Tampereen yliopisto

Miikael Lehtimäki

UNCOVERING THE COMPLEX GENETIC ARCHITECTURE OF HUMAN PLASMA LIPIDOME

Master of Science Thesis
Faculty of Information Technology and Communication Sciences
Examiners: PhD Pashupati Mishra and
Prof. Terho Lehtimäki
June 2021

TIIVISTELMÄ

Lehtimäki Miikael: Ihmisen plasman lipidomin monimutkaisen geneettisen arkkitehtuurin selvittäminen
Diplomityö, 60 sivua (27 kuvaa, 3 taulukkoa 12 koodia), 8 lisätaulukkoa
Tampereen yliopisto
Tietotekniikan maisterikoulutus
Kesäkuu 2021

Tausta: Plasman lipidomin geneettisen arkkitehtuurin yksityiskohtainen ymmärtäminen antaa tarkemman kuvan rasva-aineiden metabolian säätelystä ja rasva-aineiden yhteydestä kardiometabolisiin sairauksiin. Tein työssä 437 erilaisen plasman rasva-aineen (=lipidomi) kokogenomin laajuisen assosiaatioanalyysin (GWAS) ja käytin sitä täydentävää koneälytekniikkaa ns. fenotyyppi-genotyyppi moni-moneen suhdeanalyysiä (*engl.* PGMRA), löytääkseni monimutkaisia aikaisemmin tunnistamattomia lipidifenotyyppi-genotyyppi verkostoja ja uusia vielä tuntemattomia geenejä, joilla on vaikutuksia ihmisen rasva-aineenvaihduntaan ja plasman lipidomiin.

Tavoitteet: Käyttää toisiaan täydentäviä datapohjaisia GWAS ja koneoppimismenetelmiä ja selvittää ihmisen lipidomin monimutkainen ja aikaisemmin suurimaksi osaksi tuntematon geneettinen arkkitehtuuri.

Aineisto ja menetelmät: Aineisto koostui 1426 iältään 30–45 vuotiaasta Suomalaisesta miehestä ja naisesta (53 %), jotka osallistuivat Lasten ja nuorten Sepelvaltimotaudin Riskitekijät seurantatutkimukseen (*engl.* Young Finns study). GWAS tehtiin 437:lle erityyppiselle plasman rasva-aineelle käyttäen 546,677 genotyypattua yhden emäsmuutoksen aiheuttamaa DNA geenivarianttia eli geenimuunnosta (*engl*. single nucleotide polymorphisms SNPs). Plasman lipidomi määritettiin massaspektrometria (LC-MS/MS) tekniikalla ja genotyypien määritys tehtiin Illuminan mikroarray teknologialla käyttäen Illuminan Custom 670K genotyypitys sirua. Aineistosta tunnistettiin PGMRA:llä biklusteroimalla sen henkilöihin liittyviä SNP- ja lipidifenotyyppiryhmiä, jotka liittyivät tilastollisesti merkitsevästi toisiinsa hypergeometrisessa testissä (vertailtavat ryhmät sisälsivät samoja henkilöitä). Näin tunnistettujen tilastollisesti merkitsevien geeniryhmien biologista toimintaa tutkittiin edelleen bioinformaattisella geeniryhmien rikastusanalyysillä (*engl.* GSEA) GSEA:lla analysoitiin yhteensä 28922 biologisiin prosesseihin ja geeniontologiaan perustuvaa geeniryhmää. Työn analyyseissä käytettiin sekä R-kielistä ohjelmistoympäristöä ja koodistoja sekä PLINK-ohjelmaa (koodit 1–12 esitetty tekstissä).

Tulokset: Lipidominlaajuisessa GWAS analyysissä löysimme 266 näihin 437 plasman rasva-aineeseen tilastollisesti merkittävästi ($P < 5 \times 10^{-8}$) yhteydessä olevaa SNP:iä. PLINK-ohjelmalla tehdyssä regressioanalyysissä löytyi lisäksi 18,370 erillistä nominaalisesti merkitsevää ($P < 5 \times 10^{-4}$) SNP-rasva-aine assosiaatiota. PGMRA jatkoanalyysissä käytimme näitä nominaalisesti merkitseviä SNP:ejä ja lipidifenotyyppi dataa.

PGMRA-analyysissä aineistosta löytyi 93 tilastollisesti merkittävää suhdetta biklusteroimalla saatujen (genotyyppi-henkilö) vs. (lipidifenotyyppi-henkilö) ryhmien välillä. Näiden biklustereiden välisiin suhteisiin sisältyi yhteensä 5977 eri SNP:iä. Viimeisimpään genominrakennereferenssiin (ensemble assembly GRCh37, versio 102) sijoitettuina nämä SNP:t paikallistuivat 3164 eri geenilokukseen, jotka assosioituivat tilastollisesti merkitsevästi plasman rasva-aineenvaihduntaa kuvaaviin lipidiryppäisiin. Näistä 93 ryhmästä 35 oli erillisiä, eli niihin ei sisältynyt yhtään samaa SNP:iä tai henkilöä.

Näistä 35 erillisestä geneettisestä lipidomialaryhmästä 18:ta ryhmän sisältämät geenivariantit kertyivät tilastollisesti merkitsevästi GSEA:ä useisiin biologisesti merkittäviin prosesseihin tai aineenvaihduntateihin. Näiden biologisten prosessien ja aineenvaihduntateiden välittäminä löydetyt geenivariantit voivat säädellä ja vaikuttaa plasman lipidiprofiileihin.

Päätelmät: Ihmisen plasman lipidomiin vaikuttaa yli 3164 eri geeniin paikallistuva geneettinen variaatio ja ihmiset voidaan sen suhteen luokitella 35 erilaiseen geneettiseen lipidifenotyyppialaryhmään. Tämän työn uutuusarvona on sen tulosten lisäksi se, että siinä on ensimmäisen kerran sovellettu GWAS-PGMRA-GSEA menetelmien yhdistelmää ihmisen

lipidomin monimutkaisen geenitaustan tutkimiseen. Työn tuloksena on löydetty joukko uusia lipidomin säätelyyn, sen biologisiin prosesseihin ja aineenvaihduntateihin vaikuttavia tilastollisesti merkitsevästi liittyviä geenivariantteja ja geenejä. Tutkimuksen tulokset täytyy kuitenkin vielä varmistaa toisessa vastaavassa, mutta riippumattomassa aineistossa.

Avainsanat: Genominlaajuinen assosiaatiotutkimus (GWAS), PGMRA, lipidomi, geenien rikastamisanalyysi, genetiikka.

ABSTRACT

Lehtimäki Miikael: Uncovering the complex genetic architecture of human lipidome.
Master of Science Thesis, 60 pages (27 figures, 3 tables, 12 codes), 8 supplementary tables
Tampere University
Master's Degree Education in Information Technology
June 2021

Background: Understanding the genetic architecture of plasma lipidome could provide better insights into regulation of lipid metabolism and its link to cardiometabolic diseases. I performed genome-wide association study (GWAS) of 437 quantified plasma lipid species followed by a use of a novel machine learning method, phenotype-genotype many to many relationships analysis (PGMRA), to uncover the hidden complex genotypic–phenotypic networks and novel genes i.e. missing inheritance for human lipid metabolisms and plasma lipidome.

Objectives: To use GWAS and a complementary data-driven machine learning methods to uncover the complex and hidden genetic architecture of human plasma lipidome.

Subjects and Methods: The study sample consisted of 1426, 30–45-year-old Finnish men and women (53 %) taking part in ongoing Young Finns study (YFS). GWAS was performed between lipidome data with 437 mass-spectrometry (LC-MS/MS) quantified plasma lipid species and 546,677 single-nucleotide polymorphisms (SNPs) genotyped using Illumina 670K custom bead chip. Genotype data consisting of nominally significant SNPs from GWAS ($p < 5x10^{-4}$) and lipidome data were further analysed with a novel machine learning method, phenotype-genotype many to many relationship analyses (PGMRA). PGMRA involved biclustering of genotype and lipidome data independently yielding SNP-subject sets and lipid-subject sets from genotype and lipidome data respectively. Then, association analysis between the SNP-subject sets and the lipid-subject sets was done by calculating overlap of subjects in each pair of sets using hypergeometric test. Biological significance of the significant SNP sets was further studied by using gene set enrichment analysis (GSEA). Using GSEA altogether 28922 Biological process Gene Ontology based gene sets were analyzed. The biostatistical analyses were done in open access R-statistical environment using R-based coding and using PLINK-program (codes 1-12 shown in the text).

Results: We identified 266 SNPs significantly associated (p-value$< 5 \times 10^{-8}$) and 18370 SNPs nominally significantly associated (p-value$< 5 \times 10^{-4}$) with 437 studied molecular lipids with traditional lipidome wide GWAS analysis over whole lipidome. Using PGMRA biclustering analysis for a subset of genotype data with these nominally significant SNPs as preselected SNPs and lipidome data as phenotypes, we found 93 statistically significant genotype-subject vs. lipid phenotype-subject group relations involving 5977 separate SNPs. After their gene annotation to latest available genome reference (ensemble assembly GRCh37, the version 102) these SNPs were located into 3164 separate gene loci. Thirty-five of the significant SNP sets did not share any SNP or subject, therefore representing 35 genetically distinct lipidomic profiles i.e. genetic lipidomic subgroups. GSEA of the SNPs involved in 18 out of these 35 distinct genotype-lipidome relations revealed several statistically significantly enriched biological processes and pathways through which the identified SNPs can influence plasma lipid profiles.

Conclusions: Human plasma lipidome has 35 genetically distinct subject subgroups and are influenced by genetic variations in 3164 genes via several biological processes and pathways. The novelty of the work, in addition to its results, is that in the first time we apply the GWAS-PGMRA-GSEA method combination in unravelling the complex genetics of plasma lipidome. The results of the present study should however be replicated in other corresponding and independent data set.

Keywords: Genome wide association study, PGMRA, lipidome, GSEA, genetics.

The originality of this thesis has been checked using the Turnitin OriginalityCheck service, order 1606086384.

FOREWORDS

This Master thesis was carried out at the Faculty of Medicine and Health Technology, the Tampere University and the Department of Clinical Chemistry in Fimlab Laboratories during years 2019-2021 under Master's in information technology program in the Faculty of Information Technology and Communication Sciences, the Tampere University.

First and foremost, I want to thank my supervisor professor Terho Lehtimäki who introduced me to scientific research and to the field of bioinformatics and gave me this highly interesting Master Thesis topic. I sincerely thank my second supervisor PhD, bioinformatician, Pashupati Mishra for his valuable comments and advice concerning bioinformatic and statistics issues during this project. I also want to thank the personnel of Department of Clinical Chemistry and all study group members and especially PhD student Leo-Pekka Lyytikäinen for expertise help in GWAS studies and PhD student Binisha Hamal Mishra for helping in GSEA analyses and Professor Reijo Laaksonen and Zora Biosciences organizing the lipidomic analysis of Young Finns study.

I want to express my sincere gratitude to Academy professor Olli Raitakari and professor Mika Kähönen and all other excellent collaborators in Young Finns study team. Especially I want to thank professor Igor Zwir and Professor Robert Cloninger from University of Washington (St. Louis, USA). Special thanks go to their team member bioinformatician Coral De Val for his bioinformatic advice and help in the final stages of the project.

This work was financially supported by Foundation for the Promotion of Laboratory Medicine (for M.L), Tampere University Hospital Supporting Foundation (for M.L) and Finnish Foundation for Cardiovascular Research, Academy of Finland (grant 322098), the European Union's Horizon 2020 research and innovation programme under grant agreements No 848146 for To Aition and grant agreement 755320 for TAXINOMISIS; and Finnish Society of Clinical Chemistry.

Tampere, 15.6.2021

CONTENTS

LIST OF ABBREVIATIONS

CVD	cardiovascular disease
DNA	deoxyribonucleic acid
Eigenlipid	constructed from the 1st PC of a group of lipids
EU	European Union
GSEA	gene set enrichment analysis
GWAS	genome-wide association study
HDL	high-density lipoprotein
LDL	low-density lipoprotein
Locus	group of things
MetS	metabolic syndrome
NA	missing value
PC	principal component
gPC	genetic principal component
PCA	principal component analysis
PGMRA	phenotype-genotype many-to-many relationship analysis
RNA	ribonucleic acid
miscRNA	miscellaneous RNA
miRNA	micro-RNA, small single stranded non-coding RNA
lncRNA	long non-coding RNA
lincRNA	long intergenic non-coding RNA
snoRNA	short nucleolar RNA
snRNA	small nuclear RNA
rs number	SNP identification code
SKAT	sequence kernel association test
SNP	single nucleotide polymorphisms
TC	total cholesterol
TG	triglycerides
TUNI	Tampere University
URL	uniform resource locator, webpage address

1. INTRODUCTION

Cardiovascular diseases (CVDs) are significant causes of death worldwide – nearly 16,7 million people died of these diseases (1). In Finland, cardiovascular disease caused every fifth death (2). In EU, CVDs cause 40% of all deaths and are estimated to cost EU economy almost 200 billion euros a year (Figure 1). The ageing European population is heavily affected by CVDs as this malady develops over the lifetime and takes various forms with the potential culmination in CVD death.

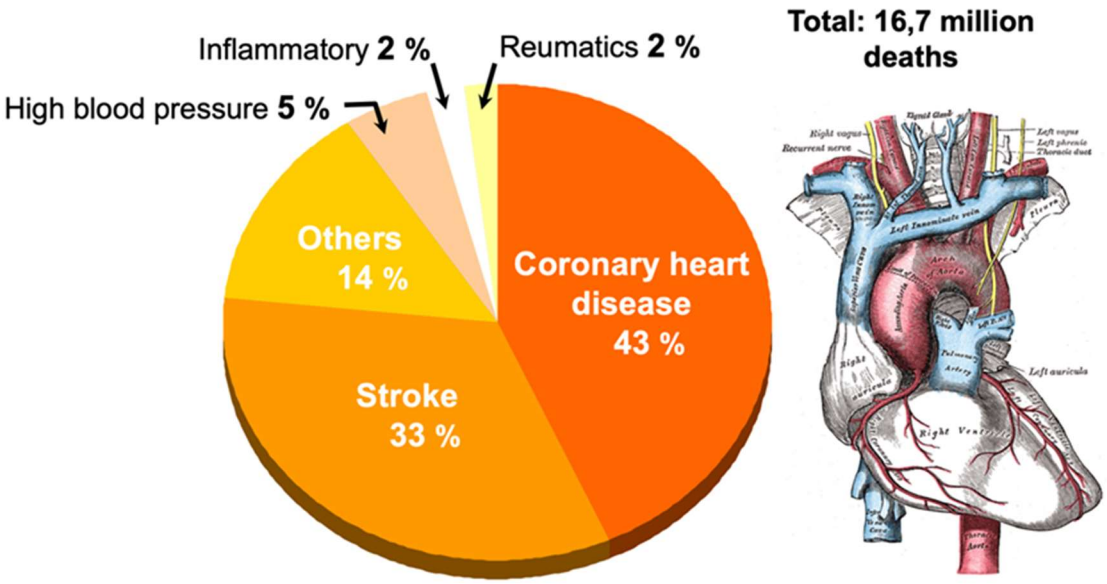

Figure 1. Pie chart of WHO cardiovascular deaths in 2015. Modified from ***(1)***.

Atherosclerosis (Figure 2), the underlying pathology behind majority of CVDs is a heterogeneous and multifactorial disease with roots varying from genetics to lifestyle factors. The rising prevalence of atherosclerosis is primarily due to the increase of the age of the population and of the occurrence of variable risk factors, such as hypertension, dyslipidemia, obesity and insulin resistance, clustering together as metabolic syndrome (MetS). Lipidomic has revealed, ceramides, phospholipids and other lipid species which are hypothesized to be associated with many of the central atherosclerosis processes such as lipoprotein aggregation, uptake of lipoprotein and accumulation of cholesterol

within macrophages, production of superoxide anions and expression of different cytokines and inflammation (3). Therefore, monitoring ratios and plasma concentrations of selected ceramides and phospholipids as well other lipid species may provide insights into the metabolic regulation of such events (3).

Plasma total cholesterol, LDL-cholesterol (LDL-C) and HDL-cholesterol (HDL-C) concentrations have been used for risk prediction of cardiometabolic disease outcomes.

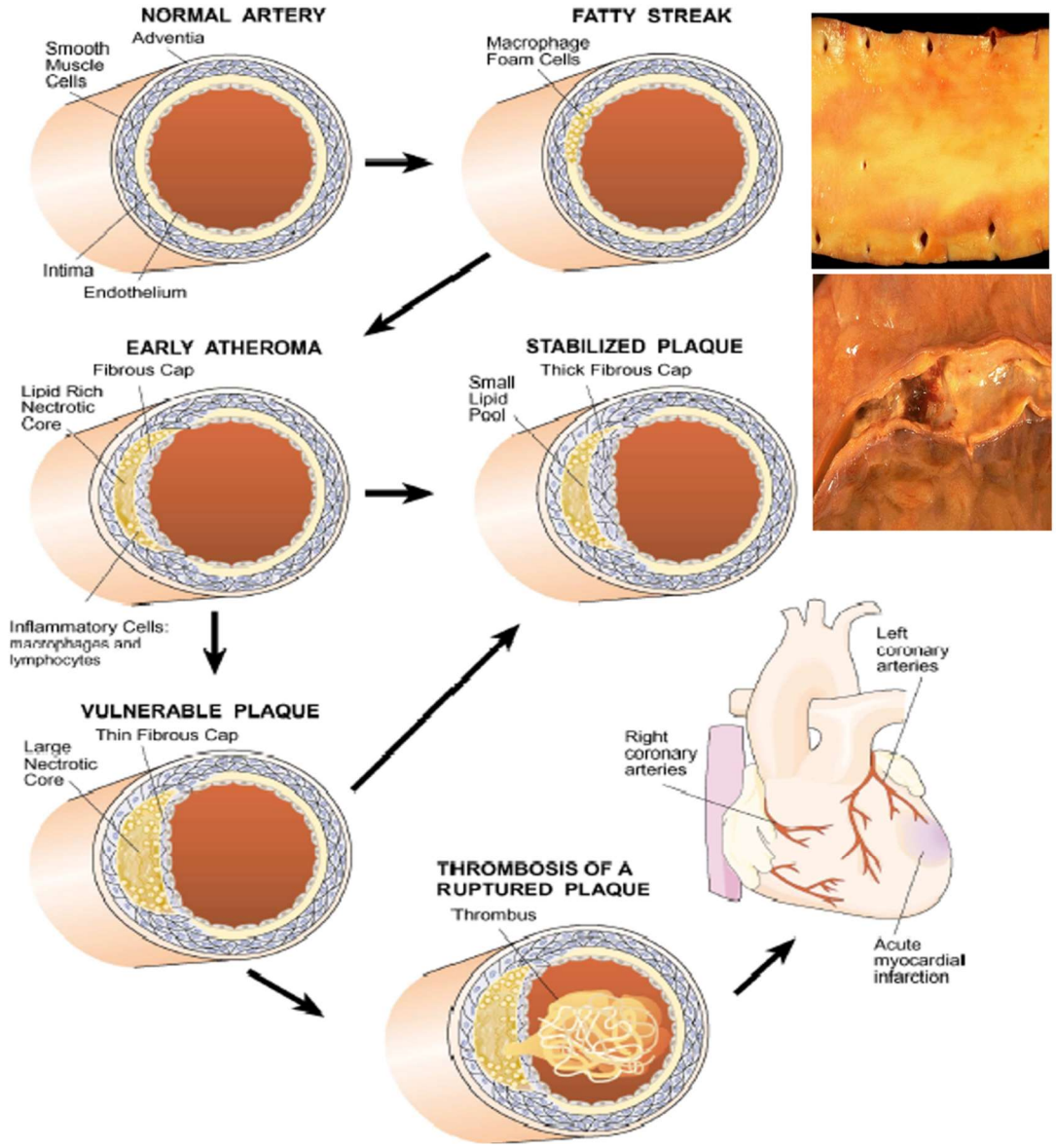

Figure 2. Shows the development of atherosclerosis and subsequent coronary artery disease. Modified from ***(98)***.

LDL-C has become the main therapeutic target in the management of patients with cardiometabolic outcomes, such as MetS, type 2 diabetes (T2D) and cardiovascular diseases with atherosclerotic etiology (*i.e*. coronary artery disease, stroke and peripheral artery disease). However, the increase in the number of patients with cardiometabolic

outcomes and identification of their residual cardiometabolic risks demands more detailed lipid analyses both for diagnostic purposes and for monitoring the efficacy of prescribed therapy.

In Finland, the relative proportion of aging people in population and expected life expectancy are both increasing and therefore the importance of prevention of CVDs is emphasized. Inheritable traits are known to increase the risk of coronary artery disease, but the role of heritable traits in the beginning and during development of atherosclerosis are not yet known well enough. As the gene-based personalized care-plans become more common, it is important to clarify the connection between genetic diseases such as atherosclerosis and the whole lipidome (hundreds of different fat molecules).

Data science is an important part of research that involves big genetic and molecular datasets. Development, evaluation and implementation of new analysis methods are in a key position, as the amount of data and capabilities of computers are increasing (4). The connection of clinical lipids as a risk factor to coronary artery disease is well researched (5). However, there is no previous study where the plasma lipidome has been studied using the new bioinformatic PGMRA (phenotype-genotype many-to-many relationship analysis) method as in the present work. Neither has genotype-lipidome clusters, obtained with such method, been studied for their connection to cardiometabolic diseases, such as obesity, metabolic markers, early subclinical or clinical atherosclerosis or type 2 diabetes.

Instead, there are several studies done using the classical GWAS method to identify the genetics of basic lipids (6-8), as well the genetics of nuclear magnetic resonance (NMR) measured plasma metabolites (9, 10).

Only one classical lipidome-wide GWAS study has been (11), in that study 141 lipid species were available for plasma lipidome analysis. It has been proposed that single nucleotide polymorphisms (SNPs) discovered by genome-wide association studies (GWAS) account for only a small fraction of the genetic variation of complex traits in human population (12). The remaining unexplained variance or missing heritability is thought to be due to marginal effects of many loci with small effects and has eluded attempts to identify its sources (13).

The new (PGMRA) method enhances the traditional genome-wide and linear regression based GWAS method by identifying relevant SNPs within subsets of studied population. The traditional GWAS identifies relevant SNPs across the entire population and therefore can miss the SNPs that are specific to only distinct sub-populations. (13-16)

The method is a good fit for studies where the studied expression of phenotype is multi-molecular, such as the plasma lipidome in this work, which includes the concentration measurements of 437 different quantified plasma lipid molecules. Using PGMRA method, multi-molecular phenotype data such as lipidome can be analysed against whole human genomes including around 40-400 million gene variants (i.e. 1000G/TOP-Med imputed GWAS data).

Clustering is another suitable method for multidimensional datasets, which allows improved statistical power as clustering reduces number of statistical tests (for multiple testing correction) from individual molecules to cluster of molecules, when compared to a more traditional subject-wise linear regression methods of analysis, like GWAS. The new method, PGMRA-pipeline, differs from previous linear regression method in that it is non-linear, allowing to study, in addition to the main genotypic effects, their interactions with other genes and many environment reflecting factors (like our personality or plasma lipidome). Our research team has already successfully utilized the new method for finding the “missing” genetics of temperament, character and personality and found 972 related genes, which together are a near complete explanation for the whole inheritance of temperament, character, and personality. (14-16)

In this thesis, my aim is, for the first time, to apply this complementary bioinformatic GWAS-PGMRA-GSEA platform for uncovering detailed lipidome-wide genetic architecture and through extensive bioinformatic gene annotation, define the role of this common variation in the regulation of human lipid metabolism and content of plasma lipidome.

2. REVIEW OF LITERATURE

2.1 Human genome and classification of genetic variation

The Trans-Omics for Precision Medicine (TOPMed) program seeks to elucidate the genetic architecture and disease biology of heart, lung, blood, and sleep disorders, with the goal of improving diagnosis, treatment, and prevention (17). TOPMed Imputation Server (TOPMed Imputation Server (nih.gov) is a database and reference panel of 97 256 deeply sequenced human genome samples and 410,323,831 genetic variants (381,343,078 SNVs and 28,980,753 indels) distributed across the 22 autosomes and the X chromosome (17).

Careful analysis of the human genome and its coding and non-coding variants in diverse populations continue to give more information on the mutational history and evolution of the human genome (18), increased information on structural coding and non-coding variations (19) and the development of new algorithms and techniques for detecting such rare and common structural variations (20).

2.2 Human lipidome and lipid classification

A lipid is a macro biomolecule used to store energy, serving as a signal both inside and outside a cell and used in cell membranes as a structural component, outside of biology and biochemistry, it also sees use in various forms of industry, such as food and cosmetics.

Lipids are characterized by needing non-water solvent and include fats, oils and hormones. Humans and other mammals have the own biosynthetic pathways for breaking down and synthesizing some lipids, but some essential lipids and fatty acids can only be obtained from their diet (21).

Lipids are an essential part of life, playing three major roles and comprise around one-third of all human body metabolites (22). Lipids are often divided into eight categories, which are then further divided into classes, subclasses and sometimes into subclasses of subclasses (23) (Figure 3). These classifications arise both in detection methods and data display tools (23), with machine learning being used to classify both known and novel lipids (24).

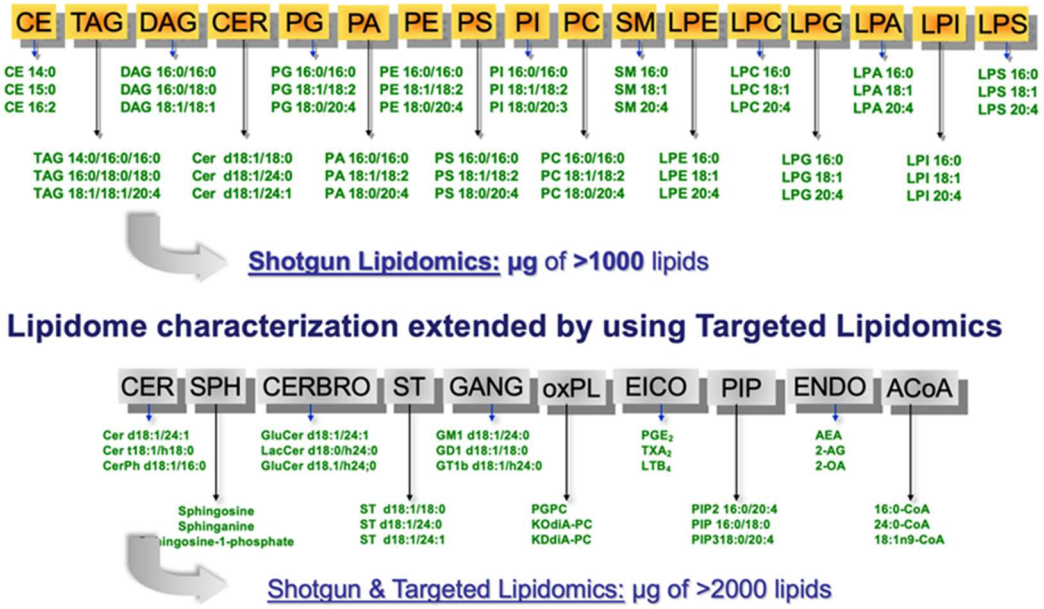

Figure 3. Shows an example in depth characterization of plasma lipidome using either shotgun or targeted lipidomic mass-spectrometry (LC MS/MS) based platforms. (Permission from R. Laaksonen)

The establishment of lipidomics standards (https://lipidomics-standards-initiative.org) and LIPIDS Initiatives (https://www.lipidmaps.org) are both advancing the standardization of lipidomics methodologies and annotations (22). The massive amounts of data in lipidomics and its analysis can easily allow for systemic errors to creep in, which has led to multiple strategies being developed to handle them (25). The lipidomic structure of the brain is an important area of study (26) and may have important connections to Alzheimer's disease (27) yet more new things are still being discovered about the development (28) and population (29) differences in lipids. In collaboration, we have widely used these methods to find novel lipidomics markers like ceramides and phospholipids species which have also been commercialized as Cardiovascular Event Risk Test (CERT) or in Finnish as HERTTA tests (https://zora.fi/?lang=fi) and can be used in the prediction of many cardiovascular disease outcomes and death.

These new lipidomic markers also improve CVD detection beyond their classical risk factors (30, 31). Plasma lipidomic architecture is also shared by subclinical markers of

osteoporosis and atherosclerosis (32), thus they can be used also as biomarkers for primary prevention purposes.

In circulation, the lipids are transported in different lipoprotein fractions to and from peripheral tissues to the liver and other organs (Figure 4).

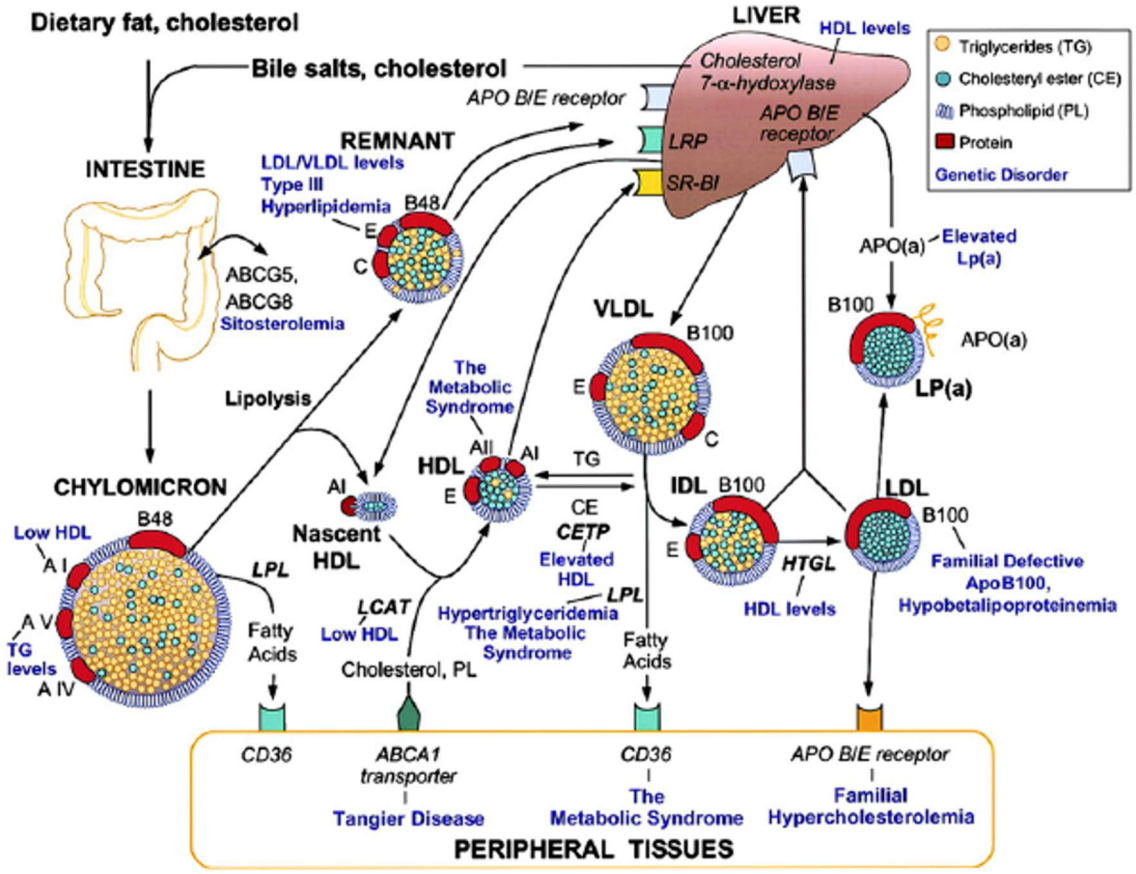

Figure 4. Lipoprotein pathways and lipid metabolism. ***(99)***. Abbreviations: ABCG5/8, ATP-binding cassette sub-family G member 5/8; ABCA1, ATP-binding cassette transporter A1; apolipoproteins AI, AII, AIV, AV, C, E, B48, B100; Lp(a), lipoprotein (a); LPL, lipoprotein lipase; HTGL, hepatic triglyceride lipase; LCAT, lecithin–cholesterol acyltransferase; CD36, cluster of differentiation 36 or fatty acid translocase; SRB1, scavenger receptor B1; LRP, LDL receptor related protein or apoB/E receptor; LDL, low-density lipoprotein; VLDL, very low-density lipoprotein; IDL, intermediate-density lipoprotein; HDL, high-density lipoprotein; apo B/E, apolipoprotein B/E receptor or LDL-receptor.

2.3 DNA genotyping array technology

Whole-genome genotyping provides an overview of the entire genome, enabling genome-wide discoveries and associations (33). Using high-throughput next-generation sequencing (NGS) and microarray technologies, researchers can obtain a deeper understanding of the genome, providing insight into the functional consequences of genetic variation. Array genotyping is a widely used tool that enables the assessment of several millions of genetic markers in thousands of individuals while also being cost-effective

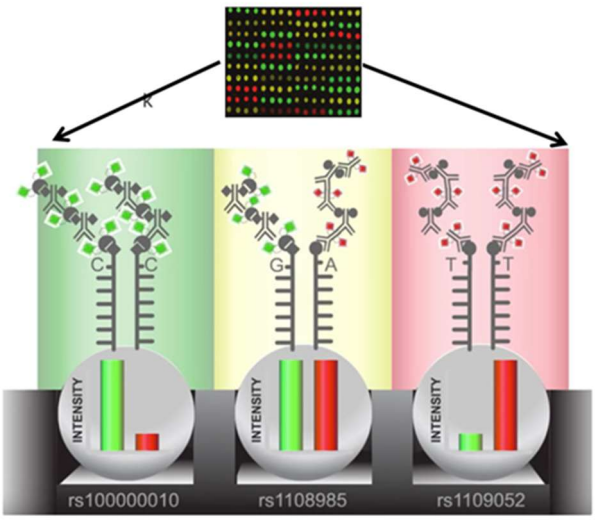

Figure 5. *In the chip array each genotype locus alleles are coloured by one of two fluorescent labelling dyes. Heterozygous and homozygous genotypes are differentiated by the relative intensity of these two colours*. Modified from ***(33)***.

(34). Recently, advancements have been made in creating more portable genotyping devices and supposedly cheap tests, the devices which include USB attachments and smartphones (35). The principle of microarray genotyping and individual well colour readouts is given in Figure 5. One chip array can contain millions of this kind of microwells, allowing millions of simultaneous genotyping.

Within the last year, multiple studies have been made on the subject (36-40). Some approaches have been made to assess various challenges, including repetitive DNA elements and accurate copy numbers quantification (41) and opportunities for possible use of long-term stored serum samples, that are generally understood to provide insufficient amounts of DNA (42).

2.4 DNA/RNA sequencing techniques

While whole-genome microarrays can currently interrogate over 4 million markers per sample, NGS-based whole-genome sequencing provides a comprehensive base-by-base method for interrogating the 3.2 billion bases of the human genome (33). Each technology offers unique advantages in price, data analysis, and throughput depending on study goals (33). Deep sequencing technologies have greatly improved the study of

transcriptomes and genomes, allowing the investigation of posttranscriptional mechanisms as RNA editing and splicing at unprecedented throughput and resolution (43). Advancements have also been made in creating novel methods for delivering nucleic acid probes to living cells to better image telomerase RNA (44).

Sequencing technologies are still improving (45), with massively parallel sequencing and now nanopores, a technique to sequence DNA without synthesizing or amplification, has seen successful use (46).

2.5 Lipid determination with LC-MS/MS based techniques

Liquid chromatography tandem mass spectrometry (LC-MS/MS) emerged as an analytical technology with a wide number of applications, an overview of its use in lipidomics is show in Figure 6. When compared to gas chromatography-mass spectrometry (GC-MS), LC-MS/MS is applicable to a larger number of analytes and easier to use (47). It has since seen a wide variety of use (36, 48-50).

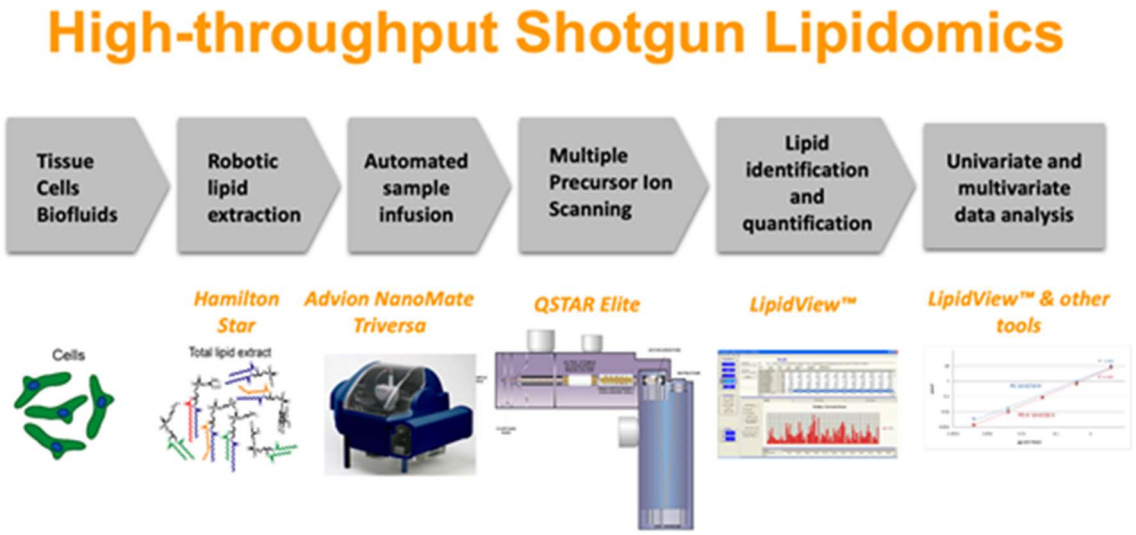

Figure 6. Shows a typical analytic pipeline of shotgun lipidomics. ***(100)***

LC-MS/MS operates with a combination of chromatography and multiple quadrupole mass spectrometers (Figure 7). First the chromatographic system separates the different components, the first quadrupole ionizes the molecules, selected molecular ions are then fragmented in the second one, selectively isolated by the third and final quadrupole for measurement by a detector (www.mccrone.com).

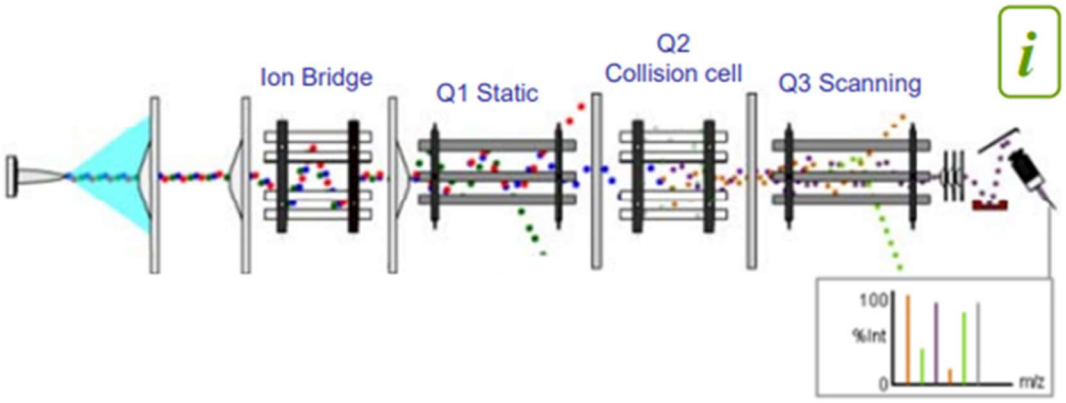

Figure 7. The principle of tandem mass spectrometry. ***(101)***

This series (11) of processes provides highly sensitive detection, which can be as sensitive as several parts per billion and are consistently in the part per million range.

2.6 Genome wide association study (GWAS)

A genome-wide association study (GWA study, or GWAS), with aliases whole genome association study (WGA study, or WGAS), is an observational study to find variants associoted with a chosen trait, performed with a genetic variants genome-wide and from different individuals to find possible associations. GWA studies typically focus on associations between single-nucleotide polymorphisms (SNPs) and traits like major human diseases (https://www.ebi.ac.uk/gwas/). A single-nucleotide polymorphism (SNP, pronounced snip) is a single nucleotide variation occurring in a DNA sequence, these differences can exist between different members of the same species or in paired chromosomes in a single subject, the different nucleotides are adenine (A), thymine (T), cytosine (C), or guanine (G)(https://isogg.org/wiki/Single-nucleotide_polymorphism). (Figure 8).

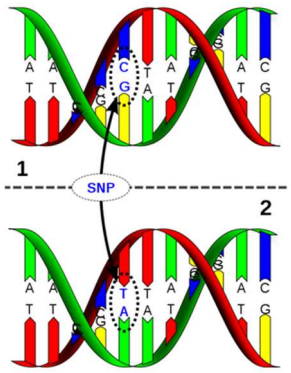

Figure 8. DNA-SNP model ***(102)***

An example illustration of a Manhattan plot (Figure 9.) depicting the chromosome spread of SNPs found in GWAS analysis of the AcylCarnatine 18:2 lipid, each dot represents a SNP, with the X-axis showing genomic location and Y-axis showing association level (P-values).

GWA studies compare the DNA of control participants to cases DNA having varied phenotypes for a particular trait or disease. These participants may be cases (people with a disease) and controls (without the disease) (51), or they may be people with different values (phenotypes) for a particular trait, for example high vs. low LDL-cholesterol (52).

GWA studies classify participants by the clinical manifestation, called phenotype-first, instead of their genetic manifestation. Participants give samples of DNA that are then analysed using SNP arrays. If an allele, a type of the variants for genetic traits, is found to be more frequent, it is called to be associated with the phenotype (clinical trait or disease being investigated. SNPs found this way are associated and are then marked in the regions in the human genome that could influence the disease risk. (https://en.wik-

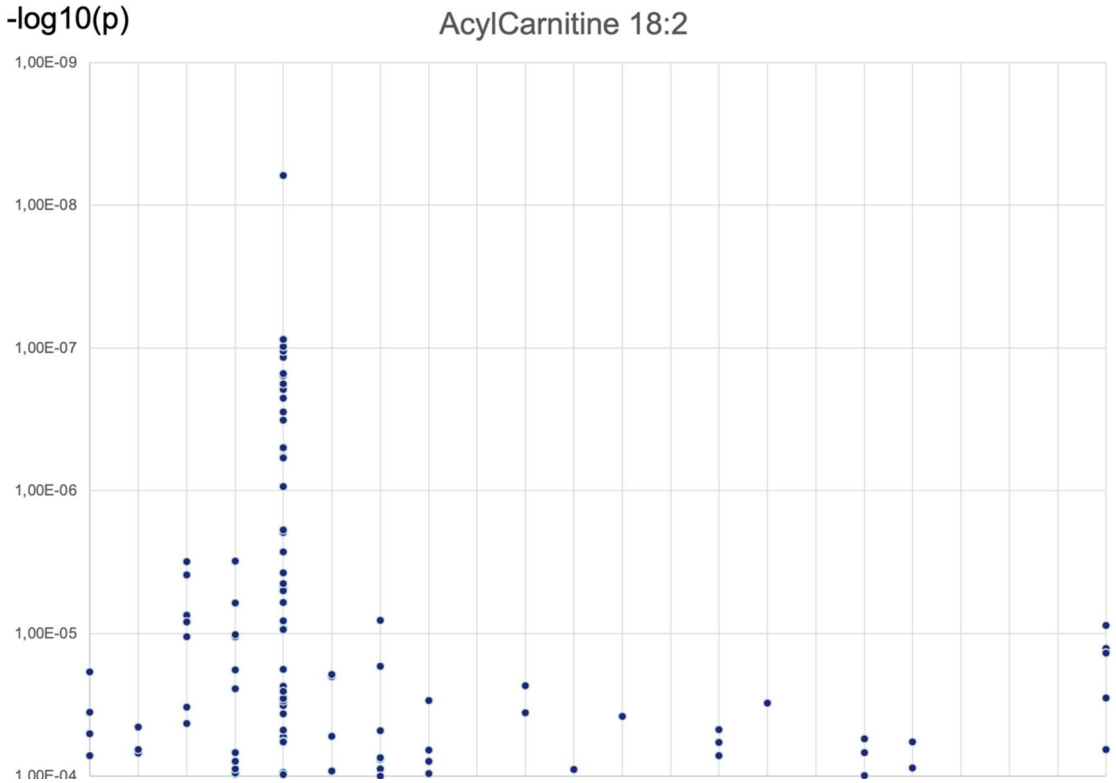

Figure 9. Manhattan plot of SNPs for a lipid in the GWAS analysis. The x-axis is chromosome, y-axis is p-value in scientific notation.

ipedia.org/wiki/Genome-wide_association_study). Since the discovery of the human genome (53) and the correlation between nearby genetic variations (54), many advances have been made. For example, with the extension of the number of genotypes since first HapMap database (55), advancements in statistical methods (56), and since the first genome wide association study (57) thousands GWAS studies for many different traits have been performed.

A postulation of GWAS is that individuals are predisposed to complex diseases by carrying multiple alleles with small independent effects that can combine, manifesting as, among others, coronary artery disease.

The ability of GWAS is limited in finding rare novel variants that occur in just few families, as compared to more laborious sequencing techniques which allows also the discovery of these novel and rare variants (58).

GWA studies identify variants in DNA and SNPs, this can find associations with diseases and can help narrow down causalit y among genes. The first successful GWA study was published in 2002 and it studied myocardial infarction (59). After that, the 2021-06-08 version of the GWAS Catalog contains 5106 publications and 258738 associations (https://www.ebi.ac.uk/gwas/).

2.7 Biostatistical analysis methodology

PGMRA is a web server-based machine learning program, that uses a generalised factorization method to identify SNP sets and phenotype sets from GWAS data and uncover relations among them. PGMRA avoid possible bias by and distinguishes itself by (12): (i) unsupervised machine learning, a grouping strategy using no previous knowledge and not consider the status of subjects in the data to form datasets; (ii) subjects, SNPs and features can belong to more than one relation; (iii) SNPs within a SNP set can be located anywhere in the genome; (iv) dimensionality of phenotype features is not reduced; (v) have no predefined SNP sets, phenotype sets or relations; (vi) relations between phenotype and SNP sets are found using the probability of subject intersection, without considering disease status; (vii) disease risk is estimating in a unbiased way by adding disease state afterwards into each relation and analysing the frequency of cases, relatives and controls (12, 13).

3. THEORY

3.1 Transcription, translation, and gene expression regulation

Cells make all the proteins needed by the body with transcription and translation using information stored in DNA. DNA and RNA are built from four bases (C-G, A-T in DNA and A-U in RNA as shown in Figure 10). The DNA pieces for a gene are copied in the cell nucleus as messenger RNA (mRNA). The DNA template information is carried by the mRNA to the cytoplasm. Information brought by the mRNA sequence is used to make proteins, i.e. in cytosol the mRNAs are then “translated” into proteins by large cell machines called ribosomes. In translation process, transfer RNA (tRNA) brings amino acids one-by-one to the ribosome as the mRNA proceeds through the ribosome in a matching codon sequences with amino acids specific tRNAs (multiple codons can code the same amino acid). The joining of tRNA to the mRNA strand codons (nucleotide triplets) creates a growing protein chain by adding amino acids and joining together in a sequence until mRNA sequence ends to termination codon (stop codons UAA, UAG or UGA) and the ribosome will release the protein Figure 10.

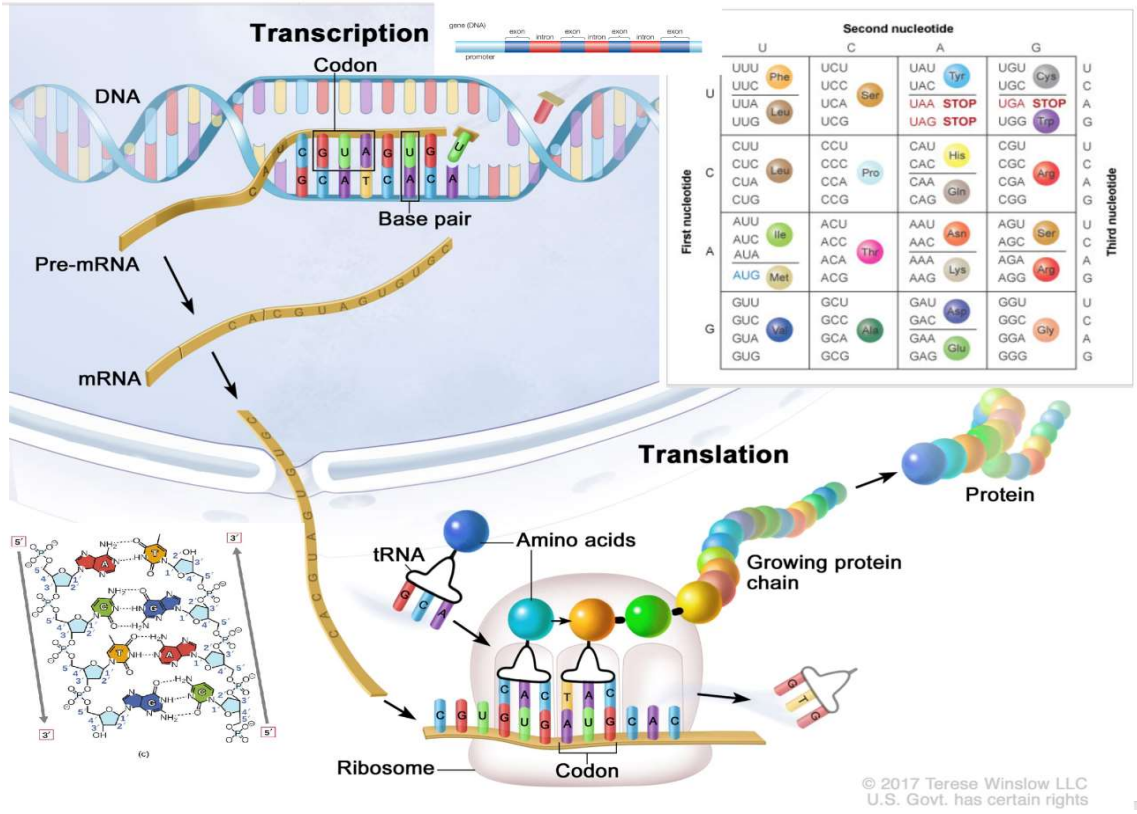

Figure 10. Gene transcription and translation process and a codon translation table. Modified from ***(103)***. Abbreviations: mRNA, messenger RNA; tRNA, transfer RNA.

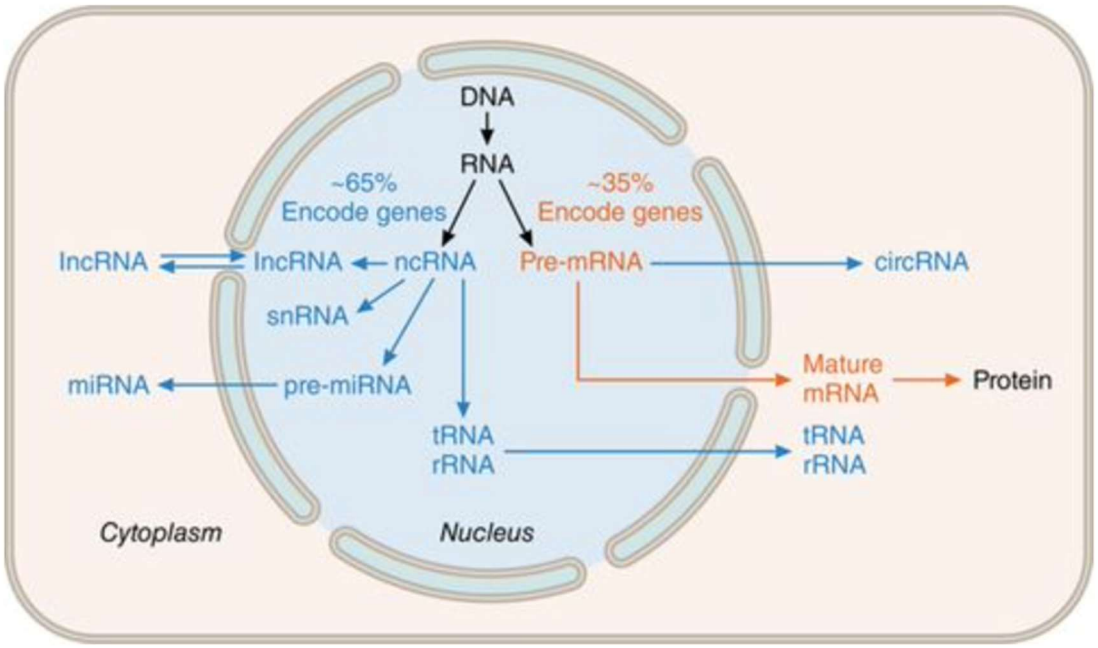

Figure 11 Transmission of genetic information. Primary transcripts give rise to protein-coding messenger RNAs (mRNAs) and a large variety of noncoding RNAs. mRNAs are further translated into protein. However, most noncoding RNA molecules can be broadly subdivided into long noncoding RNA (lncRNA), circular RNA (circRNA), microRNA (miRNA), transfer RNA (tRNA), ribosomal RNA (rRNA), small nuclear RNA (snRNA), etc, which act in the nucleus or cytoplasm ***(60)***.

The transcriptome-wide massive parallel sequencing has led to the discovery that most parts of mammalian genomes are actively transcribed into RNA, but only 2% to 3% of the genome, is further translated into protein (60). The majority of RNAs represents a heterogeneous group of noncoding RNAs comprising ribosomal RNA, transfer RNA, lncRNA, microRNA (miRNA), circular RNA, and other small RNA molecules (Figure 11-13) (61). Many lncRNAs, 5'UTRs, and pseudogenes are translated, and some are likely to express functional proteins as shown by Ji, Z et al. using a new bioinformatic method to analyse ribosome profiling data, found that 40% of lncRNAs and pseudogene RNAs expressed in human cancer related cells are translated (61). The translation efficiency of cytoplasmic lncRNAs in these cancer cells are nearly comparable to that of mRNAs, suggesting that cytoplasmic lncRNAs are engaged by the ribosome and translated. While most peptides generated from lncRNAs may be highly unstable byproducts without function (61), but at least a small number of the peptides translated have changed little over the course of evolution (61).

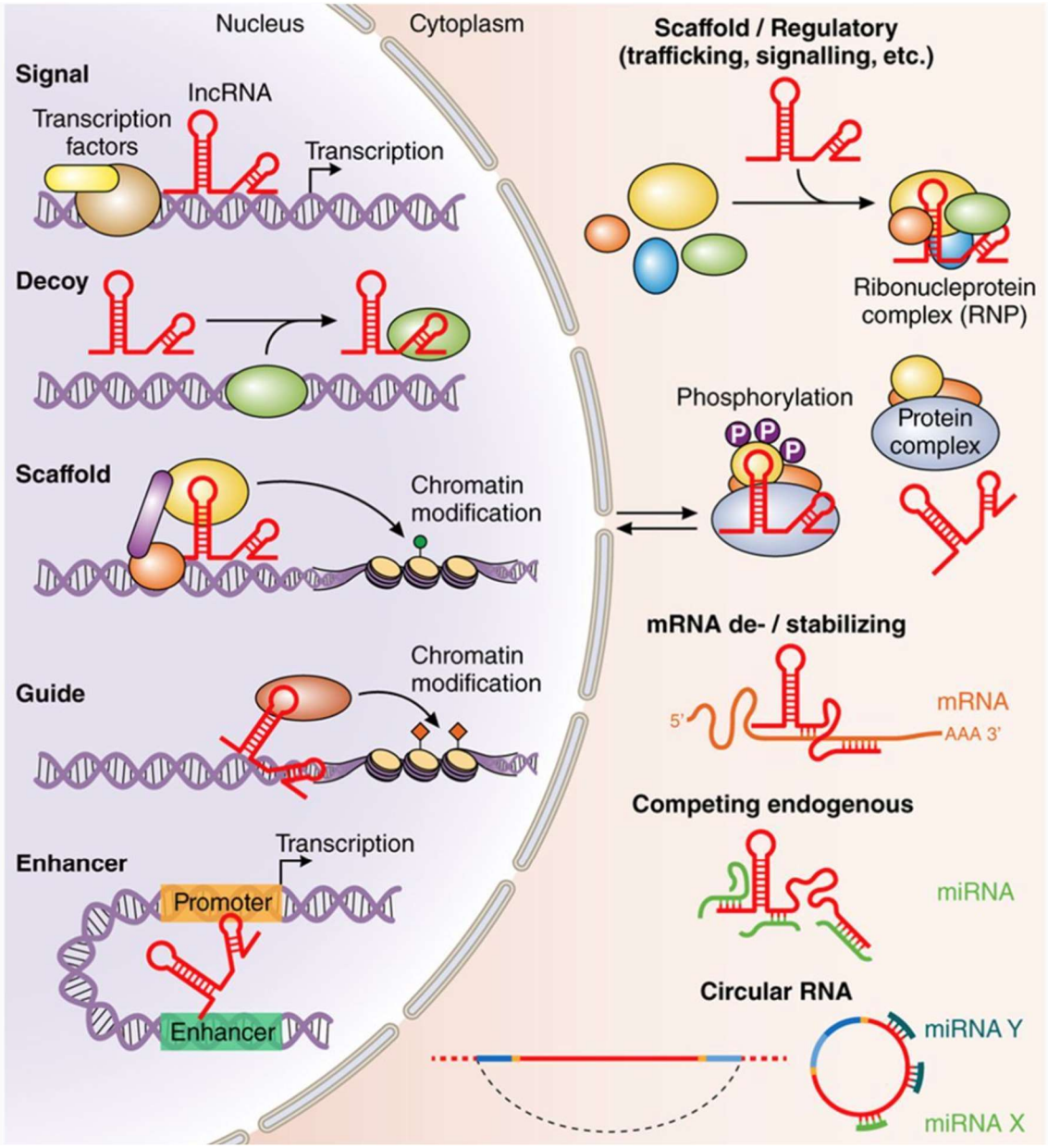

Figure 12 **LncRNAs have versatile modes of action in the nucleus and the cytoplasm.** Nuclear-localised lncRNA regulate gene expression in various modes such as in a response to stimuli (signal), sequester transcription factors/protein complex (decoy), bring together multiprotein complexes (scaffold) or guide transcription factors/protein complex to specific target site (guide) to activate or repress transcriptional and induce chromosomal looping to increase association between enhancers and promoter region (enhancers). Cytoplasmic lncRNAs (linear or circular) can stabilise ribonucleoprotein complexes, regulate mRNA stability or sponge miRNAs, thus controlling translational events. Further regulatory functions may involve protein signalling (e.g. phosphorylation status) and trafficking. LncRNA indicates long noncoding RNA, and miRNA, microRNA. ***(60)***.

LncRNA was observed to not have many homologues if the species are separated by more than 50 million years, suggesting that new lncRNA appear with high frequency (62), with multiple different mechanisms, including duplication, loss of coding potential of pro-

tein-coding genes, formation of new transcriptional units following a transposable element (TE, transposon, or jumping gene) integration, mutations that stabilize cryptic transcripts by enhancing splicing and exaptation of sequences that were previously non-coding (62).

Genomic and transcriptomic analyses are revealing that as much as 85% of the human genome is transcribed (63). The role of lncRNA as a regulator of transcription has been gaining traction, with many proposed mechanisms for accomplishing this, including signal, decoy and guide RNAs (Figure 12)(63).

LncRNAs are molecules defined by lack of computationally, protein-coding potential, and come in many diverse classes. With current RNA-sequencing and epigenomic methods and technologies, the discovery rate of new lncRNA genes is easily outpacing the efforts to characterize them, due to, among other reasons, the many experimental difficulties in studying lncRNA in comparison to protein-coding ones (64).

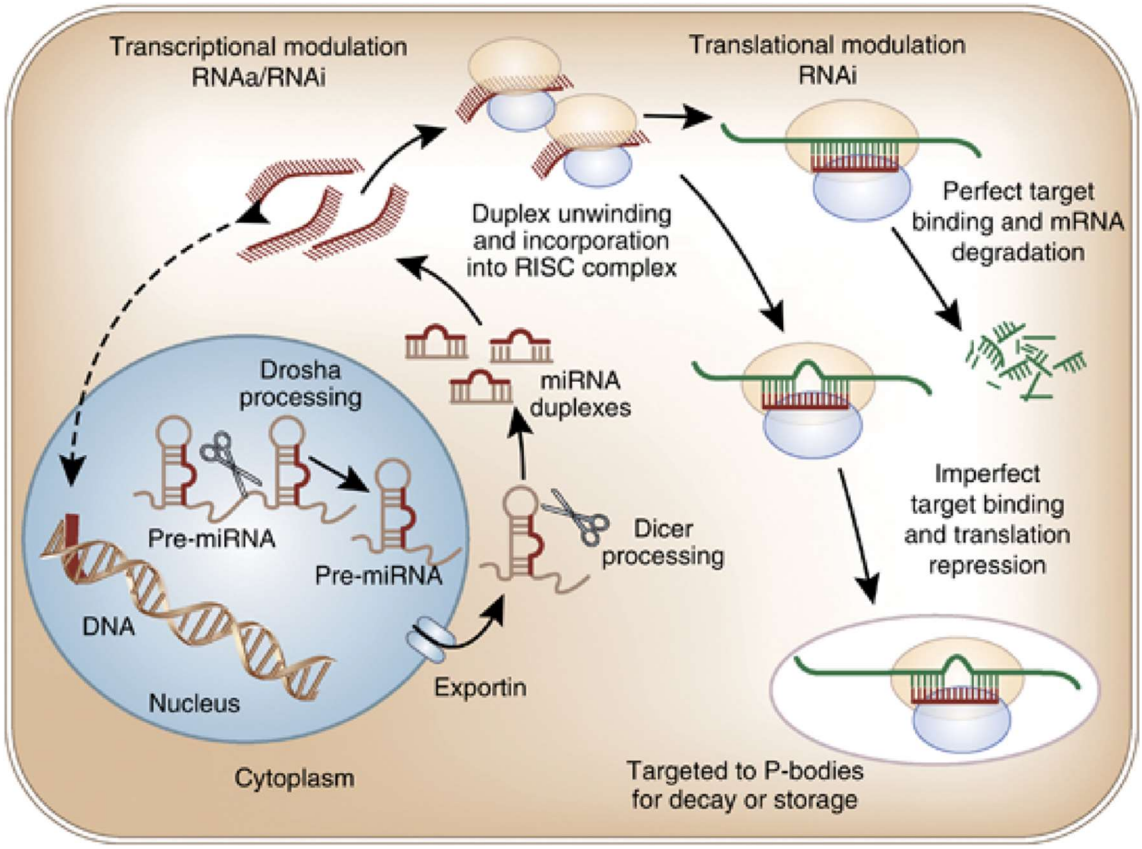

Figure 13 MicroRNA (miRNA)—biogenesis and function. The biosynthesis of miRNAs, as well as their activity in translational repression (RNA interference (RNAi)) and transcriptional modulation (RNAa or RNAi), is represented diagrammatically. Pre-miRNA, precursor miRNA; RISC, RNA-induced silencing complex; RNAa, RNA activation ***(65)***.

MicroRNAs (Figure 13) are endogenous, short (20–22 nucleotides in length) non-coding RNAs that control the expression of many genes. MicroRNAs have been associated with several physiological activities, such as tissue development, lipid metabolism, cell differentiation, apoptosis, and stem cell division (65). In humans, approximately 70% of known miRNAs reside in non-protein-coding regions of the genome, while the remaining 30% are transcribed from the intergenic regions (65).

3.2 Theory of GWAS

Genome wide associating study, GWAS in short, is a method of comparing the genomes between different groups such as between sick and healthy, case-control groups, and attempts to identify SNPs, single nucleotide polymorphisms, that occur more often in those with a disease than those without (https://medlineplus.gov/genetics/understanding/genomicresearch/gwastudies/). The analysis is performed using software such as PLINK v1.90b3q 64-bit (https://www.cog-genomics.org/plink2) (66), there is also a newer v2.x of PLINK which is computationally faster.

GWAS is based on linear regression model which is a statistical technique to investigate association between one or several explanatory variables such as SNPs and an outcome variable such as a disease. Regression can be used in both explanatory and predictive models, a common example being the connection between sale of fans and alcohol use to smoking. (67)

Multivariable regression can be shown using the following formula.

$$Y = a + b_1X_1 + b_2X_2$$

Where Y is the factor to be explained (outcome), let a be the default constant, X_1 and X_2 the explaining variables and b_1 and b_2 their regression multipliers (coefficients).

3.2.1 Limitations

GWAS must adopt a high level of significance to account for multiple tests, lowering the detection power due to high signal thresholds, leading to GWAS missing a large amount of proposed heritability in the genome. This leads to that GWAS being unlike to ever identify all genetic determinants for complex traits. The clinical value of GWAS is also limited due to missing heritability and that genetic screening of an entire populations is unfeasible (68).

3.3 Theory of phenotype-genotype many-to-many relationship analysis

Now PGMRA functions without open-source code as a web server, implemented using PHP (PHP: Hypertext Pre-processor), it uses a bash script to communicate with the different biclustering and post processing implementations (13). The biclustering methods are written in C but are used through their respective Perl or R wrappers. The PGMRA implementation uses several R-project packages (12):

- pheatmap for the heatmap graphs

- latticeExtra, akina, tgp, animation and plotrix for 3D risk graphs

- rpart and rpart.plot for classification trees

- SKAT for the statistical analysis of SNP sets

- biclust, fabia and BicARE for the Cheng&Church, FABIA and FLOC biclustering methods, respectively

3.3.1 Clustering

Clustering is a proven approach to the analysis of large data sets, like tens of thousands of gene expressions, it is also known as one-way clustering, in which gene data can be gathered based on their profiles.

The purpose of clustering is to group elements of the data, trying to optimize either homogeneity, where the groups information is as same as possible or separation, where the information of the groups is as dissimilar to each other as possible. Clustering has brought significant results in genetic research (69-72), but one of its assumption is that that all members of an elements share a similar function.

Clustering is performed separately either on the rows or columns of a data matrix. It is called biclustering when performed on both dimensions at the same time, looking for sub-matrixes that can overlap. Fuzzy, possibilistic and probabilistic clustering comprise the conceptual field of so-called soft clustering methods (73).

3.3.2 Biclustering

This technique clusters both rows and columns simultaneously, as opposed to clustering only rows or only columns. Let Y be a m × n matrix. The goal of biclustering now is to find subgroups of rows and columns which are as similar as possible to each other and as different as possible to the rest (74, 75).

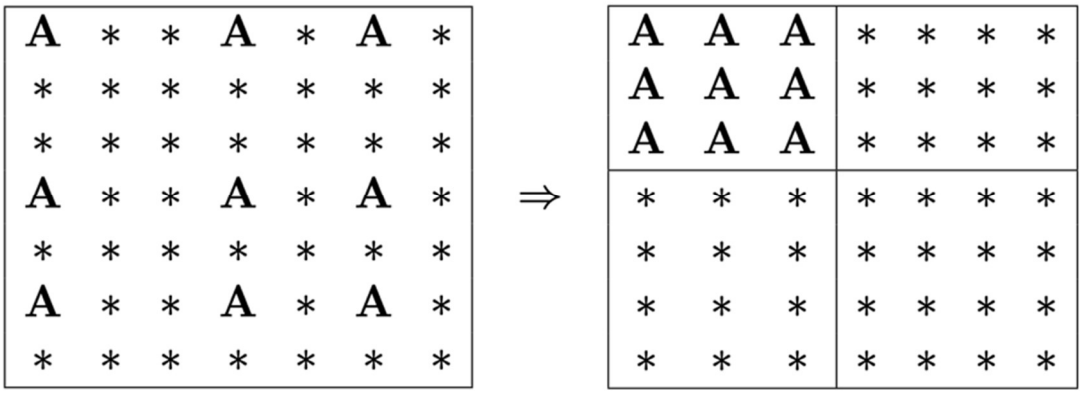

Figure 14. *Biclustering finds objects (subjects) and variables with a similar value A and reports them as a bicluster (submatrix*). ***(104)***

This basically comes down to clustering on both the row and column dimension simultaneously and while clustering methods on one dimension derive a global model, biclustering algorithms will produce a local model (75).

For example, in clustering algorithms each row in a cluster is defined over all the columns, however a row in a bicluster is selected using only a subset of columns. Going back to the matrix Y, this corresponds to looking for submatrices with a high similarity of elements. This submatrix above is what is called a bicluster (Figure 14)(75).

Biclustering functions by attempting to minimize the mean squared residue score, but this can lead to simple one gene, one condition sets as that more easily satisfy the assumptions. A way to overcome this is, defining A(I,J) as a δ - bicluster if H(I, J) ≤ δ. Assuming that this threshold indicates strong similarity, you can confine the search to find large δ-biclusters (76).

When searching for matrixes, the algorithm first starts by removing rows and columns which have the greatest H values, so that H (I, J) < δ, then adding rows and columns until it is no longer possible to do so without H exceeding δ. The remaining sub-matrix is declared a bicluster if the matrix is empty then none were found. The algorithm is completely deterministic, finding the same biclusters every time, thus for it to find more than one, the previous bicluster must be masked, this is done by filling the positions of the bicluster with random values, these new values are unlikely to form patterns and are first candidates for removal on subsequent runs of the algorithm (77).

The BiclustGUI R Package - 1.1.3 developed by De Troyer Ewoud serves an easy user-friendly interface in R Commander for the users to different types of biclustering analyses and their visualisation (see below) (75).

3.3.3 NMF- nonnegative matrix factorization

Nonnegative matrix factorization (NMF) is a technique for acquiring information from high-dimensional data, it has been successfully used in, among others, signal processing, face recognition and text mining. NMR theory development has resulted in various applications with multiple algorithms and methods on commercial platforms, one if which is R Commander (Figure 15). Free applications tend to have hurdles in their usage, such as requiring programming skills, limiting their use in the wised research community (77).

NMF, also known as positive matrix factorization and nonnegative matrix approximation (78), is a technique with wide application in both clustering and classification. It is a

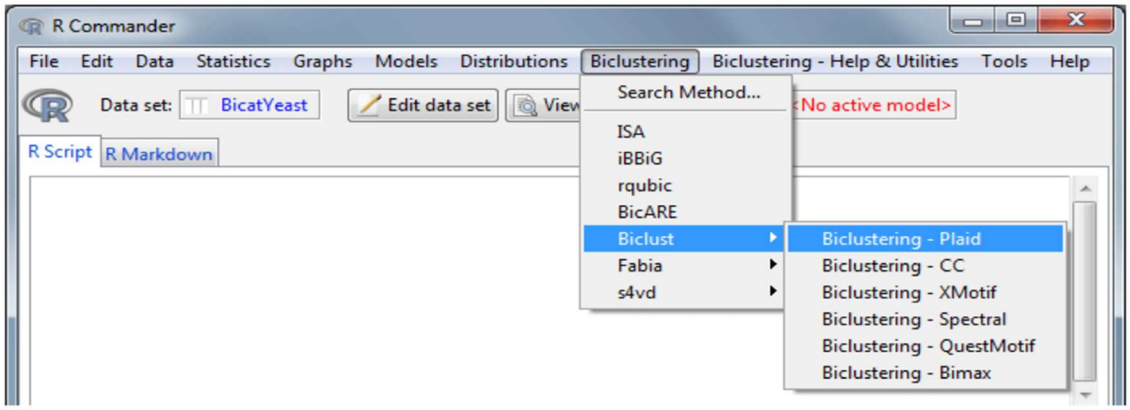

Figure 15. Screen capture of R-commander biclustering options.

method for approximating nonnegative high dimensional data in a low dimensional space (79). Figure 16 shows an example of a biclustering algorithm. Below is a quoted review of the nonnegative matrix factorization.

Given a nonnegative matrix X ∈ Rm × n, each column of X is a data sample. The NMF algorithm aims to mate this matrix by the product of two nonnegative matrices U ∈ Rm × k and V ∈ Rk × n (80). To achieve this, the following objective function is minimized:

$$O = ||X - UV||_F^2 \; s.t. \; U \geq = 0, V \geq = 0 \qquad (1)$$

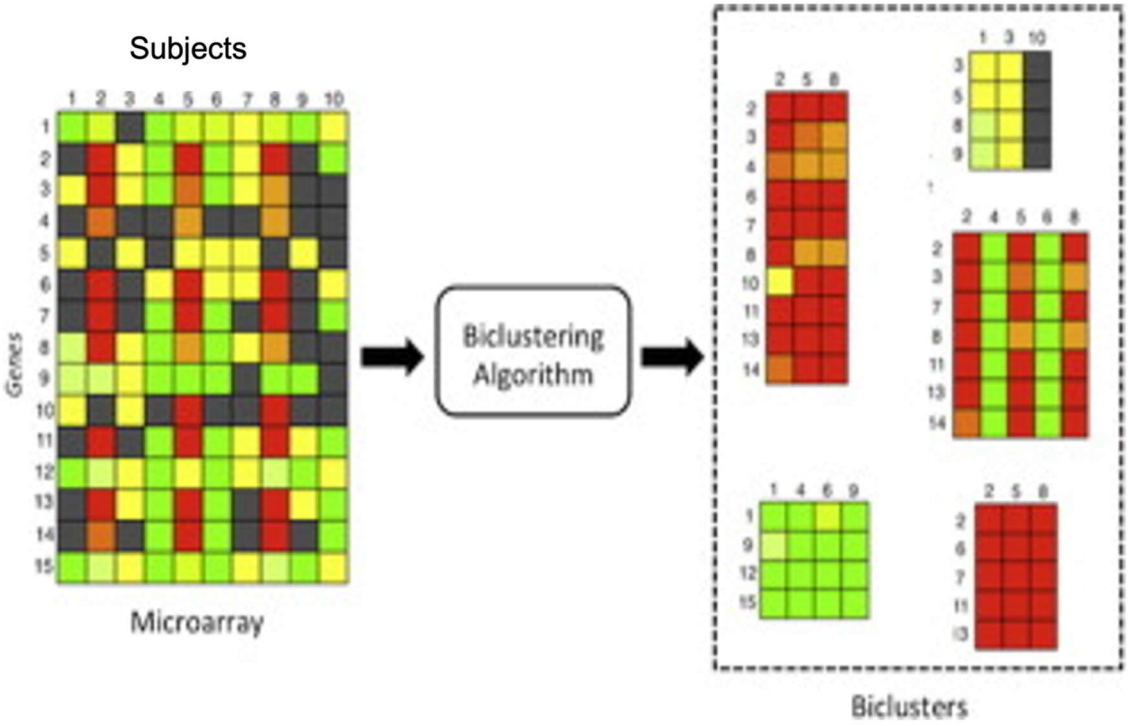

Figure 16 Shows how biclustering algorithm works in principle by picking up groups with similar features (= subject x gene SNP biclusters). ***(105)***

where $||.||_F$ denotes the Frobenius norm (79).

saNMF is an essential part of PGMRA pipeline which is a commercial platform (no open code available). Our objective is to apply this method in the biostatistical analysis of lipidome- and genome-wide omic data and prove the suitability of fuzzy FNMF method as a part of solving complex, especially non-linear biological problems. For that purpose, Gaujoux R and Seoighe C (2010). “A flexible R package for nonnegative matrix factorization.” (77) have developed a package for the R/BioConductor platform. The package ports public code to R and is structured to enable users to easily modify and/or add algorithms. It includes several published NMF algorithms and initialization methods and facilitates the combination of these to produce new NMF strategies. Commonly used benchmark data and visualization methods are provided to help in the comparison and interpretation of the results (77). Documentation, source code and sample data are available from: R Latest stable release from CRAN (77): http://cran.r-project.org/package=NMF. PGMRA pipeline uses Factorization method (NMF) proposed and termed Fuzzy NMF (FNMF) and first introduced by our collaborators (13) as a web-based server. FNMF allows overlapping among sub-matrices and detection of outliers and is implemented as the default option.

3.3.4 Sequence kernel association SKAT-test

Sequence kernel association test (74) or SKAT is used in PGMRA pipeline to test for associations between SNP sets and phenotypes. SKAT is a kernel association test which quantify the similarity between pairs of subjects and tests if this similarity is associated with trait similarity. Kernel methods relate covariate outcomes in terms of pairwise similarities and are related to distance-based multivariate regression methods (81, 82), with differences in the way the p-value is computed, kernel methods calculate it analytically while distance-based methods mainly calculate by permutation methods. Kernel methods are a family of flexible representative approaches for aggregative variant association analysis and advances in these methods have extended its use to a diverse array of study designs and outcome types (82).

3.3.5 Hypergeometric test

A hypergeometric distribution describes a probability of drawing a set number of 'good items' from a collection of 'good' and 'bad' items in a certain number of draws, more formally, the probability of drawing exactly k desired people from a population of N with desired attribute K.

$$\Pr(X = k) = f(k; N, K, n) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}, \quad (2)$$

The hypergeometric test checks for over-representation, where the p-value is the probability of randomly drawing k or more successes in n draws, or under-representation, where the p-value is the probability of drawing k or fewer successes (83).

4. OBJECTIVES

To perform lipidome-wide GWAS analysis first between 437 separate plasma lipid species and 546,677 genotyped SNPs and then between five separate eigenlipid classes derived from PCA analysis (i.e. used as surrogate markers for these 437 lipids) and 546,677 genotyped SNPs.

The specific objectives of this work were:

1. To use complementary PGMRA analysis for lipidome-wide GWAS uncovering the detailed genetic architecture of a human plasma lipidome by studying the significance of nominally associated ($p < 5x10^{-4}$) 18,370 SNPs revealed by lipidome-wide GWAS (= 437 separate GWAS) using PGMRA-analysis to find the true significance of these hidden “grey area” SNPs in the regulation of plasma lipidome.
2. To study weather PCA analysis can be used to find surrogate markers called hereafter eigenlipids for the whole lipidome, to lower the time-consuming computational load as compared to lipidome-wide GWAS analysis.
3. To discover new genetic base for lipidome classification and study the biological significance of significant lipidome associated SNPs set by using gene set enrichment analysis.

5. STUDY SUBJECT AND METHODS

5.1 Study subjects and risk factor follow-up data

5.1.1 The Cardiovascular Risk in Young Finns Study (YFS)

The Cardiovascular Risk in Young Finns Study (YFS) is a Finnish longitudinal general population study on the evolution of cardiovascular risk factors from childhood to adulthood (84). The study began in 1980, when 3,596 children and adolescents aged 3, 6, 9, 12, 15 and 18 years were randomly selected from five university hospital catchment areas in Finland (31). In 2007, 2,200 participants aged 30-45 years attended the 27-year follow-up. Of these subjects, we included to final analysis those for whom the GWAS genotype data and the lipidomic parameters and covariate data were available. Therefore, 1,426 participants contributed to the final association analyses of GWAS and plasma lipidome profile and final PGMRA analysis. Figure 17 shows the general description of The Young Finns Study 40-year-follow-up and its three-generation study. The more general aims of the study are presented in project internet pages (https://youngfinnsstudy.utu.fi) (84).

5.1.2 YFS ethical issues

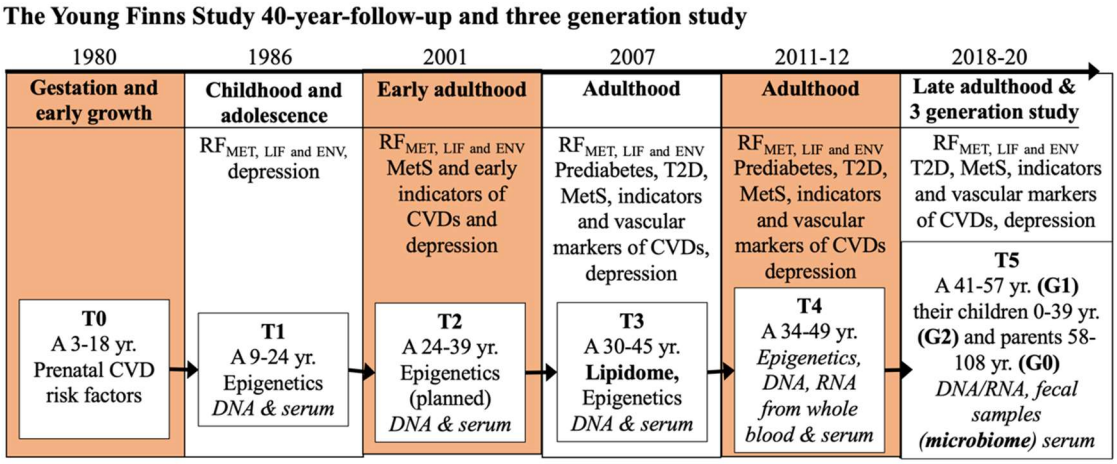

Figure 17. Young Finns study follow-up study timepoints (T0-T5) between 1980-2020. Abbreviations:T2D, type 2 diabetes; MetS, metabolic syndrome; RF, risk factor; MET, metabolic; LIF, life-style; ENV, environmental; CVD, cardiovascular diseases; G, generation G0,G1 and G2.

The YFS was approved by the 1st ethical committee of the Hospital District of Southwest Finland and by local ethical committees (1st Ethical Committee of the Hospital District of Southwest Finland, Regional Ethics Committee of the Expert Responsibility area of Tampere University Hospital, Helsinki University Hospital Ethical Committee of Medicine, The Research Ethics Committee of the Northern Savo Hospital District and Ethics Committee of the Northern Ostrobothnia Hospital District) (10, 31). The study protocol of each study phase corresponded to the proposal by the World Health Organization. All present subjects gave written informed consent, and the study was conducted in accordance with the Helsinki declaration. At prior follow-ups of YFS, informed consent of every participant under the age of 18 was obtained from a parent and/or legal guardian (10, 31).

5.2 Genotyping for GWAS

From whole blood samples of YFS the genomic DNA was extracted from peripheral blood leukocytes using a commercially available kit and Qiagen BioRobot M48 Workstation according to the manufacturer's instructions (Qiagen, Hilden, Germany) (85). Genotyping was performed at the Welcome Trust Sanger Institute using a custom-made Illumina Human 670k BeadChips. Genotypes were determined using the Illuminus clustering algorithm. Fifty-six samples failed the Sanger genotyping pipeline QC criteria (i.e. duplicated samples, heterozygosity, low call rate, or Sequenom fingerprint discrepancies)(85). Three samples were removed due to a low genotyping call rate (< 0.95) and 54 samples were excluded for possible relatedness (pi.hat > 0.2). A total of 11,766 single SNPs were excluded based on the variation from Hardy-Weinberg equilibrium (HWE) test ($p \leq 1.0 \times 10^{-6}$), 7,746 SNPs failed the missingness test (call rate < 0.95) and 34,596 SNPs failed the frequency test (MAF < 0.01). After quality control there were 2,443 samples and 546,677 genotyped SNPs available for further analysis (85).

5.3 Lipidome-wide analysis with mass-spectrometry

Lipids in the plasma were separated using the previously described method (86)**.** Lipidomic analysis were performed before with an unfocused mass-spectrometry (LC-MS/MS) method from the plasma (Zora Biosciences Oy, Espoo, Finland). The method is described in more detail in (87).

The analysis was made with a hybrid-triple quadrupole/linear ion trap utilizing mass-spectrometry (QTRAP 5500, AB Sciex, Concord, Canada), equipped with ultra-high-performance liquid chromatography (Nexera-X2, Shimadzu, Kyoto, Japan) (31). Chromatographic separation of the lipidomic screening platform was performed on an Acquity BEH C18, 2.1 × 50 mm id. 1.7 µm column (Waters Corporation, Milford, MA, USA). The data were collected using a scheduled multiple reaction monitoring algorithm and the data were processed using Analyst and MultiQuant 3.0 software (AB Sciex) (31, 88). The list of studied 437 lipids and their annotations are in Table 6S.

5.4 Phenotype-genotype many to many relationship analysis

This work uses a new bioinformatic statistical method framework PGMRA (13)(Figure 18), which our research team has previously used in cooperation with the developers (14-16), a screen capture of the web server can be seen in Figure 19.

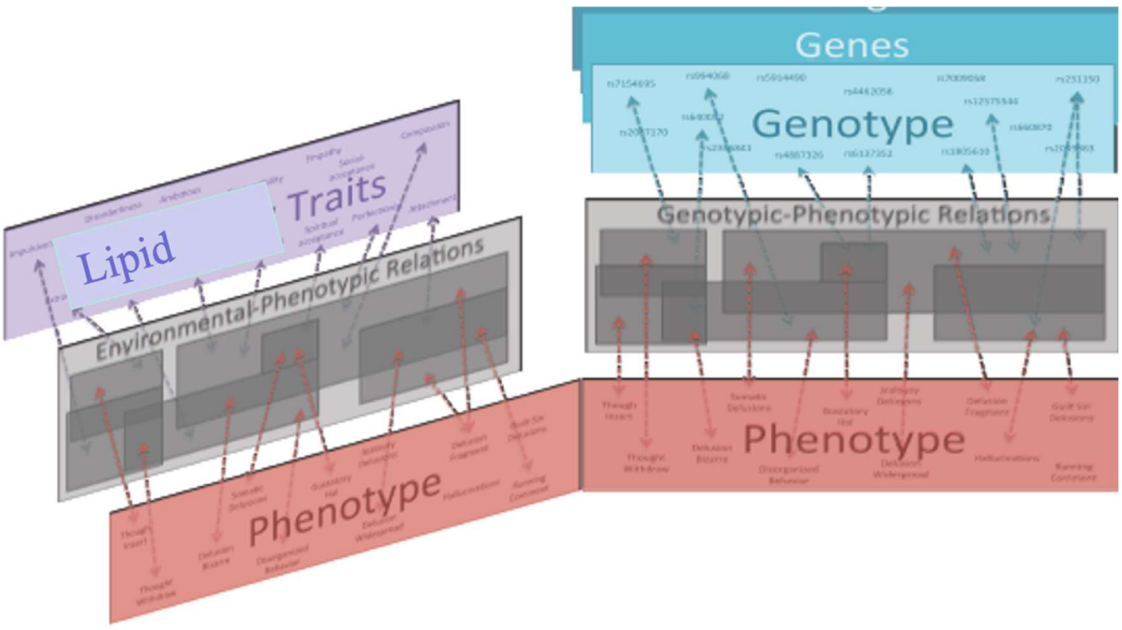

Figure 18. PGMRA framework. Modified from ***(12)***

It is based on NMF (nonnegative matrix factorization), where clusters are made of both phenotype data (phenotype features x subjects) and genotype data (SNPs x subjects), the data is first factorised using Fuzzy NMF (13) (89), after which the significance of overlapping subject (ID) is tested over each genotype x phenotype co-clusters. To do so,

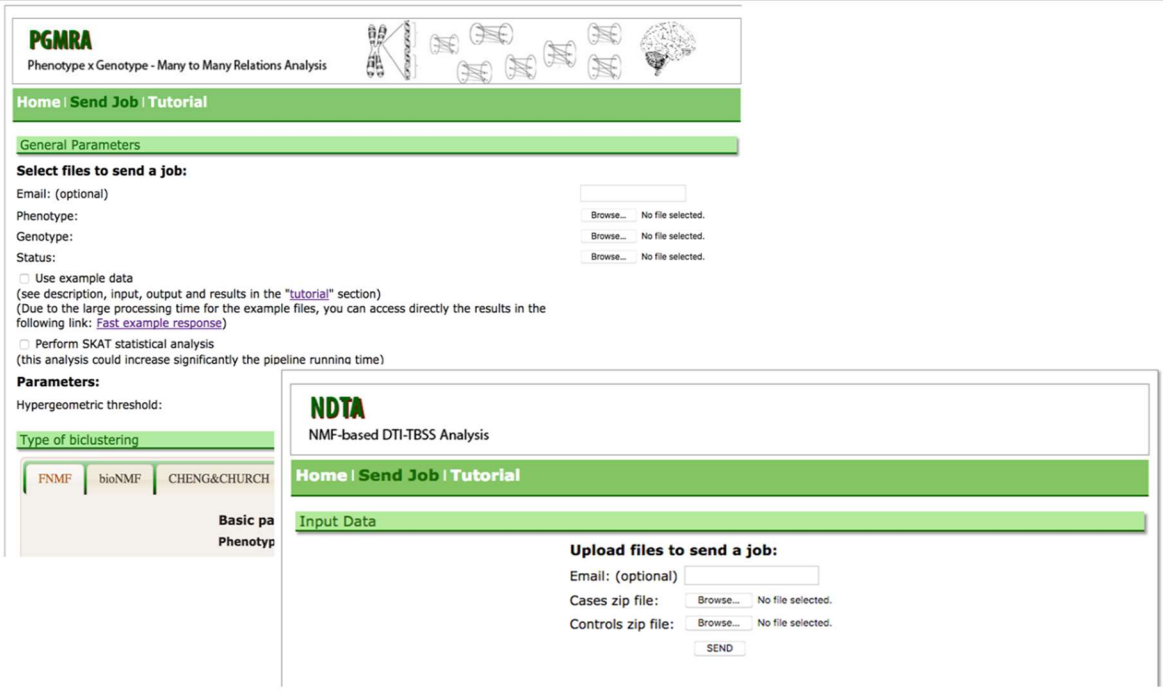

Figure 19. Screen capture of the PGMRA website

the algorithm uses a co-cluster test based on hypergeometric statistics, where the coherence between two bicluster is evaluated by their common observations. Then, coherent bicluster are encoded as relational bicluster or simply relations. (13) **(**Figure 20**).**

Then, by incorporating a posteriori the subject status within each relation, we can establish the risk surface of a disease (in our research risk for subclinical atherosclerosis) in an unbiased mode.

The analysis is performed using R, Python and C programming languages, with artificial intelligence based counting algorithms and from them coded wide capable calculating platform. This platform produces the virtualization of the results, which is easy to interpret and understand. In this work, many-phenotype (P) from the 437 quantitative lipid molecules measured using the LC-MS/MS method from plasma and the genome wide SNP data as the explainer, from which the pre-selected nominally significant (gene variants

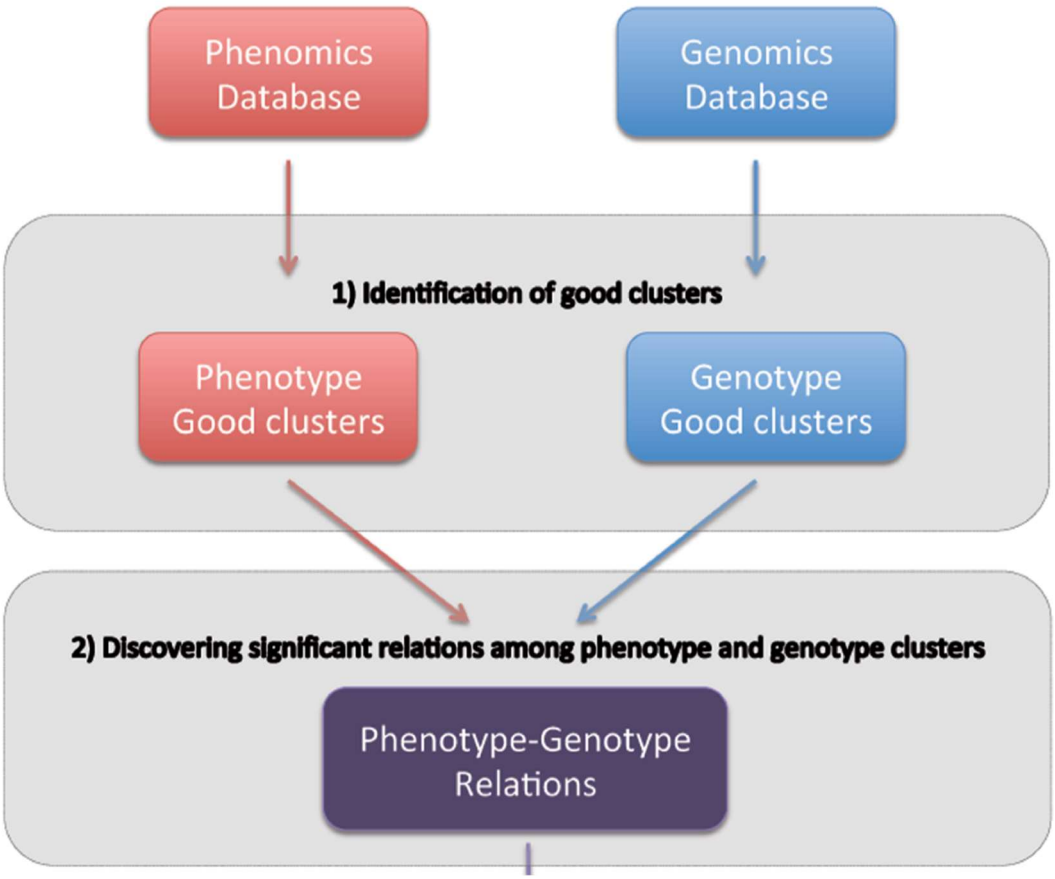

Figure 20. Clusters are formed and evaluated from data; good clusters are then analysed in how they relate. Modified from ***(13)***.

(SNPs) are chosen. These gene variants are picked by performing a GWAS analysis for each one of the lipidome containing lipid species.

Concentrations of the 437 molecular lipid species in the data were measured from plasma using LC-MS/MS-method. In this bioinformatic work, before the actual PGMRA analysis, a GWAS analysis was performed for each of the 437 lipids, from where the genetic variants for the PGMRA were chosen.

For this reason, due to a hard requirement on our available computing capacity we had to limit the individual (437+5) GWAS analyses to just genotyped 546,677 SNPs instead of available 40 million (1000 genome reference) imputed SNPs.

5.5 Gene set enrichment analysis

The gene set enrichment analysis (GSEA) was done using overrepresentation method implemented in the *clusterProfiler* R package (doi: 10.1089/omi.2011.0118), chosen since we just had a list of genes to which the SNPs were mapped. The analysis was done using a R-program and codes 11 and 12 shown below in chapter g. Altogether 28922 Biological process Gene Ontology based gene sets were analyzed.

6. BIOSTATISTICAL ANALYSES

6.1 Data pre-processing for biostatistical analysis

The data for the study was analysed on a standard desktop computer (AMD Ryzen 7 3700X, 32 GB ram, Windows 10). The data consists of three parts, covariate data, lipidomic data and SNP data. The data was read using R Statistical package v. 4.0.2 (http://www.r-project.org), together with RStudio v. 1.3.1073 (https://rstudio.com/) (**Code 1**).

```
staualab07 <-read_sas("staualab07.sas7bdat")

lipidAnnotation  <-  as.data.frame(read.csv2("YFS_Zora_lipids_annotation.csv"))

lipido <- read.csv("Young_Finns_lipid.csv", header = T, sep = ";")

lipidNames <- read.table("lipidlist.csv", sep = ";")
```

Code 1. *Loading data to memory.*

Lipids were coded differently in the files, differences such as using "_", ":" or ".", and were then renamed to match the data file naming scheme. This was primarily done using gsub() function, which searches for a given set of characters from a string and replaces it with another given set (**Code 2**).

```
valist<-c("tutkno", gsub("_", ".", lipidAnnotation[lipidAnnotation$Lipid_class1==lclass,]$SAS_DATA_NAME))

if (lclass == "Glycerophospholipid"){
    valist <- valist[-2:-77]
    ##The replaced list of names for the lipids
    valist <- c(valist , "LPC.14.0_sn1", …, "LPE.P.20.0")
  }
```

Code 2. *Renaming lipids to match the data.*

The lipidomic data was read into memory in string format and is changed back to numeric with a use of lapply() and gsub(), which apply a given function to every entry in a list (**Code 3**).

```
#Copy the lipid data from the list of chosen lipids
pheno <- lipido[,valist]

#Changing the string values back to numeric using lapply and gsub, substituing ',' for '.'.

pheno[,2:length(names(pheno))]<-laply(
pheno[,2:length(names(pheno))],          function(x)          as.numeric(gsub(",",".",x)))}
```

Code 3. *Using lappy() and gsub to switch "," with ".".*

6.2 Dimensionality reduction of lipidomic data with Principal Component Analysis

The LC-MS/MS based analysis results of plasma lipids were classified to five eigenlipid groups, Fatty_acyl, Sterol_lipid, Sphingolipid, Glycerolipid and Glycerophospholipid by dimensional reduction using PCA. The specific classification (annotation) of all analysed lipids is found in Supplemental Table 1S.

PCA was performed in a for() loop within a try({}), repeating so for each of the above five groups separately and due to the try({}), if something in a loop fails, it will simply move to the next loop instead of interrupting the code, with a warning message printed to the console. One of the first things the loop performs is the data pre-processing mentioned previously and the following check (**Code 4**), where the data is checked for missing value (NA, not available) values, the numeric conversion is confirmed and the lipidomic data is

```
for (val in names(pheno)[2:length(names(pheno))]){
  n <- length(pheno$tutkno) - length(pheno[!is.na(pheno[, val]), val])
  ##print(sprintf("%s NA: %s n: %s", val, n, length(pheno[,val])));

  ## IF value not numeric, like in the case of TUTKP11 variable, skip to next loop
  if(!is.numeric(pheno[, val])) next



  #If there are more than 0 non-NA values, add to p the non-NA values. Then scale the values in p$val amongs each other.
  if (n > 0){
    pn <- p[!is.na(p[,val]),]
    p <- pn
    rm(pn)
  }
  p[,val] <- scale(p[,val])
}
```

Code 4. *Checks the data for NA (=missing) values, confirm numeric nature and scale the lipidomic data.*

normalised using the scale() generic function from the base R package.

Then in preparation for the PCA analysis, subject ID numbers are stored into row-names, the ID value row is removed and the PCA analysis is run (**Code 5**), after the analysis is done, a temporary empty data frame is made, to which the first principal component (PC1) is extracted from the PCA results, the subject ID values column is renamed, and the data is then merged to an out of loop object by subject ID.

```
#Add tutkno ID to the rownames as the prcomp did not like the ID
values. Then remove tutkno ID entry.
row.names(p) <- p$tutkno
p <- p[,-1]

#PCA analysis
pca <- prcomp(p,scale=T, rank. = 5)

#Refresh p for use as a data from for extracting the data from the pca
analysis.
#Tehn pick the PC1 entry values as numeric.
p <- data.frame()

#Extract the rowname information that holds the subject ID values
p <- data.frame(as.numeric(names(pca$x[,"PC1"])))

#Gather the data to p, into a column called by the lclass name cur-
rently ongoing in the loop
p[,lclass] <- pca$x[,"PC1"]
names(p)[names(p) == "as.numeric.names.pca.x..PC1.."] <- "tutkno"

#Merge data into an out-of-loop data frame by tutkno ID values.
collection <- merge(collection, p, by.x = "tutkno", by.y = "tutkno",
sort = F, all = T)
```

Code 5. *PCA analysis and data extraction.*

6.3 Genome-wide association analysis of human lipidome

Illumina 670k custom-built chip was used in the Young Finns study genotyping. All GWAS analyses were done by using PLINK v1.90b3q 64-bit (29 May 2015) and using genotyped data.

The data was first preformatted, a .fam file was made with subject IDs that matched the ones in use and the files required for the GWAS analysis were loaded, the pheno.link object contains the pre-calculated genetic principal components (PC1-10) for the staualab7 data (**Code 6**).

The code is run a loop, with separate scripts made for the lipid groups and the larger individual lipids, but the main point stays the same, first the data for a GWAS run is gathered, renamed and checked for NA values, then a text file is created with the lipid data and all the covariate data (**Code 7**). The nlist loop element is a list of names for the lipids or the lipid groups and is written in the txt file as well as used to name it.

```
linkkiID <- read.delim("ID_Linkki_laseri.txt")

staualab07_lim <- staualab07[,c("Tutkno07", "ika07", "TUTKNO80", "SP",
"bmi07", "diabet207", "fhlaak07")]

staualab7_lim   <-   merge(   staualab07_lim,   pheno_link[,c("tno",
paste("PC", 1:10, sep = ""))],
        by.x = " Tutkno07", by.y = "tno", all = F, sort = F
)
```

Code 6. *File load for GWAS.*

```
for (name in nlist){
  combined  <-  merge(staualab07_lim,  p,  by.x  =  "Tutkno07",  by.y  =
"tutkno", sort = F, all = T)

  combined <- combined[,c("TUTKNO80","Tutkno07", "SP", name,
                          "bmi07", "ika07", "diabet207", "fhlaak07")]

  names(combined)[names(combined)=="TUTKNO80"] <- "FID"
  names(combined)[names(combined)=="Tutkno07"] <- "IID"

  combined  <-  merge(combined,  linkkiID,  by.x  =  "FID",  by.y  =
"tutk_nro", all = F, sort = F)

  combined[,"FID"] <- combined$Sanger_ID
  combined[,"IID"] <- combined$Sanger_ID

  names(combined)[names(combined)=="SP"] <- "sex"
  names(combined)[names(combined)=="ika07"] <- "age"
  names(combined)[names(combined)=="bmi07"] <- "bmi07"

  combined <- combined[!is.na(combined[, name]),]
  combined <- combined[!is.na(combined[,"bmi07"]),]
  combined <- combined[!is.na(combined[,"sex"]),]
  combined <- combined[!is.na(combined[,"diabet207"]),]
  combined <- combined[!is.na(combined[,"fhlaak07"]),]

  ##Switch numbers for sex
  combined[,"sex"] <- ifelse(combined[,"sex"]==1, 2, 1)

  phenolist<- paste("GWAS_Lipid_Loop/pheno_", name, ".txt", sep = "")
  write.table(cbind(combined[,c("FID","IID",  "sex",  name,  "age",
"bmi07", "diabet207", "fhlaak07")]),
```

```
row.names=F,col.names=T,dec=".",            sep="\t",file=phenolist,
quote=FALSE)

```

Code 7. *In-loop file creating for GWAS run.*

The latter half of the loop creates two strings, one to hold all covariates used and then the actual command for the GWAS run, which is run directly from RStudio using system (), which allows one to run system commands (**Code 8**).

```
  covarName <- paste("sex, age, bmi07, diabet207, fhlaak07, sep = "")

  command <- paste("plink --bfile LASERI_20091110 --linear hide-covar
mperm=100 --pheno ",
                   phenolist," --pheno-name ",
                   name," --covar ", phenolist,
                   " --covar-name ", covarName, " --pfilter 1e-4 --
out GWAS_Lipid_Loop/OutputB/",
                   name,".txt --threads 7", sep = "")
  system(command)
```

Code 8. *GWAS command creation, execution and loop end.*

6.4 Lipidome-genotype many-to-many relationchip analysis

The PGMRA analysis was performed in PGMRA server by the group that originally developed it (13). I did the preselection of nominally associated SNPs ($p < 5 \times 10^{-4}$) for PGMRA analysis, and it was based on GWAS of 437 original lipids (see chapter 6.3). These analyses resulted a subset of 18 370 nominally associated SNPs that were then used in PGMRA analysis. This list of preselected SNPs (extract snps.txt) for PGMRA analysis was extracted among YFS study SNP data (bfile LASERI_20091110) using the (**Code 9**) below.

Additionally, Manhattan plots were made from the GWAS data using **Code 10,** which

```
plink --bfile LASERI_20091110 --extract snps.txt --make-bed --out LA-
SERI09_subset
```

Code 9. *Subsetting the selected SNP data from GWAS .bed/.bim/.fam format into their own files.*

was done using the package ‘manhattanly’(version 0.2.0) and later ‘qqman’ (version 0.1.8).

6.5 Gene set enrichment analysis

```
library('manhattanly')

manhattanly(collectedOutput, snp = "SNP")
manhattanly(groupOutput[!duplicated(groupOutput$SNP),], snp = "SNP",
title = "Eigenlipid SNPs")
   manhattanly(groupOutput[groupOutput$LCLASS == 'Fatty_acyl',], snp =
"SNP", gene = "LCLASS")

```

Code 10. *Doing Manhattan plots from the genome-wide SNP-data*

The group of 35 relations was subjected to a gene set enrichment analysis that enriched 18 out of the 35, done using **Code 11,** this was done using packages ‘EnsDb.Hsapines.v79’(2.99.0) and ‘clusterProfiler’ (3.16.1). The results where then formed into a bar plot by **Code 12**.

```
dat<- read.csv("Relationdata_fromResults.txt",sep = " ",header = F)
ens_id<-dat$V4


#BiocManager::install("EnsDb.Hsapiens.v79")
library(EnsDb.Hsapiens.v79)
BiocManager::install("clusterProfiler")
library(clusterProfiler)
# see this if error with annotation db 'options(connectionObserver
= NULL)'
gene_symbol <- ensembldb::select(EnsDb.Hsapiens.v79, keys= ens_id,
keytype = "GENEID", columns = c("GENEID","SYMBOL"))

relations<-read.csv("Relation35snp_RelationTable.txt",sep    =    "
",header = T)
uniq <- unique(relations$Relation)

relations_tbl <- list()

for(i in 1:length(uniq)){
  relations_tbl[[i]]   <-   relations[grep(uniq[i],relations$Rela-
tion),]
}

############### attach gene symbol to SNPs ####
library(clusterProfiler)
enrich_list <- list()

##### replace i with values from 1 to 35 below (you can use for loop
but this requires some fixing in enrich_list as some relations don't
have any pathways. I did it manually)
res <- relations_tbl[[i]]
tmp <- dat_relations[match(res$SNP,dat_relations$V1),]
sum(res$SNP==tmp$V1)==nrow(res)
res$ensID <- tmp$V4
enrich_list[[i]]       <-       enrichGO(gene=res$ensID,       Or-
gDb=org.Hs.eg.db,ont= "BP", pAdjustMethod = "BH", pvalueCutoff   =
0.05, qvalueCutoff=0.05, keyType='ENSEMBL',readable=TRUE)@result

save(enrich_list,file="enrich_list.RData")
```

Code 11. *Pathway analysis by using gene set enrichment method.*

```
load("enrich_list.RData") ## this data was generated using codes from 'manhattan_pathway.R'

## isolate relations with significant GO terms. For example,
R184 <- enrich_list[[34]][which(enrich_list[[34]]$p.adjust<0.05),]
R184 <- R184[!duplicated(R184$geneID),]

# combine all tables/relations
dat <- rbind(R104,R112,R119,R125,R128,R136,R160,R162,R18,R184,R188,R30,R34,R49,R56,R76,R77,R79,R94)

dat$Genotype_lipidome_relations <- c(rep("R104",nrow(R104)),rep("R112",nrow(R112)),rep("R119",nrow(R119)),rep("R125",nrow(R125)),rep("R128",nrow(R128)),rep("R136",nrow(R136)),rep("R160",nrow(R160)),rep("R162",nrow(R162)),rep("R18",nrow(R18)),rep("R184",nrow(R184)),rep("R188",nrow(R188)),rep("R30",nrow(R30)),rep("R34",nrow(R34)),rep("R49",nrow(R49)),rep("R56",nrow(R56)),rep("R76",nrow(R76)),rep("R77",nrow(R77)),rep("R79",nrow(R79)),rep("R94",nrow(R94)))

# remove duplicates
dat1 <- dat[!duplicated(dat$Description),]

#  avoid ggplot ordering description
dat1$Description <- factor(dat1$Description, levels = dat1$Description)
dat1$p.adjust <- -log10(dat1$p.adjust)

ggplot(dat1,   aes(x=Description,   y=p.adjust,   fill=Genotype_lipidome_relations)) +
  geom_bar(stat="identity") + xlab("Biological processes") + ylab("-log10(P-values) adjusted with Benjamini & Hochberg method") +
coord_flip()

# or,

legend_title <- "Genotype-lipidome relations"
ggplot(dat1,   aes(x=Description,   y=p.adjust,   fill=Genotype_lipidome_relations))
 + geom_bar(stat="identity") + xlab("Biological processes")
+ ylab("-log10(P-values) adjusted with Benjamini & Hochberg method")
+theme(text = element_text(size=15), axis.text.x = element_text(angle=90, hjust=1))


```

Code 12. *Gene set enrichment analysis result plot drawing after gene set enrichment analysis.*

6.6 Lipidome associated SNPs, annotation to genes and gene functions

The Ensembl Variant Effect Predictor (VEP) is an open-source tool for the analysis and annotation of genomic variants, for both coding and non-coding regions (90). It is a tool-set with multiple interfaces and options for configuring your analysis and has access to a large collection of genomic annotation with many different options for specific requirements. The Ensembl Variant Effect Predictor is free to use and is available as a web interface and a command line tool (90)(Figure 21).

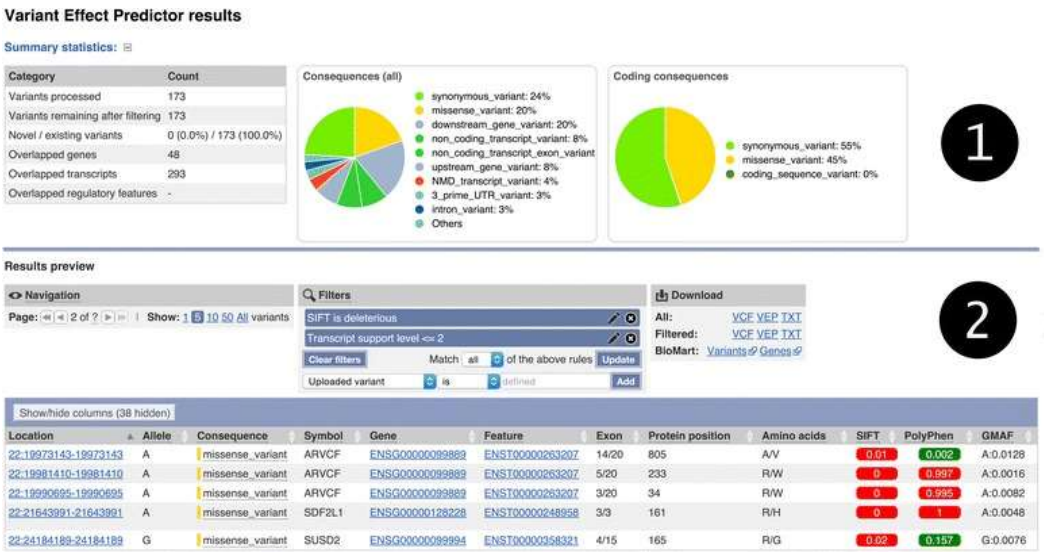

Figure 21. A typical VEP Web results page. Section (1) gives summary pie charts and statistics. Section (2) contains a preview of the results table with navigation, filtering and download options. The preview table contains hyperlinks to genes, transcripts, regulatory features, and variants in the Ensembl browser. The results can be downloaded in CF, text, or custom VEP file formats ***(90)***.

7. RESULTS

7.1 Dimensionality reduction of lipidomic data with PCA and GWAS analysis of eigenlipids

The GWAS of five PCA derived eigenlipids (in chapter 6.2) serving as a surrogate marker for all plasma lipids is given in Figure 22. showing combined Manhattan plot for these five eigenlipids.

We performed GWAS for the five eigenlipid groups (named Fatty_acyl, Sterol_lipid, Sphingolipid, Glycerolipid and Glycerophospholipid groups) derived after principal component analysis (see PCA analysis chapter 6.2) using age, sex, and BMI as covariates. In this analysis for the five eigenlipids (using p-value threshold < 5 x 10^{-4}), we found a total of 751 separate SNP - eigenlipid associations (Figure 22).

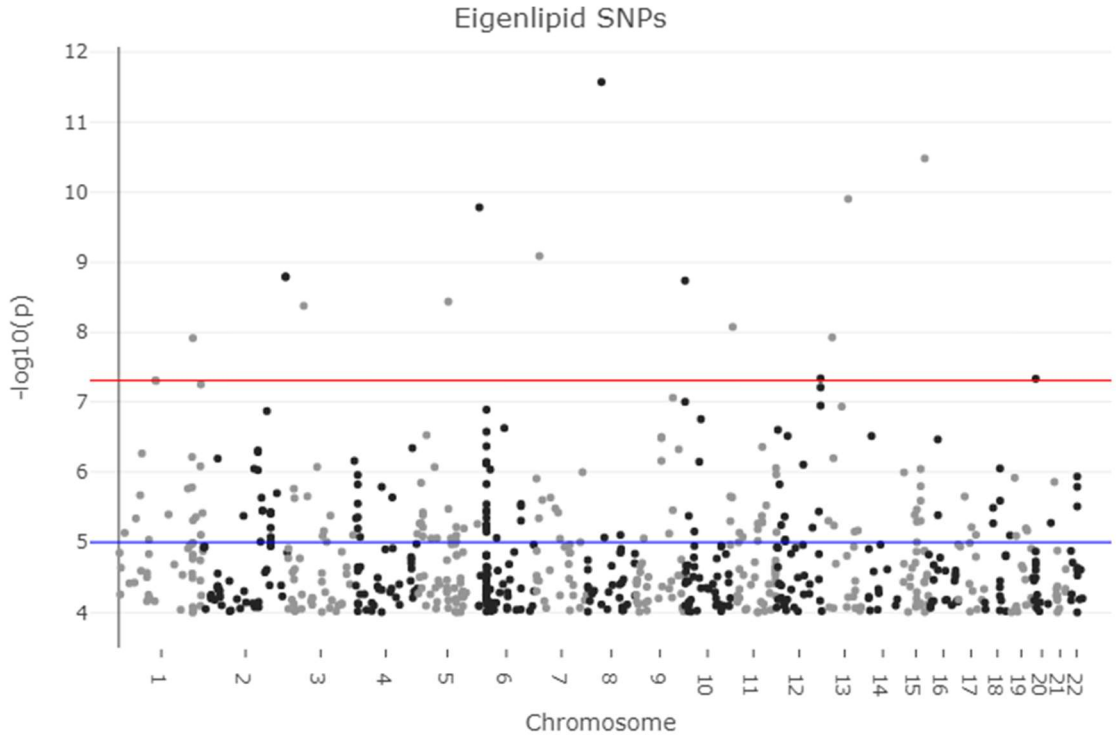

Figure 22. Manhattan plot of SNPs from the five eigenlipid GWAS analysis. X-axis indicates the Chromosome numbers (1-22) and y-axis SNP related p-values, transformed to -log10(p) scale.

7.2 Lipidome-wide GWAS analysis

Two different statistical regression models, i.e. GWAS studies (a and b) were performed for all 437 separate lipid species using PLINK **a)** using sex, age, BMI, type 2 diabetes, lipid medication as covariates and model **b)** adding also the first 10 genetic principal components (PC1-10) as covariates.

The GWAS for individual 437 lipid species run without the PC's (model a), resulted ~48 000 nominally (p-level < 5 x 10^{-4}) associated SNP's (Figure 23), which increased to 51 707 SNPs when also the genetic PCs1-10 were added as additional covariates to the otherwise similar analysis (model b). Of these 51 707 total SNP-lipid-trait associations, only 18 370 were related to separate SNPs. From these SNPs, 634 and 266 were significant at genome-wide level p-level < 5 x 10^{-7} and < 5 x 10^{-8}, respectively. The total number of significant (p-level < 5 x 10^{-7}) associations between all studied lipid traits (n=437) and these separate 634 significant SNPs was 4482. Similarly, the total number of significant (p-level < 5 x 10^{-8}) associations between all studied lipid traits (n=437) and these separate 266 significant SNPs was 2340.

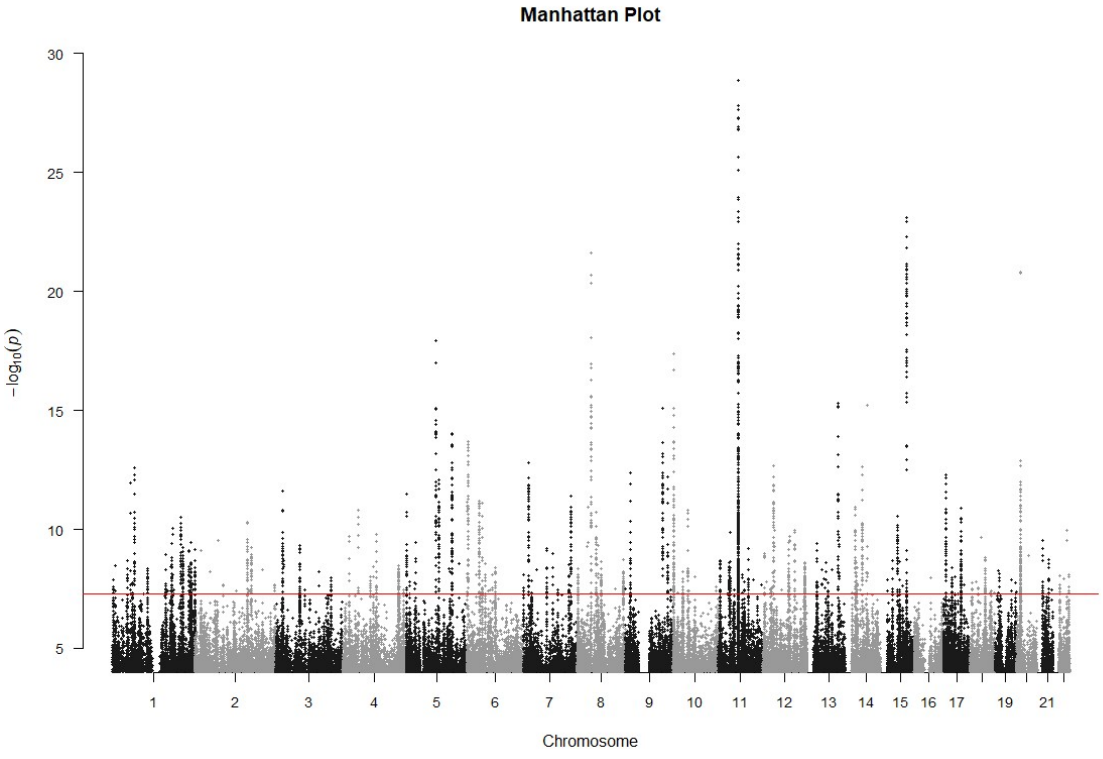

Figure 23. Manhattan plot of the lipidome-wise GWAS results, in figure those above red line are nominally significant SNPs (p-value < 5 x 10^{-4}).

From these genome-widely significant 634 SNPs (at p-level < 5 x 10^{-7}) only 32 was shared (5 %) with the ones found to be associated in the grouped eigenlipid analysis at the same p-level < 5 x 10^{-7} (Figure 22 in chapter 7.1).

7.3 PGMRA analysis and new genetic lipidome classification

The PGMRA analysis identified a total of 209 both “phenotype X subject” and “genotype X subject” groups, or biclusters to be more exact. These were filtered using a SKAT test resulting in 71 of the 209 phenotype-subject biclusters surviving and 153 of the 209 genotype-subject biclusters surviving. Relationship analysis between these two groups of biclusters gave a total 189 suitable relations of which 93 were found significant by a further hypergeometric testing. These significant relations comprise of a total 5977 different SNPs that were in 3164 different genes, the analysis flow is shown in Figure 24.

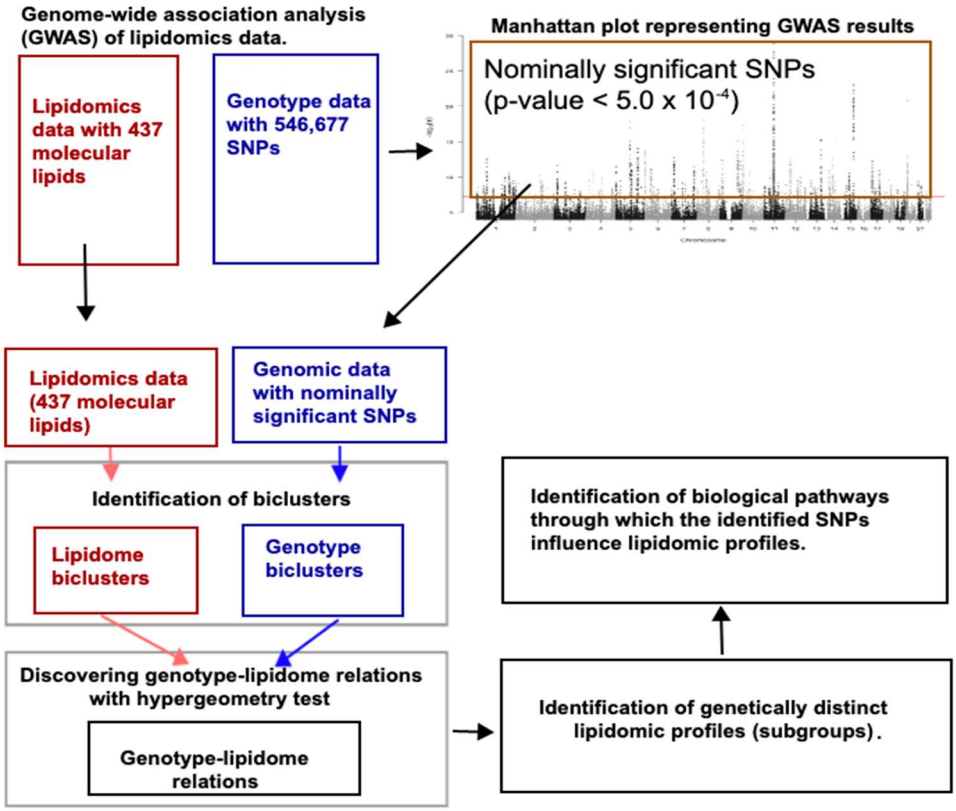

Figure 24. Represents the outline of PGMRA analysis. The PGMRA analysis resulted following data output and results called lipids x subject-sets, SNPs x subjects sets and their relations. GWAS-PGMRA analysis flow image was made using PDFescape online software, with no code available in an open format.

7.3.1 Lipidome biclusters (lipids x subject sets)

Lipid biclusters (lipid x subject) are subset of people who are similar based on subset of their plasma lipid concentrations. PGMRA identified 71 optimized phenotypic lipid biclusters in our YFS lipidomic data. Supplementary Table 2S. shows the detailed lipid names of each of the lipid phenotypic biclusters.

7.3.2 Genotypic biclusters (SNPs x subjects sets)

Genotypic biclusters (SNPs x subject) are subset of people who are similar based on subset of their genome SNPs. PGMRA identified 153 optimized genotypic SNP biclusters in our YFS genotypic data. Supplementary Table 3S. shows the detailed rsnumbers for each SNP of the identified SNP biclusters and the Table 4S statistical associations of each of these 153 SNP sets with subjects.

7.3.3 Bicluster (lipids x subjects) to (SNPs x subjects) relations

Supplementary Tables 2-5 show results from bicluster (lipids) to bicluster (SNPs) relations. There are 93 pairs of genotype-lipid biclusters with statistically significant relation. The significance between related groups was calculated with/using hypergeometric test (i.e. overlap of participants between these two biclusters i.e. between the two datasets).

PGMRA analysis resulted in 209 biclusters for both (lipid x subject) and (SNP x subject), filtering of these groups by SKAT reduced them to 71 lipid (Supplemental Table 5S) and 153 SNP biclusters (Supplemental Table 4S). The hypergeometric test found 93 significant ($p<0.01$ see supplementary Table 6S) relations between them and 17 relations of them had a p-level $< 5 \times 10^{-4}$. These 93 statistically significant relations identified of 5977 different SNPs that cover 3164 different genes (listed in Supplemental Table 7S). Thirty-five of the relations contained distinct SNPs representing genetic lipidomic subgroups (Table 1).

Table 1. The 35 significant and distinct lipid x gene SNP set bicluster relations after PGMRA and hypergeometric testing.

Relation	numPatients	numFeatsAY(SNPs)	numFeatsBY(Phenovars)	Hypergeometric
R0	23	59	82	0.007083508
R103	5	193	28	0.000467118
R104	5	10	36	0.00992867
R107	7	232	30	4.24E-05
R112	7	34	17	0.000284391
R119	9	34	30	0.000461113
R122	10	253	17	2.08E-08
R125	5	58	22	0.001375996
R128	7	105	16	0.005995584
R136	5	67	50	0.005221617
R143	5	108	17	0.001004533
R153	7	13	82	0.001494418

R160	6	16	14	0.000280794
R162	6	54	12	0.000162383
R164	7	10	16	0.008086693
R170	6	74	16	0.001615117
R18	7	39	50	0.000756494
R184	21	12	82	1.73E-05
R186	6	53	22	0.000759614
R188	14	19	22	0.000830085
R30	8	466	17	0.008708178
R34	6	10	20	0.008313783
R45	11	114	16	0.003974201
R48	10	108	17	6.04E-06
R49	7	58	13	0.008219131
R54	6	90	16	0.003897149
R56	7	8	7	0.007580586
R59	6	34	25	8.80E-05
R72	7	32	19	0.001422033
R76	5	22	17	0.00885272
R77	5	274	17	0.006299747
R79	10	42	22	0.002996717
R81	6	71	25	5.00E-05
R93	7	50	17	0.000725517
R94	5	7	23	0.001703721

7.4 Defining the role of common variation in the genomic and biological architecture of human lipidome

7.4.1 The annotation of lipidome associated genetic variation

Ensemble data base and Ensembl Variant Effect Predictor (VEP) was used as basis in the gene/SNP annotations (http://www.ensembl.org/index.html) (90).

The Supplemental Table 7S summarizes the genes and gene regions affected by lipidomic associated SNPs. There are a total 3614 different gene loci which we associated to lipidome regulation.

We did an exhaustive annotation for all of these SNPs/genes using the most up to date release of ensemble assembly GRCh37, the version 102 (GRCh37; homo_sapiens_core_102_37 on ensembldb.ensembl.org, November 2020 © EMBL-EBI). In the Supplementary tables 1-8 you will find information about all the discovered SNP sets, their consequences (all possible), the genes affected, their function, location, and many other features. Each discovered new SNP sets were annotated as Figure 25 shows an example below.

The genetic variation is classified (annotated) according to their consequence types as follows:

- Missense variant
- Synonymous variant

- non-coding transcript exon variant
- intron-variant
- intergenic variant

In general, the found SNP data (n=5977) indicates that there are a great number of protein-coding genes affected (n=1944), 312 pseudogenes, 58 miRNA genes, 393 lincRNA genes, 37 miscRNA genes, 19 snoRNA genes, 28 snRNA genes, 7 rRNA genes, and many up and downstream gene regions affected (protein gene regulatory regions).

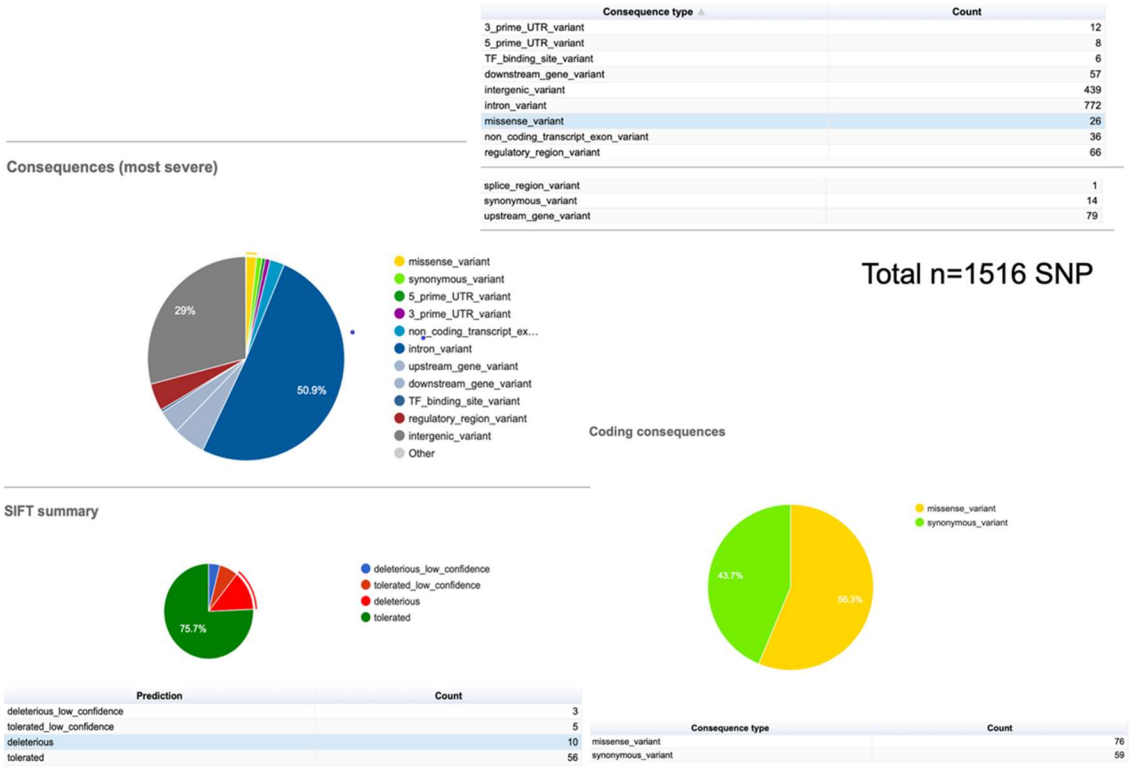

Figure 25. Annotation results of SNP sets, example (g.3.3 SNP set) including 1516 SNPs. Screenshot from ***(90)***.

The genes associated significantly with lipidome in PGMRA analysis were found to be in all autosomal chromosomes 1-22. Most of the genes i.e. 257 genes were in chromosome 1 and 246 in chromosome 2 see Figure 26**,** the chromosomal distribution of lipidome regulating genes.

There are also some SNP sets where most SNPs fall into blind regions of the genome,

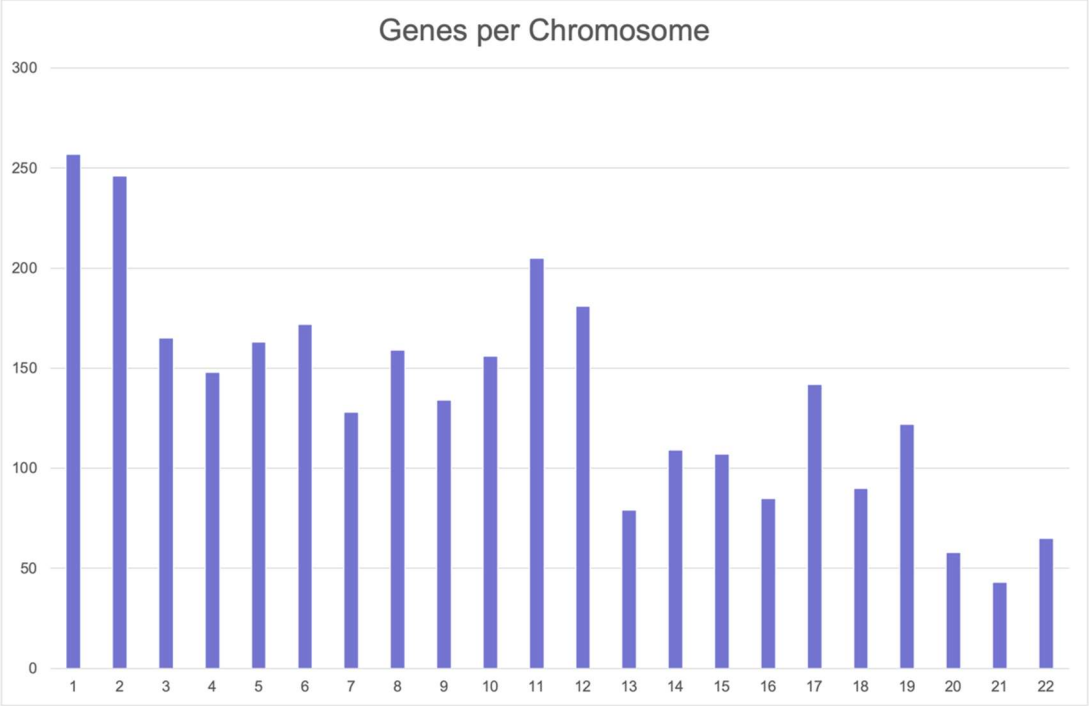

Figure 26. X-axis indicates chromosome number and y-axis number of lipidome associated genes.

this means a genomic location where very much nothing is currently known. The annotation analysis of all discovered lipidomic SNPs resulted a huge amount of information, both SNP-set and SNP wise manner, which is given in Supplementary tables 7-8.

7.4.2 The role of lipidome associated coding variants

From all SNPs 375 variants were protein changing missense-variants (Table 7S). In SIFT and PolyPhen based functionality analysis 19 of these missense variants were suggested to be deleterious/probably/possibly damaging. The gene location of these functional variants is shown in Table 2.

Table 2. Sift and PolyPhen,19 damaging missense variants.

SNP	Location	Allele	SYMBOL	Codons	SIFT	PolyPhen
rs10846018	12:7832722-7832722	A	SLC2A14	cGg/cTg	deleterious(0)	probably_damaging(0.961)
rs10888390	1:150755063-150755063	A	CTSS	Cgg/Tgg	deleterious(0.04)	possibly_damaging(0.521)
rs11225089	11:101961859-101961859	A	CEP126	tCc/tAc	deleterious(0)	probably_damaging(0.917)
rs1135889	17:75930040-75930040	A	FBF1	gGt/gTt	deleterious(0)	possibly_damaging(0.642)
rs11568658	13:95210754-95210754	A	ABCC4	Ggg/Tgg	deleterious(0)	probably_damaging(0.985)
rs1609860	12:103654676-103654676	A	STAB2	cCc/cAc	deleterious(0)	probably_damaging(0.999)

rs1806931	19:15728555-15728555	T	OR10H2	tCc/tTc	deleterious(0.01)	possibly_damaging(0.595)
rs1919127	2:27578626-27578626	C	C2orf16	gTg/gCg	deleterious(0.05)	possibly_damaging(0.775)
rs2235638	16:1523889-1523889	T	IFT140	gCg/gAg	deleterious(0)	possibly_damaging(0.636)
rs2240227	19:15741432-15741432	A	OR10H3	Ctc/Atc	deleterious(0.02)	possibly_damaging(0.844)
rs2302948	19:48592808-48592808	G	SULT2B1	Ctg/Gtg	deleterious(0)	probably_damaging(0.922)
rs272893	5:132327369-132327369	G	SLC22A4	aTa/aGa	deleterious(0)	possibly_damaging(0.675)
rs3739407	8:17755366-17755366	G	MTUS1	Tgt/Cgt	deleterious(0.03)	possibly_damaging(0.667)
rs3742303	13:30646969-30646969	T	USPL1	Cct/Tct	deleterious(0.04)	possibly_damaging(0.68)
rs3765623	18:3086067-3086067	T	MYOM1	Gat/Aat	deleterious(0.02)	possibly_damaging(0.885)
rs6413419	10:133532171-133532171	T	CYP2E1	Gtc/Ttc	deleterious(0)	probably_damaging(1)
rs6700677	1:42757818-42757818	T	P3H1	Gga/Aga	deleterious(0)	probably_damaging(1)
rs8181512	11:5821126-5821126	T	OR52N2	cAt/cTt	deleterious(0.02)	probably_damaging(0.988)
rs8480	1:151760859-151760859	G	MRPL9	gAa/gCa	deleterious(0.04)	possibly_damaging(0.655)

Annotation analysis of lipidomic related SNPs showed that there are a great number of protein-coding genes that are affected (n=1944) (Table 7S).

From these synonymous variations which are defined as codon substitutions that do not change the encoded amino acid, were previously thought to have no effect on the properties of the synthesized protein(s) (91). However, now mounting evidence shows that these “silent” variations can have a significant impact on protein expression and function and should no longer be considered “silent” (91, 92). These synonymous codon substitutions can perturb co-translational protein folding in vivo and impair cell fitness (92).

7.4.3 The role of lipidome associated non-coding RNAs

Table 3 below show the genes of the 58 miRNAs that were found in PGMRA, the full list of non-coding genes is included in the supplemental Table 7S.

Table 3 Stable ID, chromosome and gene names of the lipidome associated 58 miRNAs found by PGMRA.

Gene stable ID	Chromosome/ scaffold name	Karyotype band	Gene name	EntrezGene description
ENSG00000199065	9	p24.1	MIR101-2	microRNA 101-2
ENSG00000199127	8	p22	MIR383	microRNA 383
ENSG00000207601	11	q12.2	MIR611	microRNA 611
ENSG00000207951	2	q32.1	MIR561	microRNA 561
ENSG00000212037	14	q12	AL049831.1	
ENSG00000216058	3	q28	MIR944	microRNA 944
ENSG00000221401	7	q32.1	AC025594.1	

ENSG00000221531	2	p23.3	AC074091.1	
ENSG00000221670	13	q12.3	AL596092.1	
ENSG00000221703	11	p15.4	MIR302E	microRNA 302e
ENSG00000221753	8	q22.2	MIR1273A	microRNA 1273a
ENSG00000221762	2	q22.1	AC092786.1	
ENSG00000222117	9	q31.1	AL391867.1	
ENSG00000222326	11	q12.2	MIR1908	microRNA 1908
ENSG00000222331	3	p22.1	AC104434.1	
ENSG00000222482	7	q22.1	AC005071.1	
ENSG00000222602	13	q13.3	AL136160.1	
ENSG00000222805	19	p13.11	AC010319.1	
ENSG00000223148	19	q12	AC011478.1	
ENSG00000223243	10	q22.3	AC074323.1	
ENSG00000238728	16	p13.11	MIR1972-1	microRNA 1972-2
ENSG00000238957	10	q11.21	AL512640.1	
ENSG00000251772	20	q13.31	AL117380.2	
ENSG00000252748	14	q32.11	AL096869.1	
ENSG00000253012	10	q21.3	AC022538.1	
ENSG00000253037	6	q21	AL109947.2	
ENSG00000254324	8	q24.3	MIR151A	microRNA 151a
ENSG00000263649	6	p21.32	MIR3135B	microRNA 3135b
ENSG00000264022	1	q21.2	AL732363.1	
ENSG00000264094	8	p11.1	AC022616.1	
ENSG00000264180	HSCHR6_MHC_APD	p21.32	MIR3135B	microRNA 3135b
ENSG00000264200	11	q22.3	MIR4693	microRNA 4693
ENSG00000264209	3	p21.31	AC104448.1	
ENSG00000264244	HSCHR6_MHC_DBB	p21.32	CR753846.4	
ENSG00000264283	11	p15.3	AC025300.1	
ENSG00000264382	2	p13.3	AC007881.2	
ENSG00000264412	HSCHR6_MHC_COX	p21.32	MIR3135B	microRNA 3135b
ENSG00000264553	1	q21.3	MIR4257	microRNA 4257
ENSG00000264645	16	q23.1	AC010528.1	
ENSG00000264749	12	q23.3	AC011313.1	
ENSG00000264770	HSCHR6_MHC_SSTO	p21.32	MIR3135B	microRNA 3135b
ENSG00000265083	6	p25.1	MIR3691	microRNA 3691
ENSG00000265140	11	q23.2	MIR4301	microRNA 4301
ENSG00000265207	HSCHR6_MHC_MANN	p21.32	BX927160.1	
ENSG00000265224	2	q22.2	AC012353.1	
ENSG00000265377	HSCHR6_MHC_QBL	p21.32	AL773543.2	
ENSG00000265450	13	q34	MIR4502	microRNA 4502
ENSG00000265520	8	p22	MIR548V	microRNA 548v
ENSG00000265974	HG14_PATCH	p11.2	MIR3147	microRNA 3147
ENSG00000266072	13	q14.3	MIR5693	microRNA 5693
ENSG00000266168	7	p11.2	MIR3147	microRNA 3147
ENSG00000266189	17	q25.3	MIR3186	microRNA 3186
ENSG00000266303	HG271_PATCH	q25.3	MIR3186	microRNA 3186
ENSG00000266320	22	q12.3	MIR3909	microRNA 3909
ENSG00000266429	9	p21.2	AL442639.1	
ENSG00000266704	12	q24.23	MIR4498	microRNA 4498
ENSG00000266761	20	q13.2	MIR3194	microRNA 3194
ENSG00000266807	6	q23.2	MIR548H5	microRNA 548h-5
ENSG00000266851	HSCHR6_MHC_MCF	p21.32	CR759848.2	
ENSG00000270758	HG1287_PATCH	q21.1	AC239859.1	

7.4.4 Gene set enrichment analysis of lipidome associated SNP-set

We used gene set enrichment analysis to further elucidate the biological significance of lipidome associated SNP-sets. In the case of the significant 35 distinct genotype-lipidome relations, 18 SNP sets out of the 35 were enriched in 52 various biological processes (FDR<0.05) shown in Figure 27. The bar plot (Figure 27) contains all these 52 significant gene sets. The colours in the figure represents which relations they belong to.

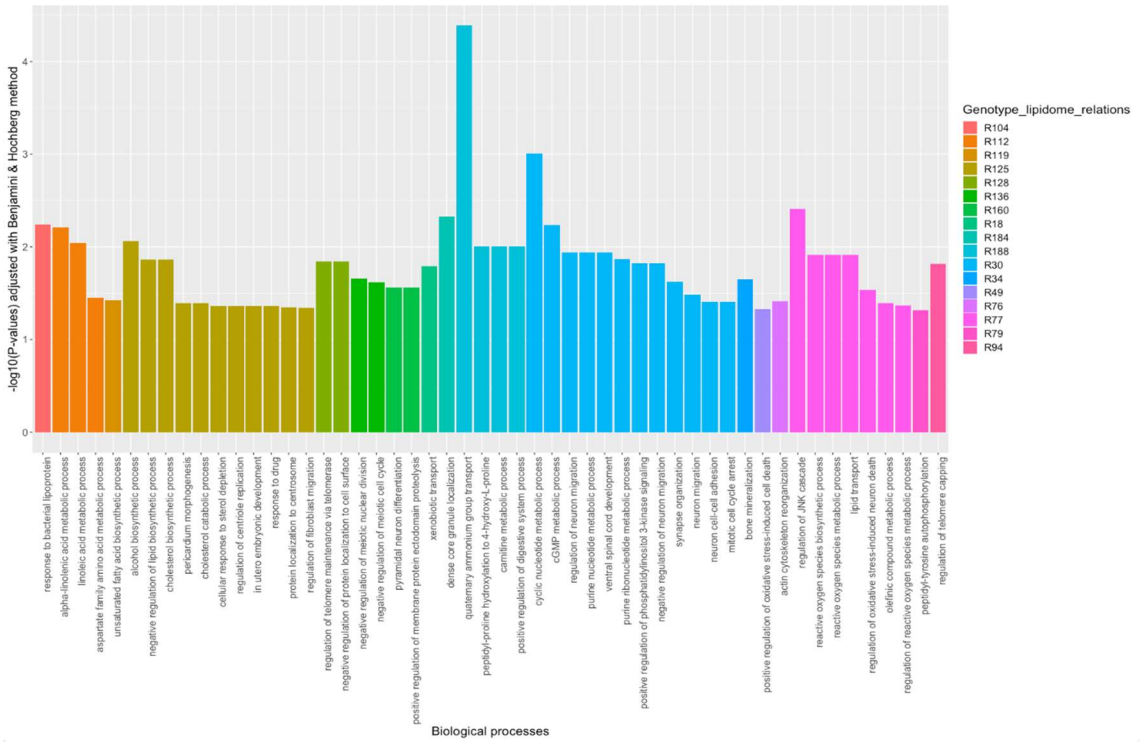

Figure 27. Biological processes (x-axis) from Gene Ontology database that are significantly (adjusted p-value < 0.05, y-axis) enriched in SNPs sets from the 18 out of the 35 genotype-lipidome relations.

8. DISCUSSION

The novelty of this study was that we have applied the combined GWAS-PGMRA-GSEA-pipeline for the first time in revealing the complex genetic background of human lipidome and its biological significance. This extended our understanding significantly beyond the current knowledge in the field, and we concluded that human plasma lipidome has at least 35 genetically distinct subgroups and are influenced by gene variations in 3164 genes via several biological processes.

The major idea to use the dimensionality reduction of lipidomic data by PCA was to test weather more computer intensive GWAS analysis of 437 separate lipids species could be replaced with this simpler reduced GWAS analysis of the five eigenlipids. However, the comparison of the results showed that only about 10 % of the lipidome associated SNPs are the same at p-value 5 x 10^{-7}. Therefore, for the final PGMRA analysis, we preselected SNPs based on traditional lipid-wise GWAS analysis, resulting a total of 18 370 different nominally associated lipidome SNPs.

A previous GWAS study (11) performed on 141 lipid species found 518 variants for 42 lipid species at statistical significance threshold of p-value= 1.5 x 10^{-9}. They found a total of 3754 associations (at p-level 5.0 x 10^{-8}) between SNPs and all studied lipid species in the study, with 821 different SNPs, associating with 35 different genes. In the present study, the SNPs involved in the PGMRA relations revealed 3164 different gene associations, a previous classical GWAS study (11) found 35 genes and this study found and replicated 13 of those 35 previous ones. The largest difference between this and the earlier study (11) comes from using quantitative measurements for the lipids and a wider LC-MS/MS based analysis platform of 437 lipids as compared to their 141. For comparison, while their GWAS found 3754 separate SNP associations located in 35 different genes with 141 lipids at p-level 5.0 x 10^{-8} and using ~9.3 million imputed genetic markers, the GWAS in this study found 266 different SNP-lipid association between 437 lipids (p-level 5.0 x 10^{-8}) and ~500 thousand genotyped (not imputed) genetic markers but located in 3164 different gene locus. Thus, the repeat of our GWAS-PGMRA analysis with imputed SNP data is warranted but requires the use of supercomputers.

The largest published GWAS study so far (93) with ~600 000 participants and 32 million genetic markers but looking only the four clinical lipid traits (HDL-cholesterol, LDL-cholesterol, total cholesterol and triglycerides) identified 826 independent lipid genome-wide significant variants. From these, 118 were novel (at p-level 5.0 x 10^{-8}) loci and also 268

previously identified (93, 94) loci were replicated. Further gene annotation using the SNP data of these loci and finding the related genes with Ensmbl VEP, resulted in a total of 535 different genes. From these novel findings, 18 of the 118 genes were replicated in the present study and of the previously identified 386 loci, 78 genes were replicated in a substantially lower number of subjects (~ N=2000 vs. ~600.000) using novel GWAS-PGMRA approach and unimputed SNP data.

The application of the novel person-centred methods proposed here will allow identification of distinct groups of genetic variants that contribute synergistically or additively to the risk of dyslipidaemias.

Here we searched for the effects of specific combinations of many genetic variants occurring simultaneously in the same persons, rather than the averaged effects of individual genes among many patients (“all cases”).

Furthermore, our approach describes complex patterns of interaction ("lock and key" genome-phenome combinations (95) between such sets of genetic variations and equally complex sets of individual features (i.e. 437 different lipids) that each person exhibits (96).

Given the degree of complexity of the plasma lipidome regulation, it’s very long developmental trajectory and its ongoing ability to change its own structure and function to adapt to environmental challenges, a reductionist characterization of lipid disorders as binary categories (case vs control) is virtually assured to miss nuances in the genetic, biological and behavioural descriptors of normal and abnormal lipoprotein metabolism functions. We therefore started from the premise that a better segmentation (identification of boundaries) between/among patients can be achieved by embracing complexity and incorporating as many features as possible *simultaneously* in an unbiased clustering process, without prejudging on what features, how many features or how many clusters are optimal.

This study, consequently, of this novel statistical approach, identified 93 statistically significant genotype-lipidome relations with 5977 SNPs located in 3164 different genes. Thirty-five of the significant relations contained distinct SNPs representing genetic lipidomic subgroups. While the more classical GWAS results would have independently only yielded 266 SNPs of statistical significance (p-value<5×10^{-8}), the GWAS-PGMRA yielded 5977 SNPs, greatly amplifying the existing GWAS based genetic results, i.e. in the SNPs in grey zone and revealing thus the hidden genetic architecture of lipidome.

Our present work is a ground-breaking resolution of the missing heritability problem and suggested to be applied in many other physical disorders beyond the case-control stratification. This is the first step in the direction of new genetic based classification of dyslipidemias but requires additional studies.

Every study like this has strength and limitations. First, results from this study could not be replicated due to the lack of comparable cohort's data. Thus, the major limitation in this study was that we have not replicated our results in another independent population-based cohort. However, to solve this limitation, the search of suitable replication cohorts is ongoing, and comparativeness requires the use of same lipidomic and genotypic platforms, thus for this reason not too many replication cohorts are available.

9. FURTHER DIRECTIONS

Using otherwise similar bioinformatic GWAS-PGMRA-GSEA pipeline but replacing the genotyped SNP-data set with substantially larger a 1000G or TOPMed (Imputation Server (nih.gov)) imputed gene variant data with 10-400 million SNPs is the next step for our analysis.

However, the use of imputed SNP-set will require significantly more computing time and use of supercomputers instead of home PC which was used for this MSc thesis.

A wider challenge for our team is to develop an open access version of GWAS-PGMRA-GSEA pipeline suitable for a wider scientific audience. This kind of new complementary biostatistical approaches is urgently needed.

In further studies, we also aim to link statistically found lipidomic SNP-set to the risk of various available phenotypes using additional SKAT-analysis. For example, connections of the found genome-lipidome clusters on cardiometabolic diseases like prediabetes, metabolic syndrome, type 2 diabetes, subclinical atherosclerosis, cognition, and different types of ICD-10 dyslipidaemia diagnosis is warranted.

10. SUMMARY AND CONCLUSIONS

Total of 442 (437 for lipids + 5 for PCA components) GWAS analyses were done in this project. We used non-imputed genotyped SNP data to avoid imputation related bias in genotyping.

In traditional lipidome wide GWAS analysis over whole lipidome, we identified 266 separate SNPs significantly associated at statistical significance threshold of p-value $<5 \times 10^{-8}$ and 337 SNPs at p-value $< 5 \times 10^{-7}$ and 18370 SNPs nominally significantly associated (p-value$<5 \times 10^{-4}$) with 437 studied molecular plasma lipids.

In PGMRA biclustering analysis, we used these nominally significant, preselected SNPs and lipid-phenotypes and found 93 statistically significant genotype-subject vs. lipid-phenotype group relations among between 71 subject-phenotype and 153 subject-genotype

clusters. The significant relations involved 5977 separate SNPs that after their gene annotation to latest available genome reference (ensemble assembly GRCh37, the version 102) were located into 3164 separate gene loci.

Thirty-five of the significant 93 SNP sets did not share any SNP or subject, therefore they represent 35 genetically distinct lipidomic profiles i.e. genetic lipidomic subgroups.

In GSEA SNP sets involved in 18 out of these 35 distinct genotype-lipidome relations were statistically significantly enriched in several biological processes throughout which the identified SNPs i.e. gene variations can influence to serum lipid profiles. Our results have also been published as a poster (97).

In the alternative eigenlipid based GWAS analysis, the eigenlipid classes were formed using PCA-analysis, by classifying the 437 lipids into 5 PCA (eigenlipids) groups. In their GWAS analysis we found 751 separate SNP-lipid associations at nominal p-value < $5x10^{-4}$ and 40 at p-value < $5x10^{-7}$. From these 32 of 40 were found also among the larger lipidome-wide GWAS analysis of different 337 SNPs ($p < 5x10^{-7}$), covering only 9,5 % (32/337) of the total SNPs found in wider lipid-wise GWAS.

We conclude that eigenlipid based GWAS analysis significantly reduces the total number of significant SNPs as compared to traditional GWAS analysis and therefore cannot be used as surrogate of lipidome-wides GWAS analysis without losing essential genetic information.

In this study, we have applied the combined GWAS-PGMRA-GSEA analysis pipeline for the first time in revealing the complex genetic background human lipidome and its biological significance. We conclude that human plasma lipidome has at least 35 genetically distinct subgroups and that the discovered genetic variants are influenced through 3164 genes via several biological processes.

In bioinformatic methodology perspective, the used of this novel bioinformatic approach reduces the missing heritability problem in lipidomic plasma patterns by uncovering the fraction of heritability distributed into independent networks of interacting genes that affect heterogeneous subsets of subjects and is missed by global averaging subjects across binary categories (cases vs controls).

11. REFERENCES

1. The top 10 causes of death. www.who.int/: 2018 May 24.

2. Tilastokeskus - Kuolemansyyt 2015 [Internet].: Helsinki: Tilastokeskus; 2015 [updated Jan; cited Nov 25, 2020]. Available from: https://www.stat.fi/til/ksyyt/2015/ksyyt_2015_2016-12-30_kat_001_fi.html.

3. Hilvo M, Meikle PJ, Pedersen ER, Tell GS, Dhar I, Brenner H, et al. Development and validation of a ceramide- and phospholipid-based cardiovascular risk estimation score for coronary artery disease patients. Eur Heart J. 2020 Jan 14;41(3):371-80.

4. Tetko IV, Engkvist O, Koch U, Reymond J, Chen H. BIGCHEM: Challenges and Opportunities for Big Data Analysis in Chemistry. Molecular informatics. 2016 Dec;35(11-12):615-21.

5. Mundra PA, Shaw JE, Meikle PJ. Lipidomic analyses in epidemiology. Int J Epidemiol. 2016 Oct;45(5):1329-38.

6. Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, et al. Discovery and refinement of loci associated with lipid levels. Nat Genet. 2013 Nov;45(11):1274-83.

7. Surakka I, Horikoshi M, Mägi R, Sarin A, Mahajan A, Lagou V, et al. The impact of low-frequency and rare variants on lipid levels. Nat Genet. 2015 Jun;47(6):589-97.

8. Bentley AR, Sung YJ, Brown MR, Winkler TW, Kraja AT, Ntalla I, et al. Multi-ancestry genome-wide gene-smoking interaction study of 387,272 individuals identifies new loci associated with serum lipids. Nat Genet. 2019 Apr;51(4):636-48.

9. Kettunen J, Demirkan A, Würtz P, Draisma HHM, Haller T, Rawal R, et al. Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. Nat Commun. 2016 Mar 23;7:11122.

10. Karjalainen J, Mononen N, Hutri-Kähönen N, Lehtimäki M, Juonala M, Ala-Korpela M, et al. The effect of apolipoprotein E polymorphism on serum metabolome - a population-based 10-year follow-up study. Sci Rep. 2019 Jan 24;9(1):458.

11. Tabassum R, Rämö JT, Ripatti P, Koskela JT, Kurki M, Karjalainen J, et al. Genetic architecture of human plasma lipidome and its link to cardiovascular disease. Nat Commun. 2019 Sep 24;10(1):4329.

12. Machine Learning for schizophrenia study [Internet]; 2015 [updated Feb; cited 8 Jun 2021]. Available from: https://dasci.es/research/outstanding-cientific-research/schizophrenia/.

13. Arnedo J, del Val C, de Erausquin GA, Romero-Zaliz R, Svrakic D, Cloninger CR, et al. PGMRA: a web server for (phenotype × genotype) many-to-many relation analysis in GWAS. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W142-9.

14. Zwir I, Arnedo J, Del-Val C, Pulkki-Råback L, Konte B, Yang SS, et al. Uncovering the complex genetics of human character. Molecular psychiatry. 2018 Oct 3;25(10):2295-312.

15. Zwir I, Arnedo J, Del-Val C, Pulkki-Råback L, Konte B, Yang SS, et al. Uncovering the complex genetics of human temperament. Molecular psychiatry. 2018 Oct 2;25(10):2275-94.

16. Zwir I, Mishra P, Del-Val C, Gu CC, de Erausquin GA, Lehtimäki T, et al. Uncovering the complex genetics of human personality: response from authors on the PGMRA Model. Molecular psychiatry. 2019 Mar 18;25(10):2210-3.

17. Taliun D, Kessler MD, Carlson J, Taliun SAG, Kang HM, Tian X, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature. 2021;590(7845):290-9.

18. Bergström A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, et al. Insights into human genetic variation and population history from 929 diverse genomes. Science. 2020 Mar 20;367(6484).

19. Ho SS, Urban AE, Mills RE. Structural Variation in the Sequencing Era: Comprehensive Discovery and Integration. Nat Rev Genet. 2020 Mar;21(3):171-89.

20. Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. Genome Biol. 2019 Jun 3;20(1):117.

21. lipid | Definition, Structure, Examples, Functions, Types, & Facts [Internet]. [cited Jun 7, 2021]. Available from: https://www.britannica.com/science/lipid.

22. O'Donnell VB, Ekroos K, Liebisch G, Wakelam M. Lipidomics: Current state of the art in a fast moving field. Wiley Interdiscip Rev Syst Biol Med. 2020 Jan;12(1):e1466.

23. Fahy E, Cotter D, Sud M, Subramaniam S. Lipid classification, structures and tools. Biochim Biophys Acta. 2011 Nov;1811(11):637-47.

24. Taylor R, Miller RH, Miller RD, Porter M, Dalgleish J, Prince JT. Automated structural classification of lipids by machine learning. Bioinformatics. 2015 Mar 1;31(5):621-5.

25. Li B, Tang J, Yang Q, Li S, Cui X, Li Y, et al. NOREVA: normalization and evaluation of MS-based metabolomics data. Nucleic Acids Res. 2017 Jul 3;45(W1):W162-70.

26. Hancock SE, Friedrich MG, Mitchell TW, Truscott RJW, Else PL. The phospholipid composition of the human entorhinal cortex remains relatively stable over 80 years of adult aging. Geroscience. 2017 Feb;39(1):73-82.

27. Proitsi P, Kim M, Whiley L, Simmons A, Sattlecker M, Velayudhan L, et al. Association of blood lipids with Alzheimer's disease: A comprehensive lipidomics analysis. Alzheimers Dement. 2017 Feb;13(2):140-51.

28. Yu Q, He Z, Zubkov D, Huang S, Kurochkin I, Yang X, et al. Lipidome alterations in human prefrontal cortex during development, aging, and cognitive disorders. Mol Psychiatry. 2020 Nov;25(11):2952-69.

29. Tkachev A, Stepanova V, Zhang L, Khrameeva E, Zubkov D, Giavalisco P, et al. Differences in lipidome and metabolome organization of prefrontal cortex among human populations. Sci Rep. 2019 Dec 4;9(1):18348.

30. Hilvo M, Meikle PJ, Pedersen ER, Tell GS, Dhar I, Brenner H, et al. Development and validation of a ceramide- and phospholipid-based cardiovascular risk estimation score for coronary artery disease patients. Eur Heart J. 2020 Jan 14;41(3):371-80.

31. Karjalainen J, Mononen N, Hutri-Kähönen N, Lehtimäki M, Hilvo M, Kauhanen D, et al. New evidence from plasma ceramides links apoE polymorphism to greater risk of coronary artery disease in Finnish adults. J Lipid Res. 2019 Sep;60(9):1622-9.

32. Mishra BH, Mishra PP, Mononen N, Hilvo M, Sievänen H, Juonala M, et al. Lipidomic architecture shared by subclinical markers of osteoporosis and atherosclerosis: The Cardiovascular Risk in Young Finns Study. Bone. 2020 Feb;131:115160.

33. Illumina Microarray Technology [Internet]. [cited Jun 7, 2021]. Available from: https://emea.illumina.com/science/technology/microarray.html.

34. Zajac GJM, Fritsche LG, Weinstock JS, Dagenais SL, Lyons RH, Brummett CM, et al. Estimation of DNA contamination and its sources in genotyped samples. Genet Epidemiol. 2019 Dec;43(8):980-95.

35. Tortajada-Genaro LA, Yamanaka ES, Maquieira Á. Consumer electronics devices for DNA genotyping based on loop-mediated isothermal amplification and array hybridisation. Talanta. 2019 Jun 1;198:424-31.

36. Rodríguez-Morató J, Pozo ÓJ, Marcos J. Targeting human urinary metabolome by LC-MS/MS: a review. Bioanalysis. 2018 Apr 1;10(7):489-516.

37. Sok P, Lupo PJ, Richard MA, Rabin KR, Ehli EA, Kallsen NA, et al. Utilization of archived neonatal dried blood spots for genome-wide genotyping. PLoS One. 2020;15(2):e0229352.

38. Kumar D, Chhokar V, Sheoran S, Singh R, Sharma P, Jaiswal S, et al. Characterization of genetic diversity and population structure in wheat using array based SNP markers. Mol Biol Rep. 2020 Jan;47(1):293-306.

39. Grewal S, Hubbart-Edwards S, Yang C, Devi U, Baker L, Heath J, et al. Rapid identification of homozygosity and site of wild relative introgressions in wheat through chromosome-specific KASP genotyping assays. Plant Biotechnol J. 2020 Mar;18(3):743-55.

40. LaBarre BA, Goncearenco A, Petrykowska HM, Jaratlerdsiri W, Bornman MSR, Hayes VM, et al. MethylToSNP: identifying SNPs in Illumina DNA methylation array data. Epigenetics Chromatin. 2019 Dec 20;12(1):79.

41. Morton EA, Hall AN, Kwan E, Mok C, Queitsch K, Nandakumar V, et al. Challenges and Approaches to Genotyping Repetitive DNA. G3 (Bethesda). 2020 Jan 7;10(1):417-30.

42. Rounge TB, Lauritzen M, Erlandsen SE, Langseth H, Holmen OL, Gislefoss RE. Ultralow amounts of DNA from long-term archived serum samples produce quality genotypes. Eur J Hum Genet. 2020 Apr;28(4):521-4.

43. Lo Giudice C, Pesole G, Picardi E. High-Throughput Sequencing to Detect DNA-RNA Changes. Methods Mol Biol. 2021;2181:193-212.

44. Li X, Yin F, Xu X, Liu L, Xue Q, Tong L, et al. A facile DNA/RNA nanoflower for sensitive imaging of telomerase RNA in living cells based on "zipper lock-and-key" strategy. Biosens Bioelectron. 2020 Jan 1;147:111788.

45. Kono N, Arakawa K. Nanopore sequencing: Review of potential applications in functional genomics. Dev Growth Differ. 2019 Jun;61(5):316-26.

46. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol. 2018 Apr;36(4):338-45.

47. Vogeser M, Parhofer KG. Liquid chromatography tandem-mass spectrometry (LC-MS/MS)--technique and applications in endocrinology. Exp Clin Endocrinol Diabetes. 2007 Oct;115(9):559-70.

48. Keevil BG. LC-MS/MS analysis of steroids in the clinical laboratory. Clin Biochem. 2016 Sep;49(13-14):989-97.

49. Tyanova S, Temu T, Carlson A, Sinitcyn P, Mann M, Cox J. Visualization of LC-MS/MS proteomics data in MaxQuant. Proteomics. 2015 Apr;15(8):1453-6.

50. van den Broek I, Sobhani K, Van Eyk JE. Advances in quantifying apolipoproteins using LC-MS/MS technology: implications for the clinic. Expert Rev Proteomics. 2017 Oct;14(10):869-80.

51. Nelson CP, Goel A, Butterworth AS, Kanoni S, Webb TR, Marouli E, et al. Association analyses based on false discovery rate implicate new loci for coronary artery disease. Nat Genet. 2017 Jul 17;49(9):1385-91.

52. Evangelou E, Warren HR, Mosen-Ansorena D, Mifsud B, Pazoki R, Gao H, et al. Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. Nat Genet. 2018 Sep 17;50(10):1412-25.

53. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature. 2001 Feb 15;409(6822):928-33.

54. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, et al. The structure of haplotype blocks in the human genome. Science. 2002 Jun 21;296(5576):2225-9.

55. A haplotype map of the human genome. Nature. 2005 Oct 27;437(7063):1299-320.

56. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet. 2010 Jul;11(7):499-511.

57. Klein RJ, Zeiss C, Chew EY, Tsai J, Sackler RS, Haynes C, et al. Complement factor H polymorphism in age-related macular degeneration. Science. 2005 Apr 15;308(5720):385-9.

58. Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. Natural selection has driven population differentiation in modern humans. Nat Genet. 2008 Mar;40(3):340-5.

59. Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, Tsunoda T, et al. Functional SNPs in the lymphotoxin-[alpha] gene that are associated with susceptibility to myocardial infarction. Nat Genet. 2002 Dec 1;32(4):650.

60. Bär C, Chatterjee S, Thum T. Long Noncoding RNAs in Cardiovascular Pathology, Diagnosis, and Therapy. Circulation. 2016 Nov 08;134(19):1484-99.

61. Ji Z, Song R, Regev A, Struhl K. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. Elife. 2015 Dec 19;4:e08890.

62. Ulitsky I. Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. Nat Rev Genet. 2016 Oct;17(10):601-14.

63. Fang Y, Fullwood MJ. Roles, Functions, and Mechanisms of Long Non-coding RNAs in Cancer. Genomics Proteomics Bioinformatics. 2016 Feb;14(1):42-54.

64. Quinn JJ, Chang HY. Unique features of long non-coding RNA biogenesis and function. Nat Rev Genet. 2016 Jan;17(1):47-62.

65. Chandrasekaran K, Setyowati K, Sepramaniam S, Armugam A, Wintour E, Bertram J, et al. Role of microRNAs in kidney homeostasis and disease. Kidney international. 2012 January 11;81:617-27.

66. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. American journal of human genetics. 2007 Sep;81(3):559-75.

67. LINEAARISET REGRESSIOMALLIT [Internet]; 2010 [updated Jun 17; cited Nov 30, 2020]. Available from: http://myy.haaga-helia.fi/~taaak/m/regressio.pdf.

68. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. Nat Rev Genet. 2019 Aug;20(8):467-84.

69. Andreopoulos B, An A, Wang X, Schroeder M. A roadmap of clustering algorithms: finding a match for a biomedical application. Briefings in bioinformatics. 2008 Dec 6;10(3):297-314.

70. Oghabian A, Kilpinen S, Hautaniemi S, Czeizler E. Biclustering Methods: Biological Relevance and Application in Gene Expression Analysis. PloS one. 2014 Mar 20;9(3):e90801.

71. Hu CW, Kornblau SM, Slater JH, Qutub AA. Progeny Clustering: A Method to Identify Biological Phenotypes. Scientific reports. 2015 Aug 12;5(1):12894.

72. Wallace T, Sekmen A, Wang X. Application of Subspace Clustering in DNA Sequence Analysis. Journal of computational biology. 2015 Oct 1;22(10):94-952.

73. Ruspini EH, Bezdek JC, Keller JM. Fuzzy Clustering: A Historical Perspective. MCI. 2019 Feb;14(1):45-55.

74. A Toolbox for Bicluster Analysis in R [Internet]; 2008 [cited Jan 25, 2021]. Available from: https://docplayer.net/7182007-A-toolbox-for-bicluster-analysis-in-r.html.

75. RcmdrPlugin.BiclustGUI: 'Rcmdr' Plug-in GUI for Biclustering [Internet].: Comprehensive R Archive Network (CRAN); 2020 [updated Jul; cited Jun 13, 2021]. Available from: https://CRAN.R-project.org/package=RcmdrPlugin.BiclustGUI.

76. Analysis of Biological Networks:
Network Modules – Clustering and Biclustering [Internet]; 2006 [updated Nov 23; cited Nov 30, 2020]. Available from: http://www.cs.tau.ac.il/~roded/courses/bnet-a06/lec05.pdf.

77. Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. BMC bioinformatics. 2010 Jul 2;11(1):367.

78. Generalized Nonnegative Matrix
Approximations with Bregman Divergences [Internet]; 2005 [cited Dec 3, 2020]. Available from: https://papers.nips.cc/paper/2005/file/d58e2f077670f4de9cd7963c857f2534-Paper.pdf.

79. Nonnegative Matrix Factorization
Clustering on Multiple Manifolds [Internet]; 2010 [updated Sep 27; cited Dec 3, 2020]. Available from: https://www.cs.purdue.edu/homes/lsi/AAAI_2010_NMF_Clustering_MM.pdf.

80. Mingyi He, Feng Wei, Xiuping Jia. Globally maximizing, locally minimizing: Regularized Nonnegative Matrix Factorization for hyperspectral data feature extraction. IEEE; Jun 2012.

81. Pan W. Relationship between Genomic Distance-Based Regression and Kernel Machine Regression for Multi-marker Association Testing. Genet Epidemiol. 2011 May;35(4):211-6.

82. Larson NB, Chen J, Schaid DJ. A Review of Kernel Methods for Genetic Association Studies. Genet Epidemiol. 2019 Mar;43(2):122-36.

83. Berkopec A. HyperQuick algorithm for discrete hypergeometric distribution. Journal of discrete algorithms (Amsterdam, Netherlands). 2007;5(2):341-7.

84. Raitakari OT, Juonala M, Rönnemaa T, Keltikangas-Järvinen L, Räsänen L, Pietikäinen M, et al. Cohort profile: the cardiovascular risk in Young Finns Study. Int J Epidemiol. 2008 Dec;37(6):1220-6.

85. Smith EN, Chen W, Kähönen M, Kettunen J, Lehtimäki T, Peltonen L, et al. Longitudinal genome-wide association of cardiovascular disease risk factors in the Bogalusa heart study. PLoS Genet. 2010 Sep 9;6(9):e1001094.

86. Mamtani M, Kulkarni H, Wong G, Weir JM, Barlow CK, Dyer TD, et al. Lipidomic risk score independently and cost-effectively predicts risk of future type 2 diabetes: results from diverse cohorts. Lipids in health and disease. 2016 Apr 4;15(1):67.

87. Braicu EI, Darb-Esfahani S, Schmitt WD, Koistinen KM, Heiskanen L, Pöhö P, et al. High-grade ovarian serous carcinoma patients exhibit profound alterations in lipid metabolism. Oncotarget. 2017 Nov 28;8(61):102912-22.

88. Mishra BH, Mishra PP, Mononen N, Hilvo M, Sievänen H, Juonala M, et al. Uncovering the shared lipidomic markers of subclinical osteoporosis-atherosclerosis comorbidity: The young Finns study. Bone (New York, N.Y.). 2021 Jun:116030.

89. bioNMF: a versatile tool for non-negative matrix factorization in biology [Internet].: BioMed Central; 2006 [cited Nov 30, 2020]. Available from: https://search.datacite.org/works/10.5167/uzh-24.

90. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016 Jun 6;17(1):122.

91. Edwards NC, Hing ZA, Perry A, Blaisdell A, Kopelman DB, Fathke R, et al. Characterization of Coding Synonymous and Non-Synonymous Variants in ADAMTS13 Using Ex Vivo and In Silico Approaches. PLOS ONE. 2012 Jun 29;7(6):e38864.

92. Walsh IM, Bowman MA, Soto Santarriaga IF, Rodriguez A, Clark PL. Synonymous codon substitutions perturb cotranslational protein folding in vivo and impair cell fitness. Proc Natl Acad Sci U S A. 2020 Feb 18;117(7):3528-34.

93. Klarin D, Damrauer SM, Cho K, Sun YV, Teslovich TM, Honerlaw J, et al. Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. Nat Genet. 2018;50(11):1514-23.

94. Liu DJ, Peloso GM, Yu H, Butterworth AS, Wang X, Mahajan A, et al. Exome-wide association study of plasma lipids in >300,000 individuals. Nat Genet. 2017 Dec;49(12):1758-66.

95. Houle D, Govindaraju DR, Omholt S. Phenomics: the next challenge. Nat Rev Genet. 2010 Dec;11(12):855-66.

96. Arnedo J, Svrakic DM, Del Val C, Romero-Zaliz R, Hernández-Cuervo H, Fanous AH, et al. Uncovering the hidden risk architecture of the schizophrenias: confirmation in three independent genome-wide association studies. Am J Psychiatry. 2015 Feb 1;172(2):139-53.

97. Mishra BH, Lehtimäki M, Mishra PP, Arnedo J, del Val C, Zwir I, et al. Uncovering the complex genetic architecture of human lipidome. Session Epidemiology and dyslipidemias poster no 1202. 89th EAS Congress May 30- June 2, 2021, Helsinki, Finland.

98. Lusis AJ. Genetics of atherosclerosis. Trends in Genetics. 2012 June 1;28(6):267-75.

99. Valentina Pallottini. Role of mevalonate pathway in the central nervous system. Development Biology. Université de Strasbourg, 2017. English. NNT : 2017STRAJ096. tel-01764119

101. Chromacademy LCMS Intro [Internet]; 2014 [updated Feb 15; cited Jun 7, 2021]. Available from: http://www.ecs.umass.edu/eve/background/methods/chemical/Openlit/Chromacademy%20LCMS%20Intro.pdf.

102. Sukhumsirichart W. Polymorphisms. In: Genetic Diversity and Disease Susceptibility; 2018; doi: 10.5772 / intechopen.76728.

103. Definition of transcription - NCI Dictionary of Genetics Terms - National Cancer Institute [Internet]; 2012 [updated Jul 20 2020; cited Jun 7, 2021]. Available from: https://www.cancer.gov/publications/dictionaries/genetics-dictionary/def/transcription.

104. Data Mining Algorithms In R/Clustering/Biclust - Wikibooks, open books for an open world [Internet]; 2020 [updated Apr 16; cited Jun 7, 2021]. Available from: https://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Clustering/Biclust.

105. Pontes B, Giráldez R, Aguilar-Ruiz JS. Journal of biomedical informatics. 2016 Aug 31;57:163-80.