



FACULTY OF MEDICINE

Master's Programme in Public Health

Salivary amylase gene copy number and plasma metabolites in the Malmö diet and cancer cohort study

Joshua Mendis

Supervisor: Emily Sonestedt

Department of Clinical Sciences, Malmö

June 2023

Abstract

Background: The number of copies of the salivary amylase gene (*AMY1*) a person carries is directly proportional to the amount of salivary amylase they produce. Salivary amylase is responsible for the first steps in chemical digestion by breaking starches into smaller intermediate molecules known as metabolites. *AMY1* copy number variation (CNV) may play a moderating role in plasma (blood) metabolite concentrations. Quantifying associations between *AMY1* CNV and metabolite concentrations may help to understand its role in disease risk and adiposity moderation.

Aim: To determine whether any plasma metabolites were associated with *AMY1* CNV, and whether any of those metabolites were in turn associated with body mass index (BMI) measurements.

Methods: Metabolomic data (822 metabolites), *AMY1* copy number, and anthropometric data were available among 826 participants from the Malmö diet and cancer cohort study (MDCS). Elastic net regression (ENR) was used to identify the most important plasma metabolite concentrations associated with *AMY1* CNV. The top metabolites were run through linear regression models against *AMY1* CNV and BMI.

Results: Elastic net regression (ENR) of metabolomic data from the Malmö Diet and Cancer cohort study (MDCS) revealed 18 metabolites that were associated with *AMY1* CNV. Linear regression of metabolite concentration against *AMY1* CNV revealed one metabolite, theophylline that showed a significant association with *AMY1* CNV when adjusted for age and sex ($p < 0.001$). In addition, 7 other metabolites, 1-oleoyl-2-docosahexaenoyl-GPC, 2-hydroxyisobutyrate/2-hydroxybutyrate, hydantoin-5-propionate, 2-hydroxy-4(methylthio)butanoic acid, glucose, 2-oxoarginine, and pipecolate were found to have significant associations with BMI ($p < 0.001$). At a population level, *AMY1* copy number was not significantly associated with BMI.

Conclusion: Elastic net regression of metabolomic data from the MDCS revealed 1 metabolite associated with *AMY1* CNV, and 7 associated with BMI. Their biochemical associations with BMI must be elucidated in future research.

Key Words: Metabolites, diet, obesity, adiposity, *AMY1*, salivary amylase, elastic net regression, copy number, gene.

Table of contents

Abbreviations	4
1-Introduction	5-8
1.1: Salivary Amylase and Starches.....	5
1.2: AMY1 and copy number variation	5
1.3: The Metabolome and Metabolomics.	6
1.4: Current perspectives regarding the impact of AMY1 CNV	7
1.5: Links between AMY1 and BMI.....	8
2-Methods	8-13
2.1: The Malmö Diet and Cancer Cohort Study (MDCS)	8
2.2: Dietary and Lifestyle assessment.....	9
2.3: AMY1 CNV measurement	9
2.4: Plasma metabolite measurement	10
2.5: Elastic Net Regression	10
2.6: Analysis of data.....	12
2.7: BMI sensitivity analysis	12
2.8: Data Visualization and multicollinearity.....	12
2.9: Ethical Considerations.....	13
3-Results	13-15
3.1: Population Results	13
3.2: Elastic Net regression output.....	14
3.3: Metabolite Collinearity.....	14
3.4: AMY1 analysis	14
3.5: BMI sensitivity analysis	14
3.6: Profile Score sensitivity analysis	15
3.7: Summary	15
4-Discussion	15-26
4.1: glucose	16
4.2: theophylline.....	17
4.3: 1-oleoyl-2-docosahexaenoyl-GPC..	17
4.4: 2-hydroxybutyrate/2- hydroxyisobutyrate.....	18
4.5: hydantoin-5-propionate	19
4.6: 2-hydroxy-4-(methylthio) butanoic acid.....	20
4.7: 2-oxoarginine	20
4.8: pipecolate	21
4.9: Metabolites, AMY1 CNV and microbiota	21
4.10: Validity of methods.....	22
4.11: Limitations	24
5-Public Health context for research .	26-29
5.1: Burden of noncommunicable disease	26
5.2: Preventative care using metabolomics	26
5.3: Future Metabolomic studies	27
6-Conclusion	28-29
7-References	30-35
8-Tables and Figures	36-47
9-Popular Science summary	48
10-Acknowledgments	49

Abbreviations

AMY1- Gene that codes for salivary amylase

ANOVA- Analysis of variance

BMI- Body mass index

CNV- Copy number variation

CVD- cardiovascular disease

ddpcr- Droplet digital polymerase chain reaction

ENR- Elastic net regression

HDL- High density lipoprotein

LPC- Lysophosphatidylcholine

MDC-CC- Malmö diet and cancer cardiovascular cohort

MDCS- Malmö diet and cancer cohort study

METs - metabolic equivalents

MSE-Mean standard error

PC- Phosphatidylcholine

SCFA-Short chain fatty acid

T2D- Type 2 diabetes

1. Introduction

1.1: Salivary Amylase and Starches

Salivary amylase is an enzyme found in the oral cavity of humans (1). Salivary amylase belong to a class of enzymes known as alpha amylases due to the type of the bonds in carbohydrates that they break (2, 3, 4). Carbohydrates generally represent a large portion of diet composition and are metabolized into structures that can be used by the body in cellular respiration (5, 6). Certain diet patterns with varying levels of carbohydrate consumption are associated with differing risks for noncommunicable diseases, though the biological mechanisms mediating these differences are not well understood (7).

Starches are long chain carbohydrates that are made up of two main types of compounds, amylopectin and amylose (8). Bonded amylose and amylopectin are known as starch granules, water insoluble structures that are highly varied in size (9). Amylopectin is composed of glucosyl units bonded through alpha 1,4 and alpha 1,6 linkages. These two types of chemical bonds result in formation of highly branched structures composed of short and long chains of glucosyl units and comprise 65%-85% of the entire molecule (8). Amylose is also made of glucosyl units with alpha 1,4 linkages, though it is more linear in shape compared to amylopectin (8).

Salivary amylase initiates starch digestion in the mouth, and is active throughout the digestive tract (10). Salivary amylase breaks the alpha 1,4 linkages in amylose and amylopectin to yield intermediate molecules made of maltose units including maltooligosaccharides, isomaltooligosaccharides, and maltotriose (8). These complexes are further hydrolyzed by enzymes such as maltase, ultimately yielding glucose to be used in cellular respiration (8).

1.2: AMY1 and copy number variation

A gene is a sequence of nucleotide base pairs on a chromosome, alterations of DNA base pairs result in changes to genetic information, creating the phenotypic diversity seen throughout species (11, 12). Insertions or deletions of more than 1000 base pairs on a chromosome are known as structural variations. A copy number variation is where an entire sequence coding for a gene is duplicated or deleted (13). *AMY1B* is the gene that encodes salivary amylase, this gene is found within all three domains of life and is highly expressed in the salivary glands of mammals (3, 9, 14, 15). There are a multitude of amylase coding genes in the human genome; *AMY1A*, *AMY1B*,

AMY1B, *AMY2A*, and *AMY2B* (9). The gene that codes for salivary amylase, *AMY1B*, was produced through the duplication of the pancreatic amylase gene *AMY2B* (9). Copy number variation has occurred with the *AMY1B* gene in the human genome, resulting in a high variability in the number of copies an individual carries (14). Humans carry between 2-17 *AMY1* copies, in contrast to chimpanzees who only carry 2 copies (3, 16, 17). Throughout the course of human evolution, the *AMY1B* gene has undergone several major duplication events (18). Human populations exhibit vastly different *AMY1* CNV, with higher mean *AMY1* copy number found in groups with a generally higher level of starch intake (4, 14, 15, 16, 17). This suggests a selective pressure for high *AMY1* copy number with human migration (4, 17, 19). Evolutionary biologists have theorized that increased *AMY1* CNV is associated with development of cooking or agricultural skills (4, 6, 18, 19). As human diet patterns shifted, starch consumption also increased, leading to an evolutionary advantage in carrying increased numbers of *AMY1* (4, 6, 18, 19).

1.3: The Metabolome and Metabolomics

Advances in mass spectroscopy and nuclear magnetic resonance technology have allowed for the quantification of large numbers of chemical compounds found within the human body. This has created several areas of “omic” study all based in differing phases of the translation of DNA into functional products (20). All the metabolic intermediates within a sample are known as the “metabolome” and the study of these intermediates is known as metabolomics (20). Human metabolomes are studied in stool, cell cultures, saliva, and blood plasma (21). The metabolome can be thought of as a snapshot of the metabolic state of the participant and demonstrates how an individual is responding to a variety of exposures. These exposures could include long-term diet patterns, medications, and lifestyle choices like smoking, alcohol consumption, or exercise (20). Changes to the metabolome can be caused by a variety of external factors including exposure to certain pollutants/chemicals (22). Changes in metabolite concentrations over time have been associated with increased risks for certain CVDs (23). Increased adiposity has been shown to alter plasma metabolite concentrations over time, and previous metabolomic studies have quantified these changes in lipids, amino acids, and other biochemical intermediates (24). Ottosson *et al.* concluded that certain metabolites concentration changes were associated with the onset of type 2 diabetes and coronary artery disease (25). These findings make metabolomics a useful method to elucidate the underlying processes that lead to the onset of noncommunicable diseases.

1.4: Current perspectives regarding the impact of AMY1 CNV

AMY1 CNV has an additive effect on initial digestion of starch molecules (3, 14, 17). As each copy of the *AMY1* gene contains the coding necessary for translation of salivary amylase, the amount of copies a person carries is directly proportionate to the concentration of amylase enzyme present in their saliva (17). A person who carries a low number of *AMY1* copies may not be able to metabolize starches as efficiently as a person with a high *AMY1* copy number creating downstream effects on digestion (14). Low *AMY1* CNV may also be associated with increased risk for diabetes and increased adiposity in certain populations (19, 26).

Hamid *et al.* investigated the association between *AMY1* CNV and starch intake on glucose homeostasis and risk for type 2 diabetes (T2D) (14). Studying 3624 individuals from the MDC, they concluded that individuals with the highest *AMY1* CNV and the lowest starch intake had the highest blood insulin levels (14). They also found an inverse association between starch intake and T2D in individuals with 10 or more copies of *AMY1* (14). Within the group of individuals with more than 10 copies of *AMY1* who also consumed high levels of starch, there was a lower overall risk for type 2 diabetes (14).

Mandel *et al.* investigated the role of high *AMY1* CNV in post-prandial glycemic responses to starch in a group of 14 people (27). They concluded that individuals with high *AMY1* copy numbers had more efficient responses to starch ingestion, resulting in higher blood insulin levels after food consumption as compared to the low *AMY1* copy number group (27). They hypothesized that individuals with low amylase activity may be at a higher risk of insulin resistance or diabetes if those individuals have long-term diet patterns with high starch contents (27). Marquina *et al.* studied the relationship between low *AMY1* CNV and cardiometabolic risk factors in 57 participants (28). They found that participants with low *AMY1* CNV had increased levels of CVD related biomarkers when compared to those with higher *AMY1* CNV (28). The study also hypothesized that in high adiposity individuals, low *AMY1* CNV was associated with increased CVD risk (28).

1.5: Links between *AMY1* and BMI

The prevalence of having increased adipose or fat tissue has reached epidemic proportions, tripling in many populations since 1975 (29). In 2016, at least 1.6 billion people were classified as overweight or obese, with overweight defined as a body mass index (BMI) of 25 to 29, and obesity meaning a BMI of 30 or above (29). BMI as a trait has been known to have a large genetic component, with some twin and family-based studies estimating that up to 80% of its variance is attributable to genetic factors (30, 31). Rukh *et al.* investigated the role of *AMY1* CNV and starch intake on adiposity risk and BMI (32). With 4800 individuals from the MDCS, Rukh was not able to find an association between *AMY1* CNV and BMI or body fat percentage (32). The study did find a significant association between *AMY1* CNV and starch intake on BMI (32).

Falchi *et al.* reported a significant association between low *AMY1* copy number and increased BMI in a group of 6200 European adults (33). In a group of 61 participants with a history of childhood onset adiposity, Viljakainen found a significant association between *AMY1* CNV and BMI/ whole body fat percentage in females (34). Marcovecchio *et al.* observed a significant association between low *AMY1* CNV and increased BMI/ waist circumference in a group of 354 Italian boys but found no association in 390 girls of the same age range (35). Grammatikopoulou *et al.* quantified *AMY1* CNV in 43 end stage renal patients, and found *AMY1* CNV to be associated with body weight, waist/ hip circumference, and fat mass (19).

Taken together, these studies have come to differing conclusions regarding the direct role of *AMY1* on BMI. This study aims to determine if any measured metabolites were associated with *AMY1* CNV, and to then test whether any of those metabolites were related to BMI (see figure 1).

2. Methods

2.1: The Malmö Diet and Cancer Cohort Study (MDCS)

This evaluation will be done through statistical analyses of data from the Malmö diet and cancer cohort study (MDCS). The MDCS is a prospective cohort study with the initial scope of understanding the relationship between long-term diet patterns and varying health outcomes of interest (36, 37). This ongoing study was initiated in 1991 and recruited 30,447 men and women in southern Sweden who were born between 1923 and 1945 (37). During the intake period, individuals underwent baseline examinations between 1991 and 1996 (37). Participants had

detailed lifestyle traits, physiological measurements, and long-term disease measured at baseline (37). In addition, disease outcomes have linked to the cohort through various registers. Participants who underwent baseline examinations between 1991 and 1994 were randomly selected to join the cardiovascular subcohort, the MDC-CC, resulting in a group of around 6000 participants who would be recalled for further examinations and blood sampling (38). At the time of sampling, the age range of participants was 45-68 years of age, with a median age of 58 years. This resulted in an initial data set comprising 1083 people. After removal of participants with incomplete information, the final study population comprised of 826 participants (46% male, 54% female).

2.2: Dietary and Lifestyle assessment

Anthropometric data, dietary habits and lifestyle factors patterns were recorded and collected at baseline (32, 36, 39). Habitual dietary habits were assessed through a 7-day food diary, a 168-item food frequency questionnaire, and an interview to confirm data from recorded materials (36, 37). BMI was defined as weight in kilograms divided by height in square meters (see table 1). Mean daily food intake was calculated based on the information available in the food diary, questionnaire/ interview, and then combined to form nutrient intake values (32). Nutrient intake values were generated through the Malmö diet and cancer food and nutrient database (32). This data was then used to calculate micronutrient consumption, total consumed energy (kJ), and proportion of total energy from starch (see table 1) (32, 36, 37). Starch intake was calculated by subtracting monosaccharides and disaccharides from total carbohydrate intake (32). Lifestyle/socioeconomic information was collected through a questionnaire, with education being broken into elementary (<8 y), primary or secondary (9–10 y), upper secondary (11–13 y), further education without a degree, or university degree (see table 1) (32). Smoking was divided into current smokers (including irregular smoking habits), non, and ex-smokers (see table 1) (32). Alcohol consumption was broken into sex specific quintiles with participants who reported no consumption of alcohol in the questionnaire categorized as zero consumers (see table 1) (32). Physical activity was measured in metabolic equivalents (METs) through interviews and self reported questionnaires (see table 1) (32).

2.3: AMY1 CNV measurement

AMY1 CNV was determined through droplet digital polymerase chain reaction (ddPCR) using the QX200 AutoDG ddPCR system from Bio-rad Laboratories (32, 36). PCRs were run following the

manufacturer's instructions as outlined by Li *et al.*, and Rukh *et al.* (32, 36). The following reagents were used for each assay of sample plates: 11 mL of ddPCR Supermix for Probes (no 2'-deoxyuridine 5'-triphosphate); 0.8 mL of *AMY1* assay (assay ID dHsaCP1000594); 0.8 mL of adaptor related protein complex 3 beta 1 assay (assay ID dHsaCP1000001); 1 mL of HaeIII (3 U/mL); 4 mL of DNA (5 ng/mL); 4.4 mL of distilled water and automated droplet generation oil. After amplification, plates were stored at 4°C and were read by a Bio-Rad QX200 Droplet Reader (32).

2.4: Plasma metabolite measurement

Fasting blood plasma samples were taken from participants from the cardiovascular sub-cohort of the MDC-CC and stored at -80°C (14, 36). In total, 1083 blood samples underwent Liquid Chromatography-Tandem Mass Spectroscopy to reveal 1372 biochemicals. Using the Metabolon database, 1104 of these compounds were determined to be of known identity, and 268 compounds of unknown structural identity (36, 40). These metabolites were classified into a various biochemical pathways and sub pathways, useful for identifying their role in downstream functions in the body (41).

The 268 unknown metabolites were not included in analysis as there was insufficient information regarding their associated biochemical pathways. In addition, 282 of the known metabolites were excluded due to a high level of missing data (greater than 75%). The metabolites were batch-normalized, log-transformed, and standardized to Z-scores. Extreme values (above or below 5 SD from the mean) were recoded to the threshold values for each metabolite prior to the statistical analyses. This data was then merged with the patients *AMY1* CNV data to produce the final dataset comprising of 826 individuals with 822 complete metabolite profiles that were ready for statistical analyses (see figure 2).

2.5: Elastic Net Regression

Associations between the *AMY1* CNV and concentrations of blood plasma metabolites were ascertained through models created by gaussian linear regressions with elastic net penalties. Elastic net regression (ENR) involves testing many regression models to find the optimal penalties to fit the training data set. ENR models have tuning parameters that transform regression coefficients for variables of interest. This is useful in metabolomic datasets as the tuning parameters make the

regression coefficients more uniform, helping to avoid collinearity. This allows one to determine the most effective tuning parameters for a robust model that can accurately represent outcomes in the test data set.

The first tuning parameter is the alpha tuning parameter, whose function is to regularize complex regression models by introducing penalties to restrict large regression coefficients. This parameter is an essential component for dealing with multicollinearity in regression models and can help reduce model overfitting. Overfitting to the training data may lead to low accuracy representation of the test data, creating a low model with a low level of robustness. The alpha tuning parameter is used to determine what type of regression is most effective. If alpha is set to 0, the model will use a type of regression known as ridge regression (42). When alpha is set to 1, the model will use a type of regression known as lasso (42). In ridge models the regression coefficients are squared, while in lasso models, the absolute value of coefficients is used. Elastic net regressions generally use an alpha value between 0 and 1 which defines the degree of combination between the two model types of ridge and lasso regression (42).

The second tuning variable is lambda, which is used as a “shrinkage” parameter that takes regression coefficient values that are past a specific threshold and resets them to a common value (43). If the lambda penalty is set to 0, then no coefficients are reset, and if set to infinity, then all coefficients would be converted to the same common value (43). A regression model is created in the training portion of the dataset, and then tested in the remaining data.

Elastic net regression analysis was conducted in Rstudio 2022.12.0, using “*caret*”, “*ggplot*” and “*glmnet*” packages (44, 45, 46, 47). Blood plasma metabolite data for 822 metabolites from 826 participants were randomly grouped into training/ testing groups for analysis (70% training / 30% testing). An elastic net model was generated with a 10-fold cross validation framework to regress *AMY1* CNV to the 822 metabolites of interest (7). The optimal alpha parameter between 0 and 1 was tested in 0.1 increments, while the optimal lambda tuning parameter was determined in 0.0001 increments (48). Tuning parameters were tested using a 10-fold cross validation system that would effectively determine the best possible model parameters for the robust prediction of metabolite plasma concentrations that were associated with *AMY1* CNV (48). The selection of alpha and lambda tuning parameters was done in the training set of the data and the most effective tuning parameters were then applied to the test set. Elastic net regression of the training set revealed that

the optimal training values were 1.0 for the alpha parameter and 0.2 for the lambda parameter. The measure of effectiveness of the model was the mean standard error (MSE) between the *AMY1* CNV and plasma metabolite concentrations. The model was then applied to the test set and the top metabolites with the lowest MSE consistently identified through the 10-fold cross validation of the model were selected. Once the model was applied to the test set, a metabolite profile score was generated (7). This score was the weighted sum of the strongest regression coefficients of the metabolites identified through ENR. These scores were combined to create metabolite profile scores that were regressed against *AMY1* CNV and BMI (see figure 3).

2.6: Analysis of data

Normalized data was run through an elastic net training model to reveal the top metabolites associated with *AMY1* CNV. The metabolites were then run through separate linear regression models against *AMY1* CNV. There were two linear models for *AMY1*, the unadjusted model with *AMY1* only, and an *AMY1* model adjusting for age and sex. Metabolites with regression models that passed the threshold for multiple testing ($p < 0.0026$) were included in the metabolite profile score. The metabolite profile score was then analyzed through linear regression models against *AMY1* and BMI.

2.7: BMI sensitivity analysis

Current literature regarding the influence of *AMY1* CNV on BMI is inconsistent in terms of their direct association(32, 49). BMI was used as a sensitivity analysis to test the strength of metabolite concentration association against various lifestyle determinants. There were three linear models run for BMI, an unadjusted model, a model adjusting for age and sex, and a model adjusting for age, sex, smoking status, alcohol consumption, education, total energy consumed, total starch intake, and physical activity level. Metabolites with regression models that passed the threshold for multiple testing ($p < 0.0026$) were included in the metabolite profile score.

2.8: Data Visualization and multicollinearity

The R package “Metaboanalyst” was used to visualize the associations between metabolite concentrations with *AMY1* CNV and BMI (50). Visualization of both the metabolite concentration and metabolite profile score data was done through ANOVA analysis. *AMY1* copy number was divided into quartiles with group 1 including 1-4 copies, group 2 including 5-6 copies, group 3

including 7-9 copies, and group 4 including 10 and greater copies (14). BMI was also separated into quartiles outlined by the American center for disease control (51). Group 1 included BMI is less than 18.5 (underweight range), group 2 included those with BMI between 18.5 to <25 (healthy weight range), group 3 included those with BMI from 25.0 to <30 (overweight range), group 4 included those with a BMI of 30.0 or higher (obesity range) (51). The package was also used to assess multicollinearity between the top metabolites selected through ENR. Pearson correlation heatmaps of top metabolites were also conducted to measure the level of multicollinearity within the data. ANOVA graphs were created to visualize the relationships between *AMY1* CNV, BMI, and metabolite concentrations. The metabolite profile score of significant metabolites was also analyzed through ANOVAS against *AMY1* CNV and BMI. *AMY1* CNV data from the population of 826 participants was also run against corresponding BMI measurements in a linear model to determine if there was a significant association between the two variables of interest.

2.9: Ethical Considerations

Informed written consent was provided by all participants at the time of data collection/ interviews. Participants could withdraw from the study at any point in time, and their data would not be used for the remainder of the study. Protocol for data collection was approved by the regional ethics committee at Lund University faculty of medicine (LU 51-90) and was carried out in accordance with the Helsinki declaration (3, 36). Data contained no personal information that could lead to identification of subjects and was stored on private servers.

3. Results

3.1: Population results

The population used in our study consisted of 826 individuals from the MDCS with a mean age of 58 years. Mean *AMY1* copy number was found to be 6.5, and BMI was found to be 25.9. Mean energy intake was found to be 2372 kcal per day, with mean starch intake to be 24.5% of the total kcal consumed (see table 1). Linear regression of population BMI data against corresponding *AMY1* copy number values was not found to be significant.

3.2: Elastic Net Regression output

ENR of metabolomic data from the MDCS revealed 19 metabolites that were associated with *AMY1* CNV one of which was omitted due to an association level of 0 (see figure 4). The remaining 18 compounds were implicated in a variety of biochemical pathways including Glycerophospholipid metabolism, pyrimidine metabolism, arachidonic acid metabolism, glycolysis/ gluconeogenesis, propanoate metabolism, lysine degradation, alpha-linolenic acid metabolism, retinol metabolism, linoleic acid metabolism, and sphingolipid metabolism (see table 2, figure 4).

3.3: Metabolite Collinearity

Two relationships between metabolites had moderate levels of collinearity, the first being between 1-palmitoyl-2-docosahexaenoyl and 1-oleoyl-2 docosahexaenoyl, with a correlation coefficient of 0.51 (figure 5). The second collinear relationship between metabolites was between glucose and 2-hydroxybutyrate, with a correlation coefficient of 0.17048 (figure 5).

3.4: *AMY1* analysis

Out of the 18 metabolites identified through ENR, 12 metabolites demonstrated nominal significance in their unadjusted association to *AMY1* CNV ($p < 0.05$). However, when the threshold p value of 0.0026 that adjusts for multiple testing errors is considered, only theophylline remains statistically significant (see table 3). In the *AMY1* CNV model which controlled for age, and sex, theophylline showed a strong positive association with *AMY1* CNV ($p < 0.001$) (See tables 3-4).

3.5: BMI sensitivity analysis

Out of the 18 metabolites identified through ENR, 9 metabolites demonstrated nominal significance in their unadjusted association to BMI ($p < 0.05$). Unadjusted linear regression models controlling for BMI found that 1-oleoyl-2-docosahexaenoyl-GPC, 2 hydroxybutyrate/2-hydroxyisobutyrate, hydantoin-5-propionate, 2-hydroxy-4-(methylthio) butanoic acid, glucose, and 2-oxoarginine passed the threshold p value of 0.0026 adjusting for multiple testing errors (< 0.001) (see table 4). BMI linear regression models adjusting for age and sex found significant associations for 1-oleoyl-2-docosahexaenoyl-GPC ,2-hydroxybutyrate/2-hydroxyisobutyrate, hydantoin-5-propionate, 2-hydroxy-4-(methylthio) butanoic acid, glucose, and 2-oxoarginine that

passed the threshold p value of 0.0026 which adjusted for multiple testing errors ($p < 0.001$) (see table 5). Additionally, a BMI model that adjusted for age, sex, smoking status, alcohol consumption, education, total energy consumed, total starch intake, and physical activity level was generated. These models found that 1-oleoyl-2-docosahexaenoyl-GPC, 2-hydroxybutyrate/2-hydroxyisobutyrate, 2-hydroxy-4 (methylthio)butanoic acid, glucose, and 2-oxoarginine passed the threshold p value of 0.0026 that adjusted for multiple testing errors ($p < 0.001$) (see tables 5-7).

Individual ANOVAs were run with metabolite concentrations against quartiles of *AMY1* copy number and BMI to visualize concentration changes. There were statistically significant differences in metabolite concentrations against BMI quartiles in Glucose, Hydroxybutyrate, oxoarginine, hydantoin, 2-hydroxy-4(methylthio)butanoic acid, pipecolate, and 1-oleoyl-2-docosahexaenoyl-GPC (See figure 6).

3.6: Profile score sensitivity analysis

Metabolites that demonstrated significant associations with either BMI or *AMY1* CNV in linear models were used in the metabolite profile score. The profile score was then run through linear regressions against quartiles of *AMY1* CNV and BMI and were found to have significant associations with both ($p < 0.001$). ANOVAs were also run with both individual metabolite concentrations and metabolite profiles scores (See figures 6i, 6j).

3.7: Summary

The top 18 metabolites identified from the MDCS all had nominal associations with *AMY1* CNV or BMI ($p < 0.05$). 1 metabolite, theophylline was significantly associated with *AMY1* CNV ($p < 0.001$) in a linear regression model adjusted for age and sex (see table 4). 7 metabolites; 1-oleoyl-2-docosahexaenoyl-GPC, 2-hydroxybutyrate, hydantoin-5-propionate, 2-hydroxy-4(methylthio)butanoic acid, glucose, 2-oxoarginine, and pipecolate were associated with BMI ($p < 0.001$) in linear models adjusting for various lifestyle factors (see tables 5-7).

4. Discussion

ENR of the metabolites from the MDCS revealed 18 metabolites that had nominal associations with *AMY1* CNV (see figure 4). These metabolites were from diverse families of compounds, including lipids, carbohydrates, vitamins, xenobiotics, and amino acids. These metabolites were

also involved in a variety of biochemical pathways, with the largest general category being fatty/amino acid metabolism (see table 2). Of particular interest in the selected metabolites was maltose and glucose, the former being the unit that makes up long chain starches, and the latter being the base unit of energy necessary for cellular respiration.

Linear regression models of the top 18 metabolites against *AMY1* CNV revealed theophylline to be the only metabolite to pass the multiple testing significance threshold when adjusted for age and sex ($p < 0.0026$).

Sensitivity analyses of the top 18 metabolites against BMI was done in 3 different models adjusting for various lifestyle determinant. This revealed 7 metabolites, 1-oleoyl-2-docosahexaenoyl-GPC, 2-hydroxyisobutyrate, hydantoin-5-propionate, 2-hydroxy-4-(methylthio) butanoic acid, glucose, and 2-oxoarginine that had significant associations with BMI that exceed multiple testing error.

The regression coefficients of metabolites with significant linear regression models were used to create the weighted metabolite profile score. Linear regressions of the metabolite profile against *AMY1* and BMI remained significant past the multiple testing threshold. When measured at a population level, *AMY1* copy number was found to be associated with BMI, this is in line with previous conclusions drawn from research within the MDCS (3, 32).

4.1: glucose

Glucose is a monosaccharide made of six carbons and an aldehyde group; it is the base unit needed for cellular metabolism in all living organisms (52). Besides salivary amylase, a variety of enzymes are responsible for the breakdown of larger polysaccharide units throughout the body and can be readily absorbed throughout the digestive tract (52). The proportion of blood glucose and the ability to regulate those levels are important factors for disease risk in humans. The inability to reduce blood glucose levels could be attributed to a lack of insulin in the body, or the inability of beta cells in the pancreas to produce insulin. Prolonged levels of high blood glucose can lead to a variety of issues like chronic pancreatitis and insulin resistance, both of which are precursors to the onset of diabetes (52). Glucose plasma concentration was found to be significantly associated with BMI in the MDC. Since salivary amylase breaks down complex starches into smaller carbohydrates, with the goal of yielding glucose molecules, the strong association between glucose concentrations and *AMY1* CNV seems to be a confirmation of the ENR methods used for metabolite

analysis. Glucose intolerance is known to be positively associated with BMI as well as waist circumference, and this trend may be what is being reflected within the MDC sub cohort (53). The ubiquity of glucose within human metabolic systems means that there are many factors involved that have not been accounted for from a statistical standpoint, and no concrete conclusions can be drawn regarding its apparent association with BMI within the MDC.

4.2: theophylline

Theophylline, or 1,3-dimethyl-7H-purine-2,6-dione, is created from a xanthine base with two methyl groups located at positions 1 and 3 (54). Theophylline is of a molecular class known as methylxanthines, which include caffeine (1,3,7-trimethylxanthine) and theobromine (3,7-dimethylxanthine) (55). This class of alkaloids differ in the number of methyl groups that are attached to the main xanthine ring (56). Theophylline was found to have a significant association with *AMY1* CNV but was not found to have any significant associations with BMI. This compound is naturally occurring in a variety of foods and beverages, including black tea, maté, coffee, and chocolate (54, 57). Theophylline is absorbed in the gastrointestinal tract and is further metabolized in the liver by multiple cytochrome P450 enzymes (54). Historically, this compound has also been used to treat respiratory diseases due to its ability to relax smooth muscle and pulmonary blood vessels through two main pathways, phosphodiesterase inhibition and adenosine receptor antagonism (54). Of particular interest in this study is theophylline's anti-inflammatory response, which can be attributed to its capacity for competitive nonselective phosphodiesterase inhibition (58). The nonselective nature of its inhibitory properties allows for theophylline to be associated with many inflammatory pathways, as well as associations with inflammation-related genes (54). In this context, the compound has recently been studied in the use of treatment for COVID-19 symptoms (54). Although theophylline was associated with *AMY1* CNV at a statistically significant level, there is little evidence for known interactions between the two elements in human metabolic processes.

4.3: 1-oleoyl-2-docosahexaenoyl-GPC

1-oleoyl-2-docosahexaenoyl-GPC is from a class of molecules known as phosphatidylcholines (52, 59). The main components of this type of molecule are two fatty acid chains bonded to a glycerol backbone through an ester linkage, and a phosphodiester linkage connecting another hydroxyl group to choline (52, 59). 1-oleoyl-2-docosahexaenoyl-GPC was found to have

significant associations with BMI, but not with *AMY1* CNV. Phosphatidylcholines are a diverse group of molecules due to the number of possible combinations in which fatty acid chains of diverse length can bond multiple positions along the glycerol backbone (52, 59). These molecules are a component of cell membranes in plants and animals and have been hypothesized as an effective biomarker for early cardiometabolic disease risk in individuals with elevated BMI (59). These types of metabolites were also known to be positively associated with visceral and cardiac fat deposits (59). Baek *et al.* investigated the metabolomes of visceral fat areas of 112 overweight Korean adults and found 12 significant phosphatidylcholine-based metabolites in high-fat areas of the body compared to low-fat areas (60). In an American context studying 40 obese women, Scherer *et al.* also came to the same conclusion regarding the localization of phosphatidylcholine-based metabolites in adipose tissue (61). A recent literature review of 21 human studies by Bellot *et al.* investigated phosphatidylcholine (PC) and lysophosphatidylcholine (LPC) as adiposity biomarkers. The review also identified a significant negative association between this metabolite and BMI in two studies included in the review (22, 59). Both Bagheri *et al.* and Berber *et al.* found negative associations or lower concentrations of LPC and PCs in higher BMI groups (22, 62). This negative association was also found in our study and passed the threshold for multiple error testing ($P < 0.001$). This finding was in keeping with most findings in the review, demonstrating a lower concentration of PC and LPC's as BMI increased; however, this trend however was inconsistent between differing research contexts (59). The biochemical mechanism that facilitates the relationship between increased BMI and low PC/ LPC concentration is not yet known and provides a novel area for future research (59).

4.4: 2-hydroxybutyrate/2-hydroxyisobutyrate

2-hydroxybutyrate is a short chain fatty acid (SCFA) created as a by-product of alpha-ketobutyrate by the catabolism of threonine and methionine via lactate dehydrogenase or alpha-hydroxybutyrate dehydrogenase (52, 63). 2-hydroxybutyrate was found to have significant associations with BMI in the MDC but had no significant associations with *AMY1* CNV. The compound is found in hepatic tissue in many organisms, and many studies have recently demonstrated its robust role as a biomarker for isolated impaired fasting glucose, an important physiological state that precedes the development of type 2 diabetes (52, 64). This relationship has been documented in multiple metabolomic studies, utilizing multiple methods of detection, in both short- and long-term contexts

(65). A high power metabolomic study was conducted by Cobb *et al.* in which targeted analysis of 23 metabolites was completed for 4053 individuals in 3 European cohorts. The study found that 2-hydroxybutyrate was a robust, selective biomarker for impaired glucose tolerance (64). The researchers stated that this method of glucose intolerance detection could eventually be used in lieu of the commonly used oral glucose tolerance test (64, 65). To determine if any metabolites could be biomarkers for type 2 diabetes, Peddinti *et al.* conducted a non-targeted metabolomic study in a subset of the Botnia protective study, comprising of 543 non-diabetic individuals, 146 of which contracted type 2 diabetes after a 10 year follow-up period (66). The researchers were able to conclude that 2-hydroxybutyrate was significantly associated with a future progression to type 2 diabetes and that levels of 2-hydroxybutyrate elevation were detectable years before the onset of the disease (66). A targeted metabolomic study done by Tricò *et al.* in adolescents was also able to conclude that elevated 2-hydroxybutyrate levels were a novel method for detection of reduced insulin sensitivity and beta-cell dysfunction, ultimately leading to an increased risk for type 2 diabetes (67). The results found in this study were consistent with previous literature, in which a significant positive association between 2-hydroxybutyrate levels and BMI was found.

4.5: hydantoin-5-propionate

Hydantoin-5-propionate is a heterocyclic SCFA that is a derivative of imidazolidine with ketone groups at position 2 and 4 (52). The compound belongs to a group known as hydantoins, which are molecules with the same parent ring structure found in imidazoline. Hydantoin-5-propionate is classified as a primary, or essential, metabolite meaning that is necessary for certain metabolic functions (52). In a Qatari study of 996 adults, Yousri *et al.* reported hydantoin-5-propionate as a novel biomarker for BMI in their non-targeted metabolomic study (68). The study's aim was to identify biomarkers associated with type 2 diabetes, adiposity, retinopathy, dyslipidemia, and elevated lipoprotein levels, and concluded that hydantoin-5-propionate was only significantly associated with BMI (68). Other metabolomics literature that refers to hydantoin-5-propionate found no association between hydantoin-5-propionate with BMI or CVD risk. There is, however, mention of its role as a novel biomarker for high protein diets (69, 70, 71). Pang *et al.* investigated metabolomic profiles associated with both non-alcoholic fatty liver disease, and BMI, and did not identify hydantoin as a significant biomarker (69). Rebholz *et al.* conducted a randomized clinical trial with randomly assigned exposure groups to differing levels of protein (71). Using a non-

targeted metabolomics approach, the researchers identified hydantoin-5-propionate as a biomarker for increased protein intake (71). Kim *et al.* confirmed this finding and was able to conclude that hydantoin-5-propionate had a strong association with high protein intake as opposed to high carbohydrate-based diets (70). Our study also found a significant positive association between hydantoin-5-propionate concentrations and BMI.

4.6: 2-hydroxy-4-(methylthio)butanoic acid

2-hydroxy-4-(methylthio) butanoic acid is from a class of molecules known as thia fatty acids, which are distinguished by a sulfur atom bonded to various positions on the molecule (52). Of particular interest is the metabolic efficiency of these compounds moderated by the positioning of the sulfur atom on the molecule (72). Gudbrandsen *et al.* investigated the metabolic effect of thia fatty acids in rat liver models and found that mitochondrial beta oxidation rates varied extensively with thia fatty acids (72). Beta oxidation is the set of chemical processes used to break down fatty acid molecules to forms usable in cellular respiration. The researchers found that the ability of thia fatty acids to be broken down was moderated by how far the sulfur group was from the carboxyl group (72). Gudbrandsen found that 2-hydroxy-4-(methylthio) butanoic acid underwent low levels of beta oxidation compared to thia fatty acids with sulfur localized to the 5-position or 7-position (72). This is an important trait from a biomarker perspective, as less readily metabolized molecules like 2-hydroxy-4-(methylthio) butanoic acid would accumulate within tissues and could be useful for analysis. Lind *et al.* compared the metabolomic profiles of overweight and normal weight individuals in 3 population-based samples in Sweden totalling over 11,000 individuals (23). The study observed major differences in metabolite profiles between obese and normal weight individuals, stating that there appears to be a “deterioration” of the metabolomic profile in obese individuals, and an increased association with metabolomic profiles that were related to insulin sensitivity (23). The paper also states that 2-hydroxy-4-(methylthio) butanoic acid had a significant positive association with BMI, which our study also confirmed (23).

4.7: 2-oxoarginine

2-oxoarginine is a guanidino based compound that originates from the catabolism of the amino acid arginine (52). Al-Sulaiti *et al.* investigated the metabolomic signatures of 107 individuals (75 obese and 32 normal weight individuals) (73). The researchers were able to identify metabolites that could be used to identify the differences between two stages of diabetes onset: insulin

sensitivity, the stage in which there is normal physiological responses to insulin; and insulin resistance, in which the development of diabetes is highly likely (73). The researchers were able to show a statistically significant decrease in 2-oxoarginine concentrations in insulin resistant individuals and individuals with type 2 diabetes when compared to insulin sensitive individuals. The researchers hypothesized that measuring temporal shifts in metabolite concentrations associated with diabetes progression could be helpful to quantify the underlying mechanisms that lead to disease onset (73). Perng *et al.* investigated the metabolomic profiles of 524 normal, overweight, and obese adolescents who had been organized into 4 groups based on BMI and type 2 diabetes (T2D) risk (74). Perng and colleagues found that 2-oxoarginine was associated with the high BMI/ high T2D risk groups (74). The positive association between 2-oxoarginine and BMI was also found within our study.

4.8: pipecolate

Pipecolate is a metabolite associated with the catabolism of the amino acid lysine. The D-isomer is only produced through bacterial metabolism in the digestive tract, though other forms have been associated with consumption of black soybeans (24, 52). A metabolomic study involving 136 healthy women investigated 206 metabolites and their associations to muscle strength, muscle mass, and BMI (75). The researchers were able to demonstrate an association between pipecolate and BMI, hypothesizing that amino acid metabolism and downstream biomarkers could play a role in measuring muscle regulation (75). Through our ANOVA analysis, pipecolate was found to have significantly different concentrations in BMI quartiles. This relationship was not found in linear regressions of the same data. For this reason, it was not included in the metabolite profile score for further analysis.

4.9: Metabolites, AMY1 CNV, and microbiota

The proportion of *AMY1* related metabolites with amino and fatty acid derivatives may connect salivary amylase to downstream digestion processes involving gut microbiota. Hasegawa *et al.* investigated the impact of salivary and pancreatic amylase gene copy numbers on diabetes, obesity, and functional profiles of gut microbiomes in a Japanese context (76). They were able to conclude that *AMY1* copy number was associated with BMI, and that certain genus' of gut bacteria were significantly correlated with *AMY1* copy number (76). Some of these genus, like *dialister* have been associated with gut carbohydrate metabolism in other contexts (77). This provides a potential

mechanism for *AMY1*'s role in moderating metabolite concentrations and could provide an area of future research in relation to the metabolites identified in this study.

In a Mexican context, León-Mimila et al. investigated the role of low *AMY1* copy number on gut microbiota/ obesity, finding associations with *prevotella* bacteria and increased adiposity (49). The researchers found that individuals with low *AMY1* copy number had higher BMI and altered gut microbiota, namely in the genus *prevotella* (49). This species of bacteria was also known to be associated with the breakdown of complex carbohydrates. The mechanisms that this bacteria use to digest these molecules could then influence plasma metabolite concentrations through various biochemical pathways.

May *et al.* outlined the role of gut microbial derived SCFAs and their effect on adiposity (78). They found that fatty acids with short carbon chains like propionate, and butyrate play important roles in biochemical pathways involving gut microbiota (78). They also hypothesized that increased production of such metabolites reduced adiposity risk (78). The authors concluded that SCFAs may play an important role in adipose tissue regulation, and that these metabolites could play an important role in treatment and prevention noncommunicable disease risk. This is in line with our findings of multiple *AMY1* CNV related SCFAs that were identified through ENR of metabolite data from the MDCS. The links between the identified SCFAs and gut microbiota may be a part of the mechanism(s) that links *AMY1* CNV to certain metabolite plasma concentrations.

4.10: Validity of methods

The use of elastic net regression in this study was favoured due to the collinearity of metabolomic data. Saltelli stated that “Sensitivity Analysis is the study of how the uncertainty in the output of a model can be attributed to different sources of uncertainty in the model input” (79). Sensitivity analyses through linear regression models were used to assess the ENR output. In the scope of our study, this was done by adjusting for age, sex, and lifestyle/ diet factors to determine if the association between plasma concentration and *AMY1*/ BMI was still significant. This does not however, carry enough statistical significance to allow for any large-scale conclusions applicable to diverse populations in a public health context. For these claims to be made, external validation of the metabolite profile scores generated in this study must be done to evaluate the models’ robustness and replicability. Instead, these models point us towards biomarkers that could have potential use for early diagnosis and preventative care if validated in other research contexts.

Adjustment of linear models with various lifestyle factors also help to deduce the strength of the association between the biomarker and exposure of interest when accounting for the presence of confounding variables.

The association of *AMY1* with BMI risk is a novel area of research within the field of nutritional epidemiology. Studies in varying populations involving carbohydrate intake and BMI/ CVD risk have shown differing results, depending on the demographic of interest. This can lead to incorrect claims regarding the moderation of diet patterns that may not be applicable to groups that differ significantly from the research population.

The evolution of *AMY1* CNV and its role in starch digestion is also a point of contention in literature. In 2007, Perry *et al.* found a positive association between *AMY1* CNV and salivary amylase protein levels, and this paper has been cited in most research involving *AMY1* CNV in the last 16 years (4). Perry *et al.* argued that with the onset of agriculture, an increase in plant based foods consumption in human populations acted as a selective force for increased *AMY1* CNV (4). Hardy *et al.* also hypothesized that the incorporation of cooking food acted as a selective pressure towards increased *AMY1* copy number. The authors reported that the increased energy availability in cooked starches would allow for increased brain tissue development in early *homo sapien* lineages (6). One of the main reasons Hardy *et al.* came to this conclusion is due to the increased efficiency in *salivary amylase* digestion of cooked starches compared to raw crystalline starches (6).

Fernandez *et al.* argued that the role of *AMY1* may not be as essential as previously thought in the overall process of starch digestion (9). The authors argue that there were other enzymes throughout the digestive tract, like *sucrase-isomaltase* and *maltase-glucoamylase* that have the ability to digest entire starch granules into base glucose units for cellular respiration (9). The paper ultimately comes to several conclusions, one of the most important being that since there is limited standardization in how *AMY1* CNV it is hard to assess the validity of *AMY1* CNV correlations other studies have arrived at. The difficult to effectively quantify CNV in genes is something discussed in many publications, with some suggesting that only rough copy number estimates were used in previous CNV association studies (80). The other main conclusion the study arrives at is that starch digestion may not have been the main driving force in driving *AMY1* CNV selection. The paper came to this conclusion after outlining that salivary amylase has limited interaction with

the starch molecules, producing intermediates not directly usable for metabolic processes (9). Fernandez *et al.* also questioned the role of *AMY1* CNV in glucose homeostasis and insulin resistance. The study points to a common thread in previous literature that demonstrates an inverse trend between insulin resistance risk and *AMY1* CNV (3, 9, 14).

4.11: Limitations

Profile scores including all significant metabolites were found to have associations with BMI and *AMY1* CNV quartiles in the study population. This significance was seen in both linear regression and ANOVA analysis. Linear models were run with various factors to determine the extent of their confounding influence over the data. Although some of these variables were accounted for in these models, many more physiological and environmental factors that come into play with BMI moderation. The robustness of the selected metabolites as biomarkers for BMI or CVD disease risk has not yet been evaluated. To remediate this, other forms of statistical validation, ideally in a differing research cohorts need to be completed.

Without external validation of the metabolite profile scores and individual metabolites, it remains to be seen whether the tuning parameters had effectively minimized model overfitting. The issue of multicollinearity of variables also remains significant when considering the complexity of the human metabolome. The ethnic homogeneity of the research population must also be considered when determining how applicable these findings could be to differing populations. Walford *et al.* investigated metabolomic profiles associated with diabetes incidence in over 3000 participants with elevated BMI in the US (81). The researchers were able to identify specific metabolites associated with diabetes onset and were able to remain confident in the robustness of the findings due to the high level of diversity in their research population (81). A notable issue inherent in the study design is the high ethnic homogeneity of the MDCS cohort. Claims regarding a diet pattern's ability to alleviate disease risk must be validated in multiple contexts to determine the robustness of the hypotheses regarding the underlying biochemical mechanisms involved. It has been demonstrated that different ethnicities have altered levels of intermediate metabolites in pathways related to carbohydrate metabolism (82). This is of particular interest in population based metabolomic studies with high levels of ethnic homogeneity, and in determining the validity of clinical claims made in only one context. Further population based metabolomic studies must be

conducted in diverse populations or externally validated in a cohort with a differing ethnic background to determine the robustness of claims regarding novel biomarkers (81).

Since the metabolite profile, and the metabolites of interest have not been validated outside of the original MDCS sub-cohort, no claims can be made regarding the reproducibility of these results. The relationship between *AMY1* CNV and BMI is still a disputed matter, as the effect of *AMY1* CNV is vastly different when comparing different populations and contexts. The interaction between *AMY1* CNV and BMI in moderating an individual's plasma metabolite concentrations could be used to further understand the biochemical pathways leading to the onset of CVDs. Metabolomic studies provide one tool that, in concert with other lifestyle measurements, could be useful in the early detection and prevention of CVD. The association of several metabolites to *AMY1* and BMI highlight the complex structure of downstream metabolic processes in the human body, and the necessity for multi-level testing. Several metabolites were found to be associated with either *AMY1* CNV or BMI, highlighting the multifactorial role of metabolites in CVD risk. Although the associations with *AMY1* were nominal, the results demonstrate the importance of the relationship between long-term diet patterns and genetic factors on disease risk. Although BMI is not the most accurate indicator of health, metabolites associated with elevated BMI could play an important role in preventative care to reduce noncommunicable disease risk. The relationship between some plasma metabolite concentrations does appear to be moderated on some level by *AMY1* CNV and lifestyle determinants, though the underlying mechanisms that connect them is still unknown.

Another limitation of this study is the use of BMI as an indirect measure of adiposity. BMI is a ratio between weight and height and cannot distinguish between the proportion of muscle mass and adipose mass within an individual. This led to an assumption in our study that at the population level, increased BMI meant a higher proportion of adipose tissue. By using an indirect measure of adiposity, this study failed to directly quantify any association between potential biomarkers and the exposure of interest (high levels of adipose tissue). This also meant that there may have been a proportion of individuals within the data who were misclassified as overweight or obese by their BMI. To determine the validity of biomarkers associated with adipose tissue, future studies will have to find direct measures to quantify increased adiposity in populations. Sun *et al.* conducted a systematic review of the tri-ponderal mass index as a more effective measure of adiposity in

children and adolescents, and found it to be a better measure than BMI in some contexts when used with other measures like waist circumference (83). The trio of waist circumference, waist to height ratio, and BMI measurement in future data collection efforts in the MDCS may lead to more direct studies of association between biomarkers and adiposity. Future studies that involve the sampling of visceral fat tissue may also be a more effective method for determining the robustness of biomarkers associated with increased adiposity.

5. Public Health context of research

5.1 Burden of non communicable disease

Increased adiposity has been linked to a heightened risk for cardiovascular diseases (CVD) in many populations. CVDs account for one third of all annual deaths worldwide, with diabetes prevalence in 20-to-79-year old's being 10.2% in 2021 and expected to rise to 12.2% by 2045 (84). When an individual has the combined risk factors of diabetes, high blood pressure, and obesity, they can be clinically diagnosed with “metabolic syndrome”, a physiologic state that creates a high risk for CVDs (85). A systematic review of 23 longitudinal and cohort studies involving childhood adiposity and CVD risk was conducted by Umer *et al.* (86). They concluded that increased adiposity in childhood was significantly associated with increased systolic blood pressure and reduced high density lipoprotein (HDL) levels in adults (86). Umer also hypothesized that elevated adult BMI may be a mediating factor in CVD risk (86). Skinner *et al.* concluded that severe adiposity in children was linked to an increased exposure to cardiometabolic risk factors like high diastolic/systolic blood pressure, low HDL cholesterol, and high triglyceride/glycated hemoglobin levels (87). Yang *et al.* studied a population of 68,603 Chinese adults and concluded that long periods of being overweight, or obese were associated with an increased lifetime risk for CVD (88). CVD risk is heavily dictated by these variables and plasma biomarkers provide a measure for understanding the physiological state necessary for the onset of these diseases (89).

5.2 Preventative care using metabolomics

Target 3.4 of the sustainable development goals outlined by the world health organization states that by 2030, deaths related to noncommunicable diseases should be reduced by one third through treatment and prevention (90). Reducing CVD risk is a multifactorial issue from a public health perspective and is paramount to achieving target 3.4. Lifestyle factors like exercise, smoking/

alcohol, and medication consumption all contribute to an individual's risk for developing CVDs (91, 92). Compounding these variables are socioeconomic barriers an individual may face when trying to consistently access healthy foods (92). Such barriers may include (but are not limited to) a lack of knowledge about the importance of a healthy diet, a lack of availability of healthy food options, high prices for nutritious foods, social/ cultural norms surrounding certain foods, marketing influences, and individual food preferences (91, 92). An individual's genetic makeup also plays a role in noncommunicable disease risk. This risk is individual, and generalizations about disease risk based on specific genes are difficult to quantify at a population level. Genetic variability in populations creates the necessity of external validation to determine the robustness of general claims regarding noncommunicable disease risk. Understanding the contribution of *AMY1* copy number to BMI is an important part of CVD prevention. For most human beings, starches, including grains/rice are a major source of energy, making it important to understand the impact and mechanism of *AMY1* CNV on metabolite concentration and obesity. At a population level, understanding the genetic variability of this gene, and how it interacts with differing diet and lifestyle determinants is an important part of disease risk management and preventative care. Finding biomarkers that are associated with or moderated by *AMY1* CNV would be beneficial from a preventative medicine perspective. Measuring metabolites known to be associated with elevated BMI/CVD risk before disease onset, could allow for detection and lifestyle changes that could lower disease risk or potentially prevent disease onset entirely. For example, people with low *AMY1* copy numbers could moderate long-term diet patterns to adjust starch intakes and lower their risks for certain non-communicable diseases. Undetected cases of diabetes are thought to account for 30-60% of all cases worldwide, and the main method of diagnosis (OGTTs) are not standard practise in routine medical checks (53). Finding robust biomarkers for early detection and prevention of CVDs is of the utmost importance, and metabolites provide a robust method to track changes in a person's metabolic state.

5.3 Future metabolomic studies

Quantification and statistical analyses of the 268 unknown metabolites discovered in our research population may also provide insights into the links between *AMY1* and BMI, as there may have been stronger biomarkers present within that group of metabolites.

The 8 metabolites that had significant associations with *AMY1* CNV or BMI have the potential to become novel biomarkers for CVD risk. Previous studies have demonstrated that changes in metabolite profiles over time are an effective way to identify early onset stages of noncommunicable diseases. These changes in metabolic profiles could be used for prevention or early intervention to reduce disease risk and impact. The continuous measurement of biomarkers in at risk individuals may then become an effective protocol for CVD detection, and prevention in populations, along with diet and lifestyle modifications (81). Population based research in which long-term diet patterns are recorded, like the MDC can provide high statistical power to determine the existence and valid novel biomarkers. These studies should be validated in external cohorts to determine the extent to which genetic differences influence metabolite blood plasma concentration. Future research using the data found in this study could involve the creation of long-term cohort studies measuring these metabolites in healthy populations, to determine their robustness as biomarkers for elevated adiposity and CVD risk. *AMY1* may play a role in BMI risk moderation, but the extent to which this effect can be effectively quantified remains to be seen. From a public health perspective, the findings in this study demonstrate the usefulness of metabolomic studies at the population level. Monitoring these biomarkers at a population level could lead to earlier disease detection, and a lower overall prevalence of CVD with time (93, 94). The scope of future research could also expand to include cardiovascular disease outcomes to determine whether an association between metabolite concentrations and certain cardiovascular diseases. The use of ENR methods to find robust biomarkers can be seen as a starting point to understand the intermediate physiological effects of diet patterns. Metabolomics could also be used to better understand diet related disease risks at both individual and population levels.

6. Conclusion

Elastic net regression of metabolomic data from the Malmö diet and cancer cohort study revealed 18 metabolite associated with *AMY1* CNV. Of the 18 metabolites analyzed through linear models, only theophylline had a statistically significant association with *AMY1* CNV. Sensitivity analyses of the 18 metabolites against various lifestyle determinants revealed 7 more metabolites with significant associations to BMI. These metabolites were related to a variety of biochemical pathways related to carbohydrate metabolism. The relationships between the identified metabolites and BMI need to be further quantified in other research contexts to understand their interaction

with *AMY1* (see figure 7). If validated in other research contexts, the identified metabolites could play important roles as biomarkers for elevated BMI and noncommunicable disease risk.

7. References

1. Peyrot des Gachons C, Breslin PA. Salivary Amylase: Digestion and Metabolic Syndrome. *Curr Diab Rep*. 2016;16(10):102.
2. Carpenter D, Mitchell LM, Armour JAL. Copy number variation of human AMY1 is a minor contributor to variation in salivary amylase expression and activity. *Human Genomics*. 2017;11(1):2.
3. Farrell M, Ramne S, Gouinguenet P, Brunkwall L, Ericson U, Raben A, et al. Effect of AMY1 copy number variation and various doses of starch intake on glucose homeostasis: data from a cross-sectional observational study and a crossover meal study. *Genes & Nutrition*. 2021;16(1).
4. Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, et al. Diet and the evolution of human amylase gene copy number variation. *Nature Genetics*. 2007;39(10):1256-60.
5. OECD. Obesity and the Economics of Prevention 2010.
6. Hardy K, Brand-Miller J, Brown KD, Thomas MG, Copeland L. The Importance of Dietary Carbohydrate in Human Evolution. *The Quarterly Review of Biology*. 2015;90(3):251-68.
7. Wang F, Baden MY, Guasch-Ferré M, Wittenbecher C, Li J, Li Y, et al. Plasma metabolite profiles related to plant-based diets and the risk of type 2 diabetes. *Diabetologia*. 2022;65(7):1119-32.
8. Nakamura Y. Starch. *Metabolism and Structure*. 1st 2015. ed: Springer Japan; 2015.
9. Fernandez CI, Wiley AS. Rethinking the starch digestion hypothesis for AMY1 copy number variation in humans. *AMERICAN JOURNAL OF PHYSICAL ANTHROPOLOGY*. 2017;163(4):645-57.
10. Fried M, Abramson S, Meyer JH. Passage of salivary amylase through the stomach in humans. *Digestive Diseases and Sciences*. 1987;32(10):1097-103-103.
11. Ionita-Laza I, Rogers AJ, Lange C, Raby BA, Lee C. Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis. *Genomics*. 2009;93(1):22-6.
12. McCarroll SA, Altshuler DM. Copy-number variation and association studies of human disease. *Nat Genet*. 2007;39(7 Suppl):S37-42.
13. Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, et al. Copy number variation: New insights in genome diversity. *GENOME RESEARCH*. 2006;16(8):949-61.
14. Hamid AK, Andersson-Assarsson J, Ericson U, Sonestedt E. Interaction Effect Between Copy Number Variation in Salivary Amylase Locus (AMY1) and Starch Intake on Glucose Homeostasis in the Malmö Diet and Cancer Cohort. *Frontiers in nutrition EpiHealth: Epidemiology for Health EXODIAB: Excellence of Diabetes Research in Sweden*. 2021;7.
15. Ju L, Pan Z, Zhang H, Li Q, Liang J, Deng G, et al. New insights into the origin and evolution of α -amylase genes in green plants. *Scientific Reports*. 2019;9(1):4929.
16. Carpenter D, Mitchell LM, Armour JAL. Copy number variation of human AMY1 is a minor contributor to variation in salivary amylase expression and activity. *HUMAN GENOMICS*. 2017;11:2.
17. Atkinson FS, Hancock D, Petocz P, Brand-Miller JC. The physiologic and phenotypic significance of variation in human amylase gene copy number. *The American Journal of Clinical Nutrition*. 2018;108(4):737-48.
18. Perry GH, Kistler L, Kelaita MA, Sams AJ. Insights into hominin phenotypic and dietary evolution from ancient DNA sequence data. *Journal of Human Evolution*. 2015;79:55-63.

19. Grammatikopoulou MG, Gkiouras K, Markaki AG, Gkouskou KK, Aivaliotis M, Stylianou K, et al. AMY1 diploid copy number among end-stage renal disease patients. *Hormones: International Journal of Endocrinology and Metabolism*. 2020;19(3):369-76.
20. Ute R. *Metabolomics*. [Elektronisk resurs]: InTech; 2012.
21. Pinu FR, Beale DJ, Paten AM, Kouremenos K, Swarup S, Schirra HJ, et al. Systems biology and multi-omics integration: Viewpoints from the metabolomics research community. *Metabolites*. 2019;9(4).
22. Bagheri M, Farzadfar F, Qi L, Yekaninejad MS, Chamari M, Zeleznik OA, et al. Obesity-Related Metabolomic Profiles and Discrimination of Metabolically Unhealthy Obesity. *J Proteome Res*. 2018;17(4):1452-62.
23. Lind L, Salihovic SASL, Sundström J, Elmståhl S, Hammar U, Dekkers K, et al. Metabolic Profiling of Obesity With And Without The Metabolic Syndrome: A Multi-Sample Evaluation. *Journal of Clinical Endocrinology and Metabolism*. 2022;107(5):1337-45.
24. Kim MJ, Yang HJ, Kim JH, Ahn CW, Lee JH, Kim KS, et al. Obesity-related metabolomic analysis of human subjects in black soybean peptide intervention study by ultraperformance liquid chromatography and quadrupole-time-of-flight mass spectrometry. *J Obes*. 2013;2013:874981.
25. Ottosson F, Smith E, Melander O, Fernandez C. Altered asparagine and glutamate homeostasis precede coronary artery disease and type 2 diabetes. *Journal of Clinical Endocrinology and Metabolism* EXODIAB: Excellence of Diabetes Research in Sweden EpiHealth: Epidemiology for Health. 2018;103(8):3060-9.
26. Mayneris-Perxachs J, Mousa A, Naderpoor N, Fernández-Real J-M, de Courten B. Low AMY1 Copy Number Is Cross-Sectionally Associated to an Inflammation-Related Lipidomics Signature in Overweight and Obese Individuals. *Molecular Nutrition & Food Research*. 2020;64(11):1901151.
27. Mandel AL, Breslin PA, Mandel AL, Breslin PAS. High endogenous salivary amylase activity is associated with improved glycemic homeostasis following starch ingestion in adults. *Journal of Nutrition*. 2012;142(5):853-8.
28. Marquina C, Mousa A, Naderpoor N, De Courten B, Belski R, Banaharis H. Increased inflammation and cardiometabolic risk in individuals with low AMY1 copy numbers. *Journal of Clinical Medicine*. 2019;8(3).
29. WHO. Obesity and overweight Fact Sheet. Geneva World Health Organization; 2021.
30. Bergin J, Neale M, Eaves L, Martin N, Heath A, Maes H. Genetic and Environmental Transmission of Body Mass Index Fluctuation. *Behavior Genetics*. 2012;42(6):867-74.
31. Maes HH, Neale MC, Eaves LJ. Genetic and environmental factors in relative body weight and human adiposity. *Behavior genetics*. 1997;27(4):325-51.
32. Rukh G, Ericson U, Andersson-Assarsson J, Orho-Melander M, Sonestedt E. Dietary starch intake modifies the relation between copy number variation in the salivary amylase gene and BMI. *The American Journal of Clinical Nutrition*. 2017;106(1):256-62.
33. Falchi M, El-Sayed Moustafa JS, Takousis P, Pesce F, Bonnefond A, Andersson-Assarsson JC, et al. Low copy number of the salivary amylase gene predisposes to obesity. *Nature Genetics*. 2014;46(5):492-7.
34. Viljakainen H, Andersson-Assarsson JC, Armenio M, Pekkinen M, Pettersson M, Valta H, et al. Low Copy Number of the AMY1 Locus Is Associated with Early-Onset Female Obesity in Finland. *PLoS ONE*. 2015;10(7):1-13.

35. Marcovecchio ML, Florio R, Verginelli F, De Lellis L, Capelli C, Verzilli D, et al. Low AMY1 Gene Copy Number Is Associated with Increased Body Mass Index in Prepubertal Boys. *PLoS ONE*. 2016;11(5):1-9.
36. Huiping L, Yan B, Yaogang W, Emily S. Starch intake, amylase gene copy number variation, plasma proteins, and risk of cardiovascular disease and mortality. *BMC Medicine*. 2023;21(1):1-11.
37. Mattisson I, Wirfalt E, Aronsson CA, Wallstrom P, Sonestedt E, Gullberg B, et al. Misreporting of energy: prevalence, characteristics of misreporters and influence on observed risk estimates in the Malmo Diet and Cancer cohort. *BRITISH JOURNAL OF NUTRITION*. 2005;94(5):832-42.
38. Brunkwall L, Jönsson D, Ericson U, Hellstrand S, Kennbäck C, Östling G, et al. The Malmö Offspring Study (MOS): design, methods and first results. *Eur J Epidemiol*. 2021;36(1):103-16.
39. Bao X, Borné Y, Muhammad IF, Nilsson J, Lind L, Melander O, et al. Growth differentiation factor 15 is positively associated with incidence of diabetes mellitus: the Malmö Diet and Cancer–Cardiovascular Cohort. *Diabetologia*. 2019;62(1):78-86.
40. Koen FD, Sergi S-B, Gabriel B, Christoph N, Ulf H, Diem N, et al. An online atlas of human plasma metabolite signatures of gut microbiome composition. *Nature Communications*. 2022;13(1):1-12.
41. Ottosson F, Ericson U, Almgren P, Smith E, Brunkwall L, Hellstrand S, et al. Dimethylguanidino Valerate : A Lifestyle-Related Metabolite Associated With Future Coronary Artery Disease and Cardiovascular Mortality. *Journal of the American Heart Association* EXODIAB: Excellence of Diabetes Research in Sweden EpiHealth: Epidemiology for Health. 2019;8(19).
42. Fan Y, Tang CY. Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*. 2013;75(3):531-52.
43. Chand S. On tuning parameter selection of lasso-type methods - a monte carlo study. *IEEE*; 2012. p. 120-9.
44. Team RC. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2022.
45. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag 2016.
46. Kuhn M. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*. 2008;28(5):1 - 26.
47. Friedman JH, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*. 2010;33(1):1 - 22.
48. Drouin-Chartier J-P, Hernández-Alonso P, Guasch-Ferré M, Ruiz-Canela M, Li J, Wittenbecher C, et al. Dairy consumption, plasma metabolites, and risk of type 2 diabetes. *The American Journal of Clinical Nutrition*. 2021;114(1):163-74.
49. León-Mimila P, Villamil-Ramírez H, López-Contreras BE, Morán-Ramos S, Macías-Kauffer LR, Acuña-Alonzo V, et al. Low Salivary Amylase Gene (AMY1) Copy Number Is Associated with Obesity and Gut Prevotella Abundance in Mexican Children and Adults. *Nutrients*. 2018;10(11):1607.
50. Xia J, Psychogios N, Young N, Wishart DS. MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic Acids Res*. 2009;37(Web Server issue):W652-60.
51. Fryar CD CM, Afful J. Prevalence of overweight, obesity, and severe obesity among adults

aged 20 and over: United States 1960–1962 through 2017–2018.: Division of Health and Nutrition Examination Surveys, NCHS Health E-Stats.; 2021.

52. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, et al. HMDB: the Human Metabolome Database. *Nucleic Acids Res.* 2007;35(Database issue):D521-6.

53. Saaristo TE, Barengo NC, Korpi-Hyövälti E, Oksa H, Puolijoki H, Saltevo JT, et al. High prevalence of obesity, central obesity and abnormal glucose tolerance in the middle-aged Finnish population. *BMC Public Health.* 2008;8(1):423-30.

54. Montaña LM, Sommer B, Gomez-Verjan JC, Morales-Paoli GS, Ramírez-Salinas GL, Solís-Chagoyán H, et al. Theophylline: Old Drug in a New Light, Application in COVID-19 through Computational Studies. *International Journal of Molecular Sciences.* 2022;23(8):4167-.

55. Juan MS. Methylxanthine Content in Commonly Consumed Foods in Spain and Determination of Its Intake during Consumption. *Foods.* 2017;6(12):109-.

56. Martínez-López S, Sarriá B, Gómez-Juaristi M, Goya L, Mateos R, Bravo-Clemente L. Theobromine, caffeine, and theophylline metabolites in human plasma and urine after consumption of soluble cocoa products with different methylxanthine contents. *Food Research International.* 2014;63(Part C):446-55.

57. The methylxanthine beverages and foods: Chemistry, consumption, and health effects: Progress in Clinical and Biological Research. Vol. 158. Edited by G. A. Spiller. Alan R. Liss, Inc., New York, 1984. pp. xii + 413. £60.00. ISBN 0-8451-5008-1. *Food and Chemical Toxicology.* 1985;23(10):949-.

58. Jilani TN, Preuss CV, Sharma S. Theophylline. *StatPearls.* Treasure Island (FL): StatPearls Publishing

Copyright © 2022, StatPearls Publishing LLC.; 2022.

59. Bellot PENR, Moia MN, Reis BZ, Pedrosa LFC, Tasic L, Barbosa F, Jr., et al. Are Phosphatidylcholine and Lysophosphatidylcholine Body Levels Potentially Reliable Biomarkers in Obesity? A Review of Human Studies. *Molecular nutrition & food research.* 2023:e2200568.

60. Baek SH, Kim M, Kim M, Lee JH, Kang M, Yoo HJ, et al. Metabolites distinguishing visceral fat obesity and atherogenic traits in individuals with overweight. *Obesity.* 2017;25(2):323-31-31.

61. Scherer M, Montoliu I, Qanadli SD, Collino S, Rezzi S, Kussmann M, et al. Blood plasma lipidomic signature of epicardial fat in healthy obese women. *Obesity (Silver Spring).* 2015;23(1):130-7.

62. Barber MN, Risis S, Yang C, Meikle PJ, Staples M, Febbraio MA, et al. Plasma lysophosphatidylcholine levels are reduced in obesity and type 2 diabetes. *PLoS One.* 2012;7(7):e41456.

63. Gall WE, Beebe K, Lawton KA, Adam K-P, Mitchell MW, Nakhle PJ, et al. α -Hydroxybutyrate Is an Early Biomarker of Insulin Resistance and Glucose Intolerance in a Nondiabetic Population. *PLOS ONE.* 2010;5(5):e10883.

64. Cobb J, Eckhart A, Motsinger-Reif A, Carr B, Groop L, Ferrannini E. α -Hydroxybutyric Acid Is a Selective Metabolite Biomarker of Impaired Glucose Tolerance. *DIABETES CARE.* 2016;39(6):988-95.

65. Rodríguez-García M, Fernández-Varo G, Hidalgo S, Rodríguez G, Martínez I, Morales-Ruiz M, et al. Validation of a Microwave-Assisted Derivatization Gas Chromatography-Mass Spectrometry Method for the Quantification of 2-Hydroxybutyrate in Human Serum as an Early Marker of Diabetes Mellitus. *Molecules.* 2022;27(6).

66. Peddinti G, Cobb J, Yengo L, Froguel P, Kravic J, Balkau B, et al. Early metabolic markers identify potential targets for the prevention of type 2 diabetes. *Diabetologia EXODIAB: Excellence of Diabetes Research in Sweden*. 2017;1-11.
67. Tricò D, Prinsen H, Giannini C, de Graff R, Juchem C, Li F, et al. Elevated α -Hydroxybutyrate and Branched-Chain Amino Acid Levels Predict Deterioration of Glycemic Control in Adolescents. *Journal of Clinical Endocrinology & Metabolism*. 2017:N.PAG-N.PAG.
68. Yousri NA, Suhre K, Yassin E, Al-Shakaki A, Robay A, Elshafei M, et al. Metabolic and Metabo-Clinical Signatures of Type 2 Diabetes, Obesity, Retinopathy, and Dyslipidemia. *Diabetes*. 2022;71(2):184-205.
69. Pang Y, Kartsonaki C, Lv J, Millwood IY, Fairhurst-Hunter Z, Turnbull I, et al. Adiposity, metabolomic biomarkers, and risk of nonalcoholic fatty liver disease: a case-cohort study. *The American journal of clinical nutrition*. 2022;115(3):799-810.
70. Kim H, Miller ER, Coresh J, Appel LJ, Rebholz CM, Lichtenstein AH, et al. Plasma Metabolites Associated with a Protein-Rich Dietary Pattern: Results from the OmniHeart Trial. *Molecular Nutrition and Food Research*. 2022;66(6).
71. Rebholz CM, Zheng Z, Coresh J, Grams ME, Appel LJ, Sarnak MJ, et al. Serum metabolites associated with dietary protein intake: Results from the Modification of Diet in Renal Disease (MDRD) randomized clinical trial. *American Journal of Clinical Nutrition*. 2019;109(3):517-25-25.
72. Gudbrandsen OA, Dyrøy E, Bohov P, Skorve J, Berge RK. The metabolic effects of thia fatty acids in rat liver depend on the position of the sulfur atom. *Chemico-Biological Interactions*. 2005;155(1/2):71-81.
73. Al-Sulaiti H, Diboun I, Agha MV, Mohamed FFS, Atkin S, Dömling AS, et al. Metabolic signature of obesity-associated insulin resistance and type 2 diabetes. *Journal of Translational Medicine*. 2019;17(1):348.
74. Perng W, Rifas-Shiman SL, Sordillo J, Hivert MF, Oken E. Metabolomic Profiles of Overweight/Obesity Phenotypes During Adolescence: A Cross-Sectional Study in Project Viva. *Obesity*. 2020;28(2):379-87-87.
75. Zhao Q, Shen H, Su KJ, Tian Q, Zhao LJ, Qiu C, et al. A joint analysis of metabolomic profiles associated with muscle mass and strength in Caucasian women. *Aging (Albany NY)*. 2018;10(10):2624-35.
76. Hasegawa T, Kakuta M, Yamaguchi R, Sato N, Mikami T, Murashita K, et al. Impact of salivary and pancreatic amylase gene copy numbers on diabetes, obesity, and functional profiles of microbiome in Northern Japanese population. *Scientific Reports*. 2022;12(1):7628.
77. Lin K, Zhu L, Yang L. Gut and obesity/metabolic disease: Focus on microbiota metabolites. *MedComm (2020)*. 2022;3(3):e171.
78. May KS, den Hartigh LJ. Gut Microbial-Derived Short Chain Fatty Acids: Impact on Adipose Tissue Physiology. *Nutrients*. 2023;15(2).
79. Saltelli A. Sensitivity analysis for importance assessment. *Risk Anal*. 2002;22(3):579-90.
80. Usher CL, Handsaker RE, Esko T, Tuke MA, Weedon MN, Hastie AR, et al. Structural forms of the human amylase locus and their relationships to SNPs, haplotypes and obesity. *Nature genetics*. 2015;47(8):921-5.
81. Walford GA, Florez JC, Gerszten RE, Ma Y, Clish C, Wang TJ. Metabolite profiles of diabetes incidence and intervention response in the diabetes prevention program. *Diabetes*. 2016;65(5):1424-33-33.

82. Vasishta S, Ganesh K, Umakanth S, Joshi MB. Ethnic disparities attributed to the manifestation in and response to type 2 diabetes: insights from metabolomics. *Metabolomics*. 2022;18(7):45.
83. Sun J, Yang R, Zhao M, Bovet P, Xi B. Tri-Ponderal Mass Index as a Screening Tool for Identifying Body Fat and Cardiovascular Risk Factors in Children and Adolescents: A Systematic Review. *Frontiers in Endocrinology*. 2021;12.
84. Fact Sheet: Cardiovascular diseases (CVDs) [Internet]: World Health Organization 2021 [cited 2022 Oct 5th]. Available from: [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)).
85. Jean-Pierre D. Metabolic Syndrome. [Elektronisk resurs]: MDPI AG; 2021.
86. Amna U, George AK, Lesley EC, Peter G, Kim EI, Christa LL. Childhood obesity and adult cardiovascular disease risk factors: a systematic review with meta-analysis. *BMC Public Health*. 2017;17(1):1-24.
87. Skinner AC, Perrin EM, Moss LA, Skelton JA. Cardiometabolic Risks and Severity of Obesity in Children and Young Adults. *The New England journal of medicine*. 2015;373(14):1307-17.
88. Yang Y, Song L, Wang L, Li D, Chen S, Wu S, et al. Effect of body mass index trajectory on lifetime risk of cardiovascular disease in a Chinese population: A cohort study. *Nutrition, Metabolism and Cardiovascular Diseases*. 2023;33(3):523-31.
89. Afshin A, Forouzanfar MH, Reitsma MB, Sur P, Estep K, Lee A, et al. Health Effects of Overweight and Obesity in 195 Countries over 25 Years. *The New England journal of medicine*. 2017;377(1):13-27.
90. NCD Countdown 2030: worldwide trends in non-communicable disease mortality and progress towards Sustainable Development Goal target 3.4. *Lancet (London, England)*. 2018;392(10152):1072-88.
91. Yu E, Malik VS, Hu FB. Reprint of: Cardiovascular Disease Prevention by Diet Modification: JACC Health Promotion Series. *Journal of the American College of Cardiology*. 2018;72(23):2951-63.
92. Satija A, Yu E, Willett WC, Hu FB. Understanding nutritional epidemiology and its role in policy. *Advances in Nutrition*. 2015;6(1):5-18-.
93. Lee S, Zhang C, Kilicarslan M, Brian, Bjornson E, Björn, et al. Integrated Network Analysis Reveals an Association between Plasma Mannose Levels and Insulin Resistance. *Cell Metabolism*. 2016;24(1):172-84.
94. Parnell LD, Noel SE, Bhupathiraju SN, Smith CE, Haslam DE, Zhang X, et al. Metabolite patterns link diet, obesity, and type 2 diabetes in a Hispanic population. *Metabolomics*. 2021;17(10):1-12.

8. Figures and Tables

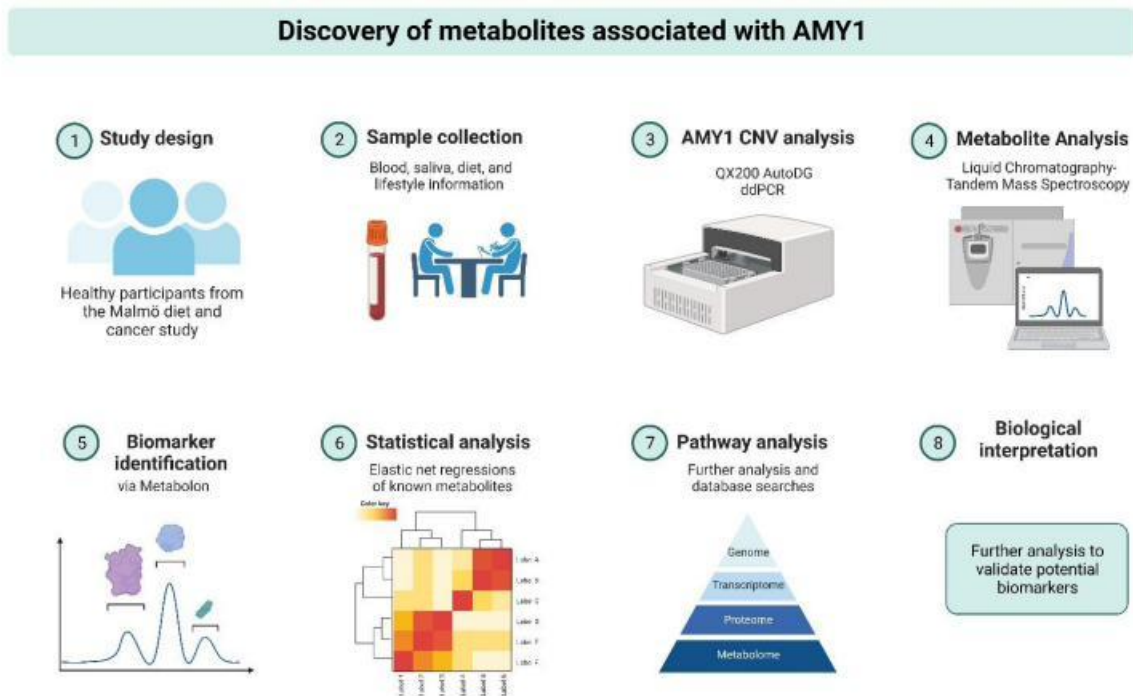


Figure 1. General overview of methods used in analysis of salivary amylase gene (AMY1) copy number variation (CNV) and metabolite data from the Malmö diet and cancer cohort study (MDCS). The image was generated using biorender.

What data was included in the analysis?

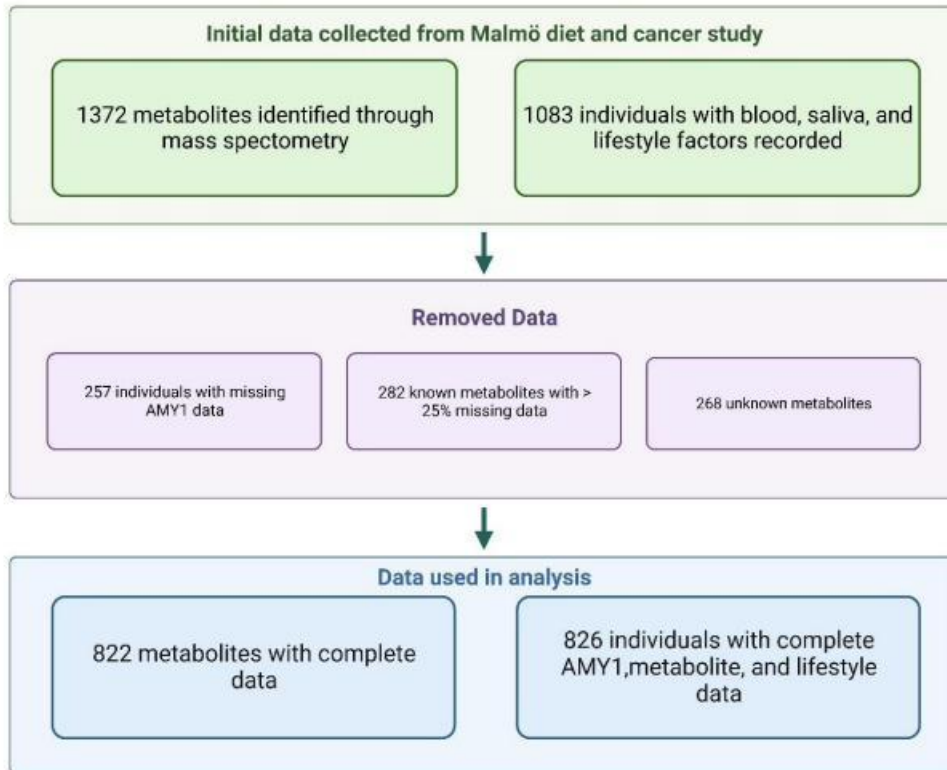


Figure 2. Overview of metabolite data selection process from the Malmö diet and cancer cohort study (MDCS).

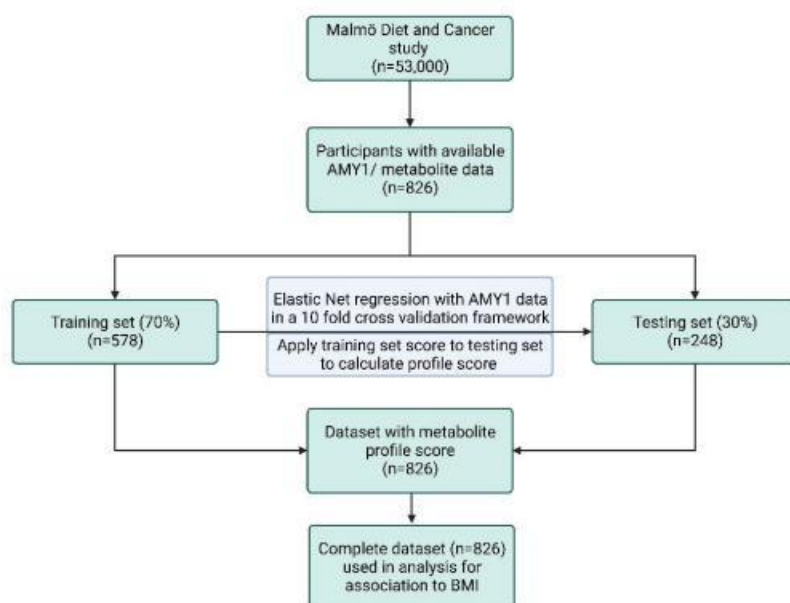


Figure 3. Schematic of elastic net methods used to select metabolites associated with salivary amylase gene (AMY1) copy number variation (CNV) from the Malmö diet and cancer cohort study (MDCS).

Table 1. Population demographic information of 826 participants from the Malmö diet and cancer cohort study (MDCS).

	N	%
Female sex	443	54
Physical Activity, MET-h/wk		
7.5-15	106	12.8
15-25	198	24.1
25-50	306	37.1
Above 50	134	16.1
Smoking Status		
Smoker	240	29.1
Former smoker	291	35.2
Never smoker	295	35.7
Education		
Elementary (<8 y)	388	46.9
Primary or secondary (9–10 y)	226	27.4
Upper secondary	72	8.7
Further education without a degree	57	6.9
University degree	83	10.1
Alcohol Consumption		
Non-consumers	55	6.7
Quantile 1	172	20.8
Quantile 2	153	18.5
Quantile 3	145	17.6
Quantile 4	163	19.7
Quantile 5	138	16.7
	Mean	SD
Body Mass Index, kg/ m	25.9	4.03
Age, years	58.0	6.01
AMY1 CNV	6.52	2.34
Starch, E%	24.5	4.53
Total energy intake, kcal/ day	2372	697

Met-h/wk:Metabolic equivalent hours/ week

AMY1: salivary amylase gene

CNV:copy number variation

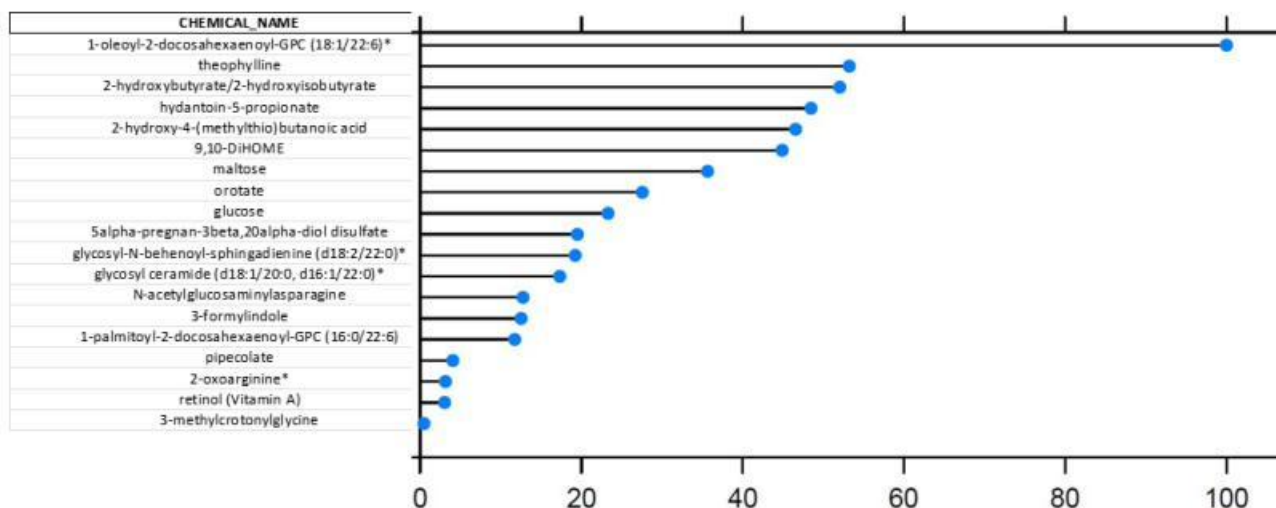


Figure 4. Graphical representation of the top 19 metabolites associated with salivary amylase gene (AMY1) copy number variation (CNV) identified through elastic net regression of metabolite data from the Malmö diet and cancer cohort study (MDCS). The numerical scale indicates the level of association with AMY1 CNV.

Table 2. Top 18 metabolites identified as being associated with salivary amylase gene (AMY1) copy number variation (CNV) from the Malmö diet and cancer cohort study (MDCS) using elastic net regression.

CHEMICAL NAME	SUPER PATHWAY	SUB PATHWAY
1-oleoyl-2-docosaheptaenoyl-GPC	Lipid	Phosphatidylcholine (PC)
theophylline	Xenobiotics	Xanthine Metabolism
2-hydroxybutyrate/2-hydroxyisobutyrate	Amino Acid	Glutathione Metabolism
hydantoin-5-propionate	Amino Acid	Histidine Metabolism
2-hydroxy-4-(methylthio)butanoic acid	Amino Acid	Methionine, Cysteine, SAM and Taurine Metabolism
9,10-DiHOME	Lipid	Fatty Acid, Dihydroxy
maltose	Carbohydrate	Glycogen Metabolism
orotate	Nucleotide	Pyrimidine Metabolism, Orotate containing
glucose	Carbohydrate	Glycolysis, Gluconeogenesis, and Pyruvate Metabolism
5alpha-pregnan-3beta,20alpha-diol disulfate	Lipid	Progestin Steroids
glycosyl-N-behenoyl-sphingadienine	Lipid	Hexosylceramides (HCER)
glycosyl ceramide	Lipid	Hexosylceramides (HCER)
N-acetylglucosaminylasparagine	Carbohydrate	Aminosugar Metabolism
3-formylindole	Xenobiotics	Food Component/Plant
1-palmitoyl-2-docosaheptaenoyl-GPC	Lipid	Phosphatidylcholine (PC)
pipecolate	Amino Acid	Lysine Metabolism
2-oxoarginine	Amino Acid	Urea cycle; Arginine and Proline Metabolism
retinol (Vitamin A)	Cofactors and Vitamins	Vitamin A Metabolism

Table 3. Unadjusted linear regressions of metabolite concentrations identified through elastic net regression of data from the Malmö diet and cancer cohort study (MDCS) against salivary amylase gene (AMY1) copy number variation (CNV). The unadjusted model only includes AMY1 CNV. The adjusted model controls for age, and sex.

CHEMICAL NAME	UNADJUSTED		ADJUSTED	
	BETA COEFFICIENT	P VALUE	BETA COEFFICIENT	P VALUE
1-oleoyl-2-docosahexaenoyl-GPC	.017	0.184	-.061	0.116
theophylline	.044	0.004*	.182	< 0.001**
2-hydroxybutyrate/2-hydroxyisobutyrate	.042	0.003*	.041	0.354
hydantoin-5-propionate	.008	0.507	.001	0.988
2-hydroxy-4-(methylthio)butanoic acid	.028	0.023*	-.005	0.891
9,10-DiHOME	-.035	0.015*	-.004	0.938
maltose	.024	0.062	.046	0.268
orotate	.019	0.038*	.022	0.455
glucose	.028	0.046*	.1	0.021*
5alpha-pregnan-3beta,20alpha-diol disulfate	-.029	0.046*	.006	0.904
glycosyl-N-behenoyl-sphingadienine	.025	0.053	-.033	0.407
glycosyl ceramide	.032	0.023*	.01	0.816
N-acetylglucosaminylasparagine	.039	0.011*	.082	0.094
3-formylindole	.012	0.323	-.010	0.797
1-palmitoyl-2-docosahexaenoyl-GPC	.033	0.017*	.043	0.304
pipecolate	-0.009	0.531	.041	0.355
2-oxoarginine	-.035	0.017*	-.082	0.078
retinol (Vitamin A)	.030	0.026*	-.006	0.88

* Indicates a significance level <0.05

**Indicates a significance level passing the multiple testing threshold level <0.0026

Table 4. Unadjusted linear regressions of metabolite concentrations identified through elastic net regression of data from the Malmö diet and cancer cohort study (MDCS) against Body Mass Index (BMI). The unadjusted model only includes BMI. Model 2 model controls for age, and sex. Model 3 adjusts for age, sex, smoking status, alcohol consumption, education, total energy consumed, total starch intake, and physical activity level.

CHEMICAL NAME	MODEL 1 (UNADJUSTED)		MODEL 2		MODEL 3	
	BETA COEFFICIENT	P VALUE	BETA COEFFICIENT	P VALUE	BETA COEFFICIENT	P VALUE
1-oleoyl-2-docosahexaenoyl-GPC	-.033	<.001**	-.036	<.001**	.039	<.001**
theophylline	.025	0.004*	.025	0.006*	.024	0.006*
2-hydroxybutyrate/2-hydroxyisobutyrate	.067	<.001**	.064	<.001**	.064	<.001**
hydantoin-5-propionate	.028	<.001**	.023	<.001**	.021	0.002**
2-hydroxy-4-(methylthio)butanoic acid	.031	<.001**	.028	<.001**	.027	<.001**
9,10-DiHOME	-.012	0.163	-.012	0.137	-.014	0.093
maltose	.014	0.066	.014	0.075	.013	0.082
orotate	.017	0.002**	.014	0.009*	.012	0.026*
glucose	.074	<.001**	.072	<.001**	.074	<.001**
5alpha-pregnan-3beta,20alpha-diol disulfate	-.015	0.086	-.011	0.178	-.010	0.237
glycosyl-N-behenoyl-sphingadienine	-.014	0.065	-.017	0.016**	-.020	0.006**
glycosyl ceramide	-.009	0.248	-.014	0.092	-.015	0.062
N-acetylglucosaminylasparagine	.001	0.929	-.002	0.845	2.11E-5	0.998
3-formylindole	.006	0.41	.005	0.453	.007	0.361
1-palmitoyl-2-docosahexaenoyl-GPC	.010	0.214	.006	0.454	.008	0.314
pipecolate	-.018	0.029*	-.016	0.054	-.014	0.087
2-oxoarginine	.049	<.001**	.049	<.001**	.048	<.001**
retinol (Vitamin A)	.001	0.868	-.002	0.806	1.096E-5	0.999

* Indicates a significance level <0.05

**Indicates a significance level passing the multiple testing threshold level <0.0026

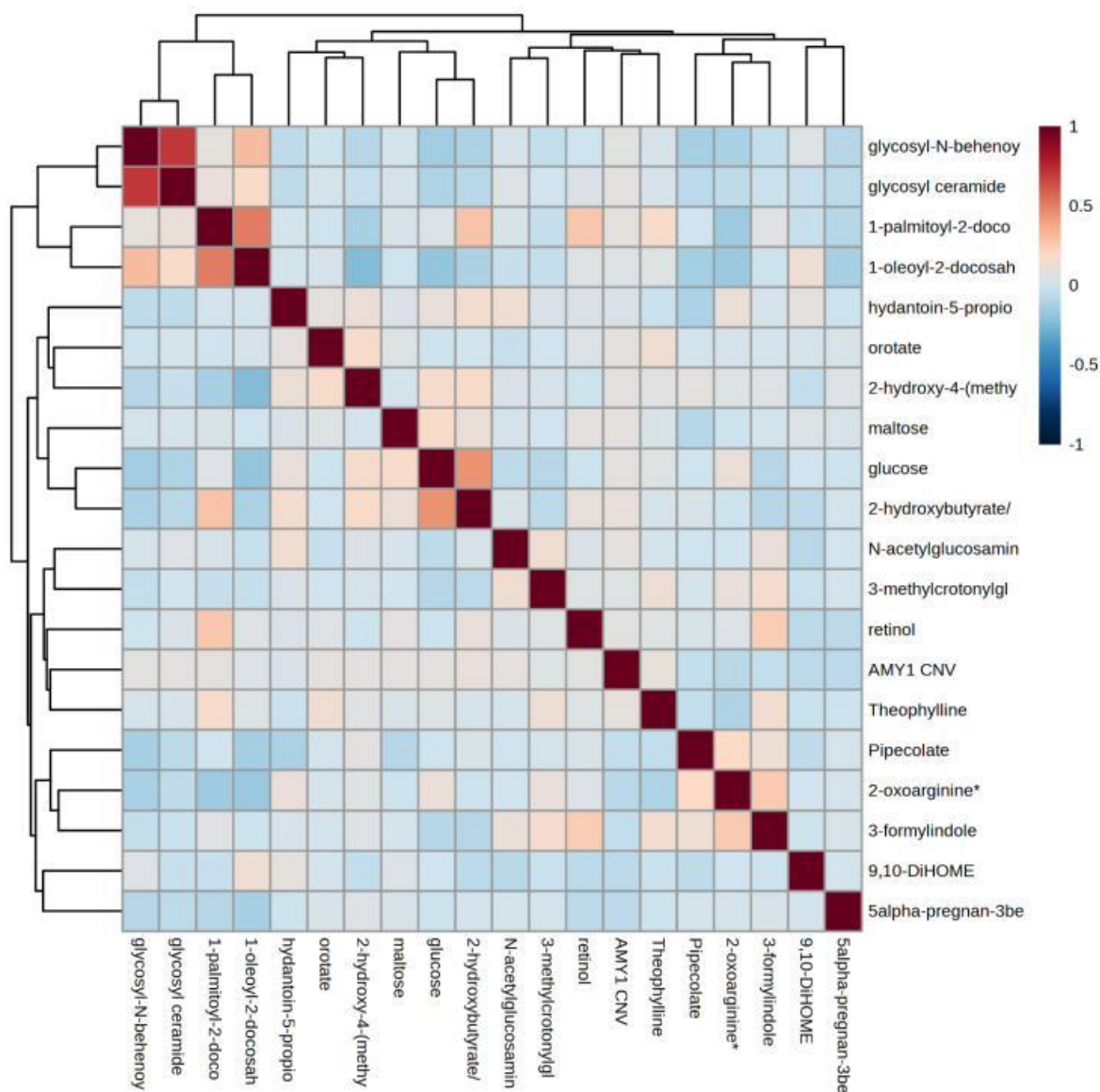


Figure 5. Pearson regression coefficient plot of the top metabolites associated with salivary amylase gene (*AMY1*) copy number variation (CNV) identified through elastic net regression of data from the Malmö diet and cancer cohort study (MDCS). The image was generated through the R package “metaboanalyst”. Colour denotes the association directions (red indicates a positive direction and blue indicates a negative direction) as well as magnitudes (the darker the colour the stronger the association).

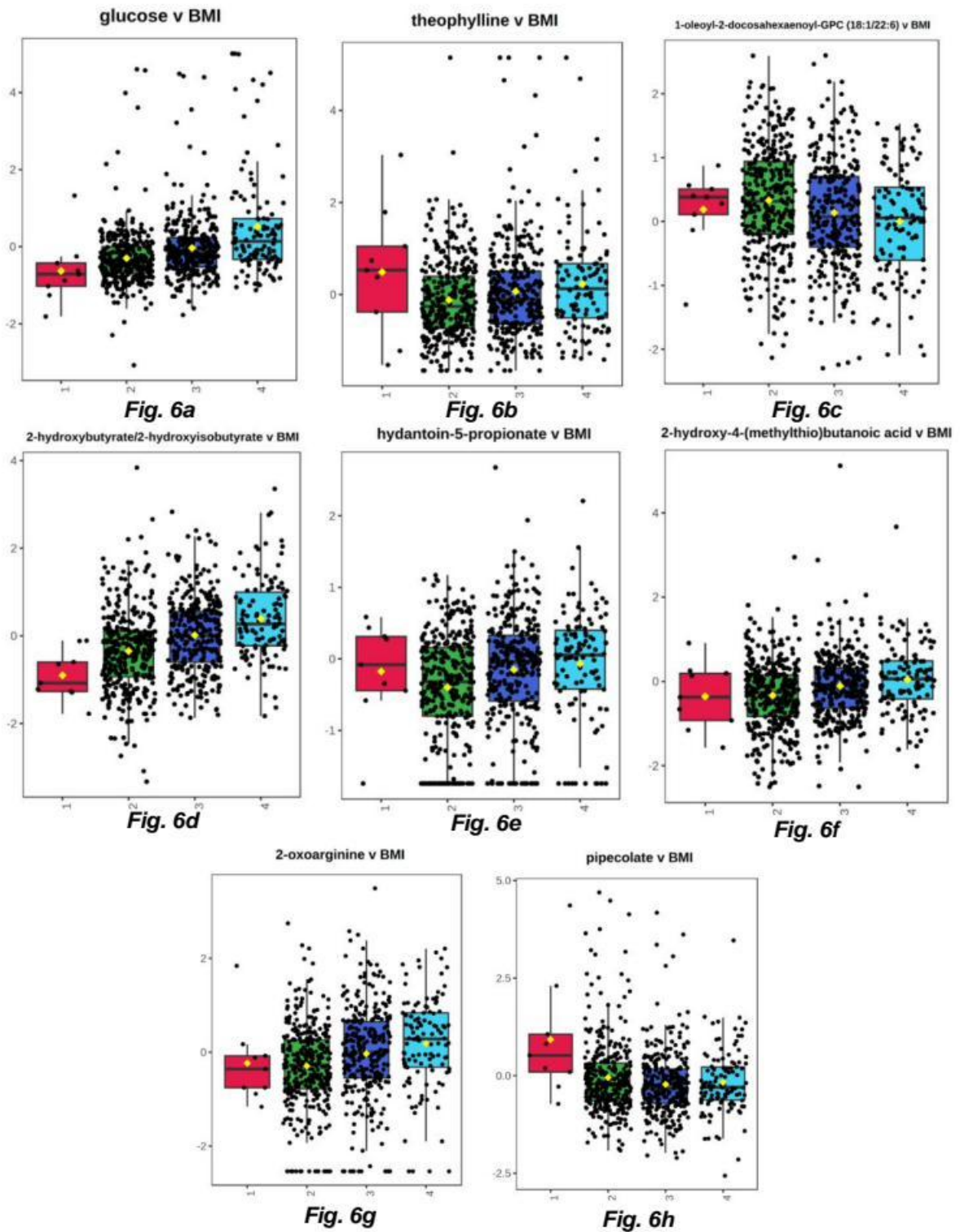


Fig 6a-h: ANOVA analysis of plasma metabolite concentrations against salivary amylase gene (*AMY1*) copy number and body mass index (BMI) quartiles in the Malmö diet and cancer cohort study (n=826).

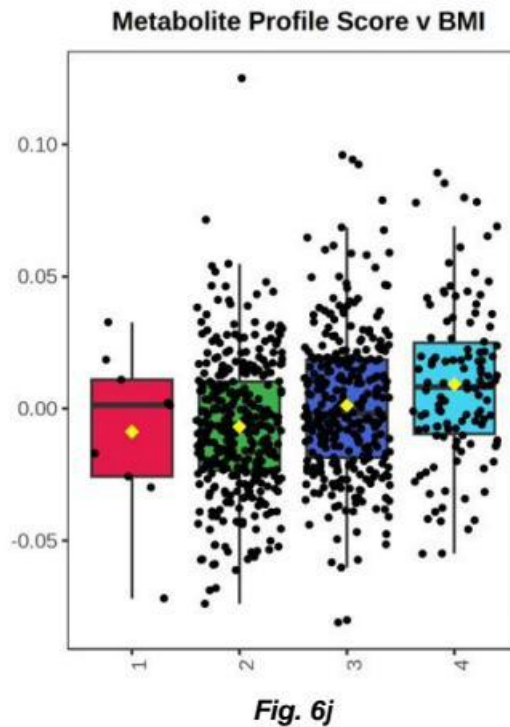
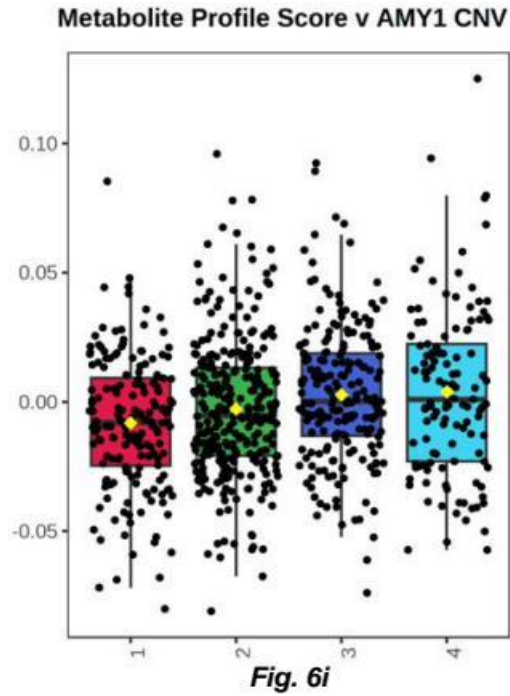


Figure 6i-j: ANOVA analysis of the metabolite profile score generated through elastic net regression of plasma metabolite concentration from a sub cohort of the Malmö diet and cancer cohort study against salivary amylase gene (AMY1) copy number (figure i) and body mass index BMI (figure j) in quartiles.

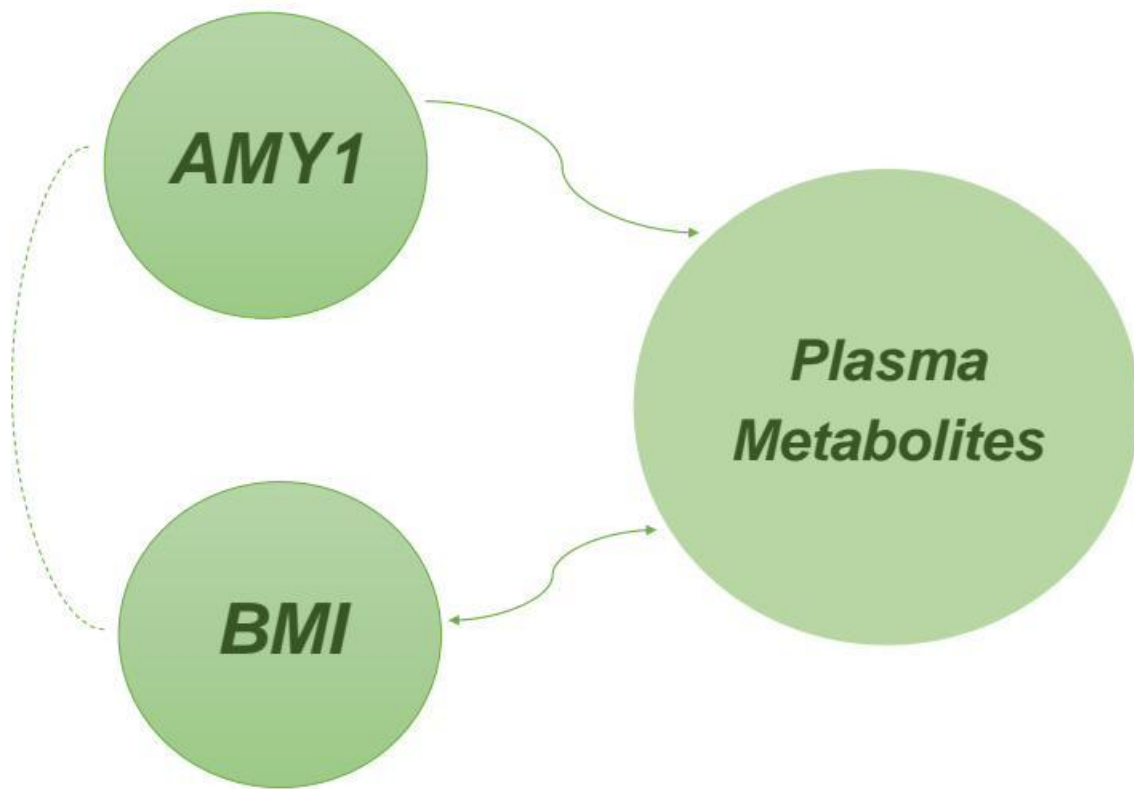


Figure 7. Schematic of the indirect relationship between body mass index (BMI), salivary amylase gene (AMY1) copy number variation (CNV), and plasma metabolite concentrations in a sub cohort of the Malmö diet and cancer cohort study.

9. Popular science summary

Salivary amylase is an enzyme responsible for the initial digestion of starches in the mouth. Salivary amylase creates intermediate molecules known as metabolites that are further broken down throughout the digestive tract. This enzyme is encoded by the gene *AMY1*, and humans can carry anywhere between 2 and 17 copies of this gene. The number of *AMY1* genes a person carries is directly proportional to the concentration of salivary amylase found in their saliva. Some research has suggested that *AMY1* copy number could influence a person's risk for increased adiposity, or CVD depending on their diet. Machine learning methods were used to create a predictive model for plasma metabolite concentrations. This model was used to determine if any metabolites were moderated by *AMY1* copy number. At a population level, our study found that there was no association between *AMY1* copy number and BMI. However, metabolites selected by the model displayed nominally significant associations with either *AMY1* CNV ($p < 0.05$). The model identified several metabolites that were also found to be significantly associated with BMI. This suggests that the metabolites identified could play important roles as biomarkers in the mechanisms that link *AMY1* to BMI and CVD risk. Biomarkers are links between the genes that moderate metabolism and the physiological state of a person, which is guided by their dietary and lifestyle habits. Understanding these links will help to identify the mechanisms at play between our genomes, our diets, and their interaction in disease risk moderation.

10. Acknowledgments

I would like to thank Yan Borné and all of Emily Sonestedt's research group for the opportunity to work with this data. I have thoroughly enjoyed learning more about this field, and I know that I have gained new insights as a researcher. This thesis has been written in 9 cities and 4 countries, and the support I received from family and friends along the way was vital.