ANALYSIS OF FASTNCA ALGORITHM BASED ON TRANSCRIPTIONAL REGULATION OF BREAST CANCER

Anne Rainey
Southern University and A&M College
Email: anne_rainey_00@subr.edu

Abstract: Single Nucleotide Polymorphisms (SNPs) Can Control Transcription Factors TF (transcription factor) on the allele-specific binding, in order to Control the expression of specific genes, quantitatively deduce TF The activity and its regulatory strength will play an important role in the analysis of differential genes and their regulatory effects. Book Research using Rapid Network Composition Analysis fast NCA (fast network component analysis) algorithm to derive breast cancer BC (breast cancer) poor different expression TF activity and its effect on target gene TG (transcription gene) and construct its transcriptional regulatory network. At the same time, considering the micro Arrayed gene expression data and next-generation sequencing technologies. This study adopts the method of comparing the two kinds of data with differential genes. Fusion method to explore the function of transcriptional regulation. Molecular biology analysis found that the shared significant TF regulated by the same or not same TG have participated with The biological processes and pathways closely related to BC pathogenesis have also proved that through the fusion analysis of multiple data, it is possible to make up for the single data Insufficient data, more comprehensive and full exploration BC pathogenic mechanism.
Keywords: breast cancer, single nucleotide polymorphism, rapid network component analysis, transcriptional regulation

INTRIDUCTION

Breast cancer (breast cancer, BC) is the middle hair of women in the world today One of the most frequently occurring tumors. lead toBC There are various factors, and the incentives are very complex, so forBC The pathogenic mechanism is still unclear. close In recent years, the annual incidence of breast cancer in my country has gradually increased. The gene transcriptional regulatory network is composed of transcription factors (transcription factor, TF) andComplex biomolecular networks formed by genes (Barabasi and Oltvai, 2004), which is the display of life functions at the level of gene expression. Therefore, the baseDue to the exploration of changes in expression levels and transcriptional regulation mechanisms, it is important for us toResearch on the pathogenesis of disease plays an important role.

In recent years, studies have found that single nucleotide polymorphisms (single nu- Genetic variation in the form of cleotide polymorphism (SNP) affects TF expression activity and its effect on target gene (transcription gene, TG) Manipulating intensity to induce differential gene expression (Ye et al., 2009). another On the one hand, with the development of high-throughput technology, high-throughput transcriptomics data such as microarrays and second-generation sequencing RNA-seq has become a detective A tool for detecting changes in gene expression. However, hybridization-based microarrays Columns are restricted to known sequences and cannot detect new RNA, the result of research The results are largely inconsistent; in contrast, RNA-seq utilizes Efficient next-generation DNA sequencing method (Qi et al., 2011), with With the characteristics of digital signal and high sensitivity, the dynamic range of detection is obtained improve. However, the current RNA-seq The sample size needs to be relatively large. That is, biological analysis with a small number of samples is more difficult. Therefore, when applying On the one hand, the two technologies not only overlap and compete but also have complementary advantages. This study will use microarray data and RNA-seq Construct transcripts separately Regulatory networks for more comprehensive analysis BC Gene expression changes. due to micro Arrays are usually assumed to follow a lognormal distribution, while RNA-seq follow Discrete distribution, this study chooses limma-voom Extract differential genes (Soneson and Delorenzi, 2013), relatively immune to outliers sound, the calculation speed increases. Considering the network composition analysis (network component analysis, NCA) with computational instability and multiple Due to the limitations of local solutions, this study takes the expression of the extracted differential genes data and TF control TG Information fusion and application to fast network Component analysis (fast network component analysis, FastNCA), The computational complexity is greatly reduced (Chang et al., 2008). Finally find out same in both data TF and its regulation TG carry out biological analysis found that the same TG Regulated same or different TF have participated with BC -related biological processes and pathways: cell interaction, chromosome analysis Isolation, signal transduction pathways, cell cycle pathways, etc. Visible, the number of microarrays According to and RNA-seq Complementary, multi-data Fusion analysis can relatively compensate single data for multiple reasons The errors caused by the following conditions can be obtained to obtain more comprehensive and sufficient results, which are useful for future exploration.

RESULTS AND ANALYSIS

Based OnLimma - Voom Differential Gene Extraction

The microarray data set used in this study was obtained from the National Institute of Biology Datasets in the Gene Expression Database of the Technical Information Center GSE42568, In this study, 17 healthy samples, 45 non-metastatic diseased sample book. RNA-seq is from TCGA (The cancer genome altas). This dataset contains 28 healthy samples and 206 sick sample book. Each sample in the above two data sets contains 42,451 and 20,503 Gene expression data corresponding to a gene probe. Because The limma-voom package converts raw count data to logarithms count, untreated count data. first convert the count for log-cpm value, then normalized, and finally generate precision weights. In this study, choose logFC (Denotes the expression between the test conditions log2 times change) ≧ 2 or logFC ≦ - 2, in a typical statistical analysis, make use p -value0.05 As a threshold, to screen differentially expressed genes.

micro array number according to pass superior described Threshold strip piece sieve select common get 1 142 genes, a total of 1 699 genes were screened out by RNA-seq differentially expressed Gene. Both data share 509 identical differential genes (Fig. 1).

Based On Fast Nca Algorithmic Regulatory Network Construction

to find out BC Two kinds of data play a major role before and after illness TF and its pair TG Regulatory strength information, this study through fast NCA Algorithmic inference TF Activity and TF right TG control strength. Selected for research In the website BIOBASE (http://www.gene-regulation.com) of TRANSFAC database, which contains a total of 67 375 strip TF-TG regulatory relationship. In order to select with BC related key TF, Ben The study combined the differentially expressed genes extracted above with TF-TG database match. run Fast NCA The prerequisite for the algorithm is to construct two Conditional input: (1) matrix [E], representing the original BC gene expression data The target gene expression profile of ; (2) The initial connectivity matrix [C], which represents TF right Regulatory relationship of TG, if TF regulation TG, then the connection matrix in The element value is 1 ; if there is no regulatory relationship, the initial value is 0. For microarray data, the filter yields 204 transcription factors, a total of 398 pairs of regulatory relationships. In order to facilitate the follow-up research, this paper selects the The number of control genes (≧ 4) is more 17 indivual TF. Similarly, for RNA-seq data, using the same method to select regulatory genes 5 of 16 indivual TF, total have 193 control relationship.

Finally, the fast NCA Algorithm obtained TF active matrix, tuning control matrix and TG Raw expression data were changed accordingly, using Cytoscape software constructs transcriptional regulatory networks for control and diseased samples Network Diagram (Figure 2; picture 3). In order to be able to visualize TF and TG of Changes, the data of the construction of the regulatory network have been normalized, TF expression activity and TG The expression value is averaged after normalization Worth it, TF-TG The positive and negative regulatory effects are determined by the regulatory matrix TF right TG The positive and negative control values are expressed qualitatively. The square nodes in the figure represent TF, circular nodes represent TG, nodes red and green indicate TF or TG The level of expression is high or low, red means high expression level, green means The expression level is low, and the color depth indicates the higher or lower expression level, even Line red indicates TF right TG is a positive regulation, and green is a negative regulation.

It can be seen from the constructed network that most TF and TG sick The expressions before and after have changed (Fig. 2; picture 3). one TF Can to control multiple TG, a TG can also be multiple TF regulated, that is TG The expression is subject to one or more TF Combined effect of expression activity.

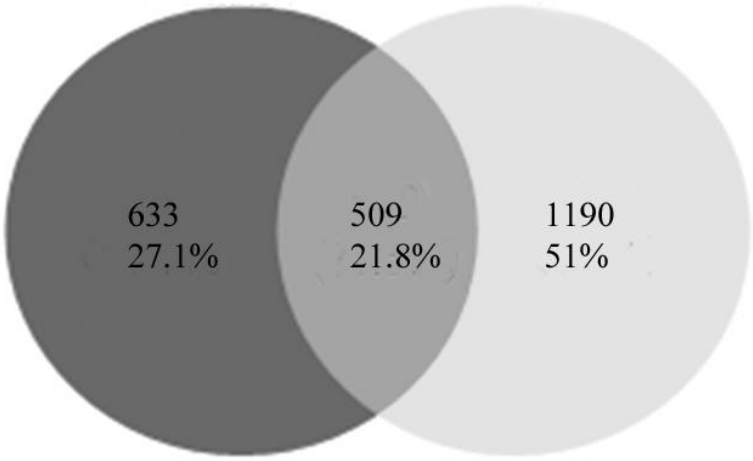


Figure 1. Veen of differentially expression genes in Microarray and RNA-seq

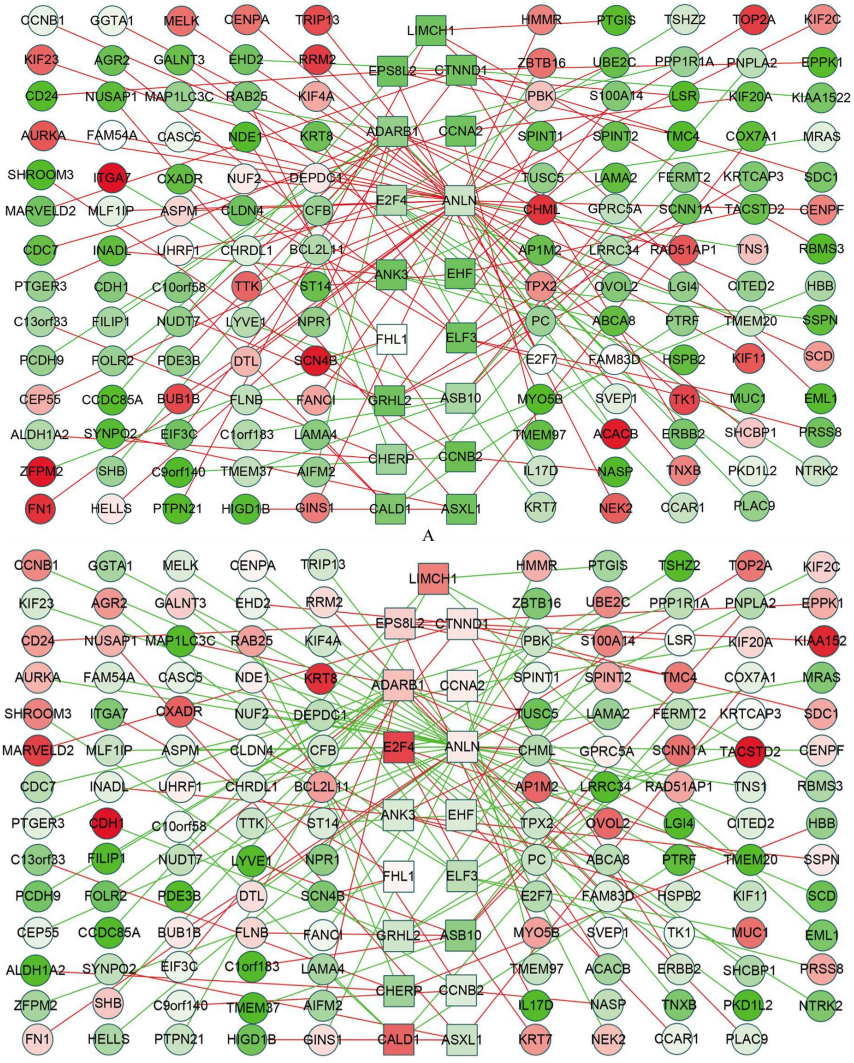
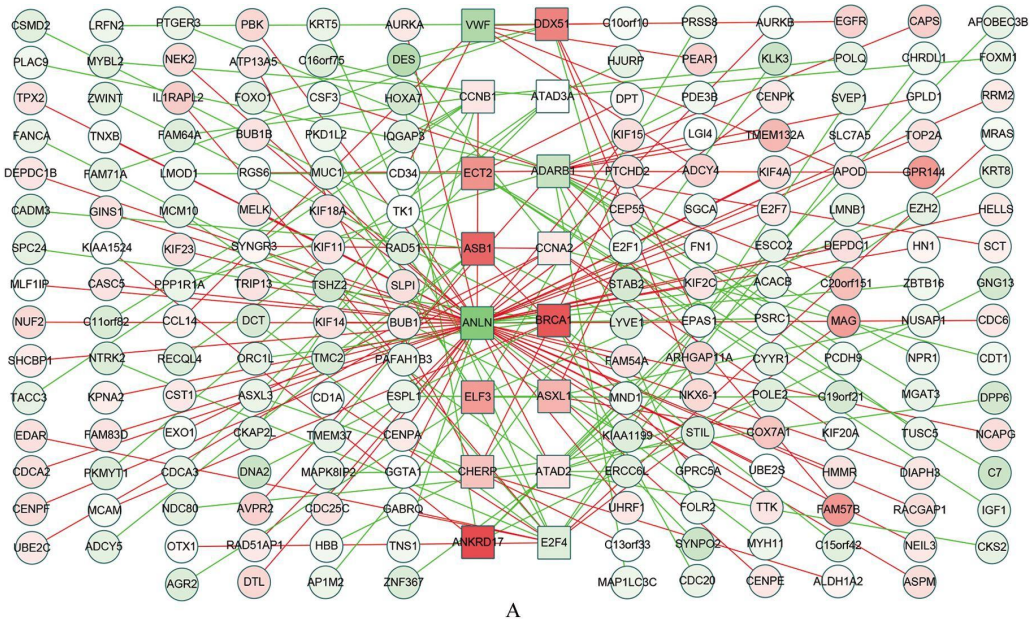
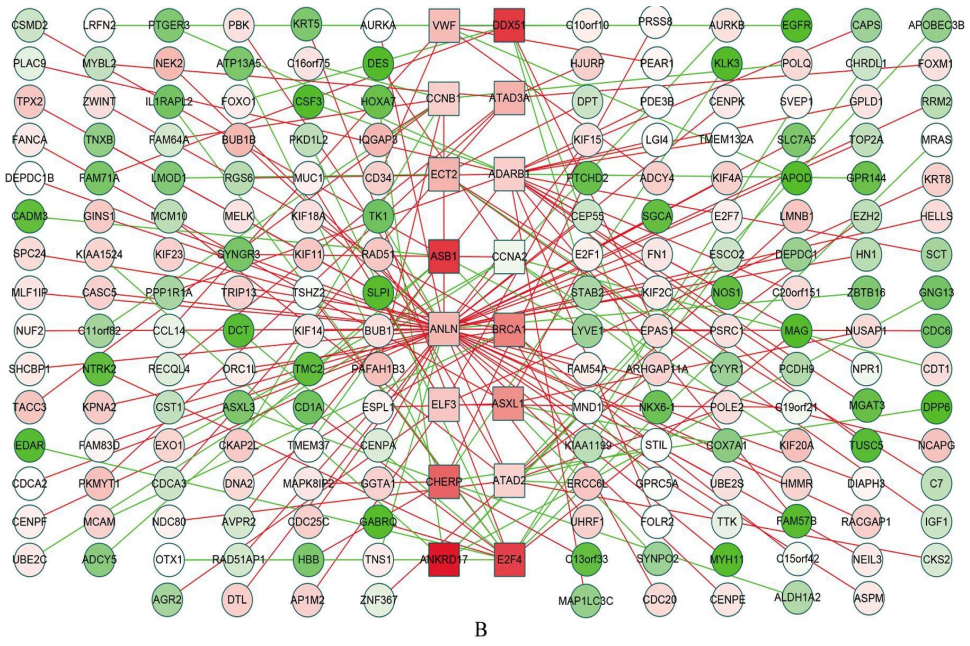


Figure 2. Transcriptional regulatory network of microarray data

Note: A: control samples of microarray data; B: Affected samples of microarray data





B

Figure 3. Transcriptional regulatory network of RNA-seq
Note: A: control samples of RNA-seq; B: Affected samples of RNA-seq

DISCUSS

Due to the two sets of data involved TF and TG Too many, this study chooses Take the common of two sets of data TF analyzed, and through FunRich right target gene go analyze. Common transcription factors share 7 each: ADARB1, ANLN, E2F4, CHERP, ELF3, CCNA2, ASXL1.

pass go Biological processes and biological pathways under analysis Now, in the microarray data, 7 indivual TF Regulation TG participated in 16 indivual biological process and 273 biological pathway, RNA-seq middle TG Co-parameter with 15 a biological process and 317 biological pathways involved in the same

The biological processes and biological pathways are 13 one and 245 one, show The common biological process with strong affinity ($p \leqq 0.05$).

The difference between the two data TG may participate in the same organism learning process, while a TG Can be involved in many different biological processes Cheng.

ADARB1 is a RNA Specific adenosine deaminase (Flanagan et al., 2009), which encodes an enzyme that, through the adenosine site-specific Heterotropic deamination is responsible for the glutamate receptor subunit B before miRNA edit. RNA Editing plays a role in tumor formation and also May interfere with editing pri-mRNA treatment to alter the target site or edit directly miRNA sequence to help or hinder miRNA Function. Research shows that some miRNA Regulation of breast cancer transcription at the post-transcriptional level The expression of transplantation-related genes is closely related to the invasion and metastasis of breast cancer close. Common target genes regulated in both datasets PDE3B Involved Pathway: insulin. As a factor that promotes the development of cancer cells, islets mitogen promotion The occurrence of MAPK (Bishop et al., 2014), increasing the mitotic rate of cells and affecting the development of tumors. The insulin pathway plays an important role in the pathogenesis of breast cancer.

ANLN is a protein required for cytokinesis (Magnusson et al., 2016). ANLN is associated with an increased risk of breast cancer recurrence Genes, in cancer development, growth, ANLN Increased expression value; Overexpression of ANLN not only induced cell growth but also enhanced cell cell migration ability. in the microarray data and RNA-seq Regulated Gene PBK and KIF15 Both are involved in a biological process: signal transduction. Many cellular events and physiological responses are involved in tumor development, signaling Abnormal expression of transduction pathways has a significant impact on it. MAPK Signal Transduction pathways are closely related to cell survival and anti-apoptosis (Yao Qing et al., 2004), in CHERP Among the genes regulated, MA PK8IP2 Involved This pathway, the abnormal expression of signaling molecules in the pathway, promotes cell The excessive proliferation of cells affects their normal transformation. Studies have found that signal transduction Abnormal conduction pathways are closely related to a variety of tumors.

Some findings suggest that E2F4 may regulate E2F dependency switch play an important and unique role in transcription and cell growth (Wang et al., 2000). with other E2F gene is different, E2F4 throughout the cell cycle Expressed constitutively, even in resting cells. E2F4 is suffering from The expression value was significantly increased in disease samples, and the overexpression of E2F4 was found in some aspect simulated when Rb Occurs when loss of function E2F4 activation of Rb Loss of function leads to increased apoptosis and tumorigenesis.

CHERP of surface reach live sex exist suffer from sick Sample Book middle quilt suppress system, CHERP is

associated with neuroblastoma proliferation and apoptosis, in microarray Genes regulated in the data BCL2L11 Involved in colorectal cancer, non-alcoholic fatty liver disease and other pathways; target gene CHML Involved in Alzheimer's Haimer's disease pathway; ASXL1 Regulated target gene COX7A 1 Involved Parkinson's disease, Huntington's disease and other pathways. These findings suggest that BC and AD or other diseases are not independent of each other, their pathogenesis can be There can be an internal connection.

ELF3 Target genes regulated in microarray data CLDN4 reference and cell adhesion molecule pathway, expression of cell adhesion factors and cancer

There were significant correlations between differentiation, invasion, and metastasis of the disease, as a function of estimated mammary log-counts. The LOWESS curve is statistically robust and provides a trendline through most standard deviations. It can be used as a reference index for judging the metastasis and prognosis of breast cancer; exist RNA-seq target genes regulated in C19orf21 involved in metabolism Pathways, metabolic disorders increasingly threaten human health, estrogen, aromatization Both enzymes contribute to the initiation and progression of postmenopausal breast cancer. adenylate activation Protein kinases are negative regulators of aromatase, through metabolic pathways that can capable of inhibiting cancer cell proliferation. It can be seen that metabolism and the development of breast cancer closely related.

CCNA2 common target gene UBE2C Participated in the pathway: Ubiquitin-mediated proteolysis. ubiquitin-proteasome system-mediated protein The research on the relationship between hydrolysis and the pathogenesis of breast cancer is a research topic in recent years. hotspot. The proteasome distinguishes and degrades ubiquitinated protein substrates substances to change their levels (Deng Shishan, 2008, Journal of North Sichuan Medical College, 23(6): 553-556). Ubiquitination in breast cancer tissue compared with benign cpm value y gi and associated weights input to limma in the standard linear modeling and empirical Bayesian differential expression analysis pipelines.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

Barabasi AL, and Oltvai ZN, 2004, network biology: understand ing the cell's functional organization, Nature Reviews Ge- netics, 5(2): 101-113.

Bishop EA, Lightfoot S., Thavathiru E., and Benbrook DM, 2014, Insulin exerts direct effects on carcinogenic transform- mation of human endometrial organotypic cultures, Cancer Investigation, 32(3): 63-70.

Chang C., Ding Z., Hung YS, and Fung PC, 2008, Fast net work component analysis (fastnca) for gene regulatory net- work reconstruction from microarray data, Bioinformatics, 24(11): 1349-1358.

Flanagan JM, fun JM, Henderson S., Wild L., Carey N., and Boshoff C., 2009, Genomics screen in transformed stem cells reveal RNASEH2A, PPAP2C, and ADARB1 as puta- tive anticancer drug targets, Molecular Cancer Therapeutics, 8(1): 249-260.

Law CW, Chen Y., Shi W., and Smyth GK, 2014, Voom: pre- decision weights unlock linear model analysis tools for RNA-seq read counts, Genome Biology, 15(2): R29.

M a gn u ss o n K., G r e m e l G., Ryan L., Po n t e n V., U h l e n M., D i m berg A., Jirstrom K., and Ponten F., 2016, ANLN is a prog- nostic biomarker independent of Ki - 67 and essential for cell cycle progression in primary breast cancer, BMC Cancer, 16(1): 904

Qi YX, Liu YB, and Rong WH, 2011, RNA-Seq and its ap- applications: a new technology for transcriptomics, Yichuan (Hereditas), 33(11): 1191-1202.

Soneson C., and Delorenzi M., 2013, A comparison of methods For differential expression analysis of RNA-seq data, BMC Bioinformatics, 14(1): 91.

Wang D., Russell JL, and Johnson DG, 2000, E2F4 and E2F1 have similar proliferative properties but different apoptotic and oncogenic properties in vivo, Molecular & Cellular Bi- ology, 20(10): 3417-3424.

Yao Q., Luo JR, Chen JH, Zhang JL, Yuan SF, Ling R., and Wang L., 2004, Expression and activation of MAPK path- way signaling molecules in human breast cancer cell lines, Xibao Yu Fenzi Mianyixue Zazhi (Chinese Journal of Cellular and Molecular Immunology), 20 (3): 328-330.

Ye C., Galbraith SJ, Liao JC, and Eskin E., 2009, Using net- work component analysis to dissect regulatory networks me- diated by transcription factors in yeast, PLos Computational Biology, 5(3): e1000311.

Zheng LH, Zhang MM, Zhao YH, and Liu YJ, 2014, Progress in the development of metabolic and breast cancer, Guoji Waikexue Zazhi (International Journal of Surgery), 41(6): 420-423.