAs single expansions network, ARN - aggregation of riboproteins single expansion network, ASRN - aggregated snoRNAs and riboproteins single expansion network. Among a total of 396 total snoRNA present in the expansions, only few of them demonstrated correlation with their host.

		Relative frequence					
		RN	SN	SRN	ASN	ARN	ASRN
SNORA74A MEG8	ENSG00000202048 ENSG00000225746	-	0.1547	0.1630	0.1791	-	0.1791
SNORD114-20 SNHG4	ENSG00000200959 ENSG00000281398	-	0.8923	0.9934	0.7921	-	0.7921

Table 3: This table reports the two most correlated snoRNA - host interactions found among the networks. The first gene name in the cell corresponds to the snoRNA gene name, the second to its host. The same order has been kept for the Ensembl labels.
The first item of the last six columns corresponds to the specific network: RN - riboproteins network expansion, SN - snoRNAs network expansion, SRN - snoRNA and riboproteins network expansion, ASN - aggregation of snoRNAs single expansions network, ARN - aggregation of riboproteins single expansion network, ASRN - aggregated snoRNAs and riboproteins single expansion network. It is noticeable that even the second correlation interaction frequency is relatively low. Also, we underline that both the hosts MEG8 and SNHG4 belong to the lncRNA family.

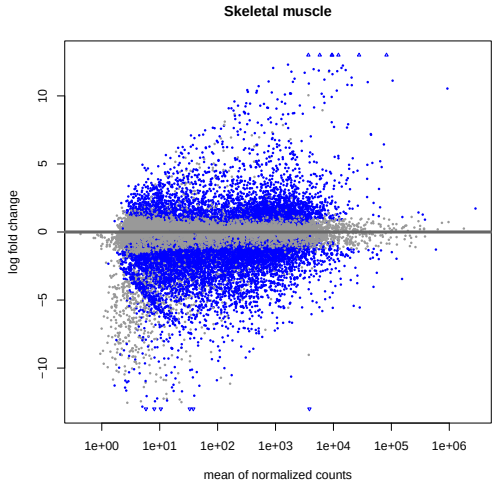
We quickly performed a standard statistical test to compare the snoRNAs and hosts correlations directly from the TPM table with an unsigned correlation coefficient trough WGCNA (*Weighted Correlation Network Analysis*). This test assign a correlation from zero to one between two genes. This approach discovered 339 snoRNA-host interactions with threshold equal to 0 on the Pearson correlation coefficient, 86 with threshold equal to 0.05, 3 interaction with 0.5 and only one with a cutoff set at 0.9. Without compare the exact values for the cutoffs (since are applied on different coefficients), we notice that the results of this test are slightly higher than the ones coming from the network, but with no significant margin. Also, we underline that the snoRNA - hosts genes found correlated by this test are different from the ones stored in the networks.

4.2 Differential Expression Analysis

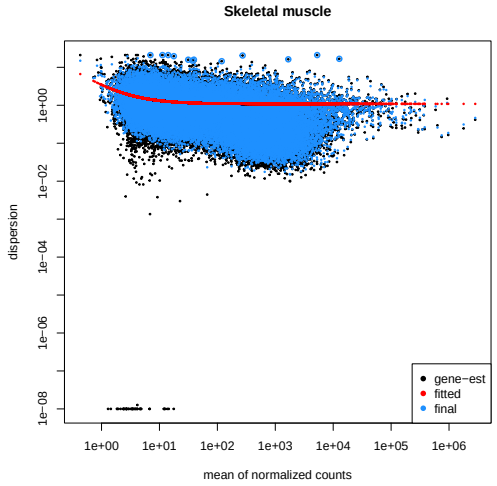
Gene expression profiles were available, in the form of a raw counts matrix, for 61915 genes in seven different types of tissue each in triplicate (21 total samples). To understand if there are tissue-specific snoRNAs in healthy tissues we approached the differential expression analysis in a one-vs-all manner.

4.2.1 One-vs-all Approach

The design of the one-vs-all approach for DESeq() was a simple formula: $design = \ tissue$, where tissue is a vector containing the tissue to be considered as reference level and as second level *other*, which identifies the six remaining tissues.
The quality of the analyses was assessed by computing the MA plot and the dispersion plot. For the sake of simplicity, in figure 8a and 8b are reported the two plots only for the analysis of Skeletal Muscle vs all, while the remaining plots are reported in the Supplementary Data file.



(a) This MA plot show in blue the significant differentially expressed genes (up- and down-regulated) while in black the unsignificant DE genes. Notice that the genes with lower average mean are called differentially expressed as much as the ones with higher average mean.



(b) DESeq2 computes the dispersion for each gene based on its expression level and observed variance. The inverse relationship between the mean and the dispersion is slightly visible through the fitted red line.

Figure 8

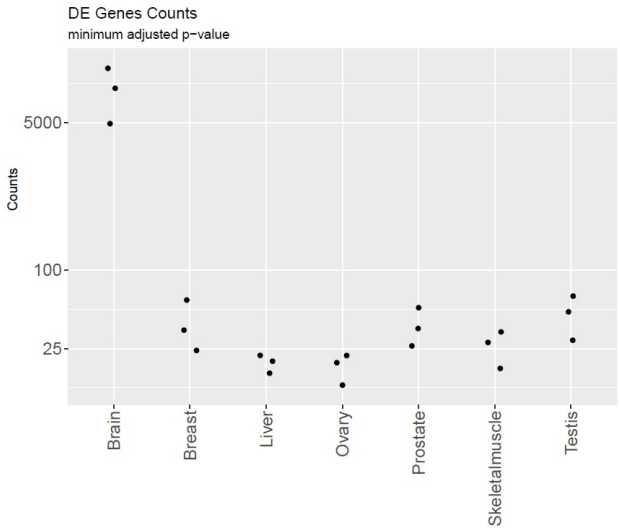


Figure 9: Counts of the most significantly DE genes (minimum adjusted p-value) across all samples are shown in the plot. The brain counts are substantially higher than the remaining six tissues counts.

In figure 9 we notice consistently higher counts for genes expressed in *brain*. To avoid having all the analyses driven by this tissue, we decided to perform the one-vs-all DEA excluding the three brain samples. However, the results do not differ in accuracy, except for an around 0.06% difference in the number of outliers detected by DESeq, which is why we decided to continue with the initial analysis retaining all seven tissues.

The results are summarized in figure 10. The counts are reported with a log transformation to highlight the relative abundance of snoRNAs and host genes compared to the total number of DE genes. The distribution of snoRNAs looks uniform if we normalize them over the total number of DE genes; however, brain shows the highest number of DE snoRNAs together with testis and ovary as expected [2].

4.3 Over-representation Analysis

The Over-representation analysis was performed using the r package `clusterProfiler` firstly on host genes of tumor related snoRNAs over the total snoRNAs present in our raw-counts matrix, and lastly on each list of snoRNAs generated from the one-vs-all differential expression analysis output. For each run a p-value cutoff of 0.10 was chosen to consider an enrichment significant, and the Benjamini-Hochberg method was selected for the p-value adjustment.

Figure 11a shows the results of the first OR Analysis. The enriched pathway are very generic, but still coherent with the known snoRNAs functions.

For what concerns the results of the DE snoRNAs host genes ORA the prostate came out empty, and breast show very low gene ration for only two slightly enriched pathway. Brain shows a more complex enriched environment, but the terms are still not specific as for the ones resulted from the tumor-related snoRNAs host genes. In fact, even in figure 11b we see terms such as: nucleolus, protein binding and RNA binding, as the most enriched terms.

The fact that tumor-related and tissue-specific snoRNA host genes present very similar enriched terms, even if generic, could be useful in more in-depth analysis on tumor-related snoRNAs roles, as they seem to match the ones in healthy state of tissues.

This analysis could be probably refined testing for actual independence among the genes (for

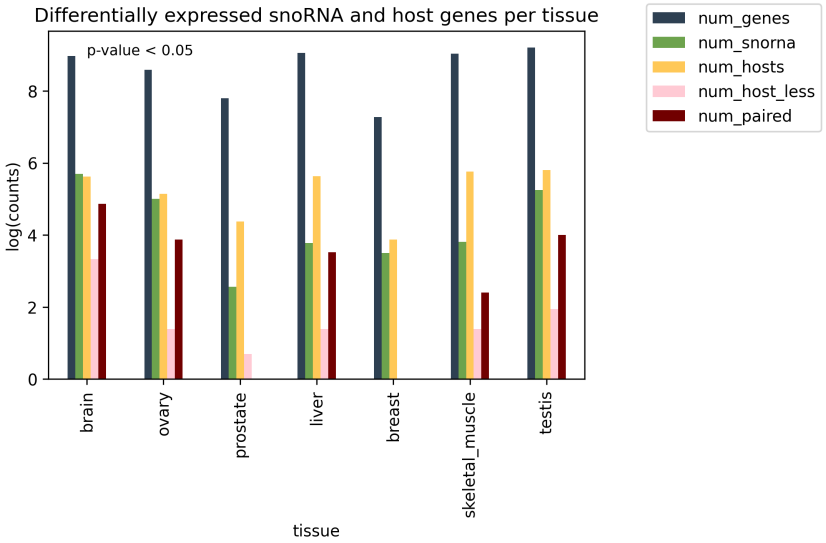


Figure 10: For each tissue is reported the number of significantly (p-value < 0.05) differentially expressed: genes, snoRNAs, host genes, snoRNAs without their host in the list (host-less), and snoRNAs paired with their host gene.

instance using Pearson Correlation), but it is unlikely that the results would be much better. We tried to not remove the duplicates (meaning considering the host as many times as one of its snoRNA appears in the list) but as expected the results are very skewed since some lnc-genes are hosts of up to 80 snoRNAs.

5 Conclusion

SnoRNAs are implicated in a lot of functions in eukaryotic cells, even in crucial ones (i.e., rRNA modifications), yet their abundance pattern across tissues and their alternative function or dis-regulation effect in diseases has not been studied in depth. In this project we started from already existing RNA-seq data from seven healthy human tissues to firstly investigate the relative abundance of snoRNAs and of hosts genes for each tissue, from which we demonstrated that brain, ovary and testis are the tissues with the highest snoRNAs expression compared with the other tissues. From these results we derived the pathway enrichment analysis, performed as over-representation analysis, from which we confirmed the ubiquitous influence of snoRNAs in processes from the nucleoar to the cytoplasmic environment, including fundamental roles such as: translation, RNA binding, and protein binding. Our second goal was to perform snoRNAs expansions to unveil new possible connection to pathways or even single genes, to build a network of possible interactions that could be useful in the study of diseases, such as tumor or neuro-degenerative diseases. From the network expansions trough gene@home project we obtained networks of possible interactions, we performed validation and we looked for snoRNA-canonical target interactions. The validation occurred by looking for canonical interactions (coming from literature and experimental proofs): no network levelled up to a sufficient score to be considered reliable for in depth analyses. The snoRNA-hosts interactions analyses produces a really low number of correlated pairs.

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] Ritchie M, E Phipson, B Wu, D Hu, Y Law, Shi W, Smyth G K . limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Research*, 2015.

[4] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 2014.

[5] 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.

[6] Seal RL, Braschi B, Gray K, Jones TEM, Tweedie S, Haim-Vilmovsky L, Bruford EA. Genenames.org: the HGNC resources in 2023. *Nucleic Acids*, 2022.

[7] Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W . Biomart and bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 2005.

[8] Enrico Blanzieri et al. A computing system for discovering causal relationships among human genes to improve drug repositioning, 2021.

[9] Kalisch, Markus and Buehlmann, Peter. Estimating high-dimensional directed acyclic graphs with the pc-algorithm, 2005.

[10] Francesco Asnicar et al. Nes2ra: Network expansion by stratified variable subsetting and ranking aggregation, 2016.

[11] Francesco Asnicar et al. Onegene: Regulatory gene network expansion via distributed volunteer computing on boinc, 2019.

[12] Lestrade L., Weber M.J. . snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res*, 2006.

[13] Krogh N., Jansson M.D., Hafner S.J., Tehler D., Birkedal U., Christensen-Dalsgaard M., Lund A.H., Nielsen H.. Profiling of 2-O-Me in human rRNA reveals a subset of fractionally modified positions and provides evidence for ribosome heterogeneity. *Nucleic Acids Res.*, 2016.

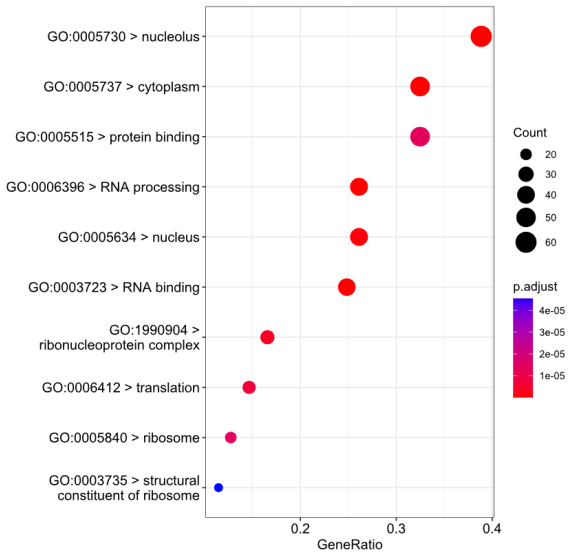
[14] Liang J, Wen J, Huang Z, Chen XP, Zhang BX, Chu L. Small Nucleolar RNAs: Insight Into Their Function in Cancer. *Front Oncol.*, 2019.

Attachment 1

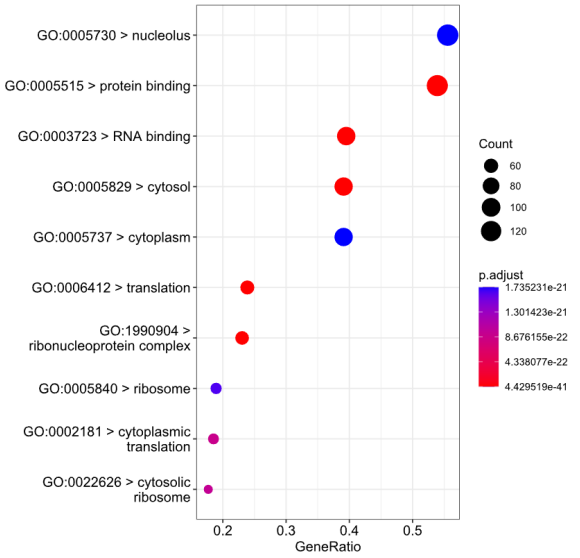
Post filter	Network	Threshold	Total canonical	Found canonical	Fraction canonical
(1, 0.2)	RN	0	621	0	0
(1, 0.2)	SN	0	621	9	0.01
(1, 0.2)	SN	0.2	621	5	0.01
(1, 0.2)	SN	0.5	621	4	0.01
(1, 0.2)	SN	0.9	621	1	0
(1, 0.2)	SRN	0	621	9	0.01
(1, 0.2)	SRN	0.2	621	5	0.01
(1, 0.2)	SRN	0.5	621	5	0.01
(1, 0.2)	SRN	0.9	621	1	0
(1, 0.2)	ARN	0	621	0	0
(1, 0.2)	ASN	0	621	61	0.1
(1, 0.2)	ASN	0.2	621	16	0.03
(1, 0.2)	ASN	0.5	621	8	0.01
(1, 0.2)	ASN	0.9	621	1	0
(1, 0.2)	ASRN	0	621	61	0.1
(1, 0.2)	ASRN	0.2	621	16	0.03
(1, 0.2)	ASRN	0.5	621	8	0.01
(1, 0.2)	ASRN	0.9	621	1	0
(5, 0.4)	RN	0	481	0	0
(5, 0.4)	SN	0	481	8	0.02
(5, 0.4)	SN	0.2	481	5	0.01
(5, 0.4)	SN	0.5	481	4	0.01
(5, 0.4)	SN	0.9	481	1	0
(5, 0.4)	SRN	0	481	8	0.02
(5, 0.4)	SRN	0.2	481	5	0.01
(5, 0.4)	SRN	0.5	481	5	0.01
(5, 0.4)	SRN	0.9	481	1	0
(5, 0.4)	ARN	0	481	0	0
(5, 0.4)	ASN	0	481	57	0.12
(5, 0.4)	ASN	0.2	481	16	0.03
(5, 0.4)	ASN	0.5	481	8	0.02
(5, 0.4)	ASN	0.9	481	1	0
(5, 0.4)	ASRN	0	481	57	0.12
(5, 0.4)	ASRN	0.2	481	16	0.03
(5, 0.4)	ASRN	0.5	481	8	0.02
(5, 0.4)	ASRN	0.9	481	1	0

Post filter	Network	Threshold	Total canonical	Found canonical	Fraction canonical
(10, 0.4)	RN	0	463	0	0
(10, 0.4)	SN	0	463	8	0.02
(10, 0.4)	SN	0.2	463	5	0.01
(10, 0.4)	SN	0.5	463	4	0.01
(10, 0.4)	SN	0.9	463	1	0
(10, 0.4)	SRN	0	463	8	0.02
(10, 0.4)	SRN	0.2	463	5	0.01
(10, 0.4)	SRN	0.5	463	5	0.01
(10, 0.4)	SRN	0.9	463	1	0
(10, 0.4)	ARN	0	463	0	0
(10, 0.4)	ASN	0	463	57	0.12
(10, 0.4)	ASN	0.2	463	16	0.03
(10, 0.4)	ASN	0.5	463	8	0.02
(10, 0.4)	ASN	0.9	463	1	0
(10, 0.4)	ASRN	0	463	57	0.12
(10, 0.4)	ASRN	0.2	463	16	0.03
(10, 0.4)	ASRN	0.5	463	8	0.02
(10, 0.4)	ASRN	0.9	463	1	0

Table 4: This table reports for each network - for a set of thresholds and a set of filtering parameters on the TPM table - the total canonical interaction found, the overall canonical interactions present in the dataset and the fraction of the detected ones.
The first column represent the filtering criterion (1, 0.2) means that the filtering occurred with 1 TPM in at least 20% of the samples, (5, 0.4) % TPM in at least 40% of samples and (10, 0.4) stands for 10 TPM in at least 40% of the samples.
RN - riboproteins network expansion, SN - snoRNAs network expansion, SRN - snoRNA and riboproteins network expansion, ASN - aggregation of snoRNAs single expansions network, ARN - aggregation of riboproteins single expansion network, ASRN - aggregated snoRNAs and riboproteins single expansion network.
Among a total of 396 total snoRNA present in the expansions, only few of them demonstrated correlation with their host.



(a) Enriched terms plot depicts the enrichment scores (p-values) and gene ratio through color and dimension of the dots. The most enriched terms are reported on the y-axis. This plot, together with the next ones, do not highlight some specific pathway, rather they enriched terms are generic (i.e., nucleolus, cytoplasm, protein binding, ...)



(b) Enrichment plot for snoRNAs host genes derived from the DE analysis. Here is reported only the plot for brain-vs-all run. The enriched terms cover a very complex and fundamental part of protein expression, from nucleolar to cytoplasmic environment.

Figure 11