Using Data from the Cancer Genome Atlas to Investigate Molecular Events Related to Ovarian Cancer

Xiaoye Liu, Gang Fang, Michael Steinbach

Abstract

Ovarian Cancer is the most lethal gynecologic malignancy. Analyzing the molecular events related to ovarian cancer helps understand the pathogenesis of ovarian cancer from a genetic point of view. As miRNA singletons have been found significantly related to ovarian cancer and many other cancers, miRNAs have been recognized as an important riboregulator of gene expression. However, little is known about how pairs of miRNA expression profiles associate with ovarian cancer. In our analysis, we explored the combinatorial effects of miRNA pairs on regulating gene expression. We assessed the non-additive interaction between miRNA pairs on gene expression of patients that carry high grade ovarian cancer. We demonstrate how different miRNAs collectively contribute to ovarian cancer. We will illustrate two examples of miRNA pairs, hsa.miR.937 & hsa.let.7b and hsa.miR.1277& hsa.miR.485.3b, that we found exhibit non-additive interaction pattern on affecting gene expression of the patients with high grade Ovarian Cancer.

Introduction:

Ovarian cancer is the second most common gynecologic cancer in the United States and sixth most common cancer in women worldwide[1,2,3]. In 2010 alone, there were 13,850 deaths from of ovarian cancer with an estimated 21,888 new cases diagnosed [4]. Although 90% of the patients with early-stage ovarian cancer could successfully survive 5 years or longer after initial diagnosis, only 21% of the patients who were diagnosed at advanced-stage could achieve that [2]. Due to the lack of robust methods for early detection, 19% of all ovarian cancer is diagnosed at early stage [5]. Thus, more understanding on the pathogenesis of ovarian cancer is needed for helping to improve the early stage detection diagnosis.

Analyzing the molecular events related to ovarian cancer give the best help on understanding the pathogenesis of ovarian cancer from a genetic point of view. In order to identify the genetic alteration associates with the malignant phenotype of ovarian, many researchers start investigation on analyzing how miRNA expression profiles contribute to ovarian cancer. miRNA is a small and new class of non-coding RNAs that plays an important role in cell-cycle progression, tissue differentiation and organ development [5,7]. Many miRNA singletons have been found significantly related to many other cancers, such as breast cancer, pancreatic cancer and lung cancer. [7,8,9] Several recent studies have also successfully identified that particular miRNA singletons such as miR-21, miR-125a, miR-125b[3], miR-200a, miR-141, miR-199a [5], are significantly differentially or over expressed in the miRNA expression profile of ovarian cancer patients. Although miRNA has been recognized as a riboregulator of gene expression [3] and many miRNA singletons have been discovered as essential cancer bio-makers, little is known about how pairs of miRNA expression profiles associate with ovarian cancer.

In this report, we represent the result of genome-wide miRNA expression profiles in a large set of high grade Ovarian Serous Cystadenocarcinoma (OSC) tissue, a type of highly aggressive epithelial ovarian tumor that accounts for 90% of all ovarian cancer [6]. We assessed the non-additive interaction between miRNA pairs on gene expression of patients that carry OSC. We demonstrate how different miRNAs collectively contributes to this high aggressive OSC.

Material and Methods:

An overview of the methods used can be found in Research Process Chart in Figure1.

Data from The Cancer Genome Atlas (TCGA):

We acquired the OSC miRNA expression data, gene expression data and clinical data from TCGA data portal. (Available at: http://tcga-data.nci.nih.gov/tcga/)

The level three OSC data from TCGA were analyzed for both miRNA expression and gene expression. In TCGA, level three data represents the data which are background corrected, molecular abnormalities interpreted and normalized.

miRNA Expression Data: The miRNA microarray platform of miRNA expression data is Agilent 8*15K performed by University of North Carolina(UNC).The measurements for 799 miRNAs were included in the miRNA expression data in our analysis.

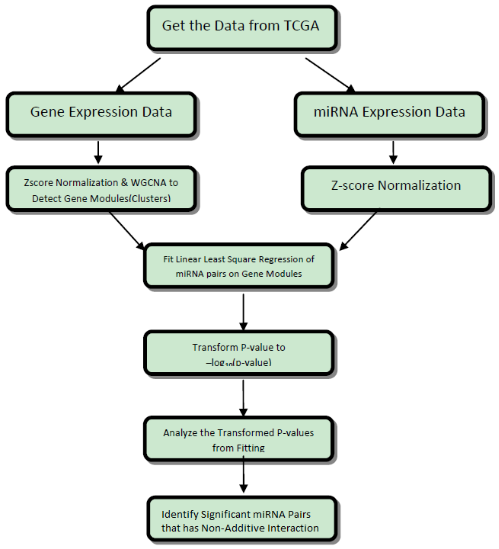
Get the Data from TCGA
Gene Expression Data
miRNA Expression Data
Zscore Normalization & WGCNA to Detect Gene Modules(Clusters)
Z-score Normalization
Fit Linear Least Square Regression of miRNA pairs on Gene Modules
Transform P-value to –log10(p-value)
Analyze the Transformed P-values from Fitting
Identify Significant miRNA Pairs that has Non-Additive Interaction

Figure 1: Research Process Chart.

Gene Expression Data: The platform of gene expression data is Affymetrix HT Human Genome U133 Array Plate Set performed by Massachusetts Institute of Technology (MIT). We have included 12042 genes in our analysis.

We have averaged the expression values of the duplicated samples from the same patients in miRNA and gene expression data. We have also removed patients whose tumor stage information is null. 493 patients who have been annotated having OSC stage II to IV in the clinical data are finally selected in our analysis. (Table 1)

Zscore Normalization:

Before any methods are applied, both miRNA and gene expression data are pre-processed by zscore normalization: subtracting the mean and dividing by standard deviation for data standardization.

Construct Gene Co-expression Network:

We performed Weighted Gene Co-expression Network Analysis (WGCNA) [10-13] on OSC gene expression data. The purpose of applying WGCNA is to attempt to identify co-expressed genes and to combine the high correlated genes into gene modules (clusters). Since gene modules might associate with biological pathways [14], combining genes into modules is an effective scheme for reducing the dimension of gene expression data from a level of thousands into a smaller set of biologically meaningful gene clusters.

Table1: Data Information

*clinical data were last updated on August 10th 2010

Type	Platform	No. of Patients	No. of Singletons
miRNA-Expressions	UNC__H-miRNA_8x15Kv2	491	799
Gene-Expressions	BI__HT_HG-U133A	491	12042
Clinical*	_	491	_

In WGCNA, each gene is considered a node in the network. The distance between a pair of gene nodes is first determined by their pair-wise Pearson correlation value. The gene network is constructed based on all pair-wise Pearson correlation values between the gene nodes in the network. The pair-wise Pearson correlation was calculated across all patients for all genes in the gene expression data matrix. In order to emphasize the large Pearson correlation value, a power β ≥ 1, known as soft threshold as against hard threshold in un-weighted network, needs to be chosen to raise the power of the absolute value of Pearson correlation by the formula:

$$\text{adj}ij = |\text{cor}(xi, xj)|\ \hat{}\beta$$

This adj*ij* is defined as the adjacency of the genes in an unsigned weighted gene co-expression network.

Picking a proper value for the soft threshold β is an essential step in WGCNA. In order to allow the network to keep the continuous nature of the gene co-expression information, β should be chosen obeying the criterion of scale-free topology. [10] In this way, a weighted network exhibits its advantages on the robustness in hierarchical clustering analysis [10,14].

As we applied WGCNA to gene expression data in our analysis, we chose soft power threshold β = 4, which is the smallest value reaching the level of 0.9 on the scale independence. (Figure 2)

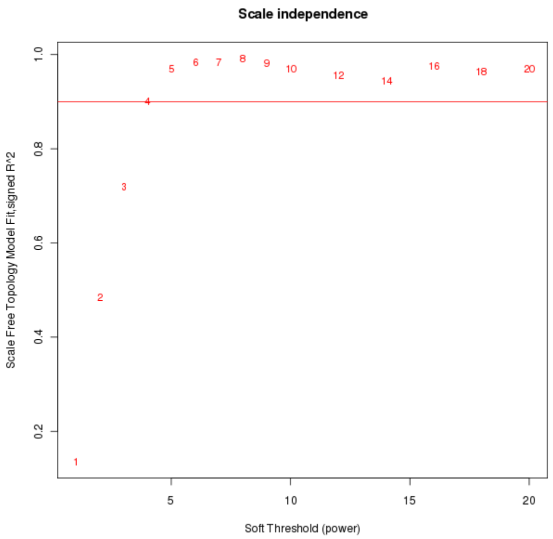

Figure2 : the scale free topology and soft-power selection

Taking the adjacency as input, the dissimilarity of a Topological Overlap Matrix (dissTOM) is computed. The dissTOM computation minimizes the effect of noise and unauthentic associations. dissTOM is the final input for the module detecting in WGCNA.[10]

As we expect to analyze gene modules with a relatively large size, we set the minimum module size to be 30 when performing WGCNA. This allows all gene modules returned have a size no less than 30 genes. After combining dynamic gene modules with a correlation higher than 0.8, all genes have been assigned to 80 modules . The final results are module sizes ranging from 1530 to 33.(79 modules for signed genes and 1 module for unsigned ones). A graphic illustration is shown in figure 3. Module memberships were assigned to each gene in forms of numeric labels and color labels.

In WGCNA, a module eigengene was computed for each gene module as the new representative of all genes in that module. It refers to the first principal component of a gene module. [10, 14] Its value summarizes the gene expression level of a certain gene module. We consider the new eigengene values as the new quantitative traits of genes in our analysis.

Linear Regression Based Test for Finding Non-additive Interaction:

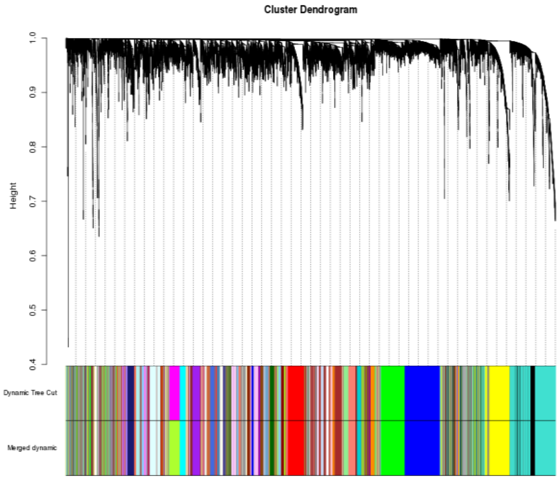

Figure3 : gene module detection. The color bar on the bottom level is the final module selection after merging the similar modules.

In order to identify the level of non-additive interaction of each pair of miRNAs, we performed linear least square regression fitting to each miRNA pairs for each gene module.

$$y = \beta0 + \beta1 * x1 + \beta2 * x2 + \beta12 * x1 * x2 + \varepsilon \quad (*)$$

*y stands for the module eigengene data; gene expression
*x1 stands for miRNA1, x2 stands for miRNA2
*x1*x2 stands for the interaction term
* β 0 is the base,β1, β2 and β12 are the coefficient of each term
*ε represents the error term

Since our primary goal is assessing the non-additive interaction between miRNA pairs on the expressions of gene modules, the test of epistasis was performed in our analysis. We computed F-statistics to compare the full model (*) to the model which only includes the additive terms [17]. Such statistical computation allows us to assess significance of the non-additive interaction of each pair directly.

$$\begin{cases} \text{Ho: } \beta12=0 & \text{purely additive model} \\ \text{Ha: } \beta12\neq0 & \text{full model } (*) \end{cases}$$

We have collected a matrix of all p-values from F-statistic computation in the model comparison for all possible miRNA pair combinations on gene modules. There are 80 sets of p-value matrices returned and each set is a symmetric matrix with zeros on the diagonal and all 799 miRNA as rows and columns. We have transformed all such matrices by taking –log10 of the P-value in every cell (-log10PV). (Figure 4)

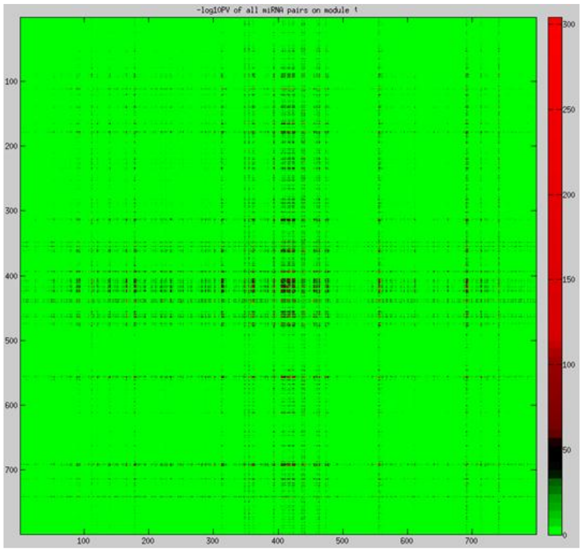

Figure4: An example for illustrating the symmetry –log10PV matrix of all pairs on gene module 1.

Since each matrix is symmetric, both the upper and the lower triangular contains the same -log10PV information. We selected the upper triangular for each of 80 –log10PV matrices. We break every selected part into a vector, where each cell in that vector represents the p-value of corresponding pair of miRNAs. We have combined all such vectors into a matrix with 80 columns and 318801rows. (-log10PV matrix)

Identify non-additive interaction miRNA pairs:
Since we are comparing the full model with the purely additive model, a lower P-value would suggest that the interaction effect is making a significant contribution. This is because we could not ignore the effect of the interaction if we reject the null hypothesis. Equivalently, after transforming PV to –log10PV, a high –logPV would support this conclusion as well.

For identifying the significant miRNAs, we took the row maximum of the –log10PV matrix. We filtered out the low –log10PV pairs and kept the ones that have relatively high (-log10PV>6) –log10PV as target pairs for further graphical identification analysis. In the graphical analysis, for every target pair, we used scatter plot to show the pattern of normalized miRNA expression for the two selected miRNAs, one for each axis. We used color for all scatters to indicate the level of gene expressions of their selected best gene module. The best gene module refers to the module where the target pair has achieved their maximum –log10PV. In this way, we see how the expression of the two miRNAs affects the level of the gene module expressions.

However, due to the high skewness and low variation of the eigengene value in some of the 80 gene modules, the non-additive pattern for some of the pairs we found is difficult to distinguish on their best gene module. Therefore, we have filtered out the gene modules that have IQR < 0.065 in order to illustrate the non-additive pattern clearly. (Figure 5) There are 9 qualified modules selected.

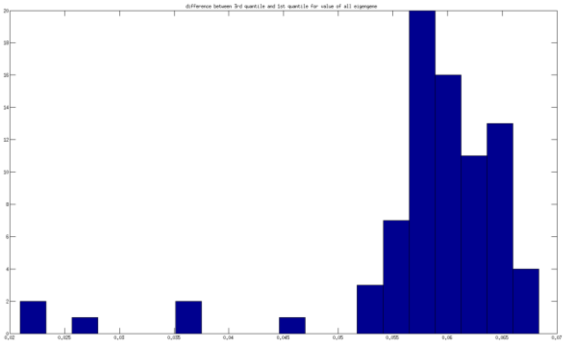

Figure 5: Histogram of IQR for each gene module

False Discover Rate (FDR) computation:
We used FDR computation to correct for the multiple comparisons. [15,16] Across all 9 gene modules, we found 942 pairs show significant to the gene expression trait.

By setting the –log10PV threshold to 6, there are 25 pairs left.(Figure6)

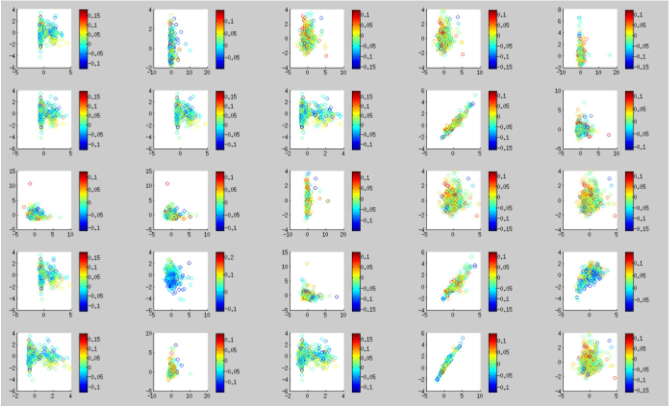

Figure 6: Brief graphic overview of 25 pairs.

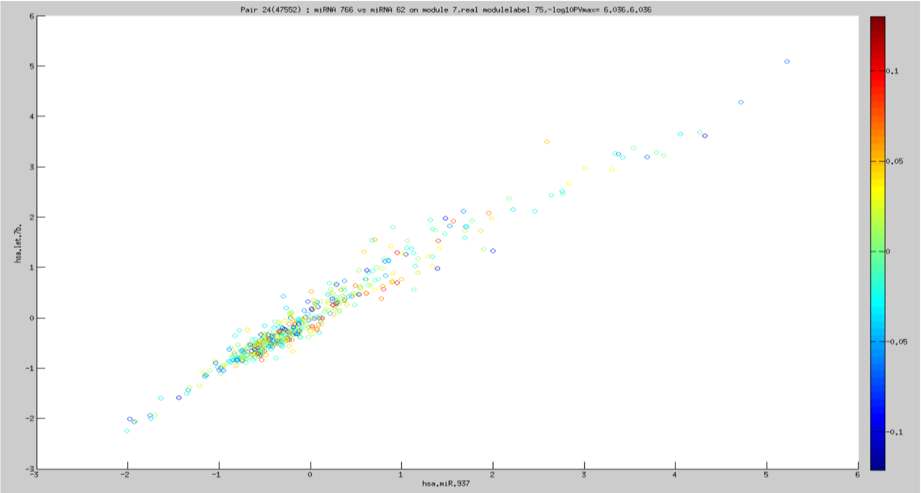

Figure 7: hsa.miR.937 and hsa.let.7b on gene module 7 (Pair 1)

From those 25 pairs, we have successfully found several that exhibit non-additive pattern on affecting eigengene value. We will be presenting two pairs that have a significant pattern as examples.

Pair1: hsa.miR.937 and hsa.let.7b on gene module 7

In the hsa.miR.937 and hsa.let.7b, the –log10PV is 6.036 on their best eigengene -- gene module 7 of the 9 qualified gene modules (The original gene module label is 75 before IQR filtering). This high –log10PV indicates that the PV for the corresponding model comparison test is very small and indicates the significance of the interaction effect of the two miRNAs.

From figure 7, we can see that the corresponding expression value of the two miRNAs over 493 patients shows a linear pattern. However, the non-additive interaction is illustrated by color of figure 7. The color, which indicates the eigengene expression level, does not exhibit a linear trend related to the miRNA expression values. In other words, we could not find the value of eigengene increasing or decreasing as the expression of miRNAs monotonically changes. It means the miRNAs are not additively controlling the expression values on their best eigengene, which best proves that their non-additive interaction exists on affecting module7.

Pair2: hsa.miR.1277 and hsa.miR.485.3b on gene module 4.

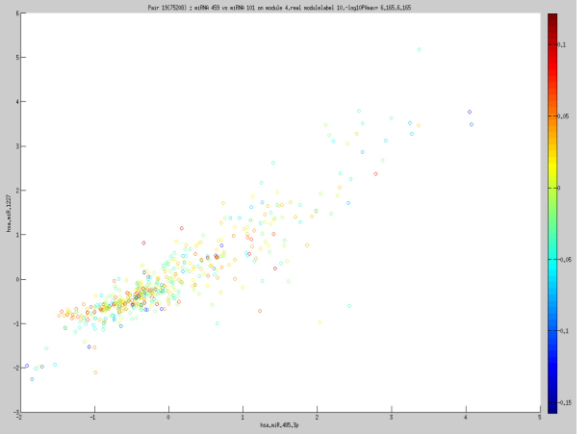

Figure 8: hsa.miR.1277 and hsa.miR.485.3b on gene module 4

In this selected pair, hsa.miR.1277 and hsa.miR.485.3b, the –log10PV is 6.165 on their best eigengene -- gene module 4 of the 9 qualified gene modules (The original gene module label is 10 before IQR filtering). Similarly, the –log10PV is large, which indicates that the PV for the corresponding model comparison test is very small. We fail to reject adding interaction term in the full model.

In this pair, we do not find additive effects of the miRNAs on their best eigengene, which proves their non-additive interaction on the eigengene expression exists.

Conclusion:

We have successfully discovered the existence of non–additive interactions between miRNAs on affecting the gene expression of patients with ovarian cancer. Although the significance of miRNA singletons on regulating gene expression is recognized today, there are seldom studies investigating on the pair effects of miRNAs in ovarian cancer. Through our analysis, we have successfully discovered the existence of non–additive interactions between miRNAs that affect gene expression of patients with ovarian cancer. This discovery gives strong support for further analysis, such pathways analysis relates to non-additive interaction effects of miRNAs on ovarian cancer or other cancers.

References:

1. "Ovarian Cancer- Get the Facts about Gynecologic Cancer." Centers for Disease Control and Prevention. Mar. 2009. Web.04.Mar. 2011. <http://www.cdc.gov/cancer/ovarian/pdf/Ovarian_FS_0308.pdf>.
2. Douglas D. Taylor, Cicek Gercel-Taylor, MicroRNA signatures of tumor-derived exosomes as diagnostic biomarkers of ovarian cancer, Gynecologic Oncology, Volume 110, Issue 1, July 2008, Pages 13-21, ISSN 0090-8258, DOI:10.1016/j.ygyno.2008.04.033.(http://www.sciencedirect.com/science/article/pii/S0090825808003430)
3. Nam EJ, Yoon H, Kim SW, Kim H, Kim YT, Kim JH, et al.MicroRNA expression profiles in serous ovarian carcinoma. Clin Cancer Res 2008;14:2690–5.< http://clincancerres.aacrjournals.org/content/14/9/2690.full.html>
4. Altekruse SF, Kosary CL, Krapcho M, Neyman N, Aminou R, Waldron W, Ruhl J, Howlader N, Tatalovich Z, Cho H, Mariotto A, Eisner MP, Lewis DR, Cronin K, Chen HS, Feuer EJ, Stinchcomb DG, Edwards BK (eds). SEER Cancer Statistics Review, 1975-2007, National Cancer Institute. Bethesda, MD, http://seer.cancer.gov/csr/ 1975_2007/, based on November 2009 SEER data submission, posted to the SEER web site, 2010. <http://seer.cancer.gov/statfacts/html/ovary.html >
5. Iorio MV, Visone R, Di Leva G, Donati V, Petrocca F, Casalini P, et al.MicroRNA signatures in human ovarian cancer. Cancer Res 2007;67:8699–707.< http://cancerres.aacrjournals.org/content/67/18/8699.short>
6. Xiu-Qin Li, Shu-Lan Zhang, Zhen Cai, Yuan Zhou, Tian-Min Ye, Jen-Fu Chiu, Proteomic identification of tumor-associated protein in ovarian serous cystadenocarinoma, Cancer Letters, Volume 275, Issue 1, 8 March 2009, Pages 109-116, ISSN 0304-3835, DOI: 10.1016/j.canlet.2008.10.019. <http://www.sciencedirect.com/science/article/pii/S0304383508008173>
7. Marilena V. Iorio, Manuela Ferracin, Chang-Gong Liu, Angelo Veronese , et al. MicroRNA Gene Expression Deregulation in Human Breast Cancer Cancer Res August 15, 2005 65:7065-7070; doi:10.1158/0008-5472.CAN-05-1783
8. Mark Bloomston, Wendy L. Frankel, Fabio Petrocca, Stefano Volinia, Hansjuerg Alder, John P. Hagan, Chang-Gong Liu, Darshna Bhatt, Cristian Taccioli, Carlo M. Croce.Preliminary Communication MicroRNA Expression Patterns to Differentiate Pancreatic Adenocarcinoma From Normal Pancreas and Chronic Pancreatitis JAMA. 2007;297(17):1901-1908.doi:10.1001/jama.297.17.1901
9. Nozomu Yanaihara, Natasha Caplen, Elise Bowman, Masahiro Seike, Kensuke Kumamoto, Ming Yi, Robert M. Stephens, Aikou Okamoto, Jun Yokota, Tadao Tanaka, George Adrian Calin, Chang-Gong Liu, Carlo M. Croce, Curtis C. Harris, Unique microRNA molecular profiles in lung cancer diagnosis and prognosis, Cancer Cell, Volume 9, Issue 3, March 2006, Pages 189-198, ISSN 1535-6108, DOI: 10.1016/j.ccr.2006.01.025. <http://www.sciencedirect.com/science/article/pii/S153561080600033X>
10. Zhang B, Horvath S: A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol 2005,4(1):. Article17
11. Horvath S, Zhang B, Carlson M, Lu KV, Zhu S, Felciano RM, Laurance MF, Zhao W, Qi S, Chen Z, et al.: Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. Proc Natl Acad Sci USA 2006, 103(46):17402-17407.
12. Horvath S, Dong J: Geometric interpretation of gene coexpression network analysis. PLoS Comput Biol 2008, 4(8):e1000117.
13. Langfelder P, Horvath S: WGCNA: an R package for weighted gene co-expression network analysis. BMC Bioinformatics 2008, 9(1):559.
14. Saris Christiaan Horvath, Steve van Vught, Paul van Es, Michael,Blauw, Hylke Fuller, Tova,et al 2009 1 10.1186/1471-2164-10-405. Weighted gene co-expression network analysis of the peripheral blood from Amyotrophic Lateral Sclerosis patients http://www.biomedcentral.com/1471-2164/10/405>
15. Storey JD (2002) A direct approach to false discovery rates. J R Stat Soc [SerB] 64: 479–498.
16. Storey JD, Tibshirani R (2003) Statistical significance for genome-wide studies. Proc Natl Acad Sci U S A 1009440–944
17. Storey JD, Akey JM, Kruglyak L, 2005 Multiple Locus Linkage Analysis of Genomewide Expression in Yeast. PLoS Biol 3(8): e267. doi:10.1371/journal.pbio.0030267