GENE-GENE INTERACTION ANALYSIS IN ALZHEIMER

Rishi Yadav, Ravi Bhushan Mishra

Computer Science & Engineering, IIT BHU (Varanasi)

ABSTRACT

Genome-wide transcription profiling is a powerful technique in studying disease susceptible footprints. Moreover, when applied to disease tissue it may reveal quantitative and qualitative alterations in gene expression that give information on the context or underlying basis for the disease and may provide a new diagnostic approach. However, the data obtained from high-density microarrays is highly complex and poses considerable challenges in data mining. Past researches prove that neuro diseases damage the brain network interaction, protein- protein interaction and gene-gene interaction. A number of neurological research paper also analyze the relationship among damaged part. Analysis of gene-gene interaction network drawn by using state-of-the-art gene database of Alzheimer's patient can conclude a lot of information. In this paper we used gene dataset affected with Alzheimer's disease and normal patient's dataset from NCBI databank. After proper processing the .CEL affymetrix data using RMA, we use the processed data to find gene interaction outputs. Then we filter the output files using probe set filtering attributes p-value and fold count and draw a gene-gene interaction network. Then we analyze the interaction network using GeneMania software.

KEYWORDS

Gene-gene interaction, Dementia, Alzheimer's disease, RMA, GeneMania

1. INTRODUCTION

Alzheimer's disease (AD) is a chronic, incurable, irreversible neurodegenerative disorder and multifaceted disease which along with other neurodegenerative diseases, represents the largest area of unmet need of modern medical science[2]. The cognitive decline caused by this disorder leads to behavioral change, loss of memory and thinking skills and ultimately effects the simpler task performance. The disease begins with mild deterioration and ultimately worse in neurodegenerative type of dementia. Diagnosis of Alzheimer's disease is possible but it is very costly, time-taken and medically infeasible. It requires very careful medical assessment such as patient history, behavioral pattern analysis and different physical and neurobiological exam.

Bioinformatics field providing a great support to modern medical science. Bioinformatics is the application of tools of computation and to the capture and interpretation of biological data. A lot of computational methods have been used to assist modern biology. Computational tools are very efficient in recognizing pattern of diseases, handing big sample data, analyzing and interpreting the sample data. Neurodegenerative diseases like Alzheimer's disease (AD), Parkinson's diseases etc. fallows a great patterns in patients, which can be analyzed by machine learning techniques, computational tools and software.

DOI: 10.5121/ijcsit.2018.10307

Alzheimer's disease neuronal dysfunction mainly causes due to failure of functional integration and damage in neural connectivity network **[8].** Brain is the center of human nervous system which is made of 100 billion nerves that communicate in trillions of connections called synapses. In past three decades, a rich study on recognizing neural pattern have provided plenty of knowledge about defected neural connection in AD. A number of computational tools and software have developed to collect this information **[7].**

Sarraf et al.,[8] used convolutional neural network (CNN) to classify Alzheimer brain and normal healthy brain. They used MRI slices of Alzheimer patient and normal healthy brain from ADNI databank and pre-processed the image using the standard modules of FMRIB library v5.0. Images were labelled for binary classification of Alzheimer's Vs normal dataset and these labelled images were converted to Imdb storage databases for higher throughput to be fed into Deep learning platform. The study used CNN and famous architecture LeNet-5 and successfully classified fMRI data of Alzheimer's patient from normal controls. The accuracy of test data on training data reached 96.85 %, which was trained and tested with huge number of images. This study also showed a novel path to use more complicated computational architecture and classifiers to increase the efficient and effective pre-clinical diagnosis of AD.

There are lot of genetic factors that influence common and complex traits human behavior. The characterization of the effects of those factors is both a goal and challenge for modern geneticists. In the last couple of years, the field has been revolutionized by the success of genome-wide association (GWA) studies **[10] [11].** Gene is a basic physical and functional unit of heredity. Gene made up of DNA, act as instructions to make molecules called protein. The one goal of modern geneticists is to identify genes with specific DNA sequence variations that increase or decrease susceptibility of disease **[5].** In past years many researches focused on finding genetic architecture of diseases. Gene-gene interaction or epistasis is a ubiquitous component of the genetic architecture of common human diseases. **Musameh et al., [4]** aimed to investigate the role of gene-gene interactions in common variants in candidate cardiovascular genes in coronary artery disease (CAD).

Meng et al., [3] explored gene-gene interactions that have the potential to partially fill the missing heritability. Though in the previous studies interesting genes have been found, they identified such interesting genes which had no strong main effect on hypertension but which having indirect effect on hypertension prone genes. The paper provides evidence that the genome-wide gene-gene interaction analysis has the possibility to identify new susceptibility genes, which can provide more knowledge and insights into the genetic pattern of blood pressure regulation and its effect on hypertension.

Gilbert-Diamond et al., [5] introduced the basics and importance of gene-gene interaction in identifying disease susceptibility genes. The paper analyzed several statistical and computational methods for characterizing and detecting genes with effects that are dependent on other genes. They focused on genetic association studies of discrete and quantitative traits because most of the methods for detecting gene-gene interactions have been developed by using these methods. The paper introduced the gene-gene interaction, use of gene-gene interaction and challenges in this research field. They described methods for detecting gene-gene interaction in association studies of discrete traits like logistic regression and multifactor dimensionality reduction. In association studies of quantitative traits methods like linear regression and combinational partitioning method

also discussed. Lastly paper also proved the advantage of genome wide approach in place of traditional candidate-gene approach.

Gene-phenotype relationship is so complex that machine learning approaches have a considerable appeal as a strategy foe modelling interactions. A number of such methods have been developed and applied in recent years with some modest success. **Koo et al., [1]** presented an overview of several machine learning techniques to solve gene-gene interaction problem in genetic epidemiology. Traditional statistical methods are not efficient due to the curse of high dimensionality of data and occurrence of multiple polymorphism. This paper gives an overview of machine learning techniques support vector machine (SVM), random forest (RFs) and neural networks (NNs) and its application in detecting gene environment interactions. They also analyzed machine learning techniques which are most suitable to implement on gene-gene interaction like genetic programming neural network (GPNN), back propagation neural network (BPNN), grammatical evolution neural network (GENN), Recursive feature addition (SVM-RFA), Recursive feature elimination (SVM-RFE), local search (SVM- Local), genetic algorithm (SVM- GA) and mutual information network guided RF method (MINGRF). Lastly this paper also discussed the strengths and weakness of all the variants of Support vector machine, random forest and neural networks in implementing and detecting gene-gene interaction environment in complex human diseases.

Nowadays genome-wide association studies have offered thousands of single nucleotide polymorphism (SNPs). This study is important to unravel the genetic basis to complex multifactorial diseases. The greatest challenge is to discover gene-gene interactions among large amount of data having multiple polymorphisms. Thus various promising software have been developed for epistasis interactions. **Sirava et al., [2]** developed a tool for modeling, analyzing and visualizing biochemical pathways and networks of genomes and array data. The software tool BioMiner is based on new comprehensive, extensible and reusable data model called BioCore. The paper presented two applications, PathFinder and PathViewer. PathFinder predicting biochemical pathways by comparing groups of related organisms based on sequence similarity and successfully tested in number of experiments. PathViwer is an application for visualizing metabolic networks and supports the graphical comparison of metabolic networks of different organisms. **Koo et al., [12]** gives an overview on the software that had been used to detect gene-gene interactions that bring the effect on common and multifactorial diseases. Lastly in this paper sources, link, strength and weakness of the software that has been widely used in detecting epistasis interactions in complex human diseases also analyzed.

We used state-of-the-art gene dataset id GDS4758 from The National Centre for Biotechnology Information (NCBI), which provides free access to biomedical and genomic datasets. Dataset consists of 79 sample counts postmortem brain tissues pathologically diagnosed as having AD and normal patient. Dataset is in .CEL affymatrix format. We used most efficient method for pre-processing of microarray data, normalization in R using Robust Multi-array average or Robust Multi-chip average (RMA) method. We use AltAnalyze software for microarray data processing, which includes several basic expression. Statistics are comprised of following steps- rawp, adjp, log-fold, fold change, ANOVA rawp, ANOVA adjp and max log-fold. Then we use probe set filtering in output files of AltAnalyze using p-value (<0.05) and fold count (>2). We get 20 interesting genes. Then we analyze 20 interesting gene network using GeneMania and draw relevant conclusions from these gene-gene interaction network.

The paper's organization is as fallows. Apart from introduction, Section II deals with problem description, Section III deals with proposed methods including data acquisition, data pre-processing, microarray analysis, Section IV deals with results description and Section V represents Conclusion & Discussion.

2. PROBLEM DESCRIPTION

Early diagnosis of neuro disease like Alzheimer's disease (AD) via medical assessment is very costly and time consuming. Gene-gene interaction or epistasis is a ubiquitous component of the genetic architecture of neuro disease. The purpose of this paper is to analyze some interesting findings from gene-gene interaction network of AD patient. We use gene dataset GDS4758 from NCBI as input to gene-gene interaction network.

3. PROPOSED METHOD

3.1 DATA ACQUISITION

The National Centre for Biotechnology information (NCBI) provided access to biomedical and genomic datasets and related information. We take dataset record id GDS4758 from (https://www.ncbi.nlm.nih.gov/geo/download/?acc=GDS4758). This dataset having 79 sample counts of postmortem brain tissues from male and female Hisayama residents pathologically diagnosed as having Alzheimer's disease (AD). Dataset also having normal brain tissues data. Dataset is collected by using the Affymetrix scanner software, which produces a number of file when a gene chip is scanned. Two of these are .CHP and the .CEL files. Dataset GDS4758 contains .CEL files that contains probe level intensities. It contains the data extracted from probes on an Affymetrix gene chip and can store thousands of data points. We used GDS4748 .CEL format 79 sample counts as input dataset.

3.2 DATA PRE-PROCESSING

Microarray gene expressions are little noisy. There are many sources of noise in microarray experiments like different amount of RNA used for labelling and hybridization, imperfectness on the array surface, imperfect synthesis of the probes and difference in hybridization conditions etc. Systematic differences due to noise in place of true biological variability should be removed in order to make biologically meaningful conclusions about the data. There are many techniques for the pre-processing of microarray data**. R hasan et al., [13]** discussed some of the most important preprocessing techniques applicable to Affymetrix microarray platform.

One of the most efficient method for pre-processing microarray data is normalization in R using Robust Multi-array average or Robust Multi-chip average (RMA). RMA method for computing an expression measure begins by computing back-ground correlated perfect match intensities for each perfect match cell on every GeneChip. Background-corrected intensity is calculated for each PM probe in such a way that all background corrected intensities are positive. After that base-2 logarithm of the background- corrected intensity is calculated for each probe. This will male data more skewed and more normally distributed and provide an equal spread of up and down regulated expression ratios. Next step is for quantile normalization to correct for variation between the arrays. It equalizes the data distributions of the arrays and make the samples

completely comparable. After that last step is probe normalization to correct for variation between probe sets, which equalizes the behavior of the probes between the arrays and combines normalized data values of probes from a probe set into a single value for the whole probe set. We use R studio for RMA normalization of .CEL sample counts of GDS4758.

3.3 MICROARRAY ANALYSIS

Microarray gene analysis comprises a number of computational steps. Such analysis having huge amount and diversity of data. There are number of software which analyze microarray gene-gene interactions and produces interaction outputs. We used AltAnalyze - an extremely user friendly and open-source analysis toolkit that can be used for a broad range of genomics analysis. First of all Altanalyze also normalize affymetrix .CEL files using RMA and expression values are computed based on corresponding probe set annotations in the downloaded CDF file which is automatically recognized and downloaded by AltAnalyze.

3.3.1 GENE EXPRESSION ANALYSIS

Several basic expression statistics are performed for any user specified pairwise comparisons (e.g., AD versus normal) and between all groups in the user dataset. These statistics are comprised of following steps- rawp, adjp, log-fold, fold change, ANOVA rawp, ANOVA adjp and max log-fold. Log-fold is log2 fold calculated by geometric subtraction of the experimental from the control groups for each pairwise comparison of AD and normal. The fold change is non-log2 transformed fold value. The max log-field means the log2 fold value between the lowest group mean and the highest group expression mean. The rawp is one one-way analysis of variance p-value calculated for each pairwise comparisons.

3.3.2 RECIPROCAL JUNCTION ANALYSIS

Alternative exon regulation is determined by comparing normalized expression of one exon in two or more conditions for an exon-level analysis. Normalized expression means the expression of an individual divided by the gene-expression value. Altanalyze first identifies reciprocal junctions or pair of exon-junctions to identify alternative exons from junction analysis, where one measuring the inclusion of an exon and the other measuring exclusion of the exon. Altanalyze used ASPIRE, Linear regression and Ri-PSI algorithms for reciprocal junction analysis.

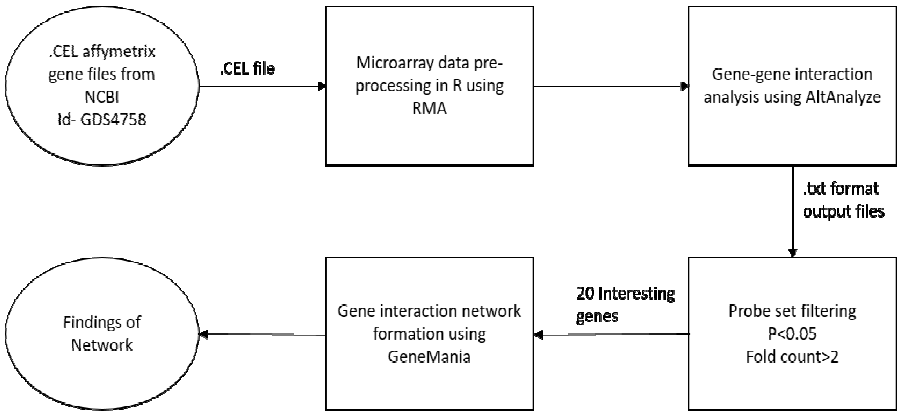


Figure-1 Flowchart of Proposed method

3.4 PROBE-SET FILTERING

During the process of routine analysis of microarray gene interactions, a multiple testing adjustment is certainly warranted due to large number of hypothesis test. Filtering allows us for the reduction in number of tests and a corresponding increase in analysis power. We use p- value, midas p-value, fold count and si p-value as a filter during microarray analysis.

p-value helps in determining significance of microarray gene-gene interaction analysis. P-value is the level of marginal significance within a statistical hypothesis test representing the probability of the occurrence of the given event. In most of the microarray analysis p-value is set to be (< 0.05). p-value (< 0.05) indicates strong evidence against the null hypothesis, so we can reject the null hypothesis. A large p-value (>0.05) indicates weak evidence against the null hypothesis, so we will fail to reject the null hypothesis. Thus p-value approach to hypothesis testing uses the calculated probability to determine whether there is evidence to reject the null hypothesis. In our research we set p-value thresholding as p-value <0.05.

Microarray detection of alternative splicing (Midas) p-value calculate an alternative splicing score based on Midas method. Si p-value means significant p-value. We used filtering limits of <0.05 for both, Midas p-value and si p-value. Another important filter in microarray analysis is fold change. Fold change is a measure describing how much a quantity changes going from an initial to final value. As per [14] fold count (>1.5) or (>2) produces most interesting genes in microarray analysis. Dalman et al., [15] analyzed different combinations of p-value and fold counts in microarray analysis and concluded to use fold count (>2) and p-value (<0.05). We use same filters thresholding in our research work.

4. RESULTS

Microarray gene-gene interaction analysis using Altanalyze produces a five output files in delimited text format that can be opened in Microsoft Excel for better visualization. These files consists of a lot of gene-gene interaction information of input .CEL dataset. File #1 reports gene expression values for each sample and group in your probeset input expression file. The values are derived from probe sets that align to regions of a gene that are common to all transcripts and thus are informative for transcription (unless all probe sets are selected - see "Select expression analysis parameters", above) and expressed above specified background levels. Along with the raw gene expression values, statistics for each indicated comparison (mean expression, folds, t-test p-values) will be included along with gene annotations for that array, including putative microRNA binding sites. This file is analogous to the results file you would have with a typical, non-exon microarray experiment and is saved to the folder "ExpressionOutput".

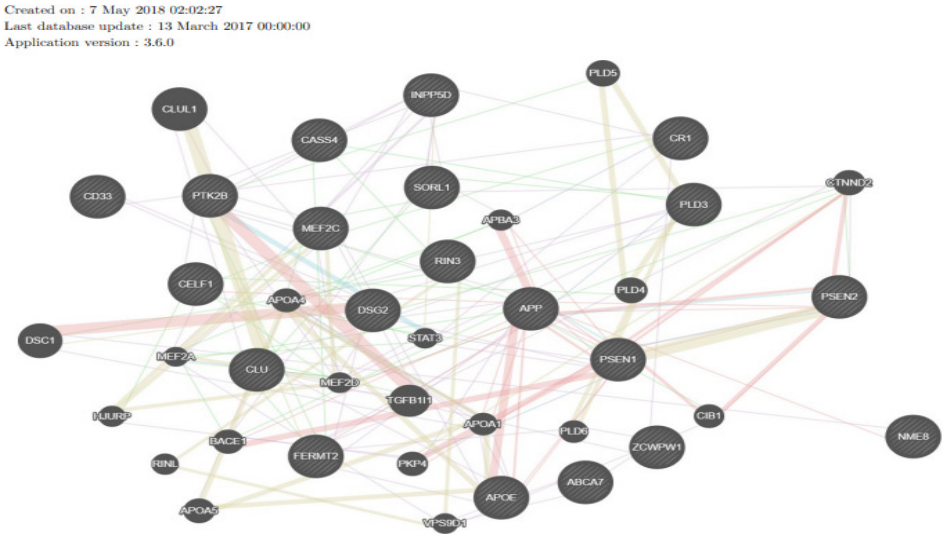
Results from files #2-5 are produced from all probe sets that may suggest alternative splicing, alterative promoter regulation, or any other variation relative to the constitutive gene expression for that gene (derived from comparisons file). Each set of results correspond to a single pair-wise comparison (e.g., cancer vs. normal) and will be named with the group names you assigned (groups file).

File #2 reports probe sets that are alternatively regulated, based on the user defined splicing-index score and p-value. For each probe set several statistics, gene annotations and functional predictions are provided. Files #4 and #5 report over-representation results for protein domains

(or other protein features) and microRNA-binding sites, predicted to be regulated by AltAnalyze. These files include over-representation statistics and genes associated with the different domains or features predicted to be regulated.

File #6 includes the number of genes alternatively regulated, differentiatially expressed and the mean number of protein residues differening between predicted alternative isoforms. This file is useful in comparing results between different pair-wise comparison files. Originally file#2 contains 206 interesting genes analyzed from gene-gene interaction using Altanalysis. We use filters to find out most interesting genes related to Alzheimer's disease dataset. We set p-value (<0.05), Midas p-value (<0.05), si p-value (<0.05) and fold count (>2). On imposing these filters in file#2 , we get 20 most interesting genes related to gene-gene interaction analysis of Alzheimer's disease dataset GDS4758. Now we can draw and analyze the gene-gene interaction network of these most interesting 20 genes to conclude some interesting findings. **Koo et al., [12]** gives an overview of software for visualizing gene-gene network with their respected strength and weaknesses.

We use GeneMania - a flexible, user-friendly web interface for generating hypothesis about gene function, analyzing gene list and prioritizing genes for functional assays. GeneMania takes gene symbols or NCBI gene IDs as input to draw network. We gave 20 Homo Sapiens gene symbols of file#2 as input. All the 20 genes got validation from GeneMania database.



(**Figure-2** Gene-gene interaction network drawn by using GeneMania Software. It consists of 20 interesting genes (having cross-hatched circle) found after proper pre-processing, analysis and filtering. figure also consists 20 relevant genes (having solid circle) having most connectivity with interesting genes)

We get gene-gene interaction network as seen in figure 2. This network consists of 40 genes. 20 genes (having cross-hatched circle) are the most interesting genes from the output file#2 and remaining 20 genes (having solid circle) are relevant genes having most connectivity with interesting genes. Different type of interactions like physical interactions, co-expression, shared protein domains, pathways etc. are represented by different colored connector lines. We can select or deselect to which interactions are displayed. Gene-gene interaction network for input

genes having 33.6% physical interactions with relevant genes. This shows high physical interaction of most interesting 20 genes. All 40 genes having 175 total connector links.

GeneMania also displays functions associated with genes in the network and their FDR and coverage (as number of genes annotated with that function in the network versus number of genes annotated with that function in the genome). In our gene-gene interaction network functions triglyceride-rich lipoprotein particle remodeling and very low density particle remodeling are most effected functions from these 20 interesting genes having coverage ratio 4 out of 11 genes. GeneMania also provide facility of relocation and adjustment of network.

From GeneMania, we export the gene-gene interaction network related pdf and text format files. File GeneMania report provides gene interaction details of the network with gene interaction ranking. Relevant genes CLUL1, TGFB1I1 and PLD5 are top 3 interacting genes of the network.

5. DISCUSSION & CONCLUSION

We used state-of-the-art gene dataset id GDS4758 from The National Centre for Biotechnology Information (NCBI), which provides free access to biomedical and genomic datasets. Dataset consists of 79 sample counts postmortem brain tissues pathologically diagnosed as having AD and normal patient. Dataset is in .CEL affymatrix format. We used most efficient method for pre-processing of microarray data, normalization in R using Robust Multi-array average or Robust Multi-chip average (RMA) method. We use AltAnalyze software for microarray data processing, which includes several basic expression. Statistics are comprised of following steps- rawp, adjp, log-fold, fold change, ANOVA rawp, ANOVA adjp and max log-fold. Then we use probe set filtering in output files of AltAnalyze using p-value (<0.05) and fold count (>2). We get 20 interesting genes. Then we analyze 20 interesting gene network using GeneMania and draw relevant conclusions from these gene-gene interaction network.

From the gene-gene interaction network using GeneMania we draw these conclusions- (a) Out of 20 interesting genes some genes are highly correlated with relevant genes. CLUL1, DSC1 and TGFB1I1 are top 3 genes having highest correlation values. Thus these homo sapiens genes required extra care and medical research. (b) Every genotype having special effects on phenotype or body functions. Using GeneMania we analyze the most effected functions are triglyceride-rich lipoprotein particle remodeling and very low-density lipoprotein particle remodeling, which consists 4 interesting genes out of 11 genes. This can help in biological study of Alzheimer's disease.

REFERENCE

[1] Koo, Ching Lee, Mei Jing Liew, Mohd Saberi Mohamad, Mohamed Salleh, and Abdul Hakim. "A review for detecting gene-gene interactions using machine learning methods in genetic epidemiology." BioMed research international 2013 (2013).

[2] Širava, M., T. Schäfer, Markus Eiglsperger, Michael Kaufmann, Oliver Kohlbacher, Erich Bornberg-Bauer, and Hans-Peter Lenhof. "BioMiner—modeling, analyzing, and visualizing biochemical pathways and networks." Bioinformatics 18, no. suppl_2 (2002): S219-S230.

[3] Meng, Ying, Susan Groth, Jill R. Quinn, John Bisognano, and Tong Tong Wu. "An Exploration of Gene-Gene Interactions and Their Effects on Hypertension." International journal of genomics 2017 (2017).

[4] Musameh, Muntaser D., William YS Wang, Christopher P. Nelson, Carla Lluís-Ganella, Radoslaw Debiec, Isaac Subirana, Roberto Elosua et al. "Analysis of gene-gene interactions among common variants in candidate cardiovascular genes in coronary artery disease." PloS one10, no. 2 (2015): e0117684.

[5] Gilbert-Diamond, Diane, and Jason H. Moore. "Analysis of gene-gene interactions." Current protocols in human genetics(2011): 1-14.

[6] Cordell, Heather J. "Detecting gene–gene interactions that underlie human diseases." Nature Reviews Genetics 10, no. 6 (2009): 392.

[7] Biju, K. S., S. S. Alfa, Kavya Lal, Alvia Antony, and M. Kurup Akhil. "Alzheimer's Detection Based on Segmentation of MRI Image." Procedia Computer Science 115 (2017): 474-481.

[8] Sarraf, Saman, and Ghassem Tofighi. "Classification of alzheimer's disease using fmri data and deep learning convolutional neural networks." arXiv preprint arXiv:1603.08631 (2016).

[9] Caspi, Avshalom, Ahmad R. Hariri, Andrew Holmes, Rudolf Uher, and Terrie E. Moffitt. "Genetic sensitivity to the environment: the case of the serotonin transporter gene and its implications for studying complex diseases and traits." Focus 8, no. 3 (2010): 398-416.

[10] Wellcome Trust Case Control Consortium. "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." Nature 447, no. 7145 (2007): 661.

[11] Easton, Douglas F., Karen A. Pooley, Alison M. Dunning, Paul DP Pharoah, Deborah Thompson, Dennis G. Ballinger, Jeffery P. Struewing et al. "Genome-wide association study identifies novel breast cancer susceptibility loci." Nature 447, no. 7148 (2007): 1087.

[12] Koo, Ching Lee, Mei Jing Liew, Mohd Saberi Mohamad, Abdul Hakim Mohamed Salleh, Safaai Deris, Zuwairie Ibrahim, Bambang Susilo, Yusuf Hendrawan, and Agustin Krisna Wardani. "Software for detecting gene-gene interactions in genome wide association studies." Biotechnology and bioprocess engineering 20, no. 4 (2015): 662-676.

[13] R Hasan, Ahmed, John E Pattison, and Alex Hariz. "Pre-processing of affymetrix gene chip microarray data." Current Bioinformatics 5, no. 4 (2010): 270-279.

[14] Vaes, Evelien, Mona Khan, and Peter Mombaerts. "Statistical analysis of differential gene expression relative to a fold change threshold on NanoString data of mouse odorant receptor genes." BMC bioinformatics 15, no. 1 (2014): 39.

[15] Dalman, Mark R., Anthony Deeter, Gayathri Nimishakavi, and Zhong-Hui Duan. "Fold change and p-value cutoffs significantly alter microarray interpretations." In BMC bioinformatics, vol. 13, no. 2, p. S11. BioMed Central, 2012.