# miRHiC: enrichment analysis of miRNA targets in Hierarchical gene Co-expression signatures

## 1. Overview

miRHiC was proposed to infer the perturbed miRNA regulatory networks in cancer by incorporating the hierarchically organized co-expression information of the differentially expressed genes: firstly, the hierarchical gene co-expression signatures were established by clustering the differentially expressed genes based on pairwise gene co-expression correlations; then the miRNA target gene set enrichment was analyzed across the hierarchical co-expression signatures; and finally, a permutation test was used to estimate the statistical significance of the enrichment (Figure 1). **This document provides a detailed introduction for miRHiC**, including the data pre-processing steps, the core algorithm and the software implementation.

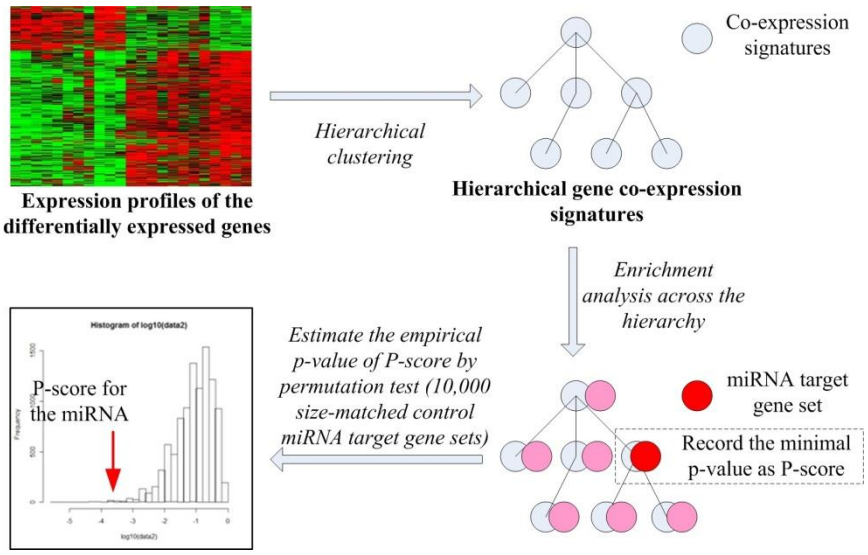


**Figure 1**. The flowchart of miRHiC. In the first step, the differentially expressed genes were clustered as hierarchical gene co-expression signatures; then, the most significant enrichment of the miRNA target gene set was found across the hierarchical signatures; and finally, a permutation test was used to estimate the empirical p-value of the enrichment.

## 2. Get the hierarchical gene co-expression signatures

### 2.1 Data pre-processing: extract the gene expression profiles of the significantly differentially expressed genes

Only the differentially expressed genes will be used to establish the co-expression signatures. Any software, such as LIMMA or SAM, can be used to identify the differentially expressed genes. Or, simply use t-test (paired) to do this step. The p-values should be multiple testing adjusted (BH correction). The gene expression profiles usually show large changes between cancer and adjacent normal

tissues. The cutoffs should be set relatively stringent to select the significantly differentially expressed genes (we usually use adjusted p-value < 0.0001 as cutoff). Then, extract the expression values (or fold changes for matched samples) of the identified differentially expressed genes for the following hierarchical gene co-expression clustering. The miRHiC package does not integrate the analysis for the significantly differentially expressed genes. It should be done by users.

**2.2 Average linkage hierarchical gene co-expression clustering**

At first, the pairwise gene co-expression correlations (Spearman's correlation is used in the package) are calculated between all differentially expressed genes. Then, average linkage hierarchical clustering is implemented. To reduce the noises caused by poorly correlated genes, the hierarchical clustering is stopped at a loose correlation cutoff with z-score 0.52 (about p-value 0.3). This cutoff shows few influences on the results: for the LUC dataset, when this z-score cutoff changes from 0.3 to 0.9 by step 0.1, the hierarchy was stopped at the same place. The z-score of Spearman's rank correlation is calculated using Fisher's transformation as the following formula ($N$ is the number of sample, $r$ is the correlation and $z_r$ is the z-score following a standard normal distribution): $z_r = \sqrt{\frac{N-3}{1.06}}\frac{1}{2}\ln\frac{1+r}{1-r}$. This cutoff should be manually set before clustering (see User Guide for how to set it).

**2.3 Get the co-expression signatures from the gene co-expression hierarchy**

We extracted the gene co-expression signatures (clusters) at different scales by traversing the co-expression hierarchy from leaf to root according to the following rules (the correlation is decreasing and the size of signatures is increasing when traversing the hierarchy from leaf to root):

1) A new cluster is *initiated*, if the cluster contains the minimum number of genes $N^i$ ($N^i$ is arbitrarily set as 10% of the number of differentially expressed genes);
2) A cluster is *extended*, if it merges a branch with less than $N^e$ genes ($N^e$ is set as the maximum of 20 and 10% of the current cluster size);
3) A cluster is *extended*, if it merges another cluster which is very close (normalized cluster distance is less than 0.5) and the records of the two merged clusters are deleted;
4) A cluster is *extracted* at current scale, if the cluster will merge a branch with more than $N^e$ genes or it will merge another cluster with large distance. And a new cluster is *initiated* as the merged cluster.

A binary tree algorithm was implemented to do above steps.

**3. Analyze the miRNA target gene set enrichments in the hierarchical gene co-expression signatures**

**3.1 Data pre-processing: discretize the TargetScan scores**

miRNAs and their target genes (the miRNAs from the same family are merged as a single item) were extracted from TargetScan database (v6.2) [1,2]. A gene was regarded as a target of one miRNA, if the gene contains at least one conserved

predicted miRNA binding site in its 3'-UTR. And the summarized context score (a negative score measuring miRNA-target regulation strength or confidence, provided by TargetScan) was recorded for each miRNA-target pair. Then, we discretized the context scores into *K* levels: all miRNA-target pairs were sorted according to their context scores in decreasing order (the pairs ranked on the top have the lowest regulation strength) and the discretized score for the miRNA-target pair with rank *r* was defined as: $s = 1 + b[rK/N]$. It means the first 1/*K* miRNA-target pairs have lowest score 1, while the last 1/*K* pairs have highest score 1+*b*(*K*-1). According to ref.[3], *K* is set as 5 and *b* as 3 in this study.

**3.2 Generate the randomized control miRNA target gene sets**

To estimate the empirical p-values of the target gene set enrichments, randomized control miRNA target gene sets should be generated. Complete randomization without any restriction usually causes over-optimistic estimation. Like FAME, the control miRNA target gene sets were generated by bipartite graph based random permutation of the miRNA-target pairs with the same discretized scores but keeping the sizes of all target gene sets. This kind of stringent permutation procedure can generate the control miRNA target gene sets which preserve the statistical properties much better [3]. The bipartite graph based random permutation of the miRNA-target pairs is time-consuming. It is suggested that the generated control miRNA target gene sets should be locally stored for future use. At least 10,000 times control miRNA target gene sets should be generated for stable p-value estimation.

**3.3 Calculate the miRNA target gene set enrichment in single signature**

For the *j*-th gene co-expression signature in the hierarchy, we can find the overlapped genes between the signature (denoted as $S_j$) and the *i*-th miRNA target gene set (denoted as $T_i$), and then calculate the raw enrichment score by summing the discretized TargetScan scores (see the details of the score discretization in the following section) of the overlapped genes for *i*-th miRNA:

$$ES_{ij} = \sum_{g \in T_i \cap S_j} s_{ig} .$$

The p-value $p_{ij}$ for this enrichment was estimated by examining the enrichment scores $ES_{ij}(r)$ of 10,000 size-matched random control miRNA target gene sets (the randomization method was presented in the following section):

$$p_{ij} = \frac{\#\left\{r \mid ES_{ij}(r) \geq ES_{ij}, r = 1, 2, \cdots, 10000\right\}}{10000} .$$

**3.4 Find the most significantly enriched signatures**

After getting the enrichments in all hierarchical gene co-expression signatures, the *P*-score $P_i$ for the *i*-th miRNA was calculated as the p-value of the most significant enrichment:

$$P_i = \min_j p_{ij} .$$

The *P*-score was used to measure the miRNA target gene enrichment across the whole gene co-expression hierarchy.

**3.5 Estimate the statistical significance of the most enriched signatures**

The *P*-score is the minimal of a set of p-values, so it is not uniformly distributed along 0~1 (biased to 0). It cannot be directly used to measure the statistical significance of enrichment. Again, we used permutation test to estimate the statistical significance of the P-score: the P-scores $P_i(r)$ of 10,000 size-matched control miRNA target gene sets were calculated according to above steps; and the empirical p-value $p_i$ for the P-score $P_i$ was calculated as:

$$p_i = \frac{\#\left\{r \mid P_i\left(r\right) \le P_i, r = 1, 2, \cdots, 10000\right\}}{10000}.$$

The empirical p-value $p_i$ was used to measure the statistical significance of miRNA target gene set enrichment across the whole hierarchical gene co-expression signatures. To correct the multiple test, fdrtool was used to calculate the *q*-values according to the empirical p-values [4].

## 4. Implementation and Availability

The package is written as Perl scripts. The package can smoothly run under most OS without installing any additional Perl modules. The hierarchical gene co-expression clustering requires large memory and is very time-consuming. For a gene expression profiles with ~5000 genes and ~100 samples, the clustering program requires 16GB memory and will run about one week on single core of AMD Opteron™ 6212 (64bit, 2.6GHz). All the scripts and example files are provided via miRHiC website (http://bioinfo.au.tsinghua.edu.cn/member/jgu/miRHiC).

## References


1. Friedman RC, Farh KK, Burge CB, Bartel DP (2009) Most mammalian mRNAs are conserved targets of microRNAs. Genome Res 19: 92-105.
2. Lewis BP, Burge CB, Bartel DP (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 120: 15-20.
3. Ulitsky I, Laurent LC, Shamir R (2010) Towards computational prediction of microRNA function and activity. Nucleic Acids Res 38: e160.
4. Strimmer K (2008) fdrtool: a versatile R package for estimating local and tail area-based false discovery rates. Bioinformatics 24: 1461-1462.