A General Workflow to Analyze Key Cancer-Related Genes Based on NCBI illustrated by the case of gastric cancer

Zhikuan Quan

School of Statistics, Beijing Normal University, Beijing 100875, China

201511011114@bnu.edu.cn

Abstract

Background: In bioinformatic and clinical analysis, bioinformaticians often need to identify the key cancer-related genes, find out the signaling pathway and analyze the survival parameters in the development of cancer. Recent study has shown that the differentially expressed genes played an important role in finding out the disease-related genes. The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information, which gives opportunities to use this database to apply genetic analysis. Methods: In this study, a general workflow to find out the key cancer-related genes and signaling pathway based on NCBI and gene expression omnibus was conducted. In this workflow, we implement four basic steps. The first step is to get the data from NCBI. And then, we select a subset of differentially expressed genes (DEGs) which is used to apply GO and KEGG pathway analysis in order to find out some specific cancer-related genes. The final step is to focus on survival analysis with specific cancer-related genes in KM-Plotter. Results: Gastric cancer is one of the most common tumors in the digestive tract. The workflow of exploring cancer-related genes was established based on the NCBI database, which takes gastric cancer as an example, new genes related to gastric cancer were found through the significance test, the analysis of GO and KEGG signal pathways and the survival analysis of patient samples. Conclusions: Through public database of GEO, we conduct a comprehensive analysis of the cancer-related gene on cancer groups and control groups. And then we analyze the key genes involved in biological processes and signaling pathways. In order to comprehensively analyze a certain gene expression in tumor characteristics and significance, we use the survival analysis tools to verify the genetic effect on cancer patients' overall survival time. It provides theoretical support and guidance to further related researches about the gene.

Keywords

Cancer-related genes; NCBI; Survival Analysis.

1. Background

In bioinformatic and clinical analysis, bioinformaticians often need to identify the key cancer-related genes, find out the signaling pathway and analyze the survival parameters in the development of cancer. Recent study has shown that the differentially expressed genes played an important role in finding out the disease-related genes. [1] The National Center for Biotechnology Information is an important and useful tool to science and health that offers plenty of ways to biomedical and genomic information, which gives opportunities to use this database in order to apply genetic analysis. [2] Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data which uses the R statistical programming language with open sources and developments, which means it will be an excellent tool to apply bioinformatic analysis. Through NCBI database, there are numerous bioinformatics analysis using the data combined with clinical trials, which not only applies

evolutional methods with genome data in NCBI, but also exploits the dataset to distinguish diseases-related genes. [3] Based on the previous researches on how to find out specific cancer-related genes, the National Center for Biotechnology Information database enables researchers analyze with high-throughput genetic data to create new theory and workflow. In this article, a general workflow to analyze key caner-relate genes based on NCBI was constructed, which is based on diverse methods from getting the data in NCBI to applying survival analysis with specific new genes which is likely to be seen as key cancer-related gene.

2. Methods

2.1 Get the data from NCBI

The National Center for Biotechnology Information (NCBI) boosts the development of science and health by providing many datasets and resources about biomedical and genomic information. The most important tool is Gene Expression Omnibus (GEO) which is a public functional genomics data library which supports compliant data submissions. Array-based and sequence-based data are used to restore in GEO. [4] In the GEO tools, it offers many accesses to use online analysis models to help users search for relative materials and download data about some experiments and related gene expression profiles.

The first step in general workflow is to get the data from NCBI, which combines and re-analyzes some public specific disease data sets with previous researches so that it will reveal previously unknown relations of specific genes and the development of diseases in the data. The datasets usually contain some information of expressional level of diverse genes in different organs, which means we can apply more statistical analysis based on the data.

2.2 Select a subset of differentially expressed genes

After downloading the data from GEO database, significant analysis of microarrays (SAM) was used to select a subset of differentially expressed genes (DEGs). We will compare two or more conditions to identify genes with significant differential expression.

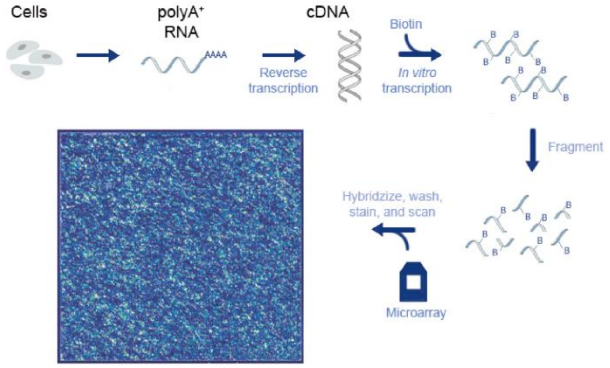

Fig.1 The overview of GEO Data The process of getting the raw data from cells which contains diverse expressional information. We get the Chip grayscale data from NCBI and acquire expression level data through quality controlling and data cleaning.

There are two statistical methods to analyze the differentially expressed genes. [5] We can use normal t-test to analyze it but it has some drawbacks. The drawbacks of common t-statistic: Genes with small fold change but even smaller within-group variability, may be highly significant from a statistical, but not a biological point of view; Small sample estimates of standard deviation can be very noisy. In order to deal with the flaw of statistical testing, we will take the empirical Bayes test. In Bayes methods, we will use the data to come up with a sensible guess for the shape and mean parameters. It can used R with limma package to apply statistical test. [6] In the dataset, we use logarithmic transformation to handle the raw data, which means positive logFC indicates the Upregulation and the negative logFC indicates the Downregulations.

2.3 Gene Ontology Analysis

When we select the differentially expressed genes, we need to find out what is the specific genes that deserve to be analyzed. Gene Ontology is the outline for the model of biology. [7] The GO defines concepts or classes used to describe gene function, and the relationships between these concepts. In GO analysis methods, it divides diverse functions into three aspects: (1). Molecular Function: molecular activities of gene products; (2). Cellular Component: where gene products are active; (3). Biological Process: pathways and larger processes made up of the activities of multiple gene products. The most common use of the Gene Ontology annotations is for interpretation of large-scale molecular biology experiments, sometimes called "omics" experiments. These experiments measure either: gene products especially for RNA and proteins; mutations in the DNA sequence of genes and small molecules metabolized by proteins. Thus, they can all be related to gene function. One of the main uses of the GO is to perform enrichment analysis on gene sets. For instance, given a set of genes that are up-regulated under certain conditions, an enrichment analysis will find out which GO terms are over-represented or under-represented by using annotations for that gene set.

We will apply Fisher exact test in R with clusterProfiler package. In the Fisher test, the P-value is the probability or chance of seeing at least x number of genes out of the total n genes in the list annotated to a particular GO term, given the proportion of genes in the whole genome that are annotated to that GO Term. That is, the GO terms shared by the genes in the user's list are compared to the background distribution of annotation. The closer the p-value is to zero, the more significant the particular GO term associated with the group of genes is. Through these methods, we can classify genes with different function in order to focus on some specific genes.

2.4 Kyoto Encyclopedia of Genes and Genomes Pathway

Kyoto Encyclopedia of Genes and Genomes (KEGG) is a database resource which is used for understanding high-level functions and utilities of the biological system linking the networks of genes and molecules, especially for the large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies. [8] In this study, the differentially expressed genes were mainly classified by using KEGG in the public database, and the genes in Pathway were analyzed based on some kind of specific discrete distributions, such as the hypergeometric distribution. What's more the Pathway classification was significantly correlated with the experimental purpose.

2.5 Survival Analysis

In survival analysis, the survival function is able to describe the time-to-event phenomena, which measures the probability of an individual surviving beyond time x. It's defined when X is continuous:

$$S(x) = \Pr(X > x) = \int_x^{\infty} f(t)dt$$

Here the $f(x)$ is the probability density function. We will use the Kaplan-Meier Estimator to estimate the survival function. [9]

Notation:

- Let $Y_i = \sum_{j=1}^{n} I(X_j \geq t_{i-1})$ be the number of individuals who are alive at time t_i;
- $d_i = \sum_{j=1}^{n} I(t_{i-1} < X_j \leq t_i)$ be the number of deaths between t_{i-1} and t_i;
- Hazard Rate: $\frac{d_i}{Y_i}$, provide an estimate of the conditional probability that an individual who survives to just prior to time t_i experiences the event (death) at time t_i.

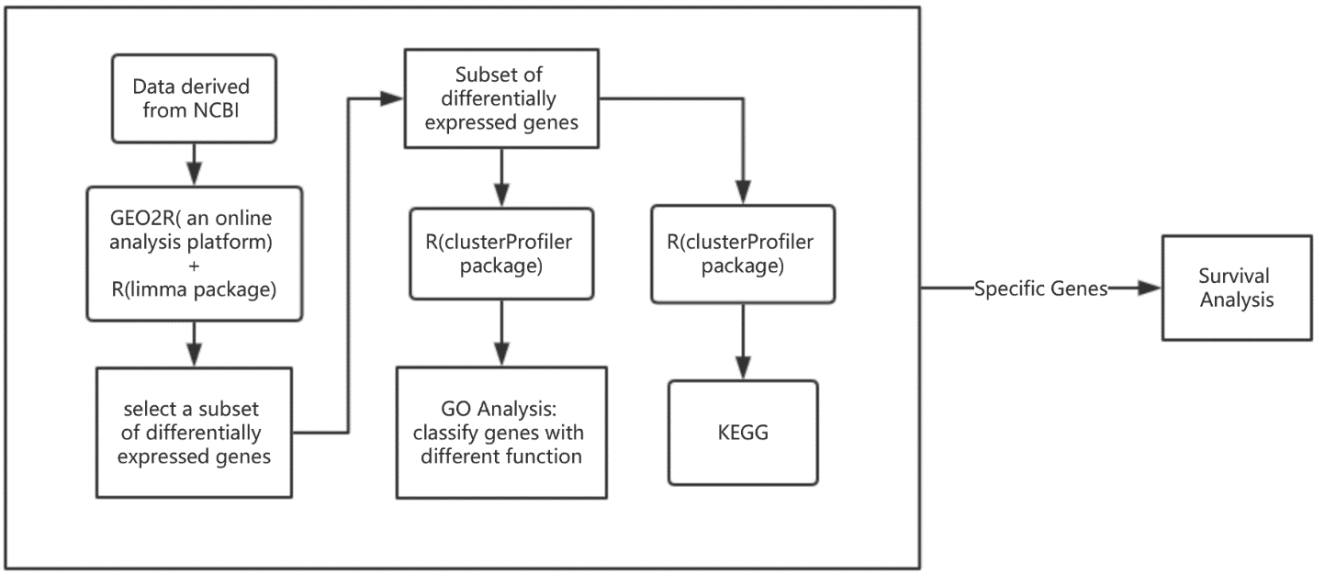

Figure 2 The General Workflow The general workflow includes five methods: (1). Get the data from NCBI; (2). Select a subset of differentially expressed genes; (3). Gene Ontology Analysis; (4). Kyoto Encyclopedia of Genes and Genomes Pathway Analysis; (5). Survival Analysis.

The Kaplan-Meier Estimator is defined as:

$$\hat{S}(t) = \begin{cases} 1 & , \text{if } t < t_1 \\ \prod_{t_i \le t} \left[1 - \frac{d_i}{Y_i}\right]^{\delta_{t_i}} & , \text{if } t_1 \le t \end{cases}$$

Clearly, when t_i is a censored survival time, $\delta_{t_i} = 0$. In contrast, when t_i is an actual failure time, $\delta_{t_i} = 1$.

We will use KM-Plotter to analyze the survival function in the development of cancer to verify the expression of differentially expressed genes and analyze the effect of key genes on survival time of cancer patients. The KM-Plotter is an online analysis software with some major cancers like gastric cancer or breast cancer. [10] What's more, we can analyze the prognostic value of target gene for overall survival of cancer patients, which provides some extra clinical information specific cancer-related genes.

2.6 General Workflow

In this study, a general workflow to find out the key cancer-related genes and signaling pathway based on NCBI and gene expression omnibus was conducted. In this workflow, we implement five basic steps. It includes: (1). Get the data from NCBI; (2). Select a subset of differentially expressed genes; (3). Gene Ontology Analysis; (4). Kyoto Encyclopedia of Genes and Genomes Pathway Analysis; (5). Survival Analysis. If we can download data which includes the expression levels derived from NCBI, we can apply these methods to analyze the cancer-related genes and we can take specific genes to do survival analysis, so as to distinguish cancer-related genes preliminarily.

3. Results

We will take gastric cancer [11] as example to illustrate the general workflow. Gastric cancer is one of the most common tumors in the digestive tract. In this study, the database related to gastric cancer in GEO was used to study the genes related to gastric cancer through bioinformatics, and the characteristics and significance of genes related to the occurrence, progression and prognosis of gastric cancer were screened and predicted, providing a new idea for tumor research. In order to implement the general workflow, we use the gastric cancer data from NCBI.

The first step is to download Gastric Cancer Data GSE79973, GSE54129 and GSE13911 from NCBI. It includes the whole genome sequencing data of gastric cancer tissue and normal precancerous tissue. The data set covered different stages and different tissue types of gastric cancer. What's more the comparative test results for drug or other interventions in patients with gastric cancer were excluded.

And then we select a subset of differentially expressed genes. In the data, it includes multiple probes combined to make one expression summary per array per gene. Genes differ with respect to average log expression in each disease condition and variability within each condition.

Based on the previous research and using empirical Bayes test, [12] we set $|\log_2 FC| > \log_2 1.5$ and $p \leq 0.001$ in order to find out the differentially expressed genes: In this study, the selected genes were combined to obtain 1,318 differentially expressed genes, 720 differentially expressed genes were up-regulated and 598 differentially expressed genes were down-regulated. A total of 339 genes with intersections in two or more data sets were taken for further bioinformatics analysis. The DEGs which we want to use to analyze are shown below.

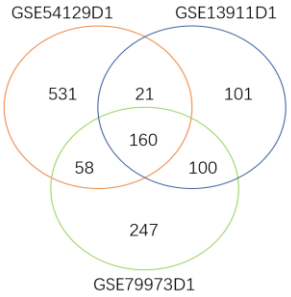

Fig. 3 The overview of DEGs in datasets we set $|\log_2 FC| > \log_2 1.5$ and $p \leq 0.001$ in order to find out the differentially expressed genes a total of 339 genes with intersections in two or more data sets were taken for further bioinformatics analysis.

We can use R to apply GO analysis. The GO enrichment analysis showed that these differentially expressed genes were mainly distributed in gastric, duodenal, colon, tendon, lung, kidney and other tissues. The enrichment degree of the first few genes is mainly involved in the biological process of digestion, drug metabolism, ketone metabolism, and collagen metabolism.

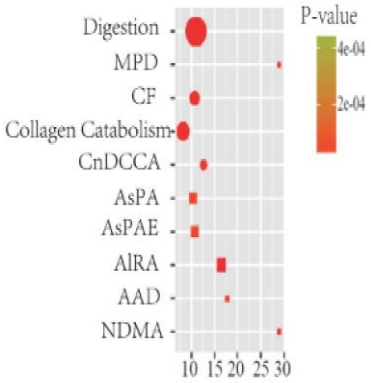

Fig.4 The result of Top-10 GO Enrichment Analysis The figure shows us different types of enrichment. The rectangle represents the Molecular Function; the triangle stands for Cellular Component; the rounded dot means Biological Process. The bigger the dot, the more differently expressed genes. According to the analysis of GO enrichment, the functional genes mainly focus on digestion, drug metabolism and steroid, retinol metabolism and other pathways. In order to illustrate the general workflow, we think about some genes related to digestion.

Classification	Main Genes
Alcohol dehydrogenase family	ADH1A 、ADH1B、ADH7
The cytochrome P450 family	CYP2C9、CYP2C18、CYP3A5、CYP2C19
The glucuronic acid transferase family	UGT2B15、UGT1A3、UGT1A8、UGT1A5、UGT1A9
Collagen family	COL10A1、COL12A1、COL1A1、COL11A1、COL1A2、COL8A1、COL4A1
Aldo-keto reductases	AKR1、AKR1C2、AKR1C1、AKR1B10、AKR7A3
The family of glutathione S- transferase	GSTA1、GSTA3
Family of age-related epithelial membrane proteins	CLDN1、CLDN3、CLDN4、CLDN7、CLDN18
Thrombospondin	THBS1、THBS2

Signal pathway analysis of differentially expressed genes in gastric cancer

Fig. 5 The Result of KEGG Analysis In this study, the identified gastric cancer differential genes included gene clusters such as cytochrome P450 family and glucuronic acid transferase family. We take GSTA3 from the family of glutathione S- transferase as example to illustrate the workflow. We select some main genes possibly related to gastric cancers and show them.

The analysis of signal pathway based on KEGG pathway analysis is consistent with the analysis of GO function. It mainly focuses on digestion, drug metabolism and steroid, retinol metabolism and other pathways. In addition, the accumulation of different genes in the pathway analysis was also significant in the interaction between the vascular endothelial growth factor signaling pathway and the extracellular matrix receptor. In this study, the identified gastric cancer differential genes included gene clusters such as cytochrome P450 family and glucuronic acid transferase family. [12] Previous studies have also confirmed that these genes are related to the occurrence and development of gastric cancer.

We take GSTA3 [13] from the family of glutathione S- transferase as example to illustrate the survival analysis.

In this study, key genes were divided into high expression group and low expression group according to the median expression of target genes for survival curve analysis. The total survival time of the high expression group was significantly lower than that of the low expression group. These genes play different roles in the occurrence and development of gastric cancer. [14] Therefore, further functional verification of selected key genes will be of great significance for studying their exact functions.

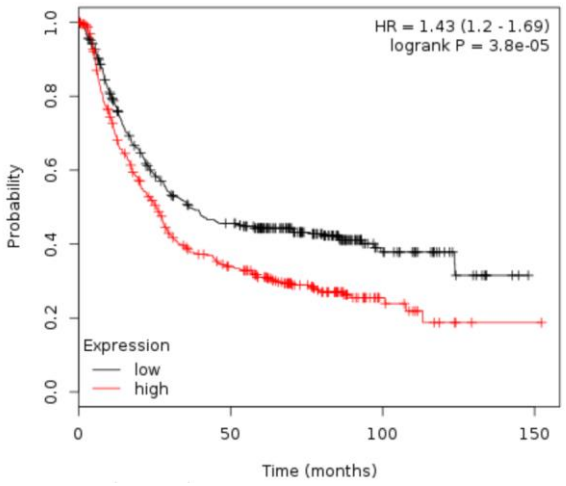

Figure 6 Survival Function In this graph, the red curve represents the high expression group and the black curve means the low expression group. The probability of survival goes down over time. The total survival time of the high expression group was significantly lower than that of the low expression group. It means that GSTA3 gene is related to the development of gastric cancer.

4. Conclusion

Through public database of GEO, we can conduct a comprehensive analysis of the cancer-related gene on cancer groups and control groups. And then we analyze the key genes involved in biological processes and signaling pathways. In order to comprehensively analyze a certain gene expression in tumor characteristics and significance, we use the survival analysis tools to verify the genetic effect on cancer patients' overall survival time. It provides theoretical support and guidance to further related researches about the gene.

References

[1] Kisand, V., & Lettieri, T. (2013). Genome sequencing of bacteria: sequencing, de novo assembly and rapid analysis using open source tools. Bmc Genomics, 14(1), 211-211.
[2] Tatusova, T. (2016). Update on genomic databases and resources at the national center for biotechnology information. Methods in Molecular Biology, 1415(609), 3.
[3] Adam M. Goldstein. (2010). The ncbi databases: an evolutionist's perspective. Evolution: education\s&\soutreach, 3(3), 451-455.
[4] Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., & Tomashevsky, M., et al. (2013). Ncbi geo: archive for functional genomics data sets—update. Nucleic Acids Research, 39(Database issue), 1005-10.
[5] Dudoit, S., Yang, Y. H., Callow, M. J., & Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. stat. sinica 12, 111-139. Statistica Sinica, 12(1), 111--139.
[6] Marchionni, L. (2010). A package to perform run gene set enrichment across genomic platforms.
[7] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., & Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. the gene ontology consortium. Nature Genetics, 25(1), 25-9.
[8] Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., & Kanehisa, M. (2000). Kegg: kyoto encyclopedia of genes and genomes. Nucleic Acids Research, 27(1), 29-34.
[9] Foldvary, N., Nashold, B., Mascha, E., Thompson, E. A., Lee, N., & Mcnamara, J. O., et al. (2000). Seizure outcome after temporal lobectomy for temporal lobe epilepsy: a kaplan-meier survival analysis. Neurology, 54(3), 630-4.
[10] Zhou, X., Teng, L., & Min, W. (2016). Distinct prognostic values of four-notch-receptor mrna expression in ovarian cancer. Tumor Biology, 37(5), 6979-6985.
[11] Genetic polymorphism of cytochrome P450 (CYP) 1A1, CYP1A2, and CYP2E1 genes modulate susceptibility to gastric cancer in patients with Helicobacter pylori infection[J].
[12] Ujjala Ghoshal, Shweta Tripathi, Sushil Kumar, Balraj Mittal, Dipti Chourasia, Niraj Kumari, Narendra Krishnani, Uday C. Ghoshal. Gastric Cancer. 2014 (2)
[13] lan R Jowsey, Stephen A Smith, John D Hayes. Expression of the murine glutathione S - transferase α3 (GSTA3) subunit is markedly induced during adipocyte differentiation: activation of the GSTA3 gene promoter by the pro-adipogenic eicosanoid 15-deoxy-Δ 12,14 -prostaglandin J 2[J]. Biochemical and Biophysical Research Communications,2003,312(4).
[14] Meyer, H. J. (2005). The influence of case load and the extent of resection on the quality of treatment outcome in gastric cancer. European Journal of Surgical Oncology, 31(6), 595-604.