CAR-NET R Shiny Tutorial

Contents

1 Overview

CAR-NET is a RShiny-based application with graphical user interface (GUI) for inferring non-coding RNA regulatory network from transcriptomic data and curated database. This application took input of both ncRNA and gene expression data generated from either bulk or single-cell RNA-seq and incorporated a list of downloaded curated databases for users to choose based on the type of ncRNA and the condition/disease of interest. It includes a preparatory step to preprocess the expression data following the standard pipeline and major analytical steps to construct the ncRNA regulatory network, identify the differential network, detect network modules, and perform pathway analysis to facilitate the biological

interpretation of the network findings. In addition, it provides visualization of the network/-modules and downloadable graphical and tabular outputs.

In this tutorial, we will go through the installation and usage of CAR-NET step by step using real data examples. The CAR-NET Shiny software and associated R package are available at https://github.com/kehongjie/CAR-NET. This tutorial can be found at xxx.

2 Preliminaries

2.1 Citing CAR-NET

Please cite this paper if you use CAR-NET:

> Ke, H., Ye, Z., Feng, L., Xu, Z., Li, E. & Ma, T. (2025+). Inferring noncoding-RNA gene regulatory network from transcriptomic data and curated database. bioRxiv

2.2 How to start CAR-NET

Requirement:

- R >= 4.0.0
- Shiny >= 1.0.0

To install the software:

1. Install the dependency packages following the instruction at https://github.com/kehongjie/CAR-NET.
2. Download the CAR-NET Shiny project from the GitHub repository by clicking on "code > Download ZIP" and extract "CAR-NET-main" to a local folder.

To start the software:

1. Open the "RunShiny.R" script in the "CAR-NET" folder.
2. Set the working directory of R to the directory which contains the Shiny project folder "CAR-NET-main".
3. Run the code `shiny::runApp('CAR-NET-main', port=9987, launch.browser=T)`.

3 Preparation

In this section, we will introduce how to prepare the data input for CAR-NET, which includes a ncRNA expression dataset, a gene expression dataset and a clinical dataset (optional).

3.1 Noncoding RNA expression data

The ncRNA expression matrix should be prepared as a comma-separated ".csv" file. Each row should be a ncRNA and each column should be a sample. The first column (row names) needs to be the ncRNA name (or ID) and the rest of columns are the expression data for samples whose IDs are annotated as column names. CAR-NET can take both microarray and RNA-seq (bulk or single cell) input for ncRNA expression data. It can be one specific type of ncRNA (e.g., miRNA) or a mixtures of multiple types.

3.2 Gene expression data

The coding gene expression matrix should be prepared as a comma-separated ".csv" file. Similar to the ncRNA dat, each row should be a ncRNA and each column should be a sample. The first column (row names) needs to be the gene name (or ID) and the rest of columns are the expression data for samples whose IDs are annotated as column names. For the gene name, please make sure they are using either HGNC Symbol or ENSG ID. CAR-NET can take both microarray and RNA-seq (bulk or single cell) input for gene expression data.

3.3 Optional clinical data

If you wish to run differential regulation analysis, you will also need to upload a clinical data in a comma-separated ".csv" file. Each row in the clinical should be a clinical variable indicating certain biological conditions (e.g., cancer stage) and these clinical variables need to be categorical.

4 Data uploading and preprocessing

In this section, we will introduce how to upload and preprocess the data.

4.1 Noncoding RNA

To start with, you can click the "Browse" button and upload their ncRNA data from your local directory. We provide a example dataset from TCGA for demonstration purpose, which can be found in the `"/CAR-NET-main/data"` directory. After uploading, there are a few things that you need to specify for preprocessing:

1. Type of ncRNA (Figure xx): which types of ncRNA you are working on. Options are either "miRNA" (micro RNA), "lncRNA" (long noncoding RNA), "other" or "mixed types". If you choose "lncRNA", we will perform annotation for your long noncoding

RNAs based on LNCipedia reference database (Volders et al., 2019) and you will need to further specify whether the naming format of you lncRNA data is “LNCipedia”, “ENSG”, “HSALN” or “HGNC”. After annotation, some of your lncRNAs might be renamed but the order in the dataset stays the same.

2. Platform (Figure xx): what platform dose your lncRNA data come from. The options are “bulk RNA-seq/microarray” and ”scRNA-seq” (single cell). This will leads to different recommended cut-off values during the preprocessing.

3. Mean cut-off (Figure xx): the cut-off values for the mean of ncRNA expression level. NcRNAs whose average expression level lower than this value will be removed. Default value is 1 for “bulk RNA-seq/microarray” and 0.1 for ”scRNA-seq” but you can determine based on the nature of your own data.

4. Variance cut-off (Figure xx): the cut-off values for the variance of ncRNA expression level. For example, if this value is set to 0.5, then only the top 50% ncRNAs with the largest variance will be kept after the preprocessing. If this value is 1, then this step will not remove any ncRNAs. We recommend a value smaller than 1 to remove some non-informative ncRNAs if the number of ncRNAs is large (e.g., 5000). Default is 1.

5. Has the data been log2 transformed (Figure xx): if no, then we will perform a log2 transformation for the ncRNA data.

After these, you can click the “Submit” button and a preview of your preprocessed ncRNA expression data will pop out on the right panel (see Figure xx).

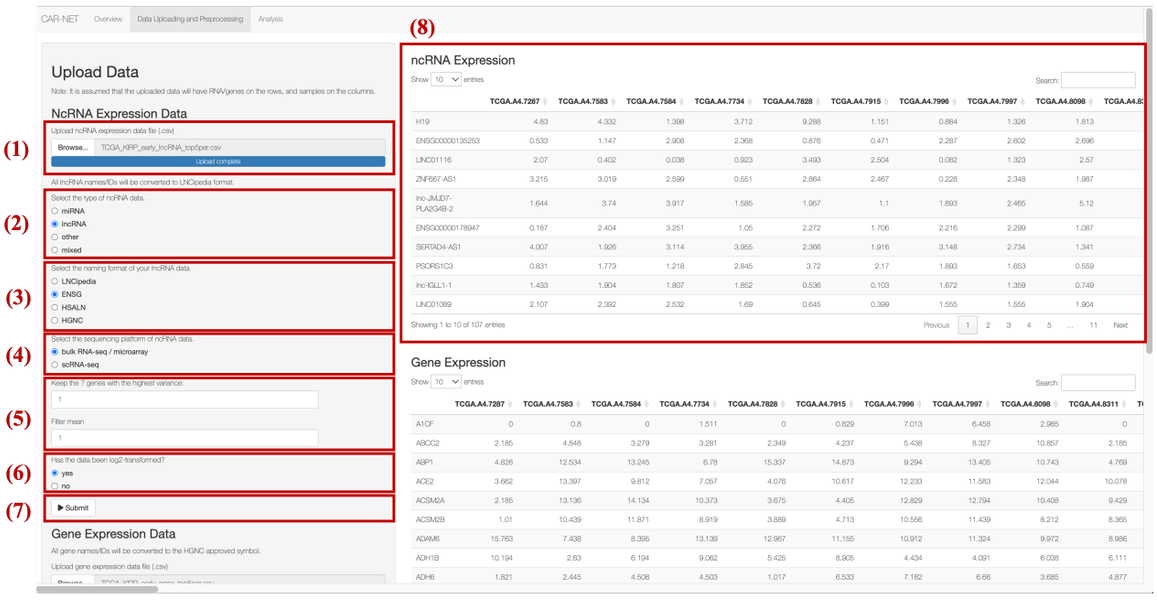

Figure 1: Uploading and preprocessing of data.

4.2 Gene

The uploading and preprocessing for gene expression data are very similar to the ones for ncRNA. After uploading, you need to specify (and refer to the previous subsection for more details):

1. Naming format of genes (Figure xx): what naming system is used for your gene expression data. It can be either "HGNC Symbol" or "ENSG ID"".
2. Platform (Figure xx): what platform dose your gene expression data come from.
3. Mean cut-off (Figure xx): the cut-off values for the mean of gene expression level.
4. Variance cut-off (Figure xx): the cut-off values for the variance of gene expression level.
5. Has the data been log2 transformed (Figure xx): if no, then we will perform a log2 transformation for the gene expression data.

Again, you can click the "Submit" button and a preview of your preprocessed gene expression data will pop out on the right panel (see Figure xx).

5 Analysis

5.1 Running CAR-NET

Tunning parameter alpha

The most important tunning parameter in CAR-NET when constructing the network is called alpha, which controls the sparsity level of overall network. It is a value between 0 and 1 and a smaller value will result in a more sparse network (similar to p-value). We provide two options in terms of choose the value of this alpha: "auto-tune" and "choose by my self" (see Figure xx).

If you use "auto-tune", an extra algorithm will be ran and an optimal alpha value will be automatically chosen and used. Note that this process will take some extra time. Otherwise, if you decide to "choose by my self", then you will need to specify a alpha value (default is 0.00001). Be careful with using very large alpha value (e.g., 0.01) as this will greatly increase the computation time.

Select external databases to be used as prior

CAR-NET provides the option of prioritizing those nodes and edges appear in the external curated database during the construction of Bayesian network. You can choose which external databases you would like to include (see Figure xx). Currently, miRNA can use miRCancer (Xie et al., 2013) and miRTarBase (Huang et al., 2022) database while lncRNA

can use EVLncRNAs (Zhou et al., 2021), LncRNA2Target (Cheng et al., 2019) and LncTarD (Zhao et al., 2023) databases.

Pathway analysis for genes in the identified network

After constructing the network, if you wish to perform pathway enrichment analysis, our software also provide this option. You can pick up which pathway datasets you want to use for pathway analysis (see Figure xx). Currently, four common pathway datasets are supported: GO, KEGG, Reactome and Biocarta.

Output

After you click the "Run CAR-NET" button, the network will be constructed and you can find a summary table of the network from the first tab in the right panel (see Figure xx). The summary table will tell you how many ncRNAs, genes, and ncRNA-gene pairs are included in the network. In addition, a table will be presented showing which ncRNA-gene pairs (level-1 edges) and gene-gene pairs (level-2 edges) are connected in the network.

If you would like to run the pathway enrichment analysis, click the "Do pathway analysis for all genes in the network"". A enrichment bubble plot of top 10 most enriched pathways will be shown (see Figure xx).

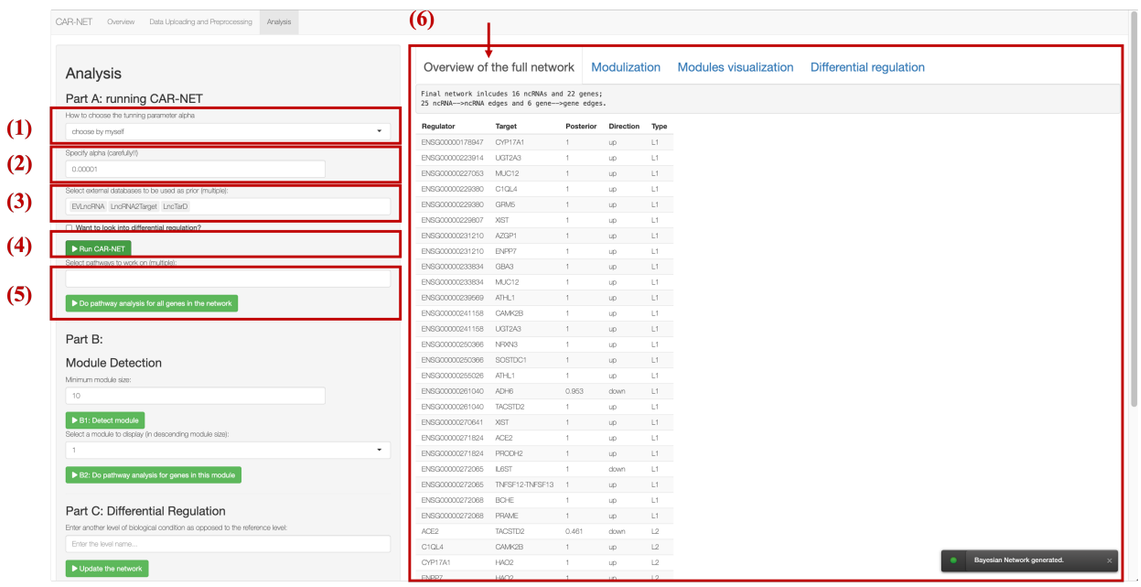

Figure 2: Run CAR-NET and overview of the full network.

5.2 Module detection

Modulization

Typically, the network that our algorithm identifies could still be relatively large and hard to visualize or interpret. For better interpretation and more in-depth investigation of the network, CAR-NET includes a implementation of a directed Louvain algorithm (Blondel

et al., 2008) to partition the identified network into smaller modules for subnetwork analysis.

You will need to specify what is the “minimum module size”” (see Figure xx) for a module to be considered. We recommend using numbers around 10 as very small modules are usually of less interest. After clicking the ”Detect module” button, a heatmap will show up in the right panel, under the tab of ”Modulization” (see Figure xx). This heatmap shows how ncRNAs and genes within the same module are connected (and rarely connected between modules). Each row is a ncRNA and each column is a gene. The color scale indicates the posterior for this ncRNA-gene edge. Each block is a module and they are sorted by module size. Due to the page limit, we only shows top modules here.

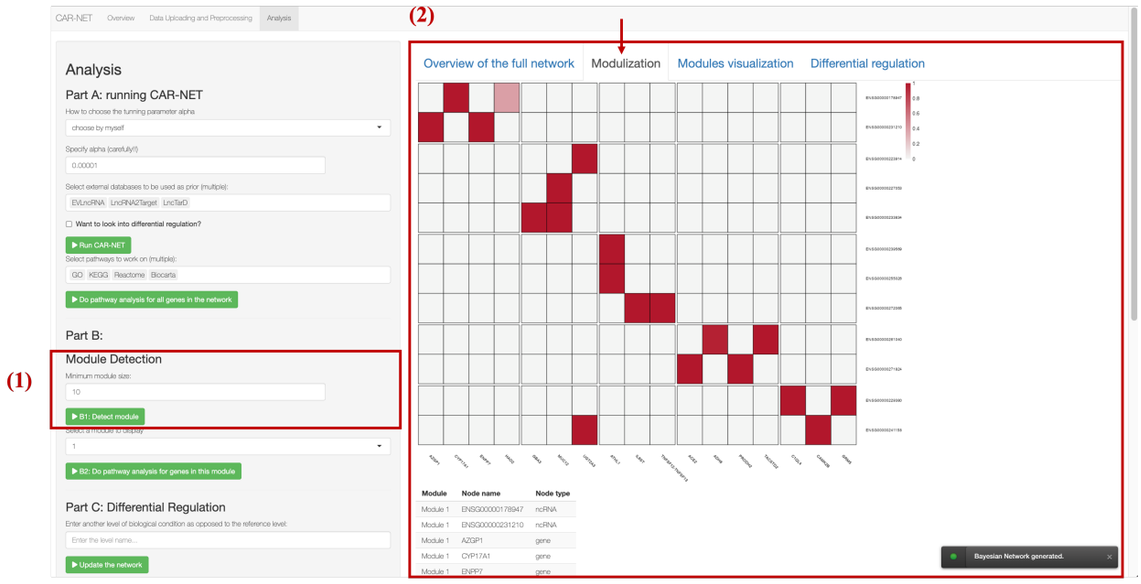

Figure 3: Detect modules.

Visualize individual module

If you want to look into each module and visualize the network, you can move to the tab of ”Modules visualization” in the right panel (see Figure xx). A network visualization will be presented for one specific module, with yellow nodes being ncRNAs and blue nodes being genes. The width of edges indicates the posterior probability of corresponding edges. To navigate between different modules, use the drop-down function as in Figure xx, where ”1” indicates the largest module and ”2” indicates the second largest module, etc.

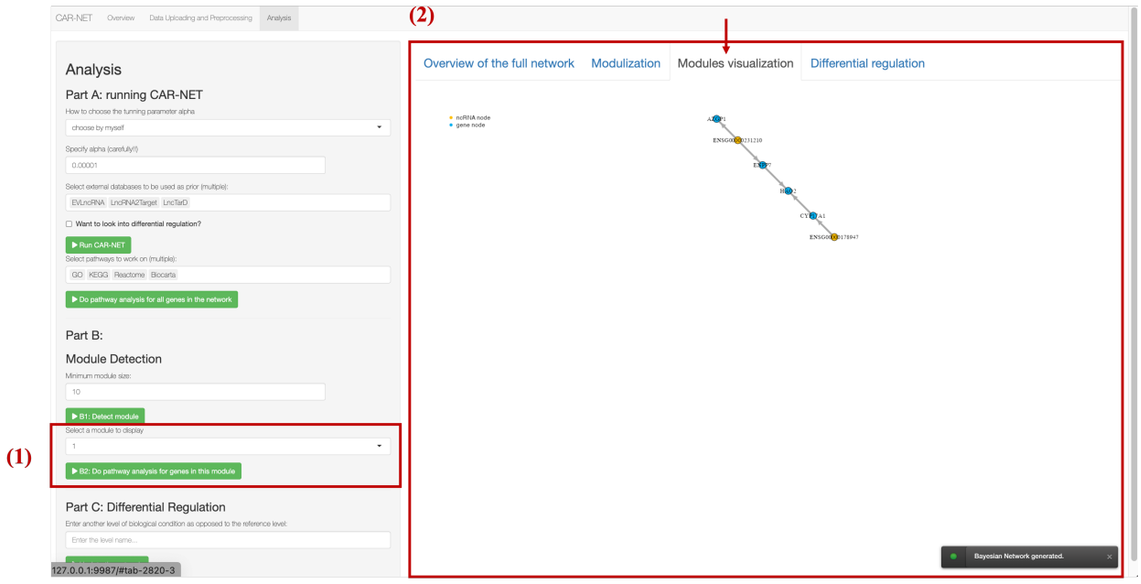

Figure 4: Visualization of individual module.

References

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

Cheng, L., Wang, P., Tian, R., Wang, S., Guo, Q., Luo, M., Zhou, W., Liu, G., Jiang, H., and Jiang, Q. (2019). Lncrna2target v2. 0: a comprehensive database for target genes of lncrnas in human and mouse. *Nucleic acids research*, 47(D1):D140–D144.

Huang, H.-Y., Lin, Y.-C.-D., Cui, S., Huang, Y., Tang, Y., Xu, J., Bao, J., Li, Y., Wen, J., Zuo, H., et al. (2022). mirtarbase update 2022: an informative resource for experimentally validated mirna–target interactions. *Nucleic acids research*, 50(D1):D222–D230.

Volders, P.-J., Anckaert, J., Verheggen, K., Nuytens, J., Martens, L., Mestdagh, P., and Vandesompele, J. (2019). Lncipedia 5: towards a reference set of human long non-coding rnas. *Nucleic acids research*, 47(D1):D135–D139.

Xie, B., Ding, Q., Han, H., and Wu, D. (2013). mircancer: a microrna–cancer association database constructed by text mining on literature. *Bioinformatics*, 29(5):638–644.

Zhao, H., Yin, X., Xu, H., Liu, K., Liu, W., Wang, L., Zhang, C., Bo, L., Lan, X., Lin, S., et al. (2023). Lnctard 2.0: an updated comprehensive database for experimentally-supported functional lncrna–target regulations in human diseases. *Nucleic Acids Research*, 51(D1):D199–D207.

Zhou, B., Ji, B., Liu, K., Hu, G., Wang, F., Chen, Q., Yu, R., Huang, P., Ren, J., Guo, C., et al. (2021). Evlncrnas 2.0: an updated database of manually curated functional long non-coding rnas validated by low-throughput experiments. *Nucleic Acids Research*, 49(D1):D86–D91.