Journal of Functional Foods 42 (2018) 380–386

Contents lists available at ScienceDirect

Journal of Functional Foods

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

Data mining of nutrigenomics experiments: Identification of a cancer protective gene signature

Roberto Martín-Hernández[a,*], Guillermo Reglero[b,c], Alberto Dávalos[d]

[a] Bioinformatics Unit, IMDEA Food Institute, CEI UAM + CSIC, Madrid 28049, Spain
[b] Sección Departamental de Ciencias de la Alimentación, Facultad de Ciencias, Universidad Autónoma de Madrid, Madrid 28049, Spain
[c] Laboratory of Food Products for Precision Nutrition, IMDEA Food Institute, CEI UAM + CSIC, Madrid 28049, Spain
[d] Laboratory of Epigenetics of Lipid Metabolism, IMDEA Food Institute, CEI UAM + CSIC, Madrid 28049, Spain

ARTICLE INFO

Keywords:
Nutrigenomics
Microarrays
Clustering
Gene signature
Anticancer

ABSTRACT

Regular consumption of certain foods has shown beneficial effects on cardiometabolic health. However, it is not clear by which molecular mechanisms they may exert their beneficial effects. Many genomic experiments available in public databases have generated gene expression data following the treatment of human cells with different food nutrients. Exploration of such data offers great possibilities for gaining insights into the molecular effects of nutrients at cellular level. In this work, we explored the genomic responses triggered by food bioactive compounds with well-known healthy properties. We show that human cell lines treated with different food compounds tend to cluster in a cell type dependent manner based on gene expression, with an influence of the physiological attributes of cells. Finally, we identify a genomic signature of 18 genes implicated in cell cycle, which may characterize a protective effect of certain food compounds against cancer. Our data provides evidence that nutrigenomic studies found in public databases can be used to discover novel signatures of gene expression and identify common mechanism of actions of food bioactive compounds.

1. Introduction

Nutritional genomics, also known as nutrigenomics, is a relatively new science which explores the effects of nutrients on the genome, proteome and metabolome. Whereas the idea of modulating human health by food intake is a millennial concept, there are great expectations on the tremendous potential this science may have to change the future of dietary guidelines in order to improve health and hence to build up a precision nutrition era (DeBusk, Fogarty, Ordovas, & Kornman, 2005).

A functional food has been defined as “any modified food or food ingredient that may provide a health benefit beyond that of the traditional nutrients it contains” (Snetselaar, 1994). During the last 20 years there have been substantial efforts to identify bioactive compounds in food which might be associated with beneficial biological activities. For example compounds such as long-chain polyunsaturated fatty acids (n-3 PUFAS), which consumption has been associated with a reduced risk of cardiovascular disease, are known to act as ligands for cellular receptors to trigger a signaling cascade that inhibits the expression of proinflammatory genes (Ferguson, 2009). Also, many natural products, extracted from foods used in human diet, have shown great potential as anti-proliferative agents on cultured cancer human cells (Gonzalez-Vallinas et al., 2013; Ramirez de Molina et al., 2015). Indeed, a wide range of drugs for treating diseases such as diabetes and cancer are derived from natural products. Interestingly, some food compounds have proved their ability to interact with the epigenome, thus modifying microRNA expression (Gil-Zamorano et al., 2014). However, the molecular mechanisms by which food bioactive compounds exert their beneficial effects are still not well understood.

Omics technologies are widely adopted to study the expression of thousands of genes and proteins at a time. These technologies generate a vast amount of gene expression data that accumulates in public repositories such as the NCBI Gene Expression Omnibus (GEO) (Barrett et al., 2013). Whereas these data remains unclassified by phenotype or experimental condition, the user interface allows easily querying and mining the database for experiments.

Other databases such as the Broad Institute's Connectivity Map (CMap) (Lamb et al., 2006) collect highly specific expression data from cell lines treated with drugs and other chemicals. Such type of transcriptomic data has previously been utilized to establish functional

Abbreviations: GEO, NCBI Gene Expression Omnibus; GES, Gene Expression Signature; TCT, Tocotrienols; LB, Lactobacillus; AMF, Amorfrutin; I3C, Indole-3-carbinol; RSM, Rosemary; CA, Carnosic Acid; WFNA, Withaferin A; SFN, Sulforaphane; RVT, Resveratrol; TRVT, Transresveratrol; FC, Fold change

* Corresponding author at: Bioinformatics Unit, IMDEA Food Institute, CEI UAM + CSIC, Ctra. De Canto Blanco 8, E-28049 Madrid, Spain.
E-mail address: roberto.martin@imdea.org (R. Martín-Hernández).

https://doi.org/10.1016/j.jff.2018.01.021
Received 30 October 2017; Received in revised form 19 January 2018; Accepted 20 January 2018

connections between drugs, genes and diseases using computational approaches. From a transcriptomic point of view, mathematical models have already been applied on gene expressions data for the identification of pathway responsiveness to drugs (Pratanwanich & Lio, 2014). Another approach allows the generation of a list of drugs triggering a similar gene expression pattern at cellular level (Lee et al., 2012) and thus possibly sharing a common mechanism of action. From a disease point of view, other approaches consider that a particular gene expression signature (GES) related to a disease might be reverted using a drug which triggers an opposite GES (Setoain et al., 2015), showing promising opportunities within the drug repositioning field (Jia et al., 2016). Artificial intelligence, and specifically deep learning algorithms, has been applied on large transcriptional response data sets with the aim of classifying various drugs to therapeutic categories solely based on their transcriptional profiles (Aliper et al., 2016). However, to the best of our knowledge there is no evidence about such approaches applied to the emerging field of nutrigenomic studies, seeking to investigate the effect of food and nutrients on gene expression.

We extracted from GEO repository all the available experiments related to nutrigenomics in human cells to survey the gene expression patterns. The correlation of gene expression patterns can show potential connections between bioactive compounds, indicating that they may share a common mechanism of action, and allowing the discovery of new potential therapeutic molecules (Lamb et al., 2006). Here we present a comprehensive data mining analysis of a set of nutrigenomics experiments extracted from GEO database. The assessment of human cell's gene expression cultured in vitro after treatment with bioactive compounds obtained from food should lead to a better characterization of the molecular mechanisms that confer a beneficial effect to certain food products.

2. Materials and methods

2.1. Data collection and analysis

Studies corresponding to nutrigenomics were identified from GEO database. Specific queries were launched containing words such as “nutrient”, “nutrition”, “natural product”, “extract” and “phytochemical”. For data corresponding to Affymetrix platforms, raw data was downloaded and normalized locally with the RMA algorithm using specific Bioconductor packages. For data generated by other platforms, the normalized matrix was directly downloaded for analysis. Gene differential expression was assessed using LIMMA package from Bioconductor.

2.2. Hierarchical clustering

A hierarchical clustering algorithm was applied using gene’s log2 fold change (FC) from each analyzed experiment as input values. A distance matrix was computed among all the experiments within the database, using the Euclidean distance as a metric. The agglomeration method of the clustering process was set to complete. Heatmaps.2 library was used for dendrogram and heatmap generation. All the statistical computations were performed using R software. To evaluate the batch effects presence, normalized gene average expression data for each experiment was used as data input for hierarchical clustering analysis.

2.3. Functional enrichment

Genecodis3 software was used for functional enrichment using default parameters and selection of GO Biological Process as target annotations.

2.4. Statistical analysis

Moderated *t*-test statistics were applied to microarray features once a linear model was fitted. Statistical significance of the overrepresented GO biological processes in our target gene list was obtained with chi-square test. False discovery rate (FDR) method was employed to adjust the obtained p-values.

3. Results

3.1. Data collection

Experimental gene expression data corresponding to nutrigenomics experiments was identified from GEO database by launching specific queries. Results were filtered in order to obtain gene expression data from *Homo sapiens* as organism, and expression profiling by array as study type. Few of these studies, corresponding to human nutritional interventions with large cohorts, were filtered out since we were strictly interested on experiments performed on cultured cells. We initially identified 71 potential GEO studies (Table S1) to be included in our analysis. Of those, 34 studies were filtered out due to different criteria such as studies corresponding to human interventions, lack of replicates in the experimental designs, expression data obtained with rare or custom arrays, and expression data corresponding to micro RNA’s. We ended up with a set of 37 GEO studies.

3.2. Gene expression analysis workflow

Experiments included in each study were carefully assessed before analysis in accordance with their experimental design, by manually assigning control and perturbation samples. That is to say, for each experiment their appropriate control was obtained within the same study. Subsequently, a common computational analysis workflow applying linear models was used to assess differential expression in each experiment. Finally, microarray features were annotated with Gene Symbol and Entrez gene identifiers to allow cross-platform data integration. Thus, we obtained a database which includes gene differential expression data from 81 comparisons among different compounds, treatments and cell types (Fig. 1) that arise from the 37 GEO studies.

3.3. Cluster analysis

The clustering has been performed using log2 fold change (FC) expression values obtained following the gene expression analysis workflow. After removing missing values and aggregating expression values for duplicate gene ID's, we proceeded to integrate gene expression data from all the microarray platforms used in our database. Our database included expression data obtained from 19 distinct microarray platforms. A first limitation is that only overlapping genes represented in all platforms could be used in our analysis. Therefore, we used the corresponding Entrez gene ID of the features screened in each platform for data integration. Indeed, gene symbols can be hard to match across platforms because of the continuous updates of gene names, as well as the many to one relationship issues where different gene symbols might correspond to the same gene.

We ended up with a log2 FC expression matrix of 15,591 genes among 81 variables (experiment comparisons). Such an expression matrix included NA values corresponding to the genes that were absent in a microarray platform. With the aim of grouping experiments which trigger similar gene expression profiles across the studied cell lines and treatments, we performed a hierarchical clustering on the nutrigenomics gene expression matrix obtained from our database (Fig. 2).

We observed in the cluster dendrogram that the most remarkable property is that, as previously observed (Lamb et al., 2006), cell lines