Gene Set Enrichment Analysis (GSEA) of differentially expressed genes in green tea against methionine-choline deficient diet in high-fat patients in the development of non-alcoholic fatty liver disease (NAFLD)

Sarbojoy Saha¹, Ishtiaque Ahamad², Shampa Barmon³

¹ Department of Biochemistry and Microbiology, North South University, Dhaka, Bangladesh. * Corresponding author. Email: sarbojoy.saha@northsouth.edu
² Department of Biochemistry and Microbiology, North South University, Dhaka, Bangladesh. Email: ishtiaque.ahammad@northsouth.edu
³ Department of Environmental Science and Management, North South University, Dhaka, Bangladesh. Email: shampab50@gmail.com

Abstract

The most common liver disorder nowadays is non-alcoholic fatty liver disease (NAFLD) and it is a progressive disease that rises in severity from steatosis to nonalcoholic steatohepatitis (NASH), fibrosis and cirrhosis to increase risk of developing hepatocellular carcinoma. It is a cause of great concern as there is an estimated seventy million Americans who are currently affected by NAFLD, and this is expected to only increase because of its association with obesity and diabetes and also a lack of therapies to keep its development and progression in check. In this particular study we performed a gene set enrichment analysis (GSEA) of differentially expressed genes in a green tea against methionine-choline deficient diet in high-fat patients in the development of non-alcoholic fatty liver disease (NAFLD). The downregulated genes were used to perform an enrichment analysis and in the ARCHS4 TFs Coexpression database the most significant gene was found to be KLF5_human_tf_ARCHS4_coexpression. In the ARCHS4 Kinases Coexpression pathway database STYK1_human_kinase_ARCHS4 Coexpression was found to be the most significant gene. And finally for the upregulated genes a similar enrichment analysis was performed and in the humancy database γ-linolenate biosynthesis_Homo sapiens_PWY-6000 gene was discovered to be the most significant one. This study has used bioinformatics tools and the Enrichr software to perform a comparative analysis of differentially expressed gene sets for high-fat patients having a diet consisting of green tea against a methionine-choline deficient diet. Green tea is known to contain several antioxidants and polyphenols which provide protection against many liver diseases.
such as non-alcoholic fatty liver disease (NAFLD). The present study simply tries to build awareness of this to the general public and allow them to learn more about certain diets which have protective effects against liver diseases. Hopefully by implementing these in their daily lifestyles the public can gain some form of protection against these types of liver disorders.

**Keywords**: non-alcoholic fatty liver disease, steatosis, NASH, gene set enrichment analysis, green tea, methionine-choline deficient diet, γ-linolenate biosynthesis_Homo sapiens_PWY-6000

**Introduction**

Nowadays in the developed world, non-alcoholic fatty liver disease (NAFLD) is the most common liver disorder (Sass, Chang, and Chopra 2005). It is a progressive disease that increases in severity from steatosis to nonalcoholic steatohepatitis (NASH), fibrosis and cirrhosis to increase hepatocellular carcinoma risk (Wong et al. 2010). There is an estimated seventy million Americans affected by NAFLD (Angulo 2007) and its prevalence is expected to rise due to its association with obesity and diabetes (Wanless and Lentz 1990), and a lack of therapies to keep its development and progression in check. No validated treatments of NAFLD exists beyond weight loss and weight management. Weight loss intervention have a poor success rate, thereby emphasizing the need to validate alternative strategies which can mitigate the disease progression to NASH (Ayyad and Andersen 2000).

Thought the mechanism leading to NASH remain unclear, the ‘two-hit’ mechanism is often used to describe its simple etiology (Day and James 1998). The ‘first-hit’ leading to liver steatosis is caused by obesity, insulin resistance and excess lipid accumulation. Then reactive oxygen species and reactive nitrogen mediated ‘second-hits’ leads to oxidative and nitratative modifications to lipids and proteins, and liver injury and NASH (Fujita et al. 2010).

Reactive species implicated in the ‘second-hit’ phase include those that have originated from mitochondria, cytochrome P-450 induction (Passeyre D, Fromenti B, Mansour A,2010) and nicotinamide adenine dinucleotide phosphate(NADPH) oxidase following inflammatory cell activation. Greater hepatic NADPH oxidase activity increased liver inflammation, injury, lipid peroxidation and fibrogenesis in a diet-induced model of NAFLD (Carmiel-Haggai, Cederbaum, and Nieto 2005). Other proinflammatory enzymes like myeloperoxidase(MPO) and inducible nitrous acid synthase, can also cause hepatocyte damage in NASH. In patients with NASH, higher accumulation of MPO-positive Kupffer was accompanied by the liver injury induced by the MPO-H₂O₂ system, as proven by greater 3-nitro-Tyrosine(N-Tyr) levels. Similarly, greater liver inducible nitrous oxide synthase in rodents having NASH was paralleled by increased reactive nitrogen species(RNS) and N-Tyr in association with liver injury, inflammation and fibrosis (Fujita K et al,2010).

Green tea has potential for use in new molecular therapies. Green tea, made from the dried leaves of the *Camelia Sinensis* plant, is a popular beverage that has been consumed for thousands of years, and Traditional Chinese medicine has emphasized its use in disease prevention and control (Suzuki, Miyoshi, and Isemura 2012). Epidemiological studies have discovered that green tea...
consumption is linked to a lower incidence of liver disease (Ui et al. 2009). This is because of the catechins found in green tea, such as epigallo-catechin-3-gallate (EGCG), a polyphenol found in green tea and green tea extract and is the most bioactive and well-studied component of green tea. EGCG has multiple health benefits (Kim, Yang, and Cho 2009) including the attenuation and prevention of liver fibrosis (Ding et al. 2015). In vitro studies have shown that EGCG can reduce liver fibrosis by inhibiting the proliferation of fibroblasts, reducing collagen distribution and upregulating the mitochondrial respiratory chain (Santamarina et al. 2015). Outside the liver disease spectrum, this catechins can attenuate tumor growth factor B9 (TGF-b) (Sriram et al. 2015).

Also, the catechins EGCG downregulates OPN expression by reducing the half-life of EGCG mRNA, EGCG-mediated mRNA degradation leads to reduced closure of wounds (Zapf et al. 2015).

One probable mechanism for EGCG-mediated OPN mRNA degradation is through microRNA (miRNA) mediated mRNA deca. miR181a targets OPN mRNA and downregulates OPN protein expression leading to a suppression of migration and adhesion in HepG2 cells (Bhattacharya et al. 2010). To note, the expression of multiple miRNAs (Arola-Arnal and Bladé 2011) is modulated by miRNA (Tsang and Kwok 2010). In this particular study, we have concentrated on three miRNAs: miR-221, miR-181, and miR-10b, as studies have shown that EGCG modulates the expression of these three miRNAs (Milenkovic et al. 2012). Furthermore, miRBase Target Database showed that three miRNA target OPN mRNA (Yamada et al. 2016).

The objective of this study is to identify differentially expressed genes from individuals having Non-Alcoholic Fatty Liver Disease to determine the efficacy of green tea on hyperlipidemic Animal Model. This was a comparative study between efficacy of 2% green tea extract and a methionine-choline deficient diet. In order to do this, gene expression dataset series GSE 77964 was used.

Materials and Methods

Ethical statement

For this study, mice were used as animal models instead of human patients. The protocol for this study was approved by the Institutional Animal Care and Use Committee (2012A00000156) at The Ohio State University. All mice were acclimatized to the temperature-, light-, and humidity-controlled facility for at least 2 weeks until WT and Nrf2-null mice were 11-12 weeks old. No harmful chemicals or substances were used on the mice during or after this study. The mice were kept in spacious, controlled environment with sufficient light, air and water. They were given the diet consisting of green tea and also the methionine-choline deficient diet on a group basis. The author and co-author were fully aware of this and gave their informed consent. There are no conflicts of interest regarding this study.

Finding out the differentially expressed Genes from GSE 77964

From the NCBI website GEO datasets were searched using the term “green tea and methionine-choline deficient diet” and reference series GSE 77964 was analyzed with GEO2R.
For GEO2R analysis two groups named “2% green tea extract” and “methionine-choline deficient diet” were defined. Four samples were included in the green tea extract group and three constituted the other group to each groups. the GEOquery (Davis and Meltzer 2007) and limma R (Smyth n.d.) packages from the Bioconductor project were used, and GEO2R analysis was performed (Smyth 2004). Top 250 differentially expressed genes were found using the Benjamini & Hochberg (false discovery rate) method (Benjamini and Hochberg 1995) P values were adjusted. Log2 transformation to the data was applied. R script used to perform the calculation was obtained from the R script tab.

**Enrichment Analysis**

Enrichment analysis of Downregulated genes was carried out using

- ARCHS4 TFs co-expression (Lachmann et al. 2017)
- ARCHS4 Kinases Coexp (Lachmann et al. 2017)

Enrichment analysis of Upregulated genes was carried out using

- Humancyc 2016 (Caspi et al. 2016)

**R Script**

```r
# Version info: R 3.2.3, Biobase 2.30.0, GEOquery 2.40.0, limma 3.26.8
# R scripts generated Tue Apr 17 06:35:20 EDT 2018

# load series and platform data from GEO

gset <- getGEO("GSE77964", GSEMatrix =TRUE, AnnotGPL=FALSE)
if (length(gset) > 1) idx <- grep("GPL20964", attr(gset, "names")) else idx <- 1
gset <- gset[[idx]]

# make proper column names to match toptable
fvarLabels(gset) <- make.names(fvarLabels(gset))

# group names for all samples
gsms <- "XXXX0000XXXXXXXXXX111XXXX"
sml <- c()
for (i in l:inchar(gsms)) { sml[i] <- substr(gsms,i,i) }

# eliminate samples marked as "X"
sel <- which(sml != "X")
sml <- sml[sel]
gset <- gset[,sel]

# log2 transform
ex <- exprs(gset)
```
qx <- as.numeric(quantile(ex, c(0., 0.25, 0.5, 0.75, 0.99, 1.0), na.rm=T))
LogC <- (qx[5] > 100) ||
if (LogC) { ex[which(ex <= 0)] <- NaN
exprs(gset) <- log2(ex) }

# set up the data and proceed with analysis
sml <- paste("G", sml, sep="")  # set group names
fl <- as.factor(sml)
gset$description <- fl
design <- model.matrix(~ description + 0, gset)
colnames(design) <- levels(fl)
fit <- lmFit(gset, design)
cont.matrix <- makeContrasts(G1-G0, levels=design)
fit2 <- contrasts.fit(fit, cont.matrix)
fit2 <- eBayes(fit2, 0.01)
tT <- topTable(fit2, adjust="fdr", sort.by="B", number=250)
tT <- subset(tT, select=c("ID","adj.P.Val","P.Value","t","B","logFC","ORF","GB_ACC","SEQUENCE" ))
write.table(tT, file=stdout(), row.names=F, sep="\t")

########################################################################
#   Boxplot for selected GEO samples
library(Biobase)
library(GEOquery)
# load series and platform data from GEO
gset <- getGEO("GSE77964", GSEMatrix =TRUE, getGPL=FALSE)
if (length(gset) > 1) idx <- grep("GPL20964", attr(gset, "names")) else idx <- 1
gset <- gset[[idx]]

# group names for all samples in a series
gsms <- "XXX0000XXXXXXXXXX111XXXX"
sml <- c()
for (i in 1:nchar(gsms)) { sml[i] <- substr(gsms,i,i) }
sml <- paste("G", sml, sep="")  set group names

# eliminate samples marked as "X"
sele <- which(sml != "X")
sml <- sml[sele]
gset <- gset[,sel]

# order samples by group
ex <- exprs(gset)[, order(sml)]
sml <- sml[order(sml)]
fl <- as.factor(sml)
labels <- c("2percent+green+tea+extract","methionine-choline+deficient+t+diet")

# set parameters and draw the plot
palette(c("#dfeaf4","#f4dfdf","#AABBCC"))
dev.new(width=4+dim(gset)[[2]]/5, height=6)
par(mar=c(2+round(max(nchar(sampleNames(gset)))/2),4,2,1))
title <- paste("GSE77964", ", ", annotation(gset), ", ", selected samples", sep = ",")
boxplot(ex, boxwex=0.6, notch=T, main=title, outline=FALSE, las=2, col=fl)
legend("topleft", labels, fill=palette(), bty="n")

Results and Discussion

<table>
<thead>
<tr>
<th>Index</th>
<th>Name</th>
<th>P-value</th>
<th>Adjusted p-value</th>
<th>Z-score</th>
<th>Combin ed score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>KLF5_human_tf_ARCHS4_coexpression</td>
<td>9.51E-09</td>
<td>0.00001</td>
<td>-1.6</td>
<td>29.5</td>
</tr>
<tr>
<td>2</td>
<td>ATOH1_human_tf_ARCHS4_coexpression</td>
<td>7.87E-08</td>
<td>0.00001657</td>
<td>1.6</td>
<td>26.69</td>
</tr>
<tr>
<td>3</td>
<td>TBX10_human_tf_ARCHS4_coexpression</td>
<td>7.87E-08</td>
<td>0.00001657</td>
<td>1.6</td>
<td>26.43</td>
</tr>
<tr>
<td>4</td>
<td>CDX2_human_tf_ARCHS4_coexpression</td>
<td>7.87E-08</td>
<td>0.00001657</td>
<td>1.5</td>
<td>25.54</td>
</tr>
<tr>
<td>5</td>
<td>ISX_human_tf_ARCHS4_coexpression</td>
<td>7.87E-08</td>
<td>0.00001657</td>
<td>1.5</td>
<td>25.18</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Index</th>
<th>Name</th>
<th>P-value</th>
<th>Adjusted p-value</th>
<th>Z-score</th>
<th>Combin ed score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>STYK1_human_kinase_ARCHS4_coexpression</td>
<td>7.87E-08</td>
<td>0.00002708</td>
<td>1.69</td>
<td>27.67</td>
</tr>
<tr>
<td>2</td>
<td>MST1R_human_kinase_ARCHS4_coexpression</td>
<td>0.00002633</td>
<td>0.00452943</td>
<td>1.63</td>
<td>17.17</td>
</tr>
<tr>
<td>3</td>
<td>EPHB3_human_kinase_ARCHS4_coexpression</td>
<td>0.00015</td>
<td>0.01032</td>
<td>1.67</td>
<td>14.71</td>
</tr>
<tr>
<td>4</td>
<td>ERBB3_human_kinase_ARCHS4_coexpression</td>
<td>0.00015</td>
<td>0.01032</td>
<td>1.67</td>
<td>14.7</td>
</tr>
</tbody>
</table>
The downregulated genes were used to conduct an enrichment analysis and in the ARCHS4 TFs Coexpression database KLF5_human_tf_ARCHS4_coexpression was found to be the most significant gene. And again for these genes in the ARCHS4 Kinases Coexpression pathway database STYK1_human_kinase_ARCHS4_coexpression was found to be the most significant gene. And for the upregulated genes a similar enrichment analysis was performed, and in the Humancy database γ-linolenate biosynthesis_Homo sapiens_PWY-6000 was discovered to be the most significant pathway. Naturally, this is so as gamma-linoleic acid is a fatty acid which is essential for humans. And so its pathway is also very significant as can be seen from the results from the Humancy database. Some other pathways are also shown in the database as can be seen here.

Gamma linoleic acid(GLA) is a polyunsaturated (18:3) ω-6 fatty acid (the 6 refers to the position of the first double bond from the methyl end of the fatty acid). It is synthesized from linoleic acid by delta$^6$ desaturase enzymes. Unlike its precursor, GLA is not an essential fatty acid, since humans possess a delta$^6$ desaturase enzyme. GLA is of major importance, being the precursor for the synthesis of several delta$^6$-desaturated fatty acids, such as arachidonic acid and eicosapentaenoic acid(EPA). Delta$^6$-desaturated fatty acids have roles in the maintenance of membrane structure and function, in the regulation of cholesterol synthesis and transport, in the prevention of water loss from the skin, and as precursors of eicosanoids, including prostaglandins and leukotrienes. Although humans possess a delta$^6$-desaturase, they lack a delta$^{12}$ and delta$^{15}$ desaturases, and
therefore linoleic acid, the precursor for gamma linoleic acid biosynthesis, is an essential fatty acid that must be obtained through the diet (Fan and Chapkin 1998).

Conclusion

In this particular study an enrichment analysis was performed for the differentially expressed genes and in the ARCHS4 TFs Coexpression database the most significant gene was found to be KLF5_human_tf_ARCHS4_coexpression. In the ARCHS4 Kinases Coexpression pathway database STYK1_human_kinase_ARCHS4_coexpression was found to be the most significant gene. And finally for the upregulated genes a similar enrichment analysis was performed and in the humancy database γ-linolenate biosynthesis_Homo sapiens_PWY-6000 gene was discovered to be the most significant one. And we know that gamma-linolenate or gamma linoleic acid (GLA) is an essential omega 6 fatty acid, synthesized from linoleic acid, itself another essential fatty acid. It is also a precursor for several other essential fatty acids such as arachidonic acid and eicosapentaenoic acid. It must be obtained from the diet, as humans lack the delta-6 desaturase enzymes needed for its synthesis. This study hopes to build some awareness about the prevalence of liver diseases in recent years and the protective measures to counteract this growing problem. This can be in the form of a healthy and balanced diet consisting of healthy options such as green tea which is explained in this study. Hopefully, the common public will become more aware of this increasing concern of non-alcoholic fatty liver disease (NAFLD) and take necessary steps to combat this problem.
REFERENCES


Ding, Yi, Xin Sun, Yuning Chen, Yue Deng, and Ke Qian. 2015. “Epigallocatechin Gallate


