Title

Proteomic and Epigenetic Biomarker Discovery to Predict Health-Related Lifestyle Traits

Author

Manickam, Joshua

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Proteomic and Epigenetic Biomarker Discovery to Predict Health-Related Lifestyle Traits

A thesis submitted in partial satisfaction

of the requirements for the degree Master of Science

in Bioinformatics

by

Joshua Zachary Manickam

2024

ABSTRACT OF THE THESIS

Proteomic and Epigenetic Biomarker Discovery to

Predict Health-Related Lifestyle Traits

by

Joshua Zachary Manickam

Master of Science in Bioinformatics

University of California, Los Angeles, 2024

Professor Matteo Pellegrini, Chair

Epigenetics and proteomics have emerged as powerful fields with the potential to transform our understanding of the relationship between lifestyle factors and health outcomes. This paper utilizes data from an exploratory, cross-sectional study conducted by Prosper DNA Inc. to identify proteomic and epigenetic biomarkers that can predict clinical lifestyle traits related to individual health. The study collected data from two groups, one representing healthy individuals and the other of unhealthy individuals with specific lifestyle characteristics. Imputation, Principal Component Analysis (PCA), and correlation analysis were performed on the proteomic and

methylation data, which resulted in 10 principal components (PCs) each and 3 correlation matrices. Linear regression models were developed using proteomic or methylation PCs as predictors for each lifestyle trait. The models were evaluated using Leave-One-Out-Cross-Validation and Pearson correlation coefficient (r) to determine significance and confirm the accuracy of the models. 19 significant proteomic models and 28 significant methylation models were identified. Notably, 15 models across the proteomic and methylation models were directly associated with the original selection criteria, such as body mass index (BMI), fitness, and dietary habits. Through a correlation analysis, PC3, PC9, and PC9 of the proteomic PCs and MPC1, MPC3, and MPC4 of the methylation PCs were selected for PC loading analysis based on the strength and significance of the correlation between a given PC and the traits associated with the biomarkers. The most influential proteins and CpG sites were extracted from the loading and passed through STRING and Cistrome, respectively, to gather protein network information and transcription factor information. For the proteomic PCs, the proteins were involved in networks relating to platelet activation/coagulation, inflammatory response, and the regulation of proteolytic activity. For the methylation PCs, the transcription factors that bind to the CpG sites showed some relation to adipogenesis and tumor suppression. Future steps should include deeper analyses of the individual proteins, CpG sites, and transcription factors to provide concrete validation for the trends observed in this study.

The thesis of Joshua Zachary Manickam is approved.

Xia Yang

Xianghong Jasmine Zhou

Matteo Pellegrini, Committee Chair

University of California, Los Angeles

2024

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

ID number	IDNM
1=life; 2=control	GROUP
1=male; 2=female	Sex
age years	Age
body weight kg	Weight (kg)
height cm	Height (cm)
body mass index	BMI
fat mass kg	fatmass
fat free mass kg	FFM
% body fat	BodyFat%
waist circumference	Waist (cm)
abdominal diameter cm	Sagittal
fat mass index	FMI
fat free mass index	FFMI
height meters	HTM
waist circ meters	WCM
Fitness score 1-10	FITNESS
Exercise days/wk category	EXERCFREQ
hand grip kg	HANDGRIPTOT
leg back strength	Leg/Back (kg)
Leg back strength (weight adjusted)	Rel. Leg/Back (per kg)
vo2max estimate	VO2max
total fruit vegetable inake	FRUIT+VEGE

Red meat intake	REDMEAT
white blood cells	WBC
hemoglobin	HBG
platelets	PLT
neutrophils	NE (ABS)
lymphocytes	LY (ABS)
monocytes	MON (ABS)
blood urea nitrogen	BUN
creatine	Creatinine
glomerular function	eGFR na
glomerular function (african-american)	eGFR aa
BUN/Creatinine Ratio	BUN/Creatinine Ratio
albumin	ALBUMIN
globulin	GLOBULIN
Albumin/globulin ratio	A/G RATIO

ACKNOWLEDGEMENTS

I would like to take this moment to thank my esteemed PI, Professor Matteo Pellegrini. Over the past three years, the wealth of knowledge I was able to glean from working as a part of his laboratory is invaluable. Throughout this project, Professor Pellegrini has been as dedicated as he has been patient, and I am endlessly grateful for being given the opportunity to work with him. I would also like the chance to thank Dr. Xia Yang and Dr. Jasmine Zhou for agreeing to join my committee and review the culmination of my work these last three years. I have admired their work over the years, so it is a privilege to be able to present my findings to them. I would also like to thank ProsperDNA Inc. for providing the proteomic and methylation data. None of this would have been possible without the research they conducted. Finally, I would like to thank my parents and my brother for their continuous, unwavering love and support.

1. INTRODUCTION

Epigenetics and proteomics are two rapidly expanding fields that have the potential to revolutionize our understanding of the relationship between lifestyle factors and health outcomes. Epigenetic modifications, such as DNA methylation, can be influenced by a range of environmental and lifestyle factors and have been shown to play a critical role in the development of various diseases. Similarly, proteomic data, which provides a comprehensive snapshot of the proteins present in a biological sample, can reveal important insights into the molecular mechanisms underlying disease development and progression. In addition to monitoring disease progression, proteomic data can also provide an understanding of the effect of lifestyle factors on protein abundance in the blood. Much like DNA methylation, lifestyle factors have been shown to play a role in proteomic changes. For instance, in a 2021 study analyzing proteomic biomarkers related to cardiovascular disease, the researchers found statistically significant associations between 60 proteins with smoking, 30 proteins with alcohol consumption, and 5 proteins with physical activity (Corlin et al. 2021). Research such as these studies exemplifies the value of analyzing proteomic data and can motivate future studies.

This was the primary motivation behind an explorational, cross-sectional study conducted by Prosper DNA Inc., a biotechnology company specializing in epigenetic analysis, that compared epigenetics and global proteomics outcomes in two groups (n=50 each) varying widely in selected lifestyle factors. In particular, they sought to understand the molecular drivers of lifestyle traits related to an individual's health, such as body mass index (BMI), diet, exercise, etc. To accomplish this, they gathered data from individuals across two groups with no current chronic or infectious disease history. The Lifestyle group consisted of 50 “healthy” individuals

who were not overweight or obese (BMI < 25 kg/m^2), non-smokers for at least three years prior, and a healthy dietary pattern. The Control group consisted of 50 “unhealthy individuals” who were overweight or obese (BMI > 25 kg/m^2) with an unhealthy dietary pattern. Prosper's data came in 35 mL blood samples, questionnaires, body composition measurements, and strength tests. The questionnaires listed questions related to lifestyle habits, physical activity levels, and mood profiles. The body composition measurements recorded values like height, weight, waist circumferences, abdominal diameter, body fat (bioelectrical impedance or BIA), etc. The strength tests consisted of handgrip and leg/back strength dynamometer tests. Finally, proteomic data was gathered from the blood samples through global proteomics, and methylation data was developed through the bisulfite conversion and subsequent cytosine estimation from DNA extracted from blood, saliva, and buccal samples. The researchers utilized this proteomic information and extracted concentrations of 870 unique proteins from the 100 samples. The methylation data was gathered in the form of beta values, which denote the percentage of methylation a particular CpG is subject to. They gathered beta values for over 130,000 sites and recorded them in a matrix separate from the proteomic data.

Using the vast amount of information assembled by Prosper DNA, we were most curious about the development of proteomic and epigenetic biomarkers that can predict clinical lifestyle traits. To accomplish this, we will develop linear regression models for each lifestyle trait, utilizing either proteomic or methylation data as the predictors. The hope of this research is to discover significant proteins/methylation sites related to clinical lifestyle traits and eventually create informed hypotheses around the biological basis for the associations we observe. The discovery of biomarkers such as these can provide critical insight into some of the molecular drivers of health and uncover possible therapy targets for managing unfavorable lifestyle traits.

2. METHODS

2.1. Data Pre-Processing

First, we received the data across 96 samples from Prosper DNA in the form of two datasets: the Prosper Proteomic Lifestyle Data and the Prosper Methylation Data. We split the first dataset into two since the lifestyle trait data and the proteomic data were on the same file. What resulted was a lifestyle trait data matrix with 94 traits across 96 samples and a proteomic data matrix with measurements of 882 proteins across 96 samples. The methylation data matrix consisted of measurements of methylation percentage across roughly 131,000 CpG sites. Using R, we used k-nearest-neighbors imputation to impute the missing data since the proteomic and lifestyle trait matrices had missing values (Gardner & Freitas, 2021). Following imputation, we performed principal component analysis (PCA) on the proteomic and methylation datasets. Since both datasets contained far more features than samples, PCA is essential for building generalizable models and avoiding overfitting the data (Pramoditha, 2021). To preserve the variability of the data while also avoiding the need for penalized regression, we shrunk each dataset down to 10 principal components (PCs) each, with the proteomic PCs denoted as PC1, PC2, PC3, etc., and the methylation PCs denoted as MPC1, MPC2, MPC3, etc.

2.2. Correlation Analysis

The next step we undertook was the assembly of correlation matrices and their subsequent correlation plots. Using cor() in R, we computed three sets of correlation matrices and their graphical representations: the proteomic-lifestyle trait correlation matrix, the

methylation-lifestyle trait correlation matrix, and the methylation-proteomic correlation matrix. The first matrix correlated the proteomic PCs with traits, the second correlated methylation PCs with traits, and the third correlated proteomic PCs with methylation PCs. This step is essential in understanding the variance of the data. Since principal components are ordered by the amount of variance they explain from the original dataset, utilizing the correlation matrices and plots can provide insight into where the variation comes from and understand whether the PCs that produce the variation have strong associations with particular traits (Kumar, 2023).

2.3. Linear Regression Model-Building, Selection, and Analysis

The next step involved building linear regression models for each trait, utilizing either proteomic or methylation PCs as predictors and the specific lifestyle trait as the response variable. This step would produce a total of 188 models: 94 models that predict lifestyle traits from the ten proteomic PCs and 94 that predict lifestyle traits from the 10 methylation PCs. Due to the small sample size, we opted to use Leave-One-Out-Cross-Validation (LOOCV) to validate and evaluate the performance of each model while maintaining the robustness across the models (Brownlee, 2020). As is common practice in the field of genetics and epigenetics, we evaluated our models using the Pearson Correlation Coefficient ® found by compiling the predictions gathered from LOOCV and correlating them to the actual values from the original datasets (Waldman, 2019). The cutoff we decided to use regarding the r value in selecting significant models was 0.3. We validated the significance by calculating the p-value of the r-value of each selected model. Given our sample size and our r-value cutoff, the p-value is always guaranteed to be less than 0.05 if the r-value is greater than or equal to 0.3, which does indicate that a particular

model is significant and is worthy of further analysis (Tredennick et al. 2021). Once we organized lists of valid proteomic and methylation models, we produced scatterplots for each model, plotting the predicted values against the actual values to visualize how closely our model predictions reflect the original data. This step would also allow us to analyze the data's spread and determine the quality and quantity of the data.

2.4. Principal Component Analysis

We inspected the loading matrices of influential PCs to identify which sets of proteins or CpG sites might contribute to changes in the lifestyle traits that our selected biomarkers model. To accomplish this, we started by computing the correlation coefficients between every PC and every selected lifestyle trait and developing heatmaps with hierarchical clustering. This was done for both sets of models, and it allowed us to identify influential PCs, characterize correlations between traits and PCs as positive or negative, and visualize natural groupings produced by the hierarchical clustering. We also created a grid of p-values of all the resulting correlation coefficients to mark correlations as significant if the p-value does not exceed the 0.05 threshold.

Once we identified the most influential PCs, we inspected the loading matrices and separated the entries based on the sign of their PC weight value. The sign of a given PC weight indicates the association of that site or protein with the PC it is contained within. For instance, an increase in the concentration of a protein with a negative loading weight results in a decrease in the overall PC value. By separating the entries and organizing the resulting data frames in order of descending absolute value, we were able to understand which proteins or CpG sites were influential and whether they were positively/negatively associated with a given PC. As such, the

subsequent analyses of these proteins and CpG sites would be performed separately on the negative and positive lists.

2.5. Transcription Factor Analysis using Cistrome

The Cistrome Project is a web-based, bioinformatic hub of transcription factor binding and histone modification data. It provides access to a range of tools, such as the CistromeDB Toolkit, that can construct a list of transcription factors that have significant binding overlap with the peak set submitted by the user. This peak set we submitted came in the form of a BED file that hosted the following information:

1. Chromosome number (chr1, chrX, etc)
2. Start position of the segment
3. End position of the segment

This information was taken from the methylation PC loading matrices. Since the loading matrices contain the column names of the original methylation matrix, we were able to extract the influential CpG sites from the loading matrices. The sites came in the format "chr1_123456", with the chromosome number and start position being clear. The end position was denoted as the start position+1. Once we aggregated that information, we assembled BED files for the PCs of interest, using the top 200 sites with the highest absolute value for both the negative and positive lists. We then submitted these files to Cistrome for analysis and received GIGGLE plots for each submission. The CistromeDB Toolkit makes use of GIGGLE software, which searches the Cistrome database to find ChIP-Seq, DNase-seq, or ATAC-seq samples that are similar to the set of binding sites submitted by the user and returns a list and plot of transcription factors that are

likely to bind to those sites. A transcription with a higher GIGGLE score, corroborated by multiple samples in the database, is expected to bind to sites in the peak set. Thus, for each methylation PC of interest, we were able to find transcription factors that bind to sites that are positively associated with the PC and negatively associated with the PC. We then took the results from Cistrome and analyzed the plots to determine which transcription factors are likely involved in each set of sites. We prioritized the transcription factors with the highest GIGGLE scores and those with multiple samples. Using NCBI and conducting literature searches, we characterized each transcription factor of interest based on its function (predicted or known) and possible associations with various diseases.

2.6. Protein Network Analysis using STRING

We used the STRING database for the proteomic PCs to identify which biological networks the proteins in the loading matrices may be involved in. This would allow us to biologically ascertain differences between healthy and unhealthy people at the protein network level. The process to extract the protein names from the ordered negative and positive data frames of each influential PC remains largely the same, aside from converting the data into a BED file. STRING's input only requires submitting a list of valid protein identifiers, which the loading matrices already held. It was a simple task of extracting the top 100 proteins and running them through STRING. Following submission, we visualized the potential networks that our significant proteins were involved in. We also received lists of GO terms corresponding to various biological processes these networks may represent. Since STRING produces a large number of potential networks, we focused on GO terms with the lowest false discovery rates

(FDR). Utilizing FDR as a filtering measure allowed us to recognize patterns within the protein sets and patterns across PCs.

3. RESULTS

3.1. 19 Significant Proteomic Models

model	rval
lm_GROUP	0.482132766497672
lm_Sex	0.521051189589965
lm_BMI	0.319935039056922
lm_fatmass	0.459196066372566
lm_BodyFat_Percentage	0.504733363570983
lm_Sagittal	0.305665533455721
lm_FMI	0.476901921551883
lm_FITNESS	0.346360249557728
lm_YEARSEXERCISE	0.351233133842298
lm_VO2max	0.44780475520097
lm_FRUIT_AND_VEGE	0.317559647032625
lm_WBC	0.798583857572712
lm_HBG	0.453684627907896
lm_PLT	0.400528180033362
lm_NE__ABS_	0.791284639073476
lm_MON__ABS_	0.319355623050436
lm_GLOBULIN	0.464259277502055
lm_A_to_G_Ratio	0.395367823889042
lm_ALKALINE_PHOS	0.32057773032278

Figure 1: This is the list of valid proteomic models and their associated r values. They begin with an arbitrary identifier ('lm_') and are followed by the specific trait that the proteomic PCs are predicting

Of the 94 proteomic models we produced, 19 satisfied the $r \geq 0.3$ threshold. Figure 1 lists the traits that are significantly associated with proteomic PCs. Out of these significant models, 7 models directly related to the selection criteria outlined in the original Prosper DNA study: BMI, Fat Mass, Fat Mass Index (FMI), Fitness, Body Fat Percentage, Years of Exercise, and Fruit and

Vegetable Intake (FRUIT_AND_VEGE). For this paper, we highlighted two proteomic models in green to provide examples of the performance of our models of interest. These models directly relate to the selection criteria mentioned previously. Figures 2 and 3 below depict the scatterplots that compare the predicted values against the actual values of the BMI and FITNESS models.

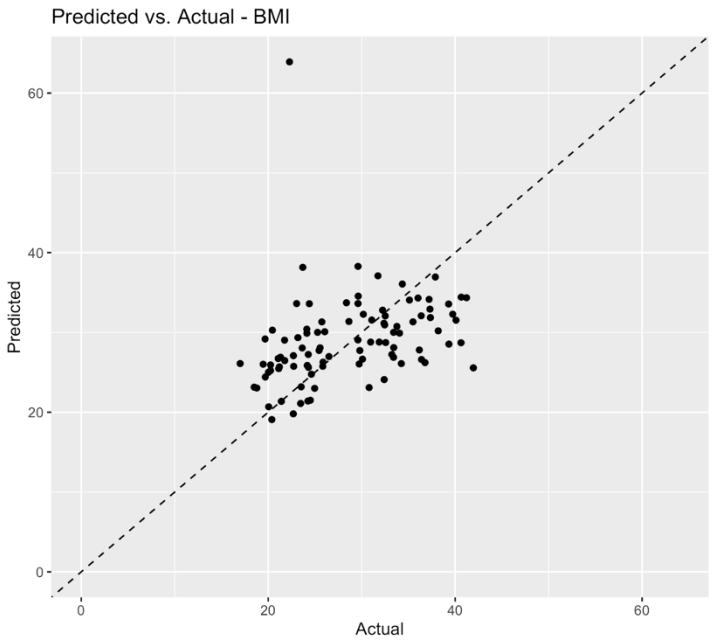

Figure 2: Performance of Proteomic PCs vs BMI model

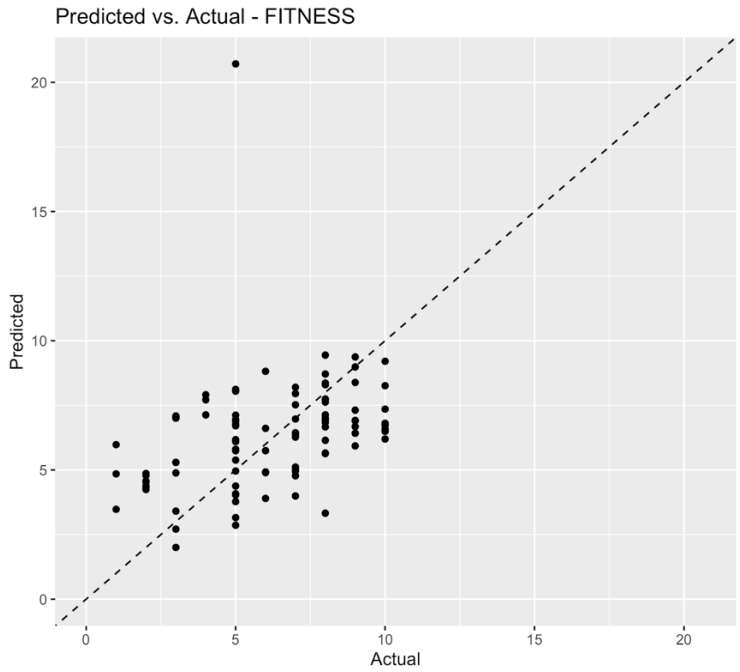

Figure 3: Performance of Proteomic PCs vs FITNESS model

Figures 4 and 5 below depict the model summaries of the models we decided to highlight. Regarding the statistics below, the Estimate column holds the β, or the coefficient, to be multiplied by the value of a particular PC within the linear regression model. The magnitude and the sign determine how much a given PC influences change in the model and in which direction change is applied, respectively (negatively or positively). The '*' in the rightmost column indicates the significance level of the β estimate. More '*' denotes a low p-value. In Figure 4, the highest magnitude β estimate was attached to PC9 and had one of the lowest p-values. In Figure 5, the highest magnitude β estimate was attached to PC3 and had one of the lowest p-values.

```
lm_BMI Summary:


Call:
lm(formula = formula, data = n_train_set)

Residuals:
     Min       1Q   Median       3Q      Max
-12.8652  -4.1814  -0.1438   3.3031  15.4760

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.65374    0.55968  51.196  < 2e-16 ***
PC1          0.06796    0.14091   0.482 0.630855
PC2          0.05817    0.18304   0.318 0.751407
PC3          0.95988    0.22425   4.280 4.92e-05 ***
PC4         -0.50249    0.25586  -1.964 0.052847 .
PC5         -1.01564    0.26807  -3.789 0.000284 ***
PC6         -0.48701    0.28550  -1.706 0.091739 .
PC7         -0.07327    0.29683  -0.247 0.805628
PC8          0.23948    0.30617   0.782 0.436324
PC9          1.13779    0.30706   3.705 0.000377 ***
PC10         0.20376    0.31263   0.652 0.516335
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 4: BMI model summary. The highlighted PC9 is the highest magnitude PC in this model

```
lm_FITNESS Summary:


Call:
lm(formula = formula, data = n_train_set)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6166 -1.6613  0.2347  1.5230  4.0505

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.152628   0.216078  28.474  < 2e-16 ***
PC1          0.054565   0.054400   1.003  0.31872
PC2         -0.017285   0.070666  -0.245  0.80737
PC3         -0.493229   0.086577  -5.697 1.76e-07 ***
PC4          0.098217   0.098780   0.994  0.32294
PC5          0.089864   0.103495   0.868  0.38771
PC6          0.264475   0.110224   2.399  0.01863 *
PC7          0.023625   0.114598   0.206  0.83717
PC8          0.005812   0.118205   0.049  0.96090
PC9         -0.365707   0.118546  -3.085  0.00276 **
PC10        -0.230085   0.120696  -1.906  0.06003 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 5: FITNESS model summary. The highlighted PC3 is the highest magnitude PC in this model

3.2. 28 Significant Methylation Models

model	rval
Mlm_Sex	0.997678158356424
Mlm_Age	0.749391582513885
Mlm_Weight__kg_	0.416716572657052
Mlm_Height__cm_	0.704283311440012
Mlm_BMI	0.326027094466076
Mlm_fatmass	0.334100016741829
Mlm_FFM	0.753992283108216
Mlm_BodyFat_Percentage	0.532219260760633
Mlm_Waist__cm_	0.366002228159308
Mlm_Sagittal	0.378486625538492
Mlm_FMI	0.401599840304744
Mlm_FFMI	0.753992283108216
Mlm_HTM	0.704283311440012
Mlm_WCM	0.366002228159308
Mlm_EXERCFREQ	0.316265110225556
Mlm_HANDGRIPTOT	0.355877144691991
Mlm_Leg_and_Back_kg	0.707937959665555
Mlm_Rel_Leg_and_Back_per_kg	0.40900422964645
Mlm_VO2max	0.612203318902866
Mlm_REDMEAT	0.361487607700574
Mlm_HBG	0.482637215106772
Mlm_NE__ABS_	0.506910062115382
Mlm_LY__ABS_	0.328745919731163
Mlm_BUN	0.345576716200666
Mlm_Creatinine	0.607014237830618
Mlm_eGFR_na	0.467045857951259
Mlm_eGFR_aa	0.466558815376031
Mlm_ALBUMIN	0.300682061985815

Figure 6: This is the list of valid methylation models and their associated r values. They begin with an arbitrary identifier ('Mlm_') and are followed by the specific trait that the methylation PCs are predicting.

For the methylation biomarkers, 28 of the models we produced satisfied the r value ≥ 0.3 threshold, and these models are listed in Figure 6. Of these selected models, 8 models directly related to the original selection criteria: BMI, Fat Mass, Free Fat Mass/Free Fat Mass Index (FFM/FFMI), Body Fat Percentage, FMI, Exercise Frequency (EXERCFREQ), Red Meat Intake (REDMEAT), and Weight (kg). As with the proteomic models, we highlighted just two methylation biomarkers that pertain to the selection criteria: the methylation BMI model and the

methylation EXERCFREQ model. Figures 7 and 8 are the corresponding scatterplots of these models.

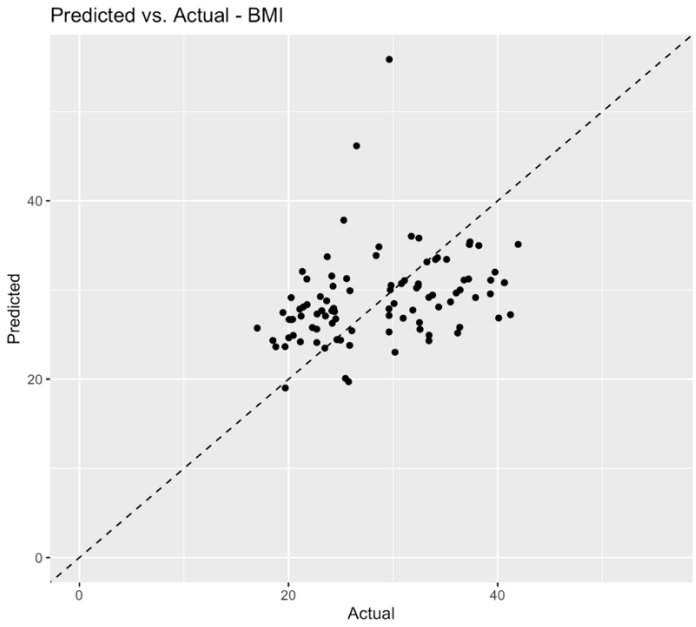

Figure 7: Performance of Methylation PCs vs BMI model

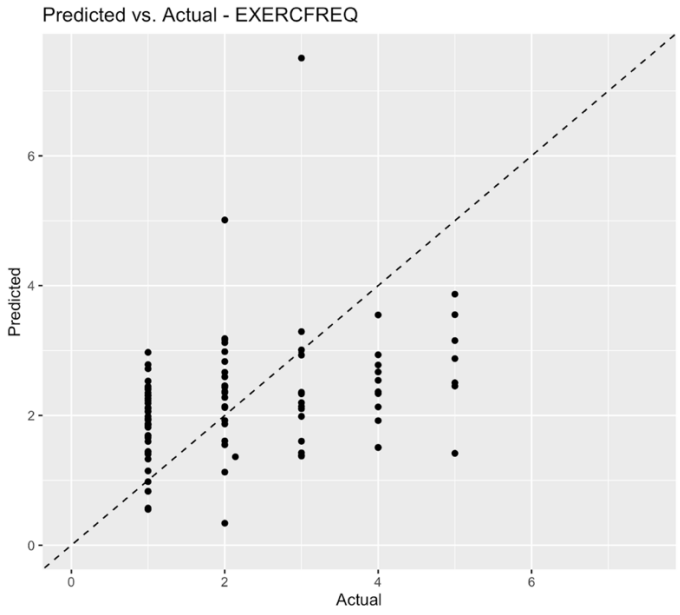

Figure 8: Performance of Methylation PCs vs EXERCFREQ model

As previously mentioned, the analysis for the model summaries remains the same for the methylation models. Figures 9 and 10 depict the methylation model summaries for the BMI and EXERCFREQ models. The highest magnitude β value for the methylation BMI model is attached to MPC4 and has one of the lowest p-values associated with it. For the EXERCFREQ model, the highest magnitude β value is attached to MPC3 and has one of the lowest p-values.

```
Mlm_BMI Summary:


Call:
lm(formula = formula, data = m_train_set)

Residuals:
    Min      1Q  Median      3Q     Max
-10.163  -4.625   0.066   3.999  13.508

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.636444   0.600317  47.702  < 2e-16 ***
MPC1        -0.007984   0.078890  -0.101 0.919627
MPC2        -0.070947   0.115082  -0.616 0.539235
MPC3        -0.488557   0.160848  -3.037 0.003179 **
MPC4        -0.722700   0.178377  -4.052 0.000113 ***
MPC5         0.098270   0.180987   0.543 0.588591
MPC6        -0.185813   0.199650  -0.931 0.354678
MPC7        -0.198344   0.223108  -0.889 0.376538
MPC8         0.036230   0.227447   0.159 0.873821
MPC9         0.567550   0.241754   2.348 0.021245 *
MPC10       -0.443855   0.242625  -1.829 0.070889 .
---
Signif. codes:  0 ‚Äò***‚Äô 0.001 ‚Äò**‚Äô 0.01 ‚Äò*‚Äô 0.05
```

Figure 9: BMI model summary. The highlighted MPC4 is the highest magnitude PC in this model

```
Mlm_EXERCFREQ Summary:


Call:
lm(formula = formula, data = m_train_set)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8586 -0.8411 -0.3350  0.8011  3.2851

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.180888   0.119746  18.213  < 2e-16 ***
MPC1        -0.008051   0.015736  -0.512 0.610254
MPC2        -0.030518   0.022956  -1.329 0.187310
MPC3        -0.115353   0.032085  -3.595 0.000546 ***
MPC4        -0.102035   0.035581  -2.868 0.005227 **
MPC5        -0.026066   0.036102  -0.722 0.472291
MPC6        -0.015169   0.039825  -0.381 0.704252
MPC7        -0.092138   0.044504  -2.070 0.041490 *
MPC8         0.022895   0.045369   0.505 0.615131
MPC9         0.052667   0.048223   1.092 0.277884
MPC10       -0.074172   0.048397  -1.533 0.129138
---
Signif. codes:  0 ‚Äò***‚Äô 0.001 ‚Äò**‚Äô 0.01 ‚Äò*‚Äô 0.05
```

Figure 10: ESERCFREQ model summary. The highlighted MPC3 is the highest magnitude PC in this model

3.3. Principal Component-Trait Correlation Heatmaps and P-value Grids

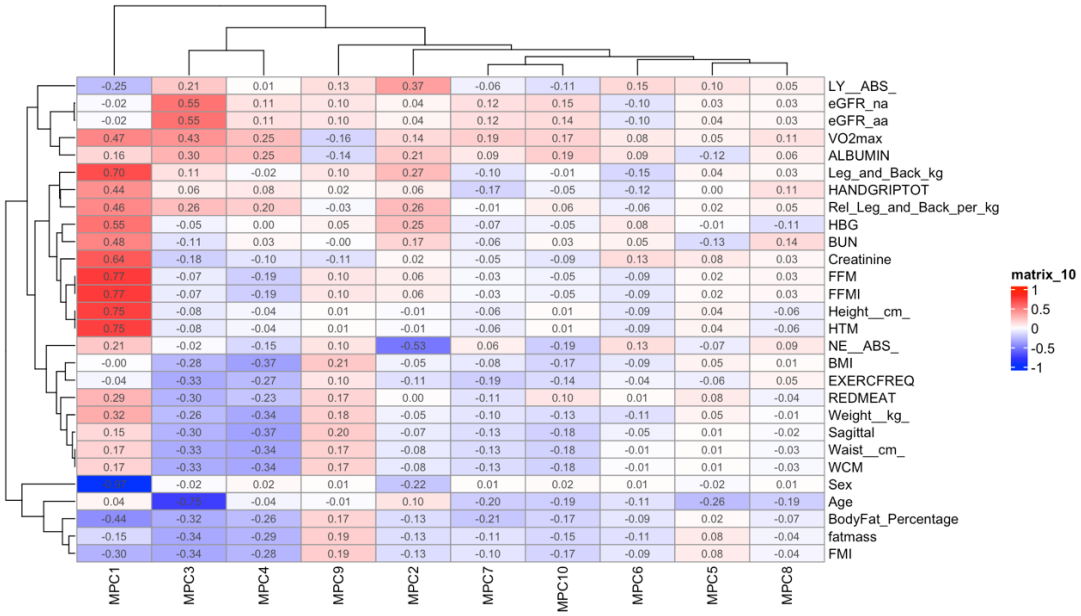

Figure 11: Correlation coefficient (r-value) heatmap of methylation PCs and traits with hierarchical clustering

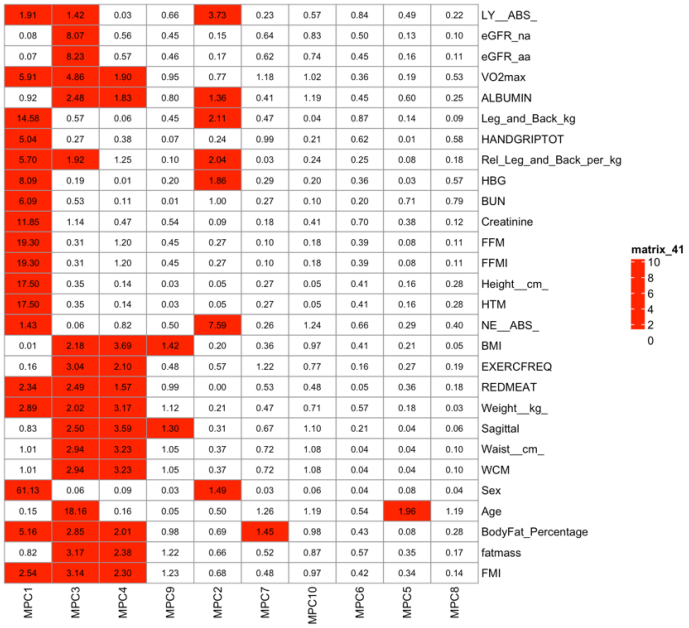

Figure 12: P-value of correlation coefficients grid to evaluate the significance of Figure 11. Red indicates significance, and cell values are -log10(p-value)

Analysis of Figure 11 indicates that MPC1, MPC3, and MP4, as suspected from the coefficient analysis, are the primary drivers across most of the proteomic biomarkers. MPC1 has significant correlations across several traits like FFM, Height, hemoglobin (HBG), etc. MPC3 and MPC4 appear closely related and are inversely associated with traits such as BMI, EXERCFREQ, and Weight. There seems to be some natural grouping of traits and PCs as a result of the hierarchical clustering. Among the PCs, MPC1, MPC3, and MPC4 are closely related and exert the most influence across the models. For the traits, there is a positively correlated group from Leg/Back (kg) to Neutrophils (NE_ABS) and a negatively correlated group from BMI to Waist Circumference in meters (WCM). The p-value grid assures these PCs and groupings are significant and worth further analysis. As such, MPC1, MPC3, and MPC4 were selected as candidates for deeper analysis in Cistrome.

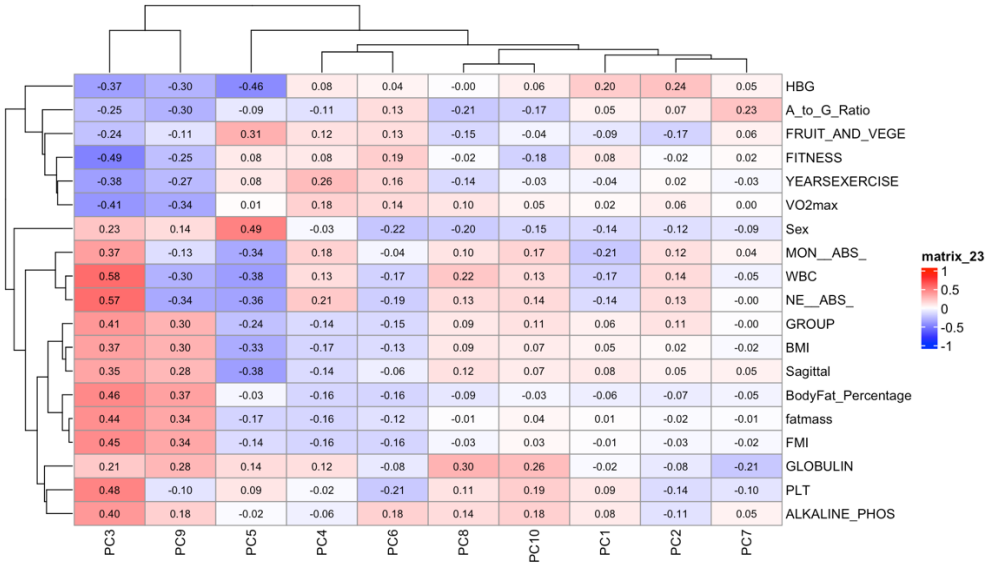

Figure 13: Correlation coefficient (r-value) heatmap of proteomic PCs and traits with hierarchical clustering

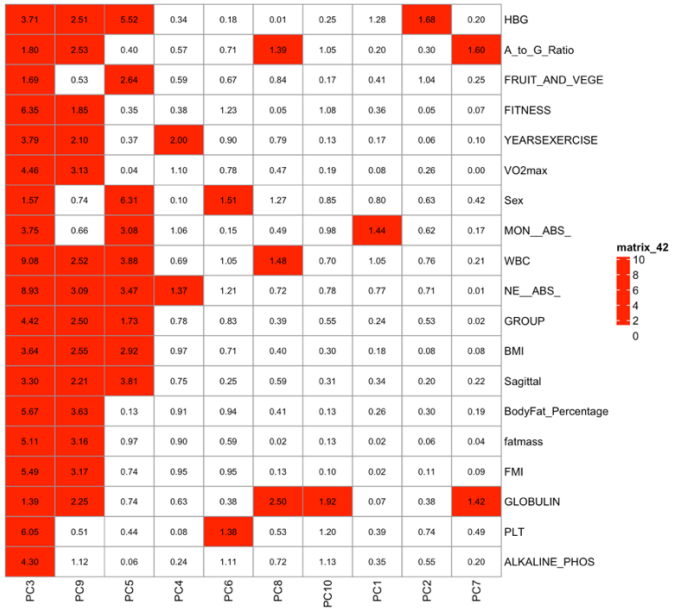

Figure 14: P-value of correlation coefficients grid to evaluate the significance of Figure 13.

Red indicates significance, and cell values are -log10(p-value)

For the proteomic PCs, PC3, PC9, and PC5 appear to be the most influential and significant PCs across most models. In particular, PC3 has significant correlations with every model. As with the methylation PCs, the hierarchical clustering revealed that the top 3 proteomic PCs were closely related. There are some notable groupings among the traits as well, namely a cluster from HBG to VO2max that is negatively correlated with PC3 and PC9, a cluster from monocytes (MON_ABS) to NE_ABS that is negatively correlated with PC5, and a cluster from GROUP to FMI that is positively correlated with PC3. All these groups and PCs were again verified by the p-value grid, resulting in PC3, PC9, and PC5 being the PCs of interest to analyze through STRING.

3.4. Cistrome Results

PC	Transcription Factor
MPC1	Positive: JARID2, EZH2, REST Negative: POLR2A, GABPA, BRPF3, ELK1
MPC3	Positive: FOSL2, REPIN1, CLOCK Negative: RYBP, TRIM24, EZH2
MPC4	Positive: H2AZ, SMARCA4, POLR2A Negative: EZH2, EP300

Table 1: Methylation PCs and the transcription factors associated with the most influential negatively/positively associated with the PC

After analyzing the GIGGLE plots produced by Cistrome, we compiled lists of transcription factors that likely bind to the important CpG sites embedded in the PC loadings. An interesting feature to note is that EZH2 appears in every PC. We cross-referenced these factors with NCBI and various medical journals to discern their functions, with the goal of uncovering some biological basis for the associations we observed.

3.5. STRING Results

The results from STRING include a table of biological processes associated with the networks developed by STRING and a protein network graph for each PC of interest. Unlike the methylation PCs, the negatively associated proteins of the proteomic PCs did not yield useful information across all three PCs of interest. Many biological processes predicted to be linked to

those networks were non-specific, relating to cell function, regulation, or metabolism. Additionally, the FDR values associated were significantly higher than the positive counterparts, indicating weaker results. As such, we decided to focus our efforts solely on the positively associated proteins.

3.5.1. PC3

#color	category	term ID	term description	observed gene count	background gene count	strength	false discovery rate
black	GO Process	GO:0006959	Humoral immune response	19	268	1.16	6.34E-13
purple	GO Process	GO:0007596	Blood coagulation	16	173	1.27	2.65E-12
blue	GO Process	GO:0070527	Platelet aggregation	11	43	1.72	3.06E-12
magenta	GO Process	GO:0030168	Platelet activation	13	97	1.43	8.22E-12
darkgreen	GO Process	GO:0006956	Complement activation	8	60	1.43	5.25E-07
limegreen	GO Process	GO:0072378	Blood coagulation, fibrin clot formation	6	24	1.71	1.86E-06
orange	GO Process	GO:0002526	Acute inflammatory response	8	80	1.31	3.23E-06
maroon	GO Process	GO:1990266	Neutrophil migration	8	92	1.25	8.06E-06
cyan	GO Process	GO:1900024	Regulation of substrate adhesion-dependent cell spreading	7	61	1.37	1.03E-05
red	GO Process	GO:0006957	Complement activation, alternative pathway	5	16	1.8	1.19E-05
yellow	GO Process	GO:0006953	Acute-phase response	6	42	1.46	2.47E-05
grey	GO Process	GO:0030593	Neutrophil chemotaxis	7	80	1.25	4.32E-05

Table 2: Biological processes associated with the positive proteins in PC3.

The colors correspond to the protein network.

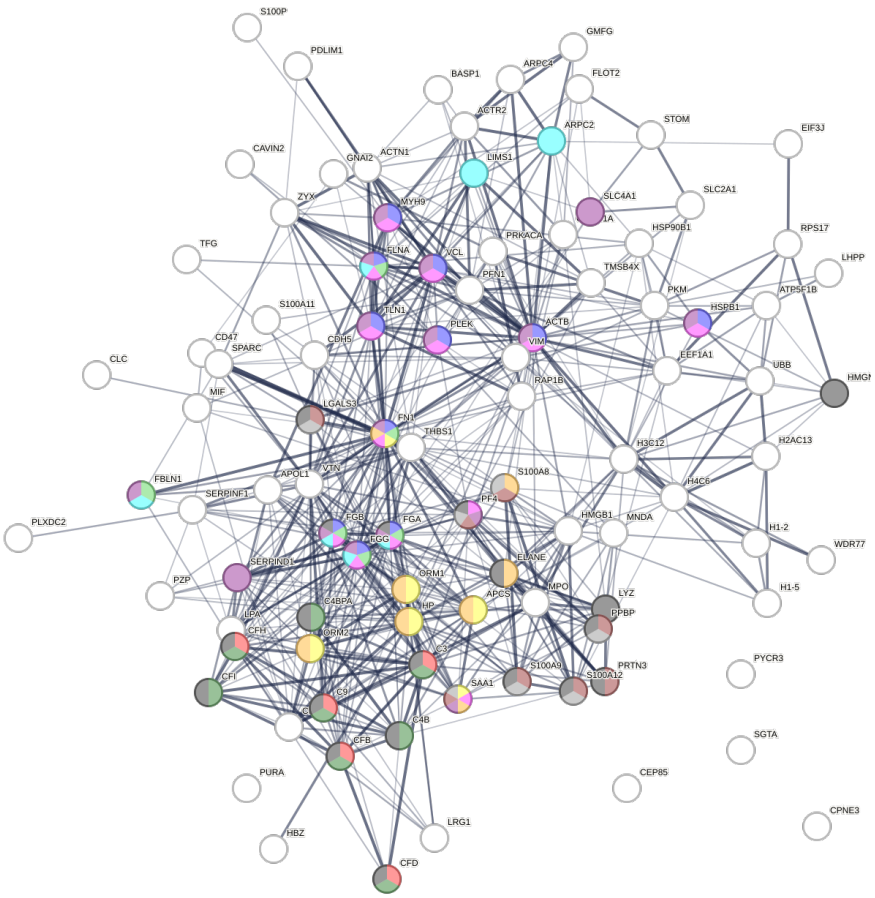

Figure 15: Protein network of PC3 proteins

On the STRING website, we selected the results with the smallest FDR values and exported the resulting table. According to these results, most of the proteins seem to be involved in coagulation and platelet activation, as well as the inflammatory immune response in the form of complement activation, neutrophil migration, and acute-phase response.

3.5.2. PC5

#color	category	term ID	term description	observed gene count	background gene count	strength	false discovery rate
limegreen	GO Process	GO:0010466	Negative regulation of peptidase activity	13	249	1.11	6.45E-07
magenta	GO Process	GO:0045861	Negative regulation of proteolysis	14	339	1.01	1.05E-06
blue	GO Process	GO:0006956	Complement activation	8	60	1.52	1.59E-06
darkgreen	GO Process	GO:0052547	Regulation of peptidase activity	15	446	0.92	1.69E-06
cyan	GO Process	GO:0030162	Regulation of proteolysis	18	739	0.78	2.55E-06
orange	GO Process	GO:0051248	Negative regulation of protein metabolic process	21	1038	0.7	2.55E-06
yellow	GO Process	GO:0010951	Negative regulation of endopeptidase activity	11	240	1.05	1.05E-05
maroon	GO Process	GO:0006950	Response to stress	36	3358	0.42	1.07E-05
purple	GO Process	GO:0044092	Negative regulation of molecular function	20	1143	0.63	3.91E-05
red	GO Process	GO:0006958	Complement activation, classical pathway	6	40	1.57	4.25E-05

Table 3: Biological processes associated with the positive proteins in PC5. The colors correspond to the protein network

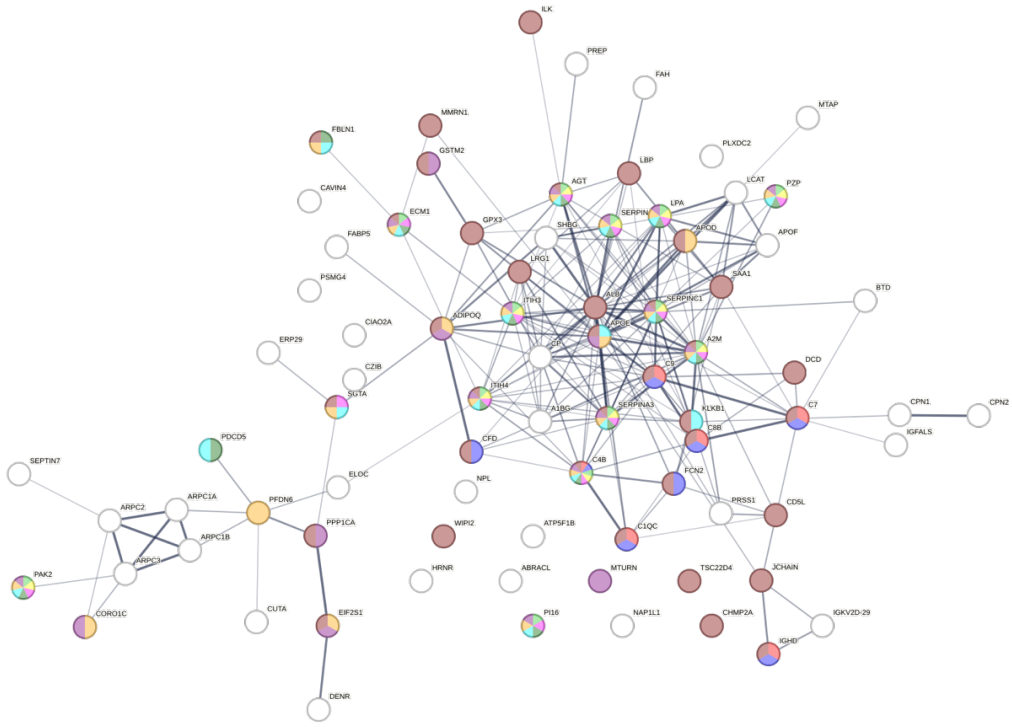

Figure 16: Protein network of PC5 proteins

The PC5 proteins seem to be most involved with protein metabolism (regulation of peptidase and endopeptidase), specifically the negative regulation of protein metabolism. There are also some mentions of complement activation, but the vast majority of the results relate to protein metabolism.

3.5.3. *PC9*

#color	category	term ID	term description	observed gene count	background gene count	strength	false discovery rate
yellow	GO Process	GO:0006956	Complement activation	17	60	1.78	3.48E-20
black	GO Process	GO:0006959	Humoral immune response	19	268	1.18	2.21E-13
lightgreen	GO Process	GO:0010466	Negative regulation of peptidase activity	16	249	1.13	1.47E-10
orange	GO Process	GO:0002526	Acute inflammatory response	11	80	1.46	5.63E-10
pink	GO Process	GO:0010951	Negative regulation of endopeptidase activity	15	240	1.12	8.5E-10
darkgreen	GO Process	GO:0006953	Acute-phase response	9	42	1.66	1.45E-09
red	GO Process	GO:0006957	Complement activation, alternative pathway	7	16	1.97	6.06E-09
blue	GO Process	GO:1903027	Regulation of opsonization	6	18	1.85	5.01E-07
lightblue	GO Process	GO:0007596	Blood coagulation	10	173	1.09	5.42E-06
maroon	GO Process	GO:0031638	Zymogen activation	7	59	1.4	8.17E-06
limegreen	GO Process	GO:1903028	Positive regulation of opsonization	5	16	1.82	1.24E-05
magenta	GO Process	GO:0051917	Regulation of fibrinolysis	5	18	1.77	1.91E-05
grey	GO Process	GO:0061045	Negative regulation of wound healing	7	69	1.33	1.96E-05
purple	GO Process	GO:0030195	Negative regulation of blood coagulation	6	46	1.44	4.23E-05
cyan	GO Process	GO:0072378	Blood coagulation, fibrin clot formation	5	24	1.64	5.65E-05

Table 4: Biological processes associated with the positive proteins in PC9. The colors correspond to the protein network

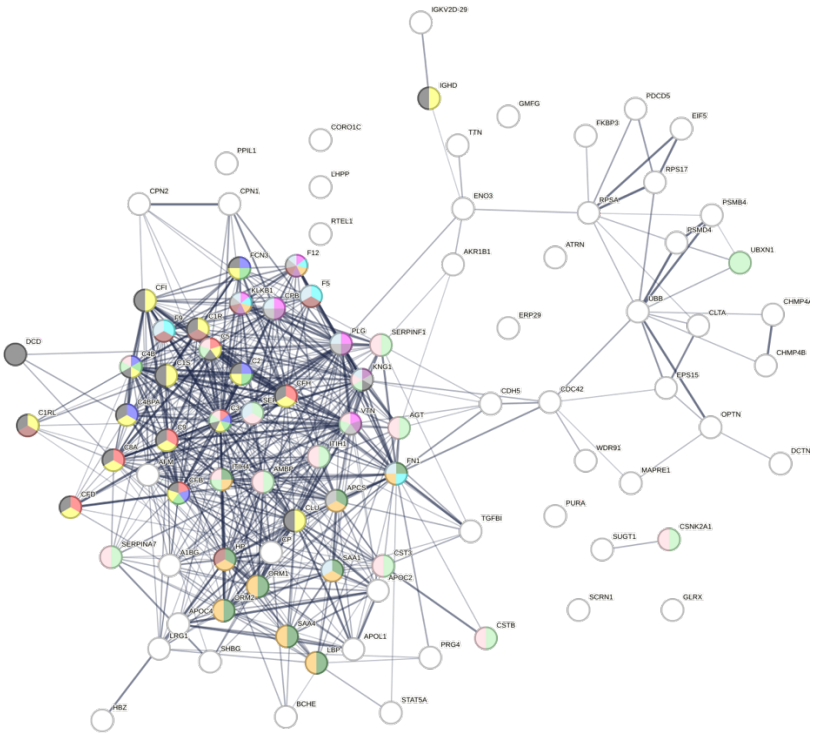

Figure 17: Protein network of PC9 proteins

The PC9 proteins appear to be involved in coagulation/platelet activation and inflammatory immune response. They also have some links to protein metabolism. This somewhat mirrors the results from the previous two PCs and provides more evidence that these three processes may be

responsible for some of the differences we observed between healthy and unhealthy people. The following section will explore analyzing these processes and their relation to health.

4. DISCUSSION & FUTURE STEPS

Through this study, we managed to develop 19 significant proteomic biomarkers and 28 significant methylation biomarkers, both of which predict specific clinical lifestyle traits. Across both sets of biomarkers, 15 biomarkers pertain directly to our original selection criteria. This may suggest significant differences in methylation percentages and protein abundances between “healthy” and “unhealthy” individuals. The scatterplots shown previously and the remaining scatterplots we developed show clustering along the diagonal, indicating our models' good performance. Additionally, we have sets of PCs from all our valid models, both proteomic and methylation, that were strongly correlated with and highly significant to our models and the lifestyle traits. To understand the biological underpinnings of our models and provide more evidence of the validity of our biomarkers, we opted to study the loadings of these PCs. The proteomic PCs we chose to investigate more closely were PC3, PC5, and PC9, and the methylation PCs we decided to examine were MPC1, MPC3, and MPC4. From there, we extracted the necessary protein and CpG site information from the loading matrices of these PCs of interest to run through STRING and Cistrome, respectively. This provided lists of transcription factors that bind to influential CpG sites and lists of GO terms denoting biological networks/processes in which our proteins are involved. Analyzing these results was a crucial step in the broader context of characterizing the downstream effects of an unhealthy lifestyle.

4.1. Protein Network Analysis

As stated, the major biological processes contributing to differences between healthy and unhealthy populations were blood coagulation and platelet activation, protein metabolism regulation, and inflammatory immune response. PC3 and PC9 proteins appear to be more specifically related to coagulation, platelet activities, and immune response, whereas PC5 proteins were more related to protein metabolism. Looking more closely at PC3 of Figure 13, we see that it is strongly positively correlated with many traits relating to fat content and overall health, such as fat mass, FMI, body fat percentage, BMI, etc. It also positively correlates with platelet (PLT) concentration, which is bolstered by downstream GO term analysis. Since these traits are positively correlated, increases in PC values and their positively associated proteins result in increases in the trait. These positively associated proteins relate to platelet activation and coagulation, so we can deduce that these biological processes positively correlate with the traits listed above. This seems to compute with some existing studies regarding the connection between obesity and elevated platelet counts, although studies disagree on whether platelet activation is actually involved in obesity (Vauclard et al., 2023). Platelet activation markers such as PMP that increase with obesity and heightened expression levels of GPV1 (which is involved in platelet degranulation) are evidence of a positive correlation (Vauclard et al., 2023). It is also well known that obese individuals have an increased risk of thrombosis due to platelet activation (Vauclard et al., 2023). The results of our study seem to support the theory of a positive association between platelet activation and our obesity-related traits (fat mass, BMI, etc.). Additionally, hemoglobin, VO2 max, exercise frequency, and fruit and vegetable intake are inversely related to PC1 and platelet activation. This makes intuitive sense since unhealthy people are less likely to exercise and typically have less balanced diets. On the molecular level,

regarding hemoglobin (and VO2 max since both are related to oxygen capacity), studies have found that lower hemoglobin concentrations promote platelet aggregation by the increased phosphorylation of signaling adapter proteins (Singhal et al. 2015).

Despite the intuitive nature of the previous analysis, the role of the inflammatory immune system and complement system activation is less clear. PC3 proteins share a strong positive association with monocytes (MON_ABS), neutrophils (NE_ABS), and white blood cells (WBC), which is bolstered by the results from STRING that show our proteins operating in networks such as neutrophil migration, acute-phase inflammation, and complement activation as seen in Figure 15. This agrees with the existing consensus about the heightened levels of inflammation that exist in obese populations. Studies have implicated complement activation as a critical marker for obesity, detailing the activation of the alternative complement pathway by obesity-induced adipose tissue, thereby creating factors that induce a pro-inflammatory state (Shim et al. 2020). The analysis becomes unclear when examining the PC5 proteins. Figure 17 shows that many of these proteins are involved in regulating protein metabolism, specifically the negative regulation of proteolysis. This could explain the research detailing the lack of muscle quality of people with obesity when compared to their healthy counterparts and lend credence to the notion that obesity can sometimes be accompanied by lower rates of protein synthesis (Freitas and Katsanos, 2022). Since the PC5 proteins seem to promote negative regulation of proteolytic processes, increases in the PC5 proteins would result in less circulating protein being broken down and absorbed from the bloodstream.

Consequently, fewer amino acids may be available to capture and reorganize for muscle synthesis. However, under this notion that the negative regulation of proteolytic processes is positively correlated with obesity, the inverse relationship between PC5 and the immune cells

(monocytes, white blood cells, and neutrophils) would contradict much of the existing understanding of the role of the immune system in obese individuals. Despite this, some evidence exists wherein neutrophil chemotaxis seems to decrease as the concentration of a pro-inflammatory (complement system activating) peptide, C5a, increases. Since PC3 appears to capture the apparent positive interaction between immune cells and obesity, PC5 might capture the less likely yet significant interaction described above. However, evidence for this interaction is sparse at best and should only be considered within the context of the broader consensus. Further research may be required to uncover the true basis of some of these interactions.

4.2. Transcription Factor Analysis

When observing the GIGGLE plots of both the positively and negatively associated CpG sites of our MPCs, we found that, for MPC1 and MPC3, the plots of their positively associated sites had significantly lower top scores. For instance, MPC3's positive sites had a maximum GIGGLE score of 200, whereas the negative sites had a maximum score closer to 2000. Only MPC4 appeared to have high scores in both sets of sites. These scores indicate the similarity of our sites to existing ChIP-seq data within the Cistrome database, so prioritizing scores with high similarity was crucial. For the purposes of this paper, the transcription factors were examined under the traditional framework, which purports that increased methylation alters the shape of DNA molecules, thereby inhibiting the binding of transcription factors.

Of the transcription factors detailed in Table 1, we decided to more closely examine ELK1, EZH2, TRIM24, and RYBP due to the strength of their GIGGLE scores and their relevance to the purpose of our study. Most of the analysis focused on the MPC3 and MPC4

transcription factors since most of MP1's transcription factors were uninformative. MPC1 does, however, contain many significant sites from the X chromosome. As such, sex-based differences are well captured by MPC1. Looking at Figure 11 and examining the loadings of MPC1, we noticed that many traits, such as creatine, blood urea nitrogen (BUN), FFM, grip strength, and height, were positively correlated with MPC1. These are all traits where we observe statistically significant discrepancies between men and women.

Additionally, the loadings revealed that the negatively associated sites were located mainly on the X chromosome and had weights with much higher absolute values across the top 200 selected compared to the positively associated counterparts. This implies the negatively associated sites were far more influential to MPC1. Since the most influential sites of MPC1 were negatively associated, increases in the methylation of these sites corresponded to decreases in the PC value, which would translate to decreases in the traits listed above. Since the increased methylation occurs on the X chromosome, MPC1 likely captures the X chromosome inactivation (XCI) in women. XCI involves the large-scale methylation of CpG islands on the X-chromosome slated for inactivation in women (Sharp et al. 2011). Thus, MPC1 may be able to predict the values of these traits based on the level of methylation recorded from the negatively associated CpG sites. As further proof of MPC1's ability to model sex-dependent traits, Sex is almost perfectly inversely correlated with MPC1, with an r-value of -0.97. Since females were denoted as 2 and males as 1 in the original study, increased methylation of influential X chromosome sites decreases the PC value, translating to an increase in the Sex "value" (i.e., closer to 2). As is evident, MPC1 efficiently captured the biological differences between men and women regarding muscle mass, strength, and size.

However, MPC1 was less efficient in modeling differences based on health, as is evident from the lower r-values that resulted from correlating MPC1 with traits like fat mass, BMI, and sagittal. MPC1 showed significant negative correlations with body fat percentage and FMI, which are relevant to the study. However, a literature dive into the MPC1's transcription factors yielded few insights into the mechanisms at play. POLR2A is non-specific in its function since it is a subunit of RNA polymerase, a protein necessary for all cell transcription. GABPA is more closely linked to mitochondrial function, and BRFP3's function is not well established. ELK1 is the only transcription factor that plays a role in adipogenesis, wherein adipocyte differentiation is heavily suppressed when ELK1 is inhibited (Pang et al. 2016). Methylation of ELK1's binding sites produces a similar inhibitory effect. Since ELK1 is negatively associated with MPC1, increased methylation at those sites may inhibit ELK1's activity. However, lower ELK1 activity relates to a decrease in adipogenesis, the opposite interaction we observe in the heatmap. Although the interaction may not agree with biological intuition, ELK1's relation to adipogenesis may still be significant due to the connection between heightened adipogenesis and obesity.

MPC3 and MPC4 behave similarly in Figure 11, negatively correlated with several key traits such as BMI, body fat percentage, fat mass, etc. Additionally, most of the transcription factors that will be examined here bind to sites that are negatively associated with either MPC3 or MPC4, indicating an increase in these key traits. Like ELK1, EZH2, which is related to MPC3 and MPC4, has been shown to play a role in adipogenesis and obesity. Although EZH2 typically relates to cell differentiation and oncogenesis, there is evidence that the inhibition of EZH2 leads to lipid accumulation in certain cancer cell lines (Yiew et al. 2019). However, there is more evidence to suggest that the inhibition of EZH2 activity leads to healthier outcomes, implying

that EZH2 plays a crucial role in adipogenesis (Wang, 2022). This contradiction may be explained by the fact that EZH2 is involved in numerous complex interactions and that the methylated CpG sites we are examining may not relate to genes involved in adipogenesis. Instead, the methylation of these sites (decrease in EZH2 binding at these sites) could relate to the silencing of tumor suppressant genes, in which EZH2 is more commonly implicated.

Further research to determine EZH2's role will be necessary if and how it relates to adipogenesis in this study. TRIM24 is another notable transcription factor that may play a role in lipid metabolism. It directly and indirectly represses hepatic lipid accumulation (Jiang et al. 2014). Thus, methylation of TRIM24's binding sites may promote lipid accumulation, a key mechanism in obesity. It is important to note these results were acquired from experiments on mice, so the effect is yet to be observed in humans. RYBP was the final transcription factor we analyzed, and it is involved in the negative regulation of a particular protein catabolic process. The function of this transcription factor seems to fit with our protein network findings regarding the negative regulation of protein catabolic activity. Like EZH2, RYBP has also been reported to affect tumor cells, wherein it has been shown to inhibit glycolysis in tumor cells and repress tumor migration. Thus, increased methylation at its binding sites may limit its tumor-suppressant activity.

Although the discovery of these transcription factors has helped inform our study, it is important to note that EZH2, RYBP, and ELK1 exhibited interactions that ultimately linked increased methylation of their binding sites to increases in obesity-related traits. However, studies have shown that inhibiting the activity of these transcription factors decreases obesity-related effects. These inconsistencies might be explained by exploring specific CpG sites and identifying which genes they might regulate. From here, we can predict binding activity more

clearly and truly uncover how these factors might increase or decrease traits relating to poor health.

4.3. Caveats and Future Steps

Through this study, we developed proteomic and epigenetic biomarkers for lifestyle traits to elucidate the biological mechanisms behind poor health and obesity. In-depth analyses of these models have provided some biological validation for the predictive power of our biomarkers. There are some caveats to these discoveries, however. In the protein network analyses, PC5, which was heavily involved in protein metabolism, was inversely related to white blood cells, monocytes, and neutrophils. During this analysis, we operated under the notion that lower protein metabolism corresponds to worse health; however, this assumption could be incorrect. Another caveat to consider would be the vast number of transcription factors not included in the analysis. Although using GIGGLE scores to prioritize transcription factors of note is efficient, it is not a perfect method for determining the significance of a transcription factor to a particular trait. Thus, a more comprehensive study of these factors could aid in understanding the mechanisms at play. Additionally, we could examine the CpG sites and determine the genes they regulate. This would provide a clearer picture of the binding behavior of our transcription factors by narrowing our search to factors that bind only to the genes we have uncovered. Including target gene information would eliminate much of the speculation behind the effects of methylating the target genes of our transcription factors.

Despite these caveats, this study presents detailed documentation of the development and biological validation of 19 proteomic and 28 methylation biomarkers. We found evidence linking

poor health to platelet activation, inflammatory immune response, protein metabolism, tumor progression, adipogenesis, and more through deep analysis of protein networks and transcription factors. This provides researchers with numerous avenues to pursue, such as testing therapeutic drugs that interact with the factors/proteins we examined or building networks to model real biological systems.

BIBLIOGRAPHY

1. Corlin, Laura, et al. “Proteomic Signatures of Lifestyle Risk Factors for Cardiovascular Disease: A Cross-Sectional Analysis of the Plasma Proteome in the Framingham Heart Study.” *Journal of the American Heart Association*, 5 Jan. 2021, www.ncbi.nlm.nih.gov/pmc/articles/PMC7955453/.

2. Brownlee, Jason. “LOOCV for Evaluating Machine Learning Algorithms.” *MachineLearningMastery.Com*, 26 Aug. 2020, machinelearningmastery.com/loocv-for-evaluating-machine-learning-algorithms/.

3. Gardner, Miranda L, and Michael A Freitas. “Multiple Imputation Approaches Applied to the Missing Value Problem in Bottom-up Proteomics.” *International Journal of Molecular Sciences*, 6 Sept. 2021, www.ncbi.nlm.nih.gov/pmc/articles/PMC8431783/.

4. Kumar, Ajitesh. “PCA Explained Variance Concepts with Python Example.” *Data Analytics*, 14 Apr. 2023, vitalflux.com/pca-explained-variance-concept-python-example/#:~:text=Related%20posts%3A-,What%20is%20Explained%20Variance%3F,which%20are%20called%20principal%20components.

5. Pramoditha, Rukshan. “How to Mitigate Overfitting with Dimensionality Reduction.” *Medium*, 5 Oct. 2021, towardsdatascience.com/how-to-mitigate-overfitting-with-dimensionality-reduction-555b755b3d66.

6. Schmidt, Andreas, et al. “Bioinformatic Analysis of Proteomics Data.” *BMC Systems Biology*, 2014, www.ncbi.nlm.nih.gov/pmc/articles/PMC4108846/.

7. Tredennick, Andrew T, et al. “A Practical Guide to Selecting Models for Exploration, Inference, and Prediction in Ecology.” *Ecology*, June 2021, www.ncbi.nlm.nih.gov/pmc/articles/PMC8187274/.

8. Waldmann, Patrik. “On the Use of the Pearson Correlation Coefficient for Model Evaluation in Genome-Wide Prediction.” *Frontiers in Genetics*, 26 Sept. 2019, www.ncbi.nlm.nih.gov/pmc/articles/PMC6781837/.

9. Jiang, Shiming, et al. “Trim24 Suppresses Development of Spontaneous Hepatic Lipid Accumulation and Hepatocellular Carcinoma in Mice.” *Journal of Hepatology*, U.S. National Library of Medicine, Feb. 2015, www.ncbi.nlm.nih.gov/pmc/articles/PMC4772153/.

10. Yiew, Nicole K H, et al. “Enhancer of Zeste Homolog 2 (EZH2) Regulates Adipocyte Lipid Metabolism Independent of Adipogenic Differentiation: Role of Apolipoprotein E.” *The Journal of Biological Chemistry*, U.S. National Library of Medicine, 24 May 2019, www.ncbi.nlm.nih.gov/pmc/articles/PMC6544862/.

11. Wang, Haixia. “Role of EZH2 in Adipogenesis and Obesity: Current State of... : Medicine.” *LWW*, Zhejiang Changzheng Vocational and Technical College, 6 Apr. 2022, journals.lww.com/md-journal/fulltext/2022/09090/role_of_ezh2_in_adipogenesis_and_obesity__current.15.aspx#:~:text=Thus%2C%20EZH2%20can%20contribute%20to,activity%20of%20catalyzing%20DNA%20methylation.

12. Pang, Lingxia, et al. “Mir-1275 Inhibits Adipogenesis via ELK1 and Its Expression Decreases in Obese Subjects.” *Jme*, Bioscientifica Ltd, 1 July 2016, jme.bioscientifica.com/view/journals/jme/57/1/33.xml.

13. Sharp, Andrew J, et al. “DNA Methylation Profiles of Human Active and Inactive X Chromosomes.” *Genome Research*, U.S. National Library of Medicine, Oct. 2011, www.ncbi.nlm.nih.gov/pmc/articles/PMC3202277/.

14. Chen, et al. “Glucose-Induced RYBP Suppresses Tumor Cell Aerobic Glycolysis and Migration.” *Biochemical and Biophysical Research Communications*, Academic Press, 9 May 2024, www.sciencedirect.com/science/article/pii/S0006291X24006259.

15. Shim K, Begum R, Yang C, Wang H. Complement activation in obesity, insulin resistance, and type 2 diabetes mellitus. World J Diabetes. 2020 Jan 15;11(1):1-12. doi: 10.4239/wjd.v11.i1.1. PMID: 31938469; PMCID: PMC6927818.

16. Daniel Ricklin, John D. Lambris; Complement in Immune and Inflammatory Disorders: Pathophysiological Mechanisms. *J Immunol* 15 April 2013; 190 (8): 3831–3838. https://doi.org/10.4049/jimmunol.1203487

17. Freitas EDS, Katsanos CS. (Dys)regulation of Protein Metabolism in Skeletal Muscle of Humans With Obesity. Front Physiol. 2022 Mar 8;13:843087. doi: 10.3389/fphys.2022.843087. PMID: 35350688; PMCID: PMC8957804.

18. Muldur, Sinan, et al. “Human Neutrophils Respond to Complement Activation and Inhibition in Microfluidic Devices.” *Frontiers*, Frontiers, 29 Oct. 2021, www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2021.777932/full.

19. Rolling CC, Barrett TJ, Berger JS. Platelet-monocyte aggregates: molecular mediators of thromboinflammation. Front Cardiovasc Med. 2023 May 15;10:960398. doi: 10.3389/fcvm.2023.960398. PMID: 37255704; PMCID: PMC10225702.

20. Vauclard, Alicia, et al. "Obesity: Effects on Bone Marrow Homeostasis and Platelet Activation." *Thrombosis Research*, vol. 231, 2023, pp. 195-205, https://doi.org/10.1016/j.thromres.2022.10.008.

21. Singhal, R. et al. Hemoglobin interaction with GP1bα induces platelet activation and apoptosis: a novel mechanism associated with intravascular hemolysis. Haematologica. 2015 Dec;100(12):1526-33. doi: 10.3324/haematol.2015.132183. Epub 2015 Sep 4. PMID: 26341739; PMCID: PMC4666328.

22. Liu, T., Ortiz, J.A., Taing, L. *et al.* Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol* **12**, R83 (2011). https://doi.org/10.1186/gb-2011-12-8-r83

23. Szklarczyk D et al. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res. 2023 Jan 6;51(D1):D638-D646. doi: 10.1093/nar/gkac1000. PMID: 36370105; PMCID: PMC9825434.