Input files

For running `SWIMmeR` with the sample project "TCGA" and the sample dataset "brca", the following files are required as input (Fig. 1).

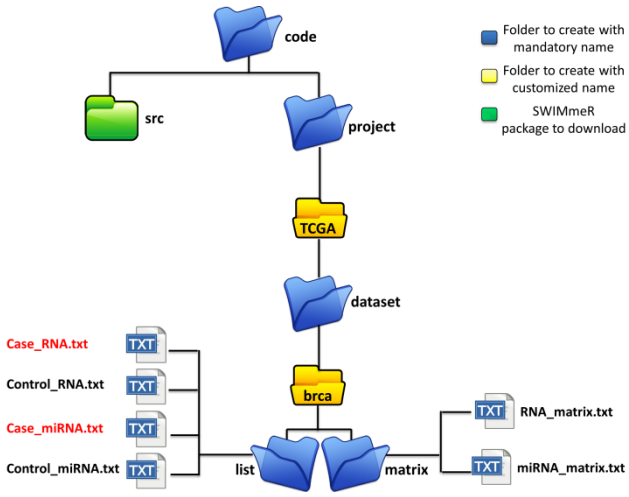

Fig. 1. Input architecture.

In *matrix* folder:

- *matrice__brca_RNASeq.txt*: a text file providing protein-coding and non-coding RNAs abundance (RNA-sequencing data matrix). The rows are the RNAs and columns are samples. The file includes row headers (TCGA gene identifiers)[1] and column headers (TCGA barcodes).
- *matrice__brca_miRNASeq.txt:* a text file providing miRNAs abundance (miRNA-sequencing data matrix). The rows are miRNAs and columns are samples. The file includes row headers (TCGA miRNA identifiers)[2] and column headers (TCGA barcodes).

In *list* folder:

- *Lista__RNASeq_Tumor__brca__4wayData.txt*: a text file that lists the barcodes concerning the condition A (tumor) for RNASeq. Each barcode exactly matches one column header of the *matrice__brca_RNASeq.txt* file [3].
- *Lista__RNASeq_Normal__brca__4wayData.txt*: a text file that lists the barcodes concerning the condition B (matched normal) for RNASeq. Each barcode exactly matches one column header of the *matrice__brca_RNASeq.txt* file.
- *Lista__miRNASeq_Tumor__brca__4wayData.txt*: a text that lists the barcodes concerning the condition A (tumor) for miRNASeq. Each barcode exactly matches one column header of the *matrice__brca_miRNASeq.txt* file.

[1] These are composite gene identifiers provided by TCGA and have the format: HGNC Gene Symbol | Entrez Gene Id. The HGNC and Entrez Gene identifiers relate to the HUGO gene nomenclature committee (http://www.genenames.org/) and the NCBI Gene (http://www.ncbi.nlm.nih.gov/gene) databases, respectively.

[2] Strictly speaking, these TCGA identifiers refer to the miRNA precursors and follow the nomenclature provided by miRBase (http://www.mirbase.org/).

[3] Note that, in this example, the analysis is restricted to all common patients between tumor and matched normal tissues of both RNA-sequencing and miRNA-sequencing data matrices.

- *Lista__miRNASeq_Normal__brca__4wayData.txt*: a text file that lists the barcodes concerning the condition B (matched normal) for miRNASeq. Each barcode exactly matches one column header of the *matrice__brca_miRNASeq.txt* file.

Output files

As output, `SWIMmeR` creates a R file (*parameters.RData*) storing all parameters set during the analysis and two folders called *filtering* and *switch* (Fig. 2).

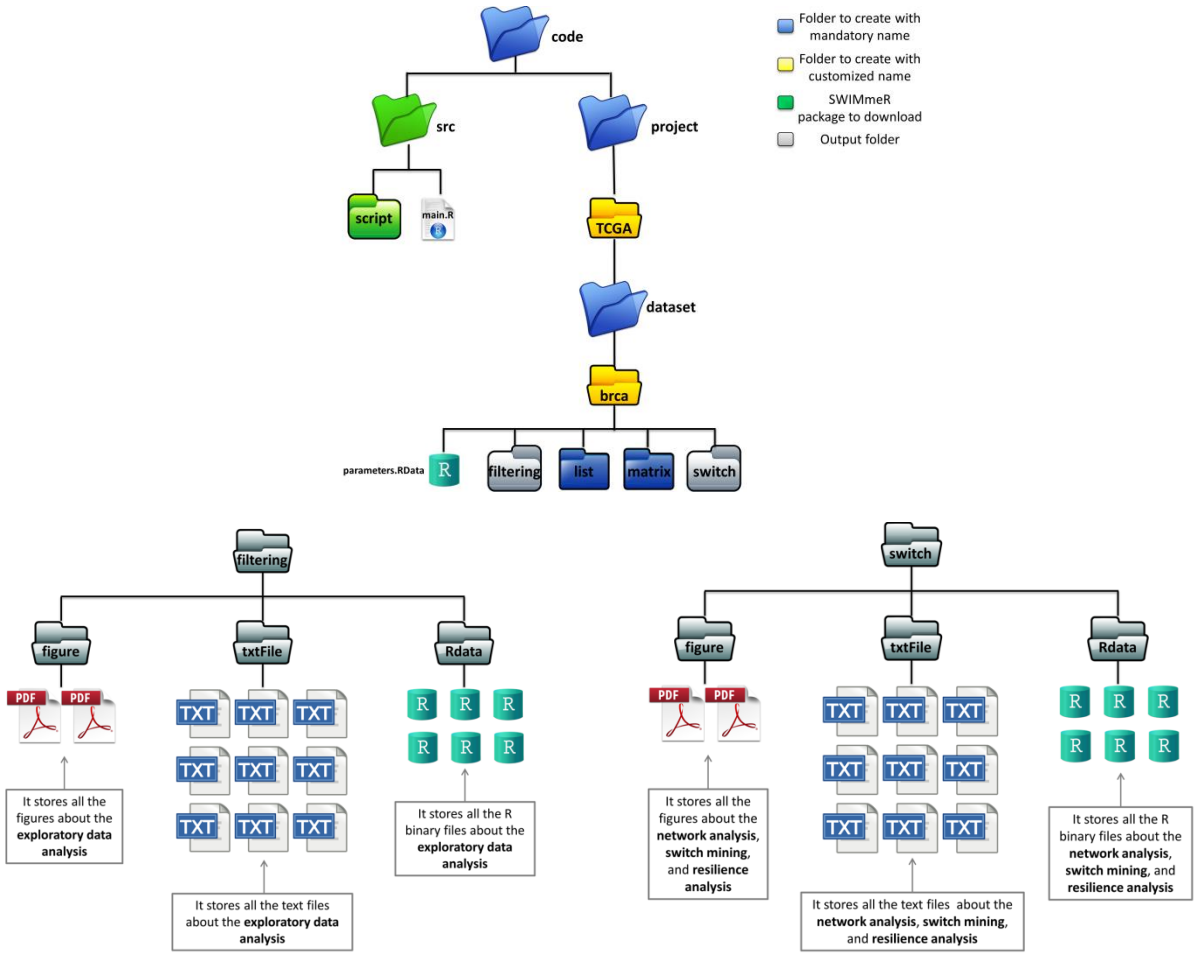

Fig. 2. Output architecture.

filtering folder

1. *figure* folder contains:
    - *heatmap_DE_RNA.pdf*: it depicts the dendrogram and the heat map of differentially expressed RNAs. The differential expressed protein-coding and non-coding RNAs are clustered according to rows (genes) and columns (samples) of the RNA-sequencing data matrix (biclustering). The colors represent different expression levels that increase from blue to yellow.
    - *heatmap_DE_miRNA.pdf*: it depicts the dendrogram and the heat map of differentially expressed miRNAs. The differential expressed miRNAs are clustered according to rows (miRNAs) and

columns (samples) of the RNA-sequencing data matrix (biclustering). The colors represent different expression levels that increase from blue to yellow.

- *VolcanoPlot.pdf*: it depicts the volcano plot for RNAs. The x-axis represents the fold-change ratio ($\log_2$ of the fold-change) that is the ratio between the average tumor expression value and the average normal expression value. The y-axis represents the adjusted p-values ($-\log_{10}$ of the p-values) of the Student's t-test. The vertical dashed red lines and the horizontal ones correspond to the selected thresholds for the fold-change and the p-values, respectively.
- *VolcanoPlot_miRNAs.pdf*: it depicts the volcano plot for miRNAs.

2. *txtFile* folder contains:

- *stat_dataFiltered.txt*: it includes the statistics computed on the RNA-sequencing data only for the differentially expressed RNAs
- *stat_dataORIG.txt*: it includes the statistics computed on the RNA-sequencing data for all the RNAs
- *stat_dataFiltered_miRNA.txt*: it includes the statistics computed on the miRNA-sequencing data only for the differentially expressed miRNAs
- *stat_dataORIG_miRNA.txt*: it includes the statistics computed on the miRNA-sequencing data for all the miRNAs

The columns of the above files contains:

1) the names of the gene
2) *IQR:* IQR value
3) *perc_zeros*: percentage of zeros
4) *logFC:* the logarithm of the ratio between the average of cancer expression values and the average of the normal expression values
5) *pval*: p-value of the statistical hypothesis Student's t-test
6) *pval_adj*: adjusted p-value (default: by using FDR method)

3. *Rdata* folder contains the files of *txtFile* folder in Rdata format

switch folder

1. *figure* folder contains:

- *APCC_distribution.pdf*: it depicts the APCC distribution of the correlation network. The curve represents the estimated probability density of the APCC for each hub of the correlation network. The APCC distribution appears to be trimodal and the three peaks correspond to (from right to left): i. party hubs, which are highly correlated with the expression of their interaction partners; ii. date hubs which show moderate co-expression with their interaction partners; iii. fight-club hubs which show average negative correlation with their interaction partners. The x-axis represents the APCC and the y-axis represents the probability density.
- *connectedComponent.pdf:* it depicted the connectivity plot of the correlation network. The x-axis represents the Pearson correlation threshold varying in the chosen range, while the y-axis represents the fraction of nodes populating the largest component. The dashed red lines correspond to the selected threshold. Note that y=1 means that all nodes fall in the largest component and thus the network is fully connected; otherwise more components exist. A reasonable choice for the correlation threshold should be the largest one for which the fraction of nodes of the largest connected component is equal to 1.

- *HeatCartography.pdf*: it depicts the Heat Cartography map. The plane is identified by two parameters: z_g (within-module degree) and K_π (clusterphobic coefficient) and it is divided into seven regions each defining a specific node role (R1-R7). High z_g values correspond to nodes that are hubs within their module (local hubs), whereas low z_g values correspond to nodes with few connections within their module (non-hubs within their communities, but they could be hubs in the network). Each node is colored according to its APCC value. Specifically, blue nodes are the fight-club hubs (i.e., showing an average negative correlation in expression with their interaction partners), and among them the ones falling in the region R4 are the switch genes.
- *heatmap_switch.pdf*: it depicts the dendrogram and heatmap of switch genes. The switch genes are clustered according to rows (switch genes) and columns (samples) of the switch genes expression data (biclustering). The colors represent different expression levels that increase from blue to yellow.
- *heatmap_switch_miRNA.pdf*: it depicts the dendrogram and heatmap of miRNAs included in the switch genes list.
- *removalNodes.pdf*: it depicts the robustness plot for the correlation network. The x-axis represents the cumulative fraction of removed nodes, while the y-axis represents the average shortest path. The shortest path between two nodes is the minimum number of consecutive edges connecting them. Each curve corresponds to the variation of the average shortest path of the correlation network as function of the removal of nodes specified by the colors of each curve.

2. *txtFile* folder contains:

- *attribute.txt*: it reports the features of all nodes in the heat cartography map. The columns of the file are the following ones:

  1) *node*: the names of the nodes in the heat cartography map
  2) *Hub*: it specifies if the nodes is (or is not) a local hub within their community
  3) *Region*: it specifies the region of the heat cartography to which the node belongs and corresponds to its universal role in the network
  4) *Type*: it specifies for each node its universal role
  5) *Total_Degree*: it specifies the node degree (i.e., the number of incoming and outgoing edges of each node)
  6) *Internal_Degree*: it specifies the node degree inside its own community
  7) *APCC*: it specifies the average of the Pearson correlation coefficients between the expression profiles of a node and those of its interaction partners
  8) *Hub_classification*: it specifies if the node is either a date or party or fight-club hub or if it is not a local hub in its community
  9) *P*: it specifies the value of the *clusterphobic coefficient* parameter
  10) *z*: it specifies the value of the *within-module degree* parameter
  11) *Cluster_ID:* it specifies the belonging cluster of each node
  12) *IQR:* it specifies the IQR value
  13) *perc_zeros:* it specifies the percentage of zeros
  14) *logFC:* it specifies the log-ratio value of the fold-change
  15) *pval:* it specifies the p-value of the statistical hypothesis Student's t-test
  16) *pval_adj:* it specifies the adjusted pvalue (default: FDR)

- *attribute-switch.txt*: it reports only the attributes of the switch genes with the same coulmns discussed above (*attribute.txt*)

- *CorrelationNetwork.txt*: it contains the correlation network with the following columns:
  1) *Source*: source node of the correlation network
  2) *Target*: target node of the correlation network
  3) *Correlation:* Pearson correlation coefficient between source and target nodes
  4) *pval:* pvalue of the correlation
  5) *pval_adj*: adjusted pvalue of the correlation
- *idx.txt*: it contains for each node of the correlation network the name and its belonging cluster
- *CartographyNetwork.txt*: it contains the nodes appearing in the heat cartography map, with the following columns:
  1) *Source*: source node of the heat cartography map
  2) *Target*: target node of the heat cartography map
  3) *Correlation:* Pearson correlation coefficient between source and target nodes
  4) *pval:* pvalue of the correlation
  5) *pval_adj*: adjusted pvalue of the correlation

Note that in general not all nodes of the correlation network appear in the heat cartography.

- *fc_switch.txt*: it contains the switch genes along with their statistics. The columns are:
  1) the name of the switch gene
  2) *IQR:* IQR value
  3) *perc_zeros*: percentage of zeros
  4) *logFC:* the logarithm of the ratio between the average of cancer expression values and the average of the normal expression values
  5) *pval*: p-value of the statistical hypothesis Student's t-test
  6) *pval_adj*: adjusted p-value (default: by using FDR method)
- *switch.txt*: it contains the names of the switch genes
- *nn_neg_switch.txt*: it is a matrix where the first column contains the names of the switch genes and each row contains the names of the nodes that are connected and negatively correlated with the corresponding switch gene
- *nn_pos_switch.txt*: it is a matrix where the first column contains the names of the switch genes and each row contains the names of the nodes that are connected and positively correlated with the corresponding switch gene
- *corr_nn_pos_switch.txt*: it contains the Pearson correlation coefficients between switch genes (reported in the first column) and their positive correlated linked nodes. The second column of this file corresponds to the average of each row.
- *corr_nn_neg_switch.txt*: it contains the Pearson correlation coefficients between switch genes (reported in the first column) and their negative correlated linked nodes. The second column of this file corresponds to the average of each row.
- *cluster_nn_pos_switch.txt*: it contains the belonging clusters of the switch genes (reported in the first column of this file) and their positive correlated linked nodes (rows)
- *cluster_nn_neg_switch.txt*: it contains the belonging clusters of the switch genes (reported in the first column of this file) and their negative correlated linked nodes (rows)

3. *Rdata* folder contains:

   - *CorrelationNetwork.RData*: it contains a data.frame with the correlation network (the same as in *CorrelationNetwork.txt*)
   - *weighted_adjMatrix.RData*: it contains the weighted adjacency matrix of the correlation network, a square and symmetric matrix, whose size is equal to the number of network nodes, and whose elements a_{ij} are equal to the correlation value if there exists an edge between the *i* and *j* node, equal to zero otherwise
   - *CartographyNetwork.RData*: it contains a data.frame with the correlation network (the same as in *CorrelationNetwork.txt*)

   If the robustness of the network has been evaluated, the following additional files are provided:

   - *removalDateHubs.RData*
   - *removalFightClub.RData*
   - *removalHubs.RData*
   - *removalNonSwitch.RData*
   - *removalPartyHubs.RData*
   - *removalSwitch.RData*
   - *removalRandomNodes.RData*

   each one including a data.frame with:

   1) *frac*: the fraction of the removed nodes (x-axes of removalNode.pdf)
   2) *mean_sp*: the averaged shortest paths (y-axes of removalNode.pdf)