Gene 661 (2018) 160–168

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Research paper

Network-based meta-analysis in the identification of biomarkers for papillary thyroid cancer

Hengqiang Zhao[a,b,*], Hehe Li[a]

[a] Department of Pancreatic Surgery, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430022, China
[b] Department of Breast and Thyroid Surgery, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430022, China

ARTICLE INFO

Keywords:
Papillary thyroid cancer
Gene expression omnibus
Microarray analysis
Differentially expressed genes

ABSTRACT

Papillary thyroid carcinoma (PTC) has been increasing across the world with incomplete understanding of its pathogenesis. We aimed to investigate gene alterations and biomarkers contributing to PTC development. A total of five eligible microarray datasets including 94 PTC and 81 normal thyroid samples were included to identify gene expression signatures. Using integrative meta-analysis of expression data (INMEX) program, we identified a total of 2699 differentially expressed genes (DEGs) (1333 overexpressed and 1366 underexpressed genes) in PTC relative to normal thyroid samples. The top 100 upregualted and downregulated DEGs identified in the meta-analysis were further validated in The Cancer Genome Atlas (TCGA) dataset for PTC with high consistency. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis revealed pathways in cancer, proteoglycans in cancer, focal adhesion, axon guidance, and ECM-receptor interaction among the top 5 most enriched pathways. Network-based meta-analysis identified *FN1* and *TRAF6* to be the most highly ranked hub genes among the overexpressed and underexpressed genes, respectively, both of which are involved in pathways in cancer. The most enriched terms for Gene Ontology (GO) of biological processes, cellular component, and molecular function were signal transduction, cytoplasm, and protein binding, respectively. Our meta-analysis comprehensively investigated DEGs, hub genes, enriched pathways and GO terms for PTC, which might provide additional approaches to explore the molecular mechanisms underlying the pathophysiology of PTC, and identify biomarkers and therapeutic targets toward PTC.

1. Introduction

Thyroid cancer is one of the fastest increased cancers worldwide and primarily involves papillary thyroid cancer (PTC) (Kitahara and Sosa, 2016). Papillary thyroid microcarcinoma (PTMC) accounts for the majority of the increase, which might be attributed to overdiagnosis and overtreatment (Vigneri et al., 2015; Lubitz and Sosa, 2016; Sanabria et al., 2017). Many potential risk factors for thyroid cancer have been explored like radiation exposure, obesity, iodine intake, thyroid-associated diseases and social economic status (Pellegriti et al., 2013; Pappa and Alevizaki, 2014; Zhao et al., 2014; Sanabria et al., 2017; Zhao et al., 2017), however, few well-established risk markers have been identified.

In the last few decades, high-throughput experiments such as gene expression microarrays have expanded our understanding of the underlying mechanisms between cancer development and genomic background. Microarray-based research has generated large databases of genome-wide gene expression data that are deposited in public archives (Rung and Brazma, 2013). Despite their great promise, the predictive performance of microarray-based studies was different (Ntzani and Ioannidis, 2003), or not robust across multiple random datasets (Michiels et al., 2005). The small sample size can lead to inflated, unrealistic results relative to tens of thousands of investigated probes, which aggravates the situation. Combining available microarray datasets from data repositories can increase the reliability and generalizability of statistical results. Meta-analysis offers the opportunity to significantly increase the statistical power, and produce more robust and accurate findings, obtaining a more precise judgement of differentially expressed genes (DEGs) (Ramasamy et al., 2008). In addition, the development of bioinformatics tools and methods for

Abbreviations: BP, biological process; CC, cellular component; DAVID, database for annotation, visualization and integrated discovery; DEGs, differentially expressed genes; ES, effect size; GEO, gene expression omnibus; GO, gene ontology; INMEX, integrative meta-analysis of expression data; KEGG, Kyoto Encyclopedia of Genes and Genomes; MF, molecular function; NUSE, normalized unscaled standard errors; PTC, papillary thyroid carcinoma; REM, random effect modeling; RLE, relative log expression; RMA, robust multi-array average; TCGA, the cancer genome atlas

* Corresponding author at: Department of Pancreatic Surgery, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430022, China.
E-mail addresses: zhaochewh@hust.edu.cn (H. Zhao), lihehe@hust.edu.cn (H. Li).

https://doi.org/10.1016/j.gene.2018.03.096
Received 6 January 2018; Received in revised form 13 March 2018; Accepted 29 March 2018
Available online 03 April 2018

microarray data drives data integration and visualization from lists of DEGs (Shannon et al., 2003; Maere et al., 2005; Pirooznia et al., 2007; Bindea et al., 2009; Xia et al., 2013; Xia et al., 2014). These tools support the prediction of functional pathways and protein-protein interactions (PPI)-incorporated network analyses (Creixell et al., 2015). Integrative meta-analysis of expression data (INMEX) allows researchers to integrate multiple microarray datasets (Xia et al., 2013; Xia et al., 2014), which has been widely used in microarray-based meta-analysis (Hounkpe et al., 2015; Santiago and Potashkin, 2015).

To get a more robust prediction of genetic background variations implicated in PTC pathophysiology, we perform the transcriptomic meta-analysis with five eligible microarray datasets to identify DEGs in PTC relative to normal thyroid tissues. Clustering analysis and PPI network were performed on these DEGs. The results may guide us to find new markers and therapeutic targets for PTC.

2. Materials and methods

2.1. Inclusion of microarray data for PTC

The gene expression data of PTC were downloaded from Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/). The search strategy was as follows: “thyroid” (title) and “homo sapiens” (organism) and “expression profiling by array” (filter). Inclusion criteria: (a) it profiled gene expression levels between PTC and normal thyroid tissues; (b) it contained 3 samples at least per group. We excluded studies examining non-human tissues or cell lines, and samples not fulfilling the microarray quality assessment. A total of five microarray studies met the inclusion criteria and were considered for subsequent analysis. GEO accession number, platform information, and sample number were recorded.

2.2. Microarray quality assessment and adjustment for batch effect

Raw data (.CEL) (downloaded from GEO) were processed using software package in R (version 3.4.2) for microarray quality assessment such as relative log expression (RLE), normalized unscaled standard errors (NUSE), and RNA degradation plots. Samples with abnormal distribution in each dataset were excluded. All datasets were preprocessed using robust multi-array average (RMA) algorithm with background-adjusted, normalized, and log transformation (Irizarry et al., 2003). Tables containing relative expression values were constructed using R and uploaded to INMEX.

The ComBat option in INMEX was used to adjust for batch effects between studies using empirical byes methods (Johnson et al., 2007). The ComBat procedure was robust to outliers in small sample size. Principal component analysis (PCA) was performed to visualize the sample clustering patterns. To address the differences in study design and platform usage, heterogeneity among microarray datasets, the effect size (ES) (the difference between two group means divided by standard deviation) in combination with random effect modeling (REM) was selected to generate more biologically consistent results (Marot et al., 2009).

2.3. Network-based meta-analysis for identification of DEGs

We conducted the microarray-based meta-analysis using INMEX to incorporate multiple gene expression datasets (Xia et al., 2013), in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines for meta-analysis (Moher et al., 2010). Each eligible dataset was uploaded, processed, and annotated to ensure the consistence of data format and class labels across all datasets. After dataset integrity check, a REM was selected for the meta-analysis according to the between-study heterogeneity based on Cochran's Q test (Xia et al., 2013). The GeneVenn web tool was used to compare differentially expressed genes (DEGs) from different microarrays and the

Table 1
Characteristics of the datasets included in the integrated analysis.

GEO ID	Platform	N vs PTC[a]	
GSE3467	GPL570; Affymetrix Human Genome U133 Plus 2.0 Array	9	8
GSE3678	GPL570; Affymetrix Human Genome U133 Plus 2.0 Array	7	7
GSE29265	GPL570; Affymetrix Human Genome U133 Plus 2.0 Array	20	20
GSE53157	GPL570; Affymetrix Human Genome U133 Plus 2.0 Array	3	14
GSE33630	GPL570; Affymetrix Human Genome U133 Plus 2.0 Array	42	45

GEO: Gene Expression Omnibus; N: normal thyroid tissue; PTC: papillary thyroid cancer.
[a] Number of eligible samples from each dataset.

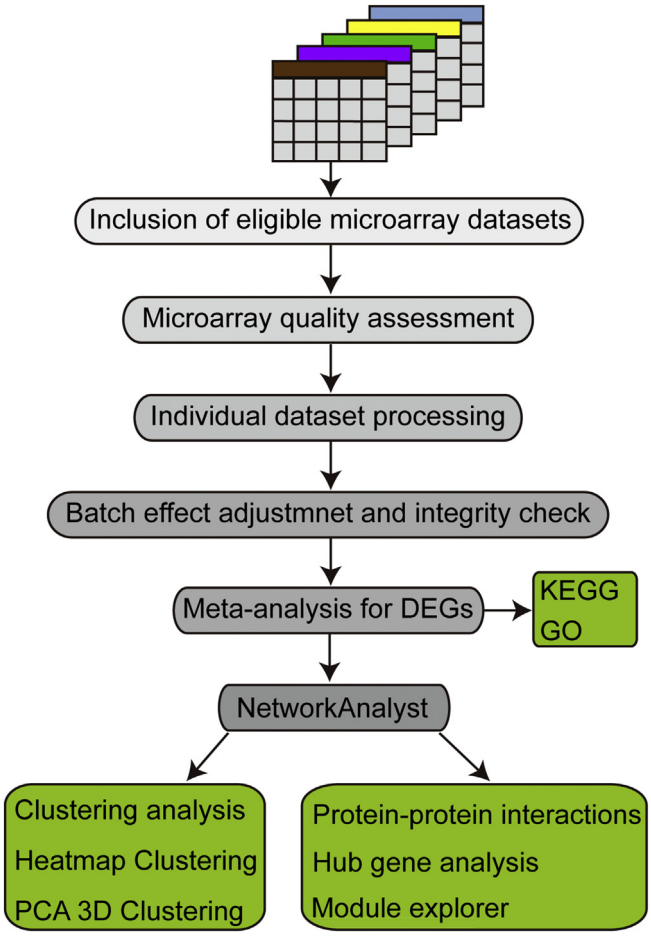

Fig. 1. Workflow of the network-based microarray meta-analysis.
The flow chart of the microarray-based meta-analysis. Five independent microarray raw data were obtained from gene expression omnibus (GEO) and underwent microarray quality assessment. The individual dataset was preprocessed in R and uploaded in INMEX where the meta-analysis was undertaken. After batch effect adjustment and dataset integrity check, the meta-analysis was performed with a random effect model to define differentially expressed genes (DEGs) in papillary thyroid cancer (PTC) relative to normal thyroid tissues. Finally, DEGs identified in the meta-analysis were visualized with clustering analysis and protein-protein interactions (PPI). In addition, Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses were performed on DEGs using Database for Annotation, Visualization and Integrated Discovery (DAVID).