이 학 박 사 학 위 논 문

Quantifying genetic effect on the longitudinal phenotypic profile

표현형 종단 자료에 대한 유전적 영향의 정량화

2020 년 2 월

서울대학교 대학원

협동과정 생물정보학과

이 동 혁

Quantifying genetic effect on the longitudinal phenotypic profile

by

Donghe Li

A thesis
submitted in fulfillment of the requirement
for the degree of Doctor of Philosophy
in
Bioinformatics

Interdisciplinary Program in Bioinformatics
College of Natural Sciences
Seoul National University
Feb, 2020

Quantifying genetic effect on the longitudinal phenotypic profile

지도교수 원 성 호

이 논문을 이학박사 학위논문으로 제출함

2020 년 2 월

서울대학교 대학원

생물정보협동과정 생물정보학 전공

이 동 혁

이동혁의 이학박사 학위논문을 인준함

2020 년 2 월

위 원 장 박 태 성 (인)

부위원장 원 성 호 (인)

위 원 손 현 석 (인)

위 원 유 연 주 (인)

위 원 박 주 현 (인)

Abstract

Quantifying genetic effect on the longitudinal phenotypic profile

Donghe Li

Interdisciplinary Program in Bioinformatics

The Graduate School

Seoul National University

The main goal is to identify the progressing effect of SNPs on the important health related phenotypic traits, and lung function specific traits by calculating SNP heritability with longitudinal data. The total 16 prominent health-related phenotypic traits were observed biennially for each subject during 10 years, and 12 spirometric measures were biennially observed for 14 years. SNP-based heritabilities for those phenotype averages and annual change were estimated. Since linear mixed models with two random effects are computationally very intensive, here, we proposed and applied two-stage model. First, the phenotypic average and annual change for each subject were estimated with a linear model, and then both regression coefficients were used as responses to estimate SNP heritability with GCTA software. This approach provides a reasonable and easy method to estimate heritability in longitudinal

data and potentially assess both heritability of the phenotypic averages and changes through several periods. In the 16 health-related phenotypes analysis, results show that that significant SNP heritability is objectively confirmed for longitudinal changes in lung function decline including FEV1 in comparison with other health-related indices. In the 12 lung function specific analysis, SNP heritabilities of the annual change rate of FEV1 % predicted and FEV1/FVC were significantly high ($h_{decline}^2$=0.105, p-value=0.004 for FEV1 % predicted; $h_{decline}^2$=0.157, p-value=7.25×10^{-5} for FEV1/FVC). In subgroup analsysis, POST FEV1/FVC ($h_{decline}^2$=0.399, p-value=0.009) were in never smokers significant high than in ever smokers.

Key words: Genome-wide association studies (GWAS), heritability, GREML, longitudinal

Student number: 2014-31030

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The background on genetic association studies

1.1.1 Overview of genome-wide association studies

Genome-wide association study (GWAS), which analyzes thousands of genetic variants across human genome to identify genetic risk factors for complex disease and traits that are common in the population(Bush and Moore 2012). The main goal of GWAS is to use genetic risk factors to predict personal disease status and to find connected biological mechanisms for developing new preventions and treatment strategies (Bush and Moore 2012). Since the success of GWAS on identifying age-related macular degeneration (AMD) risk factor gene, more than 50 thousand associations between

genotype and phenotype under genome-wide significant ($P < 5\times10^{-8}$) have been reported (Buniello et al. 2019).

As one of the methods to identify genetic risk factors, GWAS have led to insights into several important area as the architecture of disease susceptibility which through identifying the novel disease-causing genes, and clinical care which is about identifying new drug targets and disease biomarkers, and personalized medicine and personal genetic testing, which aims to provide optimized healthcare to individual patients based on their genetic information and other biological features (Bush and Moore 2012).

To identify genotype-phenotype associations, we need both genotype and phenotype dataset, which can be generated from different types of sources. In GWAS, the genome-wide single-nucleotide polymorphism (SNP) arrays that combined with imputed genotypes based on population reference panels, are generally used as genetic information. SNP arrays are highly accurate and reliable, and there are well-established analytical pipelines and tools have been developed for data analysis. As the improvement of sequencing technology, whole-genome sequencing (WGS) is also an alternative choice. Currently, even it is relatively expensive and less mature and less accurate comparing to SNP arrays, but it is possible to detect and fine-map the rare variants, even detect the ultra-rare variants that growing evidences show the rare variants or low-frequency variants contribute to the etiology of complex disease, and the trend to shift to WGS is inevitable in the near future. For the

phenotype data, it primarily consists of two classes that are categorical (often binary case/control) and quantitative data. As a disease status, which diagnosed as affected or unaffected, can be treated as a binary categorical variable, and the disease traits, like body mass index(BMI), high-density lipoprotein(HDL) and low-density lipoprotein(LDL), which are all measured in continuous values, can be regarded as quantitative data. From the statistical perspective, quantitative traits are preferred because they improve power to detect a genetic effect (Bush and Moore 2012).

1.1.2 Single SNP-based analysis in GWAS

If a well-defined phenotype has been selected for a study population, and the genotypes of the population are collected by reasonable techniques, statistical analysis is ready to perform with these data. The widely applied analysis of genome-wide association data is known as single locus statistic test, which test the association between each independent SNP and phenotype. There are variety of factors should be considered in the selection of statistical test, but the primary consideration is according to the type of phenotype data that case/control or quantitative one.

Generally, a contingency table or logistic regression would be applied to binary case/control traits. Contingency table tests examine and measure the deviation from independence that is expected under the null hypothesis that there is no association between the phenotype and genotype classes. The chi-

square test usually applied to this method. Logistic regression is an extension of linear regression using a logistic function as the outcome which predicts the probability of having case status given a genotype class. Logistic regression also can adjust the covariates and provide adjusted odds ratios as a measure of effect size (Bush and Moore 2012).

For the quantitative traits, generalized linear model (GLM) approaches are generally applied, most commonly the Analysis of Variance(ANOVA). The null hypothesis of an ANOVA using a single SNP is that there is no difference between the traits means of any genotype group. And there are some assumptions of GLM and ANOVA that the trait is following normal distribution and having the same variance within each group and the groups are independent (Bush and Moore 2012).

Except for selecting right method for analysis, in many situations, there are confounding factors (covariates) that can affect the relationship between independent variables and the outcome. Thus, we need to adjust for the covariates such as sex, age or principle component (PC) scores to reduce the spurious associations in the regression model.

The most widely used program in GWAS is PLINK, which is a freely available analysis toolkit, it has a wide range of functions, including those related to data organization, formatting, quality control, association testing and much more.

1.2 The background on heritability estimation

1.2.1 Overview of heritability estimation

Heritability is usually explained in two different aspects: one is the broad-sense heritability, which reflects all the genetic contribution to a population's phenotypic variance including additive, dominant, and epistatic, as well as parental effects, where individuals are directly affected by their parent's phenotype. The other one is the narrow-sense heritability, which only infers the proportion of the total phenotypic variance explained by the additive effect of genetic variance (Yang et al. 2010, Visscher, Hill, and Wray 2008). According to the different definition of heritability, there are also different approaches to estimate it. The former one utilizes pedigrees or twins, but one drawback of utilizing this method is that it heavily relies on the assumption regarding the cause of covariance between close relatives, which can bias the results if the assumption is false. The latter method estimates heritability with unrelated individuals and genomic data on single-nucleotide polymorphism (SNP), which is unlikely to be confounded by other environmental effects in the additive effect of genomic variance. In this study, we focus on the latter definition, estimating the narrow-sense heritability, also called SNP-based heritability (h^2_{snp}), which is explained by all SNPs used in genome-wide association study (GWAS), genotyped in unrelated individuals for complex traits and diseases. The narrow-sense heritability quantifies the aggregate genetic contribution without identifying specific effects.

1.2.2 Summary of heritability estimation methods

There are several h^2_{snp} estimation methods have been developed and still being updated nowadays. One of the most popular methods is proposed by Yang, et.al., which suggests generating genetic relatedness matrices (GRMs) to estimate genetic and phenotypic variances with restricted maximum likelihood (GREML) through linear mixed model (LMM) (Yang et al. 2010). Other popular method is LDAK, developed by Speed, et.al., which calculates a modified kinship matrix in which SNPs are weighted according to local linkage disequilibrium (LD) (Speed et al. 2012). Besides these two, other methods employ computationally efficient mixed model approaches (Loh et al. 2015), such as relating the effect sizes of SNPs from a GWAS to their degree of LD tagging (Finucane et al. 2015, Bulik-Sullivan et al. 2015), treelet covariance smoothing (Crossett et al. 2013), or using related and unrelated samples to account for rare and common variant effects (Zaitlen et al. 2013)and so on. Importantly, the fact that there are many different methodologies estimating h^2_{snp} could lead to discrepancies in estimation and even considerable biases across the different procedures. Thus, not only should estimations should be carefully interpreted, but also it is recommended to try several methods before giving a final estimation value (Ni et al. 2018, Evans et al. 2018).

1.3 Overview of GxE analysis

Even though, GWAS have achieved the clear success, the study design still has not been avoided some controversy, as the single-nucleotide variants(SNVs) identified in GWAS explain only a small fraction of the heritability of complex traits (Manolio et al. 2009), and it may represent spurious associations (McClellan and King 2010) and do not necessarily infer the causal variants and genes(Boyle, Li, and Pritchard 2017), and that GWAS will yield too many loci that may uninformative if the detected variants in all genes are implicated (Goldstein 2009). Therefore, it has been proposed to focus efforts on the analysis of rare-variants, even ultra-variants, and post-GWAS experiments that functional studies, gene network analysis and translational medicine (Tam et al. 2019).

Many researchers recognized that too much focus on main effects could become a barrier to the identification of additional genes underlying these disease traits. Increasing emphasis is being placed on gene-environment interaction analyses in recent years (Sung et al. 2014).

One of reasons to identify GxE interaction, as GxE interaction or more complex pathways involving multiple genes and environments could explain parts of missing heritability. They also can further elucidate the biological networks underlying complex disease risk and enable ″profiling″ of individuals who are at the highest risk for disease (Sung et al. 2014).

1.4 Overview of Longitudinal analysis

The progression of diseases or traits can be assessed with longitudinal study designs in which the repeatedly measured outcomes are provided. Longitudinal data allows researchers to assess temporal disease aspects, especially, compared to cross-sectional studies, longitudinal studies often have less variability and increased statistical power(Zeger and Liang 1992). But the analysis is complicated by complex correlation structures, irregularly spaced visits, missing data, and mixtures of time varying and static covariate effects (Garcia and Marder 2017). There are several methods have been developed to handle these complications, as mixed effect regression model, and it is more important to use these methods appropriately and interpret their outputs correctly.

1.5 The purpose of this study

The main purpose of this thesis is to explore the progressing genetic effect on the important health related phenotypic traits by using genome-wide association analysis and heritability estimation with longitudinal data. To overcome the analysis problem with longitudinal data, we applied two-step approach to estimate the effects of averages and longitudinal changes of phenotypic traits through periods.

In the first study, sixteen phenotypic traits associated with major health indices were observed every two years for 6,843 individuals with 10-year follow-up in a Korean community-based cohort. Average SNP heritability and longitudinal changes in the total period were estimated using a two-stage model. Average and periodic differences for each subject were considered responses to estimate SNP heritability. Furthermore, a genome-wide association study (GWAS) was performed for significant SNPs.

In the second study, twelve spirometric measures were observed every two years for 8,768 Korean adults aged 40-69 years during 14 years. Phenotypic averages and annual change were calculated for each participant, and SNP heritabilities for both were estimated by GCTA. Furthermore, we also calculated the subgroup heritibility of smoking status.

1.6 Outline of the thesis

This thesis is organized as follows. Chapter 1 is an overview of GWAS and heritability estimation on the background and the methods that the studies applied. Chapter 2 contains an overview of genetic effects quantifying analysis with longitudinal data which applied in the following studies. Chapter 3 deals with identifying the progressing effects of SNPs on 16 phenotypic traits with longitudinal data. Chapter 4 is about heritability analyses which revealed the significant effect of SNPs on lung function decline rate. At last, the summary and conclusions are presented in Chapter 5.

Chapter 2

An overview of genetic effect quantifying analysis with longitudinal data.

2.1 Challenges of genetic effect quantifying analysis with longitudinal data

The traditional way of analyzing genetic variants that influence complex traits is cross-sectional study, which usually focuses on phenotypes and covariates measurements from a single time point. Even though genetic variants are basically fixed, the quantitative disease traits and their associated risk factors would be varying over time. Recently, many genetic association studies have been performed on longitudinal cohorts to take advantage of repeat measurement of time varying variables (Wu, Hu, and Melton 2014).

There are several advantages by performing longitudinal analysis in genetic studies. First, repeated measurements can reduce type I error, increasing statistical power compared to a single measurement. Second, by analyzing longitudinal data, we can identify genetic determinants both for age of onset and subsequent progression of phenotypic traits. Finally, longitudinal studies could handle the prospective measurement of time-varying covariates that are not typically included in traditional genetic studies(Wu, Hu, and Melton 2014).

There are some challenges of genetic effect quantifying analysis with longitudinal data. One of the challenges comes from the correlated data, as measurements in longitudinal studies are correlated by design. Correlation exists between repeated measures on the same individual or the individuals from similar sites that sharing the same investigator, study protocol variations and equipment. The within-family correlation also could be a problem. Another one is computational burden. There are some advanced statistical methods developed for epidemiological studies, including generalized estimating equations (GEE) and linear mixed models (LMM) with two random effects, that account for large pedigree structure may not be available to whole-genome sequence data. Beside these challenges, missing data, irregularly spaced visits, and mixtures of time-varying and static covariate effects are problems should be considered in longitudinal study, thus, the additional statistical consideration should be accounted to solve these problems.

2.2 Review methods of longitudinal data analysis

Here, we are going to review several methods applied in longitudinal data analysis. The methods generally could be separated into traditional and modern ones. The traditional methods include ANOVA approaches like repeated measures ANOVA and multivariate ANOVA (MANOVA). The modern methods include generalized estimating equations model (GEE) and mixed effects regression (MER) (Garcia and Marder 2017).

ANOVA approaches are limited in handling irregularly timed and missing data. Repeated measures ANOVA assesses group differences over time, the group sizes can be different, but all participants must be measured at the same number of time points. The downside of repeated measures ANOVA is it assumes the measured outcomes have equal variances and covariances over time. This may be unrealistic since variances tend to increase with time and covariances decrease with increasing intervals in time. The MANOVA model, in comparison, makes no assumptions about the variance-covariance structure of the repeated measures, and thus removes misspecification concerns, but it requires fully complete data. Applying ANOVA methods to data with missing observations yields biased parameter estimates.

The limitations ANOVA approaches inspired to use the modern approaches that robustly handle challenges of longitudinal studies. Two

preferred modern methods for longitudinal data include generalized estimating equations model (GEE) and mixed effects regression (MER) (Garcia and Marder 2017). Both of the methods allow time-invariant predictors (e.g. gender, genotype) and time-varying predictors (e.g. age), and could handle irregularly timed and missing data without the need for explicit imputation(Garcia and Marder 2017).

GEE model could be applied for analyzing the regression relationship between covariates and repeated responses, but not the correlation structure of the repeated responses. When estimating the regression parameters, the correlation structure in a GEE is represented using a working, potentially incorrect model, but it still yields valid estimates without disregarding incomplete data(Garcia and Marder 2017), and it applies quasi-likelihood methods which is computationally easier than full-likelihood methods. The limitations of GEE include it cannot perform hypothesis testing since these are not directly estimated, and it cannot be used to test and compare model fits with usual methods like likelihood ratio tests (LRT), Akaike/Bayesian Information Criterions (AIC/BIC), because it focuses on regression parameters, not all model parameters(Garcia and Marder 2017).

MER models could be used for analyzing the regression relationship between covariates and repeated responses, and also the correlation structure of the repeated response. The correlations of repeated measures could be estimated by using random effects, which describing the cluster-specific trends over time. Random effects allow estimation of cluster-specific effects

useful for understanding interindividual variability in longitudinal responses and cluster-specific predictions (Garcia and Marder 2017). The MER advantages are not only could handle the limitations of GEE we listed before, but it is more robust to missing data and assumes MAR as missingness which is more general than the MCAR assumption of GEE. However, MER models still have limitations that the computational complexity, particularly with nonlinear MER, it involves time-consuming numerical integration over the random effects (Garcia and Marder 2017).

2.3 Method applied in this paper with longitudinal data analysis

We assume that the observed trait of subject *i* at time point *j* is y_{ij}, then we assume y_{ij} is a function f_i of his and her age, age_{ij}, and a measurement error with variance σ_m^2, then we have equation as follows:

$$y_{ij} = f_i(\text{age}_{ij}) + \varepsilon_{ij}, \varepsilon_{ij} \sim MVN(0, \sigma_m^2).$$

If we say f_i is simple linear regression of age_{ij}, and we centering the age by subtracting the mean of age ($\overline{\text{age}}_i$), then it can be shown as

$$f_i(\text{age}_{ij}) = \beta_{0i} + \beta_{1i}(\text{age}_{ij} - \overline{\text{age}}_i).$$

Here, β_{0i} indicates the expected phenotypic mean of subject i for the observed trait when he or she is $\overline{age}_i$ years old, and β_{1i} is the average longitudinal change in that trait. Furthermore, we apply β_{0i} and β_{1i} in linear mixed model with sex and $\overline{age}_i$ as fixed effect and g_i as the random effect of SNPs, b_i is error term. Then we have following two equations.

$$\beta_{0i} = \alpha_0^0 + \alpha_1^0 \text{sex}_i + \alpha_2^0 \overline{\text{age}}_i + g_i^0 + e_i^0,$$

$$\beta_{1i} = \alpha_0^1 + \alpha_1^1 \text{sex}_i + \alpha_2^1 \overline{\text{age}}_i + g_i^1 + e_i^1.$$

We let $\mathbf{g}^0 = (g_1^0 \quad \cdots \quad g_n^0)^t$, $\mathbf{g}^1 = (g_1^1 \quad \cdots \quad g_n^1)^t$, $\mathbf{e}^0 = (e_1^0 \quad \cdots \quad e_n^0)^t$ and $\mathbf{e}^1 = (e_1^1 \quad \cdots \quad e_n^1)^t$. If we let $\mathbf{G}$ be the genetic relationship matrix,

$$\mathbf{g}^0 \sim MVN(\mathbf{0}, \sigma_{g0}^2 \mathbf{G}), \mathbf{e}^0 \sim N(0, \sigma_{e0}^2),$$

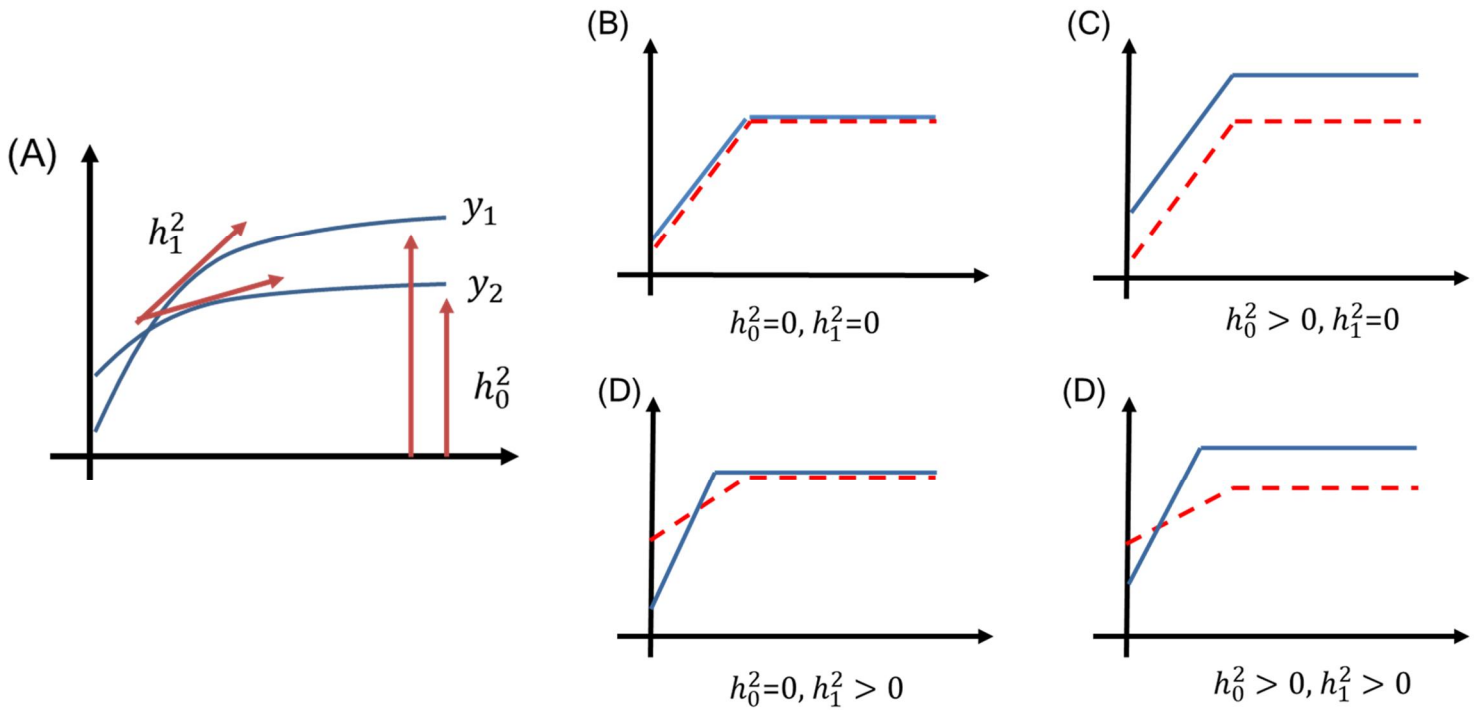
$$\mathbf{g}^1 \sim MVN(\mathbf{0}, \sigma_{g1}^2 \mathbf{G}), \mathbf{e}^1 \sim N(0, \sigma_{e1}^2).$$

Then two different relative proportions of phenotypic variances explained by the genetic components can be defined by

$$h_0^2 = \frac{\sigma_{g0}^2}{\sigma_{g0}^2 + \sigma_{e0}^2}, h_1^2 = \frac{\sigma_{g1}^2}{\sigma_{g1}^2 + \sigma_{e1}^2}.$$

h_0^2 indicates the relative proportions of phenotypic variances explained by the genetic components when he or she is $\overline{\text{age}}_i$ years old, and is equivalent to the SNP-based heritability. h_1^2 indicates the relative proportions of variances of phenotypic changes explained by the genetic components. Figure 2.1 shows the illustrative example of h_0^2 and h_1^2. We assume that effect of environment is small and phenotypes are mostly determined by genetic components. Then phenotypic distributions according to age can differ by h_0^2 and h_1^2. If h_0^2 or h_1^2 is larger than 0, then phenotypic average or annual changes are associated with genetic components.

Figure 2.1 The illustrative example of h_0^2 and h_1^2. We assume that effect of environment is small and phenotypes are mostly determined by genetic components. Then phenotypic distributions according to age can differ by h_0^2 and h_1^2.



This chapter was accepted by journal of *Genes & Genomics*
as a partial fulfillment of Donghe Li's Ph.D program.

Chapter 3

Identifying Progressing Effect of SNPs on 16 Phenotypic Traits with Longitudinal Data

3.1 Introduction

Single nucleotide polymorphism (SNP)-based heritability (h^2_{snp}) indicates the relative proportion of genetic variance explained on the basis of SNPs used for genome-wide association studies (GWASs). For h^2_{snp} estimation, the genomic restricted maximum likelihood (GREML) for linear mixed models (LMMs) is often implemented in the genetic complex trait analysis (GCTA) tool (Yang, Manolio, et al. 2011). GREML first calculates

the genetic relatedness matrices (GRM), which are used as variance-covariance matrices for random effects. The significance of estimates obtained through GREML depends on the study design; if it is applied to family-based samples, it displays pedigree-based heritability, but for unrelated subjects, it estimates h^2_{snp} (Yang et al. 2017, Kim, Lee, et al. 2015). Estimating h^2_{snp} involves considerable differences across not only methodologies but also procedures requiring careful interpretation of results (Ni et al. 2018, Evans et al. 2018). Moreover, the estimated heritability is potentially biased and misleading owing to measurement errors at various degrees. To overcome these challenges, the heritability determined from longitudinal data is more reliable than that determined from cross-sectional data. While most studies on h^2_{snp} focused on the primary effect of SNPs, significant effects of SNPs on the average annual differences indicate the SNP-by-age interaction. Numerous examples illustrate the importance of age on longitudinal changes (2000, van de Pol and Verhulst 2006, Nishimura et al. 2012). For instance, annual decline in lung function is associated with age (Kim et al. 2016), and another study reported a genetic influence on changes in both lipoprotein risk factors and systolic blood pressure over a decade (Friedlander et al. 1997). Therefore, the h^2_{snp} should be estimated on the basis of not only the mean of observed traits but also changes in the sufficient period. Hence, we applied a two-stage approach, which is a convenient method of analyzing longitudinal data by combining linear regression models

to investigate the effect of SNPs on both average and longitudinal differences in phenotypic traits.

In this study, we investigated the magnitude of the effect of SNPs on average and longitudinal differences by using both genomic data and 16 phenotypic traits associated with major health indices using a phenotype-genotype dataset of unrelated individuals in a community-based cohort and evaluated their importance. Except for baseline, each phenotype was objectively measured every 2 years for 10-year follow-up, and six repeated measurements (maximum) were obtained for each individual. For each subject, both the average phenotypic traits and their longitudinal changes were estimated via subject-specific regression analysis, using intercepts and coefficients of ages, respectively. Each h^2_{snp} value was estimated using GCTA. The results show that lung function has the only significant h^2_{snp} for longitudinal changes, while all average phenotypes of 16 traits yielded a significant h^2_{snp} value. Furthermore, the GWAS revealed certain novel genome-wide significant SNPs associated with the phenotypes analyzed herein.

3.2 Methods

3.2.1 KARE cohort data

Korea Associated Resource (KARE) data are based on a community-based epidemiological study and comprises subjects residing in Ansan (urban

area) and Ansung (rural area) in the Gyeonggi Province of South Korea (Cho et al. 2009). A baseline survey was completed in 2001–2002, and 10,030 participants aged 40–69 years were recruited. Since then, biennial repeated surveys were conducted, and the last survey were completed in 2013–2014 (Kim, Han, and Ko 2017). Six different surveys were conducted in total. These measurement periods are indicated as periods 1–6 throughout, each with a different number of subjects (period 1, 8,543 subjects [4,052 male, 4,491 female]; period 6, 5,391 subjects [2,502 male, 2889 female]). The number of overlapping subjects throughout the 6 periods was 4,306 (2,009 male, 2,297 female). Among these, subjects whose traits were measured at least thrice were considered, and 6,843 participants (3,273 male, 3,570 female) were assessed in total.

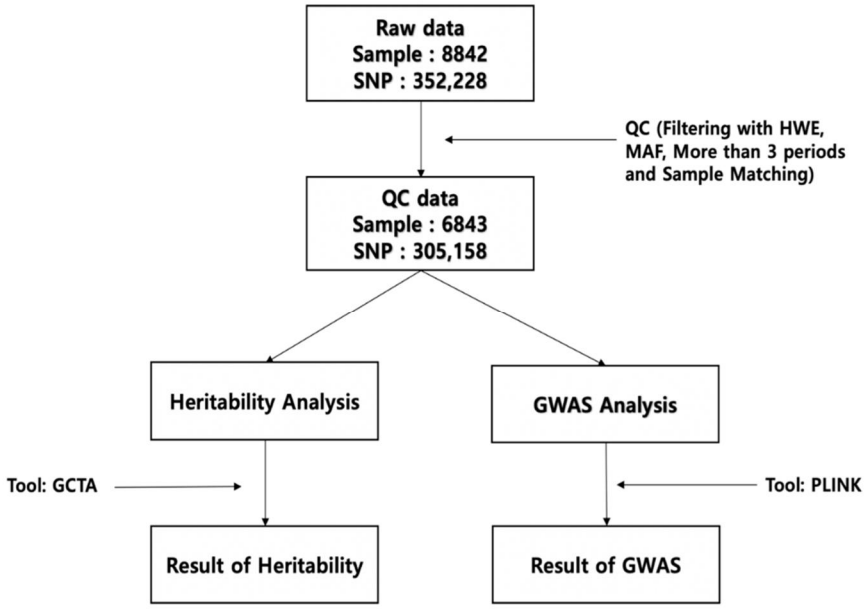
Many participant phenotypes were recorded by trained interviewers through questionnaires and clinical measurement; however, we only considered 16 quantitative traits because they were measured objectively and associated with major health indices; these were classified into four groups: anthropometric, biochemistry, cardiopulmonary, and red blood cell traits (Table 1). As glycated hemoglobin (HbA1c), fasting blood glucose (GLU0), high-density lipoprotein (HDL), triglycerides (TG), and systolic blood pressure (SBP) displayed skewed distributions, they were log-transformed and denoted by log(HbA1c), log(GLU0), log(HDL), log(TG), and log(SBP), respectively. The missing rate of HbA1c was larger than 0.5 at period 2 and

was excluded from the present analysis. For each trait, subjects with more than three measurement observations were assessed.

3.2.2 Genotype data

Genotype data for KARE cohort were obtained by using the Affymetrix Genome-Wide Human SNP array 5.0 (Cho et al. 2009). Quality control (QC) of SNPs and subjects were conducted with PLINK (Purcell et al. 2007) and ONETOOL (Song et al. 2018). We excluded SNPs with P-values from Hardy-Weinberg equilibrium (HWE) analysis $<10^{-5}$, minor allele frequencies (MAFs) <0.05, and genotype call rates <95%. Furthermore, we excluded subjects with missing genotype call rates >5% or sex-based inconsistencies. The missing genotypes for typed SNPs were imputed based on the 1000-genome sequence reference data. After quality control, 305,158 SNPs were analyzed for SNP heritability estimation and GWAS (Figure 3.1).

Figure 3.1 A schematic representation of heritability analysis and the genome-wide association study.



3.2.3 Calculation of phenotype averages and annual changes for each subject

We calculated the phenotypic averages and annual changes for each subject, and then they were used to estimate SNP heritability and for GWAS. We found significant differences of phenotypic variances among each period, and such heteroscedasticity was considered for phenotypic averages and annual changes for each subject as follows. First, we fitted the linear regression with traits belonging to the same period. Effect of sex, age, and 10 principal component (PC) scores estimated from genetic relationship matrix were adjusted by including them. If we let w_{jk} be the residual variances of trait k (k=1, …, 16) during the period j, for subject i, we fit the following linear model.

$$y_{ijk} = \beta_{ik0} + \beta_{ik1}\left(age_{ij} - \overline{age}_i\right) + \varepsilon_{ijk}, \varepsilon_{ijk} \sim N\left(0, \frac{1}{w_{jk}}\sigma_{ik}^2\right) \qquad (1)$$

Here, i indicates the ith subject, and $\overline{age}_i$ indicates the mean of ages at the observed time points. In the regression model (1), β_{ik0} indicates the expected phenotypic mean of subject i for trait k when he or she is $\overline{age}_i$ years old, and β_{ik1} is the average longitudinal change in trait k. The estimated values of β_{ik0} and β_{ik1} were used to estimate the heritability and for GWAS analysis. For convenience, both are denoted by B_0 and B_1, respectively.

3.2.4 Heritability Estimation

After we fit the Equation (1), B_0 and B_1 were separately used as responses in the SNP heritability estimation models in GCTA (Yang, Manolio, et al. 2011). The proportion of genetic variance in several chosen traits was estimated by using restricted maximum likelihood analysis, which is implemented in the GCTA. Effect of ages and sex were adjusted by including them as covariates.

3.2.5 GWAS analysis

B_0 and B_1 were also separately used as responses to identify the disease susceptibility loci for 16 different traits. Effect of ages, and sex were included as covariates. Since ages of subjects are different for different periods, the age variable was coded by $\overline{age}_i$. Furthermore, 10 PC scores estimated from the genetic relationship matrix were included as covariates to adjust the population stratification.

3.3 Results

3.3.1 Estimation of heritability

A schematic representation of heritability analysis and genome-wide association study is shown in Figure 3.1. For 16 different traits of 6,843

subjects, the mean and standard deviation values of each trait at period 1 are shown in Table 3.2 (see Table 3.1 for detailed information). Some missing values resulted in differences in the total number of subjects depending on the phenotype, and the sample sizes of those traits and descriptive statistics including sex and age are summarized.

A multidimensional scaling (MDS) plot was generated for those 6,843 subjects (Figure 3.2). As shown in Figure 3.2, subjects from the 1000 Genomes Project were also included, and the analyses were not affected by population stratification. We calculated the descriptive statistics for B_0 and B_1 (Table 3.3, see Materials and Methods for details). B_0 in equation (1) indicates the means of the predicted traits at $\overline{age}_i$ years of age. B_1 stands for the longitudinal changes in the traits of each subject. Table 3 shows that the means of B_0 are similar to those for period 1. Means of B_1 were generally closer to 0. Figure 3.3 shows the estimates of heritability with B_0 as the response in the GREML model, and the estimated heritability of height for the data peaked at 0.318 (P=1.665×10^{-16}, FDR=2.664×10^{-15}). The subsequent three highest heritability traits were total cholesterol (TCHL), log(HDL), and low-density lipoprotein (LDL), with values of 0.265 (P=3.895×10^{-12}, FDR=3.116×10^{-11}), 0.241 (P=8.911×10^{-10}, FDR=4.753×10^{-9}), and 0.222 (P=5.178×10^{-9}, FDR=1.657×10^{-8}), respectively. These three traits are cholesterol-related. The heritability of WAIST was 0.218 (P=5.016×10^{-9}, FDR=1.657×10^{-8}) and that of WEIGHT was 0.196 (P=2.046×10^{-7}, FDR=5.456×10^{-7}). For Hb, the heritability was 0.195 (P=4.926×10^{-7},

FDR=9.852×10^{-7}) and for log(TG), the value was 0.192 (P=4.419×10^{-7}, FDR=9.852×10^{-7}). The heritability of the other traits with an FDR larger than 1×10^{-6} were less than 0.19.

We compared the our estimated h^2_{snp} of 16 traits for B_0, which the phenotypic mean, with the results of Yang et al. study (Yang et al. 2013). The results are listed in Table 3.3, and we found the range of difference between the result of our study and the result of reference was about 0.02~0.17. We also estimated the h^2_{snp} of the traits for each period, and calculated the means and median of them. Figure 3.4 shows the comparison of the h^2_{snp} of phenotypic mean (B_0), result of reference and the mean of h^2_{snp} of 6 periods. Except for some traits, most of the result shows consistency.

Figure 3.5 shows the estimated h^2_{snp} for B_1, which are generally less than those for B_0, and we found that the lung function traits FVC and FEV1, WAIST, diastolic blood pressure (DBP), BMI, and log(SBP) are relatively high. The highest h^2_{snp} was observed for FEV1 (0.171) and its FDR-adjusted *P*-value was 0.0189. The heritability estimates of other traits were less than 0.1. The second highest heritability was 0.0941 for FVC, and its FDR-adjusted *P*-value was 0.166. The heritability of WAIST was also relatively higher than that of other traits. Its heritability and the FDR-adjusted *P*-values were 0.0082 and 0.0657, respectively. The higher heritability estimates of B_1 indicate that the decreasing/increasing rates are associated with genetic factors. HEIGHT displayed the highest heritability estimates for B_0; however, the

estimate for B_1 was low (0.0297). HEIGHT does not usually change since the age of 20 years, which probably attributes to the low HEIGHT value in this study. For the other traits including log(HbA1c), LDL, log(HDL), TCHL, and Hb levels, SNP heritability estimates tended towards 0.

Furthermore, we estimated the variance explained by each chromosome h_c^2 of FEV1, which displayed the highest h_c^2 in the B_1 model. Consequently, chromosome 2 accounted for the highest proportion of phenotypic variance (h_c^2=0.0397) with an albeit high standard error (Figure 3.6). We also plotted the h_c^2 against chromosome length for FEV1. There was a significant positive correlation between chromosome length and h_c^2 (r=0.58, P=0.0045) in FEV1 (Figure 3.7).

Table 3.1 Sixteen phenotypic traits associated with major health indices

Anthropomorphic Traits	Height, Waist, Weight, Body-mass index(BMI)
Biochemistry Traits	
Glucose:	Glycated hemoglobin (HbA1c), Fasting blood glucose (GLU0)
Cholesterol:	Low-density lipoprotein (LDL), High-density lipoprotein (HDL), Total cholesterol (TCHL), Triglyceride (TG)
Cardiopulmonary Traits	
Blood Pressure:	Systolic blood pressure (SBP), Diastolic blood pressure (DBP)
Lung Capacity:	Predicted forced vital capacity (FVC)%, Predicted forced expiratory volume in one second (FEV1)%, Predicted FEV1/FVC %
Red Blood Cell Traits	Hemoglobin levels (Hb)

Table 3.2 Descriptive statistics of 16 traits

Trait	Trait (Baseline)		Total (N)	Female		Age	
	Mean	SD		N	%	Mean	SD
HEIGHT(cm)	160.11	8.63	6823	3557	52.13%	51.90	8.69
WAIST(cm)	82.63	8.70	6835	3567	52.19%	51.90	8.69
WEIGHT(kg)	63.24	10.10	6822	3556	52.13%	51.90	8.69
BMI(kg/m2)	24.62	3.10	6822	3556	52.13%	51.90	8.69
HbA1c(%)	5.74	0.82	6329	3321	52.47%	51.87	8.62
GLU0(㎎/㎗)	86.73	19.41	6728	3514	52.23%	51.85	8.67
TG(㎎/㎗)	161.47	103.19	6840	3568	52.16%	51.91	8.70
LDL(㎎/㎗)	115.00	32.89	6840	3568	52.16%	51.91	8.70
HDL(㎎/㎗)	44.69	9.91	6840	3568	52.16%	51.91	8.70
TCHL(㎎/㎗)	191.92	35.09	6840	3568	52.16%	51.91	8.70
SBP(mmHg)	121.12	18.10	6843	3570	52.17%	51.91	8.70
DBP(mmHg)	80.19	11.33	6843	3570	52.17%	51.91	8.70
Hb(g/㎗)	13.61	1.57	6840	3568	52.16%	51.91	8.70
FVC(%predicted)	104.76	14.17	4291	2135	49.76%	50.37	8.17
FEV1(%predicted)	112.27	16.62	4290	2134	49.74%	50.37	8.16
FEV1/FVC(predicted)	74.89	1.77	4291	2135	49.76%	50.37	8.17

Table 3.3 Summary of B_0 and B_1 of 16 traits

TRAIT	B_0				B_1			
	MEAN	**SD**	**MIN**	**MAX**	**MEAN**	**SD**	**MIN**	**MAX**
HEIGHT	159.906	8.724	130.241	187.866	-0.060	0.139	-2.168	0.747
WAIST	83.743	8.480	58.333	121.591	0.184	0.692	-4.968	6.904
WEIGHT	62.860	9.931	30.532	105.355	-0.094	0.480	-3.739	2.657
BMI	24.531	2.992	14.197	38.831	-0.019	0.185	-1.486	1.048
log(HbA1c)	1.737	0.107	1.256	2.441	0.002	0.011	-0.093	0.157
log(GLU0)	4.538	0.149	4.260	5.733	0.012	0.018	-0.166	0.171
Log(TG)	4.834	0.423	3.584	7.189	-0.012	0.057	-0.409	0.412
LDL	120.111	25.757	11.833	281.590	0.193	4.065	-29.218	28.549
log(HDL)	3.782	0.193	3.100	4.567	-0.001	0.024	-0.211	0.135
TCHL	194.208	28.114	97.986	343.106	-0.120	4.468	-34.599	29.468
log(SBP)	4.768	0.114	4.461	5.156	-0.001	0.017	-0.122	0.086
DBP	78.252	8.239	50.639	111.556	-0.259	1.358	-12.330	8.066
Hb	13.695	1.370	7.764	18.889	0.022	0.147	-1.641	1.468
FVC	104.541	13.466	46.629	162.844	-0.090	2.478	-13.065	13.685
FEV1	111.128	16.295	38.951	184.532	-0.239	2.575	-16.022	15.620
FEV1/FVC	73.945	1.809	67.654	78.000	-0.213	0.127	-1.246	1.244

Table 3.4 Comparison between heritability of reference and cross-sectional average (B_0).

Trait	Reference (Yang *et al.*)			Cross-sectional average (B_0)		
	Sample Size	h^2_{snp} (SE)	P-value	Sample Size	h^2_{snp} (SE)	P-value
HEIGHT	7170	0.316 (0.042)	2.10E-15	6823	0.318 (0.041)	1.67E-16
WAIST	7163	0.105 (0.040)	4.10E-03	6835	0.278 (0.041)	5.02E-09
WEIGHT	7168	0.161 (0.040)	1.80E-05	6822	0.196 (0.040)	2.05E-07
BMI	7168	0.147 (0.041)	1.10E-04	6822	0.188 (0.040)	6.66E-07
HbA1c	7168	0.126 (0.040)	5.80E-04	6329	0.176 (0.044)	2.79E-05
GLU0	7006	0.112 (0.041)	2.90E-03	6728	0.152 (0.041)	0.0001042
TG	7169	0.216 (0.041)	1.50E-08	6840	0.192 (0.040)	4.42E-07
LDL	6963	0.134 (0.041)	3.80E-04	6840	0.222 (0.040)	5.18E-09
HDL	7169	0.172 (0.041)	8.50E-06	6840	0.241 (0.041)	8.91E-10
TCHL	7169	0.156 (0.040)	2.30E-05	6840	0.265 (0.041)	3.90E-12
SBP	7169	0.250 (0.041)	5.80E-11	6843	0.150 (0.039)	3.28E-05
DBP	7170	0.171 (0.041)	6.70E-06	6843	0.178 (0.039)	8.31E-07
Hb	7169	0.064 (0.039)	4.90E-02	6840	0.195 (0.041)	4.93E-07
FVC(%pred)	7009	0.226 (0.043)	2.10E-08	4291	0.107 (0.062)	0.03672
FEV1(%pred)	7007	0.134 (0.041)	4.20E-04	4290	0.119 (0.062)	0.0234
FEV1/FVC(pred)	7011	0.148 (0.041)	1.00E-04	4291	0.136 (0.063)	0.01394

Figure 3.2 Population structures identified via a multidimensional scaling (MDS) plot. This plot shows that the analyses (KARE) are not affected by population stratification. AFR, AMR, EAS, EUR, and SAS indicate African, Ad Mixed American, East Asian, European, and South Asian populations, respectively, from the 1000 Genomes Project

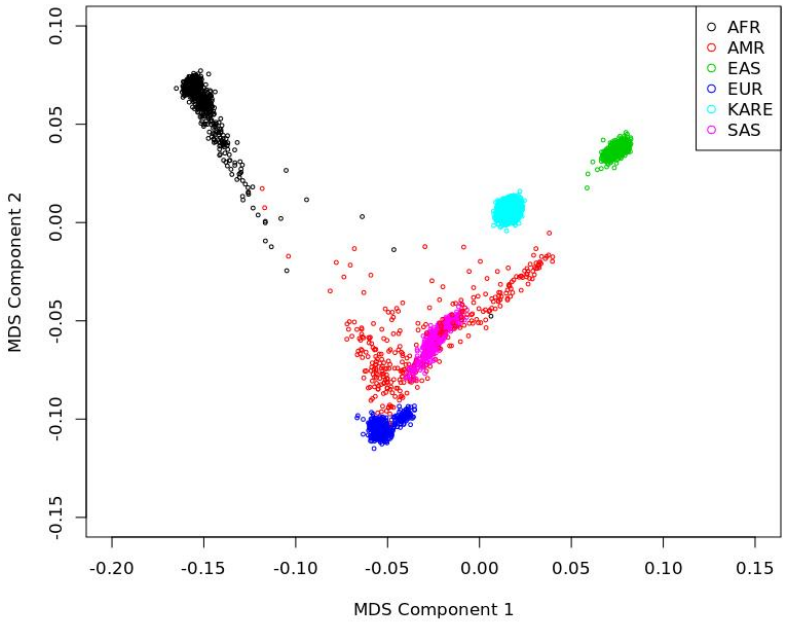


Figure 3.3 Single-nucleotide polymorphism heritability estimates of 16 traits with B_0 as the response. Error bars correspond to standard error values. The values above the error bar are *P*-values and false discovery rate (FDR; bold).

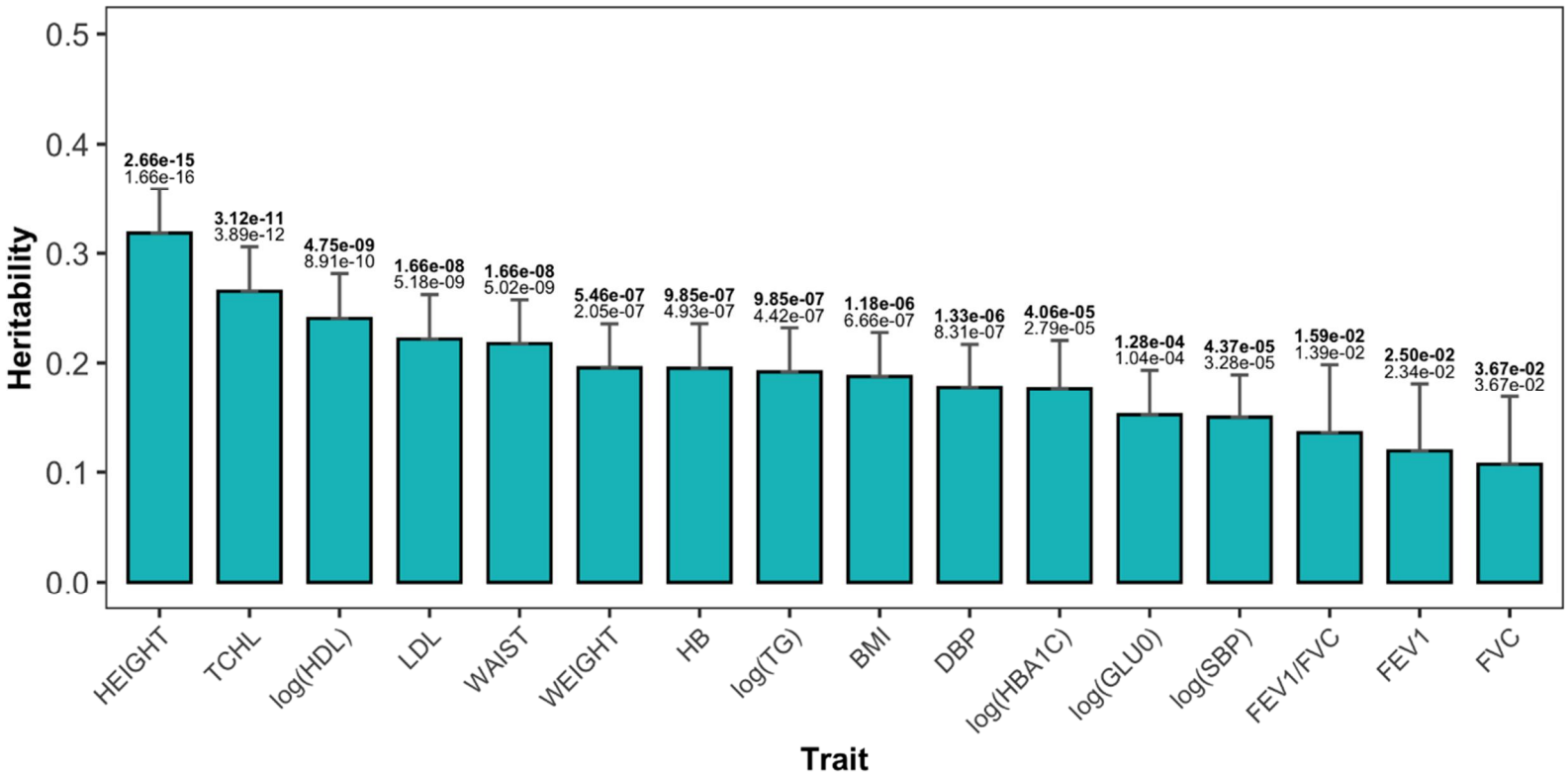


Figure 3.4 Comparison of heritability of cross-sectional average and reference paper. Red line is heritability estimation of reference paper and black solid line is estimated heritability of cross-sectional average, the black dashed line is Mean of estimated heritability of each period.

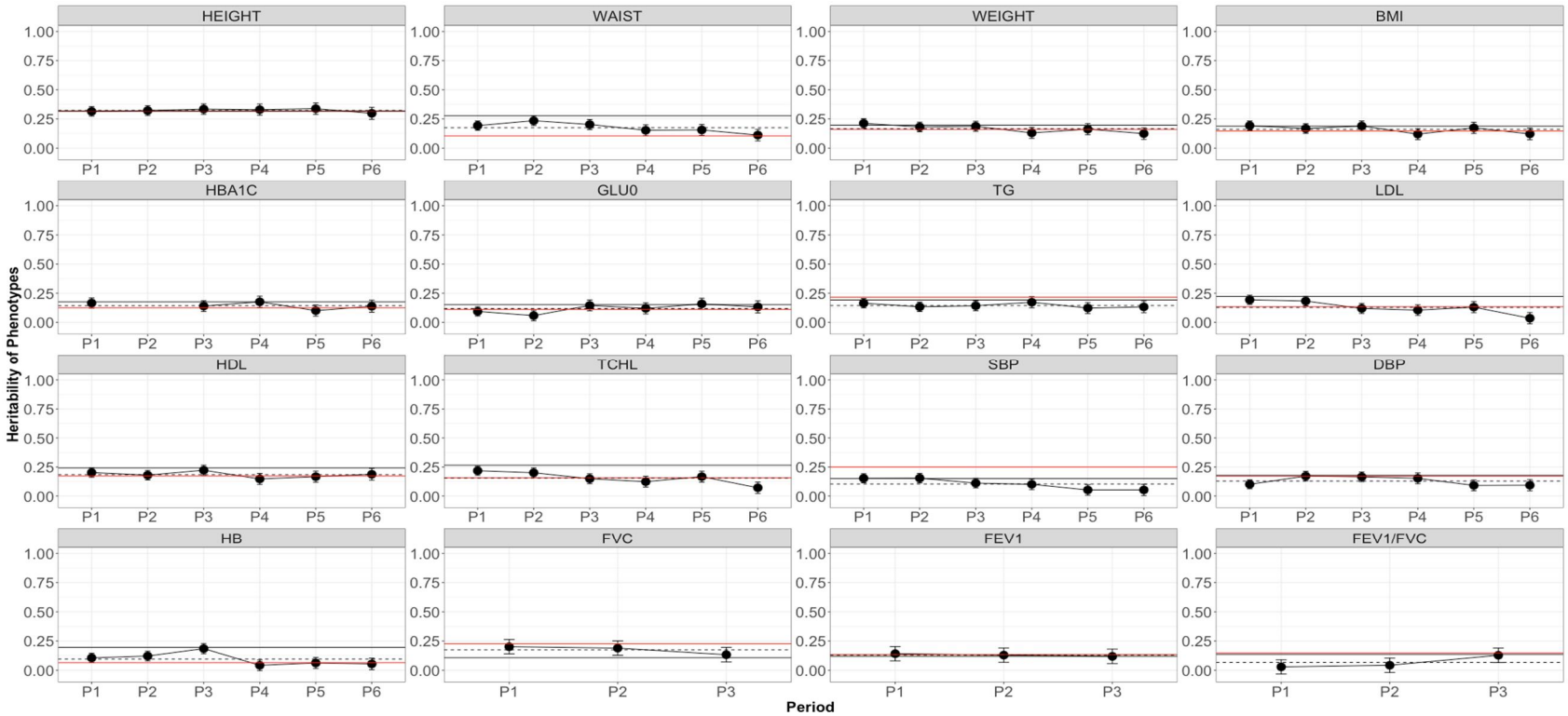


Figure 3.5 Single-nucleotide polymorphism heritability estimates of 16 traits with B_1 as the response. Error bars correspond to standard error values. The values above the error bar are *P*-values and the false discovery rate (FDR; bold), and "*" indicates significant findings at an FDR of 0.05

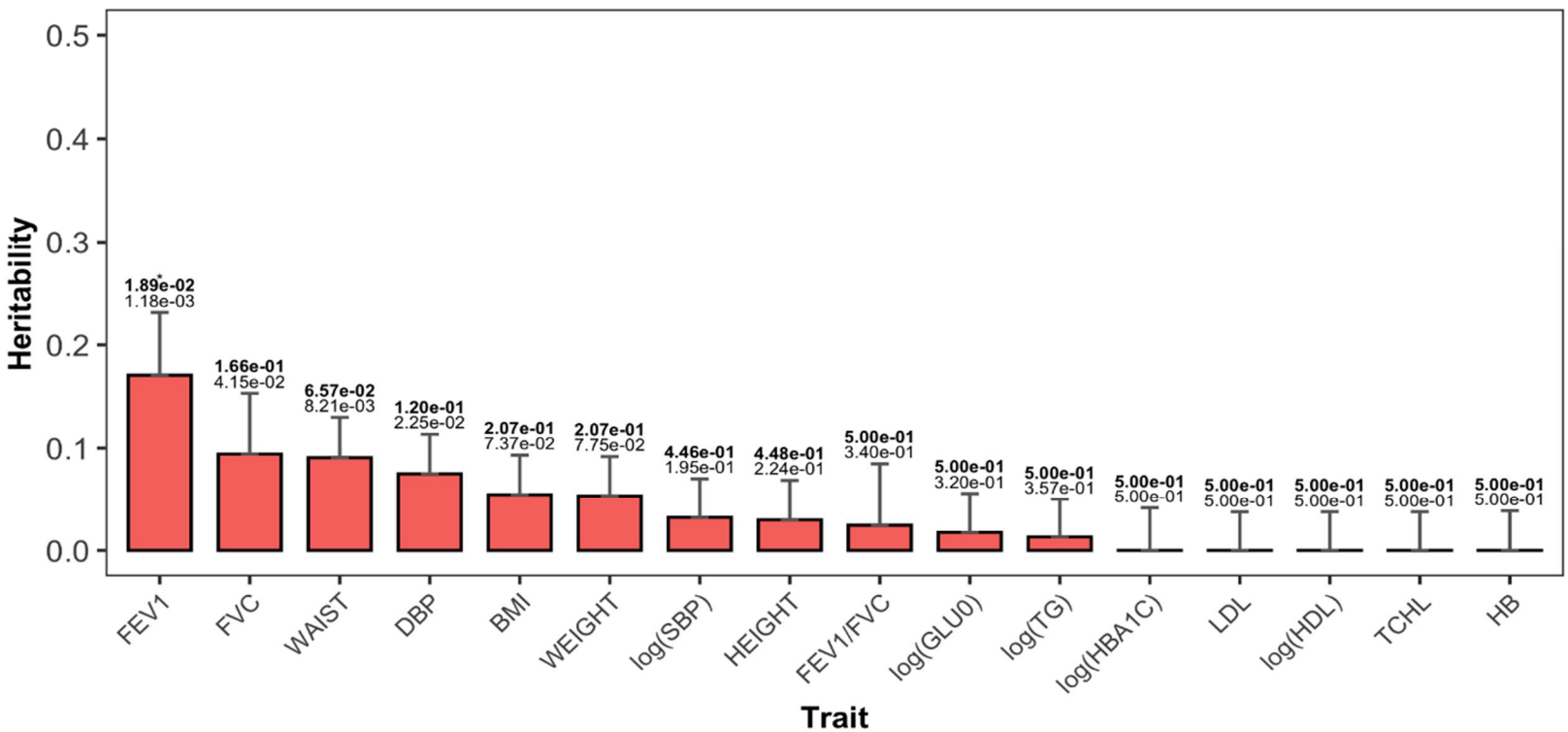


Figure 3.6 Single-nucleotide polymorphism heritability estimates of FEV1 based on chromosomes with B_1 as the response. Error bars correspond to standard error values. The values above the error bar are *P*-values and false discovery rate (FDR; bold).

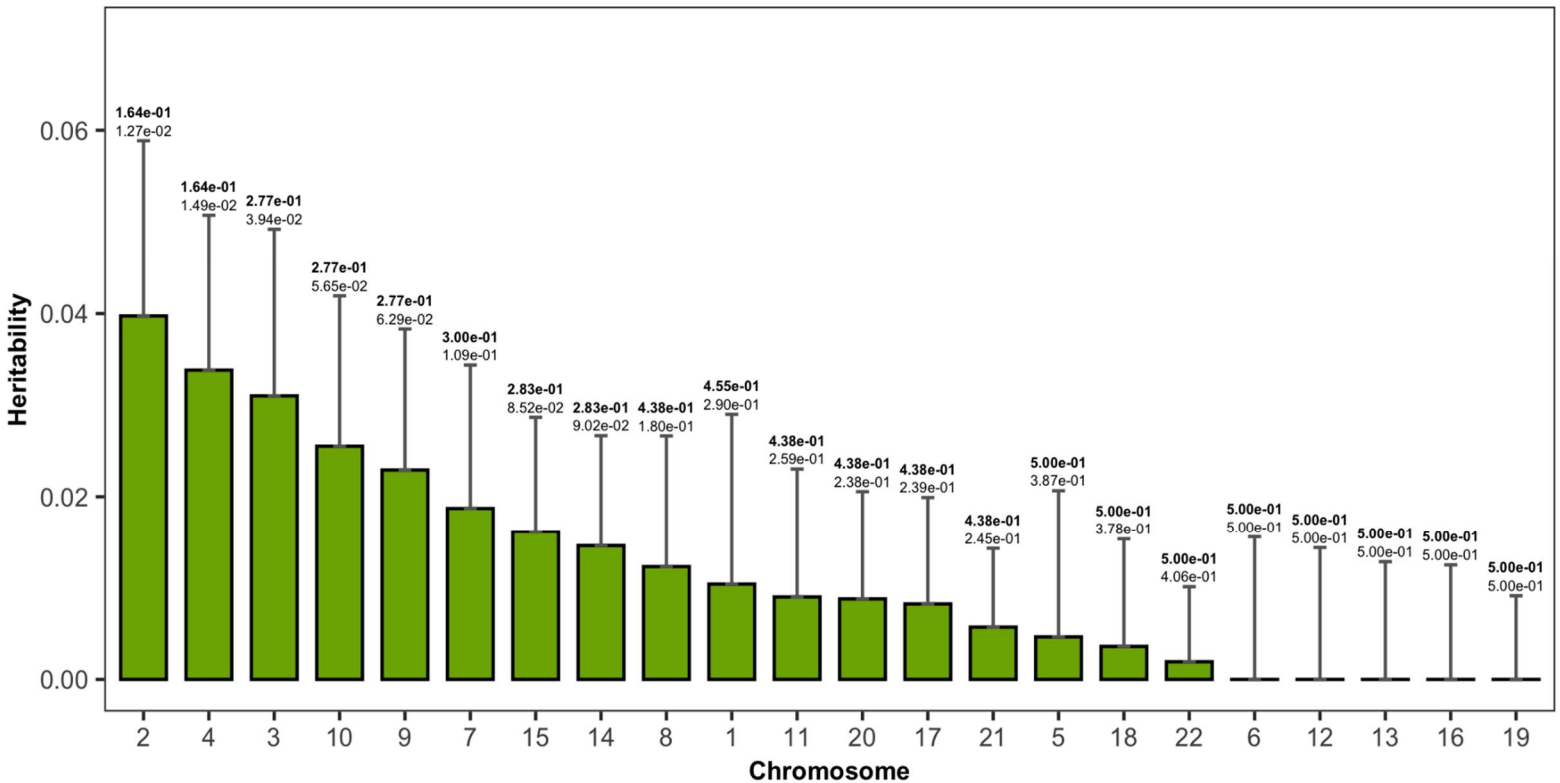
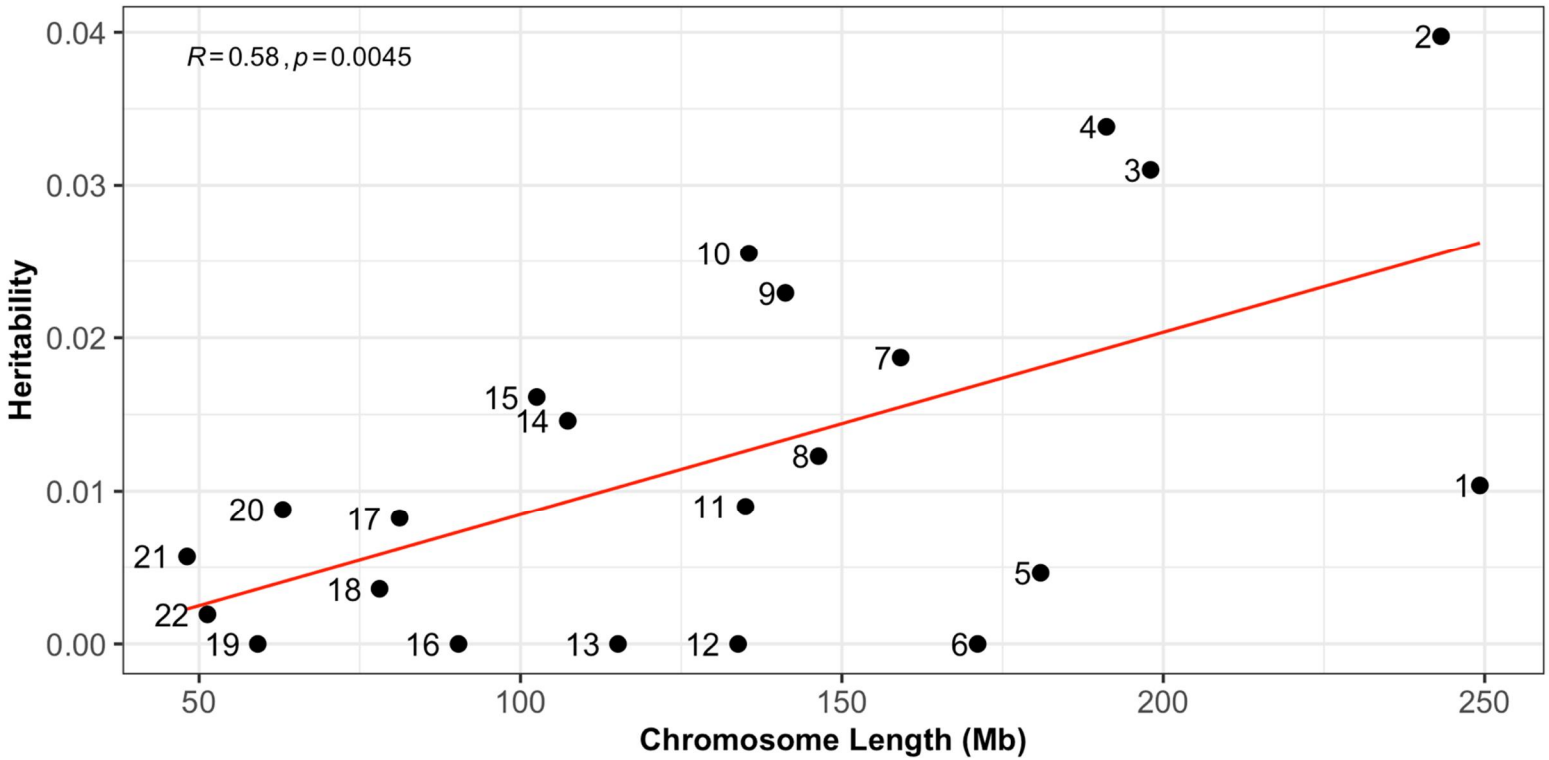


Figure 3.7 Correlation between chromosome length and estimated heritability. There was a significant positive correlation between chromosome length and heritability in FEV1



3.3.2 Genome-wide association studies

B_0 and B_1 were considered as responses for the GWAS. Tables 3.5 and 3.6 show genome-wide significant SNPs at a significance level of 1×10-7. Table 4 shows that SNPs have relatively lower P values for log(TG) and log(HDL) than any other trait. The most significant variant for log(TG) is rs6589566 in *ZPR1* with a P-value of 7.9×10-38, the lowest P-value among all 16 traits. Furthermore, *ZPR1* is associated with TG (Coram et al. 2013). The most significant variant of log(HDL) is rs16940212 with a P-value of 2.08×10-18 in *ALDH1A2*, which is associated with HDL (Spracklen et al. 2017). Certain other significant variants are significantly associated with proximal genes and with traits assessed herein. The variant rs180349 (P=8.86×10-35) of log(TG) is proximal to *BUD13*, which is associated with TG (Hoffmann et al. 2018). The variant rs17482753 (P=3.199×10-18) is proximal to *LPL*, which is strongly associated with HDL (Hoffmann et al. 2018). Herein, we also detected some de novo variants including rs4922117 (P=2.13×10-15) of log(HDL) and rs2335418 (P=3.2×10-9) of LDL, which were previously unknown; however, both their proximal genes *LPL* and *HMGCR* are significantly associated with each trait (Hoffmann et al. 2018). The Manhattan Plot and QQ plot for the model with B_0 as the response are provided in the Figures 3.8 and 3.9.

Table 5 shows the results of GWAS of B_1. Based on the results, rs2272402 (*SLC6A1*) is the most significant variant both in FEV1 (P=1.22×10-8) and FVC (P=1.40×10-9), and the *SLC6A1* enhancer is

associated with lung function. Other variants, including rs7209788 (*NARF*, P=3.36×10-7) for FEV1 and rs2668162 (*FAM19A1*, P=6.18×10-7) for FVC, have P-values less than the 1×10-6 threshold. We also found that rs4789777(*HEXDC*, P=4.599×10-6) is highly correlated with rs7209788 of FEV1. The Manhattan Plot and QQ plot for the model with B_1 as the response are provided in Figure 3.10 and 3.11.

Table 3.5 Results of the genome-wide association study with B_0 as the response. Only the significant variants with *P*-values less than 1×10^{-7} in each trait are included.

TRAIT	SNP	CHR	BP	A1	A2	GP	GENE	MAF	HWE_P	BETA	P
log(GLU0)	rs1799884	7	44229068	A	G	upstream	GCK	0.1872	0.9051	0.01889	5.62E-09
log(GLU0)	rs7754840	6	20661250	C	G	intronic	CDKAL1	0.478	1	0.01488	6.25E-09
Hb	rs5756505	22	37467354	C	G	intronic	TMPRSS6	0.4979	0.9229	0.1125	2.61E-13
Hb	rs3768751	2	46346716	G	A	intronic	PRKCE	0.1796	1	-0.113	1.79E-08
log(HbA1c)	rs7754840	6	20661250	C	G	intronic	CDKAL1	0.478	1	0.01254	5.41E-11
log(HDL)	rs16940212	15	58694020	T	G	intergenic	ALDH1A2	0.3414	0.9786	0.02988	2.08E-18
log(HDL)	rs17482753	8	19832646	T	G	intergenic	LPL(dist=7876),SLC18A1(dist=169720)	0.1241	0.1333	0.04232	3.20E-18
log(HDL)	rs4922117	8	19852586	G	A	intergenic	LPL(dist=27816),SLC18A1(dist=149780)	0.2077	0.418	0.0316	2.13E-15
LDL	rs599839	1	109822166	G	A	downstream	PSRC1	0.06456	0.1925	-5.886	1.53E-11
LDL	rs12654264	5	74648603	T	A	intronic	HMGCR	0.4758	0.8085	-2.625	1.41E-09
LDL	rs2335418	5	74603479	G	A	intergenic	ANKRD31(dist=70776),HMGCR(dist=29514)	0.4232	0.5689	-2.6	3.20E-09
LDL	rs4045166	5	74909446	G	C	intronic	ANKDD1B	0.3326	0.3137	2.727	3.55E-09
LDL	rs10942739	5	74786083	T	C	intronic	COL4A3BP	0.3325	0.276	2.709	4.61E-09
LDL	rs688	19	11227602	T	C	exonic	LDLR	0.136	0.797	3.402	6.63E-08
TCHL	rs599839	1	109822166	G	A	downstream	PSRC1	0.06456	0.1925	-6.822	1.19E-12
TCHL	rs780092	2	27743154	G	A	intronic	GCKR	0.3248	0.2948	-3.365	4.63E-11
TCHL	rs17321515	8	126486409	T	C	intergenic	TRIB1(dist=35762),LINC00861(dist=448358)	0.4425	0.04178	2.782	4.20E-09
TCHL	rs1881396	2	27844601	G	T	UTR3	ZNF512	0.3313	0.8699	-2.793	3.18E-08
TCHL	rs6861279	5	74919409	T	C	intronic	ANKDD1B	0.3386	0.1772	2.784	3.99E-08
TCHL	rs6734059	2	27808154	C	T	intronic	ZNF512	0.3357	0.8921	-2.724	6.36E-08
log(TG)	rs6589566	11	116652423	C	T	intronic	ZPR1	0.2169	0.3545	0.1113	7.90E-38
log(TG)	rs180349	11	116611827	A	T	intergenic	LINC00900(dist=980909),BUD13(dist=7059)	0.2265	0.752	0.1065	8.86E-35
log(TG)	rs10503669	8	19847690	T	G	intergenic	LPL(dist=22920),SLC18A1(dist=154676)	0.1207	0.003444	-0.08902	1.63E-16
log(TG)	rs780094	2	27741237	C	T	intronic	GCKR	0.4626	0.9806	-0.05656	2.60E-15
log(TG)	rs7115242	11	116908283	T	C	intronic	SIK3	0.2796	0.133	0.05987	8.66E-14

Table 3.6 Results of the genome-wide association study with B_1 as the response Only the variants with *P*-values less than 1×10^{-7} are included. The more variants under suggestive threshold (*P*-values less than 1×10^{-5}) are listed

TRAIT	SNP	CHR	BP	A1	A2	GP	GENE	MAF	HWE_P	BETA	P
FEV1	rs2272402	3	11075461	A	G	intronic	SLC6A1	0.07363	0.1473	-0.5823	1.22E-08
FVC	rs2272402	3	11075461	A	G	intronic	SLC6A1	0.07363	0.1473	-0.595	1.40E-09

Figure 3.8 Manhattan Plot with B_0 as response.

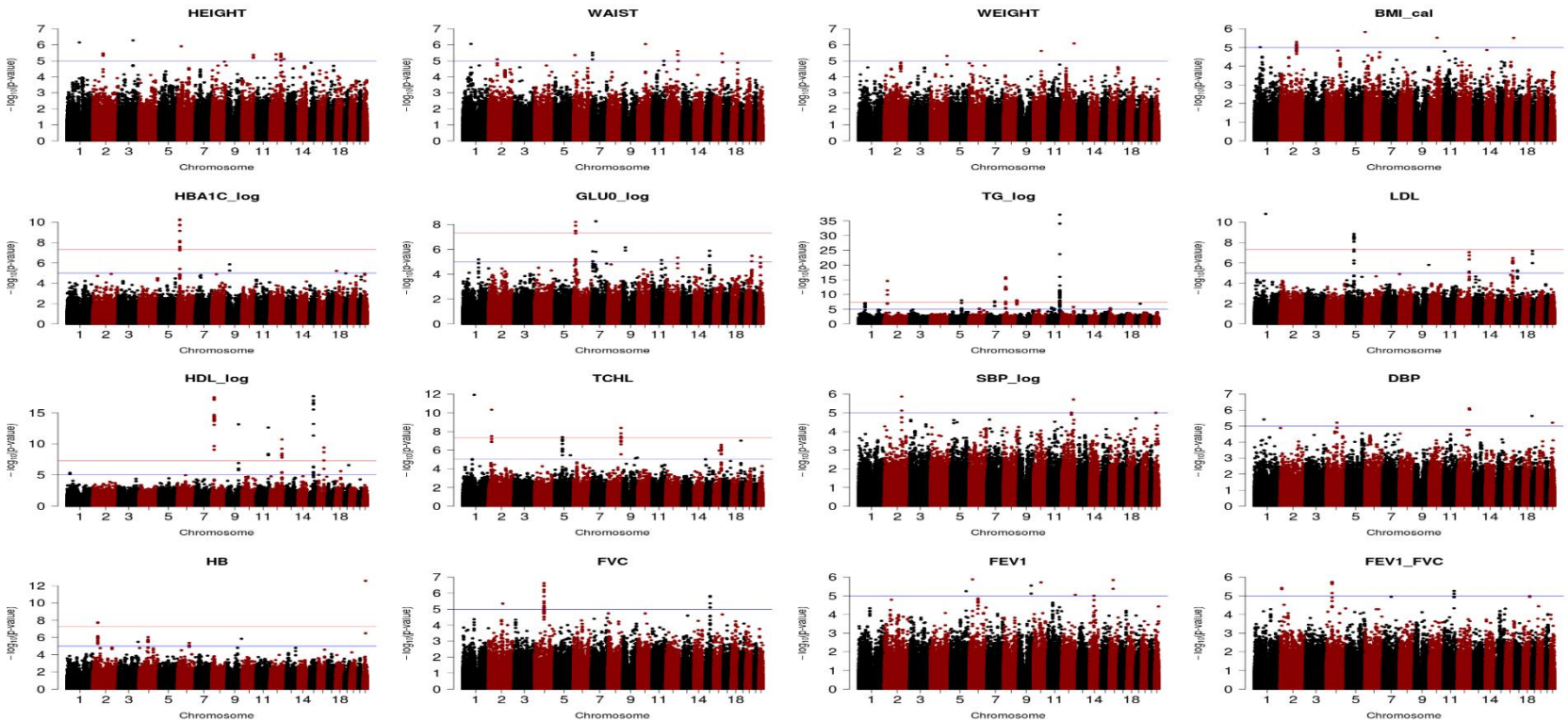


Figure 3.9 QQ plot with B_0 as response

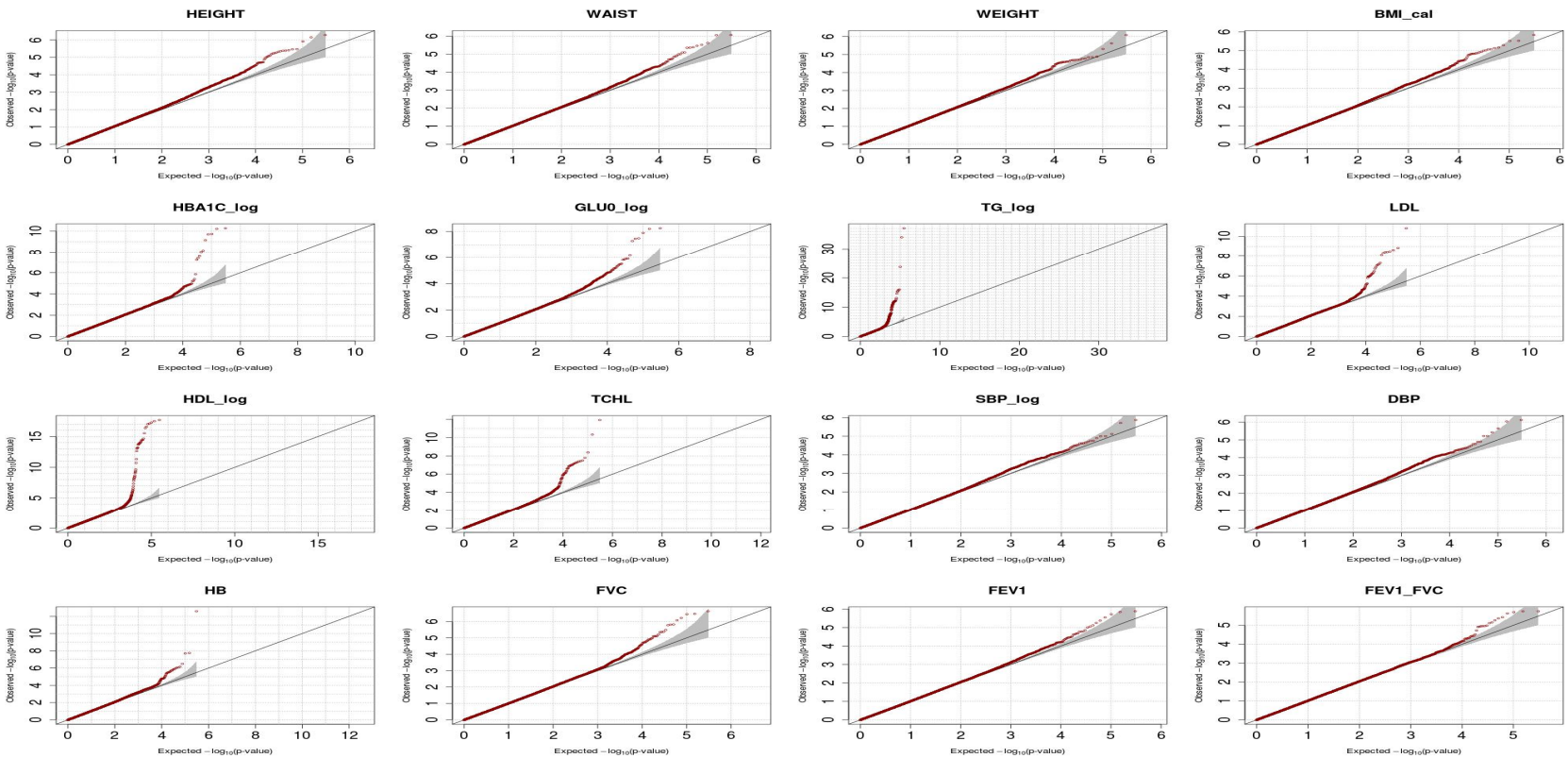


Figure 1.10 Manhattan Plot with B_1 as response.

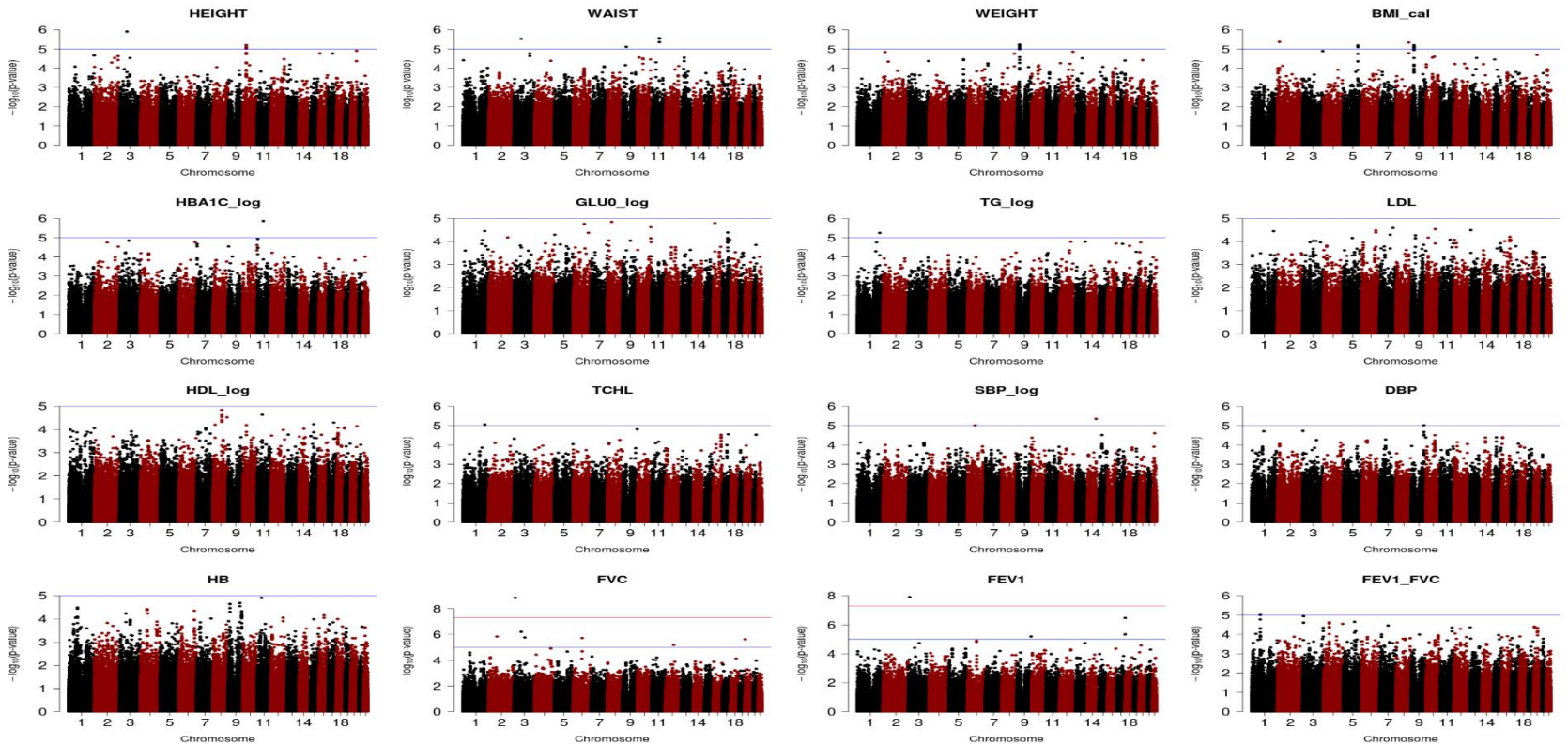
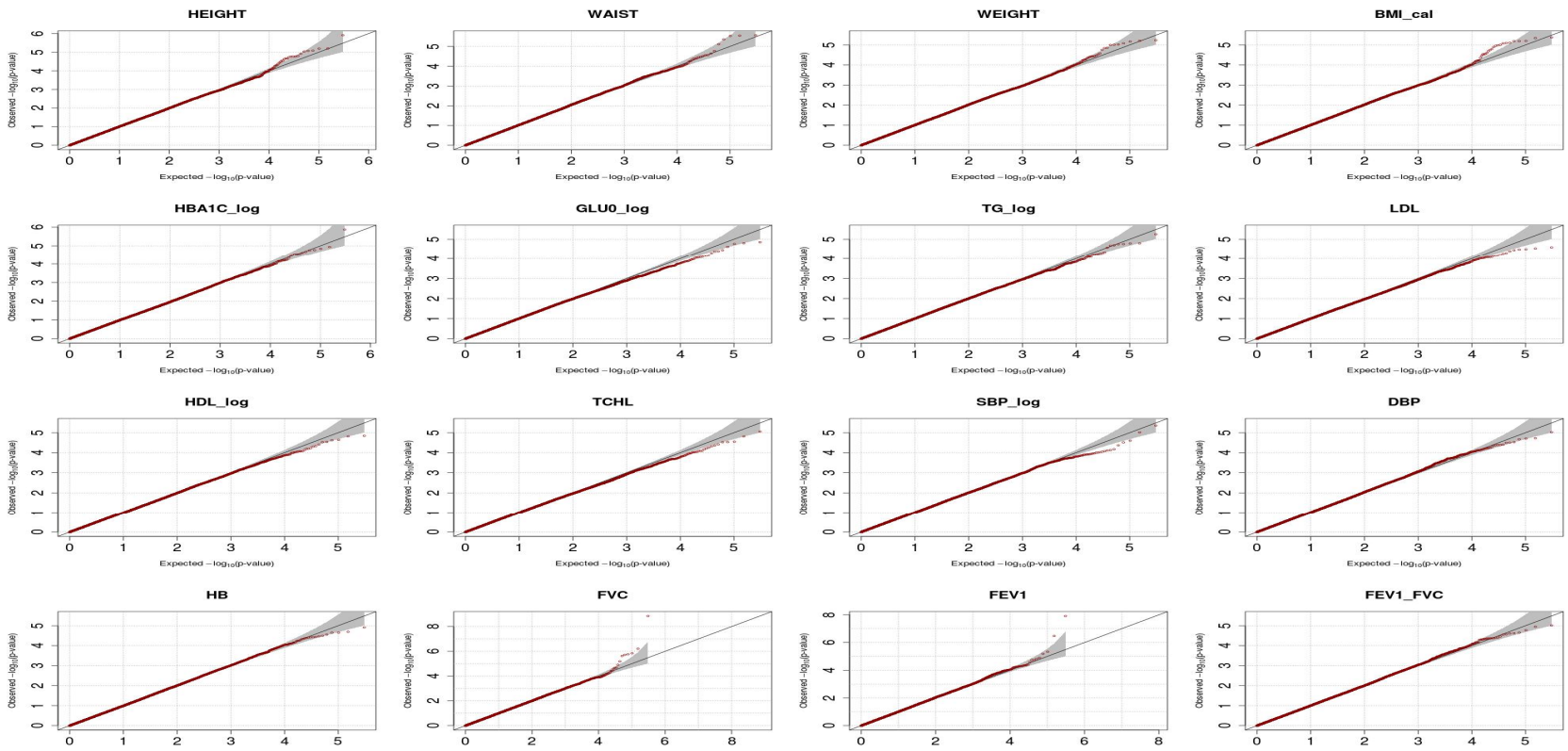


Figure 3.11 QQ plot with B_1 as response.



3.4 Discussion

In this study, SNP-based heritability estimates of 16 phenotypic traits were estimated longitudinal data with 10-year follow-up of the KARE cohort. The GCTA tool was used with a two-stage approach to determine the heritability estimate of phenotypic mean and longitudinal changes in each trait. Moreover, chromosomal heritability estimates were determined and GWAS analysis were performed using the same approach. Overall, heritability estimates within the population-based cohort including KARE are potentially lower than those of pedigree or twin studies for all 16 traits, regardless of whether the response is B_0 that phenotypic mean of traits or B_1 which stands for the changes by time of traits. For example, the heritability of height herein was estimated to be approximately 0.318 with B_0 as the response, which is lower than the conventional heritability estimate of height of approximately 0.8 based on the assumption-free model (Visscher et al. 2006). In the case of TCHL and LDL, each heritability estimate was determined to be 0.265 and 0.22, respectively, which are also lower than the heritability estimates of 0.67 and 0.69 for TCHL and LDL, respectively, on familial and pedigree analysis (van Dongen et al. 2013). The underlying reason may be explained on the basis of the missing heritability, which describes the difference in values between heritability estimated via GWAS and via familial studies (Sandoval-Motta et al. 2017). However, systemic inflation of estimated heritability estimates of polygenic phenotypes in familial studies may be confounded owing to a shared environment or environment-dependent

genetic effects (Robinson et al. 2017). Therefore, the population-based design similar to that of the present study potentially represent the average genetic effects regardless of various confounding environmental factors.

Based on the present B_0 and B_1 model, the heritabilities of B_1 are markedly lesser than those of B_0, indicating that most of the genetic variance of traits are not temporally influenced. Here, B_0 was not determined from the baseline measurements of traits but rather the average values of repeated measurements to yield a more robust and reasonable result. If baseline measurement and longitudinal changes (B_1) calculated from those were considered responses during the estimation of heritability, the estimate would have been potentially inaccurate owing to the correlation between baseline and B_1 values. Moreover, by applying a regression model to estimate the average B_0 and longitudinal changes B_1, we an independent association was observed between B_0 and B_1. Thus, more reliable estimation of heritability could be achieved.

On GWAS, the two-stage model elucidated significant variants associated with the traits and their changes in the longitudinal data. We confirmed several proven variants and identified some other significant unreported variants. In the case of the B_0 model, rs4922117 (P=2.13×10-15) of log (HDL) and rs2335418(P=3.2×10-9) of LDL were both unreported; however, their proximal genes *LPL* and *HMGCR* respectively, were significantly associated with each trait (Hoffmann et al. 2018). Furthermore,

unreported SNPs, such as rs180349, including non-coding SNPs with a significant P-value for TG, are proximal to *BUD13*, which is strongly associated with TG in the reference study (Hoffmann et al. 2018). Variants including rs17482753 also had significant P-values and was proximal to LPL, which is strongly associated with the HDL trait (Hoffmann et al. 2018). In the model with B_1 as response variable, rs2272402 (*SLC6A1*, P=1.22×10-8) was significantly associated for FEV1 and FVC. The *SLC6A1* enhancer is associated with pulmonary function. Therefore, the present results are consistent with previous findings regarding genes associated with each phenotype.

Among the 16 phenotypic traits in this study, only FEV1 displayed longitudinally significant heritability herein (Figure 3.5), thus reliably reflecting the physiological state of the lungs and airways and acting as a predictor of morbidity and mortality in the general population; FEV1 is also widely used to define chronic obstructive pulmonary disease (COPD) (Young, Hopkins, and Eaton 2007). Lung function develops in early life, peaks at a specific time point in early adulthood, and subsequently declines with age. Therefore, the decline of lung function in middle-aged and older individuals is suggested to be heritable in the general population (Gottlieb et al. 2001). However, longitudinal studies on FEV1 and FEV1/FVC have suggested several significant genetic regions that markedly differ from the numerous genetic variants associated with lung function, with FEV1 being estimated at a single time point (John et al. 2017, Tang et al. 2014). Hence, gene-

environment interactions and significant genetic heterogeneity in lung function have been observed in diseases such as asthma or COPD (Imboden et al. 2012, Hansel et al. 2013). Accordingly, the present study included the middle-aged general population with similar environmental exposure without specific lung diseases, thus suggesting the intact FEV1 decreased due to aging. Therefore, the present results show that FEV1 has significant SNP heritability for longitudinal changes (FDR=0.0012 for FEV1).

This study has several limitations. First, the analysis of new variants in the present GWAS was not replicated for other cohorts. Second, the two-stage approach is statistically inefficient, even though it is computationally fast. However, the sample size was very large, which hopefully minimized this problem. Furthermore, we considered subjects with at least three or more measurements, which potentially minimized statistical power loss. Third, gene-environment interactions were not analyzed, although the estimation of random effects in the mixed model was elusive. Fourth, GCTA itself has limitations for reasons such as data overfitting and skewed singular values (Kumar et al. 2016). Though this study optimized parameters to attain accurate results using GCTA, the sample size might have resulted in certain variations in comparison with other large studies. Furthermore, the issue regarding missing heritability was inevitable to an extent because the Affymetrix genotypic array represents only common variants for SNPs, while rare genetic SNP variants were not included herein (Bandyopadhyay, Chanda, and Wang 2017).

Despite the aforementioned limitations, this study elucidated heritability estimates via a two-stage approach using a mixed model in GCTA and a GWAS, which provides a reasonable and easy method to estimate heritability in longitudinal data and potentially assess both heritability of the phenotypic mean and longitudinal changes through several periods. Essentially, our results show that significant SNP heritability is objectively confirmed for longitudinal changes in lung function decline (i.e., FEV1) in comparison with other health-related indices. Therefore, there should be more genetic studies on longitudinal FEV1 decline in the middle-aged general population and chromosome 2, which attributes the most in genetic variance should be encouraged.

Chapter 4

Heritability analyses reveal the significant effect of SNPs on lung function decline rate

4.1 Introduction

Lung function is an important human trait, once it is damaged, it hard to reverse the condition completely, and the impaired lung function could even predict patient's future morbidity and mortality (Young, Hopkins, and Eaton 2007). Generally, spirometry is used to assess the lung function by measuring the volume or flow of air that can be inhaled and exhaled (Miller et al. 2005). It is helpful in screening general respiratory health, but on its own, it is not directly used to an aetiological diagnosis. In clinical practice, force vital capacity (FVC) and forced expiratory volume (FEV) which are two

important measurements of spirometry, have been used to evaluate physiologic status of respiratory disease, measure the effect of disease on pulmonary function and assess prognosis of pulmonary disease such as asthma, pulmonary fibrosis, cystic fibrosis and COPD (Miller et al. 2005).

Lung function in healthy persons, typically, reaches a peak level at their early ages, and then a steady decline would be followed in the rest of life. However, there is a range of lung function trajectories throughout the whole process. As an example, the person who had a failure to reach the predicted level of peak lung function in early age, would have a higher prevalence and an earlier incidence than those who did not (Agusti and Hogg 2019). Thus, longitudinal and trajectory perspective analysis is important for understanding the mechanism of lung function.

Genetic association studies have been widely applied to identify genomic regions to provide useful insights into biological mechanisms of complex diseases (Sakornsakolpat et al. 2019). Recently, genome-wide association studies (GWAS) have identified numerous genetic variants associated with lung function in cross-sectional analysis (John et al. 2017, Wilk et al. 2009, Soler Artigas et al. 2011, Loth et al. 2014). However, no genetic variants have yet been associated with rate of decline in lung function at stringent genome-wide significant level (John et al. 2017). There are some other researchers reported both cross-sectional lung function and annual decline rates in lung function are heritable by using family data (Gottlieb et al.

2001), suggesting that there is still scope for further discoveries (John et al. 2017).

Lately, estimating SNP-based heritability have proven to a powerful tool for investigating the genetic architecture of common diseases among independent population-based cohorts. The estimation is based on restricted maximum likelihood estimation (REML) in the linear mixed model (Yang et al. 2010) framework and is applied by several popular tools (Weissbrod, Flint, and Rosset 2018, Yang, Lee, et al. 2011, Speed et al. 2017). One of the most used tools, is genetic complex trait analysis (GCTA) tool (Yang, Lee, et al. 2011), which first calculates the genetic relatedness matrices (GRM) between individuals, then estimate the proportion of all the single nucleotide polymorphisms (SNPs) variance in the phenotypic variance. Previous study of lung function by Zhou et al, applied GCTA to estimate heritabilities of FEV1 and FVC/FVC in the non-Hispanic whites, were both about 37%, which consistent with estimates from family-based studies (Zhou et al. 2013).

In this study, we estimated SNP heritability of 12 most common parameters measured in spirometry, as forced vital capacity (FVC), forced expiratory volume (FEV) in one second, forced expiratory flow 25-75% (FEF 25-75) and maximal voluntary ventilation (MVV), including their pre/post measures and percent predicted values with Korean longitudinal population-based cohort data, which measured biennially for 14 years (Table 4.1). To estimate SNP heritability both phenotypic average and annual change of 12 traits, here, we proposed and applied a two-stage approach that efficiently solve longitudinal

analysis problem. And also estimated how much of phenotypic variance were explained in smoking stratified groups of these 12 traits. At last, we calculated correlation between the phenotypic averages and annual change rates of those traits.

Table 4.1 Descriptive characteristics of data

Characteristics	ALL (Baseline)	Never Smoker	Ever Smoker
Sample Size, *n*	5104	3009	2095
Female, n (%)	2692 (52.74%)	2588 (86%)	104 (5%)
Age, yr (Mean±SD)	50.91±8.15	51.19±8.3	50.54±7.94
Height, *cm* (Mean±SD)	160.3±8.53	155.9±7.02	166.6±6.26
COPD, *n* (%)	55 (10%)	15 (0.5%)	40 (1.9%)

4.2 Methods

4.2.1 Study Population and Outcome Definition

We considered the Korean Genome and Epidemiology study (KoGES) (Cho et al. 2009) which consists of participants residing in Ansan (urban area) and Ansung (rural area) in the Gyeonggi Province of South Korea. KoGES was designed to investigate genetic, environmental and behavioral risk factors of common complex diseases in Koreans and cause of death with long-term follow-up (Kim, Han, and Ko 2017). The baseline survey was completed in 2001–2002, and 10,030 participants aged 40–69 years were recruited, and then biennial repeated surveys were conducted for the same participants for 14 years. We considered 8,768 participants (4,653 male, 4,115 female) who have both genotype and phenotype information.

4.2.1 Lung functions

Here, we focused on the most common lung function phenotypes as pre/post and % predicted bronchodilator spirometry which includes forced vital capacity (FVC), forced expiratory volume in one second (FEV1), the average forced expiratory flow during the mid (25-75%) portion of the FVC (FEF25-75%) and maximal voluntary ventilation (MVV), and these phenotypes accompanying with the basic information as sex, height and smoking history.

Lung function tests were performed by a skilled technician using a portable spirometer (Vmax-2130, Sensor Medics, Yorba Linda, CA, USA) according to standardized protocols of the American Thoracic Society(1995). All participants performed prebronchodilator spirometry test until completing at least three repeated measurements and an acceptable measure was determined when the differences between the largest and the next largest FVC and FEV1 values were within 0.15l. Calibration and quality control of spirometric examinations were also performed regularly based on American Thoracic Society guidelines (1995, Kim, Kim, et al. 2015, Shin et al. 2005).

4.2.2 Genotyping, quality-control and imputation

All patients were genotyped with Affymetrix Genome-Wide Human SNP array 5.01. For quality control (QC) tests, we excluded SNPs for which the missing genotype call rates were higher than 0.05, minor allele frequencies (MAFs) were less than 0.05, and Hardy-Weinberg equilibrium (HWE) *P*-values were less than 10^{-5}; additionally, participants with missing genotype call rates higher than 0.05 or with gender inconsistencies were excluded. QC was done with PLINK (Purcell et al. 2007) and ONETOOL (Song et al. 2018). After QC tests, 5,104 participants with 305,158 markers remained.

With remaining participants and SNPs, we conducted whole-genome imputation by using SHAPEIT2 and IMPUTE2 for pre-phasing data and

genotype imputation, respectively, and the 1000 Genome Phase 3 haplotype was used as reference panel. To maintain imputation quality, we filtered out the imputed SNPs which had less than 0.5 estimated imputation "info" score. The standard QC procedure was also applied for these SNPs, and 5,104 participants with 3,352,722 SNPs were analyzed for SNP heritability estimation.

4.2.3 Statistical analysis

Cross-sectional phenotypic averages and annual change rate for each subject were calculated with two-stage method. First, a simple linear regression model for subjects of the same period with the adjustment of age for each lung function traits. Each participant was measured up to 8 times and participants with at least three measurements were considered. We found that residual variances were heterogeneous among different time points, and the inverse of the residual variances were used as weights, and for trait k and time point j we considered the following linear regression for each subject i as follows:

$$y_{ijk} = \beta_{ik0} + \beta_{ik1}\left(age_{ij} - \overline{age}_i\right) + \varepsilon_{ijk}, \varepsilon_{ijk} \sim N\left(0, \frac{1}{w_{jk}}\sigma_{ik}^2\right) \qquad (1)$$

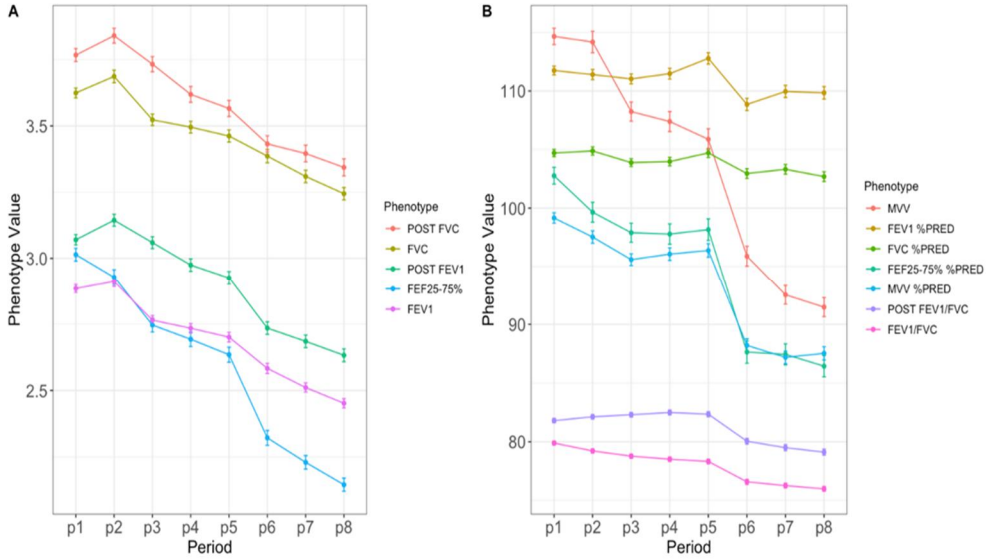
Here $\overline{age}_i$ indicates the mean of ages at the observed time points. In this model, β_{ik0} indicate the expected cross-sectional averages of subject i for trait k when he or she is $\overline{age}_i$ years old, and β_{ik1} is the annual change.

Then, the estimated values of β_{ik0} and β_{ik1} were inverse normal transformed and the SNP heritabilities, h_0^2 and h_1^2 for both were estimated with GCTA with restricted maximum likelihood method (Yang, Manolio, et al. 2011). For GCTA, $\overline{age}_i$ and sex were included as covariates. For the traits, as FVC, FEV1, FEV1/FVC, FEV25-75%, MVV, post FVC, post FEV1 and post FEV1/FVC, we also included height as covariate. We also estimated the best linear unbiased predictor of polygenic risk scores with GCTA with "--reml-pred-rand" option. The heritability estimation was also conducted for never smokers and ever smokers (consists of past smokers and current smokers) groups.

4.3 Results

From figure 4.1, we found most of 12 traits show decreasing trend through 8 periods. To identify the progressive effect of SNPs on lung function longitudinal change, first, we assessed the contribution of genetics to 12 lung function traits by estimating the SNP-based heritability for both mean and longitudinal change. Then, we estimated SNP heritability in never and ever smoking groups, separately. We also assessed the correlations between cross-sectional average and annual change rate of significant results from SNP heritability estimation.

Figure 4. 1 Mean value of 12 lung function traits in 8 periods. Since the range of 12 phenotypes were different, we divided them into A and B two groups. Most of the values shows decreasing trend through 8 periods.



4.3.1 Characteristics of Study Subjects.

Since the follow-up study performed 14 years biennially, the most of the participants visited center for 8 times including baseline, and there exists missing values in the collected data. In the view of potential bias and loss of power, we considered the data with participants who visited more than 3 times and performed analysis. The sample number of each trait are list in Table 2. And three of lung function traits, post FVC, post FEV1 and post FEV1/FVC, only Ansung data was available.

4.3.2 The SNP heritability of 12 lung function traits

To estimate the importance of genetic determinants of 12 lung function traits, we calculated the proportion of the variance after rank-based inverse normal transformation of B_0 and B_1 for each phenotype (Table 4.2). Figure 4.2(A) shows the estimation of SNP heritability with B_0 as the response in GREML model, and all the *P*-values of phenotypes are significant under FDR=0.05. The post FEV1/FVC has the largest h_0^2 ($h_0^2 = 0.325$, $P=1.16 \times 10^{-5}$), and the next is post FEV_1/FVC ($h_0^2 = 0.314$, $P=1.86 \times 10^{-5}$). FVC %predicted also gives relatively high values with $h_0^2 = 0.237$ ($P=5.36 \times 10^{-9}$). We also estimated the SNP heritability of each periods, and compared the mean of these SNP heritabilities to h_0^2 for all traits (Table 4.3 and Figure 4.3). We found h_0^2 are slightly higher than the mean of SNP heritablities of each period, it probably caused by the measurement error of

each period. Figure 4.2(B) shows the h_1^2, and it less than those for h_0^2. post FEV_1/FVC is the highest h_1^2 with value 0.176 (P=0.0099), which is followed by FEV_1/FVC with $h_1^2 = 0.158$ (P=4.91×10^{-5}). FEV_1 %predicted also has the significant h_1^2 with 0.105 (P=0.004).

For lung function traits with significant h_1^2 (FEV_1 %predicted, FEV_1/FVC, and post FEV_1/FVC), we calculate genetic correlations (ρ_g) between genetic components. Figure 4.4 shows phenotypic correlation between cross-sectional means and annual change rates of FEV_1 %predicted, FEV_1/FVC, and post FEV_1/FVC, and their correlations without any adjustment are 0.3, 0.24 and 0.22, respectively. Table 4.4 shows ρ_g and ρ_e. The former indicates the relative proportions shared between both genetic components for between corss-sectional means and annual change rates. The results show that around 50% or more of genetic components were significantly shared between them ($\rho_g = 0.5873$, P=0.0014 for FEV1% predicted; $\rho_g = 0.6279$, P=4.59×10^{-5} for FEV_1/FVC; $\rho_g = 0.466$, P=0.0219 for post FEV_1/FVC). Table 4.4 also shows the residual phenotypic correlations (ρ_g) between cross-sectional means and annual change rates. ρ_e indicates relative proportions of environmental variances shared between environmental variances for subject-specific means and annual change rates. Residual phenotypic correlations are much smaller than ρ_g ($\rho_e = 0.220$ for FEV1%predicted; $\rho_e = 0.117$ for FEV_1/FVC; $\rho_e = 0.155$ for post

FEV_1/FVC), and cross-sectional means and annual change rates may be affected by different environmental factors.

Table 4.2 Summary of cross-sectional averages and annual change rate of 12 lung function traits.

Traits	Sample Size	Cross-sectional average Mean (SD)	Annual decline rate Mean (SD)
FVC	5103	3.467 (0.832)	-0.036 (0.033)
FVC %PRED	5103	104.001 (12.974)	-0.22 (1.076)
FEV_1	5103	2.695 (0.649)	-0.04 (0.026)
FEV_1 %PRED	5103	111.028 (16.647)	-0.221 (1.218)
FEV_1/FVC	5103	77.977 (6.802)	-0.338 (0.512)
FEF25-75%	5104	2.604 (0.958)	-0.074 (0.06)
FEF25-75% %PRED	5104	95.052 (30.134)	-1.392 (2.201)
MVV	5099	103.775 (29.545)	-2.266 (1.997)
MVV %PRED	5099	93.195 (17.432)	-1.243 (1.827)
POST FVC	2706	3.62 (0.808)	-0.037 (0.025)
POST FEV_1	2707	2.932 (0.635)	-0.038 (0.021)
POST FEV_1FVC	2707	81.331 (5.74)	-0.214 (0.407)

Table 4.3 Comparison of estimated heritability of cross-sectional average and mean of estimated heritability of each period.

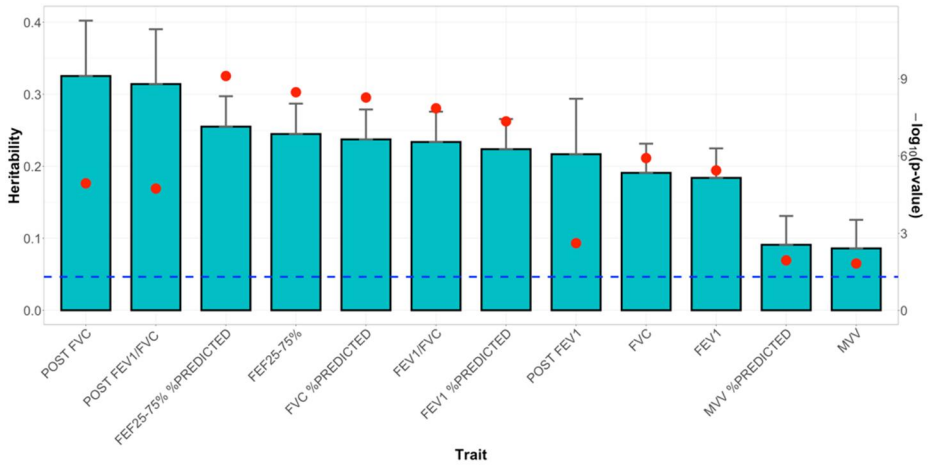
Trait	Type	Sample Size	σ^2_{pheno}	σ^2_{geno}	h^2_{snp}	S.E.
FVC	Beta0	6622	0.171	0.039	0.230	0.041
	Mean	5087.75	0.186	0.041	0.223	0.054
FVC %PRED	Beta0	6622	159.404	36.048	0.226	0.041
	Mean	5087.875	176.297	38.319	0.217	0.054
FEV1	Beta0	6622	0.134	0.027	0.203	0.041
	Mean	5087.25	0.141	0.027	0.196	0.054
FEV1 %PRED	Beta0	6622	242.160	50.577	0.209	0.042
	Mean	5087.75	257.619	52.368	0.203	0.054
FEV1/FVC	Beta0	6622	35.848	7.798	0.218	0.042
	Mean	5087.875	39.017	7.840	0.202	0.054
FEF25-75%	Beta0	6622	0.640	0.160	0.250	0.042
	Mean	5087.125	0.725	0.161	0.227	0.055
FEF25-75% %PRED	Beta0	6623	856.290	220.092	0.257	0.042
	Mean	5087.875	962.807	235.288	0.247	0.055
MVV	Beta0	6614	329.744	36.844	0.112	0.041
	Mean	5081.375	425.383	35.312	0.089	0.053
MVV %PRED	Beta0	6614	274.218	29.204	0.107	0.041
	Mean	5081.375	353.544	25.851	0.076	0.053
post FVC	Beta0	3489	0.166	0.053	0.322	0.077
	Mean	2748.5	0.172	0.052	0.300	0.099
post FEV1	Beta0	3490	0.122	0.027	0.222	0.077
	Mean	2748.625	0.126	0.026	0.204	0.099
post FEV1/FVC	Beta0	3490	27.237	7.234	0.266	0.076
	Mean	2748.75	29.149	7.626	0.263	0.099

Table 4.4 Genetic correlation of subject-specific mean and annual change rate

Traits	σ_{e0}^2	σ_{e0}^2	$\sigma_{e0}\sigma_{e1}\rho_e$	ρ_e	SE(ρ_e)	σ_{g0}^2	σ_{g1}^2	$\sigma_{g0}\sigma_{g1}\rho_g$	ρ_g	**SE(ρ_g)**	**P-value(ρ_g)**
FEV_1%PRED	0.6779	0.8798	0.1698	0.2199	0.0117	0.1954	0.1029	0.0833	0.5873	0.1713	0.0014
FEV_1/FVC	0.5927	0.8233	0.0814	0.1165	0.0121	0.1807	0.1538	0.1047	0.6279	0.1466	4.59E-05
post FEV_1FVC	0.5662	0.8178	0.1057	0.1553	0.0165	0.2593	0.1742	0.099	0.466	0.2156	0.0219

Figure 4.2 SNP heritability of 12 lung function traits. (A) SNP heritabilities of cross-sectional averages of 12 lung function traits. (B) SNP heritabilities of the annual change rate of 12 lung function traits. Error bars correspond to standard error values. The dot on the bar are P-values. Blue dash line indicates the 0.05 significant level. Red dot indicates significant findings at an FDR of 0.05.

(A)



(B)

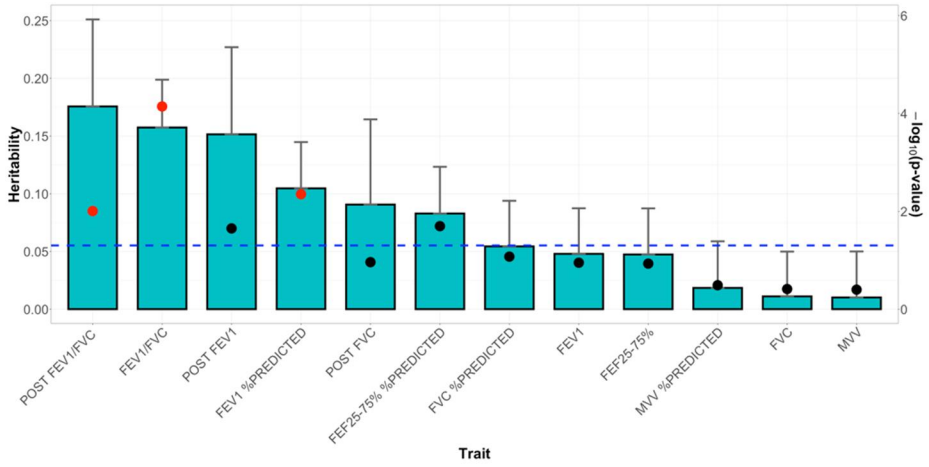


Figure 4.3 Comparison of estimated heritability of cross-sectional average and mean of estimated heritability of each period.

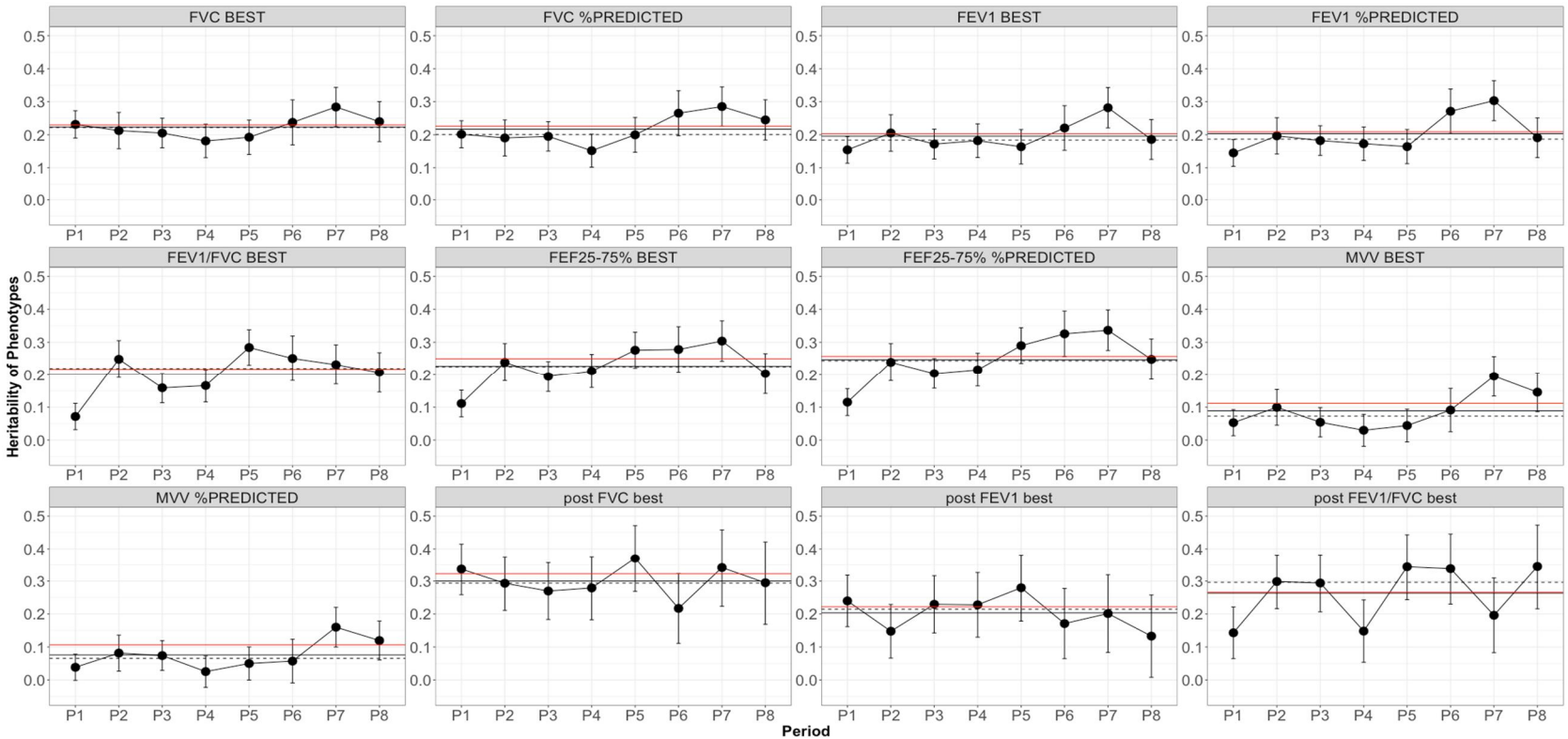
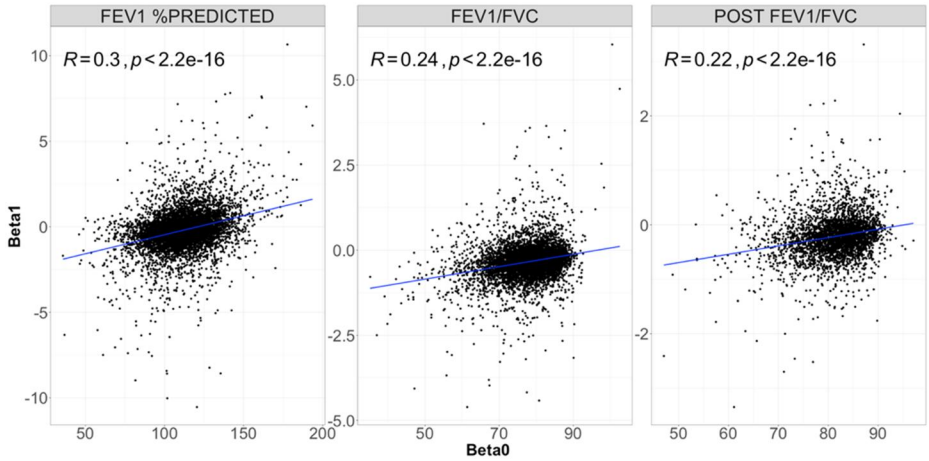
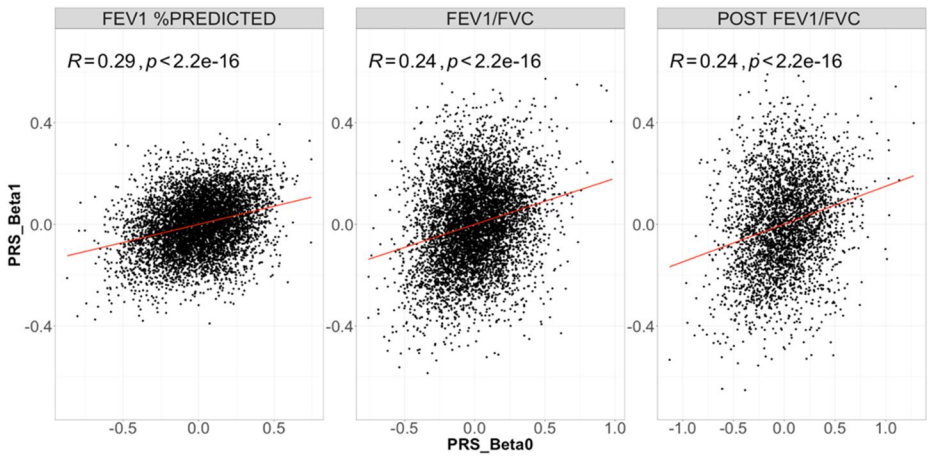


Figure 4.4 Correlation of cross-sectional averages and annual decline rates and their PRSs in FEV1 % predicted, FEV1/FVC and post FEV1/FVC. *(A)* Correlation between cross-sectional averages and annual change rate in FEV1 %predicted and FEV1/FVC. *(B)* Correlation between PRS of cross-sectional averages and PRS of annual change rate in FEV1 %predicted and FEV1/FVC.

(A)



(B)



4.3.3 Effect of smoking status on heritability of lung function traits

We stratified the data into two groups by smoking status, never and ever smoker (includes past and current smokers) groups. B_0 and B_1 were calculated in two groups separately (Table 4.5), and rank based inverse normal transformation were performed. For both groups, h_0^2 and h_1^2 were separately estimated (Figure 4.5), the estimated h_0^2 in never smoker groups are higher than those in ever smoker group except for FEV1/FVC, post FEV1/FVC and FVC %predicted. However, none of them except the never group of POST FEV_1/FVC is significant at the 0.05 significance level for h_1^2.

We also evaluate the heritability for SNP-by-smoking interaction ($h_{G\times S}^2$) for lung function traits with significant h_0^2 and h_1^2. All 12 lung function traits have significant h_0^2 but none of them have significant $h_{G0\times S}^2$ (Table 4.6). FEV_1/FVC, post FEV_1/FVC and FEV_1 % predicted have significant h_1^2 and $h_{G1\times S}^2$ were estimated for them. Table 4.6 shows that post FEV_1/FVC and FEV_1 % predicted achieve 0.05 significant level (P=0.02091 for FEV1; P=0.021158 for post FEV_1/FVC) and FEV_1/FVC is close to the significance level (P=0.079165).

Table 4. 5 Summary of cross-sectional averages and annual change rate of 12 lung function traits in ever-smoking group

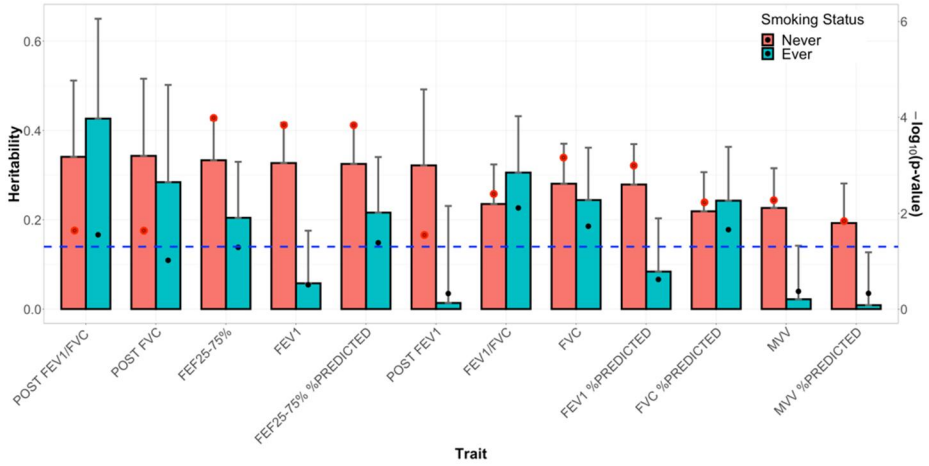
Traits	Never Smokers			Ever Smokers		
	Sample Size	Cross-sectional average Means (SD)	Annual change rates Means (SD)	Sample Size	Cross-sectional average Means (SD)	Annual change rates Means (SD)
FVC	3008	3.045 (0.642)	-0.034 (0.023)	2095	4.051 (0.661)	-0.039 (0.027)
FVC %PRED	3008	105.533 (12.978)	-0.158 (0.922)	2095	101.647 (12.066)	-0.287 (0.757)
FEV_1	3008	2.430 (0.520)	-0.036 (0.019)	2095	3.066 (0.592)	-0.046 (0.022)
FEV_1 %PRED	3008	114.953 (16.110)	-0.120 (1.086)	2095	105.474 (14.922)	-0.384 (0.852)
FEV_1/FVC	3008	79.815 (5.449)	-0.300 (0.430)	2095	75.540 (7.298)	-0.418 (0.406)
FEF25-75%	3009	2.526 (0.830)	-0.068 (0.047)	2095	2.712 (1.071)	-0.088 (0.051)
FEF25-75% %PRED	3009	99.458 (28.106)	-1.359 (1.936)	2095	88.909 (31.089)	-1.606 (1.657)
MVV	3004	92.488 (24.048)	-2.052 (1.340)	2095	119.465 (27.968)	-2.583 (1.707)
MVV %PRED	3004	93.298 (17.280)	-1.306 (1.423)	2095	93.209 (17.114)	-1.152 (1.454)
POST FVC	1524	3.181 (0.626)	-0.035 (0.018)	1182	4.144 (0.604)	-0.038 (0.023)
POST FEV_1	1525	2.628 (0.510)	-0.036 (0.014)	1182	3.296 (0.517)	-0.043 (0.018)
POST FEV_1FVC	1525	82.772 (4.862)	-0.236 (0.312)	1182	79.667 (5.949)	-0.291 (0.331)

Table 4.6 Heritability of SNP by environment interaction for 12 lung function traits

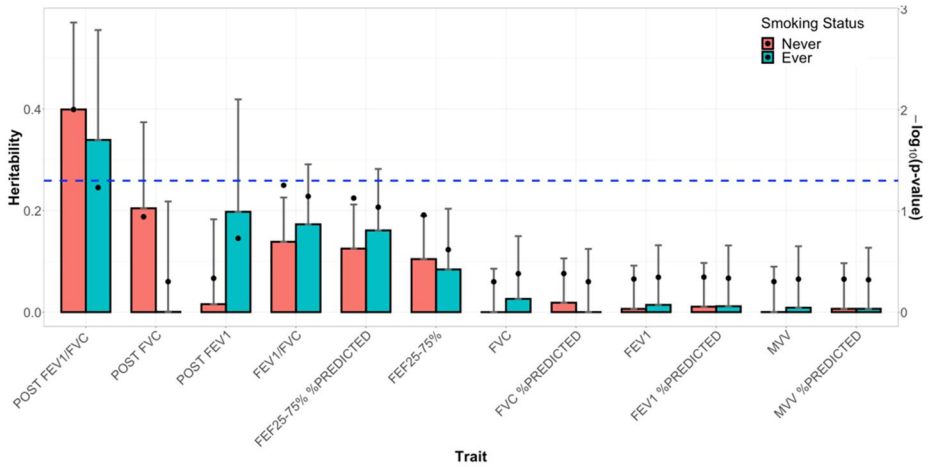
Traits	Cross-sectional average			Annual change rate		
	Vge/Vp	s.e.	P	Vge/Vp	s.e.	P
FVC	0.051622	0.097668	0.29476	-0.02459	0.101463	0.40977
FVC %PRED	-0.018434	0.094587	0.422	-0.143755	0.0973	0.09049
FEV1	-0.006611	0.097988	0.47352	-0.089591	0.090754	0.16206
FEV1 %PRED	-0.050392	0.09496	0.3002	-0.192572	0.088061	0.02091
FEV1/FVC	0.079199	0.101108	0.21902	0.138339	0.100909	0.079165
FEF25-75%	0.096592	0.102051	0.17629	0.116504	0.098932	0.10881
FEF25-75% %PRED	0.065509	0.101289	0.2634	0.090015	0.098497	0.1715
MVV	0.002662	0.098792	0.4894	-0.150229	0.094267	0.072297
MVV %PRED	0.025806	0.101696	0.4032	-0.153931	0.091747	0.061
POST FVC	-0.033354	0.181616	0.42711	-0.08897	0.189262	0.32768
POST FEV1	-0.074344	0.186164	0.35005	-0.047545	0.179968	0.39494
POST FEV1FVC	0.20578	0.190713	0.13673	0.402253	0.198761	0.021158

Figure 4.5 SNP heritability of 12 lung function traits in never and ever smokers. *(A)* SNP heritabilities of cross-sectional averages of 12 lung function traits in never and ever smokers. *(B)* SNP heritabilities of the annual change of 12 lung function traits in never and ever smokers. Error bars correspond to standard error values. The dot on the bar are P-values. Blue dash line indicates the 0.05 significant level. Red dot indicates significant findings at an FDR of 0.05.

(A)



(B)



4.4 Discussion

In the present study, we suggested two different SNP-based heritabilities of cross-sectional averages and annual change rate, and both for the 12 lung function traits were estimated. We found that heritabilities of cross-sectional averages were significant for all 12 lung function traits. For the heritability of annual change rates, post FEV_1/FVC, FEV_1/FVC and FEV_1 %predicted shows significant result, which reveals the significant effects of SNPs on lung function change rate. Then we performed stratification analysis by smoking status for both cross-sectional average and longitudinal change rate. And the heritabilites for SNP-by-smoking interaction also estimated. We found post FEV_1/FVC, FEV_1/FVC and FEV_1 %predicted show significant SNP-by-smoking interaction, inferring the amount of genetic variance would be affected by smoking conditions.

In the estimation of heritability of cross-sectional averages with all samples for the 12 traits, the estimated heritability ranges from about 9% (MVV) to 33% (post FVC). For the SNP heritability of annual change rates of 12 traits, were much lower than those of cross-sectional averages. The range approximately 1% (MVV) to 18% (post FEV_1/FVC). The heritability estimation of annual change rates of post FEV_1/FVC was the largest one among the 12 traits. And FEV_1/FVC and FEV_1 % predicted also displayed longitudinally significant heritability compared to other traits. These three traits reliably reflect the physiological state of the lungs and airways and both

are the predictor of morbidity and mortality in the general population and widely used to define choric obstructive lung disease (COPD).

Estimation of the heritability of lung function could be influenced by environmental factors. One of the important factors that influences lung function is smoking status. In this study, we found both cross-sectional averages and annual change rate showed different SNP heritability estimates between never and ever smoke groups. Among the 12 traits, FEF25-75% %predicted and FVC, the never smoker group have 10% and 4% respectively higher heritability estimation than those of ever smoker group. And FEV_1/FVC, post FEV_1/FVC and FVC % predicted, the estimated heritability in ever smoker group had slightly higher than those of never smoker group, and the difference were 7%, 9% and 3%, respectively. For the heritability estimates of annual change rates, there were no significant result under 0.05 FDR significant level. Collectively, these results indicate that smoking status does not affect much to the heritability of annual lung function decline rate in mid-aged population.

The result of genetic correlations of cross-sectional average and annual change for FEV1 % predicted, FEV1/FVC and post FEV1/FVC showed strong positive correlations. Cross-sectional mean and annual change rates consist of genetic and environment components, and positive correlations between SNP effects for cross-sectional mean and annual change rates indicates that subjects with higher genetic risk for cross-sectional means

of FEV_1/FVC, post FEV_1/FVC and FEV_1 % predicted tend to have higher genetic risk for their annual change.

One of the limitations in this study is that the limited sample sizes of some traits and in subgroup analysis, caused the large standard errors of heritability estimation in the analysis. To this problem, some previous study have GCTA (Robinson et al. 2017). Thus, to yield more stable results, more samples need to be collected in our future study.

In summary, we performed SNP heritability estimation for 12 lung function traits, by using two-stage method, which can estimate cross-sectional averages and annual change rates with Korean population based longitudinal data. We expect our work will help informing lung cancer etiology, and to discover most of the genetic variability influencing lung function related traits, large sample sizes and novel statistical approaches are required.

Chapter 5

Summary & Conclusions

Genetic effect of health-related phenotypic traits, especially lung function has been identified by multiple studies, but the progressive effect of SNPs on annual change and their interaction has remained unexplained. The main goal is to evaluate the effect of SNPs on annual change of prominent health-related phenotypic traits, and lung function related traits, by estimating SNP based heritabilities and genome-wide association analysis with longitudinal data.

In chapter 3, we analyzed sixteen phenotypic traits which is associated with major health indices, and observed every two years for 6,843 individuals with 10-year follow-up. SNP-based heritability of cross-sectional average and longitudinal changes were estimated by using the two-stage model. Cross-sectional average and longitudinal changes for each subject were considered responses to estimate SNP heritability. And genome-wide association study

(GWAS) was also performed to detecte the significant associtated SNPs. Each SNP heritability for the phenotypic averages of all sixteen traits through 6 periods (baseline and five follow-ups) were significant. Gradually, the forced vital capacity in one second (FEV1) reflected the only significant SNP heritability for longitudinal changes at a false discovery rate (FDR)-adjusted 0.05 significance level ($h^2_{snp} = 0.171$, FDR=0.0012). On estimating chromosomal heritability, chromosome 2 displayed the highest heritability upon periodic changes in FEV1. SNPs including rs2272402 and rs7209788 displayed a genome-wide significant association with longitudinal changes in FEV1 (P=1.22×10^{-8} for rs2272402 and P=3.36×10^{-7} for rs7209788). *De novo* variants including rs4922117 (near *LPL*, P=2.13×10^{-15}) of log-transformed high-density lipoprotein (HDL) ratios and rs2335418 (near *HMGCR,* P=3.2×10^{-9}) of low-density lipoprotein were detected on GWAS. Hence, significant genetic effects on longitudinal changes in FEV1 among the middle-aged general population and chromosome 2 account for most of the genetic variance.

In chapter 4, we analyzed twelve lung function traits, which observed every two years for 8,768 Korean adults aged 40-69 years during 14 years. Phenotypic average and annual change rate were calculated for each participant, and SNP heritabilities for both were estimated by GCTA. Furthermore, we also calculated the subgroup heritibility of smoking status. SNP heritabilities of the annual change rate of post FEV_1/FVC, FEV_1/FVC and FEV1 % predicted were significantly high (h^2_1=0.176, p-value=0.0099 for

post FEV_1/FVC; h_1^2=0.158, p-value=$4.91 \times 10^{-5}$for FEV_1/FVC; h_1^2=0.105, p-value=0.004 for FEV_1 %predicted). In subgroup analsysis, post FEV_1/FVC (h_1^2=0.399, p-value=0.009) were in never smokers significant high than in ever smokers. For the estimated heritability of SNP-by-smoking interaction $h_{G\times S}^2$, FEV_1/FVC, post FEV_1/FVC and FEV_1 % predicted have significant $h_{G1\times S}^2$.

In summary, the studies elucidate heritability estimates via a two-stage approach using a mixed model in GCTA and GWAS, which further determines longitudinal change effects independently with a linear model, followed by estimation of heritability using regression coefficients. This approach provides a reasonable and easy method to estimate heritability in longitudinal data and potentially assess both heritability of the phenotypic averages and annual changes through several periods. Essentially, the results show that significant SNP heritability is objectively confirmed for longitudinal changes in lung function decline including FEV1 in comparison with other health-related indices. Even in lung function specific analysis the significant genetic effect on lung function decline rate in FEV1 % predicted, FEV1/FVC and post FEV1/FVC were observed, and these traits also showed significant SNP-by-smoking interaction, inferring the amount of genetic variance would be affected by smoking conditions.

Bibliography

1995. "Standardization of Spirometry, 1994 Update. American Thoracic Society." *Am J Respir Crit Care Med* 152 (3):1107-36. doi: 10.1164/ajrccm.152.3.7663792.

2000. "Genetic epidemiologic studies on age-specified traits. NIA Aging and Genetic Epidemiology Working Group." *Am J Epidemiol* 152 (11):1003-8. doi: 10.1093/aje/152.11.1003.

Agusti, A., and J. C. Hogg. 2019. "Update on the Pathogenesis of Chronic Obstructive Pulmonary Disease." *N Engl J Med* 381 (13):1248-1256. doi: 10.1056/NEJMra1900475.

Bandyopadhyay, B., V. Chanda, and Y. Wang. 2017. "Finding the Sources of Missing Heritability within Rare Variants Through Simulation." *Bioinform Biol Insights* 11:1177932217735096. doi: 10.1177/1177932217735096.

Boyle, E. A., Y. I. Li, and J. K. Pritchard. 2017. "An Expanded View of Complex Traits: From Polygenic to Omnigenic." *Cell* 169 (7):1177-1186. doi: 10.1016/j.cell.2017.05.038.

Bulik-Sullivan, B. K., P. R. Loh, H. K. Finucane, S. Ripke, J. Yang, Consortium Schizophrenia Working Group of the Psychiatric Genomics, N. Patterson, M. J. Daly, A. L. Price, and B. M. Neale. 2015. "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies." *Nat Genet* 47 (3):291-5. doi: 10.1038/ng.3211.

Buniello, A., J. A. L. MacArthur, M. Cerezo, L. W. Harris, J. Hayhurst, C. Malangone, A. McMahon, J. Morales, E. Mountjoy, E. Sollis, D. Suveges, O. Vrousgou, P. L. Whetzel, R. Amode, J. A. Guillen, H. S. Riat, S. J. Trevanion, P. Hall, H. Junkins, P. Flicek, T. Burdett, L. A. Hindorff, F. Cunningham, and H. Parkinson. 2019. "The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019." *Nucleic Acids Res* 47 (D1):D1005-D1012. doi: 10.1093/nar/gky1120.

Bush, W. S., and J. H. Moore. 2012. "Chapter 11: Genome-wide association studies." *PLoS Comput Biol* 8 (12):e1002822. doi: 10.1371/journal.pcbi.1002822.

Cho, Y. S., M. J. Go, Y. J. Kim, J. Y. Heo, J. H. Oh, H. J. Ban, D. Yoon, M. H. Lee, D. J. Kim, M. Park, S. H. Cha, J. W. Kim, B. G. Han, H. Min, Y. Ahn, M. S. Park, H. R. Han, H. Y. Jang, E. Y. Cho, J. E. Lee, N. H. Cho, C. Shin, T. Park, J. W. Park, J. K. Lee, L. Cardon, G. Clarke, M. I. McCarthy, J. Y. Lee, J. K. Lee, B. Oh, and H. L. Kim. 2009. "A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits." *Nat Genet* 41 (5):527-34. doi: 10.1038/ng.357.

Coram, M. A., Q. Duan, T. J. Hoffmann, T. Thornton, J. W. Knowles, N. A. Johnson, H. M. Ochs-Balcom, T. A. Donlon, L. W. Martin, C. B.

Eaton, J. G. Robinson, N. J. Risch, X. Zhu, C. Kooperberg, Y. Li, A. P. Reiner, and H. Tang. 2013. "Genome-wide characterization of shared and distinct genetic components that influence blood lipid levels in ethnically diverse human populations." *Am J Hum Genet* 92 (6):904-16. doi: 10.1016/j.ajhg.2013.04.025.

Crossett, A., A. B. Lee, L. Klei, B. Devlin, and K. Roeder. 2013. "Refining Genetically Inferred Relationships Using Treelet Covariance Smoothing." *Ann Appl Stat* 7 (2):669-690. doi: 10.1214/12-AOAS598.

Evans, L. M., R. Tahmasbi, S. I. Vrieze, G. R. Abecasis, S. Das, S. Gazal, D. W. Bjelland, T. R. de Candia, M. E. Goddard, B. M. Neale, J. Yang, P. M. Visscher, M. C. Keller, and Haplotype Reference Consortium. 2018. "Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits." *Nature Genetics* 50 (5):737-+. doi: 10.1038/s41588-018-0108-x.

Finucane, H. K., B. Bulik-Sullivan, A. Gusev, G. Trynka, Y. Reshef, P. R. Loh, V. Anttila, H. Xu, C. Zang, K. Farh, S. Ripke, F. R. Day, Consortium ReproGen, Consortium Schizophrenia Working Group of the Psychiatric Genomics, Raci Consortium, S. Purcell, E. Stahl, S. Lindstrom, J. R. Perry, Y. Okada, S. Raychaudhuri, M. J. Daly, N. Patterson, B. M. Neale, and A. L. Price. 2015. "Partitioning heritability by functional annotation using genome-wide association summary statistics." *Nat Genet* 47 (11):1228-35. doi: 10.1038/ng.3404.

Friedlander, Y., M. A. Austin, B. Newman, K. Edwards, E. I. Mayer-Davis, and M. C. King. 1997. "Heritability of longitudinal changes in coronary-heart-disease risk factors in women twins." *Am J Hum Genet* 60 (6):1502-12. doi: 10.1086/515462.

Garcia, T. P., and K. Marder. 2017. "Statistical Approaches to Longitudinal Data Analysis in Neurodegenerative Diseases: Huntington's Disease as a Model." *Curr Neurol Neurosci Rep* 17 (2):14. doi: 10.1007/s11910-017-0723-4.

Goldstein, D. B. 2009. "Common genetic variation and human traits." *N Engl J Med* 360 (17):1696-8. doi: 10.1056/NEJMp0806284.

Gottlieb, D. J., J. B. Wilk, M. Harmon, J. C. Evans, O. Joost, D. Levy, G. T. O'Connor, and R. H. Myers. 2001. "Heritability of longitudinal change in lung function. The Framingham study." *Am J Respir Crit Care Med* 164 (9):1655-9. doi: 10.1164/ajrccm.164.9.2010122.

Hansel, N. N., I. Ruczinski, N. Rafaels, D. D. Sin, D. Daley, A. Malinina, L. Huang, A. Sandford, T. Murray, Y. Kim, C. Vergara, S. R. Heckbert, B. M. Psaty, G. Li, W. M. Elliott, F. Aminuddin, J. Dupuis, G. T. O'Connor, K. Doheny, A. F. Scott, H. M. Boezen, D. S. Postma, J. Smolonska, P. Zanen, F. A. Mohamed Hoesein, H. J. de Koning, R. G. Crystal, T. Tanaka, L. Ferrucci, E. Silverman, E. Wan, J. Vestbo, D. A. Lomas, J. Connett, R. A. Wise, E. R. Neptune, R. A. Mathias, P. D. Pare, T. H. Beaty, and K. C. Barnes. 2013. "Genome-wide study

identifies two loci associated with lung function decline in mild to moderate COPD." *Hum Genet* 132 (1):79-90. doi: 10.1007/s00439-012-1219-6.

Hoffmann, T. J., E. Theusch, T. Haldar, D. K. Ranatunga, E. Jorgenson, M. W. Medina, M. N. Kvale, P. Y. Kwok, C. Schaefer, R. M. Krauss, C. Iribarren, and N. Risch. 2018. "A large electronic-health-record-based genome-wide study of serum lipids." *Nat Genet* 50 (3):401-413. doi: 10.1038/s41588-018-0064-5.

Imboden, M., E. Bouzigon, I. Curjuric, A. Ramasamy, A. Kumar, D. B. Hancock, J. B. Wilk, J. M. Vonk, G. A. Thun, V. Siroux, R. Nadif, F. Monier, J. R. Gonzalez, M. Wjst, J. Heinrich, L. R. Loehr, N. Franceschini, K. E. North, J. Altmuller, G. H. Koppelman, S. Guerra, F. Kronenberg, M. Lathrop, M. F. Moffatt, G. T. O'Connor, D. P. Strachan, D. S. Postma, S. J. London, C. Schindler, M. Kogevinas, F. Kauffmann, D. L. Jarvis, F. Demenais, and N. M. Probst-Hensch. 2012. "Genome-wide association study of lung function decline in adults with and without asthma." *J Allergy Clin Immunol* 129 (5):1218-28. doi: 10.1016/j.jaci.2012.01.074.

John, C., M. Soler Artigas, J. Hui, S. F. Nielsen, N. Rafaels, P. D. Pare, N. N. Hansel, N. Shrine, I. Kilty, A. Malarstig, S. A. Jelinsky, S. Vedel-Krogh, K. Barnes, I. P. Hall, J. Beilby, A. W. Musk, B. G. Nordestgaard, A. James, L. V. Wain, and M. D. Tobin. 2017. "Genetic variants affecting cross-sectional lung function in adults show little or no effect on longitudinal lung function decline." *Thorax* 72 (5):400-408. doi: 10.1136/thoraxjnl-2016-208448.

Kim, S. J., J. Lee, Y. S. Park, C. H. Lee, H. I. Yoon, S. M. Lee, J. J. Yim, Y. W. Kim, S. K. Han, and C. G. Yoo. 2016. "Age-related annual decline of lung function in patients with COPD." *Int J Chron Obstruct Pulmon Dis* 11:51-60. doi: 10.2147/COPD.S95028.

Kim, S., H. Kim, N. Cho, S. K. Lee, B. G. Han, J. W. Sull, S. H. Jee, and C. Shin. 2015. "Identification of FAM13A gene associated with the ratio of FEV1 to FVC in Korean population by genome-wide association studies including gene-environment interactions." *J Hum Genet* 60 (3):139-45. doi: 10.1038/jhg.2014.118.

Kim, Y., B. G. Han, and G. E. S. group Ko. 2017. "Cohort Profile: The Korean Genome and Epidemiology Study (KoGES) Consortium." *Int J Epidemiol* 46 (4):1350. doi: 10.1093/ije/dyx105.

Kim, Y., Y. Lee, S. Lee, N. H. Kim, J. Lim, Y. J. Kim, J. H. Oh, H. Min, M. Lee, H. J. Seo, S. H. Lee, J. Sung, N. H. Cho, B. J. Kim, B. G. Han, R. C. Elston, S. Won, and J. Lee. 2015. "On the Estimation of Heritability with Family-Based and Population-Based Samples." *Biomed Res Int* 2015:671349. doi: 10.1155/2015/671349.

Kumar, S. K., M. W. Feldman, D. H. Rehkopf, and S. Tuljapurkar. 2016. "Correction for Krishna Kumar et al., Limitations of GCTA as a solution to the missing heritability problem." *Proc Natl Acad Sci U S A* 113 (6):E813. doi: 10.1073/pnas.1600634113.

Loh, P. R., G. Bhatia, A. Gusev, H. K. Finucane, B. K. Bulik-Sullivan, S. J. Pollack, Consortium Schizophrenia Working Group of Psychiatric Genomics, T. R. de Candia, S. H. Lee, N. R. Wray, K. S. Kendler, M. C. O'Donovan, B. M. Neale, N. Patterson, and A. L. Price. 2015. "Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis." *Nat Genet* 47 (12):1385-92. doi: 10.1038/ng.3431.

Loth, D. W., M. Soler Artigas, S. A. Gharib, L. V. Wain, N. Franceschini, B. Koch, T. D. Pottinger, A. V. Smith, Q. Duan, C. Oldmeadow, M. K. Lee, D. P. Strachan, A. L. James, J. E. Huffman, V. Vitart, A. Ramasamy, N. J. Wareham, J. Kaprio, X. Q. Wang, H. Trochet, M. Kahonen, C. Flexeder, E. Albrecht, L. M. Lopez, K. de Jong, B. Thyagarajan, A. C. Alves, S. Enroth, E. Omenaas, P. K. Joshi, T. Fall, A. Vinuela, L. J. Launer, L. R. Loehr, M. Fornage, G. Li, J. B. Wilk, W. Tang, A. Manichaikul, L. Lahousse, T. B. Harris, K. E. North, A. R. Rudnicka, J. Hui, X. Gu, T. Lumley, A. F. Wright, N. D. Hastie, S. Campbell, R. Kumar, I. Pin, R. A. Scott, K. H. Pietilainen, I. Surakka, Y. Liu, E. G. Holliday, H. Schulz, J. Heinrich, G. Davies, J. M. Vonk, M. Wojczynski, A. Pouta, A. Johansson, S. H. Wild, E. Ingelsson, F. Rivadeneira, H. Volzke, P. G. Hysi, G. Eiriksdottir, A. C. Morrison, J. I. Rotter, W. Gao, D. S. Postma, W. B. White, S. S. Rich, A. Hofman, T. Aspelund, D. Couper, L. J. Smith, B. M. Psaty, K. Lohman, E. G. Burchard, A. G. Uitterlinden, M. Garcia, B. R. Joubert, W. L. McArdle, A. B. Musk, N. Hansel, S. R. Heckbert, L. Zgaga, J. B. van Meurs, P. Navarro, I. Rudan, Y. M. Oh, S. Redline, D. L. Jarvis, J. H. Zhao, T. Rantanen, G. T. O'Connor, S. Ripatti, R. J. Scott, S. Karrasch, H. Grallert, N. C. Gaddis, J. M. Starr, C. Wijmenga, R. L. Minster, D. J. Lederer, J. Pekkanen, U. Gyllensten, H. Campbell, A. P. Morris, S. Glaser, C. J. Hammond, K. M. Burkart, J. Beilby, S. B. Kritchevsky, V. Gudnason, D. B. Hancock, O. D. Williams, O. Polasek, T. Zemunik, I. Kolcic, M. F. Petrini, M. Wjst, W. J. Kim, D. J. Porteous, G. Scotland, B. H. Smith, A. Viljanen, M. Heliovaara, J. R. Attia, I. Sayers, R. Hampel, C. Gieger, I. J. Deary, H. M. Boezen, A. Newman, M. R. Jarvelin, J. F. Wilson, L. Lind, B. H. Stricker, A. Teumer, T. D. Spector, E. Melen, M. J. Peters, L. A. Lange, R. G. Barr, K. R. Bracke, F. M. Verhamme, J. Sung, P. S. Hiemstra, P. A. Cassano, A. Sood, C. Hayward, J. Dupuis, I. P. Hall, G. G. Brusselle, M. D. Tobin, and S. J. London. 2014. "Genome-wide association analysis identifies six new loci associated with forced vital capacity." *Nat Genet* 46 (7):669-77. doi: 10.1038/ng.3011.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. Mackay, S. A. McCarroll, and P. M. Visscher. 2009. "Finding the missing heritability

of complex diseases." *Nature* 461 (7265):747-53. doi: 10.1038/nature08494.

McClellan, J., and M. C. King. 2010. "Genetic heterogeneity in human disease." *Cell* 141 (2):210-7. doi: 10.1016/j.cell.2010.03.032.

Miller, M. R., J. Hankinson, V. Brusasco, F. Burgos, R. Casaburi, A. Coates, R. Crapo, P. Enright, C. P. van der Grinten, P. Gustafsson, R. Jensen, D. C. Johnson, N. MacIntyre, R. McKay, D. Navajas, O. F. Pedersen, R. Pellegrino, G. Viegi, J. Wanger, and Ats Ers Task Force. 2005. "Standardisation of spirometry." *Eur Respir J* 26 (2):319-38. doi: 10.1183/09031936.05.00034805.

Ni, G., G. Moser, Consortium Schizophrenia Working Group of the Psychiatric Genomics, N. R. Wray, and S. H. Lee. 2018. "Estimation of Genetic Correlation via Linkage Disequilibrium Score Regression and Genomic Restricted Maximum Likelihood." *Am J Hum Genet* 102 (6):1185-1194. doi: 10.1016/j.ajhg.2018.03.021.

Nishimura, M., H. Makita, K. Nagai, S. Konno, Y. Nasuhara, M. Hasegawa, K. Shimizu, T. Betsuyaku, Y. M. Ito, S. Fuke, T. Igarashi, Y. Akiyama, S. Ogura, and Copd Cohort Study Investigators Hokkaido. 2012. "Annual change in pulmonary function and clinical phenotype in chronic obstructive pulmonary disease." *Am J Respir Crit Care Med* 185 (1):44-52. doi: 10.1164/rccm.201106-0992OC.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham. 2007. "PLINK: a tool set for whole-genome association and population-based linkage analyses." *Am J Hum Genet* 81 (3):559-75. doi: 10.1086/519795.

Robinson, M. R., G. English, G. Moser, L. R. Lloyd-Jones, M. A. Triplett, Z. Zhu, I. M. Nolte, J. V. van Vliet-Ostaptchouk, H. Snieder, Study LifeLines Cohort, T. Esko, L. Milani, R. Magi, A. Metspalu, P. K. E. Magnusson, N. L. Pedersen, E. Ingelsson, M. Johannesson, J. Yang, D. Cesarini, and P. M. Visscher. 2017. "Genotype-covariate interaction effects and the heritability of adult body mass index." *Nat Genet* 49 (8):1174-1181. doi: 10.1038/ng.3912.

Sakornsakolpat, P., M. McCormack, P. Bakke, A. Gulsvik, B. J. Make, J. D. Crapo, M. H. Cho, and E. K. Silverman. 2019. "Genome-Wide Association Analysis of Single-Breath DlCO." *Am J Respir Cell Mol Biol* 60 (5):523-531. doi: 10.1165/rcmb.2018-0384OC.

Sandoval-Motta, S., M. Aldana, E. Martinez-Romero, and A. Frank. 2017. "The Human Microbiome and the Missing Heritability Problem." *Front Genet* 8:80. doi: 10.3389/fgene.2017.00080.

Shin, C., J. Kim, J. Kim, S. Lee, J. Shim, K. In, K. Kang, S. Yoo, N. Cho, K. Kimm, and S. Joo. 2005. "Association of habitual snoring with glucose and insulin metabolism in nonobese Korean adult men." *Am J Respir Crit Care Med* 171 (3):287-91. doi: 10.1164/rccm.200407-906OC.

Soler Artigas, M., D. W. Loth, L. V. Wain, S. A. Gharib, M. Obeidat, W. Tang, G. Zhai, J. H. Zhao, A. V. Smith, J. E. Huffman, E. Albrecht, C. M. Jackson, D. M. Evans, G. Cadby, M. Fornage, A. Manichaikul, L. M. Lopez, T. Johnson, M. C. Aldrich, T. Aspelund, I. Barroso, H. Campbell, P. A. Cassano, D. J. Couper, G. Eiriksdottir, N. Franceschini, M. Garcia, C. Gieger, G. K. Gislason, I. Grkovic, C. J. Hammond, D. B. Hancock, T. B. Harris, A. Ramasamy, S. R. Heckbert, M. Heliovaara, G. Homuth, P. G. Hysi, A. L. James, S. Jankovic, B. R. Joubert, S. Karrasch, N. Klopp, B. Koch, S. B. Kritchevsky, L. J. Launer, Y. Liu, L. R. Loehr, K. Lohman, R. J. Loos, T. Lumley, K. A. Al Balushi, W. Q. Ang, R. G. Barr, J. Beilby, J. D. Blakey, M. Boban, V. Boraska, J. Brisman, J. R. Britton, G. G. Brusselle, C. Cooper, I. Curjuric, S. Dahgam, I. J. Deary, S. Ebrahim, M. Eijgelsheim, C. Francks, D. Gaysina, R. Granell, X. Gu, J. L. Hankinson, R. Hardy, S. E. Harris, J. Henderson, A. Henry, A. D. Hingorani, A. Hofman, P. G. Holt, J. Hui, M. L. Hunter, M. Imboden, K. A. Jameson, S. M. Kerr, I. Kolcic, F. Kronenberg, J. Z. Liu, J. Marchini, T. McKeever, A. D. Morris, A. C. Olin, D. J. Porteous, D. S. Postma, S. S. Rich, S. M. Ring, F. Rivadeneira, T. Rochat, A. A. Sayer, I. Sayers, P. D. Sly, G. D. Smith, A. Sood, J. M. Starr, A. G. Uitterlinden, J. M. Vonk, S. G. Wannamethee, P. H. Whincup, C. Wijmenga, O. D. Williams, A. Wong, M. Mangino, K. D. Marciante, W. L. McArdle, B. Meibohm, A. C. Morrison, K. E. North, E. Omenaas, L. J. Palmer, K. H. Pietilainen, I. Pin, O. Pola Sbreve Ek, A. Pouta, B. M. Psaty, A. L. Hartikainen, T. Rantanen, S. Ripatti, J. I. Rotter, I. Rudan, A. R. Rudnicka, H. Schulz, S. Y. Shin, T. D. Spector, I. Surakka, V. Vitart, H. Volzke, N. J. Wareham, N. M. Warrington, H. E. Wichmann, S. H. Wild, J. B. Wilk, M. Wjst, A. F. Wright, L. Zgaga, T. Zemunik, C. E. Pennell, F. Nyberg, D. Kuh, J. W. Holloway, H. M. Boezen, D. A. Lawlor, R. W. Morris, N. Probst-Hensch, Consortium International Lung Cancer, Giant consortium, J. Kaprio, J. F. Wilson, C. Hayward, M. Kahonen, J. Heinrich, A. W. Musk, D. L. Jarvis, S. Glaser, M. R. Jarvelin, B. H. Ch Stricker, P. Elliott, G. T. O'Connor, D. P. Strachan, S. J. London, I. P. Hall, V. Gudnason, and M. D. Tobin. 2011. "Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function." *Nat Genet* 43 (11):1082-90. doi: 10.1038/ng.941.

Song, Y. E., S. Lee, K. Park, R. C. Elston, H. J. Yang, and S. Won. 2018. "ONETOOL for the analysis of family-based big data." *Bioinformatics* 34 (16):2851-2853. doi: 10.1093/bioinformatics/bty180.

Speed, D., N. Cai, Ucleb Consortium, M. R. Johnson, S. Nejentsev, and D. J. Balding. 2017. "Reevaluation of SNP heritability in complex human traits." *Nat Genet* 49 (7):986-992. doi: 10.1038/ng.3865.

Speed, D., G. Hemani, M. R. Johnson, and D. J. Balding. 2012. "Improved heritability estimation from genome-wide SNPs." *Am J Hum Genet* 91 (6):1011-21. doi: 10.1016/j.ajhg.2012.10.010.

Spracklen, C. N., P. Chen, Y. J. Kim, X. Wang, H. Cai, S. Li, J. Long, Y. Wu, Y. X. Wang, F. Takeuchi, J. Y. Wu, K. J. Jung, C. Hu, K. Akiyama, Y. Zhang, S. Moon, T. A. Johnson, H. Li, R. Dorajoo, M. He, M. E. Cannon, T. S. Roman, E. Salfati, K. H. Lin, X. Guo, W. H. H. Sheu, D. Absher, L. S. Adair, T. L. Assimes, T. Aung, Q. Cai, L. C. Chang, C. H. Chen, L. H. Chien, L. M. Chuang, S. C. Chuang, S. Du, Q. Fan, C. S. J. Fann, A. B. Feranil, Y. Friedlander, P. Gordon-Larsen, D. Gu, L. Gui, Z. Guo, C. K. Heng, J. Hixson, X. Hou, C. A. Hsiung, Y. Hu, M. Y. Hwang, C. M. Hwu, M. Isono, J. J. Juang, C. C. Khor, Y. K. Kim, W. P. Koh, M. Kubo, I. T. Lee, S. J. Lee, W. J. Lee, K. W. Liang, B. Lim, S. H. Lim, J. Liu, T. Nabika, W. H. Pan, H. Peng, T. Quertermous, C. Sabanayagam, K. Sandow, J. Shi, L. Sun, P. C. Tan, S. P. Tan, K. D. Taylor, Y. Y. Teo, S. A. Toh, T. Tsunoda, R. M. van Dam, A. Wang, F. Wang, J. Wang, W. B. Wei, Y. B. Xiang, J. Yao, J. M. Yuan, R. Zhang, W. Zhao, Y. I. Chen, S. S. Rich, J. I. Rotter, T. D. Wang, T. Wu, X. Lin, B. G. Han, T. Tanaka, Y. S. Cho, T. Katsuya, W. Jia, S. H. Jee, Y. T. Chen, N. Kato, J. B. Jonas, C. Y. Cheng, X. O. Shu, J. He, W. Zheng, T. Y. Wong, W. Huang, B. J. Kim, E. S. Tai, K. L. Mohlke, and X. Sim. 2017. "Association analyses of East Asian individuals and trans-ancestry analyses with European individuals reveal new loci associated with cholesterol and triglyceride levels." *Hum Mol Genet* 26 (9):1770-1784. doi: 10.1093/hmg/ddx062.

Sung, Y. J., J. Simino, R. Kume, J. Basson, K. Schwander, and D. C. Rao. 2014. "Comparison of two methods for analysis of gene-environment interactions in longitudinal family data: the Framingham heart study." *Front Genet* 5:9. doi: 10.3389/fgene.2014.00009.

Tam, V., N. Patel, M. Turcotte, Y. Bosse, G. Pare, and D. Meyre. 2019. "Benefits and limitations of genome-wide association studies." *Nat Rev Genet* 20 (8):467-484. doi: 10.1038/s41576-019-0127-1.

Tang, W., M. Kowgier, D. W. Loth, M. Soler Artigas, B. R. Joubert, E. Hodge, S. A. Gharib, A. V. Smith, I. Ruczinski, V. Gudnason, R. A. Mathias, T. B. Harris, N. N. Hansel, L. J. Launer, K. C. Barnes, J. G. Hansen, E. Albrecht, M. C. Aldrich, M. Allerhand, R. G. Barr, G. G. Brusselle, D. J. Couper, I. Curjuric, G. Davies, I. J. Deary, J. Dupuis, T. Fall, M. Foy, N. Franceschini, W. Gao, S. Glaser, X. Gu, D. B. Hancock, J. Heinrich, A. Hofman, M. Imboden, E. Ingelsson, A. James, S. Karrasch, B. Koch, S. B. Kritchevsky, A. Kumar, L. Lahousse, G. Li, L. Lind, C. Lindgren, Y. Liu, K. Lohman, T. Lumley, W. L. McArdle, B. Meibohm, A. P. Morris, A. C. Morrison, B. Musk, K. E. North, L. J. Palmer, N. M. Probst-Hensch, B. M. Psaty, F. Rivadeneira, J. I. Rotter, H. Schulz, L. J. Smith, A. Sood, J. M. Starr, D. P. Strachan, A. Teumer, A. G. Uitterlinden, H. Volzke, A. Voorman, L. V. Wain, M. T. Wells, J. B. Wilk, O. D. Williams, S. R. Heckbert, B. H. Stricker, S. J. London,

M. Fornage, M. D. Tobin, G. T. O'Connor, I. P. Hall, and P. A. Cassano. 2014. "Large-scale genome-wide association studies and meta-analyses of longitudinal change in adult lung function." *PLoS One* 9 (7):e100776. doi: 10.1371/journal.pone.0100776.

van de Pol, M., and S. Verhulst. 2006. "Age-dependent traits: a new statistical model to separate within- and between-individual effects." *Am Nat* 167 (5):766-73. doi: 10.1086/503331.

van Dongen, J., G. Willemsen, W. M. Chen, E. J. de Geus, and D. I. Boomsma. 2013. "Heritability of metabolic syndrome traits in a large population-based sample." *J Lipid Res* 54 (10):2914-23. doi: 10.1194/jlr.P041673.

Visscher, P. M., W. G. Hill, and N. R. Wray. 2008. "Heritability in the genomics era--concepts and misconceptions." *Nat Rev Genet* 9 (4):255-66. doi: 10.1038/nrg2322.

Visscher, P. M., S. E. Medland, M. A. Ferreira, K. I. Morley, G. Zhu, B. K. Cornes, G. W. Montgomery, and N. G. Martin. 2006. "Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings." *PLoS Genet* 2 (3):e41. doi: 10.1371/journal.pgen.0020041.

Weissbrod, O., J. Flint, and S. Rosset. 2018. "Estimating SNP-Based Heritability and Genetic Correlation in Case-Control Studies Directly and with Summary Statistics." *Am J Hum Genet* 103 (1):89-99. doi: 10.1016/j.ajhg.2018.06.002.

Wilk, J. B., T. H. Chen, D. J. Gottlieb, R. E. Walter, M. W. Nagle, B. J. Brandler, R. H. Myers, I. B. Borecki, E. K. Silverman, S. T. Weiss, and G. T. O'Connor. 2009. "A genome-wide association study of pulmonary function measures in the Framingham Heart Study." *PLoS Genet* 5 (3):e1000429. doi: 10.1371/journal.pgen.1000429.

Wu, Z., Y. Hu, and P. E. Melton. 2014. "Longitudinal data analysis for genetic studies in the whole-genome sequencing era." *Genet Epidemiol* 38 Suppl 1:S74-80. doi: 10.1002/gepi.21829.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher. 2010. "Common SNPs explain a large proportion of the heritability for human height." *Nat Genet* 42 (7):565-9. doi: 10.1038/ng.608.

Yang, J., S. H. Lee, M. E. Goddard, and P. M. Visscher. 2011. "GCTA: a tool for genome-wide complex trait analysis." *Am J Hum Genet* 88 (1):76-82. doi: 10.1016/j.ajhg.2010.11.011.

Yang, J., T. Lee, J. Kim, M. C. Cho, B. G. Han, J. Y. Lee, H. J. Lee, S. Cho, and H. Kim. 2013. "Ubiquitous polygenicity of human complex traits: genome-wide analysis of 49 traits in Koreans." *PLoS Genet* 9 (3):e1003355. doi: 10.1371/journal.pgen.1003355.

Yang, J., T. A. Manolio, L. R. Pasquale, E. Boerwinkle, N. Caporaso, J. M. Cunningham, M. de Andrade, B. Feenstra, E. Feingold, M. G. Hayes, W. G. Hill, M. T. Landi, A. Alonso, G. Lettre, P. Lin, H. Ling, W.

Lowe, R. A. Mathias, M. Melbye, E. Pugh, M. C. Cornelis, B. S. Weir, M. E. Goddard, and P. M. Visscher. 2011. "Genome partitioning of genetic variation for complex traits using common SNPs." *Nat Genet* 43 (6):519-25. doi: 10.1038/ng.823.

Yang, J., J. Zeng, M. E. Goddard, N. R. Wray, and P. M. Visscher. 2017. "Concepts, estimation and interpretation of SNP-based heritability." *Nat Genet* 49 (9):1304-1310. doi: 10.1038/ng.3941.

Young, R. P., R. Hopkins, and T. E. Eaton. 2007. "Forced expiratory volume in one second: not just a lung function test but a marker of premature death from all causes." *Eur Respir J* 30 (4):616-22. doi: 10.1183/09031936.00021707.

Zaitlen, N., P. Kraft, N. Patterson, B. Pasaniuc, G. Bhatia, S. Pollack, and A. L. Price. 2013. "Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits." *PLoS Genet* 9 (5):e1003520. doi: 10.1371/journal.pgen.1003520.

Zeger, S. L., and K. Y. Liang. 1992. "An overview of methods for the analysis of longitudinal data." *Stat Med* 11 (14-15):1825-39. doi: 10.1002/sim.4780111406.

Zhou, J. J., M. H. Cho, P. J. Castaldi, C. P. Hersh, E. K. Silverman, and N. M. Laird. 2013. "Heritability of chronic obstructive pulmonary disease and related phenotypes in smokers." *Am J Respir Crit Care Med* 188 (8):941-7. doi: 10.1164/rccm.201302-0263OC.

초 록

유전체 정보(SNP)의 대량생산이 가능해지며 질환의 원인을 규명하고자 질환 또는 위험요인에 대한 유전체 정보를 기반을 둔 전장 유전체 연관성 분석(GWAS)은 지속해서 활발히 진행됐고, 지역 또는 인종에 따라 다양하게 나타나고 있어 국내에서도 많은 결과가 발표되고 있다. 그러나, 실제 질환과 연관 있다고 보고된 SNP 들의 설명력은 높지 않았다. 이러한 설명되지 않은 유전적 경향성(missing heritability)에 대한 문제점을 보완하기 위한 유전율 추정 방법들이 제안되고 있고, 최근에는 인구집단 기반(population-based)을 둔 유전율 추정이 많이 진행되고 있다. 현재까지 대부분 population-based 유전율 추정은 단면연구(cross-sectional study)에 집중되어 연구가 진행됐으나 반복측정자료(longitudinal data)를 이용한 유전율 추정 및 유전자-환경, 유전자-시간의 상호작용으로 인한 유전율 추정 분석은 많이 진행되지 않았다.

본 논문에서는 한국인 질병 관련 임상역학의 종단자료 및 유전체 자료를 기반 표현형에 대한 상염색체 공통변이(common variant) 유전적 영향의 추정에 목적을 두어 16 가지의 표현형에 대하여 유전율 추정 및 GWAS 를 진행하였고, 추가로 12 가지의 폐기능관련 표현형에 대하여 유전율 추정을 진행 하였다. 또 표현형과 유전변이의 상호작용으로 인한 유전율에 대한 영향을 추정하였고 종단자료 특성상 분석이 어려운 것을 해결하기 위하여 two-stage 방법론을 제안하여 특정 표현형이 시간으로 인한 변화에 연관된 유전자들을 성공적으로 발굴하였다. 본 연구는

다량의 질병 관련 표현형 종단자료의 분석에 활용될 수 있을 것으로 기대된다.

주요어: 전장유전체연관성분석, 유전율분석, 종단자료 분석, GREML

학 번: 2014-31030