보건학 석사 학위논문

Effect of host genome, microbiome and envirome on metabolic profile

대사증후군 위험요인에 대한 유전체, 마이크로바이옴 및 환경인자의 영향

2019년 6월

서울대학교 보건대학원

보건학과 보건학전공

이 윤 환

Abstract

Quantifying the associations of host genome, microbiome and envirome on metabolic profile

Yunhwan Lee

Department of Public Health

Graduate School of Public Health

Seoul National University

Background

Metabolic syndrome is a well known risk factor for cardiovascular disease. Therefore better understanding of each component of metabolic syndrome is required. Metabolic syndrome results from genetics, environmental factors, and the interaction between them. Many recent studies have shown that microbiome composition also affects the development of metabolic syndrome. In this study, the main goal is to identify and compare the associations of host genetic, metagenome, and environmental factors on metabolic syndrome components.

Methods

Each data source was prepared through quality control and imputation process considering characteristics of each data. The effect of each data source was evaluated using the heritability estimation approach and the prediction model separately.

Results

In heritability estimation, we found that 5 of the 11 phenotypes are significantly associated with metagenome-wide similarity. Metagenome source also provided a more accurate estimation than the genetics at the same sample size. In the prediction model, the contribution of each source to the prediction accuracy varied for each phenotype.

Conclusion

Through various methods, we grasped the influence of host genetic, metagenome and environmental factors on each metabolic component and quantified the relative importance of each data source. If the sample size is increased and methods are developed considering the characteristics of each data source in a further study, the effect of each data will be confirmed more accurately.

Keywords: Host genetics, microbiome, environmental factors, metabolic syndrome, heritability, prediction model

***Student Number*:** 2017-28211

CONTENTS

TABLES

FIGURES

INTRODUCTION

Metabolic syndrome (MetS) is defined as a combination of clinical conditions including abdominal obesity, high blood pressure, elevated fasting blood glucose, high triglyceride, and low concentration of HDL cholesterol. Having three or more of these risk factors will result in a diagnosis of metabolic syndrome. Patients with MetS are at 2 to 4-fold increased risk of stroke, a 3 to 4-fold increased risk of myocardial infarction (MI), and 2-fold risk of all-cause mortality compared to those without MetS regardless of a previous history of cardiovascular events [1]. It has constantly been a global health concern therefore proper understanding and management of MetS is essential.

The components of MetS are well-known to be a result of complex interactions of genetic and environmental factors. Conventional genome-wide association studies (GWAS) have revealed that multiple genomic loci with small effects contribute to the development of metabolic risk factors. A number of environmental modifiers such as caloric intake and physical activity interact with genetic risk factors [2].

Recent studies have revealed that gut microbiota impacts host metabolism and implements an essential role in the etiology of metabolic disease such as obesity, insulin resistance, type 2 diabetes, and cardiovascular disease [3]. Notably, Wang et al. have identified and validated approximately 60,000 type 2 diabetes associated markers and constructed taxonomic species-level analyses [4].

As discussed above, genome-wide association studies and metagenome-wide association studies highlighted that the development of many complex diseases including MetS can be resulted from host genetics, microbiome, the environment, and their interactions. However, the relative effect of host genetics and gut microbiome on MetS is not clear.

In this paper, we will identify the effects of host genetics, metagenome and environmental factors on MetS in two distinct approaches.

The first approach is to estimate variance for metabolic syndrome traits explained by all SNPs or all genus level. This analysis utilizes the GCTA tool [5], which was developed to measure the heritability of quantitative traits of the conceptually unrelated individuals using a relationship matrix representing the genetic similarity of individuals. We will similarly construct a metagenomic relationship matrix using the relative abundance of the genus level and estimate the proportion of phenotypic variation explained by metagenomic similarity between individuals.

The second approach is to evaluate the prediction performance of each of the host genetic and metagenome, and to see how the performance improves when combining them. The prediction accuracy was evaluated by the coefficient of determination (R^2).

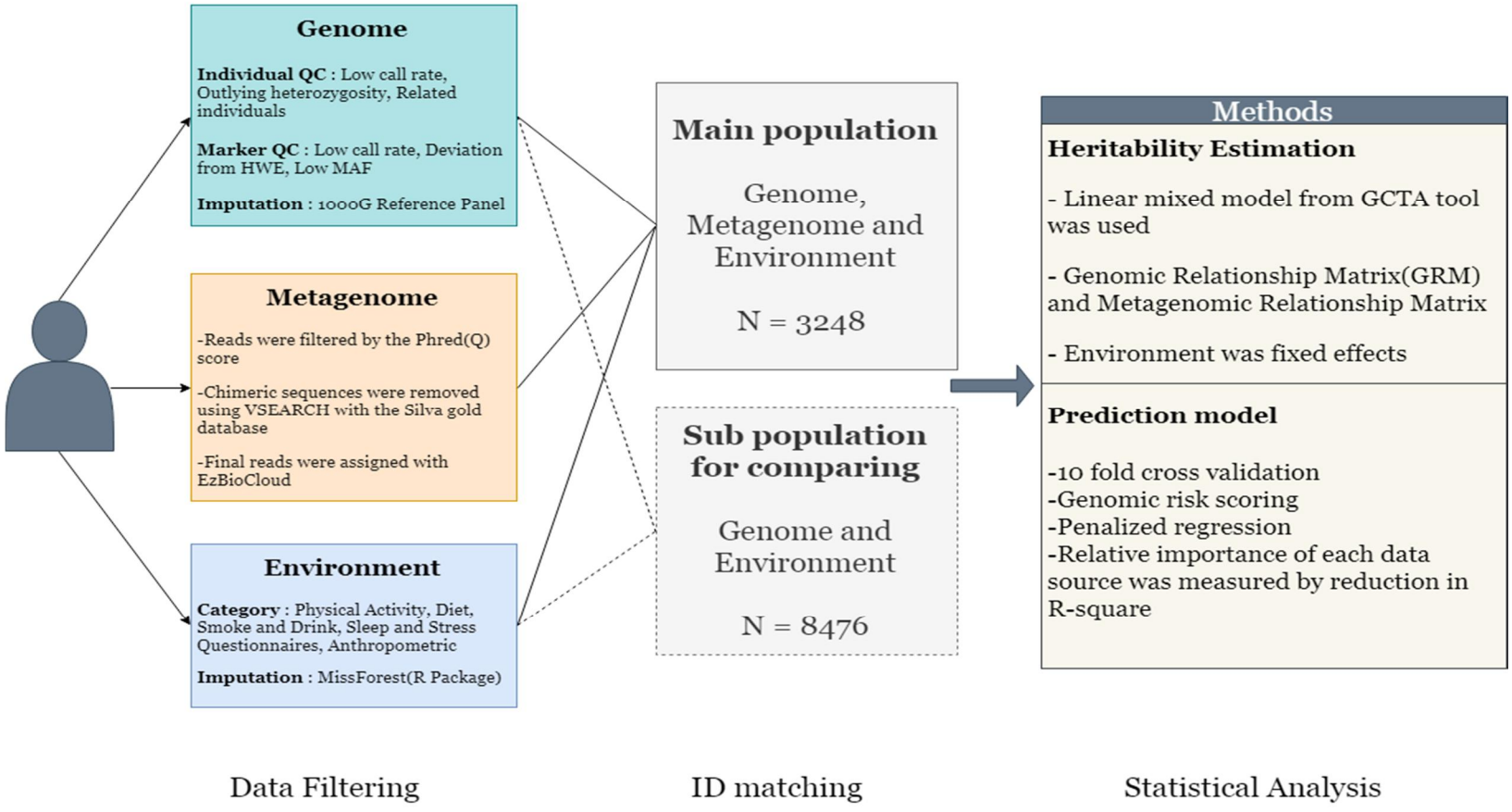

Figure 1. Study workflow

METHODS

Data description

Genome

This study was based on KARE cohort data. Initially, 10,004 individuals were genotyped for 500,568 SNPs with the Affymetrix Genome-Wide Human SNP array 5.0. For individual QC, we excluded individuals with low sex inconsistency, low call rate (call rate < 97%), outlying heterozygosity (heterozygosity rate > mean $\pm$ 3SD) and related individuals (IBS > 0.9). Then, we filtered SNPs with p-values for Hardy-Weinberg equilibrium (HWE) less than 10^{-6}, with genotype call rates less than 95% or minor allele frequencies (MAF) less than 0.01, and 352,228 SNPs were left. The quality controlled data were imputed with Impute2 [6] using 1000 Genomes data as a reference panel (total number of imputed SNPs were 3,351,033).

Metagenome

Metagenome data were produced by isolating urine-based extracelluar vesicle (EV) from 3844 samples of the Ansan cohort. Metagenomic sequencing was performed using 16S ribosomal RNA gene. The trimmed sequence pair was merged using CASPER, and the merged reads were filtered by the Phred (Q) score as described by Bokulich. After filtering, only reads with length between 350 bp and 550 bp were used, because the reads that do not satisfy this criterion are either errors or artifacts. To

identify chimeric sequences introduced by PCR or amplification, a reference-based chimera searching method was conducted using VSEARCH with the silva gold database. Next, the remaining reads were clustered into operational taxonomic units (OTUs) by open reference methods with EzBioCloud. The representative sequences in each OTU cluster were finally assigned taxonomy with UCLUST, along with the three databases(parallel_assign_taxonomy_uclust.py script on QIIME version 1.9.1) under default parameters. Relative abundances of 77 genus were used as the main variable. Alpha and beta diversity calculations were also conducted in QIIME version 1.9.1.

Environment

There are many environmental factors available including physical activity, dietary habits, stress/sleep questionnaires, smoking and sleep. However, data such as stress/sleep questionnaire and medication were removed from the analysis since the quality of these data is not guaranteed.

Specifically, previous researches have showed that the percentage of energy from fat or carbohydrate is associated with metabolic syndrome and its components, such as elevated triglycerides and total cholesterol [7]. Therefore, dietary fat intake and dietary carbohydrate intake was categorized into quintiles with the percent from each nutrient by sex.

All missing environmental variables were imputed by Miss-Forest R package [8].

Metabolic profile

Metabolic traits used in the analysis were fasting glucose, 2 hour GTT glucose, HbA1C(%), insulin level, total cholesterol, HDL, triglyceride, SBP, DBP, waist-hip ratio and body mass index. All phenotypes were inverse-normal transformed since some phenotypes seem to have skewed distribution. All missing phenotypes were imputed by Miss-Forest R-package [8].

Statistical analysis

Heritability estimation

The LMM framework was used to identify the proportion of metabolic syndrome-related traits that could be estimated from host genetics or microbiome. The analysis above was conducted through the GCTA tool (Yang et al., 2011). The tool uses genomic relationship matrix using all observed genotypes among individuals and estimates the phenotypic variance explained by all genome-wide SNPs (h^2) by the restricted maximum likelihood (REML) approach. The higher h^2 value indicates that the phenotypic variance we are interested in is well explained by genetic architecture.

The genomic relationship matrix (GRM) between individuals j and k can be estimated by following equation,

$$A_{jk} = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_{ij}-2p_i)(x_{ik}-2p_i)}{2p_i(1-p_i)}$$ (Yang et al. 2011)

where x_{ij} is the genotype frequency for the i^{th} SNP of the j^{th} individual and p_i is the population frequency of specific SNP allele.

Similarly, we experimented with various types of metagenomic relationship matrix (MRM) and estimated the phenotypic variance explained by microbiome variance component. We denoted phenotypic fraction of microbiome h_m^2 in the context of h^2.

As the relationship matrix, a covariance matrix was constructed using inverse normal transformed-relative abundance for each genus. Metagenomic relationship matrix was calculated as $M = \frac{XX'}{m}$,

where x_{ij} in matrix X is the inverse-normal transformed abundance of j_{th} genera in individual i and m is the total number of genus.

Thus, it is similar to the genomic relationship, indicating the similarity between individuals according to the abundance of specific genus. In addition, we also considered a covariance matrix by encoding absence/presence indicator (0 or 1) of relative abundance. To adopt the best fitting model among the MRMs, the differences in the Akaike information criterion (ΔAIC) between null model (without metagenomic component) and full model was compared. The AIC defined as 2v- 2ln(likelihood), where v is the number of variance components.

Smoking and drinking, physical activity, and nutrient intake were also contained in the model as covariates, and the variance explained by them was notated h_e^2 .

Prediction model

We applied a prediction model as another way of comparing host genome and urine microbiome without bias. We considered two methods: a polygenic risk score and penalized regression according to data sources.

Data source	Model	Method
Basic features	Y ~age + sex	Linear regression
Basic + Nutrient	Y ~age + sex + nutrient intake	Linear regression
Basic + Genome	Y ~age + sex + nutrient intake + PRS	Adding polygenic risk score to linear regression
Basic + Metagenome	Y ~age + sex + nutrient intake + genus	Shrinkage method (Ridge, Lasso, E-net)
Basic + Genome + Metagenome	Y ~age + sex + nutrient intake + PRS + genus	Combining PRS and ridge coefficients of genus

Table 1. Prediction model

All of the metabolic traits were used as outcome. For prediction scheme, we applied 10-fold cross validation and all samples were randomly split into 10-folds. Each fold total samples were data source. In each fold,

train data was used to obtain GWAS SNP effect or shrinkage coefficients of genus and test data was used to evaluate total performance. Prediction performances were evaluated by coefficient of determination (R^2).

Polygenic Risk Score

Calculation of polygenic risk score (PRS) consists of two procedures. Once each SNP effect size was estimated by conventional linear regression in GWAS, then individual's risk score was calculated by computing the sum of risk alleles, weighted by the effect size on the specific phenotype.

However, SNPs located within the same genomic region are in linkage disequilibrium (LD) and including these SNPs in the model alleviates the prediction performance. So we used the clumping method, which aims to select SNPs so that the most associated SNP in the same region remains in the risk profile.

Penalized regression

Penalized regression allows to construct a linear regression model that is penalized, for having too many factors in the model, by adding a constraint as called in the equation. The consequence of adding this penalty is to shrink the coefficient values towards zero. This reduces the effect of less informative variables on the phenotype, and there are methods such as Ridge, Lasso and E-net depending on the penalty term.

RESULTS

In this study we evaluated the effects of host genetics, metagenome, and environmental factors on metabolic traits via two distinct approaches: variance estimation and prediction model. The analysis included 3,248 individuals with genome, metagenome, and environmental factors available. Since large numbers of samples were lost due to the metagenome data source, the same analysis was applied to the data of 8,476 individuals having genome and environmental factors. Table 2 shows that the distributions of 8,476 samples and 3,248 samples are approximately similar.

	Matched Samples (N=3248)	**Total Population** (N=8476)
Age	48.9 ± 7.7	52.0 ± 8.8
Sex		
Male	1655(51.0%)	4486 (52.9%)
Female	1593 (49.0%)	3990 (47.1%)
Fasting Glucose	87.4 ± 20.4	87.6 ± 21.9
HbA1C(%)	5.7 ± 0.8	5.7 ± 0.8
Total cholesterol	190.9 ± 36.5	191.8 ± 35.8
Triglyceride	159.3 ± 100.8	161.2 ± 103.5
HDL cholesterol	45.2 ± 10.4	44.7 ± 10.1

Body Mass Index	24.5 ± 3.2	24.6 ± 3.1
Waist Hip ratio	0.9 ± 0.1	0.9 ± 0.1
Waist Circumference	83.7 ± 8.7	82.5 ± 8.8

Table 2. Baseline characteristics

Variance estimation

We first investigated how well metabolic traits can be inferred from the perspective of the microbiome as compared to host genetics. Prior to comparing host genetics and microbiome, we evaluated metagenomics two different metagenomics relationship matrices based on ΔAIC. Table 3 compares the differences in the AIC between null model and full model. The MRM_INT with an inverse-normal transformation of the relative abundance shows a better fit across all phenotypes although estimated variance is lower than MRM_raw.

Phenotype	**MRM_INT**		**MRM_raw**	
	h_m^2(s.d)	ΔAIC	h_m^2(s.d)	ΔAIC
Fasting Glucose	4.12(0.012)	63.262	15.36(0.039)	54.966
HBA1C%	0.6(0.007)	-1.050	2.83(0.027)	-1.096
Total Cholesterol	4.95(0.013)	110.582	15.45(0.03)	92.360
Diastolic BP	<0.01	-	<0.01	-

Systolic BP	<0.01	-	<0.01	-
Triglycerides	0.1	-2.000	1.31(0.02)	-1.264
HDL	2.15(0.008)	28.104	7.74(0.029)	23.194
BMI	0.001	-2.004	0.001	-2.068
WHR	7.75(0.005)	254.762	25.4(0.044)	238.798
Waist circumference	4.44(0.123)	87.950	16.3(0.06)	80.245

Table 3. Comparison of MRM matrix

Table 4 shows the phenotypic variance explained by each data source and their significance level. We found that 5 of the 11 phenotypes were significantly associated with microbiome composition after adjusting for age, sex and several environments, with h_m^2 values of 4.1% for fasting glucose, 5.0 % for total cholesterol, 2.1% for HDL, 7.75% for WHR, 4.85% for waist circumference.

On the other hand, only 3 out of the 11 phenotypes were significantly associated with genetic architecture, with 19.33 % for 2 hours-GTT insulin, 10.55 % for DBP, 16.91% for Triglyceride. In addition, most phenotypes showed much lower heritability values than those already presented in previous researches. We have therefore suggested the heritability of the 8,476 KARE samples available for genome data and have confirmed that the heritability estimates are significantly improved when the sample size

increased. For example, total cholesterol which is known to show 14-50% of heritability was 11% in 3,248 samples versus 16.5% in 8,476 samples.

It has also been confirmed that the estimation by microbiome data source was more accurate and reliable across most phenotypes given the same sample size.

The effect of the environmental factors estimated via the fixed effect was about 1% over the entire 11 phenotypes.

	Genome (N=3248)		**Genome** (N=8476)		**Metagenome** (N=3248)		**Environment** (N=3248)
	h^2	P-value	h^2	P-value	h_m^2	P-value	h_e^2
Fasting Glucose	7.63	0.18	11.66	5.49E-06	4.11	3.33E-16	0.21
HBA1C	2.62	0.3	7.85	7.26E-04	0.6	0.16	0.02
2H-GTT Insulin	19.33	0.008	11.09	5.08E-06	0.3	0.39	0.02
Total cholesterol	11.28	0.07	16.47	1.25E-10	4.95	0.00E00	0.02
DBP	10.55	4.37E-02	13.42	3.4E-07	<0.01	0.5	0.53
SBP	4.96	0.21	10.32	2.45E-06	<0.01	0.5	0.22
Triglyceride	16.91	0.016	17.96	3.14E-10	0.1	0.29	0.16
HDL	9.92	0.12	12.67	3.07E-06	2.15	2.04E-08	0.39
BMI	3.06	0.3	13.60	7.17E-08	<0.01	0.5	0.02
Waist-Hip Ratio	5.67	0.23	11.24	1.38E-05	7.75	0.00E00	1.69
Waist circumference	<0.01	0.49	5.6	0.003	4.85	0.00E	0.94

Table 4. Variance explained by each data source

Prediction accuracy

Next, we made prediction model with a combination of data sources. We empirically identified that BLUP – based SNP selection and penalized regression did not improve the prediction performance in genome data. Therefore, the prediction performance of the genome data was evaluated using the PRS constructed with the SNP effect of GWAS result.

The prediction accuracy from metagenome data were evaluated through ridge regression because there was little difference between penalized models (ridge, LASSO, E-net).

Basic model includes age (continuous) and sex (binary) as common covariates. The prediction accuracy using the metagenome data source was much higher in 4 out of 11 phenotypes. Phenotypes such as DBP and BMI showed low accuracy regardless of which data source was used.

We also confirmed that the prediction accuracy improves when host genetic and metagenome data are used simultaneously.

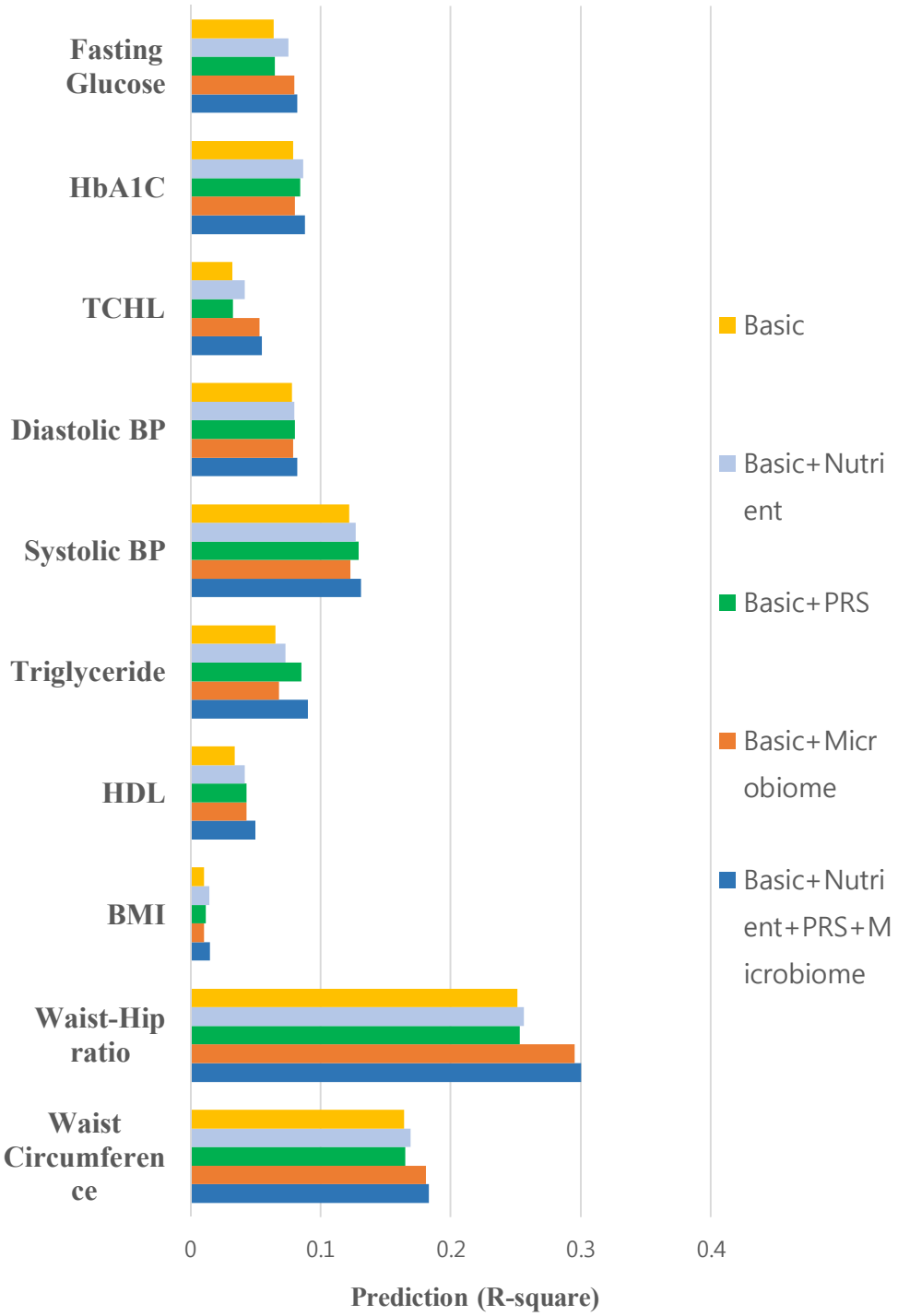

Figure 2. Phenotype prediction accuracy

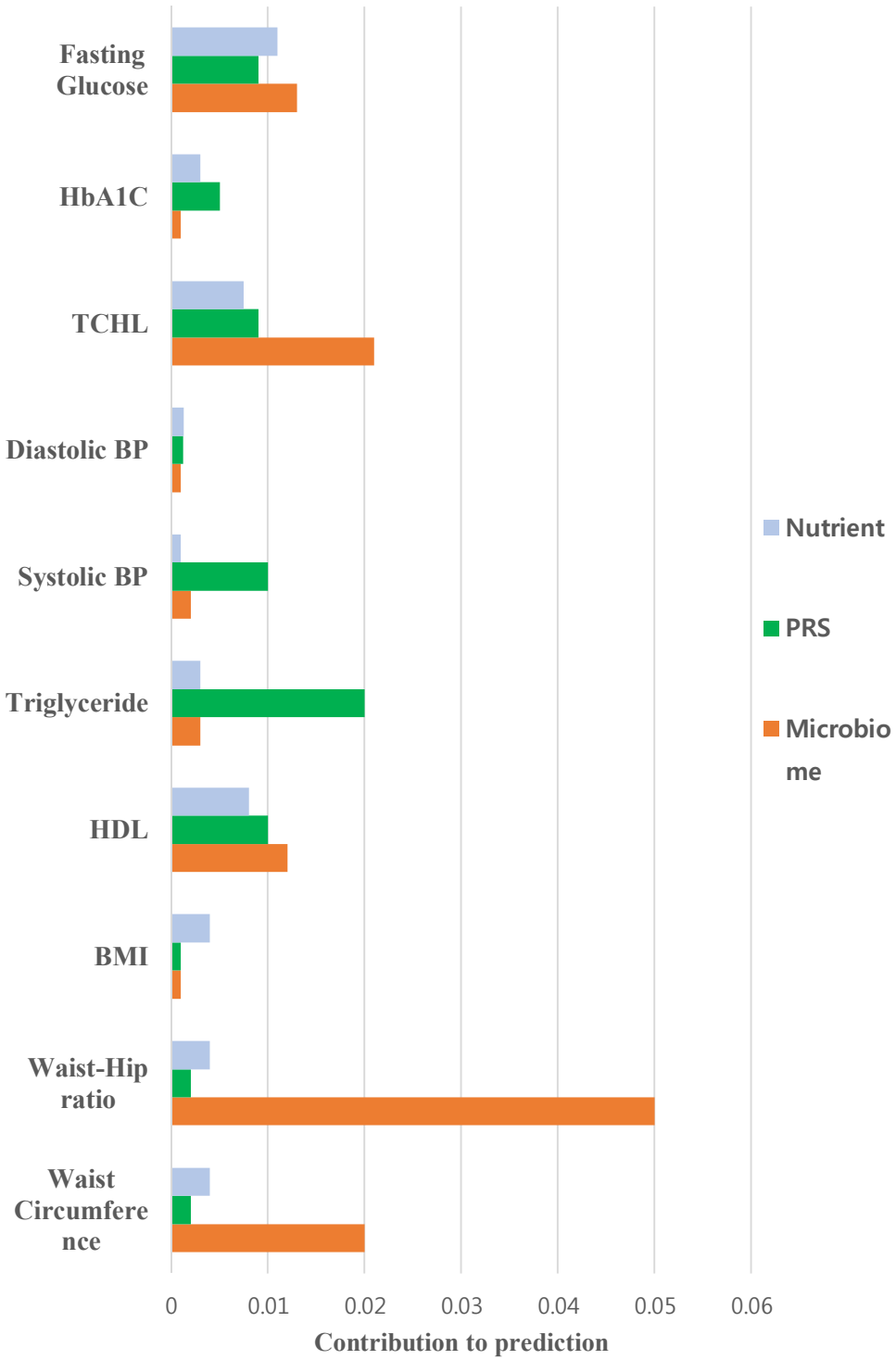

Figure 3. Contribution to prediction

DISCUSSION

In this study we integrated data sources with different features and applied various statistical techniques to provide further understanding of metabolic syndrome components. Through various experiments, we have provided some meaningful implications.

First, we identified that the association of each data source differs according to metabolic syndrome components. Notably, In the WHR and waist circumference, which is an indicator of obesity, prediction performance from microbiome was much higher than other data sources and the relative importance was also the largest. However, there may be a bias in this result since human microbiome has variability and constantly interacts with multiple environmental factors. Human microbiome affects and is affected by environmental factors.

In the components related to diabetes and dyslipidemia, the performance of each data source was different in components with similar metabolic characteristics. For example, the contribution of microbiome source was the highest in fasting glucose, while the contribution of host genome was highest in HbA1C. This difference may be due to the polygenic risk score, which is the prediction method for host genome. Because it only includes SNPs that satisfy specific p-value thresholds, SNPs with small effect were not considered. However, since we have confirmed that the performance of PRS and BLUP method is not different in 3000 sample size,

if more sufficient sample is obtained in the further studies, more accurate performance of genome will be possible.

In some components, we confirmed that the microbiome data source has a more accurate and informative effect than that of the genome data source. We also confirmed that the results were reliable by confirming that the relative effects of genome and microbiome on metabolic traits are consistent when applying different approaches.

We have also verified that the quality of urine-based metagenomics sequencing data is maintained to some extent by confirming that the urine-based data in this study had a similar result to gut microbiome. Generally, stool samples are widely used to identify the association between phenotype and gut microbiome. In previous literature performing analogous analysis scheme to ours, the microbiome data source showed better predictability than host genome in Waist-Hip ratio (WHR), waist-circumference and fasting glucose. Our study also showed that the prediction performance of microbiome source to those phenotypes was superior to host genome.

The limitation of this study is that the effects of metagenome data source are confounded by several unobserved environmental factors. It is known that host genome also influences the composition of microbiome although its association is still debating. Therefore, the heritability estimated from the microbiome source should be interpreted as the meaning of association rather than causality.

In real world, changes in metabolic profiles are also affected by

complex interactions between the host genome, microbiome and the environment, but these interactions have not been considered due to the limitations of statistical modeling. Therefore we have shown the associations of each data source with a simple assumption that there is no interaction between each data source.

In addition, due to the nature of the urine based EV, the renal function has a influence on the microbiome composition. However, we have partially confirmed that the difference in microbiome composition due to renal function is not large.

Many studies show that a genetic predictor alone will always have limited predicted power and is not of diagnostic value [9]. However, predictive power will increase if non-genetic risk factors are combined with the genetic predictors [10]. We empirically verified that combining host genome and metagenome data improves prediction ability of metabolic syndrome components. If more sample sizes are available and methods that reflect the characteristics of each data source are developed, it will provide a better understanding of metabolic syndrome and help improve the predictability through the lens of precision medicine.

REFERENCE

[1] P. M. Gorter, J. K. Olijhoek, Y. Van Der Graaf, A. Algra, T. J. Rabelink, and F. L. J. Visseren, "Prevalence of the metabolic syndrome in patients with coronary heart disease, cerebrovascular disease, peripheral arterial disease or abdominal aortic aneurysm," *Atherosclerosis*, vol. 173, no. 2, pp. 361–367, 2004.

[2] P. W. Franks, E. Pearson, and J. C. Florez, "Gene-Environment and Gene-Treatment Interactions in Type 2 Diabetes," *Diabetes Care*, vol. 36, no. 5, pp. 1413–1421, May 2013.

[3] M. T. Khan, M. Nieuwdorp, and F. Bäckhed, "Microbial Modulation of Insulin Sensitivity," *Cell Metab.*, vol. 20, no. 5, pp. 753–760, Nov. 2014.

[4] J. Wang *et al.*, "A metagenome-wide association study of gut microbiota in type 2 diabetes," *Nature*, vol. 490, no. 7418, pp. 55–60, 2012.

[5] J. Yang, S. H. Lee, M. E. Goddard, and P. M. Visscher, "GCTA: a tool for genome-wide complex trait analysis.," *Am. J. Hum. Genet.*, vol. 88, no. 1, pp. 76–82, 2011.

[6] B. Howie, J. Marchini, and M. Stephens, "Genotype imputation with thousands of genomes.," *G3 (Bethesda).*, vol. 1, no. 6, pp. 457–70, 2011.

[7] S. J. Song, J. E. Lee, W. O. Song, H. Y. Paik, and Y. J. Song, "Carbohydrate intake and refined-grain consumption are associated with metabolic syndrome in the korean adult population," *J. Acad. Nutr. Diet.*, 2014.

[8] D. J. Stekhoven and P. Bühlmann, "Missforest-Non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, pp. 112–118, 2012.

[9] N. R. Wray, J. Yang, M. E. Goddard, and P. M. Visscher, "The genetic interpretation of area under the ROC curve in genomic profiling," *PLoS Genet.*, vol. 6, no. 2, 2010.

[10] N. R. Wray, S. H. Lee, D. Mehta, A. A. E. Vinkhuyzen, F. Dudbridge, and C. M. Middeldorp, "Research Review: Polygenic methods and their application to psychiatric traits," *J. Child Psychol. Psychiatry Allied Discip.*, vol. 55, no. 10, pp. 1068–1087, 2014.