UNIVERSITY OF TURKU
Faculty of Science

Effects of the STRIP dietary intervention on blood cell transcriptome and activity of signaling pathways

Janette Rikkinen

Physiology and Genetics
Master's thesis
Credits: 30 op

Supervisors:
Juha Mykkänen
Katja Pahkala
Christina Nokkala

24.05.2023
Turku

Master's thesis

Subject: Biology, Physiology and Genetics
Authors: Janette Rikkinen
Title: Effects of the STRIP dietary intervention on blood cell transcriptome and activity of signaling pathways
Supervisors: Juha Mykkänen, Katja Pahkala and Christina Nokkala
Number of pages: 58 pages + 16 appendix pages
Date: 24.05.2023

Atherosclerotic cardiovascular diseases (CVDs) are the largest group of diseases causing deaths. CVD risk may be reduced with following the dietary and other lifestyle recommendations. Diet can affect gene expression both at an individual gene and molecular pathway levels. The STRIP study (Special Turku Coronary Risk Factor Intervention Project) is a dietary intervention, where intervention group received repeated and individualized dietary counselling from 0.7 to 20 years of age, targeting to reduce CVD risk factor levels by replacing saturated fatty acids (SAFAs) with unsaturated fat in diet. Aims of the thesis was to find out, if SAFA intake of total energy intake (SAFA E%) has an effect on blood cell transcriptome measured both at individual gene and pathway levels.

Parts from pre-collected longitudinal clinical data from the STRIP study data were used to match with blood transcriptome data from 495 STRIP participants (age 15–19 years) including mRNA-sequencing counts for 25562 genes. Clinical data were managed and analyzed with SAS. DESeq2 package in R software was used for differential gene expression (DGE) analysis which was adjusted with age, sex, and mRNA-sequencing batch. Molecular pathway enrichment analysis was conducted with two different methods (GOseq and edgeR R packages). Benjamin-Hochberg's adjusted p-value <0.05 was used to determine the statistical significance. 64 upregulated and 10 downregulated genes were found in DGE analysis using continuous SAFA E% as a predictor variable. The two molecular pathway enrichment analyses resulted one upregulated (membrane biogenesis) and two downregulated (immunoglobulin complex and spermatogenesis) pathways when SAFA E% increases.

In conclusion, the results of this thesis suggest that relative intake of dietary saturated fat affects gene expression in whole blood at both individual gene and pathway levels. Most of the results supported the prior literature, but also novel results possibly associated with cardiovascular health were observed. However, further studies are needed to establish their role in heart health.

Key words: atherosclerosis, blood cell transcriptome, CVD, differential gene expression, mRNA-seq, pathway enrichment analysis, SAFA

Pro gradu -tutkielma

Pääaine: Biologia, Fysiologian ja genetiikan linja
Tekijät: Janette Rikkinen
Otsikko: STRIP-ravintointervention vaikutukset verisolujen transkriptomiin ja signalointireittien aktiivisuuteen
Ohjaajat: Juha Mykkänen, Katja Pahkala ja Christina Nokkala
Sivumäärä: 58 sivua + liitteet 16 sivua
Päivämäärä: 24.05.2023

Ateroskleroottiset sydän- ja verisuonitaudit (CVD) ovat suurin kuolemia aiheuttava sairauksien ryhmä. Ateroskleroosin kehittymistä voi ehkäistä noudattamalla terveellisiä elämäntapoja. Ruokavalio voi vaikuttaa yksilön yksittäisten geenien ilmentymiseen ja siten myös signalointireittien aktiivisuuteen. STRIP (SepelvaltimoTaudin Riskitekijöiden InterventioProjekti) on interventiotutkimus, missä interventioryhmän osallistujat saivat yksilöllistä ravintoneuvontaa 0.7 vuoden iästä 20 ikävuoteen saakka tavoitteena vähentää CVD-riskitekijöiden tasoja korvaamalla tyydyttyneiden rasvojen (SAFA) käyttöä tyydyttymättömillä rasvoilla. Tutkielman tavoitteena oli selvittää, vaikuttaako SAFA:n osuus kokonaisenergian saannista (E%) verisolujen transkriptomiin sekä yksittäisten geenien että molekyylireittien tasolla.

Tutkielman tekemisessä käytettiin osia STRIP-tutkimuksen aikana kerätystä kliinisestä datasta, sekä verinäytteestä analysoitua lähetti-RNA-sekvensointidataa. Mukana oli 495 15–19-vuotiasta STRIP-osallistujaa sekä geenitoiminnan tasot 25 562 geenistä. Kliininen data muokattiin ja analysoitiin SAS-ohjelmistolla. R-ohjelmiston DESeq2-pakettia käytettiin analysoimaan SAFA E%:n yhteyksiä veren geeni-ilmentymään (nk. differentiaaliseen geenieksprессio; DGE). DGE-analyysi vakioitiin iän, sukupuolen ja sekvensointierän mukaan. Molekyylireittien aktiivisuutta testattiin kahdella eri analyysimenetelmällä (GOseq- ja edgeR -ohjelmistopaketit). Benjamini–Hochbergin korjattu p-arvo <0.05 määritti tilastollisen merkitsevyyden. Kun jatkuvaa SAFA E% muuttujaa käytettiin DGE-analyysissä, tulokset osoittivat 64 geenin ilmentymisen lisääntyneen ja 10 geenin ilmentymisen vähentyneen. Molekyylireittien aktiivisuusanalyyseistä löytyi yksi reitti (solukalvon biogeneesi), jonka aktiivisuus lisääntyy ja kaksi reittiä (immunoglobuliinikompleksi ja spermatogeneesi), joiden aktiivisuus vähenee SAFA E% kasvaessa.

Yhteenvetona, tämän tutkielman tulokset viittaavat ravinnon tyydyttyneiden rasvahappojen suhteellisen osuuden vaikuttavan kokoveren geenien ilmentymiseen sekä yksittäisten geenien että molekyylireittien tasoilla. Suurin osa tutkielman tuloksista tukee aikaisempaa kirjallisuutta, mutta osa tuloksista on uusia ilman aiempia yhteyksiä sydänterveyteen. Tarvitaan kuitenkin lisätutkimuksia, jotta uusien tulosten tarkempi yhteys sydänterveyteen voidaan varmentaa.

Avainsanat: Ateroskleroosi, verisolujen transkriptomi, CVD, differentiaalinen geeni ekspressio, mRNA-seq, molekyylireittien rikastumisanalyysi, SAFA

List of abbreviations

AA	arachidonic acid
B-H	Benjamin-Hochberg's
COX	cyclooxygenase
CVD	cardiovascular disease
CYP	cytochrome P450
DGE	differential gene expression
DISC	The Dietary Intervention Study in Children
DNA	deoxyribonucleic acid
E%	total energy intake
FDR	false discovery rate
FHS	The Framingham Heart Study
GO	Gene Ontology
GTF	general transcription factors
LDL	low-density lipoprotein
LOX	lipoxygenase
MD	Mediterranean diet
mRNA	messenger RNA
MSigDB	The Molecular Signatures Database
MUFA	monounsaturated fatty acid
NGS	next-generation sequencing
PCSK9	Proprotein convertase subtilisin/kexin type 9
PTGR	post-transcriptional gene regulation
PUFA	polyunsaturated fatty acid
RNA	ribonucleic acid
RNA-Seq	RNA sequencing
SAFA	saturated fatty acids
STRIP	Special Turku Coronary Risk Factor Intervention Project
T2T	Telomer-to-Telomer
TLR	Toll like receptor
VOO	Virgin olive oil
WHO	World Health Organization

Table of content

APPENDIX 2. SAS code: Standardization of continuous variable

APPENDIX 3. SAS code: Phenotypic information and t-test

APPENDIX 4. R code: DESeq2 Differential gene expression analysis

APPENDIX 5. R code: Volcano plot

APPENDIX 6. R code: GOseq analysis

APPENDIX 7. R code: edgeR analysis

APPENDIX 8. Downregulated genes from DGE analysis with continuous SAFA E%

APPENDIX 9. Upregulated genes from DGE analysis with continuous SAFA E%

1 Introduction

1.1 Atherosclerotic cardiovascular diseases

Atherosclerotic cardiovascular diseases (CVDs) are a group of diseases affecting the heart and blood vessel health. They are the largest group of diseases causing deaths. Atherosclerosis is a condition where arteries narrow with plaque, composed mainly of LDL-cholesterol and inflammatory cells, which builds up slowly inside the arteries. The surface of plaque is soft and fragile, due to which it easily ruptures which lead to a blood clot formation. Depending on the artery's size plaque itself starts to slow down blood flow but the emergence of blood clot can stop blood flow in the vein rapidly. If the artery clogs up completely, blood flow stops to the tissue in the vascular area and it leads to necrosis (Mustajoki, 2020). Even tough atherosclerosis is diagnosed usually in middle and late adulthood, it is known that its development begins in early childhood (Raitakari et al., 2022).

We have limited options to minimize the risk for developing atherosclerosis and CVDs. Both lifestyle and hereditary factors affect the risk and even though we cannot control hereditary factors, the lifestyle choices are controllable. Smoking, increased LDL-cholesterol levels in blood and increased blood pressure are the most significant risk factors for developing atherosclerosis.

1.1.1 CVD research study designs

Epidemiology of cardiovascular diseases has been studied a lot and different research study designs are presented for different study questions. Cross-sectional and longitudinal study designs are both used for research of cardiovascular diseases. Cross-sectional study collects data on large population cohort at a certain time point, thus causes and consequences are measured at the same time. Cross-sectional study can be used for example to indicate the prevalence of a disease, but it does not give information about which factors affects the illness. On the other hand, longitudinal research is following changes in a long term and the research usually lasts years or decades, providing information about the relationship between CVD risk factors. Longitudinal studies are also cohort studies as participants are selected based on certain factor or factors, for example based on their blood cholesterol levels (Ranganathan and Aggarwal, 2018).

When studying a human population, participants usually share a defining characteristic for example age or birth at certain period.

For example, cross-sectional study design has been used to find out if plasma fatty acid level has correlation to gene expression in lipid metabolism related genes in humans (Larsen et al., 2018). Because the study was cross-sectional, causality cannot be ensured, but the study provides indicative results which should be studied further, for example with intervention study to see, how change in plasma fatty acid levels affects gene expression. Cross-sectional study can be implemented more quickly than longitudinal study and therefore it might be beneficial to start with cross-sectional study design.

The Framingham Heart Study (FHS) is a longitudinal cohort study initiated in 1948 and it is still ongoing (Tsao and Vasan, 2015). FHS started with the Original cohort but has since expanded and includes now four more cohorts, Offspring cohort and Gen 3 cohort which includes children and grandchildren of the people in the Original cohort, OMNI 1 and OMNI 2 cohorts has been formed to include more ethnical diversity to the study. The Original cohort comprises two-thirds of the adult population in Framingham, Massachusetts. Almost all participants have been adults when participants to the cohorts were requirement. The three-generation family structure is the major strength of the FSH. The FSH has collected a significant amount of data over the years and has been evolved with time, cardiovascular imaging and genetics are examples of studies that has been added to the participant studies.

The Dietary Intervention Study in Children (DISC) is other example of a longitudinal cohort study but as a difference to FSH study, DISC study performs intervention on some participants. DISC aims to assess how dietary intervention effects on children ages 8–10 with elevated blood cholesterol levels. Children were randomized to either a control group or an intervention group. Children and their parents in the intervention group received dietary counseling. This study showed that children getting dietary counseling chose more often foods that were ranked as "less atherogenic" than children in the control group (Van Horn et al., 2005). FHS and DISC studies have been collecting data from participants for years or decades, but longitudinal studies can also be shorter. An eight-week dietary intervention study, conducted in Norway, showed that metabolomics has potential to expand understanding of biological and molecular effects of dietary fat quality (Ulven et al., 2019).

Research field of CVD covers adults and aged people rather comprehensively. Infants and young children were excluded from nutritional recommendations for a long time because there was a fear that low intake of cholesterol and saturated fat might affect negatively to growth and development (Simell et al., 2009). Because of this assumption intervention studies aiming to reduce CVD risk in infants and young children are rare.

1.1.2 The STRIP Study

The STRIP study (Special Turku Coronary Risk Factor Intervention Project) is a prospective, randomized trial which aims to prevent atherosclerosis beginning in infancy (Simell et al., 2009). Study participants, families of 5-month-old infants, were recruited by nurses at well-baby clinics in Turku, Finland. The infants were born between July 1989 and December 1991. At the age of 7 months, 1062 infants (56.5 % of the eligible age-cohort) were randomly allocated to a dietary intervention (n=540) or control (n=522) group. A vast number of risk factors were measured from both groups at the corresponding age points. Participants from both groups returned food diaries that were used to collect nutritional data. Based on the Nordic nutrition recommendations, the intervention group received individualized dietary counselling at 1- to 3-month intervals until 2 years of age, and thereafter biannually until 20 years of age (Matthews et al., 2019). A fixed diet was never specified, and the counselling did not attempt to reduce total fat intake. What the counselling did target in the child's diet was to replace saturated fat with unsaturated fat and to reduce the intake of cholesterol.

The STRIP study is a unique worldwide because it is the only study that has started to give dietary intervention in infancy and continued it for 20 years simultaneously collecting data with repeated measurements on a voluminous number of risk factors and subclinical markers for cardiovascular health (STRIP Study Group, 2008). The STRIP study has shown that the intervention have diverse positive effect on heart health and importantly it has proven that dietary intervention was safe and therefore it did not have negative impact to growth or development (Simell et al., 2009). In addition to the original intervention-control study design, the STRIP study has shown that achieving the goals of dietary intervention improves levels of risk factors associated with the development of cardiovascular and metabolic diseases such as blood pressure, abdominal aortic thickness

and flexibility, insulin sensitivity and serum lipids and metabolome (Laitinen et al., 2018; Laitinen et al., 2020a; Laitinen et al., 2020b; Lehtovirta et al., 2021).

1.2 Gene expression

Gene expression is a process where functional gene product is formed from an activated gene and at the conclusion of various processes a final product is synthetized. Final product may be protein or non-coding RNA. In a cell all genes are not activated at once hence not all genes are expressed at the same time. Gene expression can be regulated before the transcription starts, during it or after the transcription.

Pre-transcriptional regulation activates or represses the action of transcription. Primarily the eukaryotic gene expression is controlled at the level of initiation of transcription (Cooper, 2000) This regulation can occur in several ways. Human, as a eukaryote, have RNA-polymerases (Pols), but to work, Pols uses complex set of general transcription factors (GTFs) and specific transcription factors. Eukaryotes have three different Pols which all synthesizes different type of RNA molecules. Pol I transcribes large ribosomal RNAs (rRNA), while Pol III transcribes small rRNA, as well as transfer RNA (tRNA) and other small RNAs (sRNA), whereas Pol II transcribes all messenger RNAs (mRNA), small nuclear RNAs (smRNA), micro RNAs (miRNA) and small interfering RNAs (siRNA). GTFs are relevant in promoter recognizing, recruiting Pols, DNA unwinding, transcription start site (TSS) recognizing and as regulatory factors cooperators (Grünberg and Hahn, 2013).

Specific transcription factors work as activators or repressors depending on their effect on transcription activity. Activators are key elements in the initiation of transcription as they bind to specific DNA sequences and other regulating proteins, such as GTFs, allowing transcription initiation (Cooper, 2000). Coactivators are molecules that increase gene expression by binding to activators, thus they do not bind directly to DNA. Repressors inhibits transcription by interfering with the function of activators, GTFs or Pols. Repressors can bind to DNA sequence and obstruct binding site either completely or partially but in such way that GTF, Pol or activator cannot bind to its specific site (Cooper, 2000). Repressors can also inhibit transcription by blocking active site of activators or GTFs (Cooper, 2000). Respectively, like coactivators, the corepressors binds

to repressors which lead to activation of repressor and transcription decreases (Cooper, 2000).

In some cases gene expression can be controlled during early transcription elongation, if elongation is terminated by promoter-proximal pausing of Pol II (Adelman and Lis, 2012). After transcription, before translation, post-transcriptional gene regulation (PTGR) controls actions that happen at RNA level. PTGR processes have an influence on RNA modification, transport, and stability. Transcription produces pre-mRNA which includes protein coding parts, exons, and non-coding parts, introns. Before exiting the nucleus, pre-mRNA is modified by alternative splicing, capping and addition of poly-A tail.

Capping machinery modifies 5'-end of a transcript by adding to it a cap-0 structure which consists of N7-methylguanosine nucleoside linked to the first transcribed nucleotide via a 5'-5' triphosphate bond (Decroly et al., 2011; Shatkin, 1976). Additionally, the first nucleotide can be methylated at the 2'-O position of the ribose leading to cap-1 structure (Ramanathan et al., 2016). Even tough capping is perceived as PTGR, it actually occurs already during transcription in the nucleus after the 5'-end emerges from RNA polymerase II (Kachaev et al., 2020). Capping is important step in pre-mRNA modification, the cap protects transcript from exonucleases, recruits proteins for splicing and polyadenylation and is needed for nuclear export. In addition to these, the cap is crucial factor for ribosomal attachment in the translation initiation (Pestova et al., 2001). Respectively, the 3'-end of pre-mRNA is also modified as part of PTGR and that process is called polyadenylation. The 3'-end of transcript is cleaved to free a 3' hydroxyl in order to enable the poly-A polymerase enzyme add a poly-A tail to the RNA (Nature Education, 2014). The poly-A tail is a chain of adenine nucleotides, in mammalian cells around 250 nucleotides, that makes the mRNA more stable. Like the 5' cap, poly-A tail participates in the nuclear export.

During splicing introns are removed and the exon compound for the final mRNA is formed. Splicing is catalyzed by highly dynamic ribonucleoprotein complex, spliceosome, consisting of five small nuclear ribonucleoproteins (snRNPs) and numerous proteins (Will and Lührmann, 2011). Because of the dynamic feature of spliceosomes, same core complex works in several splicing events but still retains accuracy. Splicing is regulated by both general and specific splicing factors and by mechanisms that rely on

cis-regulatory elements that are either enhancers or silencers (Chen and Manley, 2009; Wang and Burge, 2008).

As described, gene expression is controlled in several levels which leads to some genes being expressed more than others and some even being completely shut off. Gene expression is important indicator to find out how different factors, such as drugs, nutrients, or environment, affects gene expression. Differential gene expression (DGE) analysis studies expression level differences between study population and groups or follows changes during time. To name a few DGE analysis can be applied to the field of molecular cell biology for example pharmacology and development studies, nutrigenomics and public health (Carulli et al., 1998).

1.2.1 Transcriptome-wide mRNA-sequencing and genome mapping techniques

Different mRNA-sequencing technologies can be used for example to quantify gene expression in the sample of interest. Gene expression can be studied for one gene, for whole transcriptome or something between of these two. Transcriptome means the set of all mRNA transcripts present in a cell, tissue or organism. Traditional sequencing technologies, that are based on for example hybridization or Sanger sequencing, have high costs for large genomes and other technical disadvantages that limits their use thus makes them unusable in annotating the structure of transcriptomes of large genomes (Wang et al., 2009). High-throughput next-generation sequencing (NGS) techniques has been developed to solve these problems. NGS-techniques can generate massive amount of sequence data cost-effectively and yet rapidly (Marguerat and Bähler, 2010). For transcriptome analyses NGS-approach RNA-seq was developed over a decade ago and today it is the most used analyzing for differential gene expression (Lister et al., 2008; Stark et al., 2019)

Shortly, mRNA-seq data is generated through sequencing and sequence read alignment. Sequencing requires converting RNA molecules to cDNA library. Fragments of cDNA can be sequenced from both ends with pair-end sequencing which results better alignment of the reads than what sequencing from one end, single-end sequencing, offers (Illumina, 2022). Short paired-end reads offers more robust results for differential expression analysis than long single-end reads (Freedman et al., 2020). In order to create RNA-seq

experiment that provides meaningful high-quality data, the use of single- or paired-end sequencing read, the level of replication and the sequencing read depth needs to be considered (Stark et al., 2019). The length of the read is typically 30–400 bp and it should be chosen based on sample type and application, short read are more suitable for quantitative assays and long reads for qualitative assays (Wang et al., 2009).

After the sequencing, raw short reads are mapped to genomic or transcriptomic locations by using reference genome or transcripts. When choosing a mapping tool, the most important choice is whether the transcript identification and quantification are done sequentially or simultaneously (Conesa et al., 2016). There is variety of mapping tools and according to Conesa et al. (2016) depending on the chosen tool 70–90% of RNA-seq is expected to map onto human genome and uniformity on read coverage on exons and the mapped sequences should be ascertain with quality control tools. Examples of tools used for mapping reads to reference are Rsubread and Subjunc packages in R programming language, Rsubread is sufficient for read mapping when goal is to perform a differential expression analysis and Subjunc should be used for example exon-exon junction detection or genomic mutation detection (Shi and Liao, 2023).

Unlike in whole genome sequencing, duplications, same sequences that maps to the same location in the transcriptome, are true biological signals in mRNA-seq data hence not removed as technical arrays (Stark et al., 2019). This presents a challenge because as said duplications in the mRNA-seq data are not removed because most of them are true biological signals but there may be some true technical arrays as well.

The current version of human reference genome is 38.p14 (GRCh38.p14) (National Library of Medicine, 2022). Earlier version of GRCh38 covers only 92% of human genome and even though the newest GRCh38.p14 update offers more alternative loci, which offers variation in regions that are highly heterogeneity in human genome, it is still not perfect reference model (“One pangenome to bind them all,” 2022). Human reference genome is assembled from several individuals but most of the genome originates from a single individual. Because the GRCh38 reference genome was sequenced using bacterial artificial chromosomes it contains several gaps such as underrated number of repetitive sequences (Nurk et al., 2022).

Telomere-to-Telomere sequenced reference genome, the T2T-CHM13, has solved many issues that GRCh38 genomes have had (Nurk et al., 2022). For example, T2T-CHM13 reference genome holds five chromosome arms and other additional sequences that GRCh38 was missing. Newest version of T2T reference of human genome is T2T-CHM13v2.0 which is assembled of the CHM13 cell line plus chromosome Y from NA24385 (National Library of Medicine, 2022). Right now, T2T-CHM13 is considered most complete, accurate and representative human reference genome available, but as also with GRCh38 the lack of diversity of genetic variation (Nurk et al., 2022). To solve this weakness T2T Consortium works together with the Human Pangenome Reference Consortium.

The multidisciplinary collaboration Human Pangenome Reference Consortium (HPRC) aims to create human reference genome that represents the genetic diversity of the human (Wang et al., 2022). The term pangenome has been introduced first in the field of bacteria and the pan comes from Greek meaning whole (Tettelin et al., 2005). Pangenomes are created by joining whole-genome data from multiple individuals who represent different ethnicities around the world. Because human pangenome is designed to be as diverse as possible it has been suggested that it would solve many ethical, legal and social problems that current European origin reference has (Wang et al., 2022).

1.2.2 Gene-level analysis of mRNA-seq data

Gene-level analysis is conducted after mRNA-seq approach. As mRNA-seq mapping, gene-level analysis is typically done with R programming language (Bystrykh, 2021). When sequence reads are mapped to reference genome the next step is to process read summarization with Rsubread package. Read quantification program, featureCounts that is included in Rsubread, assigns reads to genomic features or meta-features, in other words to exons or genes (Shi and Liao, 2023). Challenge with read summarization is to determine how to handle reads that overlaps with two features, but featureCounts allows users to decide whether to exclude, fully count or fractionally count such reads. Also, it can handle both paired and unpaired reads and allows users to check if both ends of paired-end reads are mapped. Read summarization produces the count table. Several filters can be added to featureCounts for further information, for example multi-mapping reads that contains more than one alignment can be excluded. As default, multi-mapping

reads are included thus number of alignments is higher than number of reads. The output of featureCounts program is a count table which holds the number of alignments for each feature, and it also works as input data for analyzing the differences in gene expression. In many applications, the number of alignments for feature is estimated from the true expression level multiplied by the length of the feature (Young et al., 2010). Multiplying highlights differences and therefore expression differences are more easily detected in longer features or features that are highly expressed.

For analyzing differences in gene expression count table is used as input data. There are few methods and several tools to carry out differential gene expression analysis from RNA-seq data. DESeq2, DESeq2, edgeR, EBSeq, DSS are negativebinomial-based approaches for gene-level analysis, whereas the voom normalization method combined with linear modeling using the limma package and the SAMseq method of the samr package represents different approaches (Love et al., 2014). These tools are examples of available tools for gene-level analysis and chosen as examples because they are frequently compared with each other and with novel methods, and most of them have performed well in those comparisons (Schurch et al., 2016).

Based on PubMed (revised 19.2.2023) DESeq2 and edgeR tools are most cited among mentioned tools with 25393 and 15756 citations (Love et al., 2014; Robinson et al., 2010). The older version of DESeq2, DESeq had 7637 citations whereas limma had 2319, EBSeq had 600 and SAMseq had 194 citations (Anders and Huber, 2010; Law et al., 2014; Leng et al., 2013; Li and Tibshirani, 2013). Differences in the citation numbers shows the level of usage between these tools. The order of most cited tools has changed from the comparison made by Schurch et al. in 2016 because some of the tools were rapidly new at that time, for example the most cited tool DESeq2 had only 197 citations on 21.12.2015 (Schurch et al., 2016).

As citation numbers shows that DESeq2 and edgeR tools are most used in the gene-level analyses, which is in line with the results from comparison of different tools. Preference to other tools edgeR and DESeq2 are suggested tools based on their ability to detect true positives and to control false discovery rate (FDR) at lower fold changes (FC) (Schurch et al., 2016). Author and maintainer of DESeq2 package Michael Love has stated that from his experience edgeR and DESeq2 methods usually reports overlapping sets of genes from gene-level analyses and that the main differences between the two methods

are in the defaults, but both of the methods offer the possibility to turn of those defaults or ad those functionalities (Love, 2016).

1.2.3 Blood cell transcriptome

Blood transcriptome is a set of all expressed genes in a blood sample on a genome-wide scale. It is important that sampling do not add bias into study and since blood sampling is routine and optimized method, the use of blood serves this goal. The limitation of using blood sample is the fact that the sample includes different blood cells thus its cell population is heterogeneity (Chaussabel, 2015). Even though blood is not totally optimal for transcriptome analysis it is widely used because collection and sample handling are relatively easy, especially in longitudinal studies with large group of participants. Because RNA degenerates easily and gene expression changes to large extent *ex vivo*, specific blood collecting tubes and isolation kits for RNA analysis has been developed. The purpose of these tubes and kits is to cause the lysis of blood cells and fixate the RNA to preserve the *in vivo* levels of RNA in the sample. The PAXgene Blood RNA Tube (Qiagen) and isolation system from same manufacturer has shown to have most consistent results compared to traditional sample collection and preparation method and to other commercial methods (Meyer et al., 2016; Rainen et al., 2002). In large-scale studies it is necessary to make sure that enough kits are available because changing sampling or extraction strategy during the study is not possible for the results to be reliable (Meyer et al., 2016).

Whole blood and certain brain tissues has shown to have the largest proportion of differentially targeted genes in the gender-divergent class (Lopes-Ramos et al., 2020). Another study has shown that in healthy adult population elderly, obese participants and men have more upregulated genes associated with inflammation and increased heme metabolism whereas younger, non-obese participants and women have more association with activated immune responses and transcription (Schmidt et al., 2020). Based on these findings, other confounding variables beside the variable of interest should be considered while designing a study and the effects of gender and age should me minimalized by using them as covariates in the analysis when blood samples are being used.

1.3 Functional characterization of gene expression: molecular pathways

Molecular pathways are sequential activity of molecules resulting a certain product or change in the cell (National Human Genome Research Institute, 2020). For example, they can lead to the cell movement, changes in transcription activity, or building new molecules, such as proteins. Analyzing molecular pathway-levels offers more information about the biology of the studied subject than what the expression change of individual gene could reveal. Molecular pathways can act locally, between nearby cells, or systemically, for example through hormones circulating in the blood. Environmental factors affect people and molecular pathways control the responses these factors have on person. If one part of the pathway does not work correctly, the result may be a disease. Often it is not relevant to the development of the disease which part of the pathway the dysfunctionality occurs. Understanding these pathways gives clues what have gone wrong when disease strikes. What makes molecular pathways so complex is the difficulty to say where one pathway ends and where second starts. Thus, many molecular pathways have not yet been discovered.

1.3.1 Studying molecular pathways

Molecular pathway-levels can be studied form the count table or DE genes can be used as input data for pathway analysis. Understanding molecular pathways gives options to diagnose, treat and prevent diseases more personalized. It is going to take a years, even decades, before personalized drugs are generally used in the treatment of patient (National Human Genome Research Institute, 2020). In the case of diseases for which several drugs are available, information of molecular pathway-levels can be used to estimate which drug would be most effective for each patient.

Molecular pathway analyses are conducted by using gene sets, which are a collection of genes that are related to specific function or disease. The Molecular Signatures Database (MSigDB) is most widely used repositories of gene sets, including signatures extracted from original articles and entire collections of sets derived from specialized resources, examples of which are Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Liberzon et al., 2011). GO analyses are systems biology techniques for highlighting biological processes by analyzing if the DE genes are over expressed in gene categories that are grouped by certain biological process, cellular component or

molecular function (Young et al., 2010). KEGG amins to compound the network of interacting molecules representing a higher order biological function as pathways (Kanehisa and Goto, 2000). Unlike the previous ones, the Hallmark gene set was generated to summarize information across multiple gene sets in order to reduce variation and redundancy (Liberzon et al., 2015).

Understanding molecular pathways that are related to different cancers could lead to pathway-level diagnosis which would offer information that could be used in the selection of therapeutic methods. Published molecular pathway-level studies are mainly focused on cancer diseases possibly due to the severity of these diseases and the therapeutic challenges. HER2 specific drug, trastuzumab, is usually recommended for HER2-positive breast cancer treatment but respond is not as wanted in many cases thus a study has proposed that activation levels of "cAMP Pathway Protein Retention" pathway could be and effective biomarker to detect if the cancer responses to trastuzumab or not and that way we could avoid situations where ineffective treatment is given to patients (Sorokin et al., 2020).

Even though molecular pathway-level studies are clearly focused on cancer research it has potential to offer information about any other disease or condition as well. As an example, a study that combined multiple cohorts found pathway-level differences in lipid droplet organization pathway and lipid droplet cellular components pathway when comparing heavy drinkers that had alcohol-related liver cirrhosis and heavy drinkers without liver diseases (Schwantes-An et al., 2021). In this study pathway-level analysis linked the association between previously found loci and new loci identified in the pathways underlying alcohol-related liver cirrhosis. Overall, in many published studies pathway-level analysis is mentioned as one of the next steps researchers could focus from vast topics. This shows that the method has a lot of potential to provide meaningful information in several research areas.

1.3.2 Molecular pathways and CVDs

The cells constituting cardiovascular system needs highly complex regulatory mechanisms to act together towards maintain homeostasis. The complexity in molecular pathways causes it to be vulnerable, chronic dysfunctionality in the pathway disturbs the

homeostasis which in turn favors the development of pathological conditions (Wheeler-Jones, 2005). G protein coupled receptors (GPCRs), protein kinase and protein phosphatase enzymes have major role in maintaining cardiovascular system at molecular pathway level (Wheeler-Jones, 2005).

The association of inflammation and atherosclerosis has been studied extensively justifiably because immune cells are components of atherosclerotic plaque. Toll like receptor signaling, NLRP3 inflammasome and PCSK9 pathways are signaling pathways mediated by inflammatory mediators and potentially involved in atherosclerosis (Kong et al., 2022). Toll like receptors (TLRs) are class of transmembrane proteins that among other functions induces the production of pro-inflammatory cytokines and an increase in the immune activation through TLR signaling pathway has been proposed to be important in the progression of atherosclerosis thus a potential biomarker to use even before any symptoms occur (Huang et al., 2011; Kong et al., 2022). Drugs inhibiting TLR-2 signaling pathway could be beneficial in the prevention of atherosclerosis based on a cell-level study (Monaco et al., 2009). TLR signaling pathway is part of the KEGG subcollection of Canonical gene set collection.

The NLRP3 inflammasomes are the complexes of cytosolic proteins that assemble in response to endogenous danger signal that is created by the presence of cholesterol crystals, which lead to the activation of caspase-1 followed proteolytic cleavage of pro-inflammatory cytokines which are involved in the pathogenesis of atherosclerosis (Koushki et al., 2021). NLRP3 inflammasomes are highly expressed in several cell types involved in the pathogenesis of atherosclerosis and several inhibitors targeting NLRP3 inflammasomes are currently in the clinical trials and they are expected to be beneficial with atherosclerosis and other diseases such as diabetes, Alzheimer's disease and stroke (Coll et al., 2022; Takahashi, 2022). Pathway responsible for negative regulation of NLRP3 inflammasome complex assembly is part of Ontology gene set collection.

Proprotein convertase subtilisin/kexin type 9 (PCSK9) is a protein that affect cholesterol homeostasis by degrading LDL receptors post-transcriptionally thus increases serums cholesterol levels but the link between PCSK9 protein and atherosclerotic lesion is not dependent on cholesterol levels thus it is not clear whether the PCSK9 has other biological functions that affects the lesion (Giunzioni et al., 2016; Tavori et al., 2013). Above, are described the associations between atherosclerosis and the three significant inflammatory

mediators but when looking pathway level, it is important to remember that even though these pathways are named after some protein or protein complex there is a lot of other molecules involved in the pathway. For example, when targeting PCSK9 pathway the drug can be targeted directly or indirectly to the PCSK9 thus the drug could block the active site of PCSK9 or its binding site from LDL receptor.

The arachidonic pathway plays important role in CVDs because arachidonic acid (AA) can be metabolized by three separate enzyme systems which results a wide range of active fatty acid mediators (Wang et al., 2021). The three enzyme systems are cyclooxygenases (COXs), lipoxygenases (LOXs) and cytochrome P450 (CYP) enzymes. COXs metabolize AA to protanoids, thromboxane and prostacyclin whereas LOXs metabolize AA to leukotrienes and HETEs. As LOXs CYP enzyme ω-hydroxylases metabolize AA to HETEs while epoxygenases metabolize AA to four EETs. Especially CYP pathway has shown potential as therapeutic target for cardiovascular diseases and the identification of metabolites function as well as other factors affecting the metabolites thoroughly leads to better management of CVDs. The arachidonic acid pathway is part of the KEGG subcollection of Canonical gene set collection.

1.4 Diet and gene expression

It has been shown that diet affects gene expression by regulating gene expression directly, through their metabolites and through signal transduction molecules (Berná et al., 2014). The study of nutrients' influence on gene expression is called nutrigenomics and constituents of food that can affect whole organism through gene expression are called bioactive components of the diet (Farhud et al., 2010; Mierziak et al., 2021). Understanding how bioactive components of the diet affects gene expression and molecular pathways will be useful information when making dietary guidelines and even when giving personalized nutritional guidance.

1.4.1 Animal models and human

Using animal models in dietary intervention studies have several advantages compared to human research, however no animal model is entirely human-like. This leads to the fact that it is important to evaluate which animal is the right animal for the study under

conduct. That can be problematic when there is no suitable animal model for the study subject of interest. Usage of model animals has several benefits compared to human studies, such as invasive tissue samples can be collected, life cycles are shorter thus allowing life-long follow-up with faster results (Baker, 2008). Commitment to long-term study which requires regular visits in a clinic is not always possible for all human participants and thus may lead to drop out of the study.

Studies that collect data from humans have more ethical limitations for tissue sample collection and takes longer time to get results from corresponding phase of life or even whole life cycle. Finnish law defines that in medical research human well-being comes before science thus benefits from the study must be greater than the harm that may be caused to participants (Laki lääketieteellisestä tutkimuksesta 1999/488). Medical research needs an approval from medical research ethics committee. Body fluids, such as saliva, urine, and feces, can be collected without causing harm. Blood sample collection is invasive but the harm that it causes is very minimal and because of this, the blood sample is very commonly used. Left over samples from medical care are way to get tissue samples, for example from brain, that could not be collected for research use only.

Experiments of dietary modifications are easier to control with animals than with humans. Examples of common animal species used in nutritional research are pig, rat and mice (Baker, 2008). It has been shown in mice, that polyunsaturated fatty acid (PUFA)-enriched diets affect gene expression of several genes in central nervous tissue at different ages and both omega-3 and omega-6 PUFAs in a diet appeared to have a complex impact to gene expression. (Kitajka et al., 2004). Gene expression of obesity associated gene (*FTO*) in young minipigs, used as a model for early stage of atherosclerosis in children, showed equal mRNA levels in cortex and cerebellum in normally fed control minipigs, whereas *FTO* expression was significantly lower in cerebellum compared to cortex in high-cholesterol fed minipigs (Madsen et al., 2009). This finding indicates that high-cholesterol diet influences *FTO* expression levels in different parts of brain. This kind of study could not be repeated in humans because brain tissue sample collection and particularly collection of invasive samples from children is against the ethical principles of medical research.

In the human studies the association between diet and gene expression is usually observed in cross-sectional or short-term longitudinal studies. Cross-sectional studies can be used

to study gene expression differences between populations with different nutritional habits or between groups formed based on information from dietary journals. Longitudinal studies can be used for observing changes through time. When dietary intervention is added to longitudinal study it offers more comprehensive information.

1.4.2 Dietary fat interventions

Intervention studies are suitable for finding out clinical or molecular impact of a single or more comprehensive dietary changes. Dietary interventions can be conducted with dietary guidance which offers information to the participant, but dietary choices are made by the participant itself, or it can be conducted using stricter dietary plan or added dietary supplement. The effects of dietary fat can be studied through intervention studies by increasing, reducing, or changing the composition of dietary fats.

It is known that dietary fat quantity and quality are risk factors for several diseases. Fatty acids in a diet can be divided into three main categories based on the presence or absence of one or more double bonds in the carbon chain. Saturated fatty acids, SAFAs, do not contain a double bond. Monosaturated fatty acids, MUFAs, contains one double bond and polyunsaturated fatty acids, PUFAs, contain more than one double bonds. WHO and Nordic Nutrition Recommendations 2012 both suggests among other things that SAFA intake should be limited less than 10 E% (Nordic Council of Ministers, 2012; World Health Organization, 2020). Butter, beef, and coconut oil are examples of foods that contain high SAFA content.

Changing dietary fat composition, by replacing SAFAs with PUFAs, in healthy participants with moderately elevated LDL-cholesterol levels increases the mRNA levels of liver X receptor α and LDL receptor measured from peripheral blood mononuclear cells, in addition to liver X receptor α target genes and genes involved in inflammation, while the mRNA levels of uncoupling protein and peroxisome proliferator-activated receptor δ were downregulated (Ulven et al., 2019). Low carbohydrate and high fat diet among normal-weight young adults seems to increase the mRNA expression of transcription factor SREBP-1 when compared to control group (Retterstøl et al., 2018). However, it is difficult to determine precisely whether the changes in mRNA expressions are due to the factor of interest, in this case of dietary fats. If possible other dietary

components and other lifestyle factors, such as exercising, smoking, and sleeping, should be kept as stable as possible during studies. Other possibility is to follow these factors as well and to use these factors as covariates to reduce bias.

Human studies with different study designs have shown consistent evidence for Mediterranean dietary lowering the risk for CVDs (Mozaffarian et al., 2011). Mediterranean diet (MD) consists mainly of vegetables, fruits, fish, olive oil, and nuts whereas the consumption of meat and saturated fats is low. The three months effects of Mediterranean diet supplemented with nuts or virgin olive oil (VOO) compared with low fat diet to the gene expression and biological pathways related to CVDs was studied as part of the PREDIMED study from peripheral blood mononuclear cells (the PREDIMED study investigators, 2013). In this intervention, four genes, *IL1*β, *IL1RN*, *TNF-α* and *ICAM*, from the atherosclerosis signaling pathway were downregulated by MD together with VOO supplement. *VEGF* gene, for example part of hypoxia pathway, was downregulated by MD with both supplements but *VEGF* related gene *HIF1*α was downregulated only by the MD with nuts and *NF- κβ* gene was downregulated by MD with VOO. Also, the *JUN* gene was downregulated by both MDs and played central role in the downregulation of seven hypertension related pathways.

1.5 Aims of the study

Aims of the study are to find out if saturated fatty acid intake affects transcriptome levels measured both at the individual gene expression level and at the pathway level. Pathway enrichment analysis is conducted by using two different methods and as they use different input data's the effects of input data form are compared.

Current scientific literature completely lacks functional genetic data from long-term infancy-onset randomized clinical interventions targeting dietary fat quality in healthy human population. Preliminary studies in STRIP have found differences between intervention and control groups in gene expression measured from whole blood samples (Mykkänen et al. unpublished, Research Center of Applied and Preventive Cardiovascular Medicine, University of Turku). However, transcriptome-wide effects of achievement of dietary fat intervention target are still unknown.

2 Data and methods

2.1 STRIP clinical and mRNA-Seq data

Parts from pre-collected data from the longitudinal STRIP study was used for the analysis. In every STRIP-study visit children and their families reported a standard 4-day food diary and the food records were reviewed for completeness and accuracy by a trained nutritionists together with the child or family. From the food diary SAFA intake, among other nutrients, was analysed with the Micro-Nutrica® food analysis software developed at the Research and Development Center of the Social Insurance Institution, Finland (Keskitalo et al., 2022). At 2007-2009, when the participants were at the age 16-19 years, a single whole blood RNA (PAXgene, BD Biosciences) sample was collected from 523 participant by a medical laboratory technologist.

At 2019 total RNA, including miRNA, was isolated from the samples using MagMAX™ for Stabilized Blood Tubes RNA isolation kit (Invitrogen). Genome-wide messenger RNA sequencing (mRNA-Seq) was done using Illumina NovaSeq 6000 platform using 50 bp paired-end sequencing with an average of 22 million pair-end reads per sample. The samples were sequenced in 8 pools and 1–3 runs per sample were performed. The RNA isolation and mRNA-Seq were done in an accredited testing laboratory (Finnish Functional Genomics Centre, Bioscience Turku). A trained bioinformatician pre-processed the ~1.9 terabytes of raw sequence data and performed gene-level quantification against human reference genome build GRCh38 using methods meeting the best practices of RNA-Seq data analysis within CSC ePouta high performance computing environment for sensitive human data (Conesa et al. 2016). Quality score profiles of the raw sequence reads were generated with FastQC using default settings (Andrews, 2023). The profiles were inspected by collating the results from all the samples using multiQC (Ewels et al., 2016). Gene expression counts were obtained using Rsubread R/Bioconductor software package 2.6.4 pipeline (Liao et al., 2019). Alignment of the sequence reads to human genome reference GRCh38/hg38 was done using align function of the Rsubread package. Matrix of gene expression counts was obtained using featureCounts function (Liao et al., 2014).

From the collected 523 whole blood samples 28 samples were excluded for example based on bad sample quality or participants having a congenital or chronic disease. The

final analysis-ready transcriptome data included 495 STRIP participants (249 girls and 246 boys) and gene counts for 25562 genes. The sample size is one of the largest ever generated in context of human dietary interventions.

2.2 Differential gene expression analysis

For the analysis necessary clinic data was filtered from the vast longitudinal STRIP datasets in SAS 9.4 software. Information from the variables age, gender, mRNA-seq batch, sequence number and SAFA E% in the time point when the participants were at the age 16–19 years was filtered from the clinic STRIP data. Achieving the goals of nutritional intervention was determined by the SAFA E% which is intake of saturated fatty acids (SAFA) of total energy intake (E%). SAS codes for data filtering presented in Appendix 1. SAFA E% was used to divide participants into two groups based on nutritional recommendations. Those that achieved to goal of SAFA E% being under 10 % formed a first group and the others whose SAFA E% was over 10 % formed the second group. Additionally, continuous SAFA E% values were standardized with proc standard command in SAS so that mean was 0 and standard deviation was 1. SAFA E% symmetricity was confirmed in SAS by checking distribution plot and skewness value. SAS codes for standardization and symmetricity check are presented in appendix 2. Distribution plot showed visually symmetrical distribution and as skewness value 0.194 was between -0.5–0.5 it also indicated that the distribution is fairly symmetrical (Piovesana and Senior, 2018). Filtered clinical data was exported to Excel and tidied up.

In addition to the aforementioned, SAS software was used to calculate means and standard deviations for several phenotypic values among study groups. With proc ttest command independent group t-test for each variable was conducted in SAS. Pooled method for t-test was used when the two-tailed significance probability (Pt > F) was < 0.05 and Satterthwaite method when the Pt > F was > 0.05. SAS codes for calculating means and standard deviations together with t-test analyses presented in appendix 3. Phenotypic mean values together with standard deviations and t-test results were merged to a one table.

Clinical data, mRNA-seq data and gene annotation data were loaded to RStudio 2022.07.2+576 which uses R-4.2.1 language. BiocManager package version 1.30.18 was

installed in order to install other packages, such as DESeq2 version 1.36.0 from the Bioconductor project (Gentleman et al., 2004; Love et al., 2014). DESeq2 method was used for the DGE analysis based on its excellent results for precision, accuracy and sensitivity compared to other methodologies for DGE analysis with RNA-seq data (Costa-Silva et al., 2017). I had mRNA-seq data from 495 STRIP participants but SAFA E% information was missing from 70 participants for unknown reasons and those participants were excluded from the differential gene expression (DGE) analysis. DGE analysis was performed for 425 participants and 25562 genes. The analysis design formula contained the variable of interest, SAFA E% groups, and covariates age, gender and mRNA-seq batch. Based on low count of DE genes, same analysis was performed by changing the variable of interest to the continuous standardized SAFA E%. R codes for differential gene expression analysis with DESeq2 are presented in appendix 4.

DESeq2 package uses negative binomial (NB) generalized linear model for the differential expression analysis using the equations below (Love et al., 2014). The model indicates the read count K_{ij} for gene i in sample j using NB distribution with fitted mean expression value μ_{ij} and gene-specific dispersion α_i. The gene-specific dispersion α_i describes the variance of counts and the fitted mean expression μ_{ij} is composed of quantity q_{ij}, proportional to the expected true concentration of fragments for sample, scaled by sample-specific size factor s_j. Below, the last equation estimates the log2 FC value with the coefficients β_i and design matrix elements $x_{j.}$ for gene i.

$$K_{ij} \sim NB(\mu_{ij}, \alpha_i)$$

$$\mu_{ij} = s_j q_{ij}$$

$$log_2(q_{ij}) = x_{j.}\beta_i$$

Positive log2 FC values referred to upregulated genes and negative values downregulated genes. When comparing the differences between groups genes that got positive log2 FC values are expressed more in the group of participants achieving the SAFA E% goal compared to the group where the goal was not achieved. With the continuous variable increasing value of the SAFA E% with positive log2 FC indicates that gene is expressed more with increasing SAFA E%. When the log2 FC value was negative the gene is expressed less when the SAFA E% value increases.

I used the mostly default methods of the DESeq2 package. The Wald test for hypothesis testing and the Benjamin-Hochberg's procedure to correct for multiple testing. Genes that passed the Wald test p-values were adjusted for multiple testing using the method of Benjamini and Hochberg that controls false discovery rate (FDR) and provides adjusted p-value (Benjamini and Hochberg, 1995). Statistical significance was determined by the Benjamin-Hochberg's adjusted p-value by setting the cutoff to < 0.05. DESeq2 package performs independent filtering and uses default cutoff 0.1 but that was changed to correspond same cutoff used with adjusted p-value (0.05).

Volcano plot was created with ggplot function of ggplot2 R package and P-value, and adjusted P-value cutoffs were added to the plot. Cutoff lines for p-value and adjusted p-value were added to the plot as well as gene symbols for DE genes with smallest p-values. Appendix 5 presents the R codes used to produce the volcano plot.

2.3 Pathway enrichment analysis

Pathway enrichment analysis was done by using two different methods. The list of DE genes were used as input data for GOseq method and read counts of 25562 genes as input data for Empirical Analysis of Digital Gene Expression Data in R (edgeR) (Robinson et al., 2010; Young et al., 2010). Hallmark pathway, KEGG sub-collection and ontology collections from The Molecular Signatures Database v2022.1.Hs (MSigDB) were included in the enrichment analyses with both analysis methods (Liberzon et al., 2011). Hallmark gene sets included 4384, KEGG sub-collection 5245 and ontology gene sets 19447 genes but there were genes that belonged to more than one category leading to the gene lists lengths being 7322 for hallmark gene set, 12797 for KEGG sub-collection and 1281285 for ontology gene set.

With GOseq method Log2 fold changes (FC) from DE analysis were used in the pathway enrichment analysis to determine if the differentially expressed genes were up- (log2FC > 0) or downregulated (log2FC < 0). The Wald test and the Benjamin-Hochberg's method was used to get the p-values and adjusted p-values which established significant results (<0.05). GOseq method uses the probability weighting function (PWF) to quantify the gene length bias present in the dataset, in other words it gives the probability that a gene will be differentially expressed based on its length (Young et al., 2017). GOseq analyses

were conducted for each gene set twice, first with downregulated genes and then with upregulated genes, so the separate-DE strategy was used. The separate-DE strategy is more powerful than all-DE strategy where all DE genes would be analyzed together (Hong et al., 2014). Horizontal bar graphs from up- and downregulated genes for each gene set was created in R with ggplot function to illustrate enriched pathways.

In R, identified which DE genes were overlapping with the significantly up or downregulated pathways found from GOseq analyses. Appendix 6 presents codes for this procedure. Code presented for one pathway collection (ontology).

To compare if input data form has an effect to the results, pathway analysis was repeated by using the second RNA-Seq data analysis method, edgeR, which uses read counts of genes as input data. Analyses with edgeR method was first produced with the class variant (SAFA E% < 10 or SAFA E% > 10) but no significant results were found so as with DE analyses continuous standardized SAFA E% was used for edgeR analyses as well. To test which genes were DE and enriched to pathways, gene set test fry was used because it considers all genes in the set and do not depend in any significance cutoff (Chen et al., 2016). The edgeR did not filter independently low expressed genes as default contrary to DESeq2, but filtering was done separately with the filterByExpr function (Law et al., 2016). Analyses were made for each gene set separately and horizontal bar graphs for each gene set was created with ggplot function. As difference to GOseq method, edgeR method used all-DE strategy for analyses. Appendix 7 presents codes for edgeR pathway analyses. Code presented for one pathway collection (Hallmark).

3 Results

Transcriptome data from 495 participants was available but information about SAFA E% in the time point of interest was missing from 70 participants thus 425 participants was filtered for the analyses. Table 1 presents key factors from the data when participants were grouped into those who had SAFA E% under 10 % (n=90) and those who had it over 10 % (n=335). Based on group sizes 26.9 % of participants achieved the SAFA E% goal. 64.4 % of those who achieved the SAFA E% goal were girls, mean age was 16.9 years and mean SAFA E% 8.5 %. 47.5 % of those who did not achieve the goal were girls, mean age was 17.0 years and mean SAFA E% 13.2 %. The key factors for heart health do not differ statistically between the study groups.

Table 1. Phenotypic values (percentage or mean value and standar deviations (sd)) among the study groups. Grouping based on SAFA E% being <10 (n=90) or >10 (n=335). Statistical significance based on p-values from independent group t-test for each variable are represented in the table. ** denoted for p-value < 0.01, **** denoted for p-value < 0.0001. Table abbreviations: body mass index (BMI), low-density lipoprotein (LDL), high-density lipoprotein (HDL), blood pressure (BP) and intake of saturated fatty acids of total energy intake (SAFA E%).

N=425	**SAFA E% < 10 (N=90)**		**SAFA E% > 10 (N=335)**	
Variable	N	% or mean (sd)	N	% or mean (sd)
Girls	58	64.4 **	159	47.5 **
Age	90	16.9 (0.7)	335	17.0 (0.7)
BMI	90	21.6 (3.8)	335	21.4 (31.1)
Total cholesterol	90	4.2 (0.7)	335	4.0 (0.7)
LDL cholesterol	90	2.5 (0.6)	332	2.4 (0.6)
HDL cholesterol	90	1.2 (0.3)	332	1.2 (0.3)
Triglycerides	90	1.0 (0.5)	335	0.9 (0.4)
Systolic BP	90	115.7 (11.9)	333	117.6 (12.9)
Diastolic BP	90	61.6 (7.3)	333	60.5 (6.7)
SAFA E%	90	8.5 (1.2) ****	335	13.2 (2.2) ****

3.1 Differential gene expression analysis

Differential gene expression analysis was made twice with DESeq2. Class variable (SAFA E% groups) was used as variable of interest in the analysis but based on the results analysis was repeated as the first but continuous SAFA E% was used as variable of interest. With the class variable four DE genes were found when B-H adjusted P-value 0.1 was used as cutoff value but only three DE gene was found when the cutoff value was set to 0.05.

Table 2 shows top 10 results from differential gene expression analysis where class variable was used as variable of interest. First three genes (*ANPEP*, *RAP1GAP* and *OTOF*) are DE genes with statistically significant B-H adj. P-value (< 0.05). *ANPEP* (alanyl aminopeptidase, membrane) is protein coding gene which is expressed mainly in intestine and pancreas tissues. *RAP1GAP* gene encodes a GTPase-activating protein which indirectly inactivates RAP1 which is part of receptor-linked signaling pathways thus controls cell differentiation and growth. *OTOF* gene encodes calcium ion sensor protein that plays important role in the synaptic vesicle-plasma membrane fusion.

Table 2. Top 10 DE genes based on the B-H adjusted P-value from the differential gene expression analysis with class variable. Table holds NCBI entrezgene ID, symbol, and description for each gene together with Log2 Fold Change, standard error, P-value, and B-H adjusted P-value.

NCBI entrezgene ID	Gene symbol	Gene description	Log2 Fold Change	SE	P-value	B-H adj. P-value
290	*ANPEP*	alanyl aminopeptidase, membrane	-0.284	0.057	6.29E-07	1.60E-02
5909	*RAP1GAP*	RAP1 GTPase activating protein	-0.875	0.191	4.44E-06	3.76E-02
9381	*OTOF*	otoferlin	1.009	0.219	4.11E-06	3.76E-02
51187	*RSL24D1*	ribosomal L24 domain containing 1	0.419	0.094	9.28E-06	5.91E-02
1290	*COL5A2*	collagen type V alpha 2 chain	0.313	0.075	2.73E-05	1.16E-01
6189	*RPS3A*	ribosomal protein S3A	0.402	0.096	2.61E-05	1.16E-01
55225	*RAVER2*	ribonucleoprotein, PTB binding 2	0.303	0.074	4.06E-05	1.42E-01
84335	*AKT1S1*	AKT1 substrate 1	-0.140	0.034	4.45E-05	1.42E-01
4782	*NFIC*	nuclear factor I C	-0.143	0.036	5.86E-05	1.49E-01
60592	*SCOC*	short coiled-coil protein	0.319	0.079	5.61E-05	1.49E-01

Second differential gene expression analysis with continuous SAFA E% gave the result of 492 DE genes when B-H adjusted P-value 0.1 was used as cutoff value and 74 DE genes when the cutoff value was set to 0.05. From the 74 DE genes ten were down and 64 were upregulated genes based on log2 FC values. The 10 downregulated genes (B-H adj. P-value < 0.05 and log2 FC < 0) and corresponding information as shown in the table 2 for the genes is presented in appendix 1. Four out of ten downregulated genes, IGLV1-47, IGKV4-1, IGKV3D-20, and IGHV3-33 are immunoglobulin genes. The other downregulated genes have roles in DNA methylation (TDG), transcription regulation

(ZSCAN29 and EED), cilia formation (IQCB1), cell-cycle regulation (UHRF2) and myogenic transcription and differentiation (TAF5L).

The 64 upregulated DE genes (B-H adj. P-value < 0.05 and log2 FC > 0) and corresponding information as shown in the table 2 for the genes is presented in appendix 2. The upregulated DE genes seem to have functions in wide range and based on the table no enrichment to specific biological functions are observed. The top 10 upregulated DE genes, based on B-H adjusted P-value, function in the immunoregulation (*FKBP2*), regulation of actin modulating proteins (*DYNLRB1*), regulation of hormone-induced cardiomyocyte hypertrophy (*EDF1*), modulating glutaredoxin activity (*SH3BGRL3*), the isomerization of specific enoyl-CoA (*ECH1*), the non-specific binding to the DNA (*BANF1*), and as a subunit of the oligosaccharyl transferase complex (*KRTCAP2*), mitochondrial complex I (*NDUFA3*), mitochondrial ATP synthase (*ATP5F1D*), and the cytochrome c oxidase (*COX6B1*).

Volcano plot (figure 1) illustrates how different DE genes are settle in relation to one another based on log2 FC and –log10(P-value). In the plot red circles present upregulated genes, blue circles downregulated genes and gray circles insignificant (B-H adj. P-value > 0.05) genes. Gene symbols of DE genes that showed strong significance are presented in the plot. Gray dashed line present cutoff value for P-value (-log10 (0.05)) and black dashed line presents value which corresponds to B-H adjusted P-value cutoff. Figure 1 illustrates the difference of chosen cutoff method to the results. For example, gene *FKBP2* had positive log2 FC value (0.08), highest –log10 (P-value), which corresponds to lowest P-value and was considered as statistically significant result.

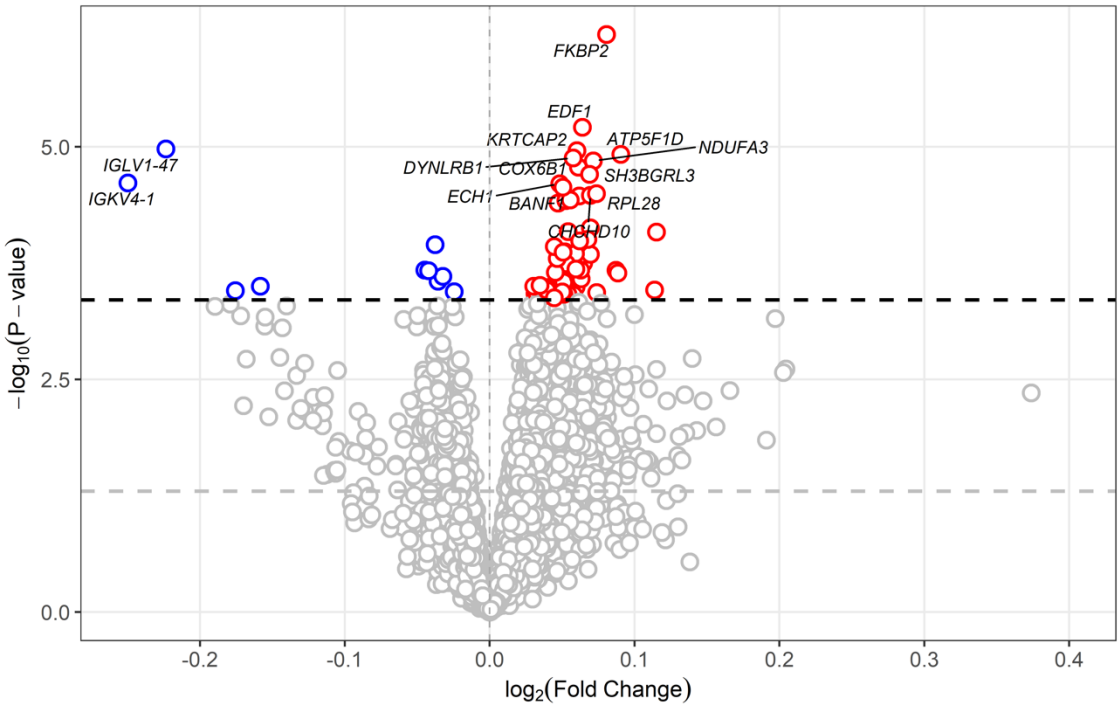

Figure 1. Volcano plot of analyzed genes where red circles are upregulated (B-H adj. P-value < 0.05 and log2 FC > 0), blue circles downregulated (B-H adj. P-value < 0.05 and log2 FC < 0) genes and gray circles insignificance genes (B-H adj. P-value > 0.5). Cutoff values for P-value and B-H adj. P-value presented with gray and black dashed lines.

3.2 Pathway enrichment analysis with GOseq

Results from pathway enrichment analyses with GOseq presented as horizontal bar graphs for each separate analysis made with MSigDB collections and also results from up and downregulated genes presented separately. Upregulated genes showed enrichment to gene set collections categories with every MSigDB collections used. Upregulated genes were enriched in two Hallmark gene set collection categories (figure 2) but neither crosses the cutoff value because B-H adj. P-value > 0.05. Out of 64 upregulated DE genes four were assigned to the HALLMARK_PI3K_AKT_MTOR_SIGNALING category which consists of 105 genes and six upregulated genes were assigned to 200 genes that consists of the HALLMARK_OXIDATIVE_PHOSPHORYLATION category.

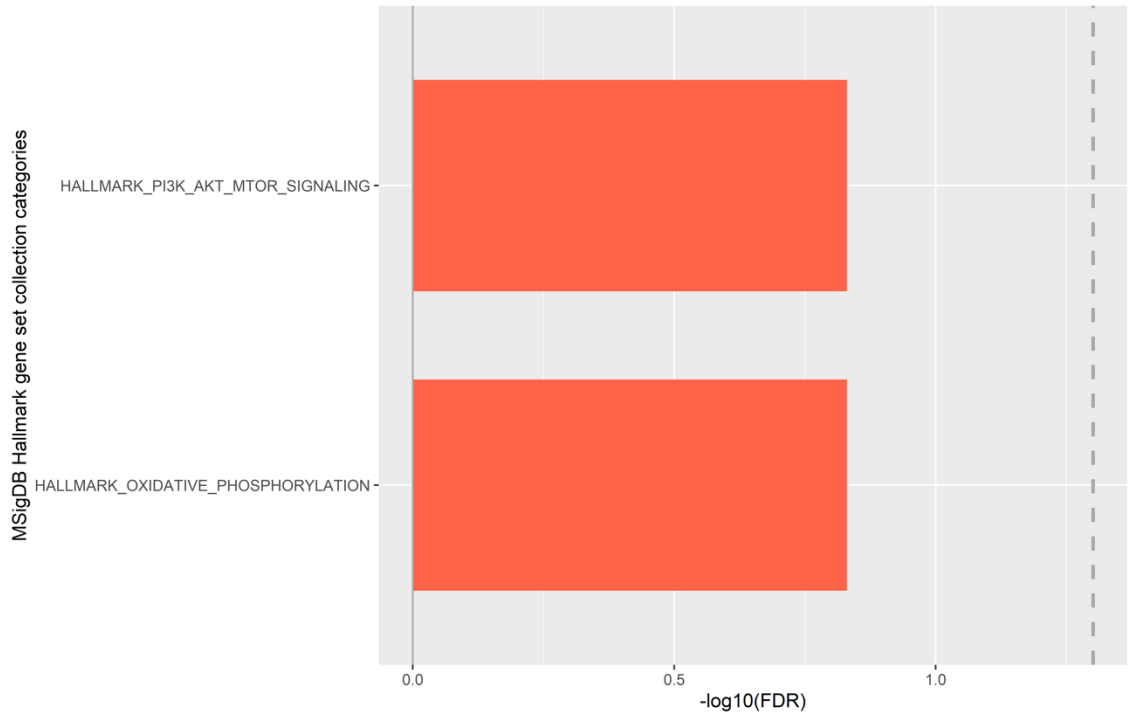

Figure 2. Enriched MSigDB Hallmark gene set collection categories from GOseq analysis made with upregulated DE genes. Red bars present –log10 values from B-H adjusted P-value for Hallmark gene set collection categories. Gray dashed line shows FDR cutoff<0.05 at -$\log_{10}$ scale.

With KEGG subset of Canonical pathways gene set collection categories upregulated DE genes were enriched to four KEGG categories (Figure 3), but none crosses the B-H adj. cutoff value. Out of 64 upregulated DE genes five genes were assigned to both KEGG_PARKINSONS_DISEASE and KEGG_HUNTINGTONS_DISEASE (consists of 113 and 172 genes respectively) categories and four genes were assigned to both KEGG_OXIDATIVE_PHOSPHORYLATION and KEGG_ALZHEIMERS_DISEASE (consists of 155 and 116 genes respectively) categories.

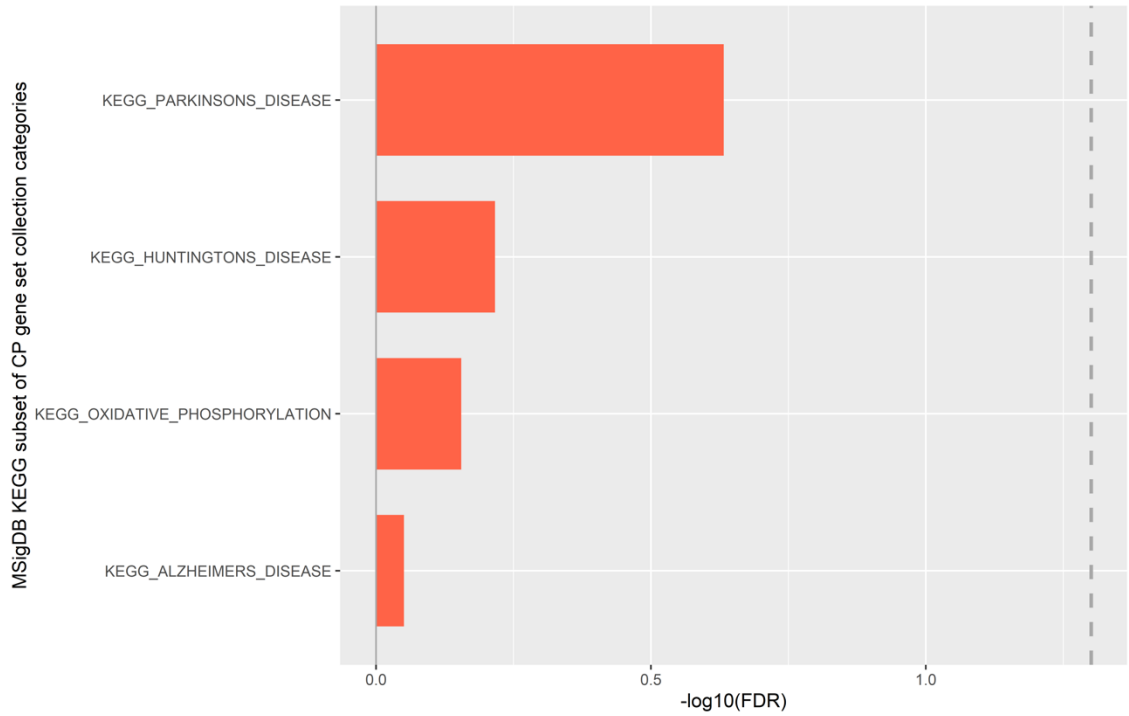

Figure 3. Enriched MSigDB KEGG subset of Canonical pathways gene set collection categories from GOseq analysis made with upregulated DE genes. Red bars present –log10 values from B-H adjusted P-value for KEGG subset gene set collection categories. Gray dashed line shows FDR cutoff<0.05 at -$\log_{10}$ scale.

Upregulated DE genes showed enrichment to 93 Ontology gene set collection categories and top 10 were included to the bar graph (Figure 4). One category, GOBP_MEMBRANE_BIOGENESIS, crossed the cutoff value thus had B-H adjusted P-value < 0.05. From the 64 upregulated DE genes five were assigned to this category which consist of 61 genes. GOMF_CARBONYL_REDUCTASE_NADPH_ACTIVITY category consists of nine genes and three upregulated DE genes were assigned to those, but B-H adjusted P-value was just over 0.05 so it was considered statistically not significantly enriched pathway. Other categories presented in the figure 4 consists of 28–1277 genes and three to 15 upregulated DE genes were assigned to these categories but neither of them showed statistical significance.

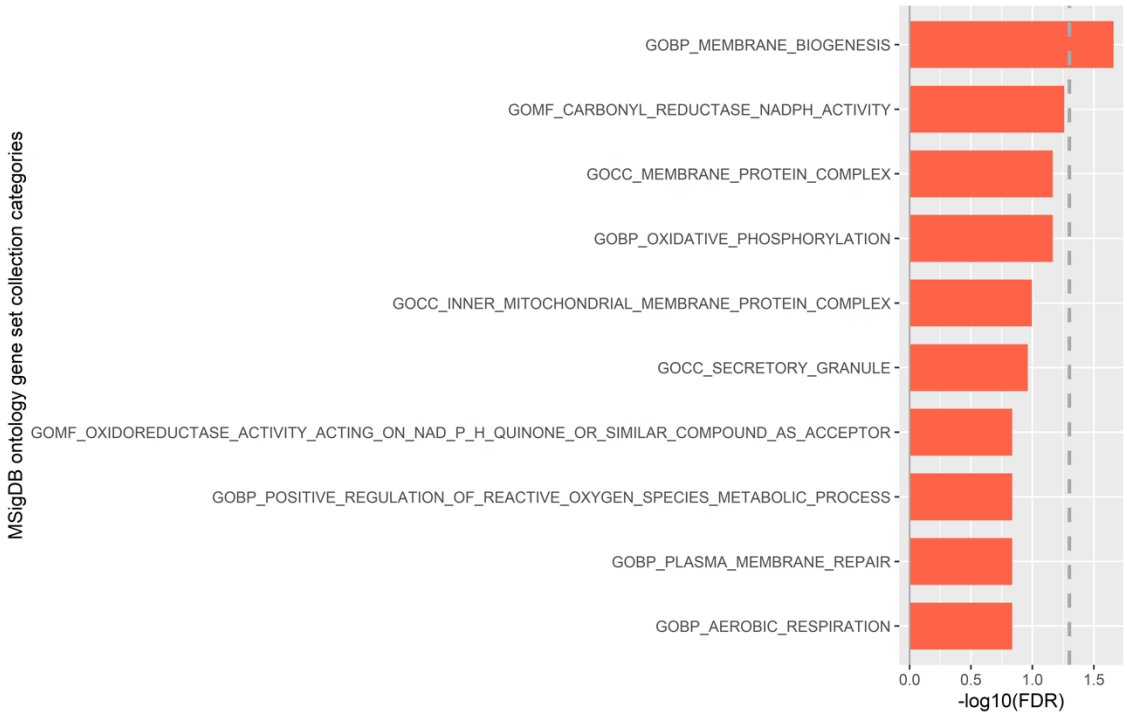

Figure 4. Top 10 enriched MSigDB ontology gene set collection categories from GOseq analysis made with upregulated DE genes. Red bars present –log10 values from B-H adjusted P-value for ontology gene set collection categories. Gray dashed line shows FDR cutoff<0.05 at -$\log_{10}$ scale.

Analyses did not show any enrichment to categories when analyses were made with downregulated DE genes for Hallmark gene set collection categories or for KEGG subset of Canonical pathways gene set collection categories. Downregulated DE genes showed enrichment to Ontology gene set collection categories (Figure 5) and GOCC_IMMUNOGLOBULIN_COMPLEX crossed the cutoff and was considered statistically significant enrichment. From the 10 downregulated DE genes four were assigned to the category which consists of 147 genes. GOMF_ANTIGEN_BINDING category consist of 158 genes and GOBP_IMMUNOGLOBULIN_PRODUCTION of 213 genes and three downregulated DE genes were assigned to each category, but they did not cross the cutoff (B-H adjusted P-values > 0.05) and was considered statistically insignificant results.

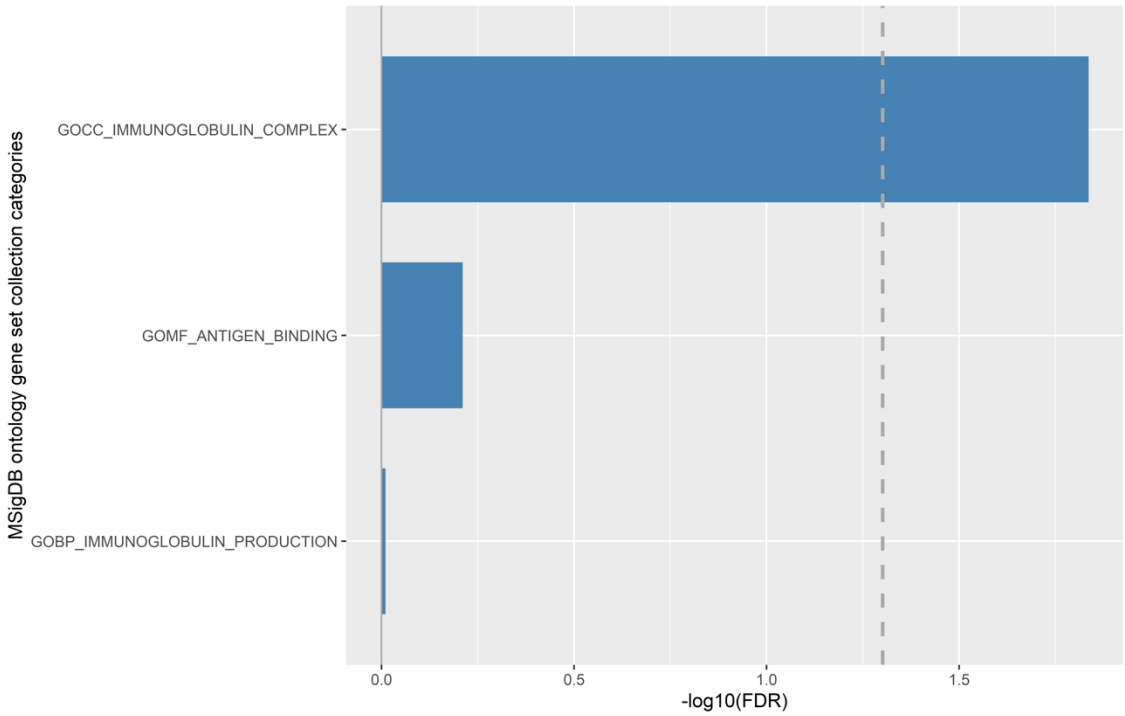

Figure 5. Enriched MSigDB ontology gene set collection categories from GOseq analysis made with downregulated DE genes. Blue bars present –log10 values from B-H adjusted P-value for ontology gene set collection categories. Gray dashed line shows FDR cutoff<0.05 at -$\log_{10}$ scale.

Table 3 recapitulates the statistically significance enriched gene set categories from GOseq analyses and presents the number of genes in each category and the number together with symbols for DE genes that overlapped with categories. Statistical significance was determined based on B-H adjusted P-value. Five upregulated genes overlapped with the 61 genes that form COBP_MEMBRANE_BIOGENESIS category and four downregulated genes overlapped with the 147 genes that form GOCC_IMMUNOGLOBULIN_COMPLEX category. The five upregulated genes overlapping with COBP_MEMBRANE_BIOGENESIS category plays roles in endosomal sorting complex required for transport III (*CHMP2A*), inhibition of PCSK9-enhanced LDLR degradation (*ANXA2*), exocytosis and endocytosis and induces the dimerization of ANXA2/p36 (*S100A10*), mediating vesicle fusion in neurons (*STX4*) and the non-specific binding to the DNA (*BANF1*). The four downregulated genes overlapping with GOCC_IMMUNOGLOBULIN_COMPLEX category are immunoglobulin genes, one that codes heavy chain variable 3-33 (*IGHV3-33*), one that codes lambda light chain variable (*IGLV1-47*) and two that codes kappa light chain variables (*IGKV3D-20* and *IGKV4-1*).

Table 3. Ontology gene set collection categories which passed B-H adjusted P-value cut off value (0.05) and symbols for overlapped DE genes. Column numInCat presents number of genes in gene set category and numDEInCat presents the number of differentially expressed genes in numInCat. P-value and B-H adjusted P-value shows the statistical significance for GOseq analyses.

Ontology gene set category	numIn Cat	numDEIn Cat	Gene symbols for overlapped DE genes	P-value	B-H adj. P-value
GOBP MEMBRANE BIOGENESIS	61	5	*CHMP2A* *ANXA2* *S100A10* *STX4* *BANF1*	1.41E-06	0.0222
GOCC IMMUNOGLOBULIN COMPLEX	147	4	*IGHV3-33* *IGLV1-47* *IGKV3D-20* *IGKV4-1*	9.33E-07	0.0147

3.3 Pathway enrichment analysis with edgeR

Results from pathway enrichment analyses for continuous SAFA E% value made with edgeR presented with horizontal bar graphs, as GOseq results, for each separate analysis made with different MSigDB collections. For the clarity of graphs categories were filtered for graphs based on FDR values. With Hallmark gene set collection categories gene counts showed changes up and down directions (figure 6) and one category, HALLMARK_SPERMATOGENESIS, crossed the cutoff thus shows statistical significance. Most of the enriched categories showed changes in up direction but none of those crossed the cutoff.

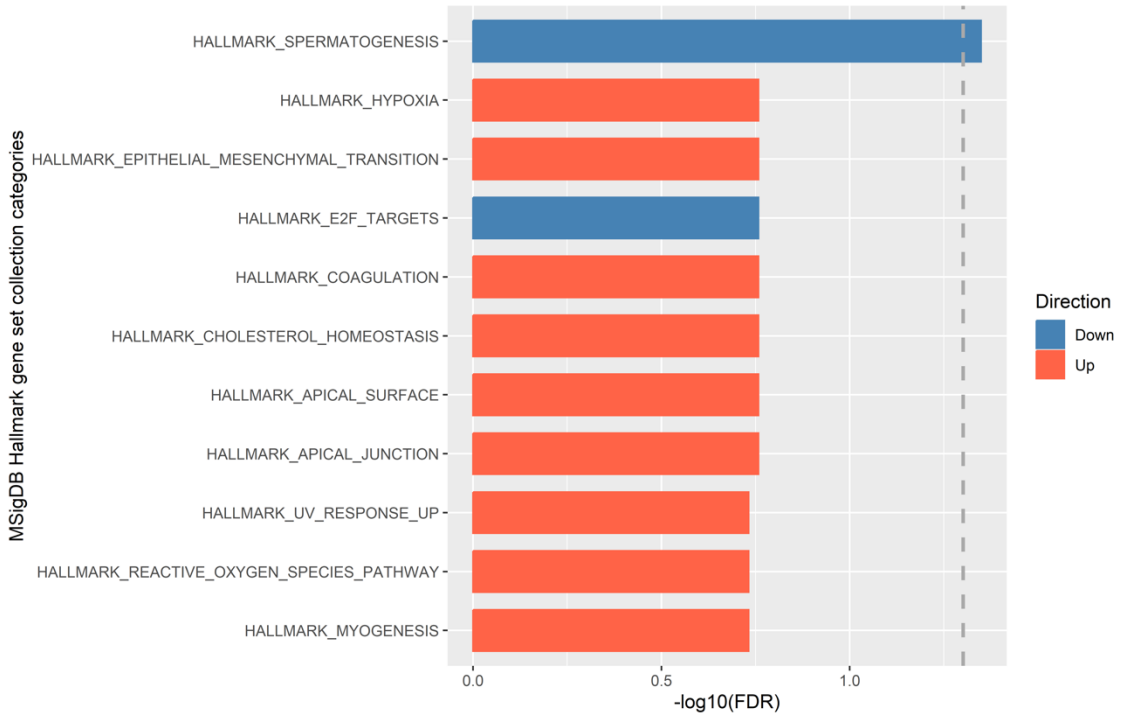

Figure 6. Enriched MSigDB Hallmark gene set collection categories which resulted B-H adjusted P-value < 0.20 from edgeR analysis. Bars present –log10 values from B-H adjusted P-value for Hallmark gene set collection categories. Color of the bar indicates the net direction of change, blue down and red up. Gray dashed line shows FDR cutoff<0.05 at -$\log_{10}$ scale.

Analyses with edgeR showed enrichment for the KEGG subset of Canonical pathway gene set collection categories and Ontology gene set collection categories (figures 7 and 8). Analyses indicated changes both up and down directions for categories at each collection. None statistically significant results were found nether with KEGG or Ontology collections even though a lot of categories showed enrichment.

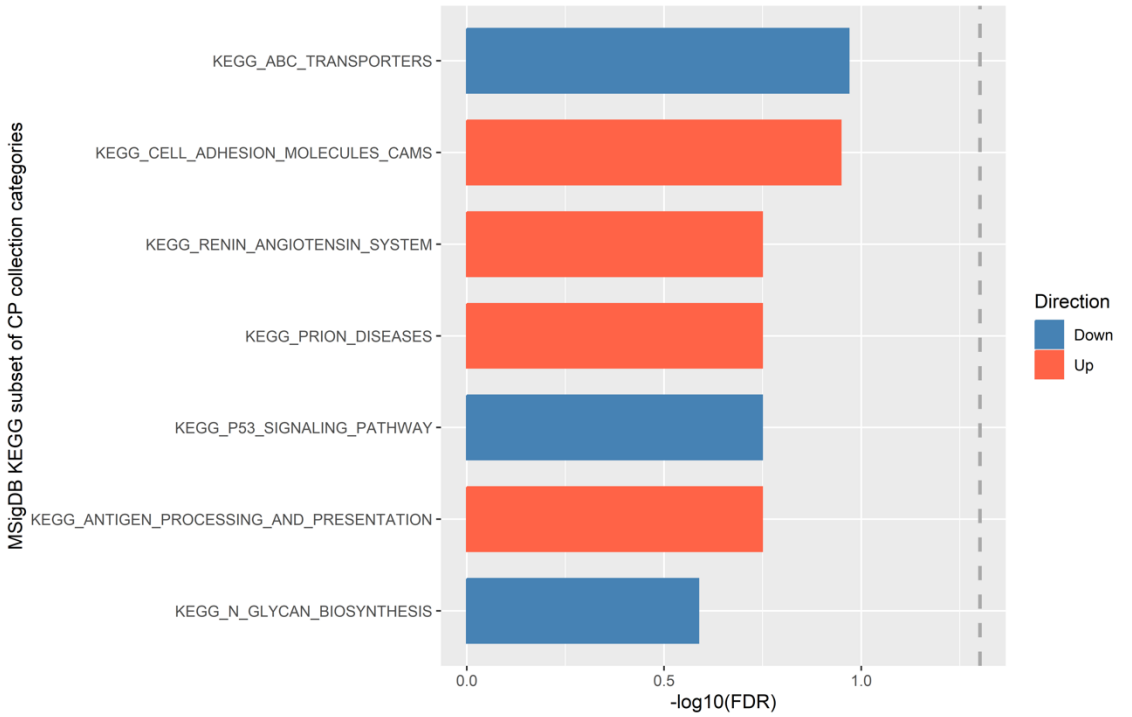

Figure 7. Enriched MSigDB KEGG subset of Canonical pathway gene set collection categories which resulted B-H adjusted P-value < 0.30 from edgeR analysis. Bars present –log10 values from B-H adjusted P-value for KEGG gene set collection categories. Color of the bar indicates the net direction of change, blue down and red up. Gray dashed line shows FDR cutoff<0.05 at -$\log_{10}$ scale.

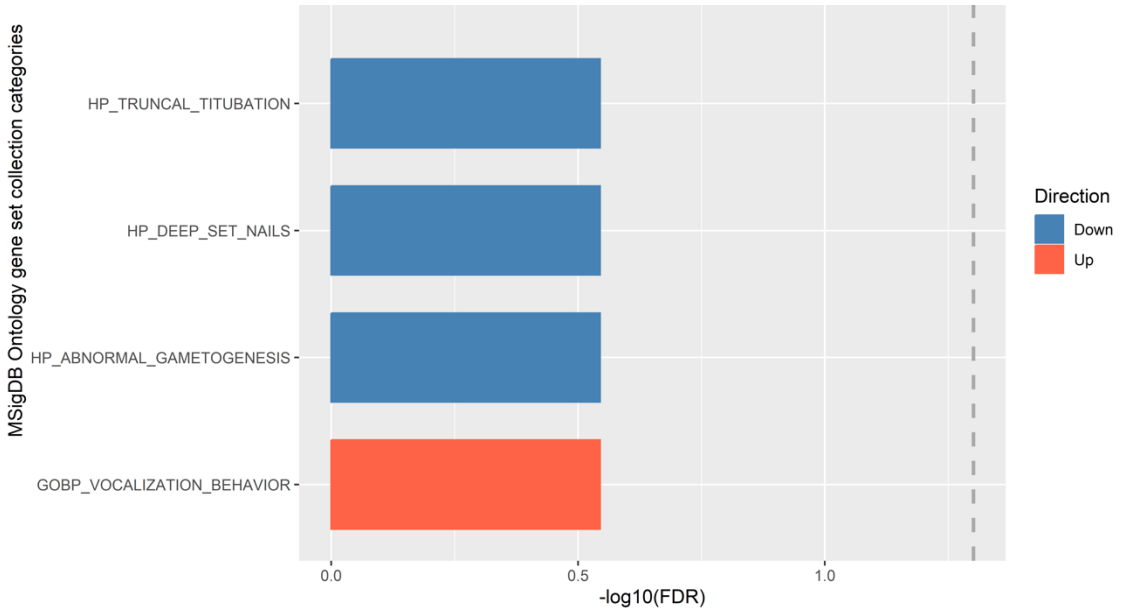

Figure 8. Enriched MSigDB Ontology gene set collection categories which resulted B-H adjusted P-value < 0.30 from edgeR analysis. Bars present –log10 values from B-H adjusted P-value for Ontology gene set collection categories. Color of the bar indicates the net direction of change, blue down and red up. Gray dashed line shows FDR cutoff<0.05 at -$\log_{10}$ scale.

4 Discussion

This thesis aims to fill the lacking knowledge about the effects that achievement of dietary fat intervention target has on transcriptome. More precisely this thesis aims to find out if saturated fatty acid intake affects transcriptome levels at the individual gene expression level or at the pathway level. Other aim was to compare if input data format has an effect on the results of pathway enrichment analysis.

4.1 Differential gene expression analysis

Differential gene expression analysis was conducted twice, first with class variable (SAFA E% groups) and then with continuous variable as variable of interest. Results from the analysis where class variable was used (Table 2.) were not expected as only three DE genes (*ANPEP*, *RAP1GAP* and *OTOF*) had statistically significant B-H adj. p-values. The following pathway analysis could not be performed with three DE genes thus continuous SAFA E% variable was used for second differential gene expression analysis which resulted 74 DE genes (Appendix 8 and 9), the amount that allows to execute the pathway analysis.

As mentioned previously, *ANPEP*, *RAP1GAP* and *OTOF* genes were differentially expressed between those who achieved the intervention goal (SAFA E% < 10) and those who did not (SAFA E% > 10). Based on the results *ANPEP* and *RAP1GAP* genes are downregulated and *OTOF* is upregulated when nutritional recommendations regarding SAFA intake is followed. In previous research, the lack of *ANPEP* was expected to protect from atherosclerosis but contrary their beliefs *ANPEP* deficiency in mice generated enlargement of atherosclerotic lesions (Devarakonda et al., 2019). It is known that reducing SAFA intake protects against CVDs thus result regarding *ANPEP* are consistent with the hypothesis Devarakonda et al. (2019) had in their research. Previously *RAP1GAP* gene has shown to be upregulated in angiotensin II induced cardiomyocyte hypertrophy which in turn may lead to heart failure (Gao et al., 2021). The results of this thesis are consistent with the result Gao et al. found in their research, since downregulation of *RAP1GAP* gene was associated with recommended SAFA intake. Based on article search upregulated *OTOF* gene do not have association with CVDs.

As already stated, second differential gene expression analysis, conducted with continuous SAFA E% values, found 74 DE genes of which 10 were downregulated (Appendix 8). Short description of each gene is presented in the results (Chapter: 3.1 Differential gene expression analysis). Six out of these ten genes have no effect on CVDs based on article search, however it is possible that there are effects but knowledge about these is still lacking. Four immunoglobulin genes, IGLV1-47, IGKV4-1, IGKV3D-20, and IGHV3-33 are expressed less when SAFA intake increases. However, it has been found in middle-age and older people that high levels of some immunoglobulins in serum are associated atherosclerosis ja other CVDs in mainly Caucasian population (Khan et al., 2023). The inconsistence compared to the results of this thesis may be caused by differences in population or age, or by the fact that Khan et al. (2023) measured protein levels from serum whereas in this thesis mRNA levels from blood cells were analyzed.

Results from the second differential gene expression analysis were unbalanced due to 64 upregulated genes but only 10 genes were downregulated. Because description of each 64 upregulated genes (Appendix 9) individually would be unnecessary in relation to the purpose of the thesis, first 40 genes from the appendix 9 are further discussed in the thesis if the gene has a connection to the heart health and CVDs based on article search.

In continuous SAFA E% analysis, *ECH1* was upregulated with increasing SAFA E%. In mice, Enoyl-CoA hydratase 1 (Ech1, encoded by *Ech1* gene) overexpression promotes browning of white adipose tissue and energy consumption leading reduction of obesity and metabolic disorders despite high-fat diet consumed (Mao et al., 2020). Another study has shown that Ech1 expression in mouse heart and aorta was increased by high-fat diet and because it takes part in fatty acid oxidation and lipid metabolism in mitochondria it is associated with the reduction in fat accumulation and lipid aggregation (Pan et al., 2023). These studies do not inform the fat composition of the diet fed for the mice. Here, results are consistent with the aforementioned findings as *ECH1* is upregulated when SAFA E% increases, suggesting it protects against the negative effects that high-fat diet has on metabolism which otherwise may lead to atherosclerosis.

Cytochrome c oxidase subunit 6B1 (encoded by *COX6B1* gene) is the last enzyme in the mitochondrial electron transport chain and protects cardiomyocytes from ischemia by reducing ROS production and cell apoptosis in rats cardiomyocytes based on hypoxia/reoxygenation injury which is used as research model for ischemia (Zhang et al.,

2019). As presented in the results *COX6B1* gene is upregulated when SAFA E% increases thus gene expression accelerates when the risk of ischemia arises.

S100 calcium binding protein A10 (encoded by *S100A10* gene) is ion channel related protein which is highly expressed in neutrophils and its involvement in blood pressure regulation has been discussed thus it has been said that further research is needed (Huan et al., 2015).

Study from Fu et al. (2022) focused to reveal cardioprotective effects of RUS protein and their results showed that RUS increases expression of Branched chain amino acid transaminase 2 (Bcat2, encoded by *Bcat* gene). Bcat2 was suggested as one of the main functional proteins protecting against myocardial ischemia in mice possibly through Keap1/Nrf2/HO-1 signaling pathway which is central inhibitory pathway in the prevention of cardiovascular diseases (Fu et al., 2022). Here, results show *BCAT2* to be upregulated when SAFA E% increases, which may indicate how the body tries to defend against ischemia when saturated fat intake is increased.

Cofilin 1 (encoded by *CFL1* gene) is actin modulating protein thus has important role in cell migration which in turn contributes to pathologies for example atherosclerosis (San Martín et al., 2008). San Martín et al. (2008) studied platelet-derived growth factor induced migration of human aortic smooth muscle cells and unravel the importance of cofilin as migration regulator. In the results, presented in this thesis, cofilin 1 is upregulated and as known high SAFA intake promotes atherosclerosis thus result is consistent with the fact that cell migration favors atherosclerosis.

Calpain small subunit 1 coding gene, *CAPNS1*, expression elevation was observed in liver whereas expression levels remained unchanged in adipose and skeletal muscle tissue in mice fed with a high-fat diet compared to mice fed with a low-fat diet (Akasu et al., 2022). Even though this study did not offer information about how diet affects the *CAPNS1* expression in blood, it supports the results presented in this thesis because differential gene expression analysis shows that increase of SAFA E% upregulates *CAPNS1* expression. Other study has presented evidence of calpains' role in diabetic cardiomyopathy in mice as they associated calpain with coronary circulation dysfunction through deleting *CAPNS1* which resulted increased coronary flow reserve (Teng et al., 2019).

Downregulation of *NDUFA13* gene in mice has been presented to offer higher tolerance to ischemia-reperfusion injury (Hu et al., 2017). Based on the results of this thesis where *NDUFA13* is upregulated when SAFA E% increases and the results presented by Hu et al. increase in SAFA E% would mean worse tolerance towards to ischemia-reperfusion injury.

Gene *HSPB1* encodes protein HSP27 which has been seen to decrease in human atherosclerotic arteries during the plaque progression (Batulan et al., 2016). However, the results of this thesis suggest that *HSPB1* is expressed more when SAFA E% increases thus are not consistent with the presented from Batulan et al. (2016).

Early atherosclerosis studied in rabbit models presented an increase in Pyruvate kinase M1/2 (encoded by *PKM* gene) in the atherosclerotic aorta, as the results from plasma were opposite, as *PKM* was downregulated, while further analysis of human plasma gave consistent results with the analysis of rabbit aortic tissue (Martin-Lorenzo et al., 2016). In the results of this thesis *PKM* is upregulated thus consistent with the previous results got from the aortic tissue of rabbit and human plasma as SAFA E% increase promotes atherosclerosis.

Profilin 1 (encoded by *PFN1* gene) expression has been previously studied in human atherosclerosis, and preclinical and clinical results suggest that *PFN1* expression is increased in human atherosclerotic plaques compared to normal vessel wall close to the plaque, thus profilin 1 dysregulation possibly has an effect on progression of atherosclerosis (Allen et al., 2021). This may indicate that STRIP participants with high SAFA E% have potential for plaque developed in their blood vessels as *PFN1* is upregulated when SAFA E% increases.

Transient receptor potential channel M2 (encoded by *TRPM2* gene) has been associated to atherosclerotic progression through several cellular processes and recently a research group showed in knockout mice that the lack of *TRPM2* reduced atherosclerotic lesions and that TRPM2 promotes hypercholesterolemia-induced atherosclerosis (Zhang et al., 2022). Zhang et al. (2022) suggests that TRPM2 could possibly promote ROS production while previously described cytochrome c oxidase subunit 6B1 reduces ROS production and controls inflammation and atherosclerotic progression whereas TRPM2 stimulates

inflammation and aggravates atherosclerosis. Upregulation of *TRPM2* in the results while SAFA E% increases is consistent with previous results from Zhang et al. (2022).

In the beginning of this chapter alanyl aminopeptidase, membrane protein (encoded by *ANPEP* gene) is described as downregulated in the class variable analysis, but analysis conducted with continuous variable suggests *ANPEP* as upregulated gene. As mentioned, Devarakonda et al. (2019) showed that ANPEP deficiency in mice promotes atherosclerosis. Based on the results from continuous variable analysis this would mean that those who had smaller SAFA E% would have faster atherosclerosis progress than those who had greater SAFA E%. These results for *ANPEP* gene are contradictory, and it cannot be established whether one of the results is reliable.

Overall, many genes that are differentially expressed in the results have connection to CVDs or factors that are known to increase the risk of CVDs. Also, almost all genes from the results that were associated to CVDs were consistent to findings from previous studies. This supports the idea that higher SAFA E% has a negative effect on hearth health.

4.2 Pathway enrichment analysis with GOseq

Pathway enrichment analysis from up- and downregulated DE gene lists was conducted with GOseq. Pathway enrichment was analyzed for pathways belonging to the Hallmark pathway, KEGG sub-collection and ontology collection. Results show significant enrichment to two pathways from ontology collection, one downregulated and one upregulated.

In the results, GOBP_MEMBARE_BIOGENESIS pathway is upregulated as five upregulated genes from DGE results overlapped with the 61 genes that comprise the pathway. Overlapped genes are *CHMP2A*, *ANXA2*, *S100A10*, *STX4* and *BANF1*. Membrane biogenesis is cellular process of membrane formation which covers molecule synthesis and arrangement. As the cell membrane is a dynamic component that regulates cell functions, it is believed today that disturbing these processes are associated with atherosclerosis (Kotlyarov and Kotlyarova, 2022). Based on literature seems that

CHMP2A and BANF1 are not connected to CVDs, but perhaps these connection to CVDs should be studied in the future.

In previous chapter *S100A10* was connected to CVDs through its possible role in blood pressure regulation as high blood pressure and upregulation of *S100A10* appeared together (Huan et al., 2015). The introduction of this thesis describes PCSK9 protein role on degrading LDL receptors thus it increases serums cholesterol levels. Annexin A2 protein (encoded by *ANXA2* gene) with S100A10 protein forms a complex that binds to CHRD, which normally offers structural integrity to PCSK9 molecule, but AnxA2·S100A10 complex bonding to CHRD causes allosteric structural change in the catalytic part of PCSK9 protein (Mayer et al., 2008). Disruption of PCSK9 function reduces LDL receptor degradation thus causing serum cholesterol level decrease. Because SAFA tends to raise blood cholesterol, upregulation of *ANXA2* and *S100A10* when the SAFA E% increases could be a normal defense mechanism of the human body against the development of atherosclerosis.

Syntaxin 4, encoded by *STX4*, belongs to SNARE protein family and is required for normal vertebrate cardiac conduction and vesicular transport (Perl et al., 2022). Perl et al. (2022) showed in vivo that SNAREs are needed for normal embryonic cardiac function and suggested that syntaxin 4 variants are associated with cardiomyopathy among several other human diseases. The literature lacks knowledge on how different gene expression levels of a normal STX4 variant affect heart health and CVDs.

Pathway GOCC_IMMUNOGLOBULIN_COMPLEX is in the results downregulated as four downregulated DEGs overlapped with the 147 genes that comprise the pathway. Overlapped genes are *IGHV3-33*, *IGLV1-47*, *IGKV3D-20* and *IGKV4-1*. Immunoglobulins are formed after six gene types of which four of them, variable, diversity, joining and constant, identify the immunoglobulins. All four genes shown in the results are variable genes and *IGHV3-33* encodes heavy chain variable while the rest encode light chain variable. It has been established that B-cell antigen receptor and T-cell receptor genes interfere with single-cell RNA sequence analyses and subsequent analyses thus should be excluded from DGE analyses (Sundell et al., 2022). Because immunoglobulins are part of B-cell and T-cell receptors, excluding immunoglobulins form the analyses should be considered. However, the literature does not mention the exclusion with the whole blood transcriptome used in this thesis. In literature these

immunoglobulin genes are not connected to CVDs, but since inflammatory cells participate in the formation of plaque in the blood vessels, the connection should possibly be studied in the future.

As mentioned above, the results have a connection with the PCSK9 pathway, which was suggested in the introduction to have an effect in the formation of atherosclerosis. The other pathways presented in the introduction have no connection to CVDs based on the results from GOseq analysis.

4.3 Pathway enrichment analysis with edgeR

Pathway enrichment analysis from read counts of 25562 genes was conducted with edgeR. Pathway enrichment was analyzed for pathways belonging to the Hallmark pathway, KEGG sub-collection and ontology collection. Results show that one pathway, HALLMARK_SPERMATOGENESIS, from Hallmark gene set is significantly downregulated. Results from other analyses do not support significant up- or downregulation to any of their pathways. Genes from this pathway are upregulated during spermatogenesis but because edgeR analysis was conducted by using different input data formation than GOseq, results lack the information about genes that cause the pathway enrichment.

Fatty acid composition in sperm membrane seems to have a connection to the sperm quality and SAFAs have negative correlation to sperm viability and positive correlation to sperm apoptosis (Collodel et al., 2020). However, this study does not show an association between the sperm membrane fatty acid composition and dietary fat composition. In another study, male rats fed with high-fat diet showed decrease in sperm quality (Luo et al., 2020). In addition, it has been shown in male rats, that dietary fatty acids can also affect positively to the sperm quality and function as PUFA and MUFA supplements counteracted the negative effects high-fat diet has on sperm cells (Ferramosca et al., 2017). Similar results have been obtained in Denmark, where the increased intake of SAFA among young men had an association with lower sperm concentration and total sperm count (Jensen et al., 2013). The result from edgeR analysis of this thesis is consistent with previous findings thus supports the belief that high SAFA E% has an effect not only on hearth health but also on reproductive health.

Even though gender was used as covariate, as both male and female participants are included in the analysis, there is possibility that the result is influenced by the inclusion of both genders as most of the genes in this pathway are located on autosomes and none in Y chromosome. The association between dietary fat and spermatogenesis could be studied in the future using only males in the analysis.

Even though there were no specific expectations or hypothesis about pathways that should come up in the analysis, except of the connection between inflammation and atherosclerosis as described in the introduction. The results from edgeR analyses in this thesis do not support these connections described in previous studies.

4.4 Comparison of pathway enrichment analysis methods

One of the aims was to compare input data forms impact on results by using two methods for pathway enrichment analysis. List of DEGs was used as input data for GOseq analysis and read counts of 25562 genes as input data for edgeR analysis. The results got from analyses conducted with different input data differ from each other thus can be stated that the format of the data affects the results obtained.

Methods have been compared previously and as Love (2016) have propound GOseq and edgeR methods usually give same results from gene-level analyses. The results from the pathway enrichment analyses were inconsistent but it cannot be determined whether the difference is due to the format of the input data or whether the choice of method affects the results. To compare more reliable, the effect of input data format edgeR could also have been used to a two-step analysis where firs the DEG list is unravel and then the pathway analysis conducted. That way edgeR would be used with two different format of input data and GOseq would be excluded from the analyses.

Other possibility could have been to execute additional differential gene expression analysis with edgeR to see if GOseq and edgeR methods would have given same results at the gene-level analyses as Love (2016) presents. Similar results from this additional edgeR analysis compared to GOseq differential gene expression analysis would have excluded the possibility that the choice of method affects the results. Based on analyses

included in this thesis can only be suggested that differences in the pathway analysis results may be due to the fact that when using DEG list as input data there is a bigger probability to errors. Every analysis can result false positives or false negatives thus conducting one analyze instead two analyses likely lowers the errors in results. Roughly can be said that here based on this assumption results from edgeR pathway enrichment analysis are more reliable than results from GOseq analysis.

4.5 General comments and sources of error

The sample size of this thesis reduced from 495 participants to 425 participants because SAFA E% data was missing form 70 STRIP participants. SAFA E% data was missing from those that did not return the food diary at the same study visit the blood sample was collected for the transcriptome analysis. Thus, true N was 14 % smaller than what was initially planned. Nevertheless, the sample size is large enough.

The mRNA data, used in the transcriptome analyses, is conducted from blood cells and because participants are young humans collecting other tissue samples would have been unethical. The inclusion of different tissues in the study would offer an opportunity to compare effects SAFA intake has on transcriptome in different tissues. Especially adipose tissue and blood vessel endothelium would be interesting to add to the analyses. This could be studied in model organisms such as mice. Cell population on blood is very heterogeneity which also has an impact on transcriptome analysis. If blood cell composition would be known the source of error could have been solved by adding the blood cell ratio in the analysis as covariate.

It is possible that intervention effect could be stronger between intervention and control groups, if control group would have been recruited outside of this longitudinal study because participants, regardless of which group they belong to in the study, may have interest on nutrition and healthy lifestyle. Also, in human dietary studies there is always possibility that participants do not provide reliable information in their food diary or improves diet during the week when the diary is filled.

Analyses in this thesis are based on alignment made against reference genome GRCh38. Today there is newer and more accurate reference T2T genome available. Upcoming

analyses should be based on newest version of GRCh38 genome or even based on T2T-CHM13 however, this would affect the validity of comparing analyzes performed at different time points.

4.6 Conclusions

In conclusion, the results of this thesis suggest that saturated fatty acid intake affects gene expression in whole blood both at the individual gene expression level and at the pathway level. Some of the results found from differential gene expression analyses supported the previous studies but the results also included genes that had no prior association to heart health thus further studies are needed to establish if they have role in heart health.

Pathway analyses resulted significant enrichment to membrane biogenesis, immunoglobulin complex and spermatogenesis pathways. These results supported partially previous studies. Inconsistencies observed between results and literature may be due to different research settings. As in the individual gene level, there is also a need for further studies around these pathways and their effects on the heart health.

Pathway enrichment analysis was conducted with two different methods using different input data forms. Results from GOseq and edgeR pathway enrichment analyses were inconsistent which could indicate that input data form influences results but could not be excluded that the difference is due only to this.

5 Acknowledgment

I would like to thank my supervisors Juha Mykkänen, Katja Pahkala and Christina Nokkala for the help and guidance I received during this thesis project. Special thanks to Juha for answering all my questions. I would like to show my appreciation to Olli Raitakari for offering me the thesis project. I would also like to thank everyone who kindly welcomed me into the research group.

Huge thanks to my friends and my family who have supported me during my studies, especially towards the end. In particular, I would like to mention Meri for offering irreplaceable peer support during countless hours in library and Jussi for making sure I remember to eat and enjoy free time alongside the project.

References

Adelman K. & Lis J.T. (2012). Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nature Reviews Genetics, 13*(10), 720. https://doi.org/10.1038/nrg3293

Akasu R., Miyazaki T., Elhussiny M.Z., Sugiura Y., Tomitsuka Y., Haraguchi S., Otsu K., Chowdhury V.S. & Miyazaki A. (2022). Calpain-mediated proteolytic production of free amino acids in vascular endothelial cells augments obesity-induced hepatic steatosis. *The Journal of Biological Chemistry, 298*(6), 101953. https://doi.org/10.1016/j.jbc.2022.101953

Allen A., Gau D. & Roy P. (2021). The role of profilin-1 in cardiovascular diseases. *Journal of Cell Science, 134*(9), jcs249060. https://doi.org/10.1242/jcs.249060

Anders S. & Huber W. (2010). Differential expression analysis for sequence count data. *Genome Biology, 11*(10), R106. https://doi.org/10.1186/gb-2010-11-10-r106

Andrews S. (2023). FastQC. A Quality Control tool for High Throughput Sequence Data [Referred 17.05.2023]. Available: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Baker D.H. (2008). Animal Models in Nutrition Research. *The Journal of Nutrition, 138*(2), 391–396. https://doi.org/10.1093/jn/138.2.391

Batulan Z., Pulakazhi Venu V.K., Li Y., Koumbadinga G., Alvarez-Olmedo D.G., Shi C. & O'Brien E.R. (2016). Extracellular Release and Signaling by Heat Shock Protein 27: Role in Modifying Vascular Inflammation. *Frontiers in Immunology, 7*, 285. https://doi.org/10.3389/fimmu.2016.00285

Benjamini Y. & Hochberg Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B : Statistical Methodology, 57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Berná G., Oliveras-López M.J., Jurado-Ruíz E., Tejedo J., Bedoya F., Soria B. & Martín, F. (2014). Nutrigenetics and nutrigenomics insights into diabetes etiopathogenesis. *Nutrients, 6*(11), 5338–5369. https://doi.org/10.3390/nu6115338

Bystrykh L. (2021). Python for gene expression. *F1000Research, 10*. https://doi.org/10.12688/f1000research.53842.1

Carulli J.P., Artinger M., Swain P.M., Root C.D., Chee L., Tulig C., Guerin J., Osborne M., Stein G., Lian J. & Lomedico P.T. (1998). High throughput analysis of differential gene expression. *Journal of Cellular Biochemistry, 30–31*, 286–296. https://doi.org/10.1002/(SICI)1097-4644(1998)72:30/31+<286::AID-JCB35>3.0.CO;2-D

Chaussabel D. (2015). Assessment of immune status using blood transcriptomics and potential implications for global health. *Seminars in Immunology, Global transcriptional regulation in the immune system, 27*(1), 58–66. https://doi.org/10.1016/j.smim.2015.03.002

Chen M. & Manley J.L. (2009). Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nature reviews. Molecular cell biology, 10*(11), 741–754. https://doi.org/10.1038/nrm2777

Chen Y., Lun A.T.L. & Smyth G.K. (2016). From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Research, 5*, 1438. https://doi.org/10.12688/f1000research.8987.2

Coll R.C., Schroder K. & Pelegrín P. (2022). NLRP3 and pyroptosis blockers for treating inflammatory diseases. *Trends in Pharmacological Sciences, 43*(8), 653–668. https://doi.org/10.1016/j.tips.2022.04.003

Collodel G., Moretti E., Noto D., Iacoponi F. & Signorini C. (2020). Fatty Acid Profile and Metabolism Are Related to Human Sperm Parameters and Are Relevant in Idiopathic Infertility and Varicocele. *Mediators of Inflammation, 2020*, 3640450. https://doi.org/10.1155/2020/3640450

Conesa A., Madrigal P., Tarazona S., Gomez-Cabrero D., Cervera A., McPherson A., Szcześniak M.W., Gaffney D.J., Elo L.L., Zhang X. & Mortazavi A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology, 17*(1), 13. https://doi.org/10.1186/s13059-016-0881-8

Cooper G.M. (2000). The Cell: A Molecular Approach, 2nd ed. *Sinauer Associates, Sunderland (MA)*. https://www.ncbi.nlm.nih.gov/books/NBK9904/

Costa-Silva J., Domingues D. & Lopes F.M. (2017). RNA-Seq differential expression analysis: An extended review and a software tool. *PLoS One, 12*(12), e0190152. https://doi.org/10.1371/journal.pone.0190152

Decroly E., Debarnot C., Ferron F., Bouvet M., Coutard B., Imbert I., Gluais L., Papageorgiou N., Sharff A., Bricogne G., Ortiz-Lombardia M., Lescar J. & Canard, B. (2011). Crystal Structure and Functional Analysis of the SARS-Coronavirus RNA Cap 2′-O-Methyltransferase nsp10/nsp16 Complex. *PLoS Pathogens, 7*(5), e1002059. https://doi.org/10.1371/journal.ppat.1002059

Devarakonda C.V., Pereira F.E., Smith J.D., Shapiro L.H. & Ghosh M. (2019). CD13 deficiency leads to increased oxidative stress and larger atherosclerotic lesions. *Atherosclerosis,*

287, 70–80. https://doi.org/10.1016/j.atherosclerosis.2019.06.901

Ewels P., Magnusson M., Lundin S. & Käller M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics, 32*(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Farhud D., Zarif Yeganeh M. & Zarif Yeganeh M. (2010). Nutrigenomics and Nutrigenetics. *Iranian Journal of Public Health, 39*(4), 1–14.

Ferramosca A., Moscatelli N., Di Giacomo M. & Zara V. (2017). Dietary fatty acids influence sperm quality and function. *Andrology, 5*(3), 423–430. https://doi.org/10.1111/andr.12348

Freedman A.H., Gaspar J.M. & Sackton T.B. (2020). Short paired-end reads trump long single-end reads for expression analysis. *BMC Bioinformatics, 21*, 149. https://doi.org/10.1186/s12859-020-3484-z

Fu F., Lai Q., Hu J., Zhang L., Zhu X., Kou J., Yu B. & Li F. (2022). Ruscogenin Alleviates Myocardial Ischemia-Induced Ferroptosis through the Activation of BCAT1/BCAT2. *Antioxidants, 11*(3), 583. https://doi.org/10.3390/antiox11030583

Gao Y., Zhao D., Xie W., Meng T., Xu C., Liu Y., Zhang P., Bi X. & Zhao Z. (2021). Rap1GAP Mediates Angiotensin II-Induced Cardiomyocyte Hypertrophy by Inhibiting Autophagy and Increasing Oxidative Stress. *Oxidative Medicine and Cellular Longevity, 2021*, 7848027. https://doi.org/10.1155/2021/7848027

Gentleman R.C., Carey V.J., Bates D.M., Bolstad B., Dettling M., Dudoit S., Ellis B., Gautier L., Ge Y., Gentry J., Hornik K., Hothorn T., Huber W., Iacus S., Irizarry R., Leisch F., Li C., Maechler M., Rossini A.J., Sawitzki G., Smith C., Smyth G., Tierney L., Yang J.Y. & Zhang J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology, 5*(10), R80. https://doi.org/10.1186/gb-2004-5-10-r80

Giunzioni I., Tavori H., Covarrubias R., Major A.S., Ding L., Zhang Y., DeVay R.M., Hong L., Fan D., Predazzi I.M., Rashid S., Linton M.F. & Fazio S. (2016). Local Effects of Human PCSK9 on the Atherosclerotic Lesion. *The Journal of pathology, 238*(1), 52–62. https://doi.org/10.1002/path.4630

Grünberg S. & Hahn S. (2013). Structural insights into transcription initiation by RNA polymerase II. *Trends in biochemical sciences, 38*(12). https://doi.org/10.1016/j.tibs.2013.09.002

Hong G., Zhang W., Li H., Shen X. & Guo Z. (2014). Separate enrichment analysis of pathways

for up- and downregulated genes. *Journal of The Royal Society Interface, 11*(92), 20130950. https://doi.org/10.1098/rsif.2013.0950

Hu H., Nan J., Sun Y., Zhu D., Xiao C., Wang Y., Zhu L., Wu Y., Zhao J., Wu R., Chen J., Yu H., Hu X., Zhu W. & Wang J. (2017). Electron leak from NDUFA13 within mitochondrial complex I attenuates ischemia-reperfusion injury via dimerized STAT3. *Proceedings of the National Academy of Sciences of the United States of America, 114*(45), 11908–11913. https://doi.org/10.1073/pnas.1704723114

Huan T., Esko T., Peters M.J., Pilling L.C., Schramm K., Schurmann C., Chen B.H., Liu C., Joehanes R., Johnson A.D., Yao C., Ying S., Courchesne P., Milani L., Raghavachari N., Wang R., Liu P., Reinmaa E., Dehghan A., Hofman A., Uitterlinden A.G., Hernandez D.G., Bandinelli S., Singleton A., Melzer D., Metspalu A., Carstensen M., Grallert H., Herder C., Meitinger T., Peters A., Roden M., Waldenberger M., Dörr M., Felix S.B., Zeller T., Vasan R., O'Donnell C.J., Munson P.J., Yang X., Prokisch H., Völker U., van Meurs J.B.J., Ferrucci L. & Levy D. (2015). A Meta-analysis of Gene Expression Signatures of Blood Pressure and Hypertension. *PLoS Genetics, 11*(3), e1005035. https://doi.org/10.1371/journal.pgen.1005035

Huang C.-.C, Liu K., Pope R.M., Du P., Lin S., Rajamannan N.M., Huang Q.-Q., Jafari N., Burke G.L., Post W., E.Watson K., Johnson C., Daviglus M.L. & Lloyd-Jones D.M. (2011). Activated TLR Signaling in Atherosclerosis among Women with Lower Framingham Risk Score: The Multi-Ethnic Study of Atherosclerosis. *PLoS One, 6*(6), e21067. https://doi.org/10.1371/journal.pone.0021067

Illumina. (2022). Paired-End vs. Single-Read Sequencing Technology [Referred 16.11.2022]. Available: https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/paired-end-vs-single-read.html

Jensen T.K., Heitmann B.L., Jensen M.B., Halldorsson T.I., Andersson A.-M., Skakkebæk N.E., Joensen U.N., Lauritsen M.P., Christiansen P., Dalgård C., Lassen T.H. & Jørgensen N. (2013). High dietary intake of saturated fat is associated with reduced semen quality among 701 young Danish men from the general population. *The American Journal of Clinical Nutrition, 97*(2), 411–418. https://doi.org/10.3945/ajcn.112.042432

Kachaev Z.M., Lebedeva L.A., Kozlov E.N. & Shidlovskii Y.V. (2020). Interplay of mRNA capping and transcription machineries. *Bioscience Reports, 40*(1), BSR20192825. https://doi.org/10.1042/BSR20192825

Kanehisa M. & Goto S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research, 28*(1), 27–30. https://doi.org/10.1093/nar/28.1.27

Keskitalo A., Munukka E., Aatsinki A., Saleem W., Kartiosuo N., Lahti L., Huovinen P., Elo L.L., Pietilä S., Rovio S.P., Niinikoski H., Viikari J., Rönnemaa T., Lagström H., Jula A., Raitakari O. & Pahkala K. (2022). An Infancy-Onset 20-Year Dietary Counselling Intervention and Gut Microbiota Composition in Adulthood. *Nutrient,s 14*(13), 2667. https://doi.org/10.3390/nu14132667

Khan S.R., Dalm V.A.S.H., Ikram M.K., Peeters R.P., van Hagen P.M., Kavousi M. & Chaker L. (2023). The Association of Serum Immunoglobulins with Risk of Cardiovascular Disease and Mortality: the Rotterdam Study. *Journal of Clinical Immunology, 43*(4), 769–779. https://doi.org/10.1007/s10875-023-01433-7

Kitajka K., Sinclair A.J., Weisinger R.S., Weisinger H.S., Mathai M., Jayasooriya A.P., Halver J.E. & Puskás L.G. (2004). Effects of dietary omega-3 polyunsaturated fatty acids on brain gene expression. *Proceedings of the National Academy of Sciences of the United States of America, 101*(30), 10931–10936. https://doi.org/10.1073/pnas.0402342101

Kong P., Cui Z.-Y., Huang X.-F., Zhang D.-D., Guo R.-J. & Han M. (2022). Inflammation and atherosclerosis: signaling pathways and therapeutic intervention. *Signal Transduction and Targeted Therapy 7*,131. https://doi.org/10.1038/s41392-022-00955-7

Kotlyarov S. & Kotlyarova A. (2022). The Importance of the Plasma Membrane in Atherogenesis. *Membranes 12*(11), 1036. https://doi.org/10.3390/membranes12111036

Koushki K., Shahbaz S.K., Mashayekhi K., Sadeghi M., Zayeri Z.D., Taba M.Y., Banach M., Al-Rasadi K., Johnston T.P. & Sahebkar A. (2021). Anti-inflammatory Action of Statins in Cardiovascular Disease: the Role of Inflammasome and Toll-Like Receptor Pathways. *Clinical Reviews in Allergy & Immunology, 60*(2), 175–199. https://doi.org/10.1007/s12016-020-08791-9

Laitinen T.T., Nuotio J., Juonala M., Niinikoski H., Rovio S., Viikari J.S.A., Rönnemaa T., Magnussen C.G., Jokinen E., Lagström H., Jula A., Simell O., Raitakari O.T. & Pahkala K. (2018). Success in Achieving the Targets of the 20-Year Infancy-Onset Dietary Intervention: Association With Insulin Sensitivity and Serum Lipids. *Diabetes Care 41*(10), 2236–2244. https://doi.org/10.2337/dc18-0869

Laitinen T.T., Nuotio J., Niinikoski H., Juonala M., Rovio S.P., Viikari J.S.A., Rönnemaa T., Magnussen C.G., Sabin M., Burgner D., Jokinen E., Lagström H., Jula A., Simell O., Raitakari O.T. & Pahkala K. (2020a). Attainment of Targets of the 20-Year Infancy-Onset Dietary Intervention and Blood Pressure Across Childhood and Young Adulthood: The Special Turku Coronary Risk Factor Intervention Project (STRIP). *Hypertension (Dallas, Tex 1979), 76*(5),

1572–1579. https://doi.org/10.1161/HYPERTENSIONAHA.120.15075

Laitinen T.T., Nuotio J., Rovio S.P., Niinikoski H., Juonala M., Magnussen C.G., Jokinen E., Lagström H., Jula A., Viikari J.S.A., Rönnemaa T., Simell O., Raitakari O.T. & Pahkala K. (2020b). Dietary Fats and Atherosclerosis From Childhood to Adulthood. *Pediatrics, 145*(4), e20192786. https://doi.org/10.1542/peds.2019-2786

Laki lääketieteellisestä tutkimuksesta. 1999/488. [Referred 17.05.2023]. Available: https://finlex.fi/fi/laki/ajantasa/1999/19990488#L2

Larsen S.V., Holven K.B., Ottestad I., Dagsland K.N., Myhrstad M.C.W. & Ulven S.M. (2018). Plasma fatty acid levels and gene expression related to lipid metabolism in peripheral blood mononuclear cells: a cross-sectional study in healthy subjects. *Genes & Nutrition, 13*(9). https://doi.org/10.1186/s12263-018-0600-z

Law C.W., Alhamdoosh M., Su S., Dong X., Tian L., Smyth G.K. & Ritchie M.E. (2016). RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. *F1000Research, 5*. https://doi.org/10.12688/f1000research.9005.3

Law C.W., Chen Y., Shi W. & Smyth G.K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology, 15*(2), R29. https://doi.org/10.1186/gb-2014-15-2-r29

Lehtovirta M., Matthews L.A., Laitinen T.T., Nuotio J., Niinikoski H., Rovio S.P., Lagström H., Viikari J.S.A., Rönnemaa T., Jula A., Ala-Korpela M., Raitakari O.T. & Pahkala K. (2021). Achievement of the Targets of the 20-Year Infancy-Onset Dietary Intervention-Association with Metabolic Profile from Childhood to Adulthood. *Nutrients, 13*(2), 533. https://doi.org/10.3390/nu13020533

Leng N., Dawson J.A., Thomson J.A., Ruotti V., Rissman A.I., Smits B.M.G., Haag J.D., Gould M.N., Stewart R.M. & Kendziorski C. (2013). EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics (Oxford, England), 29*(8), 1035–1043. https://doi.org/10.1093/bioinformatics/btt087

Li J. & Tibshirani R. (2013). Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. *Statistical Methods in Medical Research, 22*(5), 519–536. https://doi.org/10.1177/0962280211428386

Liao Y., Smyth G.K. & Shi W. (2019). The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Research, 47*(8), e47. https://doi.org/10.1093/nar/gkz114

Liao Y., Smyth G.K. & Shi W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics, 30*(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656

Liberzon A., Birger C., Thorvaldsdóttir H., Ghandi M., Mesirov J.P. & Tamayo P. (2015). The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Systems, 1*(6), 417–425. https://doi.org/10.1016/j.cels.2015.12.004

Liberzon A., Subramanian A., Pinchback R., Thorvaldsdóttir H., Tamayo P. & Mesirov J.P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics, 27*(12), 1739–1740. https://doi.org/10.1093/bioinformatics/btr260

Lister R., O'Malley R.C., Tonti-Filippini J., Gregory B.D., Berry C.C., Millar A.H. & Ecker J.R. (2008). Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. *Cell, 133*(3), 523–536. https://doi.org/10.1016/j.cell.2008.03.029

Lopes-Ramos C.M., Chen C.-Y., Kuijjer M.L., Paulson J.N., Sonawane A.R., Fagny M., Platig J., Glass K., Quackenbush J. & DeMeo D.L. (2020). Sex Differences in Gene Expression and Regulatory Networks across 29 Human Tissues. *Cell reports, 31*(12), 107795. https://doi.org/10.1016/j.celrep.2020.107795

Love M. (2016). DESeq2 or edgeR. [Referred 6.3.2023]. Available: https://mikelove.wordpress.com/2016/09/28/deseq2-or-edger/ (accessed 3.6.23).

Love M.I., Huber W. & Anders S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology, 15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Luo D., Zhang M., Su X., Liu L., Zhou X., Zhang X., Zheng D., Yu C. & Guan Q. (2020). High fat diet impairs spermatogenesis by regulating glucose and lipid metabolism in Sertoli cells. *Life Sciences, 257*, 118028. https://doi.org/10.1016/j.lfs.2020.118028

Madsen M.B., Birck M.M., Fredholm M. & Cirera S. (2009). Expression Studies of the Obesity Candidate Gene FTO in Pig. *Animal Biotechnology, 21*(1), 51–63. https://doi.org/10.1080/10495390903381792

Mao X., Huang D., Rao C., Du M., Liang M., Li F., Liu B. & Huang K. (2020). Enoyl coenzyme A hydratase 1 combats obesity and related metabolic disorders by promoting adipose tissue browning. *American Journal of Physiology-Endocrinology and Metabolism, 318*(3), E318–E329. https://doi.org/10.1152/ajpendo.00424.2019

Marguerat S. & Bähler J. (2010). RNA-seq: from technology to biology. *Cellular and Molecular Life Sciences, 67*(4), 569–579. https://doi.org/10.1007/s00018-009-0180-6

Martin-Lorenzo M., Gonzalez-Calero L., Maroto A.S., Martinez P.J., Zubiri I., de la Cuesta F., Mourino-Alvarez L., Barderas M.G., Heredero A., Aldamiz-Echevarría G., Vivanco F. & Alvarez-Llamas G. (2016). Cytoskeleton deregulation and impairment in amino acids and energy metabolism in early atherosclerosis at aortic tissue with reflection in plasma. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease, 1862*(4), 725–732. https://doi.org/10.1016/j.bbadis.2015.12.006

Matthews L.A., Rovio S.P., Jaakkola J.M., Niinikoski H., Lagström H., Jula A., Viikari J.S.A., Rönnemaa T., Simell O., Raitakari O.T. & Pahkala K. (2019). Longitudinal effect of 20-year infancy-onset dietary intervention on food consumption and nutrient intake: the randomized controlled STRIP study. *European Journal of Clinical Nutrition, 73*(6), 937–949. https://doi.org/10.1038/s41430-018-0350-4

Mayer G., Poirier S. & Seidah N.G. (2008). Annexin A2 Is a C-terminal PCSK9-binding Protein That Regulates Endogenous Low Density Lipoprotein Receptor Levels. *Journal of Biological Chemistry, 283*(46), 31791–31801. https://doi.org/10.1074/jbc.M805971200

Meyer A., Paroni F., Günther K., Dharmadhikari G., Ahrens W., Kelm S. & Maedler K. (2016). Evaluation of Existing Methods for Human Blood mRNA Isolation and Analysis for Large Studies. *PLoS One, 11*(8), e0161778. https://doi.org/10.1371/journal.pone.0161778

Mierziak J., Kostyn K., Boba A., Czemplik M., Kulma A. & Wojtasik W. (2021). Influence of the Bioactive Diet Components on the Gene Expression Regulation. *Nutrients, 13*(11), 3673. https://doi.org/10.3390/nu13113673

Monaco C., Gregan S.M., Navin T.J., Foxwell B.M.J., Davies A.H. & Feldmann M. (2009). Toll-Like Receptor-2 Mediates Inflammation and Matrix Degradation in Human Atherosclerosis. *Circulation, 120*(24), 2462–2469. https://doi.org/10.1161/CIRCULATIONAHA.109.851881

Mozaffarian D., Appel L.J. & Van Horn L. (2011). Components of a Cardioprotective Diet. *Circulation, 123*(24), 2870–2891. https://doi.org/10.1161/CIRCULATIONAHA.110.968735

Mustajoki P. (2020). Valtimotauti (ateroskleroosi). [Referred 16.11.2022]. Available: https://www.terveyskirjasto.fi/dlk00095

National Human Genome Research Institute. (2020). Biological Pathways Fact Sheet [Referred 06.03.2023]. Available: https://www.genome.gov/about-genomics/fact-sheets/Biological-

Pathways-Fact-Sheet

National Library of Medicine. (2022). Genome assembly GRCh38.p14. [Referred 17.11.2022]. Available: https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCF_000001405.40/

National Library of Medicine. (2022). T2T-CHM13v2.0. [Referred 08.03.2023]. Available: https://www.ncbi.nlm.nih.gov/assembly/GCA_009914755.4

Nature Education. (2014). poly-A tail. [Referred 16.11.2022]. Available: https://www.nature.com/scitable/definition/poly-a-tail-276/

Nordic Council of Ministers. (2012). Nordic Nutrition Recommendations 2012: Integrating nutrition and physical activity. [Referred 16.11.2022]. Available: https://www.ruokavirasto.fi/globalassets/teemat/terveytta-edistava-ruokavalio/ravitsemus--ja-ruokasuositukset/nordic-nutrition-recommendations-2012.pdf

Nurk S., Koren S., Rhie A., Rautiainen M., Bzikadze A.V., Mikheenko A., Vollger M.R., Altemose N., Uralsky L., Gershman A., Aganezov S., Hoyt S.J., Diekhans M., Logsdon G.A., Alonge M., Antonarakis S.E., Borchers M., Bouffard G.G., Brooks S.Y., Caldas G.V., Chen N.-C., Cheng H., Chin C.-S., Chow W., de Lima L.G., Dishuck P.C., Durbin R., Dvorkina T., Fiddes I.T., Formenti G., Fulton R.S., Fungtammasan A., Garrison E., Grady P.G.S., Graves-Lindsay T.A., Hall I.M., Hansen N.F., Hartley G.A., Haukness M., Howe K., Hunkapiller M.W., Jain C., Jain M., Jarvis E.D., Kerpedjiev P., Kirsche M., Kolmogorov M., Korlach J., Kremitzki M., Li H., Maduro V.V., Marschall T., McCartney A.M., McDaniel J., Miller D.E., Mullikin J.C., Myers E.W., Olson N.D., Paten B., Peluso P., Pevzner P.A., Porubsky D., Potapova T., Rogaev E.I., Rosenfeld J.A., Salzberg S.L., Schneider V.A., Sedlazeck F.J., Shafin K., Shew C.J., Shumate A., Sims Y., Smit A.F.A., Soto D.C., Sović I., Storer J.M., Streets A., Sullivan B.A., Thibaud-Nissen F., Torrance J., Wagner J., Walenz B.P., Wenger A., Wood J.M.D., Xiao C., Yan S.M., Young A.C., Zarate S., Surti U., McCoy R.C., Dennis M.Y., Alexandrov I.A., Gerton J.L., O'Neill R.J., Timp W., Zook J.M., Schatz M.C., Eichler E.E., Miga K.H. & Phillippy A.M. (2022). The complete sequence of a human genome. *Science, 376*(6588), 44–53. https://doi.org/10.1126/science.abj6987

One pangenome to bind them all. (2022). *Nature Biotechnology, 40*, 1301. https://doi.org/10.1038/s41587-022-01484-y

Pan X., Zhang X., Ban J., Yue L., Ren L., & Chen S. (2023). Effects of High-Fat Diet on Cardiovascular Protein Expression in Mice Based on Proteomics. *Diabetes, Metabolic Syndrome and Obesity, 16*, 873–882. https://doi.org/10.2147/DMSO.S405327

Perl E., Ravisankar P., Beerens M.E., Mulahasanovic L., Smallwood K., Sasso M.B., Wenzel

C., Ryan T.D., Komár M., Bove K.E., MacRae C.A., Weaver K.N., Prada C.E. & Waxman J.S. (2022). Stx4 is required to regulate cardiomyocyte Ca2+ handling during vertebrate cardiac development. *Human Genetics and Genomics Advances, 3*(3), 100115. https://doi.org/10.1016/j.xhgg.2022.100115

Pestova T.V., Kolupaeva V.G., Lomakin I.B., Pilipenko E.V., Shatsky I.N., Agol V.I. & Hellen C.U. (2001). Molecular mechanisms of translation initiation in eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America, 98*(13), 7029–7036. https://doi.org/10.1073/pnas.111145798

Piovesana A. & Senior G. (2018). How Small Is Big: Sample Size and Skewness. *Assessment, 25*(6), 793–800. https://doi.org/10.1177/1073191116669784

Rainen L., Oelmueller U., Jurgensen S., Wyrich R., Ballas C., Schram J., Herdman C., Bankaitis-Davis D., Nicholls N., Trollinger D. & Tryon V. (2002). Stabilization of mRNA expression in whole blood samples. *Clinical Chemistry, 48*(11), 1883–1890.

Raitakari O., Pahkala K. & Magnussen C.G. (2022). Prevention of atherosclerosis from childhood. *Nature Reviews Cardiology, 19*(8), 543–554. https://doi.org/10.1038/s41569-021-00647-9

Ramanathan A., Robb G.B. & Chan S.-H. (2016). mRNA capping: biological functions and applications. *Nucleic Acids Research, 44*(16), 7511–7526. https://doi.org/10.1093/nar/gkw551

Ranganathan P. & Aggarwal R. (2018). Study designs: Part 1 – An overview and classification. *Perspectives in Clinical Research, 9*(4), 184–186. https://doi.org/10.4103/picr.PICR_124_18

Retterstøl K., Svendsen M., Narverud I. & Holven K.B. (2018). Effect of low carbohydrate high fat diet on LDL cholesterol and gene expression in normal-weight, young adults: A randomized controlled study. *Atherosclerosis, 279*, 52–61. https://doi.org/10.1016/j.atherosclerosis.2018.10.013

Robinson M.D., McCarthy D.J. & Smyth G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England), 26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

San Martín A., Lee M.Y., Williams H.C., Mizuno K., Lassègue B. & Griendling K.K. (2008). Dual regulation of cofilin activity by LIM kinase and Slingshot-1L phosphatase controls platelet-derived growth factor-induced migration of human aortic smooth muscle cells. *Circulation Research, 102*(4), 432–438. https://doi.org/10.1161/CIRCRESAHA.107.158923

Schmidt M., Hopp L., Arakelyan A., Kirsten H., Engel C., Wirkner K., Krohn K., Burkhardt R., Thiery J., Loeffler M., Loeffler-Wirth H. & Binder H. (2020). The Human Blood Transcriptome in a Large Population Cohort and Its Relation to Aging and Health. *Frontiers in Big Data, 3*. https://doi.org/10.3389/fdata.2020.548873

Schurch N.J., Schofield P., Gierliński M., Cole C., Sherstnev A., Singh V., Wrobel N., Gharbi K., Simpson G.G., Owen-Hughes T., Blaxter M. & Barton G.J. (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA, 22*(6), 839–851. https://doi.org/10.1261/rna.053959.115

Schwantes-An T., Darlay R., Mathurin P., Masson S., Liangpunsakul S., Mueller S., Aithal G.P., Eyer F., Gleeson D., Thompson A., Muellhaupt B., Stickel F., Soyka M., Goldman D., Liang T., Lumeng L., Pirmohamed M., Nalpas B., Jacquet J., Moirand R., Nahon P., Naveau S., Perney P., Botwin G., Haber P.S., Seitz H.K., Day C.P., Foroud T.M., Daly A.K., Cordell H.J., Whitfield J.B., Morgan T.R. & Seth, D. (2021). Genome-wide Association Study and Meta-analysis on Alcohol-Associated Liver Cirrhosis Identifies Genetic Risk Factors. *Hepatology, 14*(5), 1920–1931. https://doi.org/10.1002/hep.31535

Shatkin A. (1976). Capping of eucaryotic mRNAs. *Cell, 9*(4), 645–653. https://doi.org/10.1016/0092-8674(76)90128-8

Shi W. & Liao Y. (2023). Rsubread/Subread Users Guide. [Referred 17.11.2022]. Available: https://bioconductor.org/packages/release/bioc/vignettes/Rsubread/inst/doc/SubreadUsersGuide.pdf

Simell O., Niinikoski H., Rönnemaa T., Raitakari O.T., Lagström H., Laurinen M., Aromaa M., Hakala P., Jula A., Jokinen E., Välimäki I., Viikari J. & STRIP Study Group. (2009). Cohort Profile: the STRIP Study (Special Turku Coronary Risk Factor Intervention Project), an Infancy-onset Dietary and Life-style Intervention Trial. *International Journal of Epidemiology, 38*(3), 650–655. https://doi.org/10.1093/ije/dyn072

Sorokin M., Ignatev K., Barbara V., Vladimirova U., Muraveva A., Suntsova M., Gaifullin N., Vorotnikov I., Kamashev D., Bondarenko A., Baranova M., Poddubskaya E. & Buzdin A. (2020). Molecular Pathway Activation Markers Are Associated with Efficacy of Trastuzumab Therapy in Metastatic HER2-Positive Breast Cancer Better than Individual Gene Expression Levels. *Biochemistry (Moscow), 85*(7), 758–772. https://doi.org/10.1134/S0006297920070044

Stark R., Grzelak M. & Hadfield J. (2019). RNA sequencing: the teenage years. *Nature Reviews Genetics, 20*(11), 631–656. https://doi.org/10.1038/s41576-019-0150-2

STRIP Study Group. (2008). A Prospective Randomized Trial of Atherosclerosis Prevention in

Childhood – The Strip Project. [Referred 16.11.2022]. Available: https://stripstudy.utu.fi/english.html (accessed 11.16.22).

Sundell T., Grimstad K., Camponeschi A., Tilevik A., Gjertsson I. & Mårtensson I.-L. (2022). Single-cell RNA sequencing analyses: interference by the genes that encode the B-cell and T-cell receptors. *Briefings in Functional Genomics, 22*(3), 263–273. https://doi.org/10.1093/bfgp/elac044

Takahashi M. (2022). NLRP3 inflammasome as a key driver of vascular disease. *Cardiovascular Research, 118*(2), 372–385. https://doi.org/10.1093/cvr/cvab010

Tavori H., Fan D., Blakemore J.L., Yancey P.G., Ding L., Linton M.F. & Fazio S. (2013). Serum PCSK9 and Cell Surface Low-Density Lipoprotein Receptor: Evidence for a Reciprocal Regulation. *Circulation, 127*(24), 2403–2413. https://doi.org/10.1161/CIRCULATIONAHA.113.001592

Teng X., Ji C., Zhong H., Zheng D., Ni R., Hill D.J., Xiong S., Fan G.-C., Greer P.A., Shen Z. & Peng T. (2019). Selective deletion of endothelial cell calpain in mice reduces diabetic cardiomyopathy by improving angiogenesis. *Diabetologia, 62*(5), 860–872. https://doi.org/10.1007/s00125-019-4828-y

Tettelin H., Masignani V., Cieslewicz M.J., Donati C., Medini D., Ward N.L., Angiuoli S.V., Crabtree J., Jones A.L., Durkin A.S., DeBoy R.T., Davidsen T.M., Mora M., Scarselli M., Margarit y Ros I., Peterson J.D., Hauser C.R., Sundaram J.P., Nelson W.C., Madupu R., Brinkac L.M., Dodson R.J., Rosovitz M.J., Sullivan S.A., Daugherty S.C., Haft D.H., Selengut J., Gwinn M.L., Zhou L., Zafar N., Khouri H., Radune D., Dimitrov G., Watkins K., O'Connor K.J.B., Smith S., Utterback T.R., White O., Rubens C.E., Grandi G., Madoff L.C., Kasper D.L., Telford J.L., Wessels M.R., Rappuoli R. & Fraser C.M. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome." *Proceedings of the National Academy of Sciences of the United States of America, 102*(39), 13950–13955. https://doi.org/10.1073/pnas.0506758102

the PREDIMED study investigators. (2013). In vivo transcriptomic profile after a Mediterranean diet in high–cardiovascular risk patients: a randomized controlled trial. *The American Journal of Clinical Nutrition, 98*(3), 845–853. https://doi.org/10.3945/ajcn.113.060582

Tsao C.W. & Vasan R.S. (2015). Cohort Profile: The Framingham Heart Study (FHS): overview of milestones in cardiovascular epidemiology. *International Journal of Epidemiology, 44*(6), 1800–1813. https://doi.org/10.1093/ije/dyv337

Ulven S.M., Christensen J.J., Nygård O., Svardal A., Leder L., Ottestad I., Lysne V., Laupsa-Borge J., Ueland P.M., Midttun Ø., Meyer K., McCann A., Andersen L.F. & Holven K.B. (2019). Using metabolic profiling and gene expression analyses to explore molecular effects of replacing saturated fat with polyunsaturated fat—a randomized controlled dietary intervention study. *The American Journal of Clinical Nutrition, 109*(5), 1239–1250. https://doi.org/10.1093/ajcn/nqy356

Van Horn L., Obarzanek E., Friedman L.A., Gernhofer N. & Barton B. (2005). Children's Adaptations to a Fat-Reduced Diet: The Dietary Intervention Study in Children (DISC). *Pediatrics, 115*(6), 1723–1733. https://doi.org/10.1542/peds.2004-2392

Wang B., Wu L., Chen J., Dong L., Chen C., Wen Z., Hu J., Fleming I. & Wang D.W. (2021). Metabolism pathways of arachidonic acids: mechanisms and potential therapeutic targets. *Signal Transduction and Targeted Therapy, 6*, 94. https://doi.org/10.1038/s41392-020-00443-w

Wang T., Antonacci-Fulton L., Howe K., Lawson H.A., Lucas J.K., Phillippy A.M., Popejoy A.B., Asri M., Carson C., Chaisson M.J.P., Chang X., Cook-Deegan R., Felsenfeld A.L., Fulton R.S., Garrison E.P., Garrison N.A., Graves-Lindsay T.A., Ji H., Kenny E.E., Koenig B.A., Li D., Marschall T., McMichael J.F., Novak A.M., Purushotham D., Schneider V.A., Schultz B.I., Smith M.W., Sofia H.J., Weissman T., Flicek P., Li H., Miga K.H., Paten B., Jarvis E.D., Hall I.M., Eichler E.E. & Haussler D. (2022). The Human Pangenome Project: a global resource to map genomic diversity. *Nature, 604*(7906), 437–446. https://doi.org/10.1038/s41586-022-04601-8

Wang Z. & Burge C.B. (2008). Splicing regulation: From a parts list of regulatory elements to an integrated splicing code. *RNA, 14*(5), 802. https://doi.org/10.1261/rna.876308

Wang Z., Gerstein M. & Snyder M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews Genetics, 10*(1), 57–63. https://doi.org/10.1038/nrg2484

Wheeler-Jones C.P.D. (2005). Cell signalling in the cardiovascular system: an overview. *Heart, 91*(10), 1366–1374. https://doi.org/10.1136/hrt.2005.072280

Will C.L. & Lührmann R. (2011). Spliceosome Structure and Function. *Cold Spring Harbor Perspectives in Biology, 3*(7), a003707. https://doi.org/10.1101/cshperspect.a003707

World Health Organization. (2020). Healthy diet [Referred 16.11.2022]. Available: https://www.who.int/news-room/fact-sheets/detail/healthy-diet

Young M.D., Davidson N., Wake M.J., Smyth G.K. & Oshlack A. (2017). goseq: Gene Ontology testing for RNA-seq datasets [Referred 20.03.2023]. Available:

https://bioconductor.org/packages/devel/bioc/vignettes/goseq/inst/doc/goseq.pdf

Young M.D., Wakefield M.J. Smyth G.K. & Oshlack A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology, 11*(2), R14. https://doi.org/10.1186/gb-2010-11-2-r14

Zhang W., Wang Y., Wan J., Zhang P. & Pei F. (2019). COX6B1 relieves hypoxia/reoxygenation injury of neonatal rat cardiomyocytes by regulating mitochondrial function. *Biotechnology Letters, 41*(1), 59–68. https://doi.org/10.1007/s10529-018-2614-4

Zhang Y., Ying F., Tian X., Lei Z., Li X., Lo C.-Y., Li J., Jiang L. & Yao X. (2022). TRPM2 Promotes Atherosclerotic Progression in a Mouse Model of Atherosclerosis. *Cells, 11*(9), 1423. https://doi.org/10.3390/cells11091423

Appendix 1. SAS code: Data filtering for DGE analysis

```
libname files 'C:\Users\jirikk\Desktop\GRADU';
libname out 'C:\Users\jirikk\Desktop\GRADU\out';

data data1;
set files.janette;
run;
proc sort data=data1;
by idf;
run;

data data2;
set files.interventio_diet;
run;
proc sort data=data2;
by idf;
run;

data data3;
set files.strip_inter_followup_long;
run;

proc sort data=data3;
by idf;
run;

/* Merge tables and filter data kept for pool table*/
data out.pool;
merge data1 (in=in2) data2 data3;
by idf i;
if in2;
keep sp seqname idf i batch data_order RF5 safa_e10;
run;

proc print data=out.pool (obs=20);
run;

/* Export table to excel formation*/
proc export data=out.pool
outfile= "\\utuhome.utu.fi\jirikk\GRADU\pool.xlsx"
dbms=xlsx
replace;
sheet="data";
run;
```

Appendix 2. SAS code: Standardization of continuous variable

```
proc univariate data=out.pool;
run;

/* Standardization of RF5 (continuous SAFA E% values) */
proc standard data=out.pool out=out.standard mean=0 std=1;
var RF5;
run;

/* Graph for visual inspection of the data distribution */
proc univariate normal plot data=out.standard;
run;
```

Appendix 3. SAS code: Phenotypic information and t-test

```
libname files 'C:\Users\jirikk\Desktop\GRADU';
libname out 'C:\Users\jirikk\Desktop\GRADU\out';

data data1;
set files.janette;
run;
proc sort data=data1;
by idf;
run;

data data2;
set files.interventio_diet;
run;
proc sort data=data2;
by idf;
run;

data data3;
set files.strip_inter_followup_long;
run;

proc sort data=data3;
by idf;
run;

/* Merge tables, deleting participants missing SAFA information and
filter data kept fot phtable table */
data out.phetable;
merge data1 (in=in2) data2 data3;
by idf i;
if in2;
if RF5=. then delete;
keep seqname idf i sp batch data_order RF5 safa_e10 bmi totkol ldlkol
hdlkol trigly SYST DIAST;
run;

data out.phetable2;
set out.phetable;
Age=i;
SAFAE10=RF5;
run;

/* Means and standard deviations for variables listed after var */
proc means data=muok.taulukko2 mean stddev maxdec=3 nmiss nolables;
class safa_e10 ;
var age bmi totkol ldlkol hdlkol trigly SYST DIAST SAFAE10;
run;

/*T-test:independent two-tailed t-test for each phenotypic variable */
proc ttest data=out.phetable2 H0=0 SIDES=2;
class safa_e10;
var Age;
run;

proc ttest data=out.phetable2 SIDES=2;
class safa_e10;
var totkol;
run;

proc ttest data=out.phetable2 SIDES=2;
```

```
class safa_e10;
var SAFAE10;
run;

proc ttest data=out.phetable2 SIDES=2;
class safa_e10;
var SYST;
run;

proc ttest data=out.phetable2 SIDES=2;
class safa_e10;
var ldlkol;
run;

proc ttest data=out.phetable2 SIDES=2;
class safa_e10;
var bmi;
run;

proc ttest data=out.phetable2 SIDES=2;
class safa_e10;
var sp;
run;

proc ttest data=out.phetable2 SIDES=2;
class safa_e10;
var batch;
run;

proc ttest data=out.phetable2 SIDES=2;
class safa_e10;
var hdlkol;
run;

proc ttest data=out.phetable2 SIDES=2;
class safa_e10;
var trigly;
run;

proc ttest data=out.phetable2 SIDES=2;
class safa_e10;
var DIAST;
run;

/* Gives group size information */
proc sql;
select safa_e10, count(*) as N
from out.phetable2
group by safa_e10;
quit;
```

Appendix 4. R code: DESeq2 Differential gene expression analysis

```
# Install the packages and tools needed
install.packages("BiocManager")
library("BiocManager")

BiocManager::install("DESeq2")
library(DESeq2)

#Specify a working directory
setwd("//utuhome.utu.fi/jirikk/GRADU")

# Load mRNA-Seq data and gene annotation data
load("strip_rnaseq_set1_analysis_data.RData")
load("strip_rnaseq_set1_gene_annotation.RData")

# Load clinical data
metadata <-read.table("uusi_jatkuvamuut.txt", header = T)

# Convert vectors to factor class
metadata$SP <- as.factor(metadata$SP)
metadata$i <- as.factor(metadata$i)
metadata$safa_e10 <- as.factor(metadata$safa_e10)
metadata$batch <- as.factor(metadata$batch)

# Sample code (seqname) as rowname
row.names(metadata) <- metadata$seqname

## Check data
# Rownames and column names have same content?
all(rownames(metadata) %in% colnames(strip_data))
# [1] TRUE

# Rownames and column names in same order?
all(rownames(metadata) == colnames(strip_data))
# [1] TRUE

# Delete from clinic data rows that miss SAFA E% values (RF5)
metaPNA <- metadata[!is.na(metadata["RF5"]),]

# Delete equivalent rows from mRNA-seq data
strip_PNA <- strip_data[,rownames(metaPNA)]

# Confirm that no NA left in data
any(is.na(metaPNA))
# [1] FALSE

# Confirm that row names in metaPNA data and column names in strip_PNA data are in same
order
all(rownames(metaPNA) == colnames(strip_PNA))
# [1] TRUE

## Analyzing RNA-seq data with DESeq2 #Creating Design matrix
dds <- DESeqDataSetFromMatrix(countData = strip_PNA,
                              colData = metaPNA,
                              design = ~ SP + i + batch + RF5)
```

```
# Running the analysis
dds <- DESeq(dds)

# Save the result
res.05 <- results(dds, alpha = 0.05)

# Summary of the results
summary(res.05)
#out of 25457 with nonzero total read count
#adjusted p-value < 0.05
#LFC > 0 (up)       : 64, 0.25%
#LFC < 0 (down)     : 10, 0.039%
#outliers [1]       : 0, 0%
#low counts [2]     : 16777, 66%
#(mean count < 222)
#[1] see 'cooksCutoff' argument of ?results
#[2] see 'independentFiltering' argument of ?results

# How many genes have adjusted p-value < 0.1?
sum(res.05$padj < 0.1, na.rm = TRUE)
#492

# How many genes have adjusted p-value < 0.05?
sum(res.05$padj < 0.05, na.rm = TRUE)
#74

# How many genes have adjusted p-value < 1.0? (needed for the volcano plot)
sum(res.05$padj < 1.0, na.rm = TRUE)
#8680

# Add gene annotation to the results
# Creating results table
resAll.05 <- as.data.frame(res.05)
# Add column ”entrezgene” and give same values as row names in resAll
resAll.05$entrezgene <- row.names(resAll.05)
# Combine resAll and annot by enterzgene
resAll.annot.05 <- merge(resAll.05, annot, by="entrezgene")

# Save all DEG results
BiocManager::install("xlsx")
library("xlsx")

write.xlsx(resAll.annot.05, file="STRIP-RF5-sp-ika-batch-adj-tulokset-kaikki.alpha.xlsx",
sheetName = "DESeq2", col.names = TRUE, row.names = F, append = FALSE)

# Save DEG results when fdr < 0.05
resSig.annot.J <- subset(resAll.annot.05, padj < 0.05)
resSig2.annot.J <- resSig.annot.J[order(resSig.annot.J$pvalue),]
write.xlsx(resSig2.annot.J, file="STRIP-RF5-sp-ika-batch-adj-tulokset-FDR05-alpha-
toisto.xlsx", sheetName = "DESeq2", col.names = TRUE, row.names = F, append = FALSE)

# Save results for volcano plot
resvolc <- subset(resAll.annot.05, padj < 1.0)
write.xlsx(resvolc, file="STRIP-RF5-sp-ika-batch-adj-tulokset-volcanoplot-alpha-toisto.xlsx",
sheetName = "DESeq2", col.names = TRUE, row.names = F, append = FALSE)
```

Appendix 5. R code: Volcano plot

```
#Download packages
require(ggplot2)
library("ggrepel")
library("data.table")
library("tidyverse")
library(dplyr)

#Specify a working directory
setwd("//utuhome.utu.fi/jirikk/GRADU")

#Load input data
strip <- fread("//utuhome.utu.fi/jirikk/GRADU/STRIP-RF5-sp-ika-batch-adj-tulokset-volcanoplot-alpha.txt")

#Add new column "significance"
#log2FC<0 and padj<0.05 = Down
strip$Significance[strip$log2FoldChange < 0 & strip$padj < 0.05] <-  "down"
#log2FC>0 and padj<0.05 = Up
strip$Significance[strip$log2FoldChange > 0 & strip$padj < 0.05] <-  "up"
#padj>0.05 = None
strip$Significance[strip$padj > 0.05] <-  "none"

#Convert character vector to factor class
strip$Significance <- as.factor(strip$Significance)

#Add new column "typeface"
strip$typeface <- sample(c("italic"))

#For Significance determine the order of factors
strip$Significance <- factor(strip$Significance, levels=rev(c("none","up","down")))

#Add new column "pvalue10" for -log10 p-values
strip$pvalue10 <- -log10(strip$pvalue)

#X-axis limits
min(strip$log2FoldChange)
# -0.249819
max(strip$log2FoldChange)
# 0.3737007

#Cutoff based on FDR cutoff
-log10(0.00044)
#3.356547

#Cutoff based on P-value
-log10(0.05)
#1.30103

#Plot design
plot <- ggplot(strip, aes(x = log2FoldChange, y = -log10(pvalue))) +
  geom_point(aes(color = Significance), shape = 21, size = 3.5, stroke =1.25, fill = "white") +
  scale_color_manual(values = c("blue", "red",    "grey")) +
  theme_bw(base_size = 13) + theme(legend.position = "none")  +
  ylab(bquote(-log[10](P-value))) +  xlab(bquote(log[2](Fold~Change))) +
```

```
  geom_text_repel(
    data = subset(strip, pvalue < 3.32E-05 | log2FoldChange < -0.6  & Significance == "down" |
log2FoldChange > 0.6),
    aes(label = symbol, fontface = typeface),
    size = 3.5,
    box.padding = unit(0.3, "lines"),
    point.padding = unit(0.3, "lines"),max.overlaps = Inf
  ) + scale_x_continuous(limits = c(-0.25, 0.4), breaks=c(-0.3, -0.2, -0.1, 0, 0.1, 0.2, 0.3, 0.4)) +
geom_vline(xintercept = 0, color="darkgrey",linetype="dashed", linewidth=0.5) +
  scale_y_continuous(breaks=c(0,2.5, 5,7.5, 10)) + geom_hline(yintercept = 3.356547,
color="black",linetype="dashed", linewidth=1) +geom_hline(yintercept =1.30103 ,
color="gray",linetype="dashed", linewidth=1) + theme(panel.background=element_blank()) +
  theme(axis.text.x = element_text(size = 12), axis.text.y = element_text(size = 12)) +
theme(panel.grid.minor = element_blank())

#print out the plot
plot

#save results
ggsave("//utuhome.utu.fi/jirikk/GRADU/STRIP-RF5-sp-ika-batch-adj-tulokset-
volcanoplot.png", device='png', dpi=600)

save.image("//utuhome.utu.fi/jirikk/GRADU/STRIP-RF5-sp-ika-batch-adj-tulokset-
volcanoplot.RData")
```

Appendix 6. R code: GOseq analysis

```
# Install packages needed
BiocManager::install("goseq")
BiocManager::install("qusage")
install.packages("tidyverse")
install.packages("ggplot2")
install.packages("forcats")
install.packages("plyr")
install.packages("ggplot")

########### GOseq with DEG list ###########
library(goseq)

# Check the number of up- and downregulated genes
resSig2.up <- resSig2.annot.J[which(resSig2.annot.J$log2FoldChange>0),] #64
resSig2.down <- resSig2.annot.J[which(resSig2.annot.J$log2FoldChange<0),] #10

################ Downregulated genes ############
# Prepare GOseq input table for downregulated genes
genes.dat.down <- data.frame(genes=resAll.annot.05$symbol,de=0)
genes.dat.down$de[genes.dat.down$genes %in% resSig2.down$symbol]=1
table(genes.dat.down$de)

genes.down <- genes.dat.down$de
names(genes.down) <- genes.dat.down$genes
table(genes.down)

## Fit probability weighting function

# Install genom wide annotation for human and reference genome
BiocManager::install("org.Hs.eg.db")
library("org.Hs.eg.db")

BiocManager::install("TxDb.Hsapiens.UCSC.hg38.knownGene")
library("TxDb.Hsapiens.UCSC.hg38.knownGene")

pwf.down <- nullp(genes.down,"hg38","geneSymbol")

head(pwf.down)
#          DEgenes    bias.data        pwf
#A1BG            0    1949.5         0.001248111
#NAT2            0    716.0          0.001287071
#ADA             0    1128.0         0.001286474
#CDH2            0    3143.0         0.001133015
#AKT3            0    1583.0         0.001277553
#TRY-GTA7-1      0    NA               NA

################ Upregulated genes ############
# Prepare GOseq input table for upregulated genes

genes.dat.up <- data.frame(genes=resAll.annot.05$symbol,de=0)
genes.dat.up$de[genes.dat.up$genes %in% resSig2.up$symbol]=1
table(genes.dat.up$de)

genes.up <- genes.dat.up$de
names(genes.up) <- genes.dat.up$genes
table(genes.up)

# Fit probability weighting function
pwf.up <- nullp(genes.up,"hg38","geneSymbol")
```

```
head(pwf.up)
#          DEgenes   bias.data        pwf
#A1BG            0   1949.5          0.001443394
#NAT2            0   716.0           0.005309832
#ADA             0   1128.0          0.004254169
#CDH2            0   3143.0          0.001084923
#AKT3            0   1583.0          0.002433998
#TRY-GTA7-1      0   NA              NA

########### GOseq with MSigDB ontology gene sets #############
# Install packages needed
BiocManager::install("limma")
library("limma")
library("qusage")

# Load Ontology gene set from MSigDB
ontol <- read.gmt("c5.all.v2022.1.Hs.symbols.gmt")

length(unlist(ontol)) # 1281285 genes
length(unique(unlist(ontol))) # 19447, means that one gene can belong to more than one category
df.ontol <- stack(ontol)
length(df.ontol$values);length(unique(df.ontol$values))
#[1] 1281285
#[1] 19447

################ Downregulated genes #############
ontol.wall.down <- goseq(pwf.down,"hg38","geneSymbol",gene2cat = df.ontol)
ontol.wall.down$over_represented_p.adj <-
p.adjust(ontol.wall.down$over_represented_pvalue,method="fdr")
ontol.wall.down$over_represented_p.adj2 <- -log10(ontol.wall.down$over_represented_p.adj)

library(ggplot2)

ggplot(ontol.wall.down, aes(x = reorder(category, over_represented_p.adj2), y =
over_represented_p.adj2)) +
  geom_bar(stat = "identity", width=0.7,colour="steelblue", fill="steelblue") + labs(x = "MSigDB
ontology gene set collection categories", y="-log10(FDR)") +
  coord_flip() + geom_hline(yintercept = 1.30103, color="darkgrey",linetype="dashed", size=1) +
  geom_hline(yintercept = 0, color="darkgrey",linetype="solid", size=0.5)

# Calculate -log10 for FDR = 0.05 cutoff and use in plotting
-log10(0.05)
# 1.30103

ontol.wall.down.plot <- ontol.wall.down[!ontol.wall.down$over_represented_p.adj2==0,]

down <- function()
{
  ggplot(ontol.wall.down.plot, aes(x = reorder(category, over_represented_p.adj2), y =
over_represented_p.adj2)) +
    geom_bar(stat = "identity", width=0.7,colour="steelblue", fill="steelblue") + labs(x = "MSigDB
ontology gene set collection categories", y="-log10(FDR)") +
    coord_flip() + geom_hline(yintercept = 1.30103, color="darkgrey",linetype="dashed", size=1) +
    geom_hline(yintercept = 0, color="darkgrey",linetype="solid", size=0.5)
}

#Save results
ggsave("STRIP_RNAseq_RF5_exp_goseq_ontology.down.2.png", down(),device='png', dpi=600)

write.xlsx(ontol.wall.down, file="STRIP_RNAseq_RF5_exp_goseq_ontology.down.2.xlsx", sheetName
= "ontology", col.names = TRUE, row.names = F, append = FALSE)
```

```
###Table of downregulated DE genes which are enriched to GOCC_IMMUNOGLOBULIN_COMPLEX pathway
allOntol_Sig <- df.ontol[df.ontol$values %in% resSig2.down$symbol,]
allOntol_Sig$ind <- as.character(allOntol_Sig$ind)
Ontol_Sig_downDEG <- subset(allOntol_Sig, ind=="GOCC_IMMUNOGLOBULIN_COMPLEX", select=c(values, ind))

#Save results
write.xlsx(Ontol_Sig_downDEG, file="Ontol_signif_down_DEGs.xlsx", sheetName = "ontoldown", col.names = TRUE, row.names = F, append = FALSE)

################ Upregulated genes ############
ontol.wall.up <- goseq(pwf.up,"hg38","geneSymbol",gene2cat = df.ontol)
ontol.wall.up$over_represented_p.adj <- p.adjust(ontol.wall.up$over_represented_pvalue,method="fdr")
ontol.wall.up$over_represented_p.adj2 <- -log10(ontol.wall.up$over_represented_p.adj)

ggplot(ontol.wall.up, aes(x=category, y=-log10(over_represented_p.adj), label="FDR")) +
  geom_bar(stat = "identity", width=0.7,colour="tomato",fill="tomato") + labs(x = "MSigDB ontology gene set collection categories", y="-log10(FDR)") +
  coord_flip() + geom_hline(yintercept = 1.30103, color="darkgrey",linetype="dashed", size=1)

ontol.wall.up.plot <- ontol.wall.up[!ontol.wall.up$over_represented_p.adj2==0,]

up <- function()
{
ggplot(ontol.wall.up.plot, aes(x = reorder(category, over_represented_p.adj2), y = over_represented_p.adj2)) +
    geom_bar(stat = "identity", width=0.7,colour="tomato",fill="tomato") + labs(x = "MSigDB ontology gene set collection categories", y="-log10(FDR)") +
    coord_flip() + geom_hline(yintercept = 1.30103, color="darkgrey",linetype="dashed", size=1) +
    geom_hline(yintercept = 0, color="darkgrey",linetype="solid", size=0.5)
}

#Save results
ggsave("STRIP_RNAseq_RF5_exp_goseq_ontology.up.2.0.7.png", up(),device='png', dpi=600)

write.xlsx(ontol.wall.up, file="STRIP_RNAseq_RF5_exp_goseq_ontology_up.2.xlsx", sheetName = "ontology", col.names = TRUE, row.names = F, append = FALSE)

###Table of upregulated DE genes which are enriched to GOBP_MEMBRANE_BIOGENESIS pathway
allOntol_Sig_up <- df.ontol[df.ontol$values %in% resSig2.up$symbol,]
allOntol_Sig_up$ind <- as.character(allOntol_Sig_up$ind)
Ontol_Sig_UP_DEG <- subset(allOntol_Sig_up, ind=="GOBP_MEMBRANE_BIOGENESIS", select=c(values, ind))

#Save results
write.xlsx(Ontol_Sig_UP_DEG, file="Ontol_signif_up_DEGs.xlsx", sheetName = "ontolUP", col.names = TRUE, row.names = F, append = FALSE)
```

Appendix 7. R code: edgeR analysis

```
# edgeR, pathway-level analysis programs and utility packages
BiocManager::install("edgeR")
BiocManager::install("goseq")
BiocManager::install("qusage") # to read rmt
install.packages("ggplot2")
install.packages("forcats")
install.packages("plyr")
install.packages("data.table")
BiocManager::install("xlsx")

# Set working directory
setwd("//utuhome.utu.fi/jirikk/GRADU")

# Load mRNA-Seq data and gene annotation data
load("strip_rnaseq_set1_analysis_data.RData")
load("strip_rnaseq_set1_gene_annotation.RData")

# Load clinic data
metadata <-read.table("uusi_jatkuvamuut.txt", header = T)

##Editing data using the same codes as used in DESeq2 analysis
metadata$SP <- as.factor(metadata$SP)
metadata$i <- as.factor(metadata$i)
metadata$safa_e10 <- as.factor(metadata$safa_e10)
metadata$batch <- as.factor(metadata$batch)

row.names(metadata) <- metadata$seqname

metaPNA <- metadata[!is.na(metadata["RF5"]),]
strip_PNA <- strip_data[,rownames(metaPNA)]

#######edgeR analysis (input data gene count data)##############
library("limma")
library("edgeR")

dds <- DGEList(counts=strip_PNA)
design <- model.matrix(~RF5 + i + SP + batch, data=metaPNA)

# Delete low expression genes
keep <- filterByExpr(dds, design)
table(keep)
# FALSE  TRUE
# 8077   17485

dds <- dds[keep, , keep.lib.sizes=FALSE]

# Normalization
dds <- calcNormFactors(dds)

# Estimate the overall dispersion for the dataset, to get an idea of the overall level of biological
variability
d <- estimateDisp(dds, design)
# Then estimate gene-wise dispersion estimates, allowing a possible trend with average count
size:
```

```
d <- estimateGLMTrendedDisp(d)
d <- estimateGLMTagwiseDisp(d)

plotBCV(d)

### Gene set analysis ###
library(qusage)
library(forcats)
library(dplyr)
library(ggplot2)

#HALLMARK GENE SET
h <- read.gmt("h.all.v2022.1.Hs.entrez.gmt")
fry.res.h <- fry(d, index=h, design=design, contrast=2)

fry.res.h$category <- row.names(fry.res.h)
fry.res.h$Direction <- as.factor(fry.res.h$Direction)

# To keep the plot clear filtering pathways
fry.res.h.plot <- fry.res.h[!fry.res.h$FDR>0.20,]

# Create plot where pathway names on y-axis and -log10(FDR) on x-axis
# Add dashed cut off line
ggplot(fry.res.h.plot, aes(x = reorder(category, -FDR), y=-log10(FDR),
label="FDR",color=Direction, fill=Direction)) +
  geom_bar(stat = "identity", width=0.7) + scale_fill_manual("Direction", values = c("Down" =
"steelblue", "Up" = "tomato")) +
  scale_colour_manual("Direction", values = c("Down" = "steelblue", "Up" = "red")) +
  labs(x = "MSigDB Hallmark gene set collection categories", y="-log10(FDR)") +
  coord_flip() + geom_hline(yintercept = 1.30103, color="darkgrey",linetype="dashed", size=1)

h.plot <- function()
{
  ggplot(fry.res.h.plot, aes(x = reorder(category, -FDR), y=-log10(FDR),
label="FDR",color=Direction, fill=Direction)) +
    geom_bar(stat = "identity", width=0.7) + scale_fill_manual("Direction", values = c("Down" =
"steelblue", "Up" = "tomato")) +
    scale_colour_manual("Direction", values = c("Down" = "steelblue", "Up" = "tomato")) +
    labs(x = "MSigDB Hallmark gene set collection categories", y="-log10(FDR)") +
    coord_flip() + geom_hline(yintercept = 1.30103, color="darkgrey",linetype="dashed",
size=1)
}

# Save results
ggsave("STRIP_RNAseq_RF5_edgeR_hallmark_pathway_FDR_0.2_results.png",
h.plot(),device='png', dpi=600)

library("xlsx")
write.xlsx(fry.res.h, file="STRIP-RF5-sp-ika-batch-adj-edgeR-Hallmark-pathway-tulokset-
kaikki.xlsx", sheetName = "edgeR", col.names = TRUE, row.names = F, append = FALSE)
```

Appendix 8. Downregulated genes from DGE analysis with continuous SAFA E%

NCBI entrezgene ID	Gene symbol	Gene description	Log2 Fold Change	SE	P-value	B-H adj. P-value
28822	*IGLV1-47*	immunoglobulin lambda variable 1-47	-0,224	0,051	1,05E-05	1,76E-02
28908	*IGKV4-1*	immunoglobulin kappa variable 4-1	-0,250	0,059	2,44E-05	1,93E-02
6996	*TDG*	thymine DNA glycosylase	-0,038	0,010	1,12E-04	4,00E-02
146050	*ZSCAN29*	zinc finger and SCAN domain containing 29	-0,045	0,012	2,11E-04	4,27E-02
9657	*IQCB1*	IQ motif containing B1	-0,042	0,011	2,15E-04	4,27E-02
8726	*EED*	embryonic ectoderm development	-0,032	0,009	2,47E-04	4,51E-02
28874	*IGKV3D-20*	immunoglobulin kappa variable 3D-20	-0,176	0,049	3,51E-04	4,75E-02
28434	*IGHV3-33*	immunoglobulin heavy variable 3-33	-0,158	0,044	3,15E-04	4,75E-02
115426	*UHRF2*	ubiquitin like with PHD and ring finger domains 2	-0,036	0,010	2,78E-04	4,75E-02
27097	*TAF5L*	TATA-box binding protein associated factor 5 like	-0,025	0,007	3,60E-04	4,75E-02

Appendix 9. Upregulated genes from DGE analysis with continuous SAFA E%

NCBI entrezgene ID	Gene symbol	Gene description	Log2 Fold Change	SE	P-value	B-H adj. P-value
2286	*FKBP2*	FKBP prolyl isomerase 2	0,081	0,016	6,20E-07	5,38E-03
83658	*DYNLRB1*	dynein light chain roadblock-type 1	0,058	0,013	1,31E-05	1,76E-02
200185	*KRTCAP2*	keratinocyte associated protein 2	0,060	0,014	1,10E-05	1,76E-02
8721	*EDF1*	endothelial differentiation related factor 1	0,064	0,014	6,15E-06	1,76E-02
4696	*NDUFA3*	NADH:ubiquinone oxidoreductase subunit A3	0,071	0,016	1,42E-05	1,76E-02
513	*ATP5F1D*	ATP synthase F1 subunit delta	0,090	0,021	1,20E-05	1,76E-02
1340	*COX6B1*	cytochrome c oxidase subunit 6B1	0,061	0,014	1,65E-05	1,79E-02
83442	*SH3BGRL3*	SH3 domain binding glutamate rich protein like 3	0,069	0,016	1,97E-05	1,90E-02
1891	*ECH1*	enoyl-CoA hydratase 1	0,048	0,011	2,49E-05	1,93E-02
8815	*BANF1*	BAF nuclear assembly factor 1	0,050	0,012	2,71E-05	1,93E-02
2014	*EMP3*	epithelial membrane protein 3	0,052	0,013	3,78E-05	1,93E-02
53827	*FXYD5*	FXYD domain containing ion transport regulator 5	0,056	0,013	3,72E-05	1,93E-02
317749	*DHRS4L2*	dehydrogenase/reductase 4 like 2	0,062	0,015	3,36E-05	1,93E-02
400916	*CHCHD10*	coiled-coil-helix-coiled-coil-helix domain containing 10	0,069	0,017	3,32E-05	1,93E-02
6158	*RPL28*	ribosomal protein L28	0,074	0,018	3,17E-05	1,93E-02
10726	*NUDC*	nuclear distribution C, dynein complex regulator	0,047	0,011	4,03E-05	1,94E-02
27243	*CHMP2A*	charged multivesicular body protein 2A	0,069	0,018	7,42E-05	3,39E-02
6281	*S100A10*	S100 calcium binding protein A10	0,054	0,014	8,13E-05	3,40E-02
723790	*H2AC19*	H2A clustered histone 19	0,115	0,029	8,23E-05	3,40E-02
9168	*TMSB10*	thymosin beta 10	0,062	0,016	1,02E-04	3,86E-02
5441	*POLR2L*	RNA polymerase II, I and III subunit L	0,068	0,017	9,87E-05	3,86E-02
587	*BCAT2*	branched chain amino acid transaminase 2	0,045	0,012	1,18E-04	4,00E-02
1072	*CFL1*	cofilin 1	0,048	0,013	1,38E-04	4,00E-02
826	*CAPNS1*	calpain small subunit 1	0,051	0,013	1,36E-04	4,00E-02
11270	*NRM*	nurim	0,052	0,014	1,34E-04	4,00E-02
51079	*NDUFA13*	NADH:ubiquinone oxidoreductase subunit A13	0,059	0,016	1,41E-04	4,00E-02
3105	*HLA-A*	major histocompatibility complex, class I, A	0,061	0,016	1,27E-04	4,00E-02
3315	*HSPB1*	heat shock protein family B (small) member 1	0,069	0,018	1,43E-04	4,00E-02
58485	*TRAPPC1*	trafficking protein particle complex subunit 1	0,047	0,012	1,58E-04	4,23E-02
2950	*GSTP1*	glutathione S-transferase pi 1	0,057	0,015	1,61E-04	4,23E-02
5315	*PKM*	pyruvate kinase M1/2	0,045	0,012	2,23E-04	4,27E-02
11230	*PRAF2*	PRA1 domain family member 2	0,051	0,014	2,02E-04	4,27E-02
10901	*DHRS4*	dehydrogenase/reductase 4	0,052	0,014	1,96E-04	4,27E-02
28956	*LAMTOR2*	late endosomal/lysosomal adaptor, MAPK and MTOR activator 2	0,053	0,014	1,91E-04	4,27E-02
5216	*PFN1*	profilin 1	0,053	0,014	1,80E-04	4,27E-02
101410538	*MMP24OS*	MMP24 opposite strand	0,055	0,015	2,19E-04	4,27E-02
7226	*TRPM2*	transient receptor potential cation channel subfamily M member 2	0,059	0,016	2,05E-04	4,27E-02
5691	*PSMB3*	proteasome 20S subunit beta 3	0,063	0,017	2,11E-04	4,27E-02
1627	*DBN1*	drebrin 1	0,065	0,017	1,77E-04	4,27E-02

290	*ANPEP*	alanyl aminopeptidase, membrane	0,087	0,024	2,12E-04	4,27E-02
6282	*S100A11*	S100 calcium binding protein A11	0,088	0,024	2,26E-04	4,27E-02
4726	*NDUFS6*	NADH:ubiquinone oxidoreductase subunit S6	0,061	0,017	2,50E-04	4,51E-02
51181	*DCXR*	dicarbonyl and L-xylulose reductase	0,063	0,017	2,62E-04	4,64E-02
7332	*UBE2L3*	ubiquitin conjugating enzyme E2 L3	0,030	0,008	3,15E-04	4,75E-02
11140	*CDC37*	cell division cycle 37, HSP90 cochaperone	0,031	0,009	3,77E-04	4,75E-02
5036	*PA2G4*	proliferation-associated 2G4	0,032	0,009	3,38E-04	4,75E-02
6810	*STX4*	syntaxin 4	0,033	0,009	3,77E-04	4,75E-02
8402	*SLC25A11*	solute carrier family 25 member 11	0,035	0,010	3,08E-04	4,75E-02
10204	*NUTF2*	nuclear transport factor 2	0,036	0,010	3,82E-04	4,75E-02
8021	*NUP214*	nucleoporin 214	0,039	0,011	3,45E-04	4,75E-02
7936	*NELFE*	negative elongation factor complex member E	0,040	0,011	3,24E-04	4,75E-02
11243	*PMF1*	polyamine modulated factor 1	0,047	0,013	3,47E-04	4,75E-02
84335	*AKT1S1*	AKT1 substrate 1	0,050	0,014	3,62E-04	4,75E-02
374882	*TMEM205*	transmembrane protein 205	0,051	0,014	2,83E-04	4,75E-02
6892	*TAPBP*	TAP binding protein	0,051	0,014	3,82E-04	4,75E-02
302	*ANXA2*	annexin A2	0,054	0,015	3,83E-04	4,75E-02
11337	*GABARAP*	GABA type A receptor-associated protein	0,057	0,016	3,79E-04	4,75E-02
10870	*HCST*	hematopoietic cell signal transducer	0,060	0,017	3,20E-04	4,75E-02
7305	*TYROBP*	transmembrane immune signaling adaptor TYROBP	0,074	0,021	3,68E-04	4,75E-02
3580	*CXCR2P1*	C-X-C motif chemokine receptor 2 pseudogene 1	0,114	0,032	3,45E-04	4,75E-02
27341	*RRP7A*	ribosomal RNA processing 7 homolog A	0,058	0,016	3,89E-04	4,76E-02
51035	*UBXN1*	UBX domain protein 1	0,042	0,012	3,95E-04	4,76E-02
1460	*CSNK2B*	casein kinase 2 beta	0,034	0,010	4,07E-04	4,83E-02
9092	*SART1*	spliceosome associated factor 1, recruiter of U4/U6.U5 tri-snRNP	0,045	0,013	4,19E-04	4,92E-02